Sparse Superposition Codes are Fast and Reliable at Rates Approaching Capacity with Gaussian Noise

Andrew R. Barron, *Senior Member, IEEE,* and Antony Joseph, *Student Member, IEEE*

For upcoming submission to *IEEE Transactions on Information Theory*, June 10, 2011.

Abstract— **For the additive white Gaussian noise channel with average codeword power constraint, sparse superposition codes provide codewords as linear combinations of subsets of vectors from a given dictionary. Both encoding and decoding are computationally feasible. An adaptive successive decoder is developed, with which communication is shown to be reliable with error probability exponentially small for all rates below the Shannon capacity.**

I. Introduction

The additive white Gaussian noise channel is basic to Shannon theory and underlies practical communication models. *Sparse superposition codes* for this channel are analyzed. Theory and practice are linked by devising fast coding and decoding algorithms and by showing sparse superposition codes from moderate size dictionaries with these algorithms achieve nearly exponentially small error probability for any communication rate below the Shannon capacity. A companion paper [10] gives reliability bounds for optimal least squares (minimum distance) decoding, whereas the present work provides comparable bounds for fast decoding algorithms. The strategy and its analysis merges modern perspectives on information theory and term selection in statistical regression.

In the familiar communication set-up, an encoder is to map input bit strings $u = (u_1, u_2, \ldots, u_K)$ of length K into codewords which are length n strings of real numbers $c_1, c_2, \ldots, c_n$ of norm expressed via the power $(1/n)\sum_{i=1}^n c_i^2$. The average of the power across the 2^K codewords is to be not more than P. The channel adds independent $N(0, \sigma^2)$ noise to the codeword yielding a received length n string Y. A decoder is to map it into an estimate $\hat{u}$ desired to be a correct decoding of u. Block error is the event $\hat{u} \neq u$. When the input string is partitioned into sections, the section error rate is the fraction of sections not correctly decoded. The reliability requirement is that, with sufficiently large n, the section error rate is small with high probability or, more stringently, the block error probability is small, averaged over input strings u as well as the distribution of Y. The communication rate $R = K/n$ is the ratio of the number of message bits to the number of uses of the channel required to communicate them.

The supremum of reliable rates of communication is the channel capacity $\mathcal{C} = (1/2)\log_2(1 + P/\sigma^2)$, by traditional information theory as in [63], [23]. This problem is also of interest in mathematics because of relationship to versions of the sphere packing problem as described in [20]. For practical coding the challenge is to achieve arbitrary rates below the capacity, while guaranteeing reliable decoding in manageable computation time.

In a communication system operating at rate R, the input bit strings arise from input sequences $u_1, u_2, \ldots$ cut into successive K bit strings, each of which is encoded and sent, leading to a succession of received length n strings Y. The reliability aim that the block error probability be exponentially small is such that errors are unlikely over long time spans. The computational aim is that coding and decoding computations proceed on the fly, rapidly, with the decoder having not too many pipelined computational units, so that there is only moderate delay in the system.

The development here is specific to the discrete-time channel for which $Y_i = c_i + \varepsilon_i$ for $i = 1, 2, \ldots, n$ with real-valued inputs and outputs and with independent Gaussian noise. Standard communication models, even in continuous-time, have been reduced to this discrete-time white Gaussian noise setting, or to parallel uses of such, when there is a frequency band constraint for signal modulation and when there is a specified spectrum of noise over that frequency band, as in [39], [36]. Solution to the coding problem, when married to appropriate modulation schemes, is regarded as relevant to myriad settings involving transmission over wires or cables for internet, television, or telephone communications or in wireless radio, TV, phone, satellite or other space communications.

Previous standard approaches, as discussed in [36], entail a decomposition into separate problems of modulation, of shaping of a signal constellation, and of coding. As they point out, though there are practical schemes with empirically good performance (including LDPC and Turbo codes), theory for practical schemes achieving capacity is lacking. In our approach, shaping is built directly into the superposition code design. With the decoder we develop, it amenable to the desired analysis, providing the first theory establishing that a practical scheme is reliable at rates approaching capacity for the Gaussian channel.

A. Sparse superposition codes:

The framework for superposition codes is the formation of specific forms of linear combinations of a given set of vectors. We have a list (or book) $X_1, X_2, \ldots, X_N$ of vectors, each with n coordinates, for which the codeword vectors take the form of superpositions

$$\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_N X_N.$$

Andrew R. Barron and Antony Joseph are with the Department of Statistics, Yale University, New Haven, CT 06520 USA e-mail: {andrew.barron, antony.joseph}@yale.edu.

Summaries [11] of portions of this paper were presented at the *IEEE International Symposium on Information Theory*, Austin, Texas, June 13-18, 2010 and St. Petersburg, Russia, August 1-5, 2011.

The vectors X_j provide the terms or components of the codewords with coefficients β_j. By design, each entry of these vectors X_j is independent standard normal and the sum of squares of the coefficients matches the power requirement P. The choice of codeword is conveyed through the coefficients, in particular through the subset of terms for which the coefficient is non-zero. The received vector is in accordance with the statistical linear model

$$Y = X\beta + \varepsilon,$$

where X is the matrix whose columns are the vectors $X_1, X_2, \ldots, X_N$ and ε is the noise vector with distribution $\mathrm{N}(0, \sigma^2 I)$. The book X is called the *design matrix* consisting of $p = N$ variables, each with n observations, and this list of variables is also called the *dictionary* of candidate terms.

For subset superposition coding a subset of terms is arranged, which we call *sent*, with L coefficients non-zero, with specified values. The message is conveyed by the choice of the subset. Denote $B = N/L$ to be the ratio of dictionary size to the number of terms sent. When B is large, it is a *sparse superposition code*, in which the number of terms sent is a small fraction $L/N = 1/B$ of the dictionary size.

Organization of the encoder and decoder is simplified by focussing on what we call a *partitioned* superposition code. The book X is split into L sections of size B, with one term selected from each, yielding L terms in each codeword. Likewise, the coefficient vector β is split into sections, with one coordinate non-zero in each section to indicate the selected term.

As discussed further below, partitioned superposition codes began in work on the capacity region of multi-user Gaussian channels [22], [30]. The L terms of our single-user code correspond via rate-splitting to the superimposed codewords of L messages in that multi-user theory. We are indebted to several ideas that arise from that work, including the particular variable power allocation and successive decoding which is a forerunner of our adaptive successive decoder. Sparse superposition models have long been used in the fields of statistical modeling, sparse signal recovery and compressed sensing. Some work adapting compressed sensing models to communications for the Gaussian channel is in [67], with a number of subsequent developments we shall discuss. Nevertheless, practical schemes with analysis proving rates up to capacity with exponentially small error probability have not previously been developed for the Gaussian channel.

For sparse superposition codes with a specified number L of terms selected, the set of permitted coefficient vectors β is not an algebraic field, that is, it is not closed under linear operations. In particular, summing two coefficient vectors with distinct sets of L non-zero entries does not yield another such coefficient vector. Hence our linear statistical model does not correspond to a linear code in the sense of traditional algebraic coding.

In the codes considered here, it is known in advance to the encoder and decoder what will be the coefficient magnitude if a term is sent. In the simplest case, the values of the non-zero coefficients are the same, with $\beta_j = \sqrt{P/L}\,1_{j\;sent}$. Optionally, the non-zero coefficient values could be $+1$ or -1 times specified magnitudes, in which case the superposition code is said to be *signed* and then the message would conveyed by the sequence of signs as well as the choice of subset. For simplicity we focus in this paper on the unsigned case in which β_j is non-negative.

Partitioning allows for variable power allocation, in which $\beta_j = \sqrt{P_j}\,1_{j\;sent}$ with P_j equal to a prescribed value $P_{(\ell)}$ for j in section ℓ, where $\sum_{\ell}^{L} P_{(\ell)} = P$. Thus $\sum_{j\;sent} P_j = P$, no matter which term is selected from each section. Set weights $\pi_{(\ell)} = P_{(\ell)}/P$ also denoted $\pi_j = P_j/P$ for j in section ℓ. For any set of terms, its *size* induced by the weights is defined as the sum of the π_j for j in that set. We study both the case of constant power allocation and the case that the power is proportional to $e^{-2\mathcal{C}\,\ell/L}$ for sections $\ell = 1, 2, \ldots, L$. These variable power allocations are used in getting the rate up to capacity.

Most convenient with partitioned codes is the case that the section size B is a power of two. Then an input bit string u of length $K = L\log_2 B$ splits into L substrings of size $\log_2 B$ and the encoder becomes trivial. Each substring of u gives the index (or memory address) of the term to be sent from the corresponding section. This makes the encoding especially straightforward. In contrast, for general subset coding, mapping from an input bit string u of length $K = \log\binom{N}{L}$ into a selection of a size L subset out of N is possible, though not so direct.

As we have said, the rate of the code is $R = K/n$ input bits per channel uses and we arrange for arbitrary R less than $\mathcal{C}$. For the partitioned superposition code this rate is

$$R = \frac{L\log B}{n}.$$

Control of the section size B and the number of sections L (and hence control of the dictionary size $N = BL$) is critical to computationally advantageous coding and decoding. If L were constant then for rate R, the section size would exponentially large $B = 2^{nR/L}$, an ultra-sparse case, as arises in the mentioned multi-user theory. In the extreme of $L = 1$, the design X is the whole single-user codebook, with its 2^{nR} columns as the codewords, but the exponential size makes its direct use impractical. At the other extreme there would be $L = K = nR$ sections, each with $B = 2$ candidate terms in subset coding or two signs of a single term in sign coding with $B = 1$; in which case X would be the generator matrix of a linear code.

Between these extremes, the section size B is free to be chosen, though typically sensible values for practice are that it be of order n or a low-order polynomial in n. The corresponding number of sections for a specified code rate is then $L = Rn/\log B$ of order $n/\log n$.

It is in this regime that we construct computationally feasible, reliable, high-rate codes with codewords corresponding to linear combinations of subsets of terms in moderate size dictionaries, with fast decoding algorithms. For the decoder we develop here, at a particular sequence of rates approaching capacity, the error probability is shown to be exponentially small in $L/(\log B)^{3/2}$.

For high rate, near capacity, we require B to be large compared to $(1+snr)^2$ and for high reliability we also require L to be large compared to $(1+snr)^2$, where $snr = P/\sigma^2$ is the signal to noise ratio.

We use a random design matrix. Entries of X are drawn independently from a normal distribution with mean zero and a variance 1 so that the codewords $X\beta$ have a Gaussian shape to their distribution and so that the codewords have average power near P. Other distributions for the entries of X may be considered, such as independent equiprobable ± 1, with a near Gaussian shape for the codeword distribution obtained by the convolutions associated with sums of terms in subsets of size L. Certain other structured designs have been studied in the signal recovery literature, as we shall discuss, but it remains open whether such can achieve rates up to capacity with analogous control of error probability.

There is some freedom in the choice of scale of the coefficients. Here we arrange the coordinates of the X_j to have variance 1 and set the coefficients of β to have sum of squares equal to P. Alternatively, the coefficient representation may be simplified by arranging the coordinates of X_j to be normal with variance P_j and setting the non-zero coefficients of β to have magnitude 1. One may use whichever of these scales is convenient to an argument at hand.

B. *Summary of findings:*

We describe and analyze a fast sparse superposition decoder by a scheme we call adaptive successive decoding.

For computation, it is shown that with a total number of simple parallel processors (multiplier-accumulators) of order $n\,B$, and total memory work space of size $n^2\,B$, it runs in a constant time per received symbol of the string Y.

For the communication rate, there are two cases. First, when the power of the terms sent are the same at P/L in each section, the decoder is shown to reliably achieves rates up to a rate $R_0 = (1/2)P/(P{+}\sigma^2)$ which is less than capacity. It is close to the capacity when the signal-to-noise ratio is low. It is a deficiency of constant power allocation with our scheme that its rate will be substantially less than the capacity if the signal-to-noise is not low.

To bring the rate higher, up to capacity, we use variable power allocation with power $P_{(\ell)}$ proportional to $e^{-2\,\mathcal{C}\ell/L}$, for sections ℓ from 1 to L, with improvements from a slight modification of this power allocation for ℓ/L near 1.

To summarize what is achieved concerning the rate, for each $B \geq 2$, there is a positive communication rate $\mathcal{C}_B$ that our decoder achieves with large L. This $\mathcal{C}_B$ depends on the section size B as well as the signal to noise ratio $snr = P/\sigma^2$. It approaches the Capacity $\mathcal{C} = (1/2)\log(1{+}snr)$ as B increases, albeit slowly. The relative drop from capacity

$$\Delta_B = \frac{\mathcal{C} - \mathcal{C}_B}{\mathcal{C}},$$

is accurately bounded, except for extremes of small and large snr, by an expression near

$$\frac{(1.5 + 1/\nu)\log\log B}{\log B},$$

where $\nu = snr/(1 + snr)$, with other bounds given to encompass accurately also the small and large snr cases.

Concerning reliability, a positive error exponent function $\mathcal{E}(\mathcal{C}_B - R)$ is provided for $R < \mathcal{C}_B$. It is of the order $(\mathcal{C}_B - R)^2\sqrt{\log B}$ for rates R near $\mathcal{C}_B$. The sparse superposition code reliably makes not more than a small fraction of section mistakes. Combined with an outer Reed-Solomon code to correct that small fraction of section mistakes the result is a code with block error probability bounded by an expression exponentially small in $L\,\mathcal{E}(\mathcal{C}_B{-}R)\sqrt{\log B}$, which is exponentially small in $n\,\mathcal{E}(\mathcal{C}_B - R)/\sqrt{\log B}$. For a range of rates R not far from $\mathcal{C}_B$, this error exponent is within a $\sqrt{\log B}$ factor of the optimum reliability exponent.

C. *Decoding sparse superposition codes:*

Optimal decoding for minimal average probability of error consists of finding the codeword $X\beta$ with coefficient vector β of the assumed form that maximizes the posterior probability, conditioning on X and Y (sometimes called the MAP estimator). This coincides, in the case of equal prior probabilities, with the maximum likelihood rule of seeking such a codeword to minimize the sum of squared errors in fit to Y. This is a constrained least squares regression problem $\min_\beta \|Y{-}X\beta\|^2$, with the constraint on the coefficient vector that it correspond to a codeword (also called minimum distance decoding). There is the concern that this exact least squares decoding is computationally impractical. Performance bounds for the optimal decoder are developed in the companion paper [10], achieving rates up to capacity in the constant power allocation case. Instead, here we develop a practical decoder for which we can still establish desired reliability and rate approaching capacity in the variable power allocation case.

The basic step of the decoder is to compute for a given vector, initially the received string Y, its inner product with each of the terms in the dictionary, as test statistics, and see which of these inner products are above a threshold. Such a set of inner products for a step of the decoder is performed in parallel by a computational unit, e.g. a signal-processing chip with $N = LB$ parallel accumulators, each of which has pipelined computation, so that the inner product is updated as the elements of the string arrive.

In this basic step, the terms that it decodes are among those for which the test statistic is above threshold. The step either selects all the terms with inner product above threshold, or a portion of these with specified total weight. Having inner product $X_j^T Y$ above a threshold $T = \|Y\|\tau$ corresponds to having normalized inner product $X_j^T Y/\|Y\|$ above a threshold τ set to be of the form

$$\tau = \sqrt{2\log B} + a,$$

where the logarithm is taken using base e. This threshold may also be expressed as $\sqrt{2\log B}\,(1{+}\delta_a)$ with $\delta_a = a/\sqrt{2\log B}$. The a is a positive value, free to be specified, that impacts the behavior of the algorithm by controlling the fraction of terms above threshold each step. We find that an ideal value of a is moderately small, corresponding to δ_a near $0.75(\log\log B)/\log B$, plus $\log(1 + snr)/\log B$ when snr

is not small. We find that $2\delta_a$ near $1.5 \log\log B/\log B$ plus $4\mathcal{C}/\log B$ constitutes a large part of the above mentioned rate drop Δ_B.

Having the threshold larger than $\sqrt{2\log B}$ implies that the fraction of incorrect terms above threshold is negligible. Yet it also means that only a moderate fraction of correct terms are found to be above threshold each step.

A fit is formed at the end of each step by adding the terms that were selected. Additional steps are used to bring the total fraction decoded up near 1.

Each subsequent step of the decoder computes updated test statistics, taking inner products of the remaining terms with a vector determined using Y and the previous fit, and sees which are above threshold. For fastest operation these updates are performed on additional computational units so as to allow pipelined decoding of a succession of received strings. The test statistic can be the inner product of the terms X_j with the vector of residuals equal to the difference of Y and the previous fit. As will be explained, we find a variant of this statistic to be somewhat simpler to analyze.

A key feature is that the decoding algorithm does not pre-specify which sections of terms will be decoded on any one step. Rather it adapts the choice in accordance with which sections have a term with an inner product observed to be above threshold. Thus one may call our class of procedures *adaptive successive decoding*.

Concerning the advantages of variable power in the partitioned code case, which allows our scheme to achieve rates near capacity, the idea is that the power allocations proportional to $e^{-2\mathcal{C}\ell/L}$ give some favoring to the decoding of the higher power sections among those that remain each step. This produces more statistical power for the test initially as well as retaining enough discrimination power for subsequent steps.

As we review, such power allocation also would arise if one were attempting to successively decode one section at a time, with the signal contributions of as yet un-decoded sections treated as noise, in a way that splits the rate $\mathcal{C}$ into L pieces each of size $\mathcal{C}/L$; however, such pre-specification of one section to decode each step would require the section sizes to be exponentially large to achieve desired reliability. In contrast, in our adaptive scheme, many of the sections are considered each step. The power allocations do not change too much across many nearby sections, so that a sufficient distribution of decodings can occur each step.

For rate near capacity, it helpful to use a modified power allocation, with power proportional to $\max\{e^{-2\mathcal{C}\frac{\ell-1}{L}}, u_{cut}\}$, where $u_{cut} = e^{-2\mathcal{C}}(1+\delta_c)$ with a small non-negative value of δ_c. Thus u_{cut} can be slightly larger than $e^{-2\mathcal{C}}$. This modification performs a slight leveling of the power allocation for ℓ/L near 1. It helps ensure that, even in the end game, there will be sections for which the true terms are expected to have inner product above threshold.

Analysis of empirical bounds on the proportions of correct detections involves events shown to be nearly independent across the L sections. The probability with which such proportions differ much from what is expected is exponentially small in the number of sections L. In the case of variable power allocation we work with weighted proportions of events, which are sums across the terms of indicators of events multiplied by the weights provided by $\pi_j = P_j/P$. With bounded ratio of maximum to minimum power across the sections, such weighted proportions agree with un-weighted proportions to within constant factors. Moreover, for indicators of independent events, weighted proportions have similar exponential tail bounds, except that in the exponent, in place of L we have $L_\pi = 1/\max_j \pi_j$, which is approximately a constant multiple of L for the designs investigated here.

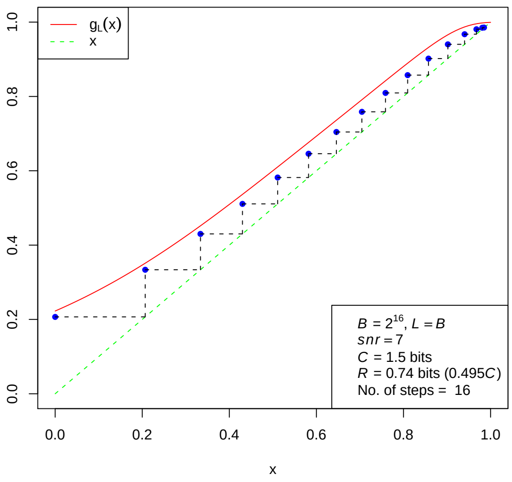

Fig. 1. Plot of the function $g_L(x)$. The dots indicate the sequence $q_{1,k}^{adj}$ for the 16 steps. Here $B = 2^{16}$, $snr = 7$, $R = 0.74$ and L taken to be equal to B. The height reached by the $g_L(x)$ curve at the final step corresponds to a weighted correct detection rate target of 0.993, un-weighted 0.986, for a failed detection rate target of 0.014. The accumulated false alarm rate bound is 0.008. The probability of mistake rates larger than these targets is bounded by 4.8×10^{-4}.

D. An Update Function:

A key ingredient of this work is the determination of a function $g_L : [0,1] \to [0,1]$, called the update function, which depends on the design parameters (the power allocation and the parameters L, B and R) as well as the snr. This function $g_L(x)$ determines the likely performance of successive steps of the algorithm. Also, for a variant of the residual-based test statistics, it is used to set weights of combination that determine the best updates of test statistics.

Let $\hat{q}_k^{tot}$ denote the weighted proportion correctly decoded after k steps. A sequence of deterministic values $q_{1,k}$ is exhibited such that $\hat{q}_k^{tot}$ is likely to exceed $q_{1,k}$ each step. The $q_{1,k}$ is near the value $g_L(q_{1,k-1})$ given by the update function, provided the false alarm rate is maintained small. Indeed, an adjusted value $q_{1,k}^{adj}$ is arranged to be not much less than $g_L(q_{1,k-1}^{adj})$ where the '*adj*' in the superscript denotes an adjustment to $q_{1,k}$ to account for false alarms.

Determination of whether a particular choice of design parameters provides a total fraction of correct detections approaching 1 reduces to verification that this function $g_L(x)$ remains strictly above the line $y = x$ for some interval of the form $[0, x^*]$ with x^* near 1. The successive values of the gap $g_L(x_k) - x_k$ at $x_k = q_{1,k-1}^{adj}$ control the error exponents as well as the size of the improvement in the detection rate and the number of steps of the algorithm. The final weighted fraction of failed detections is controlled by $1 - g_L(x^*)$.

The role of g_L is shown in Fig 1. Provision of $g_L(x)$ and the computation of its iterates provides a computational devise by which a proposed scheme is checked for its capabilities.

An equally important use of $g_L(x)$ is analytical analysis of the extent of positivity of the gap $g_L(x) - x$ depending on the design parameters. For any power allocation there will be a largest rate R at which the gap remains positive over most of the interval $[0, 1]$ for sufficient size L and B. Power allocations with $P_{(\ell)}$ proportional to $e^{-2\mathcal{C}\ell/L}$, or slight modifications thereof, are shown to be the form required for the gap $g_L(x) - x$ to have such positivity for rates R near $\mathcal{C}$.

Analytical examination of the update function shows, for large L, how the choice of the rate R controls both the size of the shortfall $1-x^*$ and the minimum size of the gap $g_L(x)-x$ for $0 \leq x \leq x^*$, as functions of B and snr. Thereby bounds are obtained on the mistake rate, the error exponent, and the maximal rate for which the method produces high reliability.

To summarize, with the adaptive successive decoder and suitable power allocation, for rates approaching capacity, the update function stays sufficiently above x over most of $[0, 1]$ and, consequently, the decoder has a high chance of not more than a small fraction of section mistakes.

E. Accounting of section mistakes:

Ideally, the decoder selects one term from each section, producing an output which is the index of the selected term. It is not in error when the term selected matches the one sent.

In a section a mistake occurs from an incorrect term above threshold (a *false alarm*) or from failure of the correct term to provide a statistic value above threshold after a suitable number of steps (a *failed detection*). We let $\hat{\delta}_{mis}$ refer to the failed detection rate plus the false alarm rate, that is, the sum of the fraction of section with failed detections and the fraction of sections with false alarms. This sum from the two sources of mistake is at least the fraction of section mistakes, recognizing that both types can occur. Our technique controls this $\hat{\delta}_{mis}$ by providing a small bound δ_{mis} that holds with high probability.

A section mistake is counted as an *error* if it arises from a single incorrectly selected term. It is an *erasure* if no term is selected or more than one term is selected. The distinction is that a section error is a mistake you don't know you made and a section erasure is one you known you made. Let $\hat{\delta}_{error}$ be the fraction of section errors and $\hat{\delta}_{erase}$ be the fraction of section erasures. In each section one sees that the associated indicators of events satisfy the property that $1_{erase}+2\,1_{error}$ is not more than $1_{failed\,detection} + 1_{false\,alarm}$. This is because an error event requires both a failed detection and a false alarm. Accordingly $2\hat{\delta}_{error} + \hat{\delta}_{erase}$ is not more than $\hat{\delta}_{mis}$, the failed detection rate plus the false alarm rate.

F. An outer code:

An issue with this superposition scheme is that candidate subsets of terms sent could differ from each other in only a few sections. When that is so, the subsets could be difficult to distinguish, so that it would be natural to expect a few section mistakes.

An approach is discussed which completes the task of identifying the terms by arranging sufficient distance between the subsets, using composition with an outer Reed-Solomon (RS) code of rate near one. The alphabet of the Reed-Solomon code is chosen to be a set of size B, a power of 2. Indeed we arrange the RS symbols to correspond to the indices of the selected terms in each section. Details are given in a later section. Suppose the likely event $\hat{\delta}_{mis} < \delta_{mis}$ holds from the output of the inner superposition code. Then the outer Reed-Solomon corrects the small fraction of remaining mistakes so that we end up not only with small section mistake rate but also with small block error probability. If $R_{outer} = 1 - \delta$ is the communication rate of an RS code, with $0 < \delta < 1$, then the section errors and erasures can be corrected, provided $\delta_{mis} \leq \delta$.

Furthermore, if R_{inner} is the rate associated with our inner (superposition) code, then the total rate after correcting for the remaining mistakes is given by $R_{total} = R_{inner}R_{outer}$, using $\delta = \delta_{mis}$. Moreover, if Δ_{inner} is the relative rate drop from capacity of the inner code, then the relative rate drop of the composite code Δ_{total} is not more than $\delta_{mis} + \Delta_{inner}$.

The end result, using our theory for the distribution of the fraction of mistakes of the superposition code, is that for suitable rate up to a value near capacity the block error probability is exponentially small.

One may regard the composite code as a superposition code in which the subsets are forced to maintain at least a certain minimal separation, so that decoding to within a certain distance from the true subset implies exact decoding.

Performance of the sparse superposition code is measured by the three fundamentals of computation, rate, and reliability.

G. Computational resource:

The main computation required of each step of the decoder is the computation of the inner products of the residual vectors with each column of the dictionary. Or one has computation of related statistics which require the same order of resource. For simplicity in this subsection we describe the case in which one works with the residuals and accepts each term above threshold. The inner products requires order nLB multiply-and-adds each step, yielding a total computation of order $nLBm$ for m steps. As we shall see the ideal number of steps m according to our bounds is not more than $2 + snr \log B$.

When there is a stream of strings Y arriving in succession at the decoder, it is natural to organize the computations in a parallel and pipelined fashion as follows. One allocates m signal processing chips, each configured nearly identically, to do the inner products. One such chip does the inner products with Y, a second chip does the inner products with the residuals from the preceding received string, and so on, up to chip m which is working on the final decoding step from the string received several steps before.

Each signal processing chip has in parallel a number of simple processors, each consisting of a multiplier and an accumulator, one for each stored column of the dictionary under consideration, with capability to provide pipelined accumulation of the required sum of products. This permits the collection of inner products to be computed online as the coordinates of the vectors are received. After an initial delay of m received strings, all m chips are working simultaneously.

Moreover, for each chip there is a collection of simple comparators, which compare the computed inner products to the threshold and store, for each column, a flag of whether it is to be part of the update. Sums of the associated columns are computed in updating the residuals (or related vectors) for the next step. The entries of that simple computation (sums of up to L values) are to be provided for the next chip before processing the entries of the next received vector. If need be, to keep the runtime at a constant per received symbol, one arranges $2m$ chips, alternating between inner product chips and subset sum chips, each working simultaneously, but on strings received up to $2m$ steps before. The runtime per received entry in the string Y is controlled by the time required to load such an entry (or counterpart residuals on the additional chips) at each processor on the chip and perform in parallel the multiplication by the associated dictionary entries with result accumulated for the formation of the inner products.

The terminology *signal processing chip* refers to computational units that run in parallel to perform the indicated tasks. Whether one or more of these computational units fit on the same physical computer chip depends on the size of the code dictionary and the current scale of circuit integration technology, which is an implementation matter not a concern at the present level of decoder description.

If each of the signal processing chips keeps a local copy of the dictionary X, alleviating the challenge of numerous simultaneous memory calls, the total computational space (memory positions) involved in the decoder is $nLBm$, along with space for LBm multiplier-accumulators, to achieve constant order computation time per received symbol. Naturally, there is the alternative of increased computation time with less space; indeed, decoding by serial computation would have runtime of order $nLBm$.

Substituting $L = nR/\log B$ and m of order $\log B$ we may reexpress $nLBm$. One sees that the total computational resource required (either space or time) is of order n^2B for this sparse superposition decoder. More precisely, to include the effect of the snr on the computational resource, using the number of steps m which arise in upcoming bounds, which is within 2 of $snr \log B$, and using R upper bounded by capacity $\mathcal{C}$, we have the computational resource of $nLBm$ memory positions bounded by $\mathcal{C} snr\, n^2 B$, and a number LBm of multiplier-adders bounded by $\mathcal{C} snr\, n\, B$.

In concert with the action of this decoder, the additional computational resource of a Reed-Solomon decoder acts on the indices of which term is flagged from each section to provides correction of the few mistakes. The address of which term is flagged in a section provides the corresponding symbol for the RS decoder, with the understanding that if a section has no term flagged or more than one term flagged it is treated as an erasure. For this section, as the literature on RS code computation is plentiful and yet undergoing continuing development, we simply note that the computation resource required is also bounded as a low order polynomial in the size of the code.

H. Achieved rate:

This subsection discusses the nature of the rates achieved with adaptive successive decoding. We achieve not only fixed rates $R < \mathcal{C}$, but also rates R up to $\mathcal{C}_B$, for which the gap from capacity is of the order near $1/\log B$.

Two approaches are provided for evaluation of how high a rate R is achieved. For any L, B, snr, any specified error probability and any small specified fraction of mistakes of the inner code, numerical computation of the progression of $g_L(x)$ permits a numerical evaluation of the largest R for which $g_L(x)$ remains above x sufficiently to achieve the specified objectives.

The second approach is to provide simplified bounds to prove analytically that the achieved rate is close to capacity, and exhibit the nature of the closeness to capacity as function of snr and B. This is captured by the rate envelope C_B and bounds on its relative rate drop Δ_B. Here we summarize contributions to Δ_B, in a way that provides a partial blueprint to later developments. Fuller explanation of the origins of these contributions arise in these developments in later sections.

The update function $g_L(x)$ is near a function $g(x)$, with difference bounded by a multiple of $1/L$. Properties of this function are used to produce a rate expression that ensures that $g(x)$ remains above x, enabling the successive decoding to progress reliability. In the full rate expression we develop later, there are quantities η, h, and ρ that determine error exponents multiplied by L. So for large enough L, these exponents can be taken to be arbitrarily small. Setting those quantities to the values for which these exponents would become 0, and ignoring terms that are small in $1/L$, provides simplification giving rise to what we call the rate envelope denoted C_B.

With this rate envelope, for $R < C_B$, these tools enable us to relate the exponent of reliability of the code to a positive function of $C_B - R$ times L, even for L finite.

There are two parts to the relative rate drop bound Δ_B, which we write as Δ_{shape} plus Δ_{alarm}, with details on these in later sections. Here let's summarize these contributions to express the form of our bound on Δ_B.

The second part denoted Δ_{alarm} is determined by optimizing a combination of rate drop contributions from $2\delta_a$, plus a term $snr/(m-1)$ involving the number of steps m, plus terms involving the accumulated false alarm rate. Using the natural logarithm, this Δ_{alarm} is optimized at m equal to an integer part of $2 + snr \log B$ and an accumulated baseline false alarm rate of $1/[(3\mathcal{C} + 1/2) \log B]$. At this optimized m and optimized false alarm rate, the value of the threshold parameter δ_a is

$$\delta_a = \frac{\log \left[m\, snr (3 + 1/2\mathcal{C}) \sqrt{\log B} / \sqrt{4\pi} \right]}{2 \log B}$$

and

$$\Delta_{alarm} = 2\delta_a + \frac{2}{\log B}.$$

At the optimized m, the δ_a is an increasing function of snr, with value approaching $0.25\log[(\log B)/\pi]/\log B$ for small snr and value near

$$\frac{.75\log\log B + 2\mathcal{C} - 0.25\log(4\pi/9)}{\log B}$$

for moderately large snr. The constant subtracted in the numerator $0.25\log(4\pi/9)$ is about 0.08. With δ_a thus set, it determines the value of the threshold $\tau = \sqrt{2\log B}(1+\delta_a)$.

To obtain a small $2\delta_a$, and hence small Δ_{alarm}, this bound requires $\log B$ large compared to $4\mathcal{C}$, which implies that the section size B is large compared to $(1+snr)^2$.

The Δ_{shape} depends on the choice of the variable power allocation rule, via the function g and its shape. For a specified power allocation, it is determined by a minimal inner code rate drop contribution at which the function has a non-negative gap $g(x)-x$ on $[0, x^*]$, plus the contribution to the outer code rate drop associated with the weighted proportion not detected $\delta^* = 1-g(x^*)$. For our determination of Δ_{shape}, we examine three cases for power allocation and, for each snr, pick the one with the best such tradeoff, which includes determination of the best x^*. The result of this examination is a Δ_{shape} which is a decreasing function of snr.

The first case has no leveling ($\delta_c = 0$). In this case the function $g(x)-x$ is decreasing for suitable rates. Using an optimized x^* it provides a candidate Δ_{shape} equal to $1/\tau^2$ plus $\xi_{\mathcal{C}}/(\tau\mathcal{C})$, where $\xi_{\mathcal{C}}$ is an explicitly given expression with value near $\sqrt{2\log(\mathcal{C}+1/2)}$ for large $\mathcal{C}$. If snr is not large, this case does not accomplish our aims because the term involving $1/\tau$, near $1/\sqrt{2\log B}$, is not small enough for our purposes. Yet with snr such that $\mathcal{C}$ is large compared to τ, this Δ_{shape} is acceptable, providing a contribution to the rate drop near the value $1/(2\log B)$. Then the total rate drop is primarily determined by Δ_{alarm}, yielding, for large snr, that Δ_B is near

$$\frac{1.5\log\log B + 4\mathcal{C} + 2.34}{\log B}.$$

This case is useful for a range of snr, where $\mathcal{C}$ exceeds a multiple of $\sqrt{\log B}$ yet remains small compared to $\log B$.

The second case has some leveling with $0 < \delta_c < snr$. In this case the typical shape of the function $g(x)-x$, for x in $[0, 1]$, is that it undergoes a single oscillation, first going down, then increasing, and then decreasing again, so there are two potential minima for x in $[0, x^*]$, one of which is at x^*. In solving for the best rate drop bound, a role is demonstrated for the case that δ_c is such that an equal gap value is reached at these two minima. In this case, with optimized x^*, a bound on Δ_{shape} is shown, for a range of intermediate size signal to noise ratios, to be given by the expression

$$\frac{2}{\nu\,\log B}\left\{2 + \log\big(\frac{1}{2} + \frac{\nu\tau}{4\mathcal{C}\sqrt{2\pi}}\big)\right\} + \frac{1}{2\log B},$$

where $\nu = snr/(1+snr)$. When $2\mathcal{C}/\nu$ is small compared to $\tau/\sqrt{2\pi}$, this Δ_{shape} is near $(1/\nu)(\log\log B)/(\log B)$. When added to Δ_{alarm} it provides an expression for Δ_B, as previously given, that is near $(1.5+1/\nu)\log\log B/\log B$ plus terms that are small in comparison.

The above expression provides our Δ_{shape} as long as snr is not too small and $2\mathcal{C}/\nu$ is less than $\tau/\sqrt{2\pi}$. For $2\mathcal{C}/\nu$ at least $\tau/\sqrt{2\pi}$, the effect of the $\log\log B$ is canceled, though there is then an additional small remainder term that is required to be added to the above as detailed later. The result is that Δ_{shape} is less than $const/\log B$ for $(2\mathcal{C}/\nu)\sqrt{2\pi}$ at least τ.

The third case uses constant power allocation (complete leveling with $\delta_c = snr$), when snr is small. The Δ_{shape} is less than a given bound near $\sqrt{2(\log\log B)/\log B}$ when the snr is less than twice that value. For such sufficiently small snr this Δ_{shape} with complete leveling becomes superior to the expression given above for partial leveling.

Accordingly, let Δ_{shape} be the best of these values from the three cases, producing a continuous decreasing function of snr, near $\sqrt{2(\log\log B)/\log B}$ for small snr, near $(1+1/snr)\log\log B/(\log B)$ for intermediate snr, and near $1/2\log B$ for large snr. Likewise, the Δ_B bound is $\Delta_{shape} + \Delta_{alarm}$. In this way one has the dependence of the rate drop on snr and section size B.

Thus we let $\mathcal{C}_B$ be the rate of the composite sparse superposition inner code and Reed-Solomon outer code obtained from optimizing the total relative rate drop bound Δ_B.

Included in Δ_{alarm} and Δ_{shape}, which sum to Δ_B, are baseline values of the false alarm rates and the failed detection rates, respectively, which add to provide a baseline value δ^*_{mis}, and, accordingly, our Δ_B splits as δ^*_{mis} plus $\Delta_{B,inner}$, using the relative rate drop of the inner code. As detailed later, this δ^*_{mis} is typically small compared to the rate drop sources from the inner code.

In putting the ingredients together, when R is less than $\mathcal{C}_B$, part of the difference $\mathcal{C}_B$–R is used in providing slight increase past the baseline to determine a reliable δ_{mis}, and the rest of the difference is used in setting the inner code rate to insure a sufficiently positive gap $g(x)-x$ for reliability of the decoding progression. The relative choices are made to produce the best resulting error exponent $L\,\mathcal{E}(\mathcal{C}_B - R)$ for the given rate.

I. Comparison to optimal least squares:

It is appropriate to compare the rate achieved here by our practical decoder with what is achieved with theoretically optimal, but possibly impractical, least squares decoding of these sparse superposition codes, subject to the constraint that there is one non-zero coefficients in each section. Such least squares decoding provides the stochastically smallest distribution of the number of mistakes, with a uniform distribution on the possible messages, but it has an unknown computation time.

In this direction, the results in the companion paper [10], for least squares decoding of superposition codes, partially complement what we give here for our adaptive successive decoder. For optimum least square decoding, favorable properties are demonstrated, in the case that the power assignments P/L are the same for each section. Interestingly, the analysis techniques there are different and do not reveal rate improvement from the use of variable instead of constant power with optimal least squares decoding. Another difference is that while here there are no restrictions on B, there it is required that $B \geq L^b$ for a specified section size rate b depending only on the signal-to-noise ratio, where conveniently b tends to 1 for large signal-to-noise, but unfortunately b gets large for small snr.

For comparison with our scheme here, restrict attention to moderate and large signal-to-noise ratios, as for computational reasons, it is desirable that B be not more than a low order polynomial in L.

Let $\Delta = (\mathcal{C}-R)/\mathcal{C}$ be the rate drop from capacity, with R not more than $\mathcal{C}$. It is shown in [10] that, for a positive constant c_1, the probability of more than a fraction δ_{mis} of mistakes, with least squares decoding, is less than $\exp\{-nc_1\min\{\Delta^2,\delta_{mis}\}\}$, for any δ_{mis} in $[0,1]$, any positive rate drop Δ and any size n. This bound is better than obtained for our practical decoder in its freedom of any choice of mistake fraction and rate drop in obtaining this reliability. In particular, the result for least squares does not restrict Δ to be larger than Δ_B and does not restrict δ_{mis} to be larger than a baseline value of order $1/\log B$.

It shows that n only needs to be of size $[\log(1/\epsilon)]/[c_1\min\{\Delta^2,\delta_{mis}\}]$ for least squares to achieve probability ϵ of at least a fraction δ_{mis} mistakes, at rate that is Δ close to capacity. With suitable target fractions of mistakes, the drop from capacity Δ is not more than $\sqrt{(1/c_1 n)\log 1/\epsilon}$. It is of order $1/\sqrt{n}$ if ϵ is fixed; whereas, for ϵ exponentially small in n, the associated drop from capacity Δ would need to be at least a constant amount.

An appropriate domain for comparison is in a regime between the extremes of fixed probability ϵ and a probability exponentially small in n. The probability of error is made nearly exponentially small if the rate is permitted to slowly approach capacity. In particular, suppose B is equal to n or a small order power of n. Pick Δ of order $1/\log B$ to within iterated log factors, arranged such that the rate drop Δ exceeds the envelope Δ_B by an amount of that order $1/\log B$. We can ask, for a rate drop of that moderately small size, how would the error probability of least squares and the practical method compare? At a suitable mistake rate, the exponent of the error probability of least squares would be quantified by $n/(\log B)^2$ of order $n/(\log n)^2$, neglecting loglog factors. Whereas, for our practical decoder the exponent would be a constant times $L(\Delta-\Delta_B)^2(\log B)^{1/2}$, which is of order $L/(\log B)^{1.5}$, that is, $n/(\log n)^{2.5}$. Thus the exponent for the practical decoder is within a $(\log n)^{0.5}$ factor of what is obtained for optimal least squares decoding.

J. Comparison to the optimal form of exponents:

It is natural to compare the rate, reliability, and code-size tradeoff that is achieved here, by a practical scheme, with what is known to be theoretically best possible. What is known concerning the optimal probability of error, established by Shannon and Gallager, as reviewed for instance in [58], and recently refined in [3], is that the optimal probability of error is exponentially small in an expression $n\,\mathcal{E}(R)$ which, for R near $\mathcal{C}$, matches $n\Delta^2$ to within a factor bounded by a constant, where $\Delta = (\mathcal{C}-R)/\mathcal{C}$. Per [3], this behavior of the exponent remains valid for Δ down to the order remaining larger than $1/\sqrt{n}$. The reason for that restriction is that for Δ as small as order $1/\sqrt{n}$, the optimal probability of error does not go to zero with increasing block length (rather it is then governed by an analogous expression involving the tail probability of the Gaussian distribution, per [58]).

Our bounds for our practical decoder do not rely on asymptotics, but rather finite sample bounds available for all choices of L and B and inner code rates $R \leq \mathcal{C}_B$, with blocklength $n = (L\log B)/R$. Our overall error probability bound is exponentially small in an expression of the form $L\,\min\left\{\Delta, \Delta^2\sqrt{\log B}\right\}$, provided R is enough less than $\mathcal{C}_B$ that the additional drop from $\mathcal{C}_B - R$ is of the same order as the total drop Δ. Consequently, the error probability is exponentially small in

$$n\,\min\left\{\frac{\Delta}{\log B}, \frac{\Delta^2}{\sqrt{\log B}}\right\}.$$

Focussing on the Δ for which the square term is the minimizer, it shows that the error probability is exponentially small in $n(\mathcal{C}-R)^2/\sqrt{\log B}$, within a $\sqrt{\log B}$ factor of optimal, for rates R for which $\mathcal{C}_B - R$ is of order between $\log\log B/\log B$ and $1/\sqrt{\log B}$.

An alternative perspective on the rate and reliability tradeoff as in [58], is to set a small block error probability ϵ and seek the largest possible communication rate R_{opt} as a function of the codelength. They show for n of at least moderate size, this optimal rate is near

$$R_{opt} = \mathcal{C} - \frac{\sqrt{V}}{\sqrt{n}}\sqrt{2\log 1/\epsilon},$$

for a constant V they identify, where if ϵ is not small the $\sqrt{2\log 1/\epsilon}$ is to be replaced by the upper ϵ quantile of the standard normal. For small ϵ this expression agrees with the form of the relationship between error probability ϵ and the exponent $n(\mathcal{C}-R_{opt})^2$ stated above. The rates and error probabilities we achieve with our practical decoder have a similar form of relationship but differ in three respects. One is that we have the somewhat smaller $n/\sqrt{\log B}$ in place of n, secondly our constant multipliers do not match the optimal V, and thirdly our result is only applicable for ϵ small enough that the rate drop is made to be at least Δ_B.

From either of these perspectives, we see that to gain provable practicality we pay the price of needing blocklength larger by a factor of $\sqrt{\log B}$ to have the same performance as would be optimal without concern for practicality.

K. On the signal alphabet and shaping:

From the review [36], as we have said, the problem of practical communication for additive Gaussian noise channels, has been decomposed into separate problems, which in addition to modulation, include the matters of choice of signal alphabet, of the shaping of a signal constellation, and of coding. Our approach merges the signal alphabet and constellation into the coding. The values of codeword symbols that arise in our codes are those that can be realized via sums of columns of the dictionary, one from each section in the partitioned case. Some background on signalling facilitates discussion of relationship to other work.

By choice of signal alphabet, codes for discrete channels have been adapted to use on Gaussian channels, with varying degrees of success. In the simplest case the code symbols take on only two possible values, leading to a binary input channel, by constraining the the symbol alphabet to allow only the values $\pm\sqrt{P}$ and possibly using only the signs of the Y_i. With such binary signalling, the available capacity is not more than 1 and it is considerably less than $(1/2)\log(1+snr)$, except in the case of low signal-to-noise ratio. When considering snr that is not small it is preferable to not restrict to binary signalling, to allow higher rates of communication. When using signals where each symbol has a number M of levels, the rate caps at $\log M$, which is achieved in the high snr limit even without coding (simply infer for each symbol the level to which the received Y_i is closest). As quantified in Forney and Ungerboeck [36], for moderate snr, treating the channel as a discrete M-ary channel of particular cross-over probabilities and considering associated error-correcting codes allows, in theory, for reasonable performance provided $\log M$ sufficiently exceeds $\log snr$ (and empirically good coding performance has been realized by LDPC and turbo codes). Nevertheless, as they discuss, the rate of such discrete channels remains less than the capacity of the original Gaussian channel. Recent quantification of how small is the gap between the capacity and the mutual information of an M-symbol input distribution is in [74] and [1].

To bring the rate up to capacity, the codeword choices should form a properly shaped multivariate signal constellation, that is, the codeword vectors should approximate a good packing of points on the $n-$dimensional sphere of squared radius dictated by the power. An implication of which is that, marginally and jointly for any subset of codeword coordinates, the set of codewords should have empirical distribution not far from Gaussian. Such shaping is likewise a problem for which theory dictates what is possible in terms of rate and reliability [36], but theory has been lacking to demonstrate whether there is a moderate or low complexity of decoding that achieves such favorable rate and error probability.

Our sparse superposition code automatically takes care of the required shaping by using linear combinations of subsets a given set of real-valued Gaussian distributed vectors. For high snr, in the role of $\log M$ being large compared to $\log snr$ is in our case replaced by having L large and having $\log B$ large compared to $\mathcal{C}$. These sparse superposition codes are not exactly well-spaced on the surface of an $n-$sphere, as inputs that agree in most sections would have nearby codewords. Nevertheless, when coupled with the Reed-Solomon outer code, sufficient distance between codewords is achieved for quantifiably high reliability.

L. Relationships to previous work:

We point out several directions of past work that connect to what is developed here. There is some prior work concerning computational feasibility for reliable communications near capacity for certain channels.

In some cases, instead of fixing snr and considering rates less than $\mathcal{C} = (1/2)\log(1 + snr)$, the codes are equivalently examined by fixing a target rate R, which would be capacity for $snr^* = e^{2R} - 1$, and examining reliability for all snr at least snr^*.

Turbo codes, also called parallel permuted convolutional codes, have empirically good performance [13], e.g., for rate $1/2$ codes for snr not much above what would correspond to capacity. Typically they have a signed binary input sets, nevertheless, these codes are intended for additive noise channels such as Gaussian and the decoder makes use of the real-valued received symbols in its iterations [42]. The presence of some low weight codewords [57] is demonstrated to produce a floor to the bit error probability that is an impediment to the demonstration of reliability that will scale favorably with increasing code-size n. Increasing the number of parallel codes improves the minimum distance as shown in [43], nevertheless, demonstration that the error probability is exponentially small with n for all rates less than capacity (i.e. for all $snr > snr^*$) remains elusive, even for these moderate target rates. The fact that current designs are for rate targets less than 1 need not be an ultimate impediment as one can use the superposition strategy and rate splitting to produce much higher rates of the same fractional drop from capacity in high snr cases as hold in low snr cases.

Building on Gallager's *low density parity check codes* [38], iterative decoding algorithms based on statistical belief propagation in loopy networks have been *empirically* shown in various works to provide reliable and moderately fast decoding at rates near the capacity for various discrete channels, and mathematically proven to provide such properties in the special case of the binary erasure channel in [51], [52]. Theory for expander codes for the binary symmetric channel is found in [64], [75], [6], [7] including error exponents for reliability at rates up to capacity in [6]. Error exponents for randomly filled low density generator matrices is demonstrated in [48]. Though substantial steps have been made in the analysis of belief networks (message passing algorithms), as summarized for instance in [61], there is not mathematical proof of the desired computational properties and reliability properties at rates near capacity for the Gaussian channel.

An approach to reliable and computationally-feasible decoding, originally restricted to binary signaling, is in the work on *channel polarization*. Error probability is demonstrated there at a level exponentially small in $n^{1/2}$ for fixed rates less than the binary signaling capacity. In contrast for our scheme, the error probability is exponentially small in $n/(\log B)^{0.5}$ and hence exponentially small in $n^{1-\epsilon}$ for any $\epsilon > 0$ and communication is permitted at higher rates, approaching capacity for the Gaussian noise channel. In recent work Abbe and Telatar [2] adapt channel polarization to achieve the sum rate capacity for m user binary input multiple-access channels, with specialization to single-user channels with 2^m inputs. Subsequent to the initial development of the methods of the present paper, Abbe and one of us [1] have recently demonstrated discrete near-Gaussian signalling that adapts channel polarization to the Gaussian noise channel.

The analysis of *concatenated codes* in Forney [35] is an important forerunner to the development of code composition we give here. For the theory, he paired an outer Reed-Solomon code with concatenation of optimal inner codes of Shannon-Gallager type, while, for practice, he paired such an outer Reed-Solomon code with binary inner codes based on linear combinations of orthogonal terms (for target rates K/n less than 1 such a basis is available), in which all binary coefficient sequences are possible codewords. A refinement of the Forney theory is in [6], using binary expander inner codes for the binary symmetric channel paired with an outer Reed-Solomon code.

A challenge concerning theoretically good inner codes is that the number of messages searched is exponentially large in the inner codelength. Forney made the inner codelength of logarithmic size compared to the outer codelength as a step toward practical solution. However, caution is required with these strategies. If the rate of the inner code has a small relative drop from capacity, $\Delta = (\mathcal{C}{-}R)/\mathcal{C}$, then for moderate reliability the inner codelength would need to be of order at least $1/\Delta^2$. So with these the required outer codelength becomes exponential in $1/\Delta^2$.

To compare, for the Gaussian noise channel, our approach provides a practical decoding scheme for the inner code. We permit use of inner and outer codelengths that are comparable to each other. One can draw a parallel between the sections of our code and the concatenations of Forney's inner codes. A key difference is our use of superposition across the sections and the simultaneous decoding of these sections. Challenges remain in the restrictiveness of the relationship of the rate drop Δ to the section sizes. Nevertheless, particular rates are identified as practical and near optimal.

Ideas of superposition codes, including rate splitting, partitioning, successive decoding and variable power allocation for Gaussian noise channels, began with Cover [22] in the context of multiple-user channels, as we have said. Let's amplify that relationship here. In his broadcast channel setting what is sent is a sum of L codewords, one for each message, and the users are sorted, the last one needing to fully decode the partitioned superposition code. Here we are putting that idea to use for the original Shannon single-user problem. The purpose here of computational feasibility is different from the original multi-user purpose which was characterization of the set of achievable rates. Variable power assignments, such as we use, yield equal rate R/L for each portion of the code corresponding to a section (or individual code) size $B = 2^{nR/L}$ for R up to the total capacity. Such B would be exponentially large in n if L were fixed. Instead, we have L of order n to within a log factor, so that B is of a manageable size. Also, as we have said, adaptation rather than pre-specification of the set of sections decoded each step is key to the reliability and speed of our scheme.

The superposition code ideas originating in [22] were applied also for achieving the sum rate in Gaussian multiple-access channels, see, e.g. [30], and for random access channels [34]. In the multiple-access rate region characterizations of [62] and [18], rate splitting is in some cases applied to individual users. So the applicability of superposition of rate split codes (with associated partitioning and variable power allocation) for a single user channel at rates up to capacity has been noted, starting from both broadcast and multiple access perspectives. However, feasibility has been lacking in the absence of demonstration of reliability at high rate with superpositions from polynomial size dictionaries with fast adaptive decoders.

It is an attractive feature of our solution for the single-user channel that it should be amenable to extension to practical solution of the corresponding multi-user channels, namely, the Gaussian multiple access and Gaussian broadcast channel.

Convex optimization and compressed sensing: Our first efforts in attempting to provide a practical decoder, reliable at rates up to capacity, involved trying to adapt existing results on convex optimization, sparse approximation, and compressed sensing. With focus on rate in comparison to capacity, the potential success and existing shortcomings of these approaches are discussed here.

Relevant convex optimization concerns the problem of least squares *convex projection* onto the convex hull of a given set of vectors. If there is the freedom to multiply these vectors by a specified constant, then such convex projection is also called ℓ_1-constrained least squares, basis pursuit [19], or the Lasso [66], though there are certainly a number of relevant precursors on such optimization. Formulation as an ℓ_1-penalized least squares is popular in cases of sparse statistical linear modeling and compressed sensing in which the non-zero coefficient values are unknown, whereas ℓ_1-constrained least squares is a more natural match to our setting in which the non-zero coefficient values are known.

The idea with such optimization is to show with certain rate constraints and dictionary properties that the convex projection is likely to concentrate its non-zero coefficients on the correct subset. Completion of convex optimization to very high precision would entail a computation time in general of the order of N^3. An alternative is to perform a smaller number of iterations, such as we do here, aimed at determining the target subset.

One line of work on sparse approximation and term selection concerns a class of iterative procedures which may be called relaxed greedy algorithms (including orthogonal matching pursuit, also called the orthogonal greedy algorithm, related to forward stepwise regression) as studied in [47], [8], [56], [49], [9], [45], [70], [76], [77]. In essence, each step of these algorithms finds, for a given set of vectors, the one which maximizes the inner product with the residuals from the previous iteration and then uses it to update the linear combination. The relaxation property, in optimizing the linear combination with previous contributions, is that those contribution can be down-weighted in the presence of the new vector.

These procedures adapt to two purposes of relevance to the decoding task. First a relaxed greedy algorithm solves, to within specified precision, for the least squares *convex projection* onto the convex hull of a given set of vectors as shown in [49] and a variant of it solves for the ℓ_1 penalized least squares solution as shown in [45], with quantification of the ℓ_2 accuracy of sparse approximators obtained in k steps [8], [49], [9], [45], with no assumptions concerning the size of inner products of the vectors.

Second, results on the correctness with high probability of the k term selection with such algorithms are obtained in [70] using orthogonal matching pursuit with random designs and in [76] using forward-backward stepwise regression with an assumption that the design satisfies what is called a restricted isometry property (RIP), satisfied with high probability by the independent Gaussian and ±1 random designs. The RIP property is also satisfied by certain deterministic designs as reviewed in [14], [46]. With the constants currently available with these analyses in the additive Gaussian noise case studied in [70], [76] the application of these results does not yet permit rates up to capacity.

Each pass of a relaxed greedy algorithm is analogous to the steps of the decoder studied here. Applied to the communication task with partitioned superposition codes, the convex hull corresponds to the coefficient vectors which are nonnegative and sum to 1 in each section (a cartesian product of simplices), for which the vertices correspond to the codewords. Each step finds *in every section* the term of highest inner product with the residuals from the previous iteration and then uses it to update the linear combination. Accordingly, its computation time is linear in N times the number of iterations (and these computations are parallelizable in the same manner as previously discussed).

The essential difference between the above-mentioned iterative algorithms and our decoder is that each step we achieve relaxation by keeping the coefficients at 0 in sections for which all the inner products remain below threshold. Accordingly, section moves, when they occur, are to vertices of the constituent simplices, and there are no interior moves.

We have conducted additional analysis of convex projection (ℓ_1-constrained least squares) in the case of equal power allocation in each section. With the Gaussian design and Gaussian noise, an approximation to the projection can be characterized which has largest weight in most sections at the term sent, when the rate R is less than R_0; whereas for larger rates the weights of the projection are too spread across the terms in the sections to identify what was sent. (As that analysis is lengthy and does not get the rate up to capacity we will not include it here.) To get to the higher rates, up to capacity, one cannot use such convex projection alone. Variable section power may be necessary in the context of such algorithms. We have found it advantageous to conduct a more structured iterative decoding, which is more explicitly targeted to finding vertices, as presented here.

With this backdrop, it is natural to look further at the fields of statistical term selection, sparse signal recovery, and compressed sensing and ask whether the desired results of practical achievement of rate up to capacity could be obtained by appealing to other existing conclusions.

Adaptation of sparse signal recovery theory to communications for the Gaussian noise channel begins, as we have said, with Tropp [67] and Gilbert and Tropp [41], by a scheme which may be regarded as a sparse signed superposition code using a dictionary of specified properties (an control of incoherence, assuming small maximum pairwise inner products). Analogous signal recovery frameworks are in [27], [28] and [37]. These works examine the reliability properties in the worst case over choices of arbitrary subsets of specified size and show that subsets up to size of order near $\sqrt{n}$ can be reliably decoded by convex projection. Unfortunately, this would correspond to communication rates that are vanishingly small, of order near $1/\sqrt{n}$.

Positive rate is realized by looking at the average probability of error over random choices of subsets of a specified size. From an n by N dictionary of specified properties, random subsets up to size of order $n/\log N$ are reliably decoded. Work in this direction is in [68], [69] and especially [15], showing that a coherence property of the dictionary implies reliable determination of random subsets of that size by ℓ_1 penalized least squares. The relationship between random subsets of columns a fixed dictionary of specified properties and the subsets of columns of a Gaussian matrix is explored in [40].

Use of a random dictionary, especially Gaussian, is in [26], [16], [70], [33], [72], [73]. These works show that reliable determination of L terms from n noisy measurements, does allow L to be of order $n/\log N$, and is achieved either by orthogonal matching pursuit as in [70] or by convex optimization with an ℓ_1 control on the coefficients, though the results of [72], [73] show that the ℓ_1 constrained convex optimization does not perform as well as the information-theoretic limits. Of course, with an average over the ensemble of random dictionaries, the average case error probability over subsets of size L is the same as the average error probability for an specific such subset. Again these random dictionary conclusions translate into saying that the communication rate with procedures based on convex optimization is positive. However, the gaps between constants in the upper and lower bounds (corresponding to achievability and converse results, respectively) correspond to saying that reliability with rates up to capacity is not identified by application of these works. Our work takes it further, identifying practical strategies that do achieve up to the information-theoretic limits.

A caveat in these discussions is that the aim of much (though not all) of the work on sparse signal recovery, compressed sensing, and term selection in linear statistical models is distinct from the purpose of communication alone. In particular rather than the non-zero coefficients being fixed according to a particular power allocation, the aim is to allow a class of such coefficients and still recover their support and estimate the coefficient values. This leads to constants in the converse bounds obtained in Wainwright that are distinct from Shannon capacity, and are not overcome by specialization of his bounds to either the fixed non-zero coefficient value case or to a specific variable coefficient value assignment. Furthermore, the inferiority of the ℓ_1 constrained convex optimization that he demonstrates is also an obstacle to communication at capacity by such schemes. Another source of distinction in the rates one would obtain by apply such results and the capacity rate here is that we only require the sparse superposition decoder to identify most of the columns of the superposition, with the remaining errors corrected using the outer code.

The conclusion of the present paper concerning communication rate in the language of sparse signal recovery and compressed sensing are summarized as follows. A number of terms (columns) selected from a dictionary is linearly combined and subject to Gaussian noise. Suppose a large value B of the order of a polynomial in n is specified for the ratio of the number of variables divided by the number of terms. For signals $X\beta$ satisfying our Gaussian design, high-probability recovery of these terms (with a negligible fraction of mistakes) from the received noisy Y of length n is possible provided the number of terms L satisfies $L \geq Rn/\log B$. Our interest is in the constant R achievable by practical schemes. Accurate recovery is possible provided $R < R_0$ in the equal power allocation case. For the variable power designs we give here, recovery by other means is possible at higher R up to the capacity $\mathcal{C}$.

Though we are not fond of the term, one can call the supremum of the constants R for which a number of terms L of size $Rn/\log B$ be reliable recovered by a practical algorithm, the *compressed sensing capacity*. It can be said that for our designs, the compressed sensing capacity is equal to the Shannon Capacity.

We mention relevant converse results from the theory of joint source and channel coding subject to a Hamming distortion which applied to the problem of recovery of a specified fraction of terms in a sparse linear model in [60]. There they study the case that the ratios n/N and L/N are fixed at constant levels. If the distortion level related to the mistake rate were likewise fixed at a positive constant level, it would follow that the optimal rates would be distinct from the Shannon channel capacity. In our case the $n/N = (\log B)/(RB)$ and $L/N = 1/B$ and the distortion level are all arranged to become small with increasing B, and the Shannon capacity becomes the limiting rate for large B.

As various colleagues have pointed out, relationships of our decoder to other settings can be seen in the problems of multiuser detection [71], term selection by forward selection in regression [29], screening in high-dimensional regression [31], multiple comparison hypothesis testing [12], and the sequential detection algorithm of [55]. In these settings, when dealing with a large number of hypotheses, there is a recent focus on the need to track the fractions of false discoveries and failed detections, especially in iterative algorithms, rather than overall error probability in a one-shot analysis.

Outline of paper: After some preliminaries, section III describes the decoder. In Section IV we analyze the distributions of the various test statistics associated with the algorithm. In particular, the inner product test statistics are shown to decompose into normal random variables plus a nearly constant random shift for the terms sent. Section V demonstrates the increase for each step of the mean separation between the statistics for terms sent and terms not sent. Section VI sets target detection and alarm rates. Reliability of the algorithm is established in section VII, with demonstration of exponentially small error probabilities. Computational illustration is provided in section VIII. A requirement of the theory is that the decoder satisfies a property of accumulation of correct detections. Whether the decoder is accumulative depends on the rate and the power allocation scheme. Specialization of the theory to a particular variable power allocation scheme is presented in section IX. The closeness to capacity is evaluated in section X. Lower bounds on the error exponent are in section XI. Refinements of closeness to capacity are in section XII. Section XIII discusses the use of an outer Reed Solomon code to correct any mistakes from the inner decoder. The appendix collects some auxiliary matters.

II. Some Preliminaries

Notation: For vectors a, b of length n, let $\|a\|^2$ be the sum of squares of coordinates, let $|a|^2 = (1/n)\sum_{i=1}^n a_i^2$ be the average square and let respectively $a^T b$ and $a \cdot b = (1/n)\sum_{i=1}^n a_i b_i$ be the associated inner products. We sometimes find it more convenient to work with $|a|$ and $a \cdot b$.

Setting of Analysis: The dictionary is randomly generated. For the purpose of analysis of average probability of error or average probability of at least certain fraction of mistakes, we investigate properties with respect to the joint distribution of the dictionary and the noise.

The noise ε and the X_j in the dictionary are jointly independent normal random vectors, each of length n, with mean equal to the zero vector and covariance matrixes equal to $\sigma^2 I$ and I, respectively. These vectors have n coordinates indexed by $i = 1, 2, \ldots, n$ which may be called the time index. Meanwhile J is the set of term indices j corresponding to the columns of the dictionary, which may be organized as a union of sections. The codeword sent is from a selection of L terms. The cardinality of J is N and the ratio $B = N/L$.

Corresponding to an input, let $sent = \{j_1, j_2, \ldots, j_L\}$ be the indices of the terms sent and let $other = J - sent$ be the set of indices of all other terms in the dictionary. Component powers P_j are specified, such that $\sum_{j\ sent} P_j = P$.The simplest setting is to arrange these component powers to be equal $P_j = P/L$. Though for best performance, there will be a role for component powers that are different in different portions of the dictionary. The coefficients for the codeword sent are $\beta_j = \sqrt{P_j}\, 1_{j\ sent}$. The received vector is

$$Y \;=\; \sum_j \beta_j X_j \;+\; \varepsilon.$$

Accordingly, X_j and Y are joint normal random vectors, with expected product between coordinates and hence expected inner product $\mathbb{E}[X_j \cdot Y]$ equal to β_j. This expected inner product has magnitude $\sqrt{P_j}$ for the terms sent and 0 for the terms not sent. So the statistics $X_j \cdot Y$ are a source of discrimination between the terms.

We note that each coordinate of Y has expected square $\sigma_Y^2 = P + \sigma^2$ and hence $\mathbb{E}[|Y|^2] = P + \sigma^2$.

Exponential bounds for relative frequencies: In the distributional analysis we shall make repeated use of simple large deviations inequalities. In particular, if $\hat{q}$ is the relative frequency of occurrence of L independent events with success probability q^*, then for $q < q^*$ the probability of the event $\{\hat{q} < q\}$ is not more than the Sanov-Csiszàr bound $e^{-LD(q\|q^*)}$, where the exponent $D(q\|q^*) = D_{Ber}(q\|q^*)$ is the relative entropy between Bernoulli distributions. Some may recognize this bound by the method of types, as in [25],[23], though with a multiplicative factor of $L{+}1$ out front. That it is true without such a multiplier follows from a simple convexity argument as in [24] or by identification that the exponent of the Cramer-Chernoff bound takes the form of the relative entropy. The information theoretic bound subsumes the Hoeffding bound $e^{-2(q^*-q)^2 L}$ via the Csiszàr-Kullback inequality that D exceeds twice the square of total variation, which here is, $D \geq 2(q^* - q)^2$. An extension of the information-theoretic bound to cover weighted combinations of indicators of independent events is in Lemma 47 in the appendix and slight dependence among the events is addressed through bounds on the joint distribution. The role of $\hat{q}$ is played by weighted counts for j in $sent$ of test statistics being above threshold.

In the same manner, one has that if $\hat{p}$ is the relative frequency of occurrence of independent events with success probability p^*, then for $p > p^*$ the probability of the event $\{\hat{p} > p\}$ has a large deviation bound with exponent $D_{Ber}(p\|p^*)$. In our use of such bounds, the role of $\hat{p}$ is played by the relative frequency of false alarms, based on occurrences of j in $other$ of test statistics being above threshold. Naturally, in this case, we arrange for p and p^* both to be small, with some control on the ratio between them. It is convenient to make use of lower bounds on $D_{Ber}(p\|p^*)$, as detailed in Lemma 48 in the appendix, which include what may be called the Poisson bound $p\log p/p^* + p^* - p$ and the Hellinger bound $2(\sqrt{p} - \sqrt{p^*})^2$, both of which exceed $(p - p^*)^2/(2p)$. All three of these lower bounds are superior to the variation bound $2(p - p^*)^2$ when p is small.

III. The Decoder

From the received Y and knowledge of the dictionary, decode which terms were sent by an iterative procedure we now specify more fully.

The first step is as follows. For each term X_j of the dictionary compute the inner product with the received string $X_j^T Y$ as a test statistic and see if it exceeds a threshold $T = \|Y\|\tau$. Denote the associated event

$$\mathcal{H}_j = \{X_j^T Y \geq T\}.$$

In terms of a normalized test statistic this first step test is the same as comparing $\mathcal{Z}_{1,j}$ to a threshold τ, where

$$\mathcal{Z}_{1,j} = X_j^T Y\,/\|Y\|,$$

the distribution of which will be shown to be that of a standard normal plus a shift by a nearly constant amount, where the presence of the shift depends on whether j is one of the terms sent. Thus $\mathcal{H}_j = \{\mathcal{Z}_{1,j} \geq \tau\}$. The threshold is chosen to be

$$\tau = \sqrt{2\log B} + a.$$

The idea of the threshold on the first step is that very few of the terms not sent will be above threshold. Yet a positive fraction of the terms sent, determined by the size of the shift, will be above threshold and hence will be correctly decoded on this first step.

Let $thresh_1 = \{j \in J : 1_{\mathcal{H}_j} = 1\}$ be the set of terms with the test statistic above threshold and let $above_1$ denote the fraction of such terms. In the variable power case it is a weighted fraction $above_1 = \sum_{j \in thresh_1} P_j/P$, weighted by the power P_j. We restrict decoding on the first step to terms in $thresh_1$ so as to avoid false alarms. The decoded set is either taken to be $dec_1 = thresh_1$ or, more generally, a value $pace_1$ is specified and, considering the terms in J in order of decreasing $\mathcal{Z}_{1,j}$, we include in dec_1 as many as we can with $\sum_{j \in dec_1} \pi_j$ not more than $\min\{pace_1, above_1\}$. Let DEC_1 denote the cardinality of the set dec_1.

The output of the first step consists of the set of decoded terms dec_1 and the vector $F_1 = \sum_{j \in dec_1} \sqrt{P_j}\, X_j$ which forms the first part of the fit. The set of terms investigated in step 1 is $J_1 = J$, the set of all columns of the dictionary. Then the set $J_2 = J_1 - dec_1$ remains for second step consideration. In the extremely unlikely event that DEC_1 is already at least L there will be no need for the second step.

A natural way to conduct subsequent steps would be as follows. For the second step compute the residual vector

$$r_2 = Y - F_1.$$

For each of the remaining terms, i.e. terms in J_2, compute the inner product with the vector of residuals, that is, $X_j^T r_2$ or its normalized form $\mathcal{Z}_j^r = X_j^T r_2\,/\|r_2\|$ which may be compared to the same threshold $\tau = \sqrt{2\log B} + a$, leading to a set dec_2 of decoded terms for the second step. Then compute $F_2 = \sum_{j \in dec_2} \sqrt{P_j}\, X_j$, the fit vector for the second step.

The third and subsequent steps would proceed in the same manner as the second step. For any step k, one computes the residual vector

$$r_k = Y - (F_1 + \ldots + F_{k-1}).$$

For terms in $J_k = J_{k-1} - dec_{k-1}$, one gets $thresh_k$ as the set of terms for which $X_j^T r_k\,/\|r_k\|$ is above τ. The set of decoded terms is either taken to be $thresh_k$ or a subset of it. The decoding stops when the size of the cardinality of the set of all decoded term becomes L or there are no terms above threshold in a particular step.

A. Statistics from adaptive orthogonal components:

A variant of the above algorithm from second step onwards is described, which we find to be easier to analyze. The idea is that the ingredients Y, $F_1, \ldots, F_{k-1}$ previously used in forming the residuals may be decomposed into orthogonal components and test statistics formed that entail the best combinations of inner products with these components.

In particular, for the second step the vector G_2 is formed, which is the part of F_1 orthogonal to $G_1 = Y$. For j in J_2, the statistic $\mathcal{Z}_{2,j} = X_j^T G_2/\|G_2\|$ is computed as well as the combined statistic $\mathcal{Z}_{2,j}^{comb} = \sqrt{\lambda_1}\mathcal{Z}_{1,j} - \sqrt{\lambda_2}\mathcal{Z}_{2,j}$, where $\lambda_1 = 1 - \lambda$ and $\lambda_2 = \lambda$, with a value of λ to be specified. What is different on the second step is that now the events $\mathcal{H}_{2,j} = \{\mathcal{Z}_{2,j}^{comb} \geq \tau\}$ are based on these $\mathcal{Z}_{2,j}^{comb}$, which are inner products of X_j with the normalized vector $E_2 = \sqrt{\lambda_1}\, Y/\|Y\| - \sqrt{\lambda_2}\, G_2/\|G_2\|$. To motivate these statistics note the residuals $r_2 = Y - F_1$ may be written as $(1 - \hat{b}_1)Y - G_2$ where $\hat{b}_1 = F_1^T Y/\|Y\|^2$. The statistics we use in the variant may be viewed as approximations to the corresponding statistics based on the normalized residuals $r_2/\|r_2\|$, except that the form of λ and the analysis are simplified.

Again these test statistics $\mathcal{Z}_{2,j}^{comb}$ lead to the set $thresh_2 = \{j \in J_2 : 1_{\mathcal{H}_{2,j}} = 1\}$ of size $above_2 = \sum_{j \in thresh_2} \pi_j$. Considering these statistics in order of decreasing value, it leads to the set dec_2 consisting of as many of these as we can while maintaining $accept_2 \leq \min\{pace_2, above_2\}$, where $accept_2 = \sum_{j \in dec_2} \pi_j$. This provides an additional part of the fit $F_2 = \sum_{j \in dec_2} \sqrt{P_j}\, X_j$.

Proceed in this manner, iteratively, to perform the following loop of calculations, for $k \geq 2$. From the output of step $k-1$, there is available the vector F_{k-1}, which is a part of the fit, and for $k' < k$ there are previously stored vectors $G_{k'}$ and statistics $\mathcal{Z}_{k',j}$. Plus there is a set $dec_{1,k-1} = dec_1 \cup \ldots \cup dec_{k-1}$ already decoded on some previous step and a set $J_k = J - dec_{1,k-1}$ of terms for us to test at step k. Consider, as discussed further below, the part G_k of F_{k-1} orthogonal to the previous $G_{k'}$ and for each j not in dec_{k-1} compute

$$\mathcal{Z}_{k,j} = X_j^T G_k / \|G_k\|$$

and the combined statistic

$$\mathcal{Z}^{comb}_{k,j} = \sqrt{\lambda_{1,k}}\, \mathcal{Z}_{1,j} - \sqrt{\lambda_{2,k}}\, \mathcal{Z}_{2,j} - \ldots - \sqrt{\lambda_{k,k}}\, \mathcal{Z}_{k,j},$$

where these λ will be specified with $\sum_{k'=1}^{k} \lambda_{k',k} = 1$. These positive weights will take the form $\lambda_{k',k} = w_{k'}/s_k$, with $w_1 = 1$, and $s_k = 1 + w_2 + \ldots w_k$, with w_k to be specified. Accordingly, the combined statistic may be computed by the update

$$\mathcal{Z}^{comb}_{k,j} = \sqrt{1-\lambda_k}\, \mathcal{Z}^{comb}_{k-1,j} - \sqrt{\lambda_k}\, \mathcal{Z}_{k,j},$$

where $\lambda_k = w_k/s_k$. This statistic may be thought of as the inner product of X_j with a vector updated as $E_k = \sqrt{1-\lambda_k} E_{k-1} - \sqrt{\lambda_k} G_k/\|G_k\|$, serving as a surrogate for $r_k/\|r_k\|$. For terms j in J_k these statistics $\mathcal{Z}^{comb}_{k,j}$ are compared to a threshold, leading to the events

$$\mathcal{H}_{k,j} = \{\mathcal{Z}^{comb}_{k,j} \geq \tau\}.$$

The idea of these steps is that, as quantified by an analysis of the distribution of the statistics $\mathcal{Z}_{k,j}$, there is an increasing separation between the distribution for terms j sent and the others.

We let $thresh_k = \{j \in J_k : \mathcal{Z}^{comb}_{k,j} \geq \tau_k\}$ and $above_k = \sum_{j \in thresh_k} \pi_j$ and for a specified $pace_k$, considering these test statistics in order of decreasing value, we include in dec_k as many as we can with $accept_k \leq \min\{pace_k, above_k\}$, where $accept_k = \sum_{j \in dec_k} \pi_j$. The output of step k is the vector

$$F_k = \sum_{j \in dec_k} \sqrt{P_j}\, X_j.$$

Also the vector G_k and the statistics $\mathcal{Z}_{k,j}$ are appended to what was previously stored, for all terms not in the decoded set. From this step we provide update to the set of decoded terms $dec_{1,k} = dec_{k-1} \cup dec_k$ and the set $J_{k+1} = J_k - dec_k$ of terms remaining for consideration.

This completes the actions of step k of the loop.

To complete the description of the decoder, we will need to specify the values of w_k that determine the λ_k and we will need to specify $pace_k$. For these specifications there will be a role for measures of the accumulated size of the detection set $accept^{tot}_k = \sum_{k'=1}^{k} accept_{k'}$ as well a target lower bound $q_{1,k}$ on the total weighted fraction of correct detection (the definition of which arises in a later section), and an adjustment to it given by $q^{adj}_{1,k} = q_{1,k}/(1+f_{1,k}/q_{1,k})$ where $f_{1,k}$ is a target upper bound on the total weighted fraction of false alarms. The choices we consider take $w_k = s_k - s_{k-1}$ to be increments of the sequence $s_k = 1/(1 - x_{k-1}\nu)$ that arises in characterizing the above mentioned separation. In the definition of w_k we take x_{k-1} as either $accept^{tot}_{k-1}$ or $q^{adj}_{1,k-1}$, both of which arise as surrogates to a corresponding unobservable quantity which would require knowledge of the actual fraction of correct detection through step $k-1$.

There are two options for $pace_k$ that we describe. First, we may arrange for dec_k to be all of $thresh_k$ by setting $pace_k = 1$, large enough that it has essentially no role and $dec_k = thresh_k$, and with this option we set w_k as above using $x_{k-1} = accept^{tot}_{k-1}$. This choice yields a successful growth of the total weighted fractions of correct detections, though to handle the empirical character of w_k there is a slight cost to it in the reliability bound, not present with the second option.

For the second option, we may let $pace_k = q^{adj}_{1,k} - q^{adj}_{1,k-1}$ be the deterministic increments of the increasing sequence $q^{adj}_{1,k}$, with which it is shown that $above_k$ is likely to exceed $pace_k$, for each k. When it does then $accept_k$ equals the value $pace_k$, and cumulatively their sum $accept^{tot}_k$ matches the target $q^{adj}_{1,k}$. Likewise, for this option, w_k is set using $x_{k-1} = q^{adj}_{1,k-1}$. It's deterministic trajectory facilitates the demonstration of reliability of the decoder.

On each step k we decode a substantial part of what remains, because of growth of the mean separation between terms sent and the others, as we shall see.

The algorithm stops under the following conditions. Natural practical conditions are that L terms have been decoded, or that the weighted total size of the decoded set $accept^{tot}_k$ has reached at least 1, or that no terms from J_k are found to have statistic above threshold, so that F_k is zero and the statistics would remain thereafter unchanged. An analytical condition is the lower bound we obtain on the likely mean separation stops growing (captured through $q^{adj}_{1,k}$ no longer increasing), so that no further improvement is theoretically demonstrable by such methodology. Subject to rate constraints near capacity, our best bounds occur with a total number of steps m equal to an integer part of $2 + snr \log B$.

Up to step k, the total set of decoded terms is $dec_{1,k}$, and the corresponding fit fit_k may be represented either as $\sum_{j \in dec_{1,k}} \sqrt{P_j}\, X_j$ or as the sum of the pieces from each step

$$fit_k = F_1 + F_2 + \ldots + F_k.$$

As to the part G_k of F_{k-1} orthogonal to $G_{k'}$ for $k' < k$, we take advantage of two ways to view it, one emphasizing computation and the other analysis.

For computation, work directly with parts of the fit. The $G_1, G_2, \ldots, G_{k-1}$ are orthogonal vectors, so the parts of F_{k-1} in these directions are $\hat{b}_{k,k'} G_{k'}$ with coefficients $\hat{b}_{k,k'} = F^T_{k-1} G_{k'}/\|G_{k'}\|^2$ for $k' = 1, 2, \ldots, k-1$, where if peculiarly $\|G_{k'}\| = 0$ we use $\hat{b}_{k,k'} = 0$. Accordingly, the new G_k may be computed from F_{k-1} and the previous $G_{k'}$ with $k' < k$ by

$$G_k \;=\; F_{k-1} \,-\, \sum_{k'=1}^{k-1} \hat{b}_{k,k'}\, G_{k'}.$$

This computation entails the $n-$fold sums of products $F_k^T G_{k'}$ for determination of the $\hat{b}_{k,k'}$. Then from this computed G_k we obtain the inner products with the X_j to yield $\mathcal{Z}_{k,j} = X_j^T G_k / \|G_k\|$ for j in J_k.

The algorithm is seen to perform an *adaptive* Gram-Schmidt orthogonalization, creating orthogonal vectors G_k used in representation of the X_j and linear combinations of them, in directions suitable for extracting statistics of appropriate discriminatory power, starting from the received Y. For the classical Gram-Schmidt process, one has a pre-specified set of vectors which are successively orthogonalized, at each step, by finding the part of the current vector that is orthogonal to the previous vectors. Here instead, for each step, the vector F_{k-1}, for which one finds the part G_k orthogonal to the vectors $G_1, \ldots, G_{k-1}$, is not pre-specified. Rather, it arises from thresholding statistics extracted in creating these vectors.

For analysis, look at what happens to the representation of the individual terms. Each term X_j for $j \in J_{k-1}$ has the decomposition

$$X_j = \mathcal{Z}_{1,j} \frac{G_1}{\|G_1\|} + \mathcal{Z}_{2,j} \frac{G_2}{\|G_2\|} + \ldots + \mathcal{Z}_{k-1,j} \frac{G_{k-1}}{\|G_{k-1}\|} + V_{k,j},$$

where $V_{k,j}$ is the part of X_j orthogonal to $G_1, G_2, \ldots, G_{k-1}$. Since $F_{k-1} = \sum_{j \in dec_{k-1}} \sqrt{P_j}\, X_j$ it follows that G_k has the representation

$$G_k = \sum_{j \in dec_{k-1}} \sqrt{P_j}\, V_{k,j},$$

from which $\mathcal{Z}_{k,j} = V_{k,j}^T G_k \,/ \|G_k\|$, and we have the updated representation

$$X_j = \mathcal{Z}_{1,j} \frac{G_1}{\|G_1\|} + \ldots + \mathcal{Z}_{k-1,j} \frac{G_{k-1}}{\|G_{k-1}\|} + \mathcal{Z}_{k,j} \frac{G_k}{\|G_k\|} + V_{k+1,j}.$$

With the initialization $V_{0,j} = X_j$, these $V_{k+1,j}$ may be thought of as iteratively obtained from the corresponding vectors at the previous step, that is,

$$V_{k+1,j} \;=\; V_{k,j} - \mathcal{Z}_{k,j}\, G_k / \|G_k\|.$$

These V do not actually need to be computed, nor do we need to compute its components detailed below, but we do use this representation of the terms X_j in obtaining distributional properties of the $\mathcal{Z}_{k,j}$.

B. The weighted fractions of detections and alarms:

The weights $\pi_j = P_j / P$ sum to 1 across j in $sent$ and they sum to $B-1$ across j in $other$. Define in general

$$\hat{q}_k = \sum_{j \in sent \cap dec_k} \pi_j$$

for the step k correct detections and

$$\hat{f}_k = \sum_{j \in other \cap dec_k} \pi_j$$

for the false alarms. In the case $P_j = P/L$ which assigns equal weight $\pi_j = 1/L$, then $\hat{q}_k\, L$ is the increment to the number of correct detections on step k, likewise $\hat{f}_k\, L$ is the increment to the number of false alarms. Their sum $accept_k = \hat{q}_k + \hat{f}_k$ matches $\sum_{j \in dec_k} \pi_j$.

The total weighted fraction of correct detections up to step k is $\hat{q}_k^{tot} = \sum_{j \in sent \cap dec_{1,k}} \pi_j$ which may be written as the sum

$$\hat{q}_k^{tot} = \hat{q}_1 + \hat{q}_2 + \ldots + \hat{q}_k.$$

Assume for now that $dec_k = thresh_k$. Then these increments $\hat{q}_k$ equal $\sum_{j \in sent \cap J_k} \pi_j 1_{\mathcal{H}_{k,j}}$.

The decoder only encounters these $\mathcal{H}_{k,j} = \{\mathcal{Z}_{k,j}^{comb} > \tau\}$ for j not decoded on previous steps, i.e., for j in $J_k = (dec_{1,k-1})^c$. For each step k, one may define the statistics arbitrarily for j in $dec_{1,k-1}$, so as to fill out definition of the events $\mathcal{H}_{k,j}$ for each j, in a manner convenient for analysis. By induction on k, on sees that $dec_{1,k}$ consists of the terms j for which the union event $\mathcal{H}_{1,j} \cup \ldots \cup \mathcal{H}_{k,j}$ occurs. Because if $dec_{1,k-1} = \{j : 1_{\mathcal{H}_{1,j} \cup \ldots \cup \mathcal{H}_{k-1,j}} = 1\}$ then the decoded set $dec_{1,k}$ consists of terms for which either $\mathcal{H}_{1,j} \cup \ldots \cup \mathcal{H}_{k-1,j}$ occurs (previously decoded) or $\mathcal{H}_{k,j} \cap [\mathcal{H}_{1,j} \cup \ldots \cup \mathcal{H}_{k-1,j}]^c$ occurs (newly decoded), and together these events constitute the union $\mathcal{H}_{1,j} \cup \ldots \cup \mathcal{H}_{k,j}$.

Accordingly, the total weighted fraction of correct detections $\hat{q}_k^{tot}$ may be regarded as the same as the π weighted measure of the union

$$\hat{q}_k^{tot} = \sum_{j\ sent} \pi_j 1_{\{\mathcal{H}_{1,j} \cup \ldots \cup \mathcal{H}_{k,j}\}}.$$

Indeed, to relate this expression to the preceding expression for $\hat{q}_k^{tot}$, the sum for k' from 1 to k corresponds to the representation of the union as the disjoint union of contributions from terms sent that are in $\mathcal{H}_{k',j}$ but not in earlier such events.

Likewise the weighted count of false alarms $\hat{f}_k^{tot} = \sum_{j \in other \cap dec_{1,k}} \pi_j$ may be written as

$$\hat{f}_k^{tot} = \hat{f}_1 + \hat{f}_2 + \ldots + \hat{f}_k$$

which when $dec_k = thresh_k$ may be expressed as

$$\hat{f}_k^{tot} = \sum_{j\ other} \pi_j 1_{\{\mathcal{H}_{1,j} \cup \ldots \cup \mathcal{H}_{k,j}\}}.$$

In the distributional analysis that follows we see that the mean separation is given by an expression inversely related to $1 - \hat{q}_{k-1}^{tot} \nu$. The idea of the multi-step algorithm is to accumulate enough correct detections in $\hat{q}_k^{tot}$, with an attendant low number of false alarms, that the fraction that remains becomes small enough, and the mean separation hence pushed large enough, that most of what remains is reliably decoded on the last step.

The analysis will provide, for each section ℓ, lower bounds on the probability that the correct term is above threshold by step k and upper-bounds on the accumulated false alarms. When the snr is low and we are using constant power allocation, these probabilities are the same across the sections, all of which remain active for consideration until completion of the steps.

For variable power allocation, with $P_{(\ell)}$ decreasing in ℓ, then for each step k, the probability that the correct term is above threshold varies with ℓ. Nevertheless, it can be a rather large number of sections for which this probability takes an intermediate value (neither small nor close to one), thereby necessitating the adaptive decoding. Most of our analysis proceeds by allowing at each step for terms to be detected from any section $\ell = 1, 2, \ldots, L$.

C. An optional analysis window:

For large $\mathcal{C}$, the $P_{(\ell)}$ proportional to $e^{-2\mathcal{C}\ell/L}$ exhibits a strong decay with increasing ℓ. Then it can be appropriate to take advantage of a deterministic decomposition into three sets of sections at any given number of steps. There is the set of sections with small ℓ, which we call polished, where the probability of the correct term above threshold before step k is already sufficiently close to one that it is known in advance that it will not be necessary to continue to check these (as the subsequent false alarm probability would be quantified as larger than the small remaining improvement to correct detection probability for that section). Let $polished_k$ (initially empty) be the set of terms in these sections. With the power decreasing, this coincides with a non-decreasing initial interval of sections.

Likewise there are the sections with large ℓ where the probability of a correct detection on step k is less than the probability of false alarm, so it would be advantageous to still leave them untested. Let $untested_k$ (desirably eventually empty) be the set of terms from these sections, corresponding to a decreasing tail interval of sections up to the last section L.

The complement is a middle region of terms

$$potential_k = J - polished_k - untested_k,$$

corresponding to a window of sections, $\text{left}_k \le \ell \le \text{right}_k$, worthy of attention in analyzing the performance at step k. For each term in this analysis window there is a reasonable chance (neither too high nor too low) of it being decoded by the completion of this step.

These middle regions overlap across k, so that for any term j has potential for being decoded in several steps.

In any particular realization of X, Y, some terms in this set $potential_k$ are already in $dec_{1,k-1}$. Accordingly, one has the option at step k to restrict the active set of the search to $J_k = potential_k \cap dec_{1,k-1}^c$ rather than searching all of the set $dec_{1,k-1}^c$ not previously decoded. In this case one modifies the definitions of $\hat{q}_k^{tot}$ and $\hat{f}_k^{tot}$, to be

$$\hat{q}_k^{tot} = \sum_{j\ sent} \pi_j 1_{\{\cup_{k' \in K_{j,k}} \mathcal{H}_{k',j}\}}$$

and

$$\hat{f}_k^{tot} = \sum_{j\ other} \pi_j 1_{\{\cup_{k' \in K_{j,k}} \mathcal{H}_{k',j}\}}$$

where

$$K_{j,k} = \{k' \le k : j \in potential_{k'}\}.$$

A refined analysis given later quantifies benefits of this restriction, particularly concerning improved bounds on the total false alarms and corresponding improvement to the rate drop from capacity, when $\mathcal{C}$ is large.

IV. Distributional Analysis

In this section we describe the distributional properties of the random variables $\mathcal{Z}_k = (\mathcal{Z}_{k,j} : j \in J_k)$ for each $k = 1, 2, \ldots, n$. In particular we show for each k that $\mathcal{Z}_{k,j}$ are location shifted normal random variables with variance near one for $j \in sent \cap J_k$ and are independent standard normal random variables for $j \in other \cap J_k$.

In Lemma 1 below we derive the distributional properties of $\mathcal{Z}_1$. Lemma 2 characterizes the distribution of $\mathcal{Z}_k$ for steps $k \ge 2$.

Before providing these lemmas we define a few quantities which will be helpful in studying the location shifts of $\mathcal{Z}_{k,j}$ for $j \in sent \cap J_k$. In particular, define the quantity

$$C_{j,R} = \pi_j \, L\nu/(2R),$$

where $\pi_j = P_j/P$ and $\nu = \nu_1 = P/(\sigma^2 + P)$. Likewise define

$$C_{j,R,B} = (C_{j,R}) \, 2 \log B,$$

which also has the representation

$$C_{j,R,B} = n \, \pi_j \, \nu.$$

The role of this quantity as developed below is via the location shift $\sqrt{C_{j,R,B}}$ seen to be near $\sqrt{C_{j,R}}\,\tau$. One compares this value to τ, that is, one compares $C_{j,R}$ to 1 to see when there is a reasonable probability of some correct detections starting at step 1, and one arranges $C_{j,R}$ to taper not too rapidly to allow decodings to accumulate on successive steps.

We have two illustrative cases. For the constant power allocation case, π_j equals $1/L$ and $C_{j,R}$ reduces to

$$C_{j,R} \; = \; R_0/R,$$

where $R_0 = (1/2)P/(\sigma^2 + P)$. In this case $C_{j,R,B} = (R_0/R) \, 2 \log B$ are equal for all j. This $C_{j,R}$ is at least 1 when the rate R is not more than R_0.

For the case of power P_j proportional to $e^{-2\mathcal{C}\ell/L}$, we have $\pi_j = e^{-2\mathcal{C}(\ell-1)/L}(1 - e^{-2\mathcal{C}/L})/(1 - e^{-2\mathcal{C}})$ for each j in section ℓ, for ℓ from 1 to L. Define

$$\tilde{\mathcal{C}} = (L/2)[1 - e^{-2\mathcal{C}/L}],$$

which is essentially identical to $\mathcal{C}$, for L large compared to $\mathcal{C}$. Then for j in section ℓ we have that

$$\pi_j = (2\tilde{\mathcal{C}}/L\nu) e^{-2\mathcal{C}(\ell-1)/L}$$

and

$$C_{j,R} \; = (\tilde{\mathcal{C}}/R) \, e^{-2\mathcal{C}(\ell-1)/L}.$$

For rates R not more than $\mathcal{C}$, this $C_{j,R}$ is at least 1 in some sections, leading to likelihood of some initial successes, and it tapers at the fastest rate at which we can still accumulate decoding successes.

A. Distributional analysis of the first step:

We now are in a position to give the lemma for the distribution of $\mathcal{Z}_1$. Recall that $J_1 = J$ is the set of all N indices.

***Lemma** 1:* For each $j \in J$, the statistic $\mathcal{Z}_{1,j}$ can be represented as

$$\sqrt{C_{j,R,B}}\,[\mathcal{X}_n/\sqrt{n}]1_{j\,sent} + Z_{1,j},$$

where $Z_1 = (Z_{1,j} : j \in J_1)$ is multivariate normal $N(0, \Sigma_1)$ and $\mathcal{X}_n^2 = \|Y\|^2/\sigma_Y^2$ is a Chi-square (n) random variable that is independent of Z_1. Here recall that $\sigma_Y^2 = P + \sigma^2$ is the variance of each coordinate of Y.

The covariance matrix Σ_1 can be expressed as $\Sigma_1 = I - b_1 b_1^T$, where b_1 is the vector with entries $b_{1,j} = \beta_j/\sigma_Y$ for j in J.

The subscript 1 on the matrix Σ_1 and the vector b_1 are to distinguish these first step quantities from those that arise on subsequent steps.

Proof of Lemma 1: Recall that the X_j for j in J are independent $N(0, I)$ random vectors and that $Y = \sum_j \beta_j X_j + \varepsilon$, where the sum of squares of the β_j is equal to P.

Consider the decomposition of each random vector X_j of the dictionary into a vector in the direction of the received Y and a vector U_j uncorrelated with Y. That is, one considers the reverse regression

$$X_j = b_{1,j}\,Y/\sigma_Y \,+\, U_j,$$

where the coefficient is $b_{1,j} = \mathbb{E}[X_{i,j}Y_i]/\sigma_Y = \beta_j/\sigma_Y$, which indeed makes each coordinate of U_j uncorrelated with each coordinate of Y. These coefficients collect into a vector $b_1 = \beta/\sigma_Y$ in $\mathbb{R}^N$.

These vectors $U_j = X_j - b_{1,j}Y/\sigma_Y$ along with Y are linear combinations of joint normal random variables and so are also joint normal, with zero correlation implying that Y is independent of the collection of U_j. The independence of Y and U_j facilitates development of distributional properties of the $U_j^T Y$. For these purposes we need the characteristics of the joint distribution of the U_j across terms j (clearly there is independence for distinct time indices i).

The coordinates of U_j and $U_{j'}$ have mean zero and expected product $1_{\{j=j'}\} - b_{1,j}b_{1,j'}$. These covariances $(\mathbb{E}[U_{i,j}U_{j,j'}] : j, j' \in J)$ organize into a matrix

$$\Sigma_1 = \Sigma = I - \Delta = I - bb^T.$$

For any constant vector $\alpha \neq 0$, consider $U_j^T\alpha/\|\alpha\|$. Its joint normal distribution across terms j is the same for any such α. Specifically, it is a normal $N(0, \Sigma)$, with mean zero and the indicated covariances.

Likewise define the random variables $Z_j = U_j^T Y/\|Y\|$, also denoted $Z_{1,j}$ when making explicit that it is for the first step. Jointly across j, these Z_j have the normal $N(0, \Sigma)$ distribution, independent of Y. Indeed, since the U_j are independent of Y, when we condition on $Y = \alpha$ we get the same $N(0, \Sigma)$ distribution, and since this conditional distribution does not depend on Y, it is the unconditional distribution as well.

Where this gets us is revealed via the representation of the inner product $X_j^T Y$ as $b_{1,j}\|Y\|^2/\sigma_Y + U_j^T Y$, which can be written as

$$X_j^T Y \,=\, \beta_j\,\frac{\|Y\|^2}{\sigma_Y^2} \,+\, \|Y\|\,Z_j.$$

This identifies the distribution of the $X_j^T Y$ as that obtained as a mixture of the normal Z_j with scale and location shifts determined by an independent random variable $\mathcal{X}_n^2 = \|Y\|^2/\sigma_Y^2$, distributed as Chi-square with n degrees of freedom.

Divide through by $\|Y\|$ to normalize these inner products to a helpful scale and to simplify the distribution of the result to be only that of a location mixture of normals. The resulting random variables $\mathcal{Z}_{1,j} = X_j^T Y\,/\|Y\|$ take the form

$$\mathcal{Z}_{1,j} \,=\, \sqrt{n}\,b_{1,j}\,|Y|/\sigma_Y \,+\, Z_j,$$

where $|Y|/\sigma_Y = \mathcal{X}_n/\sqrt{n}$ is near 1. Note that $\sqrt{n}b_{1,j} = \sqrt{n}\beta_j/\sigma_Y$ which is $\sqrt{n\pi_j\nu}$ or $\sqrt{C_{j,R,B}}$. This completes the proof of Lemma 1.

The above proof used the population reverse regression of X_j onto Y, in which the coefficient $b_{1,j}$ arises as a ratio of expected products. There is also a role for the empirical projection decomposition, the first step of which is $X_j = \mathcal{Z}_{1,j}Y/\|Y\| + V_{2,j}$, with $G_1 = Y$. Its additional steps provide the basis for additional distributional analysis.

B. Distributional analysis of steps $k \geq 2$:

Let $V_{k,j}$ be the part of X_j orthogonal to $G_1, G_2, \ldots, G_{k-1}$, from which $G_k = \sum_{j \in dec_{k-1}} \sqrt{P_j}\,V_{k,j}$. It yields the representation of the statistic $\mathcal{Z}_{k,j} = X_j^T G_k/\|G_k\|$ as $V_{k,j}^T G_k/\|G_k\|$, as we have said. Amongst other matters, the proof of the following lemma determines, for $j \in J_k$, the ingredients of the regression $V_{k,j} = b_{k,j}G_k/\sigma_k + U_{k,j}$ in which $U_{k,j}$ is found to be a mean zero normal random vector independent of G_k, conditioning on certain statistics from previous steps. Taking the inner product with the unit vector $G_k/\|G_k\|$ yields a representation of $\mathcal{Z}_{k,j}$ as a mean zero normal random variable $Z_{k,j}$ plus a location shift that is a multiple of $\|G_k\|$ depending on whether j is in $sent$ or not. The definition of $Z_{k,j}$ is $U_{k,j}^T G_k/\|G_k\|$.

We maintain the pattern used in Lemma 1 and use the caligraphic font $\mathcal{Z}_{k,j}$ to denote the test statistics that incorporate the shift for j in $sent$ and the standard font $Z_{k,j}$ to denote their counterpart mean zero normal random variables before the shift.

The lemma below characterizes the sequence of conditional distributions of the $Z_k = (Z_{k,j} : j \in J_k)$ and $\|G_k\|$, given $\mathcal{F}_{k-1}$, for $k = 1, 2, \ldots n$, where

$$\mathcal{F}_{k-1} = (\|G_{k'}\|, Z_{k'} : k' = 1, \ldots, k-1).$$

This determines also the distribution of $\mathcal{Z}_k = (\mathcal{Z}_{k,j} : j \in J_k)$ conditional on $\mathcal{F}_{k-1}$. Initializing with the distribution of $\mathcal{Z}_1$ derived in Lemma 1, we provide the conditional distributions for all $2 \leq k \leq n$. The algorithm will be arranged to stop long before n, so we will only need these up to some much smaller final $k = m$. Note that J_k is never empty because we decode at most L, so there must always be at least

$(B-1)L$ remaining. For an index set which may depend on the conditioning variables, we let $N_{J_k}(0,\Sigma)$ denote a mean zero multivariate normal distribution with index set J_k and the indicated covariance matrix.

***Lemma** 2:* For $k \geq 2$, given $\mathcal{F}_{k-1}$, the conditional distribution $\mathbb{P}_{Z_{k,J_k}|\mathcal{F}_{k-1}}$ of $Z_{k,J_k} = (Z_{k,j} : j \in J_k)$ is normal $N_{J_k}(0,\Sigma_k)$; the random variable $\mathcal{X}^2_{d_k} = \|G_k\|^2/\sigma_k^2$ is a Chi-square distributed, with $d_k = n-k+1$ degrees of freedom, conditionally independent of the Z_k, where σ_k^2 depends on $\mathcal{F}_{k-1}$ and is strictly positive provided there was at least one term above threshold on step $k-1$; and, moreover, $\mathcal{Z}_{k,j}$ has the representation

$$-\sqrt{\hat{w}_k\, C_{j,R,B}}\,\left[\mathcal{X}_{d_k}/\sqrt{n}\right] 1_{j\ sent} \,+\, Z_{k,j}.$$

The shift factor $\hat{w}_k$ is the increment $\hat{w}_k = \hat{s}_k - \hat{s}_{k-1}$, of the series $\hat{s}_k$ with

$$1+\hat{w}_2+\ldots+\hat{w}_k \,=\, \hat{s}_k = \frac{1}{1-(\hat{q}_1^{adj}+\ldots+\hat{q}_{k-1}^{adj})\,\nu}$$

where $\hat{q}_j^{adj} = \hat{q}_j/(1+\hat{f}_j/\hat{q}_j)$, determined from weighted fractions of correct detections and false alarms on previous steps. Here $\hat{s}_1 = \hat{w}_1 = 1$. The $\hat{w}_k$ is strictly positive, that is, $\hat{s}_k$ is increasing, as long as $\hat{q}_{k-1} > 0$, that is, as long as the preceding step had at least one correct term above threshold. The covariance Σ_k has the representation

$$\Sigma_k = I - \delta_k\delta_k^T = I - \nu_k\,\beta\beta^T/P$$

where $\nu_k = \hat{s}_k\,\nu$. $(\Sigma_k)_{j,j'} = 1_{j=j'} - \delta_{k,j}\delta_{k,j'}$, for j,j' in J_k, where the vector δ_k is in the direction β, with $\delta_{k,j} = \sqrt{\nu_k P_j/P}\,1_{j\ sent}$ for j in J_k. Finally,

$$\sigma_k^2 = \frac{\hat{s}_{k-1}}{\hat{s}_k}\,accept_{k-1}P$$

where $accept_k = \sum_{j\in dec_k}\pi_j$ is the size of the decoded set on step k.

The proof of this lemma follows the same pattern as the proof of Lemma 1 with some additional ingredients. We put it in Appendix I.

C. The nearby distribution:

Two joint probability measures $\mathbb{Q}$ and $\mathbb{P}$ are now specified for all the $Z_{k,j}$, $j \in J$ and the $\|G_k\|$ for $k = 1,\ldots m$. For $\mathbb{P}$, it is to have the conditionals $\mathbb{P}_{Z_{k,J_k}|\mathcal{F}_{k-1}}$ specified above.

The $\mathbb{Q}$ is the approximating distribution. We choose $\mathbb{Q}$ to make all the $Z_{k,j}$, for $j \in J$, for $k = 1,2,\ldots,m$, be independent standard normal, and like $\mathbb{P}$, we choose $\mathbb{Q}$ to make the $\mathcal{X}^2_{n-k+1} = \|G_k\|^2/\sigma_k^2$ be independent Chi-square$(n-k+1)$ random variables.

Fill out of specification of the distribution assigned by $\mathbb{P}$, via a sequence of conditionals $\mathbb{P}_{Z_{k,J}|\mathcal{F}^{full}_{k-1}}$ for $Z_{k,J} = (Z_{k,j} : j \in J)$, which is for all j in J, not just for j in J_k. Here $\mathcal{F}^{full}_k = (\|G_{k'}\|, Z_{k',J} : k' = 1,2,\ldots,k)$. For the variables Z_{k,J_k} that we actually use, the conditional distribution is that of $\mathbb{P}_{Z_{k,J_k}|\mathcal{F}_{k-1}}$ as specified in the above Lemma. Whereas for the $Z_{k,j}$ with j in the already decoded set $J - J_k = dec_{1,k-1}$, given $\mathcal{F}_{k-1}$, we conveniently arrange them to have the same independent standard normal as is used by $\mathbb{Q}$. This completes the definition of the $Z_{k,j}$ for all j, and with it one likewise extends the definition of $\mathcal{Z}_{k,j}$ as a function of $Z_{k,j}$ and $\|G_k\|$ and completes the definition of the events $\mathcal{H}_{k,j}$ for all j, used in our analysis.

This choice of independent standard normal for the distribution of $Z_{k,j}$ given $\mathcal{F}_{k-1}$ for j in $dec_{1,k-1}$, is contrary to what would have arisen in the proof of 2 from the inner product of $U_{k,j}$ with $G_k/\|G_k\|$ if there one were to have looked there at such j with $1_{\mathcal{H}_{k',j}} = 1$ for earlier $k' < k$. Nevertheless, as we have said, we have freedom of choice of the distribution of these variables not used by the decoder. The present choice is a simpler extension providing a conditional distribution of $(Z_{k,j} : j \in J)$ that shares the same marginalization to the true distribution of $(Z_{k,j} : j \in J_k)$ given $\mathcal{F}_{k-1}$.

An event A is said to be determined by $\mathcal{F}_k$ if its indicator is a function of $\mathcal{F}_k$. As $\mathcal{F}_k = (\mathcal{X}_{n-k'+1}, Z_{k',J_{k'}} : k' \leq k)$, with a random index set J_k given as a function of preceding $\mathcal{F}_{k-1}$, it might be regarded as a tricky matter. Alternatively a random variable may be said to be determined by $\mathcal{F}_k$ if it is measurable with respect to the collection of random variables $(\|G_{k'}\|\,, Z_{k',j}1_{\{j\in dec^c_{1,k'-1}\}}, j \in J,\, 1 \leq k' \leq k)$. The multiplication by the indicator removes the effect on step k' of any $Z_{k',j}$ decoded on earlier steps, that is, any j outside $J_{k'}$. Operationally, no advanced measure-theoretic notions are required, as we are working with sequences of conditional densities of explicit Gaussian form.

In the following lemma we appeal to a sense of closeness of the distribution $\mathbb{P}$ to $\mathbb{Q}$, such that events exponentially unlikely under $\mathbb{Q}$ remain exponentially unlikely under the governing measure $\mathbb{P}$.

***Lemma** 3:* For any event A determined by $\mathcal{F}_k$,

$$\mathbb{P}[A] \leq \mathbb{Q}[A]e^{kc_0},$$

where $c_0 = (1/2)\log(1+P/\sigma^2)$. The analogous statement holds more generally for the expectation of any non-negative function of $\mathcal{F}_k$.

See Appendix II for the proof. The fact that c_0 matches the capacity $\mathcal{C}$ might be interesting, but it is not consequential to our argument. What matters for us is simply that if $\mathbb{Q}[A]$ is exponentially small in L or n, then so is $\mathbb{P}[A]$.

D. Logic in bounding detections and false alarms:

Simple logic concerning unions plays an important simplifying role in our analysis to lower bound detection rates and to upper bound false alarms. The idea is to avoid the distributional complication of sums restricted to terms not previously above threshold.

Here assume that $dec_k = thresh_k$ each step. Section VII-B discusses an alternative approach where we take dec_k to be a particular subset of $thresh_k$, to demonstrate slightly better reliability bounds for given rates below capacity.

Recall that with $\hat{q}_k = \sum_{j\ sent\cap J_k}\pi_j 1_{\mathcal{H}_{k,j}}$ as the increment of weighted fraction of correct detections, the total weighted fraction of correct detections $\hat{q}_k^{tot} = \hat{q}_1 + \ldots + \hat{q}_k$ up to step k is the same as the the weighted fraction of the union $\sum_{j\ sent} \pi_j 1_{\mathcal{H}_{1,j}\cup\ldots\cup\mathcal{H}_{k,j}}$. Accordingly, it has the lower bound

$$\hat{q}_k^{tot} \geq \sum_{j\ sent} \pi_j 1_{\mathcal{H}_{k,j}}$$

based solely on the step k half-spaces, where the sum on the right is over all j in $sent$, not just those in $sent \cap J_k$. That this simpler form will be an effective lower bound on $\hat{q}_k^{tot}$ will arise from the fact that the statistic tested in $\mathcal{H}_{k,j}$ is approximately a normal with a larger mean at step k than at steps $k' < k$, producing for all j in $sent$ greater likelihood of occurrence of $\mathcal{H}_{k,j}$ than earlier $\mathcal{H}_{k',j}$.

Concerning this lower bound $\sum_{j\ sent} \pi_j 1_{\mathcal{H}_{k,j}}$, in what follows we find it convenient to set $\hat{q}_{1,k}$ to be the corresponding sum $\sum_{j\ sent} \pi_j 1_{H_{k,j}}$ using a simpler purified form $H_{k,j}$ in place of $\mathcal{H}_{k,j}$. Outside of an exception event we study, this $H_{k,j}$ is a smaller set that $\mathcal{H}_{k,j}$ and so then $\hat{q}_k^{tot}$ is at least $\hat{q}_{1,k}$.

Meanwhile, with $\hat{f}_k = \sum_{j \in other \cap J_k} \pi_j 1_{\mathcal{H}_{k,j}}$ as the increment of weighted count of false alarms, as we have seen, the total weighted count of false alarms $\hat{f}_k^{tot} = \hat{f}_1 + \ldots + \hat{f}_k$ is the same as $\sum_{j\ other} \pi_j 1_{\mathcal{H}_{1,j}\cup\ldots\cup\mathcal{H}_{k,j}}$. It has the upper bound

$$\hat{f}_k^{tot} \leq \sum_{j\ other} \pi_j 1_{\mathcal{H}_{1,j}} + \ldots + \sum_{j\ other} \pi_j 1_{\mathcal{H}_{k,j}}.$$

We denote the right side of this bound $\hat{f}_{1,k}$.

These simple inequalities permit our aim to establish likely levels of correct detections and false alarm bounds to be accomplished by analyzing the simpler forms $\sum_{j\ sent} \pi_j 1_{\mathcal{H}_{k,j}}$ and $\sum_{j\ other} \pi_j 1_{\mathcal{H}_{k,j}}$ without the restriction to the random set J_k, which would complicate the analysis.

Refinement using wedges: Rather than using the last half-space $\mathcal{H}_{k,j}$ alone, one may obtain a lower bound on the indicator of the union $\mathcal{H}_{1,j} \cup \ldots \cup \mathcal{H}_{k,j}$ by noting that it contains $\mathcal{H}_{k-1,j} \cup \mathcal{H}_{k,j}$ expressed as the disjoint union of the events $\mathcal{H}_{k,j}$ and $\mathcal{H}_{k-1,j} \cap \mathcal{H}_{k,j}$. The latter event may be interpreted as a wedge (an intersection of two half-spaces) in terms of the pair of random variables $\mathcal{Z}_{k-1,j}^{comb}$ and $\mathcal{Z}_{k,j}$. Accordingly, there is the refined lower bound on $\hat{q}_k^{tot} = \sum_{j\ sent} \pi_j 1_{\mathcal{H}_{1,j}\cup\ldots\cup\mathcal{H}_{k,j}}$, given by

$$\hat{q}_k^{tot} \geq \sum_{j\ sent} \pi_j 1_{\mathcal{H}_{k,j}} + \sum_{j\ sent} \pi_j 1_{\mathcal{H}_{k-1,j}\cap\mathcal{H}_{k,j}^c}.$$

With this refinement a slightly improved bound on the likely fraction of correct detections can be computed from determination of lower bounds on the wedge probabilities. One could introduce additional terms from intersection of three or more half-spaces, but it is believed that these will have negligible effect.

Likewise, for the false alarms, the union $\mathcal{H}_{1,j} \cup \ldots \cup \mathcal{H}_{k,j}$ expressed as the disjoint union of $\mathcal{H}_{k,j}$, $\mathcal{H}_{k-1,j} \cap \mathcal{H}_{k,j}^c$, $\ldots$, $\mathcal{H}_{1,j} \cap \mathcal{H}_{2,j}^c \cap \ldots \cap \mathcal{H}_{k,j}^c$, has the improved upper-bound for its indicator given by the sum

$$1_{\mathcal{H}_{k,j}} + 1_{\mathcal{H}_{k-1,j}\cap\mathcal{H}_{k,j}^c} + \ldots + 1_{\mathcal{H}_{1,j}\cap\mathcal{H}_{2,j}^c}$$

given by just one half-space indicator and $k-1$ wedge indicators. Accordingly, the weighted total fraction of false alarms $\hat{f}_k^{tot}$ is upper-bounded by the π weighted sum of these indicators for j in $other$. This leads to improved bounds on the likely fraction of false alarms from determination of upper bounds on wedge probabilities.

Accounting with the optional analysis window: In the optional restriction to terms in the set $pot_k = potential_k$ for each step, the $\hat{q}_k$ take the same form but with $J_k = pot_k \cap dec_{1,k-1}^c$ in place of $J_k = J \cap dec_{1,k-1}^c$. Accordingly the total weighted count of correct detections $\hat{q}_k^{tot} = \hat{q}_1 + \ldots + \hat{q}_k$ takes the form

$$\hat{q}_k^{tot} = \sum_{j\ sent} \pi_j 1_{\{\cup_{k'} \mathcal{H}_{k',j}\}},$$

where the union for term j is taken for steps in the set $\{k' \leq k : j \in pot_{k'}\}$. These unions are non-empty for the terms j in $pot_{1,k} = pot_1 \cup \ldots \cup pot_k$. For terms in $sent$ we will be arranging that for each j there is, as k' increases, an increasing probability (of purified approximations) of the set $\mathcal{H}_{k',j}$. Accordingly, for a lower bound on the indicator of the union using a single set we use $1_{\mathcal{H}_{\max_{k,j},j}}$ where $\max_{k,j}$ is the largest of $\{k' \leq k : j \in pot_{k'}\}$. Thus in place of $\sum_{j\ sent} \pi_j 1_{\mathcal{H}_{k,j}}$, for the lower bound on the total weighted fraction of correct detections this leads to

$$\hat{q}_k^{tot} \geq \sum_{j \in sent \cap pot_{1,k}} \pi_j 1_{\mathcal{H}_{\max_{k,j},j}}.$$

Likewise an upper bound on the total weighted fraction of false alarms is

$$\hat{f}_k^{tot} \leq \sum_{j \in other \cap pot_1} \pi_j 1_{\mathcal{H}_{1,j}} + \ldots + \sum_{j \in other \cap pot_k} \pi_j 1_{\mathcal{H}_{k,j}}.$$

Again the idea is to have these simpler forms with single half-space events, but now with each sum taken over a more targeted deterministic set, permitting a smaller total false alarm bound.

In most of the paper we hold off on demonstration of the benefits of the wedges and of the narrowed analysis window (or a combination of both). This is a matter of avoiding complication. But we can revisit the matter to produce improved quantification of mistake bounds.

E. Adjusted sums replace sums of adjustments:

The manner in which the quantities $\hat{q}_1, \ldots, \hat{q}_k$ and $\hat{f}_1, \ldots \hat{f}_k$ arise in the distributional analysis of Lemma 2 is through the sum

$$\hat{q}_k^{adj,tot} = \hat{q}_1^{adj} + \ldots + \hat{q}_k^{adj}$$

of the adjusted values $\hat{q}_k^{adj} = \hat{q}_k/(1 + \hat{f}_k/\hat{q}_k)$. Conveniently, by Lemma 4 below, $\hat{q}_k^{adj,tot} \geq \hat{q}_k^{tot,adj}$. That is, the total of adjusted increments is at least the adjusted total given by

$$\hat{q}_k^{tot,adj} = \frac{\hat{q}_k^{tot}}{1 + \hat{f}_k^{tot}/\hat{q}_k^{tot}}$$

which may also be written

$$\hat{q}_k^{tot} - \hat{f}_k^{tot} + \frac{(\hat{f}_k^{tot})^2}{\hat{q}_k^{tot} + \hat{f}_k^{tot}}.$$

In terms of the total weighted count of tests above threshold $accept_k^{tot} = \hat{q}_k^{tot} + \hat{f}_k^{tot}$ it is

$$\hat{q}_k^{tot,adj} = accept_k^{tot} - 2\hat{f}_k^{tot} + \frac{(\hat{f}_k^{tot})^2}{accept_k^{tot}}.$$

Lemma 4: Let $f_1, \ldots, f_k$ and $g_1, \ldots, g_k$ be non-negative numbers. Then

$$\frac{g_1}{1+f_1/g_1} + \ldots + \frac{g_k}{1+f_k/g_k}$$
$$\geq \frac{g_1 + \ldots + g_k}{1 + (f_1 + \ldots + f_k)/(g_1 + \ldots + g_k)}.$$

Moreover, both of these quantities exceed

$$(g_1 + \ldots + g_k) - (f_1 + \ldots + f_k).$$

Proof of Lemma 4: Form $p_{k'} = f_{k'}/[f_1 + \ldots + f_k]$ and interpret as probabilities for a random variable K taking values k' from 1 to k. Consider the convex function defined by $\psi(x) = x/(1+1/x)$. After accounting for the normalization, the left side is $\mathbb{E}[\psi(g_K/f_K)]$ and the right side is $\psi[\mathbb{E}(g_K/f_K)]$. So the first claim holds by Jensen's inequality. The second claim is because $g/(1+f/g)$ equals $g - f/(1+f/g)$ or equivalently $g - f + f^2/(g+f)$, which is at least $g - f$. This completes the proof of Lemma 4.

This lemma is used to assert that $\hat{s}_k = 1/(1 - \hat{q}_{k-1}^{adj,tot}\nu)$ is at least $1/(1 - \hat{q}_{k-1}^{tot,adj}\nu)$. For suitable weights of combination this $\hat{s}_k$ corresponds to a total shift factor, as developed in the next section.

V. Separation Analysis

In this section we explore the extent of separation between the distributions of the test statistics $\mathcal{Z}_{k,j}^{comb}$ for j sent versus other j. In essence, for j sent, the distribution is a shifted normal. We arrange for the assignment of the weights λ used in the definition of $\mathcal{Z}_{k,j}^{comb}$, so as to approximately maximize this shift.

A. The shift of the combined statistic:

Concerning the weights $\lambda_{1,k}, \lambda_{2,k}, \ldots, \lambda_{k,k}$, for notational simplicity hide the dependence on k and denote them simply by $\lambda_1, \ldots, \lambda_k$, as elements of a vector λ. This λ is to be a member of the simplex $S_k = \{\lambda : \lambda_{k'} \geq 0, \sum_{k'=1}^{k} \lambda_{k'} = 1\}$ in which the coordinates are non-negative and sum to 1.

With weight vector λ the combined test statistic $\mathcal{Z}_{\lambda,k,j}^{comb}$ takes the form

$$\text{shift}_{\lambda,k,j}\, 1_{\{j\ sent\}} + Z_{\lambda,k,j}^{comb}$$

where

$$Z_{\lambda,k,j}^{comb} = \sqrt{\lambda_1} Z_{1,j} - \sqrt{\lambda_2} Z_{2,j} - \ldots - \sqrt{\lambda_k} Z_{k,j}.$$

For convenience of analysis, it is defined not just for $j \in J_k$, but indeed for all $j \in J$, using the normal distribution for the $Z_{k',j}$ discussed above. Here

$$\text{shift}_{\lambda,k,j} = \text{shift}_{\lambda,k}\sqrt{C_{j,R,B}}$$

where shift$_{\lambda,k}$ is

$$\sqrt{\lambda_1 \mathcal{X}_n^2/n} + \sqrt{\lambda_k \hat{w}_2 \mathcal{X}_{n-1}^2/n} + \ldots + \sqrt{\lambda_k \hat{w}_k \mathcal{X}_{n-k+1}^2/n},$$

where $\mathcal{X}_{n-k+1}^2 = \|G_k\|^2/\sigma_k^2$. This shift$_{\lambda,k}$ would be largest with $\lambda_{k'}$ proportional to $\hat{w}_{k'}\mathcal{X}_{n-k'+1}^2$.

Outside of an exception set A_h developed further below, these $\mathcal{X}_{n-k'+1}^2/n$ are at least $1-h$, with small positive h. Then shift$_{\lambda,k}$ is at least $\sqrt{1-h}$ times

$$\sqrt{\lambda_1 \hat{w}_1} + \sqrt{\lambda_k \hat{w}_2} + \ldots + \sqrt{\lambda_k \hat{w}_k}.$$

Our test statistic $\mathcal{Z}_{k,j}^{comb}$ along with its constituent $Z_{k,j}^{comb}$ arises by plugging in particular choices of $\hat{\lambda}$. Most choices of these weights that arise in our development will depend on the data and we lose exact normality of $Z_{k,j}^{comb}$. This matter is addressed using tools of empirical processes, to show uniformity of closeness of relative frequencies based on $Z_{\lambda,k,j}^{comb}$ to the expectations based on the normal distribution. This uniformity can be exhibited over all λ in the simplex S_k. For simplicity it is exhibited over a suitable subset of it.

B. Maximizing separation:

Setting $\lambda_{k'}$ equal to $\hat{w}_{k'}/(1 + \hat{w}_2 + \ldots + \hat{w}_k)$ for $k' \leq k$ would be ideal, as it would maximize the resulting shift factor $\sqrt{\lambda_1} + \sqrt{\hat{w}_2}\sqrt{\lambda_2} + \ldots + \sqrt{\hat{w}_k}\sqrt{\lambda_k}$, for $\lambda \in S_k$, making it equal $\sqrt{1 + \hat{w}_2 + \ldots + \hat{w}_k} = \sqrt{\hat{s}_k}$, where $\hat{s}_k = 1/(1 - q_k^{adj,tot}\nu)$ and $\hat{w}_{k'} = \hat{s}_{k'} - \hat{s}_{k'-1}$.

Setting $\lambda_{k'}$ proportional to $\hat{w}_{k'}$ may be ideal, but it suffers from the fact that without advance knowledge of $sent$ and $other$, the decoder does not have access to the separate values of $\hat{q}_k = \sum_{j \in sent \cap J_k} \pi_j 1_{\mathcal{H}_{k,j}}$ and $\hat{f}_k = \sum_{j \in other \cap J_k} \pi_j 1_{\mathcal{H}_{k,j}}$ needed for precise evaluation of $\hat{w}_k$. We have devised a couple of means to overcome this difficulty. The first is to take advantage of the fact that the decoder does have $accept_k = \hat{q}_k + \hat{f}_k = \sum_{j \in J_k} \pi_j 1_{\mathcal{H}_{k,j}}$, which is the weighted count of terms above threshold on step k. The second is to use computation of $\|G_k\|^2/n$ which is $\sigma_k^2 \mathcal{X}_{n-k+1}^2/n$ as an estimate of σ_k^2 with which we can unravel a reasonable estimate of $\hat{w}_k$. A third method is to use residuals as discussed in the appendix, though its analysis is more involved.

C. Setting weights $\hat{\lambda}$ based on $accept_k$:

The first method uses $accept_{k'}$ in place of $\hat{q}_{k'}^{adj}$ where it arises in the definition of $\hat{w}_{k'}$ to produce a suitable choice of $\lambda_{k'}$. Abbreviate $accept_k$ as acc_k, when needed to allow certain expressions to be suitably displayed. This $accept_k$ upperbounds $\hat{q}_k$ and is not much greater that $\hat{q}_k$ when suitable control of the false alarms is achieved.

Recall $\hat{w}_k = \hat{s}_k - \hat{s}_{k-1}$ for $k > 1$ so finding the common denominator it takes the form

$$\hat{w}_k = \frac{\hat{q}_{k-1}^{adj}\nu}{(1 - \hat{q}_{k-1}^{adj,tot}\nu)(1 - \hat{q}_{k-2}^{adj,tot}\nu)},$$

with the convention that $\hat{q}_0^{adj} = 0$. Let $\hat{w}_k^{acc}$ be obtained by replacing $\hat{q}_{k-1}^{adj}$ with its upper bound of $acc_{k-1} = accept_{k-1}$ and likewise replacing $\hat{q}_{k-2}^{adj,tot}$ and $\hat{q}_{k-1}^{adj,tot}$ with their upper bounds acc^{tot}_{k-2} and acc^{tot}_{k-1}, respectively, with $acc^{tot}_0 = 0$. Thus as an upper bound on $\hat w_k$ set

$$\hat w^{acc}_k = \frac{acc_{k-1}\,\nu}{(1-acc^{tot}_{k-2}\,\nu)(1-acc^{tot}_{k-1}\,\nu)},$$

where for $k=1$ we set $\hat w^{acc}_k = \hat w_k = 1$. For $k>1$ this $\hat w^{acc}_k$ is also

$$\frac{1}{1-acc^{tot}_{k-1}\nu} - \frac{1}{1-acc^{tot}_{k-2}\nu}.$$

Now each $accept_{k'}$ exceeds $\hat q^{adj}_{k'}$ and is less than $\hat q^{adj}_{k'} + 2\hat f_{k'}$.

Then set $\hat\lambda_{k'}$ proportional to $\hat w^{acc}_k$. Thus

$$\hat\lambda_1 = \frac{1}{1+\hat w^{acc}_2+\ldots+\hat w^{acc}_k}$$

and for k' from 2 to k we have

$$\hat\lambda_{k'} = \frac{\hat w^{acc}_{k'}}{1+\hat w^{acc}_2+\ldots+\hat w^{acc}_k}.$$

The shift factor

$$\sqrt{\hat\lambda_1}+\sqrt{\hat\lambda_2\,\hat w_2}+\ldots+\sqrt{\hat\lambda_k\,\hat w_k}$$

is then equal to the ratio

$$\frac{1+\sqrt{\hat w^{acc}_2\,\hat w_2}+\ldots+\sqrt{w^{acc}_k\,\hat w_k}}{\sqrt{1+\hat w^{acc}_2+\ldots+\hat w^{acc}_k}}.$$

From $\hat w^{acc}_{k'} \ge \hat w_{k'}$ the numerator is at least $1+\hat w_2+\ldots+\hat w_k = \hat s_k$, equalling $1/\big(1-(\hat q^{adj}_1+\ldots+\hat q^{adj}_{k-1})\nu\big)$, which per Lemma 4 is at least $1/(1-\hat q^{tot,adj}_{k-1}\,\nu)$. As for the sum in the denominator, it equals $1/(1-acc^{tot}_{k-1}\nu)$. Consequently, the above shift factor using $\hat\lambda$ is at least

$$\frac{\sqrt{1-acc^{tot}_{k-1}\nu}}{1-\hat q^{tot,adj}_{k-1}\,\nu}.$$

Recognizing that acc^{tot}_{k-1} and $\hat q^{tot}_{k-1}$ are similar when the false alarm effects are small, it is desirable to express this shift factor in the form

$$\sqrt{\frac{1-\hat h_{f,k-1}}{1-\hat q^{tot,adj}_{k-1}\,\nu}},$$

where $\hat h_{f,k}$ for each k is a small term depending on false alarms.

Some algebra confirms this is so with

$$\hat h_{f,k} = \hat f^{tot}_k\,\frac{(2-\hat f^{tot}_k/acc^{tot}_k)\nu}{1-\hat q^{tot,adj}_k\nu}$$

which is less than the value $2\hat f^{tot}_k\nu/(1-\nu)$ equal to $2\hat f^{tot}_k\, snr$. Except in cases of large snr we find this approach to be quite suitable.

To facilitate a simple empirical process argument, replace each acc_k by its value $\lceil acc_k\tilde L\rceil/\tilde L$ rounded up to a rational of denominator $\tilde L$ for some integer $\tilde L$ large compared to k. This restricts the acc_k to a set of values of cardinality $\tilde L$ and correspondingly the set of values of $acc_1,\ldots,acc_{k-1}$ determining $\hat w^{acc}_2,\ldots,\hat w^{acc}_k$ and hence $\hat\lambda_1,\ldots,\hat\lambda_k$ is restricted to a set of cardinality $(\tilde L)^{k-1}$. The resulting acc^{tot}_k is then increased by at most $k/\tilde L$ compared to the original value. With this rounding, one can deduce that

$$\hat h_{f,k} \le 2\hat f^{tot}_k\, snr \,+\, k/\tilde L.$$

Next we can proceed with defining natural exception sets outside of which $\hat q^{tot}_{k'}$ is at least a deterministic value $q_{1,k'}$ and $\hat f^{tot}_{k'}$ is not more than a deterministic value $f_{1,k'}$ for each k' from 1 to k. This leads to $\hat q^{tot,adj}_k$ being at least $q^{adj}_{1,k}$, where

$$q^{adj}_{1,k} = q_{1,k}/(1+f_{1,k}/q_{1,k})$$

and $\hat h_{f,k}$ is at most $h_{f,k} = 2f_{1,k}\,snr$, and likewise for each $k'\le k$. This $q^{adj}_{1,k}$ is regarded as an adjustment to $q_{1,k}$ due to false alarms.

When rounding the acc_k to be rational of denominator $\tilde L$, it is accounted for by setting

$$h_{f,k} = 2f_{1,k}\,snr \,+\, k/\tilde L.$$

The result is that the shift factor given above is at least the deterministic value $\sqrt{1-h_{f,k-1}}/\sqrt{1-q^{adj}_{1,k-1}\,\nu}$ near $1/\sqrt{1-q^{adj}_{1,k-1}\,\nu}$. Accordingly shift$_{\hat\lambda,k,j}$ exceeds the purified value

$$\sqrt{\frac{1-h'}{1-q^{adj}_{1,k-1}\,\nu}}\,\sqrt{C_{j,R,B}},$$

where $1-h' = (1-h_f)(1-h)$, with $h' = h+h_f-hh_f$, where $h_f = h_{f,m-1}$ serves as an upper bound to the $h_{f,k-1}$ for all steps $k\le m$.

D. Setting weights $\hat\lambda$ based on estimation of σ^2_k:

The second method entails estimation of $\hat w_k$ using an estimate of σ^2_k. To develop it we make use of the multiplicative relationship from Lemma 2,

$$\hat s_k = \hat s_{k-1}\,\frac{ACC_{k-1}}{\sigma^2_k},$$

where $ACC_{k-1} = acc_{k-1}P = \sum_{j\in dec_{k-1}} P_j$ is the unnormalized weight of terms above threshold on step $k-1$. Accordingly, from $\hat w_k = \hat s_k - \hat s_{k-1}$ it follows that

$$\hat w_k = \hat s_{k-1}\left(\frac{ACC_{k-1}}{\sigma^2_k}-1\right),$$

where the positivity of $\hat w_k$ corresponds to $ACC_{k-1}\ge\sigma^2_k$. Also

$$\hat s_k = \prod_{k'=1}^{k-1}\frac{ACC_{k'-1}}{\sigma^2_{k'}}.$$

Recognize that each $1/\sigma^2_{k'} = \mathcal{X}^2_{n-k'+1}/\|G_{k'}\|^2$. Again, outside an exception set, we may replace each $\mathcal{X}^2_{n-k'+1}$ by its lower bound $n(1-h)$, obtaining the lower bounding estimates

$$\hat w^{low}_k = \hat s^{low}_{k-1}\left(\frac{ACC_{k-1}}{\hat\sigma^2_k}-1\right),$$

where

$$\hat s^{low}_k = \prod_{k'=1}^{k-1}\frac{ACC_{k'-1}}{\hat\sigma^2_{k'}}$$

with

$$\hat{\sigma}_k^2 = \max\left\{\frac{\|G_k\|^2}{n(1-h)},\ ACC_{k-1}\right\}.$$

Initializing with $\hat{s}_1^{low} = \hat{w}_1^{low} = 1$ we again have $\hat{w}_k^{low} = \hat{s}_k^{low} - \hat{s}_{k-1}^{low}$ and hence

$$s_k^{low} = 1 + \hat{w}_2^{low} + \ldots + \hat{w}_k^{low}.$$

Set the weights of combination to be $\hat{\lambda}_{k'} = \hat{w}_{k'}^{low}/\hat{s}_k^{low}$ with which the shift factor is

$$\frac{1 + \sqrt{\hat{w}_2^{low}\hat{w}_2} + \ldots + \sqrt{\hat{w}_k^{low}\hat{w}_k}}{\sqrt{\hat{s}_k^{low}}}.$$

Using $\hat{w}_{k'} \geq \hat{w}_{k'}^{low}$ this is at least

$$\frac{1 + \hat{w}_2^{low} + \ldots + \hat{w}_k^{low}}{\sqrt{\hat{s}_k^{low}}} = \sqrt{\hat{s}_k^{low}},$$

which is $\sqrt{\hat{s}_k}$ times the square root of

$$\prod_{k'=1}^{k-1}\left(\frac{(1-h)n}{\mathcal{X}_{n-k'+1}^2}\right).$$

When using this method of estimating $\hat{w}_k$ augment the exception set so that outside it one has $\mathcal{X}_{n-k'+1}^2/n \leq (1+h)$. Then the above product is at least $[(1-h)/(1+h)]^{k-1}$ and the shift factor $\text{shift}_{\hat{\lambda},k}$ is at least

$$\sqrt{\hat{s}_k\,(1-h')} \;\geq\; \sqrt{\frac{1-h'}{1-q_{1,k-1}^{adj}\nu}},$$

where now $1-h' = (1-h)^k/(1+h)^{k-1}$. Here the additional $(1-h)$ factor, as before, is to account in the definition of $\text{shift}_{\hat{\lambda},k}$ for lower bounding the $\mathcal{X}_{n-k'+1}^2/n$ by $(1-h)$.

Whether now the $[(1-h)/(1+h)]^{k-1}$ is less of a drop than the $(1-h_f) = (1-2f_{k-1}snr)$ from before depends on the choice of h, the bound on the false alarms, the number of steps k and the signal to noise ratio snr.

Additional motivation for this choice of $\hat{\lambda}_k$ comes from consideration of the tests statistics $\mathcal{Z}_{k,j}^{res} = X_j^T res_k/\|res_k\|$ formed by taking the inner products of X_j with the standardized residuals, where res_k denotes the difference between Y and its projection onto the span of $F_1, F_2, \ldots, F_{k-1}$. It is shown in the appendix that these statistics have the same representation but with $\lambda_{k'} = w_{k'}/s_k$, for $k' \leq k$, where $s_k = \|Y\|^2/\|res_k\|^2$ and $w_k = s_k - s_{k-1}$, again initialized with $s_1 = w_1 = 1$. In place of the iterative rule developed above

$$\hat{s}_k = \hat{s}_{k-1}\,\frac{ACC_{k-1}}{\sigma_k^2} = \hat{s}_{k-1}\,\frac{ACC_{k-1}\mathcal{X}_{n-k+1}^2}{\|G_k\|^2},$$

these residual-based s_k are shown there to satisfy

$$s_k = s_{k-1}\frac{\|\tilde{F}_{k-1}\|^2}{\|G_k\|^2}$$

where $\tilde{F}_{k-1}$ is the part of F_{k-1} orthogonal to the previous $F_{k'}$ for $k' = 1, \ldots, k-2$.

Intuitively, given that the coordinates of X_j are i.i.d. with mean 0 and variance 1, this $\|\tilde{F}_{k-1}\|^2$ should not be too different from $\|F_{k-1}\|^2$ which should not be too different from $n\,ACC_{k-1}$. So these properties give additional motivation for our choice. It is also tempting to try to see whether this λ based on the residuals could be amenable to our method of analysis. It would seem that one would need additional properties of the design matrix X, such as uniform isometry properties of subsets of certain sizes. However, it is presently unclear whether such properties could be assured without harming the freedom to have rate up to capacity. For now we stick to the simpler analysis based on the estimates here of the $\hat{w}_k$ that maximizes separation.

E. Exception events and purified statistics:

Consider more explicitly the exception events

$$A_q = \cup_{k'=1}^{k-1}\{\hat{q}_{k'}^{tot} < q_{1,k'}\}$$

and

$$A_f = \cup_{k'=1}^{k-1}\{\hat{f}_{k'}^{tot} > f_{1,k'}\}.$$

As we said, we also work with the related events $\cup_{k'=1}^{k-1}\{\hat{q}_{1,k'} < q_{1,k'}\}$ and $\cup_{k'=1}^{k-1}\{\hat{f}_{1,k'} > f_{1,k'}\}$.

Define the Chi-square exception event A_h to include

$$\cup_{k'=1}^{k}\{\mathcal{X}_{n-k'+1}^2/n \leq 1-h\}$$

or equivalently $\cup_{k'=1}^{k}\{\mathcal{X}_{n-k'+1}^2/(n-k'+1) \leq (1-h_{k'})\}$ where $h_{k'}$ is related to h by $(n-k'+1)(1-h_{k'}) = n(1-h)$. For the second method it is augmented by including also

$$\cup_{k'=1}^{k}\{\mathcal{X}_{n-k'+1}^2/n \geq 1+h\}.$$

The overall exception event is $A = A_q \cup A_f \cup A_h$. When outside this exception set, the $\text{shift}_{\hat{\lambda},k,j}$ exceeds the purified value given by

$$\text{shift}_{k,j} = \sqrt{\frac{1-h'}{1-q_{1,k-1}^{adj}\,\nu}}\,\sqrt{C_{j,R,B}}.$$

Recalling that $C_{j,R,B} = \pi_j\nu L(\log B)/R$ the factor $1-h'$ may be absorbed into the expression by letting

$$C_{j,R,B,h} = C_{j,R,B}(1-h').$$

Then the above lower bound on the shift may be expressed as

$$\sqrt{\frac{C_{j,R,B,h}}{1-x\nu}}$$

evaluated at $x = q_{1,k-1}^{adj}$.

For λ in S_k, set $H_{\lambda,k,j}$ to be the purified event that the approximate combined statistic $\text{shift}_{k,j}\,1_{j\ sent} + Z_{\lambda,k,j}^{comb}$ is at least the threshold τ. That is,

$$H_{\lambda,k,j} = \{\text{shift}_{k,j}1_{j\ sent} + Z_{\lambda,k,j}^{comb} \geq \tau\},$$

where in contrast to $\mathcal{H}_{k,j} = \{\mathcal{Z}_{k,j}^{comb} \geq \tau\}$ we use a standard rather than a calligraphic font for this event $H_{\lambda,k,j}$ based on the normal $Z_{\lambda,k,j}^{comb}$ with the purified shift.

Recall that the coordinates of our λ, denoted $\lambda_{k',k}$ for $k' = 1, 2, \ldots k$, have dependence on k. For each k, the $\lambda_{k',k}$ can be determined from normalization of segments of the first k in sequences $w_1, w_2, \ldots, w_m$ of positive values. With an abuse of notation, we also denote the sequence for $k = 1, 2, \ldots, m$ of such standardized combinations $Z^{comb}_{\lambda,k,j}$ as

$$Z^{comb}_{w,k,j} = \frac{\sqrt{w_1}Z_{1,j} - \sqrt{w_2}Z_{2,j} - \ldots - \sqrt{w_k}Z_{k,j}}{\sqrt{w_1 + w_2 + \ldots + w_k}}.$$

In this case the corresponding event $H_{\lambda,k,j}$ is also denoted $H_{w,k,j}$.

Except in $A_q \cup A_f \cup A_h$, the event $\mathcal{H}_{k,j}$ contains $H_{\hat{\lambda},k,j}$ also denote as $H_{\hat{w}^{acc},k,j}$ or $H_{\hat{w}^{low},k,j}$, respectively, for the two methods of estimating $\hat{w}$.

Also, as for the actual test statistics, the purified forms satisfy the updates

$$Z^{comb}_{\lambda,k,j} = \sqrt{1-\lambda_k}Z^{comb}_{\lambda,k-1,j} - \sqrt{\lambda_k}Z_{k,j}$$

where $\lambda_k = \lambda_{k,k}$.

F. Definition of the update function:

Via $C_{j,R,B}$ the expression for the shift is decreasing in R. Smaller R produce a bigger shift and greater statistical distinguishability between the terms sent and those not sent. This is a property commensurate with the communication interest in the largest R for which after a suitable number of steps one can reliable distinguish most of the terms.

Take note for j sent that $\text{shift}_{k,j}$ is equal to

$$\mu_j(x) = \sqrt{\frac{C_{j,R,B,h}}{1 - x\nu}}$$

evaluated at $x = q^{adj}_{1,k-1}$. To bound the probability with which a term sent is successfully detected by step k, we are motivated to examine the behavior of

$$\Phi(\mu_j(x) - \tau)$$

which, at that x, is the $\mathbb{Q}$ probability of the purified event $H_{\lambda,k,j}$ for j in $sent$, based on the standard normal cumulative distribution of $Z^{comb}_{\lambda,k,j}$. This $\Phi(\mu_j(x) - \tau)$ is increasing in x.

For constant power allocation the contributions $\Phi(\mu_j(x) - \tau)$ are the same for all j in $sent$, whereas, for decreasing power assignments, one has a variable detection probability. Note that it is greater than $1/2$ for those j for which $\mu_j(x)$ exceeds τ. As x increases, there is a growing set of sections for which $\mu_j(x)$ sufficiently exceeds τ, such that these sections have high $\mathbb{Q}$ probability of detection.

The update function $g_L(x)$ is defined as the π weighted average of these $\Phi(\mu_j(x) - \tau)$ for j in $sent$, namely,

$$g_L(x) = \sum_{j\ sent} \pi_j \Phi(\mu_j(x) - \tau),$$

an L term sum. That is, it is the $\mathbb{Q}$ expectation of the weighted fraction $\sum_{j\ sent} \pi_j 1_{H_{\lambda,k,j}}$ for any λ in S_k. The idea is that for any given x this weighted fraction will be near $g_L(x)$, except in an event of exponentially small probability.

This update function g_L on $[0,1]$ indeed depends on the power allocation π as well as the design parameters L, B, R, and the value a determining $\tau = \sqrt{2\log B} + a$. Plus it depends on the signal to noise ratio via $\nu = snr/(1 + snr)$. The explicit use of the subscript L is to distinguish the sum $g_L(x)$ from an integral approximation to it denoted g that will arise later below.

VI. Detection Build-up with False Alarm Control

In this section, target false alarm rates are set and a framework is provided for the demonstration of accumulation of correct detections in a moderate number of steps.

A. Target false alarm rates:

A target weighted false alarm rate for step k arises as a bound f^* on the expected value of $\sum_{j\ other} \pi_j 1_{H_{w,j,k}}$. This expected value is $(B-1)\bar{\Phi}(\tau)$, where $\bar{\Phi}(\tau)$ is the upper tail probability with which a standard normal exceeds the threshold $\tau = \sqrt{2\log B} + a$. A tight bound is

$$\frac{1}{(\sqrt{2\log B} + a)\sqrt{2\pi}} \exp\big\{-a\sqrt{2\log B} - (1/2)a^2\big\}.$$

We have occasion to make use of the similar choice of f^* equal to

$$\frac{1}{(\sqrt{2\log B})\sqrt{2\pi}} \exp\big\{-a\sqrt{2\log B}\big\}.$$

The fact that these indeed upper bound $B\bar{\Phi}(\tau)$ follows from $\bar{\Phi}(x) \le \phi(x)/x$ for positive x, with ϕ being the standard normal density. Likewise set $f > f^*$. We express $f = \rho f^*$ with $\rho > 1$.

Across the steps k, our choice of constant $a_k = a$ produces constant $f^*_k = f^*$ with sum $f^*_{1,k}$ equal to kf^*. Furthermore, set $f_{1,k} > f^*_{1,k}$, which arises in upper bounding the total false alarm rate. In particular, we may arrange for the ratio $f_{1,k}/f^*_{1,k}$ to be at least as large as a fixed $\rho > 1$.

At the final step m, we let

$$\bar{f}^* = f^*_{1,m} = mf^*$$

be the baseline total false alarm rate, and use $\bar{f} = f_{1,m}$, typically equal to $\rho\bar{f}^*$, to be a value which will be shown to likely upper bound $\sum_{j\ other} \pi_j 1_{\cup_{k=1}^m H_{w,j,k}}$.

As will be explored soon, we will need $f_{1,k}$ to stay less than a target increase in the correct detection rate each step. As this increase will be a constant times $1/\log B$, for certain rates close to capacity, this will then mean that we need $\bar{f}$ and hence $\bar{f}^*$ to be bounded by a multiple of $1/\log B$. Moreover, the number of steps m will be of order $\log B$. So with $\bar{f}^* = mf^*$ this means f^* is to be of order $1/(\log B)^2$. From the above expression for f^*, this will entail choosing a value of a near

$$(3/2)(\log\log B)/\sqrt{2\log B}.$$

B. Target total detection rate:

A target total detection rate $q^*_{1,k}$ and the associated values $q_{1,k}$ and $q^{adj}_{1,k}$ are recursively defined using the function $g_L(x)$.

In particular, per the preceding section, let

$$q^*_{1,k} = \sum_{j\ sent} \pi_j \Phi(\text{shift}_{k,j} - \tau)$$

which is seen to be

$$q^*_{1,k} = g_L(x)$$

evaluated at $x = q^{adj}_{1,k-1}$. The convention is adopted at $k = 1$ that the previous q_{k-1} and $x = q^{adj}_{1,k-1}$ are initialized at 0. To complete the specification, a sequence of small positive η_k are chosen with which we set

$$q_{1,k} = q^*_{1,k} - \eta_k.$$

For instance we may set $\eta_k = \eta$. The idea is that these η_k will control the exponents of tail probabilities of the exception set outside of which $\hat{q}^{tot}_k$ exceeds $q_{1,k}$. With this choice of $q_{1,k}$ and $f_{1,k}$ we have also

$$q^{adj}_{1,k} = q_{1,k}/(1 + f_{1,k}/q_{1,k}).$$

Positivity of the gap $g_L(x) - x$ provides that $q^*_{1,k}$ is larger than $q^{adj}_{1,k-1}$. As developed in the next subsection, the contributions from η_k and $f_{1,k}$ are arranged to be sufficiently small that $q^{adj}_{1,k}$ and $q_{1,k}$ are increasing with each such step. In this way the analysis will quantify as x increases, the increasing proportion that are likely to be above threshold.

C. Building up the total detection rate:

Let's give the framework here for how the likely total correct detection rate $q_{1,k}$ builds up to a value near 1, followed by the corresponding conclusion of reliability of our adaptive successive decoder. Here we define the notion of correct detection being *accumulative*. This notion holds for the power allocations we study.

Recall that with the function $g_L(x)$ defined above, for each step, one updates the new $q_{1,k}$ by choosing it to be slightly less than $q^*_{1,k} = g_L(q^{adj}_{1,k-1})$. The choice of $q_{1,k}$ is accomplished by setting a small positive η_k for which $q_{1,k} = q^*_{1,k} - \eta_k$. These may be constant, that is $\eta_k = \eta$, across the steps $k = 1, 2, \ldots, m$.

There are slightly better alternative choices for the η_k motivated by the reliability bounds. One is to arrange for $D(q_{1,k} \| q^*_{1,k})$ to be constant where D is the relative entropy between Bernoulli random variables of the indicated success probabilities. Another is to arrange η_k such that $\eta_k/\sqrt{V_k}$ is constant, where $V_k = V(x)$ evaluated at $x = q^{adj}_{1,k-1}$, where

$$V(x)/L = \sum_{j\ sent} \pi_j \Phi(\mu_j(x)) \bar{\Phi}(\mu_j(x)).$$

This V_k/L may be interpreted as a variance of $\hat{q}_{1,k}$ as developed below. The associated standard deviation factor $\sqrt{V(x)}$ is shown in the appendix to be proportional to $(1 - x\nu)$. With evaluation at $x = q^{adj}_{1,k-1}$, this gives rise to $\eta_k = \eta(x)$ equal to $(1 - x\nu)$ times a small constant.

How large we can pick η_k will be dictated by the size of the gap $g_L(x) - x$ at $x = q^{adj}_{1,k-1}$.

Let x^* be any given value between 0 and 1, preferably not far from 1.

Definition: A positive increasing function $g(x)$ bounded by 1 is said to be *accumulative* for $0 \le x \le x^*$ if there is a function $gap(x) > 0$, with

$$g(x) - x \ge gap(x)$$

for all $0 \le x \le x^*$. An adaptive successive decoder with rate and power allocation chosen so that the update function $g_L(x)$ satisfies this property is likewise said to be *accumulative*. The *shortfall* is defined by $\delta^* = 1 - g_L(x^*)$.

If the update function is accumulative and has a small shortfall, we demonstrate, for a range of choices of $\eta_k > 0$ and $f_{1,k} > f^*_{1,k}$, that the target total detection rate $q_{1,k}$ increases to a value near 1 and that the weighted fraction mistakes is with high probability less than $\delta_k = (1 - q_{1,k}) + f_{1,k}$. This mistake rate δ_k is less than $1-x^*$ after a number of steps, and then with one more step it is further reduced to a value not much more than $\delta^* = 1 - g_L(x^*)$, to take advantage of the amount by which $g_L(x^*)$ exceeds x^*.

The tactic in providing good probability exponents will be to demonstrate, for the sparse superposition code, that there is an appropriate size gap. It will be quantified via bounds on the minimum of the gap or the minimum of the standardized gap, $gap(x)/(1-x\nu)$, where the minimum is taken for $0 \le x \le x^*$.

The following lemmas relate the sizes of η and $\bar{f}$ and the number of steps m to the size of the gap.

***Lemma** 5:* Suppose the update function $g_L(x)$ is accumulative on $[0, x^*]$ with $g_L(x) - x \ge gap$ for a positive constant $gap > 0$. Arrange positive constants η and $\bar{f}$ and $m^* \ge 2$, such that

$$\eta + \bar{f} + 1/(m^* - 1) = gap.$$

Suppose $f_{1,k} \le \bar{f}$ as arises from $f_{1,k} = \bar{f}$ or from $f_{1,k} = kf$ for each $k \le m^*$ with $f = \bar{f}/m^*$. Set $q_{1,k} = q^*_{1,k} - \eta$. Then $q_{1,k}$ is increasing on each step for which $q_{1,k-1} - f_{1,k-1} \le x^*$, and, for such k, the increment $q_{1,k} - q_{1,k-1}$ is at least $1/(m^* - 1)$. The number of steps $k = m - 1$ required such that $q_{1,k} - f_{1,k}$ first exceeds x^*, is bounded by $m^* - 1$. At the final step $m \le m^*$, the weighted fraction of mistakes target $\delta_m = (1 - q_{1,m}) + f_{1,m}$ satisfies

$$\delta_m \le \delta^* + \eta + \bar{f}.$$

The value $\delta_m = (1 - q_{1,m}) + f_{1,m}$ is used in controlling the sum of weighted fractions of failed detections and of false alarms.

In the decomposition of the gap, think of η and $\bar{f}$ as providing portions of the gap which contribute to the probability exponent and false alarm rate, respectively, whereas the remaining portion controls the number of steps.

The following is an analogous conclusion for the case of a variable size gap bound. It allows for somewhat greater freedom in the choices of the parameters, with η_k and $f_{1,k}$ determined by functions $\eta(x)$ and $f(x)$, respectively, evaluated at $x = q^{adj}_{1,k-1}$.

***Lemma** 6:* Suppose the update function is accumulative on $[0, x^*]$. Choose positive functions $\eta(x)$ and $\bar{f}(x)$ on $[0, x^*]$ with $gap(x) - \eta(x) - \bar{f}(x)$ not less than a positive value denoted gap'. Suppose $q_{1,k} = q^*_{1,k} - \eta_k$ where $\eta_k \le \eta(q^{adj}_{1,k-1})$ and $f_{1,k} \le \bar{f}(q^{adj}_{1,k-1})$. Then $q_{1,k} - q_{1,k-1} > gap'$ on each step for which $q^{adj}_{1,k-1} \le x^*$ and the number of steps k such that the $q^{adj}_{1,k}$ first exceeds x^* is bounded by $1/gap'$. With a number of steps $m \le 1 + 1/gap'$, the $\delta_m = (1 - q_{1,m}) + f_{1,m}$ satisfies

$$\delta_m \le \delta^* + \eta_m + f_{1,m}.$$

The proofs for Lemmas 5 and 6 are given in Appendix III. One has the choice whether to be bounding the number of steps such that $q^{adj}_{1,k}$ first exceeds x^* or such that the slightly smaller value $q_{1,k} - f_{1,k}$ first exceeds x^*. The latter provides the slightly stronger conclusion that $\delta_k \le 1 - x^*$. Either way, at the penultimate step $q^{adj}_{1,k}$ is at least x^*, which is sufficient for the next step $m = k + 1$ to take us to a larger value of $q^*_{1,m}$ at least $g_L(x^*)$. So either formulation yields the stated conclusion.

Associated with the use of the factor $(1 - x\nu)$ we have the following improved conclusion, noting that GAP is necessarily larger than the minimum of $gap(x)$.

***Lemma** 7:* Suppose that $g_L(x) - x$ is at least $gap(x) = (1 - x\nu)GAP$ for $0 \le x \le x^*$ with a positive GAP. Again there is convergence of $g_{1,k}$ to values at least x^*. Arrange positive η_{std} and m^* with

$$GAP = \eta_{std} + \frac{\log 1/(1-x^*)}{m^* - 1}.$$

Set $\eta(x) = (1 - x\nu)\eta_{std}$ and $\bar{f} \le (1-\nu)GAP'$ with $GAP' = [\log 1/(1-x^*)]/(m^* - 1)$ and set $\eta_k = \eta(x)$ at $x = q^{adj}_{1,k-1}$ and $f_{1,k} \le \bar{f}$. Then the number of steps $k = m - 1$ until x_k first exceeds x^* is not more than $m^* - 1$. Again at step m the $\delta_m = (1 - q_{1,m}) + f_{1,m}$ satisfies $\delta_m \le \delta^* + \eta_m + \bar{f}$.

Proof of Lemma 7: We have

$$q_{1,k} = g_L(q^{adj}_{1,k-1}) - \eta(q^{adj}_{1,k-1})$$

at least

$$q^{adj}_{1,k-1} + (1 - q^{adj}_{1,k-1}\nu)(GAP - \eta_{std}).$$

Subtracting $\bar{f}$ as a bound on $f_{1,k}$, it yields

$$q^{adj}_{1,k} \;\ge\; q^{adj}_{1,k-1} + (1 - q^{adj}_{1,k-1})\nu\, GAP'.$$

This implies, with $x_k = g^{adj}_{1,k}$ and $\epsilon = \nu\, GAP'$, that

$$x_k \;\ge\; (1-\epsilon)x_{k-1} + \epsilon$$

or equivalently,

$$(1 - x_k) \le (1-\epsilon)(1 - x_{k-1}),$$

as long as $x_{k-1} \le x^*$. Accordingly for such k, we have the exponential bound

$$(1 - x_k) \le (1-\epsilon)^k \le e^{-\epsilon k} = e^{-\nu\, GAP'\, k}$$

and the number of steps $k = m - 1$ until x_k first exceeds x^* satisfies

$$m - 1 \le \frac{\log 1/(1-x^*)}{\log 1/(1-\epsilon)} \le \frac{\log 1/(1-x^*)}{\nu\, GAP'}.$$

This bound is $m^* - 1$. The final step takes $q^*_{1,m}$ to a value at least $g_L(x^*)$ so $\delta_m \le \delta^* + \eta_m + f_{1,m}$. This completes the proof of Lemma 7.

The idea here is that by extracting the factor $(1-x\nu)$, which is small if x and ν are near 1, it follows that a value GAP with larger constituents η_{std} and GAP' can be extracted than the previous constant gap, though to do so one pays the price of the $\log 1/(1-x^*)$ factor.

Concerning the choice of $f_{1,k}$, consider setting $f_{1,k} = \bar{f}$ for all k from 1 to m. This constant $f_{1,k} = \bar{f}$ remains bigger than $f^*_{1,k} = kf^*$ with minimum ratio $\bar{f}/\bar{f}^*$ at least $\rho > 1$. To give a reason for choosing a constant false alarm bound, note that with $f_{1,k}$ equal to $f_{1,m} = \bar{f}$, it is greater than $f^*_{1,m} = \bar{f}^*$, which exceeds $f^*_{1,k}$ for $k < m$. Accordingly, the relative entropy exponent $(B-1)D(p_{1,k}\|p^*_{1,k})$ that arises in the probability bound in the next section is smallest at $k = m$, where it is at least $\bar{f}\,\mathcal{D}(\rho)/\rho$, where $\mathcal{D}(\rho)$ is the positive value $\rho \log \rho - (\rho - 1)$.

In contrast, one has the seemingly natural choice $f_{1,k} = kf$ of linear growth in the false alarm bound, with $f = f^*\rho$. It is also upper bounded by $\bar{f}$ for $k \le m$ and has constant ratio $f_{1,k}/f^*_{1,k}$ equal to ρ. It yields a corresponding exponent of $kf\,\mathcal{D}(\rho)/\rho$ for $k = 1$ to m. However, this exponent has a value at $k = 1$ that can be seen to be smaller by a factor of order $1/m$. For the same final false alarm control, it is preferable to arrange the larger order exponent, by keeping $D(p_{1,k}\|p^*_{1,k})$ at least its value at $k = m$.

VII. Reliability of Adaptive Successive Decoding

Here we establish, for any power allocation and rate for which the decoder is accumulative, the reliability with which the weighted fractions of mistakes are governed by the studied quantities $1 - q_{1,m}$ plus $f_{1,m}$. The bounds on the probabilities with which the fractions of mistakes are worse than such targets are exponentially small in L. The implication is that if the power assignment and the communication rate are such that the function g_L is accumulative on $[0, x^*]$, then for a suitable number of steps, the tail probability for weighted fraction of mistakes more than $\delta * = 1 - g_L(x^*)$ is exponentially small in L.

A. Reliability using the data-driven weights:

In this subsection we demonstrate reliability using the data-driven weights $\hat{\lambda}$ in forming the statistic $\mathcal{Z}^{comb}_{k,j}$. Subsection VII-B discusses a slightly different approach which uses deterministic weights and provides slightly smaller error probability bounds.

***Theorem** 8: Reliable communication by sparse superposition codes with adaptive successive decoding*. With total false alarm rate targets $f_{1,k} > f^*_{1,k}$ and update function g_L, set recursively the detection rate targets $q_{1,k} = g_L(q^{adj}_{1,k-1}) - \eta_k$, with $\eta_k = q^*_{1,k} - q_{1,k} > 0$ set such that it yields an increasing sequence $q_{1,k}$ for steps $1 \le k \le m$. Consider $\hat{\delta}_m$, the weighted failed detection rate plus false alarm rate. Then the m step adaptive successive decoder incurs $\hat{\delta}_m$ less than

$\delta_m = (1 - q_{1,m}) + f_{1,m}$, except in an event of probability with upper bound as follows:

I)
$$\sum_{k=1}^{m} \left[e^{-L_\pi D(q_{1,k} \| q^*_{1,k}) + (k-1) \log \tilde{L}) + c_0 k} \right]$$
$$+ \sum_{k=1}^{m} \left[e^{-L_\pi (B-1) D(p_{1,k} \| p^*_{1,k}) + (k-1) \log \tilde{L}} \right]$$
$$+ \sum_{k=1}^{m} e^{-(n-k+1) D_{h_k}},$$
where the terms correspond to tail probabilities concerning, respectively, the fractions of correct detections, the fractions of false alarms, and the tail probabilities for the events $\{\|G\|_k^2/\sigma_k^2 \le n(1-h)\}$, on steps 1 to m. Here $L_\pi = 1/\max_j \pi_j$. The $p_{1,k}, p^*_{1,k}$ equal the corresponding $f_{1,k}, f^*_{1,k}$ divided by $B-1$. Also $D_h = -\log(1-h) - h$ is at least $h^2/2$. Here $h_k = (nh-k+1)/(n-k+1)$, so the exponent $(n-k+1)D_{h_k}$ is near nD_h, as long as k/n is small compared to h.

II) A refined probability bound holds as in I above but with exponent
$$L \, \frac{\eta_k^2}{V_k + (1/3)\eta_k (L/L_\pi)}$$
in place of $L_\pi \, D(q_{1,k} \| q^*_{1,k})$ for each $k = 1, 2, \ldots, m$.

***Corollary** 9:* Suppose the rate and power assignments of the adaptive successive code are such that g_L is accumulative on $[0, x^*]$ with a positive constant gap and a small shortfall $\delta^* = 1 - g_L(x^*)$. Assign positive $\eta_k = \eta$ and $f_{1,k} = \bar{f}$ and $m \ge 2$ with $1 - q_{1,m} \le \delta^* + \eta$. Let $\mathcal{D}(\rho) = \rho \log \rho - (\rho - 1)$. Then there is a simplified probability bound. With a number of steps m, the weighted failed detection rate plus false alarm rate is less than $\delta^* + \eta + \bar{f}$, except in an event of probability not more than,
$$m e^{-2L_\pi \eta^2 + m[c_0 + \log \tilde{L}]} + m e^{-L_\pi \bar{f} \mathcal{D}(\rho)/\rho + m \log \tilde{L}}$$
$$+ m e^{-(n-m+1) h_m^2/2}.$$

The bound in the corollary is exponentially small in $2L_\pi \eta^2$ if h is chosen such that $(n-m+1)h_m^2/2$ is at least $2L_\pi \eta^2$ and $\rho > 1$ and $\bar{f}$ are chosen such that $\bar{f}\,[\log \rho - 1 + 1/\rho]$ matches $2\eta^2$.

Improvement is possible using II when we find that V_k is of order $1/\sqrt{\log B}$. This produces a probability bound exponentially small in $L\eta^2 (\log B)^{1/2}$ for small η.

Proof of Theorem 8 and its Corollary: False alarms occur on step k, when there are terms j in $other \cap J_k$ for which there is occurrence of the event $\mathcal{H}_{k,j}$, which is the same for such j in $other$ as the event $H_{\hat{w}^{acc},k,j}$, as there is no shift of the statistics for j in $other$. The weighted fraction of false alarms up to step k is $\hat{f}_1 + \ldots + \hat{f}_k$ with increments $\hat{f}_k = \sum_{j \in other \cap J_k} \pi_j 1_{\mathcal{H}_{k',j}}$. This increment excludes the terms in $dec_{1,k-1}$ which are previously decoded. Nevertheless, introducing associated random variables for these excluded events (with the distribution discussed in the proof of Lemmas 1 and 2), we may regard the sum as the weighted fraction of the union $\sum_{j \in other} \pi_j 1_{\cup_{k'=1}^{k} \mathcal{H}_{k',j}}$.

Recall, as previously discussed, for all such j in $other$, the event $H_{w,k',j}$ is the event that $Z^{comb}_{w,k',j}$ exceeds τ, where for each $w = (1, w_2, w_3, \ldots, w_k)$, the $Z^{comb}_{w,k',j}$ are standard normal random variables, independent across j in $other$. So the events $\cup_{k'=1}^{k} H_{w,k',j}$ are independent and equiprobable across such j. Let $p^*_{1,k}$ be their probability or an upper bound on it, and let $p_{1,k} > p^*_{1,k}$. Then $A_{f,k} = \{\hat{f}_k^{tot} \ge f_{1,k}\}$ is contained in the union over all possible w of the events $\{\hat{p}_{w,1,k} \ge p_{1,k}\}$ where
$$\hat{p}_{w,1,k} = \frac{1}{B-1} \sum_{j \in other} \pi_j \, 1_{\cup_{k'=1}^{k} H_{w,k',j}}.$$
With the rounding of the acc_k to rationals of denominator $\tilde{L}$, the cardinality of the set of possible w is at most $\tilde{L}^{k-1}$. Moreover, by Lemma 47 in the appendix, the probability of the events $\{\hat{p}_{w,1,k} \ge p_{1,k}\}$ is less than $e^{-L_\pi (B-1) D(p_{1,k} \| p^*_{1,k})}$. So by the union bound the probability of $\{\hat{f}_k^{tot} \ge f_{1,k}\}$ is less than
$$(\tilde{L})^{k-1} e^{-L_\pi (B-1) D(p_{1,k} \| p^*_{1,k})}.$$

Likewise, investigate the weighted proportion of correct decodings $\hat{q}_m^{tot}$ and the associated values $\hat{q}_{1,k} = \sum_{j \; sent} \pi_j 1_{\mathcal{H}_{\hat{\lambda},k,j}}$ which are compared to the target values $q_{1,k}$ at steps $k = 1$ to m. The event $\{\hat{q}_{1,k} < q_{1,k}\}$ is contained in $\mathcal{F}_k$ so when bounding its $\mathbb{P}$ probability, incurring a cost of a factor of e^{kc_0}, we may switch to the simpler measure $\mathbb{Q}$.

Consider the event $A = \cup_{k=1}^{m} A_k$, where A_k is the union of the events $\{\hat{q}_{1,k} \le q_{1,k}\}$, $\{\hat{f}_k^{tot} \ge f_{1,k}\}$ and $\{\mathcal{X}^2_{n-k+1}/n < 1-h\}$. This event A may be decomposed as the union for k from 1 to m of the disjoint events $A_k \cap_{k'=1}^{k-1} A^c_{k'}$. The Chi-square event may be expressed as $A_{h,k} = \{\mathcal{X}^2_{n-k+1}/(n-k+1) < 1 - h_k\}$ which has the probability bound
$$e^{-(n-k+1) D_{h_k}}.$$

So to bound the probability of A, it remains to bound for k from 1 to m, the probability of the event
$$A_{q,k} = \{\hat{q}_{1,k} < q_{1,k}\} \cap A^c_{h,k} \cap_{k'=1}^{k-1} A^c_{k'}.$$
In this event, with the intersection of $A^c_{k'}$ for all $k' < k$ and the intersection with the Chi-square event $A^c_{h,k}$, the statistic $\mathcal{Z}^{comb}_{k,j}$ exceeds the corresponding approximation
$$\sqrt{s_k} \, \sqrt{C_{j,R,B,h}} \, 1_{j \; sent} \; + \; Z^{comb}_{\hat{w}^{acc},k,j},$$
where $s_k = 1/[1 - q^{adj}_{1,k-1} \nu]$. There is a finite set of possible $\hat{w}^{acc}$ associated with the grid of values of $acc_1, \ldots, acc_{k-1}$ rounded to rationals of denominator $\tilde{L}$. Now $A_{q,k}$ is contained in the union across possible w of the events
$$\{\hat{q}_{w,1,k} < q_{1,k}\}$$
where
$$\hat{q}_{w,1,k} = \sum_{j \; sent} \pi_j 1_{\{Z^{comb}_{w,k,j} \ge a_{k,j}\}}.$$

Here $a_{k,j} = \tau - \sqrt{s_k} \, \sqrt{C_{j,R,B,h}}$. With respect to $\mathbb{Q}$, these $Z^{comb}_{w,k,j}$ are standard normal, independent across j, so the Bernoulli random variables $1_{\{Z^{comb}_{w,k,j} \ge a_{k,j}\}}$ have success probability $\bar{\Phi}(a_{k,j})$ and accordingly, with respect to $\mathbb{Q}$, the $\hat{q}_{w,1,k}$

has expectation $q^*_{1,k} = \sum_{j\; sent} \pi_j \bar{\Phi}(a_{k,j})$. Thus, again by Lemma 47 in the appendix the probability of

$$\mathbb{Q}\big\{\hat{q}_{w,1,k} < q_{1,k}\big\}$$

is not more than

$$e^{-L_\pi D(q_{1,k}\|q^*_{1,k})}.$$

By the union bound we multiply this by $(\tilde{L})^{k-1}$ to bound $\mathbb{Q}(A_{q,k})$. One may sum it across k to bound the probability of the union.

The Chi-square random variables and the normal statistics for j in $other$ have the same distribution with respect to $\mathbb{P}$ and $\mathbb{Q}$ so there is no need to multiply by the $e^{c_0 k}$ factor for the A_h and A_f contributions.

The event of interest

$$A_{q^{tot}_m} = \{\hat{q}^{tot}_m \le q_{1,m}\}$$

is contained in the union of the event $A_{q^{tot}_m} \cap A^c_{q,m-1} \cap A^c_f \cap A^c_h$ with the events $A_{q,m-1}$, A_h and A_f, where $A_h = \cup_{k=1}^m A_{h,k}$ and $A_f = \cup_{k=1}^m A_{f,k}$. The three events $A_{q,m-1}$, A_h and A_f are clearly part of the event A which has been shown to have the indicated exponential bound on its probability. This leaves us with the event

$$A_{q^{tot}_m} \cap A^c_{q,m-1} \cap A^c_f \cap A^c_h$$

Now, as we have seen, $\hat{q}^{tot}_m$ may be regarded as the weighted proportion of of occurrence the union $\cup_{k=1}^m \mathcal{H}_{k,j}$ which is at least $\sum_{j\; sent} \pi_j 1_{\mathcal{H}_{m,j}}$. Outside the exception sets A_h, A_f and $A_{q,m-1}$, it is at least $\hat{q}_{1,m} = \sum_{j\; sent} \pi_j 1_{H_{\hat{w}^{acc},m,j}}$. With the indicated intersections, the above event is contained in $A_{q,m} = \{\hat{q}_{1,m} \le q_{1,m}\}$, which is also part of the event A. So by containment in a union of events for which we have bounded the probabilities, we have the indicated bound.

As a consequence of the above conclusion, outside the event A, at step $k = m$, we have $\hat{q}^{tot}_m > q_{1,m}$. Thus outside A the weighted fraction of failed detections, which is not more than $1 - \hat{q}_{1,m}$, is less than $1 - q_{1,m}$. Also outside A, we have that the weighted fraction of false alarms is less than $f_{1,m}$. So the total weighted fraction of mistakes $\hat{\delta}_m$ is less than $\delta_m = (1 - q_{1,m}) + f_{1,m}$.

In these probability bounds the role in the exponent of $D(q\|q^*)$ for numbers q and q^* in $[0,1]$, is played the relative entropy between the Bernoulli(q) and the Bernoulli q^* distributions, even though these q and q^* arise as expectations of weighted sums of many independent Bernoulli random variables.

Concerning the simplified bounds in the corollary, by the Pinsker-Csiszar-Kulback-Kemperman inequality, specialized to Bernoulli distributions, the expressions of the form $D(q\|q^*)$ in the above, exceed $2(q-q^*)^2$. This specialization gives rise to the $e^{-2L_\pi \eta^2}$ bound when the $q_{1,k}$ and $\tilde{q}_{1,k}$ differ from $q^*_{1,k}$ by the amount η.

The $e^{-2L_\pi \eta^2}$ bound arises alternatively by applying Hoeffding's inequality for sums of bounded independent random variables to the weighted combinations of Bernoulli random variables that arise with respect to the distribution $\mathbb{Q}$. As an aside, we remark that order η^2 is the proper characterization of $D(q\|q^*)$ only for the middle region of steps when $q^*_{1,k}$ is neither near 0 nor near 1. There are larger exponents toward the ends of the interval $(0,1)$ because Bernoulli random variables have less variability there.

To handle the exponents $(B-1)D(p\|p^*)$ at the small values $p = p_{1,k} = f_{1,k}/(B-1)$ and $p^* = p^*_{1,k} = f^*_{1,k}/(B-1)$, we use the Poisson lower bound on the Bernoulli relative entropy, as shown in the appendix. This produces the lower bound $(B-1)[p_{1,k}\log p_{1,k}/p^*_{1,k} + p^*_{1,k} - p_{1,k}]$ which is equal to

$$f_{1,k}\log f_{1,k}/f^*_{1,k} + f^*_{1,k} - f_{1,k}.$$

We may write this as $f^*_{1,k}\mathcal{D}(\rho_k)$ or equivalently $f_{1,k}\mathcal{D}(\rho_k)/\rho_k$ where the functions $\mathcal{D}(\rho)$ and $\mathcal{D}(\rho)/\rho = \log\rho + 1 - 1/\rho$ are increasing in $\rho \ge 1$.

If we used $f_{1,k} = kf$ and $f^*_{1,k} = kf^*$ in fixed ratio $\rho = f/f^*$, this lower bound on the exponent would be $kf\,\mathcal{D}(\rho)/\rho$ as small as $f\,\mathcal{D}(\rho)/\rho$. Instead, keeping $f_{1,k}$ locked at $\bar{f}$, which is at least $\bar{f}^*\rho$, and keeping $f^*_{1,k} = kf^*$ less than or equal to $mf^* = \bar{f}^*$, the ratio ρ_k will be at least ρ and the exponents will be at least as large as $\bar{f}\,\mathcal{D}(\rho)/\rho$.

Finally, there is the matter of the refined exponent in II. As above proof the heart of the matter is the consideration of the probability $\mathbb{Q}\big\{\hat{q}_{w,1,k} < q_{1,k}\big\}$. Fix a value of k between 1 and m. Recall that $\hat{q}_{w,1,k} = \sum_{j\; sent} \pi_j 1_{H_{w,k,j}}$. So we wish to bound the probability of the event that the sum of the independent random variables $\xi_j = -\pi_j\,(1_{H_{w,k,j}} - \bar{\Phi}_j)$ exceeds η, where $\bar{\Phi}_j = \bar{\Phi}(\text{shift}_{k,j} - \tau) = \mathbb{Q}(H_{w,k,j})$ provides the centering so that the ξ_j have mean 0. We recognize that $\bar{\Phi}_j$ is $\bar{\Phi}(\mu_j(x)) = 1 - \Phi(\mu_j(x))$, evaluated at $x = q^{adj}_{1,k}$, and it is the same as used in the evaluation of the $q^*_{1,k}$, the expected value of $\hat{q}_{w,1,k}$, which is $g_L(x)$. The random variables ξ_j have magnitude bounded by $\max_j \pi_j = 1/L_\pi$ and variance $v_j = \pi_j^2\,\Phi_j(1-\Phi_j)$. Thus we bound $\mathbb{Q}\big\{\hat{q}_{w,1,k} < q_{1,k}\big\}$ by Bernstein's inequality, where the sums are understood to be for j in $sent$,

$$\mathbb{Q}\Big\{\sum_j \xi_j \ge \eta\Big\} \le \exp\left\{-\frac{\eta^2}{2[V/L + \eta/(3L_\pi)]}\right\},$$

where here $\eta = \eta_k$ is the difference between the mean $q^*_{1,k}$ and $q_{1,k}$ and $V/L = \sum_j v_j = \sum_j \pi_j^2\,\Phi_j(1-\Phi_j)$ is the total variance. It is V_k/L given by

$$V(x)/L = \sum_j \pi_j^2 \Phi(\mu_j(x))(1-\Phi(\mu_j(x)))$$

evaluated at $q^{adj}_{1,k-1}$. This completes the proof of Theorem 8.

If we were to use the obvious, but crude, bound on the total variance of $(\max_j \pi_j)\sum_j \pi_j 1/4 = 1/(4L_\pi)$ the result in II would be no better than the $\exp\{-2L_\pi\eta^2\}$ bound that arises from the Hoeffding bound.

The variable power assignment we shall study arranges $\Phi_j(1-\Phi_j)$ to be small for most j in $sent$. Indeed, a comparison of the sum $V(x)/L$ to an integral, in a manner similar to the analysis of $g_L(x)$ in an upcoming section, shows that $V(x)$ is not more than a constant times $1/\tau$, which is of order $1/\sqrt{\log B}$, by the calculation in Appendix IX. This produces, with a positive constant $const$, a bound of the form

$$\exp\left\{-const\,L\min\left\{\eta, \eta^2\sqrt{\log B}\right\}\right\}.$$

Equivalently, in terms of $n = (L \log B)/R$ the exponent is a constant times $n \min\{\eta^2/\sqrt{\log B}, \eta/\log B\}$. This exponential bound is an improvement on the other bounds in the Theorem 8, by a factor of $\sqrt{\log B}$ in the exponent for a range of values of η up to $1/\sqrt{\log B}$, provided of course that we arrange $\eta < gap$ to permit the required increase in $q_{1,k}$. For our best rates, we will need η to be of order $1/\log B$, to within a loglog factor, matching the order of $\mathcal{C}$–R. So this improvement brings the exponent to within a $\sqrt{\log B}$ factor of best possible.

Other bounds on the total variance are evident. For instance, noting that $\sum_j \pi_j \Phi_j(1-\Phi_j)$ is less than both $\sum_j \pi_j \Phi_j$ and $\sum_j \pi_j (1-\Phi_j)\}$, it follows that

$$V(x)/L \leq (1/L_\pi) \min\{g_L(x), 1-g_L(x)\}.$$

This reveals that there is considerable improvement in the exponents provided by the Bernstein bound for the early and later steps where $g_L(x)$ is near 0 or 1, even improving the order of the bounds there. This does not alter the fact that we must experience the effect of the exponents for steps with x near the middle of the interval from 0 to 1, where the previously mentioned bound on $V(x)$ produces an exponent of order $L\eta^2\sqrt{\log B}$.

For the above, we used data-driven weights λ, with which the error probability in a union bound had to be multiplied by a factor of $\tilde{L}^{k-1}$, for each step k, to account for the size of the set of possible weight vectors.

Below we describe a slight modification to the above procedure using deterministic λ that does away with this factor, thus demonstrating increased reliability for given rates below capacity. The procedure involves choosing each dec_k to be a subset of the terms above threshold, with the π weighted size of this set very near a pre-specified value $pace_k$.

B. An alternative approach:

As mentioned earlier, instead of making dec_k, the set of decoded terms for step k, to be equal to $thresh_k$, we take dec_k for each step to be a subset of $thresh_k$ so that its size $accept_k$ is near a deterministic quantity which we call $pace_k$. This will yield a sum $accept_k^{tot}$ near $\sum_{k'=1}^{k} pace_{k'}$ which we arrange to match $q_{1,k}$. Again we abbreviate $accept_k^{tot}$ as acc_k^{tot} and $accept_k$ as acc_k.

In particular, setting $pace_k = q_{1,k}^{adj} - q_{1,k-1}^{adj}$, the set dec_k is chosen by selecting terms in J_k that are above threshold, in decreasing order of their $\mathcal{Z}_{k,j}^{comb}$ values, until for each k the accumulated amount nearly equals $q_{1,k}$. In particular given acc_{k-1}^{tot}, one continues to add terms to acc_k, if possible, until their sum satisfies the following requirement,

$$q_{1,k}^{adj} - 1/L_\pi < acc_k^{tot} \leq q_{1,k}^{adj},$$

where recall that $1/L_\pi$ is the minimum weight among all j in J. It is a small term of order $1/L$.

Of course the set of terms $thresh_k$ might not be large enough to arrange for $accept_k$ satisfying the above requirement. Nevertheless, it is satisfied, provided

$$acc_{k-1}^{tot} + \sum_{j \in thresh_k} \pi_j \geq q_{1,k}^{adj}$$

or equivalently,

$$\sum_{j \in dec_{1,k-1}} \pi_j + \sum_{j \in J - dec_{1,k-1}} \pi_j 1_{\mathcal{H}_{k,j}} \geq q_{1,k}^{adj}.$$

Here for convenience we take $dec_0 = dec_{1,0}$ as the empty set.

To demonstrate satisfaction of this condition note that the left side is at least the value one has if the indicator $1_{\mathcal{H}_{k,j}}$ is imposed for each j and if the one restricts to j in $sent$, which is the value $\hat{q}_{1,k}^{above} = \sum_{j \in sent} 1_{\mathcal{H}_{k,j}}$. Our analysis demonstrates, for each k, that the inequality

$$\hat{q}_{1,k}^{above} > q_{1,k}$$

holds with high probability, which in turn exceeds $q_{1,k}^{adj}$. So then the above requirement is satisfied for each step, with high probability, and thence acc_k matches $pace_k$ to within $1/L_\pi$.

This $\hat{q}_{1,k}^{above}$ corresponds to the quantity studied in the previous section, giving the weighted total of terms in sent for which the combined statistic is above threshold, and it remains likely that it exceeds the purified statistic $\hat{q}_{1,k}$. What is different is the control on the size of the previously decoded sets allows for constant weights of combination.

In the previous procedure we employed random weights $\hat{w}_k^{acc}$ in the assignment of the $\lambda_{1,k}, \lambda_{2,k}, \ldots, \lambda_{k,k}$ used in the definition of $\mathcal{Z}_{k,j}^{comb}$ and $Z_{k,j}^{comb}$, where we recall that $\hat{w}_k^{acc} = 1/(1 - acc_{k-1}^{tot}\nu) - 1/(1 - acc_{k-2}^{tot}\nu)$. Here, since each acc_k^{tot} is near a deterministic quantity, namely $q_{1,k}^{adj}$, we replace $\hat{w}_k^{acc}$ by a deterministic quantity w_k^* given by,

$$w_k^* = \frac{1}{(1 - q_{1,k-1}^{adj}\nu)} - \frac{1}{(1 - q_{1,k-2}^{adj}\nu)},$$

and use the corresponding vector λ^* with coordinates $\lambda_{k',k}^* = w_{k'}^*/[1 + w_2^* + \ldots + w_k^*]$ for $k' = 1$ to k.

Earlier we demonstrated that $\hat{w}_k \leq \hat{w}_k^{acc}$, which allowed us to quantify the shift factor in each step. Analogously, we have the following result for our current procedure using deterministic weights.

***Lemma** 10:* For $k' < k$, assume that we have arranged decoding sets $dec_{1,k'}$ so that the corresponding $acc_{k'}^{tot}$ takes value in the interval $\left(q_{1,k'}^{adj} - 1/L_\pi \,,\, q_{1,k'}^{adj}\right]$. Then

$$\hat{w}_k \leq w_k^* + \epsilon_1,$$

where $\epsilon_1 = \nu/(L_\pi(1-\nu)^2) = snr(1+snr)/L_\pi$ is a small term of order $1/L$. Likewise, $\hat{w}_{k'} \leq w_{k'}^* + \epsilon_1$ holds for $k' < k$ as well.

Proof of Lemma 10: The $\hat{q}_{k'}$ and $\hat{f}_{k'}$ are the weighted sizes of the sets of true terms and false alarms, respectively, retaining that which is actually decoded on step k', not merely above threshold. These have sum $\hat{q}_{k'} + \hat{f}_{k'} = acc_{k'}$, nearly equal to $pace_{k'}$, taken here to be $q_{1,k'}^{adj} - q_{1,k'-1}^{adj}$. Let's establish the inequalities

$$\hat{q}_1^{adj} + \ldots + \hat{q}_{k-1}^{adj} \leq q_{1,k-1}^{adj}$$

and

$$\hat{q}_{k-1}^{adj} \leq q_{1,k-1}^{adj} - q_{1,k-2}^{adj} + 1/L_\pi.$$

The first inequality uses that each $\hat{q}_{k'}^{adj}$ is not more than $\hat{q}_{k'}$ which is not more than $\hat{q}_{k'} + \hat{f}_{k'}$, equal to $acc_{k'}$ which sums to acc^{tot}_{k-1} not more than $q^{adj}_{1,k-1}$. The second inequality is a consequence of the fact that $\hat{q}^{adj}_{k-1} \leq acc^{tot}_{k-1} - acc^{tot}_{k-2}$. Using the bounds on acc^{tot}_{k-1} and acc^{tot}_{k-2} gives that claimed inequality.

These two inequalities yield

$$\hat{w}_k \leq \frac{(q^{adj}_{1,k-1} - q^{adj}_{1,k-2} + 1/L_\pi)\nu}{(1 - q^{adj}_{1,k-1}\nu)(1 - q^{adj}_{1,k-2}\nu)}.$$

The right side can be written as,

$$\frac{1}{1 - q^{adj}_{1,k-1}\nu} - \frac{1}{1 - q^{adj}_{1,k-2}\nu} + \frac{1/L_\pi\nu}{(1 - q^{adj}_{1,k-1}\nu)(1 - q^{adj}_{1,k-2}\nu)}.$$

Now bound the last term using $q^{adj}_{1,k-1}$ and $q^{adj}_{1,k-2}$ less than 1 to complete the proof of Lemma 10.

Define the exception set $A_{q,above} = \cup_{k'=1}^{k-1}\{\hat{q}^{above}_{1,k'} < q_{1,k'}\}$. In some expressions we will abbreviate *above* as *abv*. Also recall the set $A_f = \cup_{k'=1}^{k-1}\{\hat{f}^{tot}_{k'} > f_{1,k'}\}$. For convenience we we suppress the dependence on k in these sets.

Outside of $A_{q,abv}$, we have $\hat{q}^{abv}_{1,k'}$ at least $q_{1,k'}$ and hence at least $q^{adj}_{1,k'}$ for each $1 \leq k' < k$, ensuring that for each such k' one can get decoding sets $dec_{k'}$ such that the corresponding $acc^{tot}_{k'}$ is at most $1/L_\pi$ below $q^{adj}_{1,k'}$. Thus the requirements of Lemma 10 are satisfied outside this set.

We now proceed to lower bound the shift factor for step k outside of $A_{q,abv} \cup A_f$.

For the above choice of $\lambda = \lambda^*$ the shift factor is equal to the ratio

$$\frac{1 + \sqrt{\hat{w}_2 w^*_2} + \ldots + \sqrt{\hat{w}_k w^*_k}}{\sqrt{1 + w^*_2 + \ldots + w^*_k}}.$$

Using the above lemma and the fact that $\sqrt{a-b} \geq \sqrt{a} - \sqrt{b}$, we get that the above is greater than or equal to

$$\frac{1 + \hat{w}_2 + \ldots + \hat{w}_k}{\sqrt{1 + w^*_2 + \ldots + w^*_k}} - \sqrt{\epsilon_1}\frac{\sqrt{\hat{w}_2} + \ldots + \sqrt{\hat{w}_k}}{\sqrt{1 + w^*_2 + \ldots + w^*_k}}.$$

Now use the fact that

$$\sqrt{\hat{w}_2} + \ldots + \sqrt{\hat{w}_k} \leq \sqrt{k}\sqrt{\hat{w}_2 + \ldots + \hat{w}_k}$$

to bound the second term by $\epsilon_2 = \sqrt{\epsilon_1}\sqrt{k}\sqrt{\nu/(1{-}\nu)}$ which is $snr\sqrt{(1{+}snr)k/L_\pi}$, a term of order near $1/\sqrt{L}$. Hence the shift factor is at least,

$$\frac{1 + \hat{w}_2 + \ldots + \hat{w}_k}{\sqrt{1 + w_2 + \ldots + w_k}} - \epsilon_2.$$

Consequently, it is at least

$$\frac{\sqrt{1 - q^{adj}_{1,k-1}\nu}}{1 - \hat{q}^{tot,adj}_{k-1}\nu} - \epsilon_2.$$

where recall that $\hat{q}^{tot,adj}_{k-1} = \hat{q}^{tot}_{k-1}/(1 + \hat{f}^{tot}_{k-1}/\hat{q}^{tot}_{k-1})$. Here we have used that $1{+}\hat{w}_2 + \ldots + \hat{w}_k$, which is $1/(1{-}\hat{q}^{adj,tot}_{k-1}\nu)$, can be bounded from below by $1/(1{-}\hat{q}^{tot,adj}_{k-1}\nu)$ using Lemma 4.

Similar to before, we note that $q^{adj}_{1,k-1}$ and $\hat{q}^{tot,adj}_{k-1}$ are close to each other when the false alarm effects are small. Hence we write this shift factor in the form

$$\sqrt{\frac{1 - \hat{h}_{f,k-1}}{1 - \hat{q}^{tot,adj}_{k-1}\,\nu}}$$

as before. Again we find that

$$\hat{h}_{f,k-1} \leq \hat{2}f^{tot}_{k-1}\,snr + \epsilon_3$$

outside of the exception set $A_{q,abv}$. Here

$$\epsilon_3 = \frac{snr}{L_\pi} + 2\epsilon_2,$$

is a term of order $1/\sqrt{L}$.

To confirm the above use the inequality $\sqrt{1-a} - \sqrt{b} \geq \sqrt{1-c}$, where $c = a + 2\sqrt{b}$. Here our $a = (q_{1,k-1} - \hat{q}^{tot,adj}_{k-1})\nu/(1 - \hat{q}^{tot,adj}_{k-1}\nu)$ and $b = \epsilon_2^2(1 - q^{tot,adj}_{k-1}\nu)$. Noting that the numerator in a is at most $(1/L_\pi + 2\hat{f}^{tot}_{k-1} - (\hat{f}^{tot}_{k-1})^2)\nu$ outside of $A_{q,abv}$ and that $0 \leq q^{tot,adj}_{k-1} \leq 1$, one obtains the bound for $\hat{h}_{f,k-1}$.

Next, recall that outside of the exception set $A_f \cup A_{q,abv}$ we have that $\hat{q}^{tot}_{k-1} \geq q_{1,k-1}$ and $\hat{f}^{tot}_{k-1} \leq f_{1,k-1}$. This leads to the shift factor being at least

$$\sqrt{\frac{1 - h_{f,k-1}}{1 - q^{adj}_{1,k-1}\,\nu}},$$

where

$$h_{f,k} = 2f_{1,k}\,snr\, +\, \epsilon_3.$$

As before, we assume a bound $f_{1,k} \leq \bar{f}$, so that $h_{f,k}$ is not more than $h_f = 2\bar{f}\,snr + \epsilon_3$, independent of k.

As done previously, we create the combined statistics $\mathcal{Z}^{comb}_{k,j}$, now using our deterministic λ^*. For j in *other* this $\mathcal{Z}^{comb}_{k,j}$ equals $Z^{comb}_{k,j}$ and for j in *sent*, when outside the exception set $A_{abv} = A_{q,abv} \cup A_f \cup A_h$, this combination exceeds

$$\sqrt{\frac{1 - h'}{1 - q^{adj}_{1,k-1}\,\nu}}\,\sqrt{C_{j,R,B}}\,1_{j\;sent}\, +\, Z^{comb}_{k,j},$$

where $(1 - h') = (1 - h)(1 - h_f)$ as before, though with h_f larger by the small amount ϵ_3. Again we have $\text{shift}_{k,j} = \sqrt{C_{j,R,B,h}/(1 - x\,\nu)}$ evaluated at $x = q^{adj}_{1,k-1}$, with $C_{j,R,B,h}$ as before.

Analogous to Theorem 8, reliability after m steps of our algorithm is demonstrated by bound the probability of the exception set $A = \cup_{k=1}^{m} A_k$, where A_k is the union of the events $\{\hat{q}^{abv}_{1,k} \leq q_{1,k}\}$, $\{\hat{f}^{tot}_{k} \geq f_{1,k}\}$ and $\{\mathcal{X}^2_{n-k+1}/n < 1{-}h\}$. Thus the proof of Theorem 8 carries over, only now we do not require the union over the grid of values of the weights. We now state the analogous theorem with the resulting improved bounds.

***Theorem** 11:* Under the same assumptions as in Theorem 8, our m step adaptive successive decoder, using deterministic pacing with $pace_k = q_{1,k} - q_{1,k-1}$, incurs a weighted fraction of errors $\hat{\delta}_m$ less than $\delta_m = f_{1,m} + (1 - q_{1,m})$, except in an event of probability not more than

$$\sum_{k=1}^{m} \left[e^{-L_\pi D(q_{1,k} \| q_{1,k}^*)) + c_0 k} \right]$$
$$+ \sum_{k=1}^{m} \left[e^{-L_\pi (B-1) D(p_{1,k} \| p_{1,k}^*)} \right]$$
$$+ \sum_{k=1}^{m} e^{-(n-k+1) D_{h_k}},$$

where the bound also holds if the exponent $L_\pi \, D(q_{1,k} \| q_{1,k}^*)$ is replaced by

$$L \, \frac{\eta_k^2}{V_k + (1/3)\eta_k (L/L_\pi)}.$$

In the constant gap bound case, with positive η and $\bar{f}$ and $m \geq 2$, satisfying the same hypotheses as in the previous corollary, the probability of $\hat{\delta}_m$ greater than $\delta^* + \eta + \bar{f}$ is not more than

$$m \, e^{-2 \, L_\pi \eta^2 \, + m c_0} \, + \, m \, e^{-L_\pi \bar{f} \mathcal{D}(\rho)/\rho} \, + \, m \, e^{-(n-m+1) h_m^2 / 2}.$$

Furthermore, using the variance V_k and allowing a variable gap bound $gap_k \leq g_L(x_k) - x_k$ and $0 < f_{1,k} + \eta_k < gap_k$, with difference $gap' = gap_k - f_{1,k} + \eta_k$ and number of steps $m \leq 1 + 1/gap'$, and with $\rho_k = f_{1,k}/f_{1,k}^* > 1$, this probability bound also holds with the exponent

$$L \, \min_k \eta_k^2 / \big[V_k + (1/3) \eta_k (L/L_\pi) \big]$$

in place of $2 L_\pi \eta^2$ and with $min_k f_{1,k} \mathcal{D}(\rho_k)/\rho_k$ in place of $\bar{f}\mathcal{D}(\rho)/\rho$, where the minima are taken over k from 1 to m.

The bounds are the same as in Theorem 9 and its corollary, except for improvement due to the absence of the factors $\tilde{L}^{k-1}$. In the same manner as discussed there, there are choices of $\bar{f}$, ρ and h, such that the exponents for the false alarms and the chi-square contributions are at least as good as for the $q_{1,k}$, so that the bound becomes

$$3m \, e^{-2 \, L_\pi \eta^2 \, + m c_0}.$$

We remark that that for the particular variable power allocation rule we study in the upcoming sections, as we have said, the update function $g_L(x)$ will seen to be ultimately insensitive to L, with $g_L(x) - x$ rapidly approaching a function $g(x) - x$ at rate $1/L$ uniformly in x. Indeed, a gap bound for g_L will be seen to take a form $gap_L = gap^* - \theta/L_\pi$ for some constant θ, so that it approaches the value of the gap determined by g, denoted gap^*, where we note that L and L_π agree to within a constant factor. Accordingly, using $gap^* - \theta/L_\pi$ in apportioning the values of η, $\bar{f}$, and $1/(m-1)$, these values are likewise ultimately insensitive to L. Indeed, we shall see that slight adjustment to the rate allows arrangement of a gap independent of L.

Nevertheless, to see if there be any effect on the exponent, suppose for a specified η^* that $\eta = \eta^* - \theta/L_\pi$ represents a corresponding reduction in η due to finite L. Consider the exponential bound

$$e^{-2 L_\pi \eta^2}.$$

Expanding the square it is seen that the exponent $L_\pi \eta^2$, which is $L_\pi (\eta^* - \theta/L_\pi)^2$, is at least $L_\pi (\eta^*)^2$ minus a term $2\theta\eta^*$ that is negligible in comparison. Thus the approach of η to η^* is sufficiently rapid that the probability bound remains close to what it would be,

$$e^{-2 L_\pi (\eta^*)^2},$$

if we were to ignore the effect of the θ/L_π, where we are using that $L_\pi (\eta^*)^2$ is large, and that η^* is small, e.g., of the order of $1/\log B$.

VIII. Computational Illustrations

We illustrate in two ways the performance of our algorithm. First, for fixed values of L, B, snr and rates below capacity we evaluate the detection rate as well as the probability of the exception set $P_{\mathcal{E}}$ using the theoretical bounds given in Theorem 11. Plots demonstrating the progression of our algorithm are also shown. These highlight the crucial role of the function g_L in achieving high reliability.

Figures 1 and 2 presents the results of computation using the reliability bounds of Theorem 11 for fixed L and B and various choices of snr and rates below capacity. The dots in these figures denotes $q_{1,k}^{adj}$ for each k and the step function joining these dots highlight how $q_{1,k}^{adj}$ is computed from $q_{1,k-1}^{adj}$. For large L these $q_{1,k}^{adj}$'s would be near $q_{1,k}$, our lower bound on the proportion of sections decoded after k passes. In this extreme case $q_{1,k}$ would match $g_L(q_{1,k-1})$, so that the dots would lie on the function.

For illustrative purposes we take $B = 2^{16}$, $L = B$ and snr values of 1, 7 and 15. For each snr value the maximum rate, over a grid of values, is determined, for which there is a particular control on the error probability. With $snr = 1$ (Fig 2), this rate R is 0.3 bits which is 59% of capacity. When snr is 7 and 15 (Fig 1 and 2) , these rates correspond to 49.5% and 43.5% of their corresponding capacities.

The error probability is controlled as follows. We arrange each of the $3m$ terms in the probability bound to take the same value, set in these examples to be $\epsilon = 10^{-5}$. In particular, we compute in succession appropriate values of $q_{1,k}^*$ and $f_{1,k}^* = k f^*$, using an evaluation of the function $g_L(x)$, an L term sum, evaluated at a point determined from the previous step, and from these we determine $q_{1,k}$ and $f_{1,k}$.

This means solving for $q_{1,k}$ less than $q_{1,k}^*$ such that $e^{-L_\pi D(q_{1,k} \| q_{1,k}^*) + c_0 k}$ equals ϵ, and with $p_{1,k}^* = f_{1,k}^*/(B-1)$, solving for the $p_{1,k}$ greater than $p_{1,k}^*$ such that the corresponding term $e^{-L_\pi (B-1) D(p_{1,k} \| p_{1,k}^*)}$ also equals ϵ. In this way, we are using the largest $q_{1,k}$ less than $q_{1,k}^*$, that is, the smallest η_k, and the smallest false alarm bound $f_{1,k}$, for which the respective contributions to the error probability bound is not worse then the prescribed value.

These are numerically simple to solve because $D(q \| q^*)$ is convex and monotone in $q < q^*$, and likewise for $D(p \| p^*)$ for $p > p^*$. Likewise we arrange h_k so that $e^{-(n-k+1) D_{h_k}}$ matches ϵ.

Taking advantage of the Bernstein bound sometimes yields a smaller η_k by solving for the choice satisfying the quadratic equation $L \eta_k^2 / [V_k + (1/3) \eta_k L / L_\pi] = \log 1/\epsilon + c_0 k$, where V_k

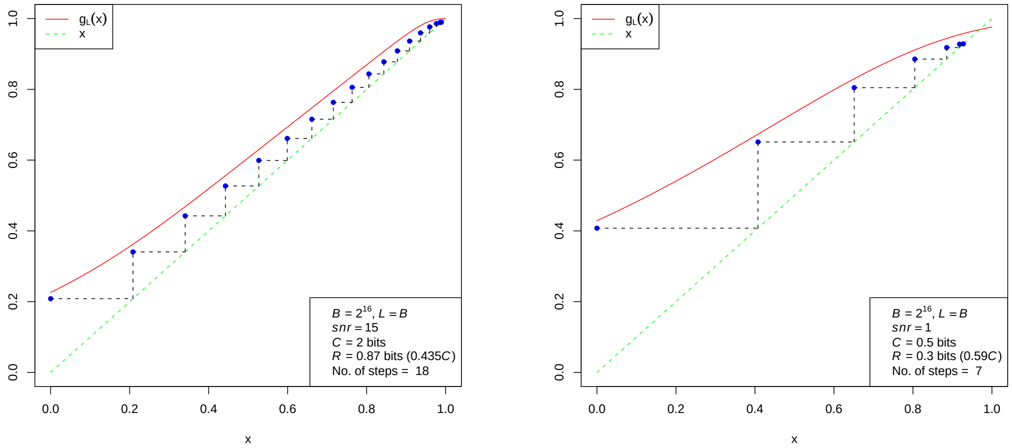

Fig. 2. Plots demonstrating progression of our algorithm. (Plot on left) $snr = 15$. The weighted (unweighted) detection rate is 0.995 (0.983) for a failed detection rate of 0.017 and the false alarm rate is 0.006. The probability of mistakes larger than these targets is bounded by 5.4×10^{-4}. (Plot on right) $snr = 1$. The detection rate (both weighted and un-weighted) is 0.944 and the false alarm and failed detection rates are 0.016 and 0.056 respectively, with the corresponding error probability bounded by 2.1×10^{-4}.

is computed by an evaluation of $V(x)$, which like $q_L(x)$ is a L term sum, both of which are evaluated at $x = q_{1,k-1}^{adj}$.

These computation steps continue as long as $(1-q_{1,k}) + f_{1,k}$ decreases, thus yielding the choice of the number of steps m.

For these computations we choose power allocations proportional to $\max\{e^{-2\gamma(\ell-1)/L}, e^{-2\gamma}(1+\delta_c)\}$, with $0 \le \gamma \le \mathcal{C}$. Here the choices of a, c and γ are made, by computational search, to minimize the resulting sum of false alarms and failed detections, as per our bounds. In the $snr = 1$ case the optimum γ is 0, so we have constant power allocation in this case. In the other two cases, there is variable power across most of the sections. The role of a positive c being to increase the relative power allocation for sections with low weights. Note, in our analytical results for maximum achievable rates as a function of B, as given in the upcoming sections, γ is constrained to be equal to $\mathcal{C}$.

Figure 3 gives plots of achievable rates as a function of B. For each B, the points on the detailed envelope correspond to the numerically evaluated maximum inner code rate for which the section mistake rate is between 9 and 10%. Here we assume L to be large, so that the $q_{1,k}$'s and f_k's are replaced by the expected values $q_{1,k}^*$ and f_k^*, respectively. We also take $h = 0$. This gives an idea about the best possible rates for a given snr and section mistake rate.

For the simulation curve, L was fixed at 100 and for given snr, B and rate values 10^4 runs of our algorithm were performed. The maximum rate over the grid of values satisfying section error rate of less than 10% except in 10 replicates, (corresponding to an estimated $P_{\mathcal{E}}$ of 10^{-3}) are shown in the plots. Interestingly, even for such small values of L the curve is quite close to the detailed envelope curve, showing that our theoretical bounds are quite conservative.

IX. Accumulative g for Finite Length Codes

The purpose of this section is to show that power allocations proportional to $e^{-2\mathcal{C}\ell/L}$ and slight modifications of it provide update functions $g_L(x)$ that are indeed accumulative for rates moderately close to capacity and to quantify how the gap and shortfall $\delta = 1 - g_L(x^*)$ depend on the rate and the section size B. Motivation for these power allocations come in part from the analysis in the appendix in which it is shown that in the saturated detection probability case arising in the limit of large τ, it is necessary for the power allocations to be near this exponential form for an iterative decoder to have accumulative update for rates up to capacity. Here our focus is quantifying the gap and shortfall in the finite L and B case.

In particular, specifics of normalized power allocation weights $\pi_{(\ell)}$ are developed in subsection A, including slight modifications to the exponential form. An integral approximation $g(x)$ to the sum $g_L(x)$ is provided in subsection B. Subsection C examines the behavior of $g_L(x)$ for x near 1, including introduction of x^* via a parameter r_1 related to an amount of permitted rate drop and a parameter ζ related to the amount of shift at x^*. For cases with monotone decreasing $g(x) - x$, as in the unmodified weight case, the behavior for x near 1 suffices to demonstrate that $g_L(x)$ is accumulative. Improved closeness of the rate to capacity is shown in the finite codelength case by allowance of the modifications to the weight via the parameter δ_c. But with this modifications monotonicity is lost. In Subsection D, a bound on the number of oscillations of $g(x) - x$ is established in that is used in showing that $g_L(x)$ is accumulative. The location of x^* and the value of δ_c both impact the mistake rate δ_{mis} and the amount of rate drop required for $g_L(x)$ to be accumulative, expressed through a quantity introduced there denoted r_{crit}. Subsection E provides optimization of δ_c. Helpful inequalities in controlling the rate drop are in subsection F. Subsection G provides optimization of the contribution to the total rate drop of the choice of location of x^*, via optimization of ζ.

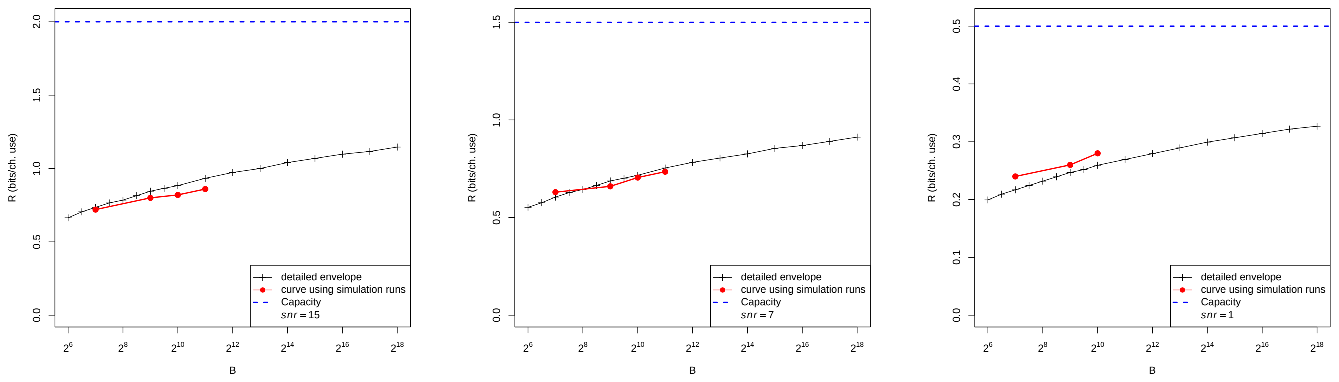

Fig. 3. Plots of achievable rates as a function of B for snr values of 15, 7 and 1. Section error rate is controlled to be between 9 and 10%. For the curve using simulation runs the rates are exhibited for which the empirical probability of making more than 10% section mistakes is near 10^{-3}.

Recall that $g_L(x)$ for $0 \le x \le 1$ is the function given by

$$g_L(x) = \sum_{j\ sent} \pi_j\, \Phi(\mu_j(x) - \tau),$$

where $\mu_j(x) = \sqrt{C_{j,R,B,h}/(1 - x\,\nu)}$. Recursively, $q_{1,k}$ is obtained from $q^*_{1,k} = g_L(x)$ evaluated at $x = q^{adj}_{1,k-1}$, in succession for k from 1 to m.

For a partitioned code, these sums for j in $sent$ are the same as the sum over j_ℓ for sections ℓ from 1 to L, with one from each section. The value of π_j is taken to be the same for terms within a section, and to vary across the sections. We denote the value for section ℓ as $\pi_{(\ell)}$. At slight risk of abuse of notation, it is also convenient to denote $C_{\ell,R} = \pi_{(\ell)}\, L\nu/(2R)$ and $C_{\ell,R,B,h} = C_{\ell,R}(1 - h')(2 \log B)$ and likewise $\mu_\ell(x) = \sqrt{C_{\ell,R,B,h}/(1 - x\,\nu)}$. In this setting, the function $g_L(x)$ is invariant to the choice of $sent$ and is expressed as

$$g_L(x) = \sum_{\ell=1}^{L} \pi_{(\ell)}\, \Phi(\mu_\ell(x) - \tau).$$

The values of $\Phi(\mu_\ell(x) - \tau)$, as a function of ℓ from 1 to L, provide what is interpreted as the probability with which the term sent from section ℓ have approximate test statistic value that is above threshold, when the previous step successfully had an adjusted weighted fraction above threshold equal to x. The $\Phi(\mu_\ell(x) - \tau)$ is increasing in x regardless of the choice of π_ℓ, though how high is reached depends on the choice of this power allocation.

A. Variable power allocations:

We consider two closely related schemes for allocating the power. First suppose $P_{(\ell)}$ is proportional to $e^{-2\mathcal{C}\ell/L}$ as motivated in the introduction. Then the weight for section ℓ is $\pi_{(\ell)}$ given by $P_{(\ell)}/P$. In this case recall that $C_{\ell,R} = \pi_{(\ell)}\, L\nu/(2R)$ simplifies to u_ℓ times the constant $\tilde{\mathcal{C}}/R$ where

$$u_\ell = e^{-2\mathcal{C}(\ell-1)/L},$$

for sections ℓ from 1 to L. The presence of the factor $\tilde{\mathcal{C}}/R$ if at least 1, increases the value of $g_L(x)$ above what it would be if that factor were not there and helps in establishing that it is accumulative.

As ℓ varies from 1 to L the u_ℓ ranges from 1 down to the value $e^{-2\mathcal{C}} = 1 - \nu$.

To roughly explain the behavior, as we shall see, this choice of power allocation produces values of $\Phi(\mu_\ell(x) - \tau)$ that are near 1 for ℓ with u_ℓ enough less than $1 - x\,\nu$ and near 0 for values of u_ℓ enough greater than $1 - x\,\nu$, with a region of ℓ in between, in which there will be a scatter of sections with statistics above threshold. Though it is roughly successful in reaching an x near 1, the fraction of detections is limited, if R is too close to $\tilde{\mathcal{C}}$, by the fact that $\mu_\ell(x)$ is not large for a portion of ℓ near the right end, of the order $1/\sqrt{2 \log B}$.

Therefore, we modify the power allocation, taking $\pi_{(\ell)}$ to be proportional to an expression that is equal to $u_\ell = \exp\{-2\mathcal{C}\frac{\ell-1}{L}\}$ except for large ℓ/L where it is leveled to be not less than a value $u_{cut} = e^{-2\mathcal{C}}(1 + \delta_c)$ which exceeds $(1 - \nu) = e^{-2\mathcal{C}} = 1/(1 + snr)$ using a small positive δ_c. This δ_c is constrained to be between 0 and snr so that u_{cut} is not more than 1. Thus let $\pi_{(\ell)}$ be proportional to $\tilde{u}_\ell$ given by

$$\max\{u_\ell, u_{cut}\}.$$

The idea is that by leveling the height to a slightly larger value for ℓ/L near 1, we arrange nearly all sections to have $\tilde{u}_\ell$ above $(1 - x\nu)$ when x is near 1. This will allow us to reach our objective with an R closer to $\mathcal{C}$. We pay a price in the required normalization, but it will be seen to be of a smaller order of $1/(2 \log B)$.

To produce the normalized $\pi_{(\ell)} = \max\{u_\ell, u_{cut}\}/(L\, sum)$, compute

$$sum = \sum_{\ell=1}^{L} \max\{u_\ell, u_{cut}\}(1/L).$$

If $c = 0$ this sum equals $\nu/(2\tilde{\mathcal{C}})$ as previously seen. If $c > 0$ and $u_{cut} < 1$, it is the sum of two parts, depending on whether $e^{-2\mathcal{C}(\ell-1)/L}$ is greater than or not greater than u_{cut}. This sum can be computed exactly, but to produce a simplified expression let's note that replacing the sum by the corresponding integral

$$integ = \int_0^1 \max\{e^{-2\mathcal{C}t}, u_{cut}\} dt$$

an error of at most $1/L$ is incurred. For each L there is a θ with $0 \le \theta \le 1$ such that

$$sum = integ + \theta/L.$$

In the integral, comparing $e^{-2\mathcal{C}t}$ to u_{cut} corresponds to comparing t to t_{cut} equal to $[1/(2\mathcal{C})] \log 1/u_{cut}$. Splitting the integral accordingly, it is seen to equal $[1/(2\mathcal{C})](1-u_{cut})$ plus $u_{cut}(1-t_{cut})$, which may be expressed as

$$integ = \frac{\nu}{2\mathcal{C}} \, [1 \, + \, D(\delta_c)/snr] \, ,$$

where $D(\delta) = (1+\delta)\log(1+\delta) - \delta$. For $\delta \ge 0$, the function $D(\delta)$ is not more than $\delta^2/2$, which is a tight bound for small δ. This $[1 + D(\delta_c)/snr]$ factor in the normalization, represents a cost to us of introduction of the otherwise helpful δ_c. Nevertheless, this remainder $D(\delta_c)/snr$ is small compared to δ_c, when δ_c is small compared to the snr. It might appear that $D(\delta_c)/snr$ could get large if snr were small, but, in fact, since $\delta_c \le snr$ the $D(\delta_c)/snr$ remains less than $snr/2$.

Accordingly, from the above relationship to the integral, the sum may be expressed as

$$sum = \frac{\nu}{2\mathcal{C}} \, \left[1 \, + \, \delta_{sum}^2\right] \, ,$$

where δ_{sum}^2 is equal to $D(\delta_c)/snr + 2\theta\mathcal{C}/(L\nu)$, which is not more than $\delta_c^2/(2snr) + 2\mathcal{C}/(L\nu)$. Thus

$$\pi_{(\ell)} = \frac{\max\{u_\ell, u_{cut}\}}{L \, sum} = \frac{2\mathcal{C}}{L\nu} \frac{\max\{u_\ell, u_{cut}\}}{1 + \delta_{sum}^2}.$$

In this case $C_{\ell,R,B,h} = (\pi_\ell L\nu/(2R))(1-h')(2\log B)$ may be written

$$C_{\ell,R,B,h} = \max\{u_\ell, u_{cut}\} \, \frac{\mathcal{C}(1-h')}{R(1 + \delta_{sum}^2)} (2\log B),$$

or equivalently, using $\tau = \sqrt{2\log B}\,(1 + \delta_a)$, this is

$$\max\{u_\ell, u_{cut}\} \, (\mathcal{C}'/R)\tau^2,$$

where

$$\mathcal{C}' \; = \; \frac{\mathcal{C}\,(1-h')}{(1+\delta_{sum}^2)(1+\delta_a)^2}.$$

For small δ_c, δ_a, and h' this is a value near the capacity $\mathcal{C}$. As we will see later, our best choices of these parameters make it less than capacity by an amount of order $\log\log B/\log B$. When $\delta_c = 0$ the $\mathcal{C}/(1+\delta_{sum}^2)$ is what we have called $\tilde{\mathcal{C}}$ and its closeness to capacity is controlled by $\delta_{sum}^2 \le 2\mathcal{C}/(\nu\, L)$.

In contrast, if δ_c were taken to be the maximum permitted, which is $\delta_c = snr$, then the power allocation would revert to the constant allocation rule, with an exact match of the integral and the sum, so that $1 + \delta_{sum}^2 \, = \, 1 + D(snr)/snr$ and the $\mathcal{C}/(1+\delta_{sum}^2)$ simplifies to $R_0 = (1/2)snr/(1+snr)$, which, as we have said, is a rate target substantially inferior to $\mathcal{C}$, unless the snr is small.

Now $\mu_\ell(x) - \tau$ which is $\sqrt{C_{\ell,R,B,h}/(1 - x\,\nu)} - \tau$ may be written as the function

$$\mu(x,u) = \left(\sqrt{u/(1-x\,\nu)} - 1\right)\tau$$

evaluated at $u = \max\{u_\ell, u_{cut}\}\,(\mathcal{C}'/R)$. For later reference note that the $\mu_\ell(x)$ here and hence $g_L(x)$ both depend on x and the rate R only through the quantity $(1 - x\,\nu)R/\mathcal{C}'$.

Note also that $\mu(x,u)$ is of order τ and whether it is positive or negative depends on whether or not u exceeds $1 - x\,\nu$ in accordance with the discussion above.

B. Formulation and evaluation of the integral $g(x)$:

The function that updates the target fraction of correct decodings is

$$g_L(x) = \sum_{\ell=1}^{L} \pi_{(\ell)} \, \Phi(\mu_\ell(x) - \tau)$$

which, for our variable power allocation with allowance for leveling, takes the form

$$\sum_{\ell=1}^{L} \pi_{(\ell)} \, \Phi\big(\mu(x, \max\{u_\ell, u_{cut}\}\mathcal{C}'/R)\big),$$

with $u_\ell = e^{-2\mathcal{C}\frac{\ell-1}{L}}$. From the above expression for $\pi_{(\ell)}$, this $g_L(x)$ is equal to

$$\frac{2\mathcal{C}}{\nu L} \sum_{\ell=1}^{L} \frac{\max\{u_\ell, u_{cut}\}}{1+\delta_{sum}^2} \, \Phi\big(\mu(x, \max\{u_\ell, u_{cut}\}\mathcal{C}'/R)\big).$$

Recognize that this sum corresponds closely to an integral. In each interval $\frac{\ell-1}{L} \le t < \frac{\ell}{L}$ for ℓ from 1 to L, we have $e^{-2\mathcal{C}\frac{\ell-1}{L}}$ at least $e^{-2\mathcal{C}t}$. Consequently, $g_L(x)$ is greater than $g_{num}(x)/(1+\delta_{sum}^2)$ where the numerator $g_{num}(x)$ is the integral

$$\frac{2\mathcal{C}}{\nu} \int_0^1 \max\{e^{-2\mathcal{C}t}, u_{cut}\} \Phi\big(\mu(x, \max\{e^{-2\mathcal{C}t}, u_{cut}\}\mathcal{C}'/R)\big) dt.$$

Accordingly, the quantity of interest $g_L(x)$ has value at least $(integ/sum)g(x)$ where

$$g(x) = \frac{g_{num}(x)}{1+D(\delta_c)/snr}.$$

Using

$$\frac{integ}{sum} \; = 1 \; - \; \frac{2\theta\mathcal{C}}{L\,\eta} \, \frac{1}{1+\delta_{sum}^2}$$

and using that $g_L(x) \le 1$ and hence $g_{num}(x)/(1+\delta_{sum}^2) \le 1$ it follows that

$$g_L(x) \ge g(x) - 2\mathcal{C}/(L\nu).$$

The $g_L(x)$ and $g(x)$ are increasing functions of x on $[0,1]$.

Let's provide further characterization and evaluation of the integral $g_{num}(x)$ for our variable power allocation. Let $z_x^{low} = \mu(x, u_{cut}\,\mathcal{C}'/R)$ and $z_x^{max} = \mu(x, \mathcal{C}'/R)$. These have $z_x^{low} \le z_x^{max}$, with equality only in the constant power case (where $u_{cut} = 1$). For emphasis we write out that $z_x = z_x^{low}$ takes the form

$$z_x = \left[\frac{\sqrt{u_{cut}\mathcal{C}'/R}}{\sqrt{1 \, - \, x\nu}} - 1\right]\tau.$$

Set $u_x = 1 - x\nu$.

***Lemma** 12: Integral evaluation.* The $g_{num}(x)$ for has a representation as the integral with respect to the standard normal density $\phi(z)$ of the function that takes the value $1{+}D(\delta_c)/snr$ for z less than z_x^{low}, takes the value

$$x \;+\; \frac{u_x}{\nu}\left(1-\frac{R}{\mathcal{C}'}\left(1+z/\tau\right)^2\right)$$

for z between z_x^{low} and z_x^{max}, and takes the value 0 for z greater than z_x^{max}. This yields $g_{num}(x)$ is equal to

$$\begin{aligned} &\big[1+D(\delta_c)/snr\big]\,\Phi(z_x^{low}) \\ &+\Big[x+\delta_R\frac{u_x}{\nu}\Big]\left[\Phi(z_x^{max})-\Phi(z_x^{low})\right] \\ &+\frac{2R}{\mathcal{C}'}\frac{u_x}{\nu}\,\frac{[\phi(z_x^{max})-\phi(z_x^{low})]}{\tau} \\ &+\frac{R}{\mathcal{C}'}\frac{u_x}{\nu}\,\frac{[z_x^{max}\phi(z_x^{max})-z_x^{low}\phi(z_x^{low})]}{\tau^2}, \end{aligned}$$

where

$$\delta_R = 1-\frac{R}{\mathcal{C}'}\left[1+1/\tau^2\right].$$

This δ_R is non-negative if $R \le \mathcal{C}'/(1+1/\tau^2)$.

In the constant power case, corresponding to $u_{cut}=1$, the conclusion is consistent with the simpler $g(x)=\Phi(z_x)$.

The integrand above has value near $x+\big(1{-}R/\mathcal{C}'\big)u_x/\nu$, if z is not too far from 0. The heart of the matter for our analysis in this section is that this value is at least x for rates $R \le \mathcal{C}'$.

Proof of Lemma 12: By definition, the function $g_{num}(x)$ is

$$\frac{2\mathcal{C}}{\nu}\int_0^1 \max\{e^{-2\mathcal{C}t},u_{cut}\}\,\Phi\big(\mu(x,\max\{e^{-2\mathcal{C}t},u_{cut}\}\mathcal{C}'/R)\big)dt,$$

which is equal to the integral

$$\frac{2\mathcal{C}}{\nu}\int_0^{t_{cut}} e^{-2\mathcal{C}t}\,\Phi\big(\mu(x,e^{-2\mathcal{C}t}\mathcal{C}'/R)\big)\,dt$$

plus the expression

$$\frac{2\mathcal{C}}{\nu}(1{-}t_{cut})u_{cut}\Phi(z_x^{low}),$$

which can also be written as $[\delta_c+D(\delta_c)]\Phi(z_x^{low})/snr$.

Change the variable of integration from t to $u=e^{-2\mathcal{C}t}$, to produce the simplified expression for the integral

$$\frac{1}{\nu}\int_{u_{cut}}^1 \Phi\big(\mu(x,u\mathcal{C}'/R)\big)du.$$

Add and subtract the value $\Phi(z_x^{low})$ in the integral to write it as $[(1{-}u_{cut})/\nu]\Phi(z_x^{low})$, which is $[1{-}\delta_c/snr]\Phi(z_x^{low})$, plus the integral

$$\frac{1}{\nu}\int_{u_{cut}}^1 \left[\Phi\big(\mu(x,u\mathcal{C}'/R)\big)-\Phi\big(\mu(x,u_{cut}\mathcal{C}'/R)\big)\right]du.$$

Now since

$$\Phi(b)-\Phi(a)=\int 1_{\{a<z<b\}}\,\phi(z)\,dz,$$

it follows that this integral equals

$$\int\!\!\int 1_{\left\{u_{cut}\le u\le 1\right\}}1_{\left\{z_x^{low}\le z\le\mu(x,u\,\mathcal{C}'/R)\right\}}\,\phi(z)\,dz\,du/\nu.$$

We switch the order of integration. In the integral, the inequality $z\le\mu(x,u\,\mathcal{C}'/R)$ is the same as

$$u \;\ge\; u_x R/\mathcal{C}'\left(1+z/\tau\right)^2,$$

which exceeds u_{cut} for z greater than z_x^{low}. Here $u_x=1{-}x\nu$. This determines an interval of values of u. For z between z_x^{low} and z_x^{max} the length of this interval of values of u is equal to

$$1 \;-\; (R/\mathcal{C}')\,u_x\left(1+z/\tau\right)^2.$$

Using $u_x=1-x\nu$ one sees that this interval length, when divided by ν, may be written as

$$x \;+\; \frac{u_x}{\nu}\left(1-\frac{R}{\mathcal{C}'}\left(1+z/\tau\right)^2\right),$$

a quadratic function of z.

Integrate with respect to $\phi(z)$. The resulting value of $g_{num}(x)$ may be expressed as

$$\begin{aligned}&[1+D(\delta_c)/snr]\,\Phi(z_x^{low})\;+\\ &\qquad\frac{1}{\nu}\int_{z_x^{low}}^{z_x^{max}}\left[1-(R/\mathcal{C}')u_x\left(1+z/\tau\right)^2\right]\phi(z)dz,\end{aligned}$$

To evaluate, expand the square $\left(1+z/\tau\right)^2$ in the integrand as $1+2\,z/\tau+z^2/\tau^2$. Multiply by $\phi(z)$ and integrate. For the term linear in z, use $z\phi(z)=-\phi'(z)$ for which its integral is a difference in values of $\phi(z)$ at the two end points. Likewise, for the term involving $z^2=1+(z^2-1)$, use $(z^2-1)=-(z\phi(z))'$ which integrates to a difference in values of $z\phi(z)$. Of course the constant multiples of $\phi(z)$ integrate to a difference in values of $\Phi(z)$. The result for the integral matches what is stated in the Lemma. This completes the proof of Lemma 12.

One sees that the integral $g_{num}(x)$ may also be expressed as

$$\begin{aligned}&\frac{1}{snr}[\delta_c+D(\delta_c)]\Phi(z_x)\;+\\ &\qquad\frac{1}{\nu}\int\left[1-\max\left\{u_x\frac{R}{\mathcal{C}'}\left(1+z/\tau\right)_+^2,u_{cut}\right\}\right]_+\phi(z)dz.\end{aligned}$$

To reconcile this form with the integral given in the Lemma one notes that the integrand here for z below z_x takes the form of a particular constant value times $\phi(z)$ which, when integrated, provides a contribution that adds to the term involving $\Phi(z_x)$.

***Corollary** 13: Derivative evaluation.* The derivative $g_{num}'(x)$ is equal to

$$\frac{\tau}{2}\left(1{+}\frac{z_x}{\tau}\right)^3\phi(z_x)\frac{R}{\mathcal{C}'}\log(1{+}\delta_c)+\int_{z_x}^{z_x^{max}}\frac{R}{\mathcal{C}'}\left(1{+}z/\tau\right)^2\phi(z)dz.$$

In particular if $\delta_c=0$ the derivative $g'(x)$ is

$$\frac{R}{\mathcal{C}'}\int_{z_x^{low}}^{z_x^{max}}\left(1+z/\tau\right)^2\phi(z)dz,$$

and then, if also $R=\mathcal{C}'/(1{+}r/\tau^2)$ with $r\ge 1$, that is, if $R\le\mathcal{C}'/(1+1/\tau^2)$, the difference $g(x)-x$ is a decreasing function of x.

Proof: Consider the last expression given for $g_{num}(x)$. The part $[\delta_c + D(\delta_c)]\Phi(z_x)/snr$ has derivative

$$z_x'[\delta_c + D(\delta_c)]\phi(z_x)/snr.$$

Use $(1+z_x/\tau) = \sqrt{u_{cut}(\mathcal{C}'/R)/(1-x\nu)}$ to evaluate z_x' as

$$z_x' = \frac{\nu}{2}\frac{1}{(1-x\nu)^{3/2}}\sqrt{u_{cut}\mathcal{C}'/R}\,\tau$$

and obtain that it is $(\nu/2)(1+z_x/\tau)^3\tau/\big(u_{cut}\,\mathcal{C}'/R\big)$. So using $u_{cut} = (1-\nu)(1+\delta_c)$ the z_x' is equal to

$$\frac{snr}{2}(1+z_x/\tau)^3\frac{\tau}{(1+\delta_c)}\frac{R}{\mathcal{C}'}.$$

Thus using the form of $D(\delta_c)$ and simplifying, the derivative of this part of g_{num} is the first part of the expression stated in the Lemma.

As for the integral in the expression for g_{num}, its integrand is continuous and piecewise differentiable in x, and the integral of its derivative is the second part of the expression in the Lemma. Direct evaluation confirms that it is the derivative of the integral.

In the $\delta_c = 0$ case, this derivative specializes to the indicated expression which is less than

$$\frac{R}{\mathcal{C}'}\int_{-\infty}^{\infty}\left(1+z/\tau\right)^2\phi(z)dz = \frac{R}{\mathcal{C}'}\left[1+1/\tau^2\right],$$

which by the choice of R is less than 1. Then $g(x) - x$ is decreasing as it has a negative derivative. This completes the proof of Corollary 13.

***Corollary** 14: A lower bound.* The $g_{num}(x)$ is at least $g_{low}(x)$ given by

$$\big[1 + D(\delta_c)/snr\big]\Phi(z_x) + \\ \frac{1}{\nu}\int_{z_x^{low}}^{\infty}\Big[1 - (R/\mathcal{C}')u_x\left(1+z/\tau\right)^2\Big]\,\phi(z)dz.$$

It has the analogous integral characterization as given immediately preceding Corollary 13, but with removal of the outer positive part restriction. Moreover, the function $g_{low}(x) - x$ may be expressed as

$$g_{low}(x) - x \;=\; (1-x\nu)\frac{R}{\nu\,\mathcal{C}'}\frac{A(z_x)}{\tau^2}$$

where

$$\frac{A(z)}{\tau^2} \;=\; \frac{\mathcal{C}'}{R} - \left(1+1/\tau^2\right) - \frac{2\tau\phi(z)+z\phi(z)}{\tau^2} \\ +\left[1+1/\tau^2 - (1-\Delta_c)(1+z/\tau)^2\right]\Phi(z).$$

with $\Delta_c = \log(1+\delta_c)$.

Optionally, the expression for $g_{low}(x) - x$ may be written entirely in terms of $z = z_x$ by noting that

$$(1-x\nu)\frac{R}{\nu\,\mathcal{C}'} \;=\; \frac{(1+\delta_c)}{snr(1+z/\tau)^2}.$$

Proof: The integral expressions for $g_{low}(x)$ are the same as for $g_{num}(x)$ except that the upper end point of the integration extends beyond z_x^{max}, where the integrand is negative, i.e., the outer restriction to the positive part is removed. The lower bound conclusion follows from this negativity of the integrand above z_x^{max}. Evaluate $g_{low}(x)$ as in the proof of Lemma 12, using for the upper end that $\Phi(z)$ tends to 1, while $\phi(z)$ and $z\phi(z)$ tend to 0 as $z \to \infty$, to obtain that $g_{low}(x)$ is equal to

$$\Big[1 + D(\delta_c)/snr\Big]\Phi(z_x) \;+\; \Big[x + \delta_R\frac{u_x}{\nu}\Big]\left[1-\Phi(z_x)\right] \\ -\,2\frac{R}{\mathcal{C}'}\frac{u_x}{\nu}\frac{\phi(z_x)}{\tau} \;-\; \frac{R}{\mathcal{C}'}\frac{u_x}{\nu}\frac{z_x\phi(z_x)}{\tau^2}.$$

Replace the $x + \delta_R u_x/\nu$ with the equivalent expression $(1/\nu)\big[1-u_x(R/\mathcal{C}')(1+1/\tau^2)\big]$. Group together the terms that are multiplied by $u_x(R/\mathcal{C}')/\nu$ to be part of A/τ^2. Among what is left is $1/\nu$. Adding and subtracting x, this $1/\nu$ is $x + u_x/\nu$ which is $x + [(u_x/\nu)R/\mathcal{C}'][\mathcal{C}'/R]$. This provides the x term and contributes the $\mathcal{C}'/R$ term to A/τ^2.

It then remains to handle $[1{+}D(\delta_c)/snr]\Phi(z_x) - (1/\nu)\Phi(z_x)$ which is $-(1/snr)[1-D(\delta_c)]\Phi(z_x)$. Multiplying and dividing it by

$$\frac{\nu\,\mathcal{C}'}{u_x R} \;=\; \frac{snr(1+z/\tau)^2}{1+\delta_c}$$

and then noting that $(1-D(\delta_c))/(1+\delta_c)$ equals $1-\Delta_c$, it provides the associated term of A/τ^2. This completes the proof of Corollary 14.

What we gain with this lower bound is simplification because the result depends only on $z_x = z_x^{low}$ and not also on z_x^{max}.

C. Values of $g(x)$ near 1*:*

From the expression for x in terms of z, we remark that when R is near $\mathcal{C}'$, the point $z = 0$ corresponds to a value of x near $1 - \delta_c/snr$. We use this relationship to establish reference values of x^* and z^* and to bound how close $g(x^*)$ is to 1.

A convenient choice of x^* satisfies $(1-x^*\nu)R = (1-\nu)\mathcal{C}'$. More flexible is to allow other values of x^* by choosing it along with a value r_1 to satisfy the condition

$$(1 - x^*\,\nu)R = (1-\nu)\mathcal{C}'/(1+r_1/\tau^2).$$

We also call the solution $x^* = x_{up}$. When r_1 is positive the x^* is increased. Negative r_1 is allowed as long as $r_1 > -\tau^2$, but we keep r_1 small compared to τ so that x^* remains near 1.

With the rate R taken to be not more than $\mathcal{C}'$, we write it as

$$R = \frac{\mathcal{C}'}{(1+r/\tau^2)}.$$

***Lemma** 15: A value of x^* near* 1. Let $R' = \mathcal{C}'/(1+r_1/\tau^2)$. For any rate R between $R'/(1+snr)$ and R', the x^* as defined above is between 0 and 1 and satisfies

$$1 - x^* \;=\; \frac{R'-R}{R\,snr} \;=\; \frac{r-r_1}{snr\,(\tau^2+r_1)} = (1-x^*\nu)\,\frac{r-r_1}{\nu\,(\tau^2+r)}.$$

It is near 1 if R is near R'. The value of z_x at x^*, denoted $z^* = \zeta$ satisfies

$$(1+\zeta/\tau)^2 = (1+\delta_c)(1+r_1/\tau^2).$$

This relationship has δ_c near $2\zeta/\tau$, when ζ and r_1 are small in comparison to τ. The $\delta_c\tau$ and r_1 are arranged, usually both positive, and of the order of a power of a logarithm of τ, just large enough that $\bar{\Phi}(\zeta) = 1 - \Phi(\zeta)$ contributes to a small shortfall, yet not so large that it overly impacts the rate.

Proof of Lemma 15: The expression $1 - x\,\nu$ may also be written $(1{-}\nu){+}(1{-}x)\nu$. So the above condition may be written $1{+}(1{-}x^*)snr = R'/R$ which yields the first two equalities. It also may be written $(1{-}x^*\nu) = (1{-}\nu)(1{+}r/\tau^2)/(1{+}r_1/\tau^2)$ which yields the third equality in that same line.

Next recall that $z_x = \mu(x, u_{cut}\mathcal{C}'/R)$ which is

$$\big(\sqrt{u_{cut}\mathcal{C}'/(R(1-x\nu))} - 1\big)\tau.$$

Recalling that $u_{cut} = (1{-}\nu)(1+\delta_c)$, at x^* it is $z^* = \zeta$ given by

$$\zeta = \big(\sqrt{(1{+}\delta_c)(1+r_1/\tau^2)} - 1\big)\,\tau,$$

or, rearranging, we may express δ_c in terms of $z^* = \zeta$ and r_1 via

$$1{+}\delta_c = (1{+}\zeta/\tau)^2/(1{+}r_1/\tau^2),$$

which is the last claim. This completes the proof of Lemma 15.

Because of this relationship one may just as well arrange u_{cut} in the first place via ζ as

$$u_{cut} = (1{-}\nu)(1+\zeta/\tau)^2/(1{+}r_1/\tau^2)$$

where suitable choices for ζ and r_1 will be found in an upcoming section. We keep also the δ_c formulation as it is handy in expressing the affect on normalization via $D(\delta_c)$.

We come now to the evaluation of $g_L(x^*)$ and its lower bound via $g_{low}(x^*)$. Since $g_L(x)$ depends on x and R only via the expression $(1 - x\nu)R/\mathcal{C}'$, the choice of x^* such that this expression is fixed at $(1{-}\nu)$ implies that the value $g_L(x^*)$ is invariant to R, depending only on the remaining parameters snr, δ_c, r_1, τ and L. Naturally then, the same is true of our lower bound via $g_{low}(x^*)$ which depends only on snr, δ_c, r_1 and τ.

***Lemma** 16: The value $g(x^*)$ is near* 1. For the variable power case with $0 \le \delta_c < snr$, the shortfall expressed by $1{-}g(x^*)$ is less than

$$\delta^* = \frac{2\tau\phi(\zeta) + \zeta\phi(\zeta) + rem}{snr\,(\tau^2 + r_1)\,[1 + D(\delta_c)/snr]},$$

independent of the rate $R \le R'$, where the remainder is given by

$$rem = \big[(\tau^2 + r_1)D(\delta_c) - (r_1 - 1)\big]\bar{\Phi}(\zeta).$$

Moreover, $g_L(x^*)$ has shortfall $\delta_L^* = 1 - g_L(x^*)$ not more than $\delta^* + 2\mathcal{C}/(L\nu)$. In the constant power case, corresponding to $\delta_c = snr$, the shortfall is

$$\delta^* = 1 - \Phi(\zeta) = \bar{\Phi}(\zeta).$$

Setting

$$\zeta = \sqrt{2\log\frac{\tau}{d\sqrt{2\pi}}}$$

with a constant d and $\tau > d\sqrt{2\pi}$, with δ_c small, this δ^* is near $2d/(snr\,\tau^2)$, whereas, with $\delta_c = snr$, using $\bar{\Phi}(\zeta) \le \phi(\zeta)/\zeta$, it is not more than $d/(\zeta\tau)$.

Proof of Lemma 16: Using the lower bound on $g(x^*)$, the shortfall has the lower bound

$$\delta^* = 1 - \frac{g_{low}(x^*)}{1{+}D(\delta_c)/snr}$$

which equals

$$\frac{1 - g_{low}(x^*) + D(\delta_c)/snr}{1 + D(\delta_c)/snr}.$$

Use the formula for $g_{low}(x^*)$ in the proof of Lemma 14. For this evaluation note that at $x = x^*$ the expression $u_x R/(\nu\mathcal{C}')$ simplifies to $1/[snr(1{+}r_1/\tau^2)]$ and the expression $x{+}\delta_R u_x/\nu$ becomes

$$1 + \frac{r_1 - 1}{snr(\tau^2 + r_1)}.$$

Consequently, $g_{low}(x^*)$ equals

$$1 \,+\, \frac{D(\delta_c)}{snr}\Phi(\zeta) \,-\, \frac{2\tau\phi(\zeta) + \zeta\phi(\zeta) - (r_1{-}1)\bar{\Phi}(\zeta)}{snr(\tau^2 + r_1)}.$$

This yields an expression for $1 - g_{low}(x^*) + D(\delta_c)/snr$ equal to

$$-\frac{D(\delta_c)}{snr}\bar{\Phi}(\zeta) \,+\, \frac{(2\tau{+}\zeta)\phi(\zeta) - (r_1{-}1)\bar{\Phi}(\zeta)}{snr(\tau^2{+}r_1)}.$$

Group the terms involving $\bar{\Phi}(\zeta)$ to recognize that this equals $[(2\tau + \zeta)\phi(\zeta) + rem]/[snr(\tau^2 + r_1)]$. Then dividing by the expression $1 + D(\delta_c)/snr$ produces the claimed bound.

As for evaluation at the choice $\zeta = \sqrt{2\log\tau/d\sqrt{2\pi}}$, this is the positive value for which $\phi(\zeta) = d/\tau$, when $\tau \ge d\sqrt{2\pi}$. It provides the main contribution with $2\tau\phi(\zeta) = 2d$. The $\zeta\phi(\zeta)$ is then $\zeta d/\tau$ which is of order $\sqrt{\log\tau}/\tau$, small compared to the main contribution $2d$.

For the remainder rem, using $\bar{\Phi}(\zeta) \le \phi(\zeta)/\zeta$ and $D(\delta_c) \le (\delta_c)^2/2$ near $2\zeta^2/\tau^2$, the $\tau^2 D(\delta_c)\bar{\Phi}(\zeta)$ is near $2\zeta\phi(\zeta) = 2\zeta b_0/\tau$, again of order $\sqrt{\log\tau}/\tau$.

For $\delta_L^* = 1 - g_L(x^*)$, using $g_L(x) \ge g(x) + 2\mathcal{C}/(L\nu)$ yields $\delta_L^* \le \delta^* + 2\mathcal{C}/(L\nu)$.

For the constant power case we use $g_L(x^*) = g(x^*) = \Phi(\zeta)$ directly, rather than $g_{low}(x^*)$. It has $\delta^* = \bar{\Phi}(\zeta)$, which is not more than $\phi(\zeta)/\zeta$. This completes the proof of Lemma 16.

***Corollary** 17: Mistake bound.* The likely bound on the weighted fraction of failed detections and false alarms $\delta_L^* {+} \eta {+} \bar{f}$, corresponds to an unweighted fraction of not more than

$$\delta_{mis} = \text{fac}\,(\delta_L^* {+} \eta {+} \bar{f})$$

where the factor

$$\text{fac} = snr\,(1{+}\delta_{sum}^2)/[2\mathcal{C}(1{+}\delta_c)].$$

In the variable power case the contribution $\delta_{mis,L}^* = \text{fac}\,\delta_L^*$ is not more than $\delta_{mis}^* + (1/L)(1{+}snr)/(1{+}\delta_c)$ with

$$\delta_{mis}^* = \frac{(2\tau{+}\zeta)\phi(\zeta) + rem}{2\mathcal{C}\,(\tau{+}\zeta)^2},$$

while, in the constant power case $\delta_c = snr$, the fac $= 1$ and $\delta^*_{mis,L}$ equals

$$\delta^*_{mis} = \bar{\Phi}(\zeta).$$

Closely related to δ^*_{mis} in the variable power case is the simplified form

$$\delta^*_{mis,simp} = [(2\tau{+}\zeta)\phi(\zeta) + rem]/2\mathcal{C}\,\tau^2,$$

for which $\delta^*_{mis} = \delta^*_{mis,simp}/(1+\zeta/\tau)^2$.

Proof of Corollary 17: Multiplying the weighted fraction by the factor $1/[L \min_\ell \pi_{(\ell)}]$, which equals the given fac, provides the upper bound on the (unweighted) fraction of mistakes $\delta_{mis} = \text{fac}\,(\delta^*_L + \eta + \bar{f})$. Now $\delta^*_L = 1 - g_L(x)$ has the upper bound

$$\frac{1 - g_{low}(x^*) + \delta^2_{sum}}{1 + \delta^2_{sum}}.$$

Multiplying by fac yields $\delta^*_{mis,L} = \text{fac}\,\delta^*_L$ not more than

$$\frac{1 - g_{low}(x^*) + \delta^2_{sum}}{(2\mathcal{C}/snr)(1{+}\delta_c)}.$$

Recall that δ^2_{sum} exceeds $D(\delta_c)/snr$ by not more than $2\mathcal{C}/(L\nu)$ and that $1 - g_{low}(x^*) + D(\delta_c)/snr$ is less than $[(2\tau{+}\zeta)\phi(\zeta) + rem]/[snr(\tau^2{+}r_1)]$. So this yields the $\delta^*_{mis,L}$ bound

$$\frac{(2\tau{+}\zeta)\phi(\zeta) + rem}{2\mathcal{C}(\tau^2{+}r_1)(1{+}\delta_c)} + \frac{(1{+}snr)}{(1{+}\delta_c)}\,\frac{1}{L}.$$

Recognizing that the denominator product $(\tau^2+r_1)(1+\delta_c)$ simplifies to $(\tau+\zeta)^2$ establishes the claimed form of δ^*_{mis}.

For the constant power case note that fac $= 1$ so that $\delta^*_{mis,L} = \delta^*_{mis}$ is then unchanged from $\delta^* = \bar{\Phi}(\zeta)$. This completes the proof of Corollary 17.

D. Showing $g(x)$ is greater than x:

This section shows that $g_L(x)$ is accumulative, that is, it is at least x for the interval from 0 to x^*, under certain conditions on r.

We start by noting the size of the gap at $x = x^*$.

***Lemma** 18: The gap at x^*.* With rate $R = \mathcal{C}'/(1 + r/\tau^2)$, the difference $g(x^*) - x^*$ is at least

$$\frac{r - r_{up}}{snr(\tau^2 + r_1)} = (1 - x^*\nu)\,\frac{r - r_{up}}{\nu(\tau^2 + r)}.$$

Here, for $0 \le \delta_c < snr$, with rem as given in Lemma 16,

$$r_{up} \;=\; r_1 \,+\, \frac{(2\tau + \zeta)\phi(\zeta) + rem}{1 + D(\delta_c)/snr}$$

while, for $\delta_c = snr$,

$$r_{up} = r_1 + snr(\tau^2 + r_1)\bar{\Phi}(\zeta),$$

which satisfies

$$\frac{r_{up}}{\tau^2} \;=\; \frac{(1 + snr\,\bar{\Phi}(\zeta))(1{+}\zeta/\tau)^2}{1 + snr} \;-\; 1.$$

Keep in mind that the rate target $\mathcal{C}'$ depends on δ_c. For small δ_c it is near the capacity $\mathcal{C}$, whereas for $\delta_c = snr$ it is near $R_0 > 0$.

If the upcoming gap properties permit, it is desirable to set r near r_{up}. Then the factor in the denominator of the rate becomes near $1{+}r_{up}/\tau^2$. In some cases r_{up} is negative, permitting $1{+}r/\tau^2$ not more than 1.

We remind that r_1, ζ, and δ_c are related by

$$1{+}r_1/\tau^2 \;=\; \frac{(1{+}\zeta/\tau)^2}{1{+}\delta_c}.$$

Proof of Lemma 18: The gap at x^* equals $g(x^*) - x^*$. This value is the difference of $1{-}x^*$ and $\delta^* = 1{-}g(x^*)$, for which we have the bounds of the previous two lemmas. Recalling that $1{-}x^*$ equals $(r{-}r_1)/[snr\,(\tau^2{+}r_1)]$, adjust the subtraction of r_1 to include in r_{up} what is needed to account for δ^* to obtain the indicated expressions for $g(x^*) - x^*$ and r_{up}. Alternative expressions arise by using the relationship that r_1 has to the other parameters. This complete the proof of Lemma 18.

Positivity of this gap at x^* entails $r > r_{up}$, and positivity of x^* requires $snr(\tau^2 + r_1) + r_1 \ge r$. There is an interval of such r provided $snr\,(\tau^2{+}r_1) > r_{up} - r_1$.

For this next two corollaries, we take the case that either $\delta_c = snr$ or $\delta_c = 0$, that is, either the power allocation is constant (completely level), or the power $P_{(\ell)}$ is proportional to $u_\ell = \exp\{-2\mathcal{C}\frac{\ell-1}{L}\}$, unmodified (no leveling). The idea in both cases is to look for whether the minimum of the gap occurs at x^* under stated conditions.

***Corollary** 19: Positivity of $g(x) - x$ with constant power.* Suppose $R = \mathcal{C}'/(1{+}r/\tau^2)$ where, with constant power, the $\mathcal{C}'$ equals $R_0(1{-}h')/(1{+}\delta_a)^2$, and suppose $\nu\,\tau \ge 2(1{+}r/\tau^2)\sqrt{2\pi}$. Suppose $r - r_{up}$ is positive with r_{up} as given in Lemma 18, specific to this $\delta_c = snr$ case. If $r \ge 0$ and if $r - r_{up}$ is less than $\nu(\tau + \zeta)^2/2$, then, for $0 \le x \le x^*$, the difference $g(x) - x$ is at least

$$gap = \frac{r - r_{up}}{snr(\tau^2{+}r_1)} = \frac{r - r_{up}}{\nu(\tau + \zeta)^2},$$

Whereas if $r_{up} < r \le 0$ and if also

$$r/\tau \ge -\sqrt{2\log\left(\nu\tau(1{+}r/\tau^2)/2\sqrt{2\pi}\right)},$$

then the gap $g(x) - x$ on $[0, x^*]$ is at least

$$\min\left\{\, 1/2 + r/(\tau\sqrt{2\pi}\,)\;,\; \frac{r - r_{up}}{\nu(\tau{+}\zeta)^2} \right\}.$$

In the latter case the minimum occurs at the second expression when

$$r < r_{up} + \nu(\tau{+}\zeta)^2\big[1/2 + r/\tau\sqrt{2\pi}\,\big].$$

This corollary is proven in the appendix, where, under the stated conditions, it is shown that $g(x) - x$ is unimodal for $x \ge 0$, so the value is smallest at $x{=}0$ or $x{=}x^*$.

From the formula for r_{up} in this constant power case, it is negative, near $-\nu\,\tau^2$, when $snr\,\bar{\Phi}(\zeta)$ and ζ/τ are small. It is tempting to try to set r close to r_{up}, similarly negative. As discussed in the appendix, the conditions prevent pushing r too negative and compromise choices are available. With $\nu\tau$ at least a little more than the constant $2\sqrt{2\pi}e^{\pi/4}$, we allow r with which the $1 + r/\tau^2$ factor becomes at best near $1 - \sqrt{2\pi}/2\tau$, indeed nice that it is not more than 1, though not as ambitious as the unobtainable $1 + r_{up}/\tau^2$ near $1-\nu$.

***Corollary** 20: Positivity of $g(x) - x$ with no leveling.* Suppose $R = \mathcal{C}'/(1{+}r/\tau^2)]$, where, with $\delta_c = 0$, the $\mathcal{C}'$ equals $\mathcal{C}(1h')/(1{+}2\mathcal{C}/\nu\, L)(1{+}\delta_a)^2$ near capacity. Set $r_1 = 0$ and $\zeta = 0$ for which $1{-}x^* = r/(snr\, \tau^2)$ and $r_{up} = 2\tau/\sqrt{2\pi}{+}1/2$ and suppose in this case that $r > r_{up}$. Then, for $0 \le x \le x^*$, the difference $g(x) - x$ is greater than or equal to

$$gap = \frac{r - r_{up}}{snr\,(\tau^2 + r_1)}.$$

Moreover, $g(x) - x$ is at least $(1 - x\nu)GAP$ where

$$GAP = \frac{r - r_{up}}{\nu(\tau^2 + r)}.$$

Proof of Corollary 20: With $\delta_c = 0$ the choice $\zeta = 0$ corresponds to $r_1 = 0$. At this ζ, the main part of r_{up} equals $2\tau/\sqrt{2\pi}$ since $\phi(0) = 1/\sqrt{2\pi}$ and the remainder rem equals $1/2$ since $\Phi(0) = 1/2$. This produces the indicated value of r_{up}. The monotonicity of $g(x) - x$ in the $\delta_c = 0$ case yields, for $x \le x^*$, a value at least as large as at x^* where it is bounded by Lemma 18. This yields the first claim.

Next use the representation of $g(x) - x$ as $(1 - x\nu)A(z_x)/[\nu(\tau^2 + r)]$, where with $\delta_c = 0$ the $A(z)$ is

$$A(z) \;=\; r - 1 - 2\tau\phi(z) - z\phi(z) + \left[\tau^2 + 1 - (\tau{+}z)^2\right]\Phi(z).$$

It has derivative which simplifies to

$$A'(z) = -2(\tau{+}z)\Phi(z),$$

which is negative for $z > -\tau$ which includes the interval $[z_0, z_1]$. Accordingly $A(z_x)$ is decreasing and its minimum for x in $[0, x^*]$ occurs at x^*. Appealing to Lemma 18 completes the proof of Lemma 20.

The above result provides a practical rate $\mathcal{C}'/\big(1{+}r/\tau^2\big)$ with r/τ^2 at least r_{up}/τ^2 nearly equal to a constant times $1/\tau$, which is near $1/\sqrt{\pi \log B}$. Nevertheless, it would be better to have a bound with r_{up} of smaller order so that for large B the rate is closer to capacity. For that reason we next take advantage of the modification to the power allocation in which it is slightly leveled using a small positive δ_c.

Monotonicity or unimodality of $g(x) - x$ or of $g_{low}(x) - x$ is used in the above proofs for the $\delta_c = 0$ and $\delta_c = snr$ cases. It what follows we develop analogous shape properties that include the intermediate case.

We use $g(x) \ge g_{low}(x)/(1 + D(\delta_c)/snr)$ so that

$$g(x) - x \ge \frac{g_{low}(x) - x - xD(\delta_c)/snr}{1 + D(\delta_c)/snr}.$$

This gap lower bound is expressible in terms of $z = z_x$ using the results of Lemma 14 and the expression for x given immediately thereafter. We have

$$g_{low}(x) - x = \frac{u_x\, R}{\nu\; \mathcal{C}'}\;\frac{A(z_x)}{\tau^2}$$

where for $R = \mathcal{C}'/(1 + r/\tau^2)$ the function $A(z)$ simplifies to

$$r - 1 - 2\tau\phi(z) - z\phi(z) + \left[\tau^2{+}1 - (1{-}\Delta_c)(\tau{+}z)^2\right]\Phi(z),$$

where $\Delta_c = \log(1 + \delta_c)$. The multiplier $u_x R/(\nu\mathcal{C}')$ is also $(1{+}\delta_c)/\big(snr(1{+}z/\tau)^2\big)$. From the expression for x in terms of z we may write

$$x = 1 - \frac{\delta_c}{snr} + \frac{(1{+}\delta_c)}{snr(1{+}z/\tau)^2}\left((1 + z/\tau)^2 - 1 - r/\tau^2\right).$$

Accordingly, we may write

$$g_{low}(x) - x - xD(\delta_c)/snr \;=\; G(z_x)$$

where $G(z)$ is the function

$$G(z) = \frac{1{+}\delta_c}{(\tau{+}z)^2}\,\frac{\tilde{A}(z)}{snr} - \frac{D(\delta_c)}{snr}(1 - \delta_c/snr)$$

with

$$\tilde{A}(z) \;=\; A(z) \;+\; \frac{\tau^2 D(\delta_c)}{snr}\left(1 - (1 + z/\tau)^2 + r/\tau^2\right).$$

In this way the gap lower bound is expressed through the function $G(z)$ evaluated at $z = z_x$. Regions for x in $[0,1]$ where $g_{low}(x) - x - xD(\delta_c)/snr$ is decreasing or increasing, have corresponding regions of decrease or increase of $G(z)$ in $[z_0, z_1]$. The following lemma characterizes the shape of the lower bound on the gap.

Definition: A continuous function $G(z)$ is said to be unimodal in an interval if there is a value z_{max} such that $G(z)$ is increasing for any values to the left of z_{max} and decreasing for any values to the right of z_{max}. This includes the case of decreasing or increasing functions with z_{max} at the left or right end point of the interval, respectively.

Likewise, with domain starting at z_0, a continuous function $G(z)$ is said to have at most *one oscillation* if there is a value $z_G \ge z_0$ such that $G(z)$ is decreasing for any values of z between z_0 and z_G, and unimodal to the right of z_G. We call the point z_G the *critical value* of G.

Functions with at most one oscillation in an interval $[z_0, z^*]$ have the useful conclusion that the minimum over the interval is determined by the minimum of the values at z_G and z^*.

***Lemma** 21: Shape properties of the gap.* Suppose the rate satisfies $R \le C'(1 + D(\delta_c)/snr)/(1 + 1/\tau^2)$. The function $g_{low}(x){-}x{-}xD(\delta_c)/snr$ has at most one oscillation in $[0,1]$. Likewise, the functions $A(z)$ and $G(z)$ have at most one oscillation for $z \ge -\tau$ and we denote their critical values z_A and z_G. For all $\Delta_c \ge 0$, these satisfy $z_A \le z_G$ and $z_A {\le} {-}\tau/2{+}1$, which is less than or equal to 0 if $\tau \ge 2$.

Moreover, if either $\Delta_c \le 2/3$ or $\Delta_c \ge 2\sqrt{2\pi}\, \textit{half}/\tau$, then z_G is also less than or equal to 0. Here *half* is an expression not much more than $1/2$ as given in the proof.

The proof of Lemma 21 is in the appendix.

Note that $\tau \ge 3\sqrt{2\pi}\, \textit{half}$ is sufficient to ensure that one or the other of the two conditions on Δ_c must hold. That would entail a value of B more than $e^{2.25\pi} > 1174$. Such size of B is reasonable, though not essential as we may choose directly to have a small value of Δ_c not more than $2/3$.

One can pin down down the location of z_G further, under additional conditions on Δ_c. However, precise knowledge of the value of z_G is not essential because the shape properties allow us to take advantage of tight lower bounds on $A(z)$ for negative z as discussed in the next lemma.

We have that $z_A \leq -\tau/2+1$ and under conditions on Δ_c that $z_A \leq -\tau/2$. For $-\tau/2+1$ to be negative, it is assumed that $\tau \geq 2$, as is the case when $B \geq e^2$. Preferably B is much larger.

***Lemma** 22: Lower bounding $A(z)$ for negative z:* In an initial interval $[-\tau, t]$, with $t=-\tau/2$ or $t=-\tau/2+1$, the function $A(z)$ is lower bounded by

$$A(z) \,\geq\, r-1-\epsilon,$$

where ϵ is $(2\tau+t)/t^2)\phi(t)$. In particular for $t=-\tau/2$ it is $(6/\tau)\phi(\tau/2)$, not more than $(3/\sqrt{\pi \log B})(1/B)^{0.5}$, polynomially small in $1/B$. Likewise, if $t=-\tau/2+1$, the ϵ remains polynomially small.

If $\Delta_c \tau \geq 4/\sqrt{2\pi}-1/\tau$, then the above inequality holds for all negative z,

$$\min_{-\tau \leq z \leq 0} A(z) \geq r-1-\epsilon$$

with $\epsilon=(6/\tau)\phi(\tau/2)$.

Finally if also $\Delta_c \geq 8/\tau^2$, then for z between $-\tau/2$ and 0, we have

$$A(z) > r-1$$

strictly greater than $r-1$ with no need for ϵ.

Proof of Lemma 22: First, examine $A(z)$ for z in an initial interval of the form $[-\tau, t]$. For such negative z one has that $A(z)$ is at least $r-1-2\tau\phi(z)$ which is at least $r-1-2\tau\phi(t)$. This is seen by observing that in the expression for $A(z)$, the $-z\phi(z)$ term and the term involving $\Phi(z)$ are positive for $z \leq 0$. So for ϵ we could use $2\tau\phi(t)$.

Further analysis of $A(z)$ permits the improved value of ϵ as stated in the lemma. Indeed, $A(z)$ may be expressed as

$$A_0(z)=r-1-(2\tau{+}z)\phi(z)-(2\tau{+}z)z\Phi(z)+\Phi(z)$$

plus an additional amount $[\Delta_c(\tau{+}z)^2]\Phi(z)$ which is positive. It's derivative simplifies as in the analysis in the previous lemma and it is less than or equal to 0 for $-\tau \leq z \leq 0$, so $A_0(z)$ is a decreasing function of z, so its minimum in $[-\tau, t]$ occurs at $z=t$.

Recall that along with the upper bound $|z|\Phi(z) \leq \phi(z)$, there is the lower bound of Feller, $|z|\Phi(z) \geq [1{-}1/z^2]\phi(z)$, or the improvement in the appendix which yields $|z|\Phi(z) \geq [1-1/(z^2{+}1)]\phi(z)$, which is $\Phi(z) \geq (|z|/(z^2{+}1)\phi(z)$, for negative z. Accordingly obtain

$$A_0(z) \geq r-1-[(2\tau{+}z)/(z^2{+}1)]\phi(z).$$

At $z=t=-\tau/2$ the amount by which it is less than $r-1$ is $[(3/2)\tau/(\tau^2/4{+}1)]\phi(\tau/2)$ not more than $(6/\tau)\phi(\tau/2)$, which is not more than $(6/\sqrt{2\pi 2\log B})(1/B)^{1/2}$. An analogous bound holds at $t=-\tau/2+1$.

Next consider the value of $A(z)$ at $z=0$. Recall that $A(z)$ equals

$$r-1-(2\tau{+}z)\phi(z)+\big[\tau^2{+}1-(1{-}\Delta_c)(\tau{+}z)^2\big]\,\Phi(z).$$

At $z=0$ it is

$$r-1-2\tau/\sqrt{2\pi}+[1+\Delta_c\tau^2]/2$$

which is at least $r-1$ if $\Delta_c\tau^2 \geq 4\tau/\sqrt{2\pi}-1$, that is, if $\Delta_c \geq 4/(\tau\sqrt{2\pi}){-}1/\tau^2$. This is seen to be greater than $\Delta_c^{**}=2/(\tau^2/4+2)$, having assumed that τ at least 2. So by the previous lemma $A(z)$ is unimodal to the right of $t=-\tau/2$, and it follows that the bound $r{-}1{-}\epsilon$ holds for all z in $[-\tau, 0]$.

Finally, for $A(z)$ in the form

$$r-1-(2\tau{+}z)\phi(z)-(2\tau{+}z)z\Phi(z)+[1+\Delta_c(\tau{+}z)^2]\Phi(z),$$

replace $-z\Phi(z)$, which is $|z|\Phi(z)$ with its lower bound $\phi(z)-(1/|z|)\Phi(z)$ for negative z from the same inequality in the appendix. Then the terms involving $\phi(z)$ cancel and we are left with the lower bound on $A(z)$ of

$$r-1+\big[1+\Delta_c(\tau{+}z)^2-(2\tau{+}z)/|z|\big]\Phi(z)$$

which is

$$r-1+\big[\Delta_c(\tau{+}z)^2+2(\tau{+}z)/z\big]\Phi(z).$$

In particular at $z=-\tau/2$ it is $r{-}1+\big[\Delta_c\tau^2/4-2\big]\Phi(-\tau/2)$ which exceeds $r-1$ by a positive amount due to the stated conditions on Δ_c. To determine the region in which the expression in brackets is positive more precisely, proceed as follows. Factoring out $\tau{+}z$ the expression remaining in the brackets is

$$\Delta_c(\tau{+}z)+2/z.$$

It starts out negative just to the right of $-\tau$ and it hits 0 for z solving the quadratic $\Delta_c(\tau{+}z)z+2=0$, for which the left and right roots are $z=[-\tau \pm \sqrt{\tau^2-8/\Delta_c}]/2$, again centered at $-\tau/2$. The left root is near $-\tau\big[1-2/(\Delta_c\tau)\big]$. So at least between these roots, and in particular between the left root and the point $-\tau/2$, the $A(z) \geq r{-}1$. The existence of these roots is implied by $\Delta_c \geq 8/\tau^2$ which in turn is greater than $\Delta_c^{**}=8/(\tau^2{+}8)$. So by the analysis of the previous Lemma, $A'(z)$ is positive at $-\tau/2$ and $A(z)$ is unimodal to the right of $-\tau/2$. Consequently $A(z)$ remains at least $r{-}1$ for all z between the left root and 0. This completes the proof of Lemma 22.

Exact evaluation of $G(z_{crit})$ is problematic, so instead take advantage for negative z of the tight lower bounds on $G(z)$ that follow immediately from the above lower bounds on $A(z)$. With no conditions on Δ_c we use $A(z) \geq r-1-\epsilon$ for $z \leq -\tau/2{+}1$ and unimodality of $A(z)$ to the right of there, to allow us to combine this with the bounds at z^*. This use of unimodality of $A(z)$ has the slight disadvantage of needing to replace $u_x=1{-}x\nu$ with the lower bound $1{-}x^*\nu$, and needing to replace $-xD(\delta_c)/snr$ with $-x^*D(\delta_c)/snr$, to obtain the combined lower bound on $G(z)$ via $A(z)$. In contrast, with conditions on Δ_c, we use directly that the minimum of $G(z)$ occurs at the minimum of the values at a negative z_G and at z^*, allowing slight improvement on the gap.

***Lemma** 23: Lower bounding $G(z)$ for negative z:* If $\Delta_c\tau \geq 4/\sqrt{2\pi}{-}1/2\tau$, then for $-\tau < z \leq 0$, setting $z'=z(1{+}z/2\tau)$, the function $G(z)$ is at least

$$(1{+}\delta_c)\,\frac{r{-}1{-}\epsilon-\big(2z'\tau{-}r\big)D(\delta_c)/snr}{(\tau{+}z)^2\,snr}-\frac{D(\delta_c)}{snr}\big(1{-}\frac{\delta_c}{snr}\big),$$

which for $r \ge (1{+}\epsilon)/(1{+}D(\delta_c)/snr)$, yields $G(z)$ at least

$$\frac{r - 1 - \epsilon + rD(\delta_c)/snr}{\tau^2\, snr} - \frac{D(\delta_c)}{snr}\big(1 - \frac{\delta_c}{snr}\big).$$

Consequently, the gap $g(x) - x$ for $z_x \le 0$ is at least

$$\frac{r - r_{down}}{snr\, \tau^2/(1{+}\delta_c)}$$

with

$$r_{down} = \frac{1{+}\epsilon + \tau^2 D(\delta_c)(1 - \delta_c/snr)/(1{+}\delta_c)}{1 + D(\delta_c)/snr},$$

less than $1{+}\epsilon + \tau^2 D(\delta_c)$. If also $\Delta_c \ge 8/\tau^2$, then the above inequalities hold for $-\tau/2 \le z \le 0$ without the ϵ.

Proof of Lemma 23: Using the relationship between $G(z)$ and $A(z)$ given prior to Lemma 21, these conclusions follow immediately from plugging in the bounds on $A(z)$ from Lemma 22.

Next we combine the gap bounds for negative z with the gap bound for z^*. This allows us to show that $g(x) - x$ has a positive gap as long as the rate drop from capacity is such that $r > r_{crit}$ for a value of r_{crit} we identify. This holds for a range of choices of r_1 including 0.

***Lemma** 24: The minimum value of the gap.* For $0 \le x \le x^*$, if $r > r_{crit}$, then the $g(x) - x$ is at least

$$\frac{r - r_{crit}}{snr\,(\tau^2{+}r_1)}.$$

This holds for an r_{crit} not more than r^*_{crit} given by

$$\max\Big\{(\tau^2{+}r_1)D(\delta_c) + 1 + \epsilon\,,\, r_1 + (2\tau{+}\zeta)\phi(\zeta) + rem\Big\},$$

where, as before, $rem = \big[(\tau^2{+}r_1)D(\delta_c) + 1{-}r_1\big]\bar\Phi(\zeta)$ and ϵ is as given in Lemma 22 with $t = -\tau/2 + 1$. Then $g_L(x) - x$ on $[0, x^*]$ has gap at least

$$gap \;=\; \frac{r - r_{crit}}{snr\,(\tau^2{+}r_1)} \;-\; \frac{2\mathcal{C}}{\nu L}.$$

Consequently, any specified positive value of gap is achieved by setting

$$r \;=\; r_{crit} \;+\; snr\,(\tau^2\; + r_1)\; [gap + 2\mathcal{C}/(L\nu)].$$

The contribution to the denominator of the rate expression $(1{+}D(\delta_c)/snr)(1{+}r_{crit}/\tau^2)$ at r_{crit} has the representation in terms of r^*_{crit} as

$$1 \;+\; (1{+}r_1/\tau^2)D(\delta_c)/snr \;+\; r^*_{crit}/\tau^2.$$

If $\Delta_c \ge 4/(\tau\sqrt{2\pi}) - 1/\tau^2$ and either $\Delta_c \le 2/3$ or $\Delta_c \ge 2\sqrt{2\pi}\, \mathit{half}/\tau$, then in the above characterization of r^*_{crit} the $D(\delta_c)$ in the first expression of the max may be reduced to $D(\delta_c)(1 - \delta_c/snr)$.

Moreover, we have the refinement that $g(x) - x$ is at least

$$\frac{1}{snr}\,\min\Big\{\frac{r - r_{down}}{\tau^2/(1{+}\delta_c)}\,,\,\frac{r - r_{up}}{\tau^2 + r_1}\Big\},$$

where r_{down} and r_{up} are as given in Lemmas 23 and 18, respectively. If also δ_c is such that the z_G of order $-\sqrt{2\log(\tau/\delta_c)}$ is between $-\tau/2$ and 0, then the ϵ above may be omitted.

For given $\zeta > 0$ we adjust r_1 to optimize the value of r^*_{crit} in the next subsection.

The proof of the lemma will improve on the statement of the lemma by exhibiting an improved value of r_{crit} that makes use of r^*_{crit}.

Proof of Lemma 24: Replacing $g(x)$ by its lower bound $g_{low}(x)/[1{+}D(\delta_c)/snr]$, the $g(x) - x$ is at least

$$gap_{low}(x) = \frac{g_{low}(x) - x[1{+}D(\delta_c)/snr]}{1 + D(\delta_c)/snr}$$

which is

$$\frac{(u_x/\nu)(R/\mathcal{C}')A(z_x)/\tau^2 - xD(\delta_c)/snr}{1 + D(\delta_c)/snr}.$$

For $0 \le x \le x^*$ the $u_x R/(\nu\mathcal{C}')$ is at least its value at x^* which is $1/[snr(1 + r_1/\tau^2)]$, so $gap_{low}(x)$ is at least

$$\frac{A(z_x)/[snr(\tau^2{+}r_1)] - x^* D(\delta_c)/snr}{1 + D(\delta_c)/snr},$$

which may also be written

$$\frac{A(z_x) - (\tau^2{+}r_1)x^* D(\delta_c)}{snr(\tau^2{+}r_1)[1{+}D(\delta_c)/snr]},$$

which by Lemma 18 coincides with $(r{-}r_{up})/[snr(\tau^2{+}r_1)]$ at $x = x^*$.

Now recall from Lemma 21 that $A(z)$ is unimodal for $z \ge t$, where t is $-\tau/2$ or $-\tau/2 + 1$, depending on the value of Δ_c. As we have seen, when Δ_c is small the $A(z)$ is in fact decreasing and then we may use $r_{crit} = r_{up}$ from the gap at x^*. For other Δ_c, the unimodality of $A(z)$ for $z \ge t$ implies that the minimum of $A(z)$ over $[-\tau, z^*]$ is equal to that of over $[-\tau, t] \cup \{z^*\}$. As we saw in Lemma 22, the minimum of $A(z)$ in $[-\tau, t]$ is given by $A_{low} = r - 1 - \epsilon$. Consequently, the $g(x) - x$ on $0 \le x \le x^*$ is at least

$$\min\left\{\frac{r - 1 - \epsilon - (\tau^2 + r_1)x^* D(\delta_c)}{snr(\tau^2 + r_1)(1 + D(\delta_c)/snr)}\,,\,\frac{r - r_{up}}{snr(\tau^2 + r_1)}\right\}.$$

Now $x^* = 1 - (r - r_1)/[snr\,(\tau^2 + r_1)]$. So $(\tau^2 + r_1)x^*$ is equal to $(\tau^2 + r_1) - (r - r_1)/snr$. Then, gathering the terms involving r, note that a factor of $1 + D(\delta_c)/snr$ arises that cancels the corresponding factor from the denominator for the part involving r. Extract the value r shared by the two terms in the minimum to obtain that the above expression is at least

$$\frac{r - r_{crit}}{snr\,(\tau^2 + r_1)}$$

where here r_{crit} is given by

$$\max\Big\{\frac{(\tau^2{+}r_1)D(\delta_c) + 1 + \epsilon + r_1 D(\delta_c)/snr}{1 + D(\delta_c)/snr}\,, r_{up}\Big\}.$$

Arrange $1{+}D(\delta_c)/snr$ as a common denominator. From the definition of r_{up} its numerator becomes $r_1[1 + D(\delta_c)/snr] + (2\tau + \zeta)\phi(z) + rem$. It follows that in the numerator the two expressions in the max share the term $r_1 D(\delta_c)/snr$. Accordingly, with $\alpha = [D(\delta_c)/snr]/[1 + D(\delta_c)/snr]$ and $1 - \alpha = 1/[1 + D(\delta_c)/snr]$, we have

$$r_{crit} = \alpha\, r_1 + (1 - \alpha) r^*_{crit},$$

with r^*_{crit} given by

$$\max\Big\{(\tau^2\!+\!r_1)D(\delta_c)+1+\epsilon\,,\, r_1+(2\tau\!+\!\zeta)\phi(\zeta)+rem\Big\}.$$

This r^*_{crit} exceeds r_1, because the amount added to r_1 in the second expression in the max is the same as the numerator of the shortfall δ^* which is positive. Hence $\alpha r_1\!+\!(1\!-\!\alpha)r^*_{crit}$ is less than r^*_{crit}. So the r_{crit} here improves somewhat on the choice in the statement of the Lemma.

Moreover, from

$$r_{crit}=\frac{r^*_{crit}+r_1D(\delta_c)/snr}{1+D(\delta_c)/snr}$$

it follows that $(1\!+\!D(\delta_c)/snr)(1\!+\!r_{crit}/\tau^2)$ is equal to

$$1+\frac{(1\!+\!r_1/\tau^2)D(\delta_c)}{snr}+\frac{r^*_{crit}}{\tau^2},$$

as claimed.

Finally, for the last conclusion of the Lemma, it follows from the fact that

$$\min_{-\tau<z\le z^*}G(z)=\min\{G(z_G),G(z^*)\},$$

invoking $z_G\le 0$ and combining the bounds form Lemmas 23 and 18. This completes the proof of Lemma 24.

Note from the form of rem and using $1-\bar\Phi(\zeta)=\Phi(\zeta)$ that $r^*_{crit}-1$ may be written

$$\max\Big\{(\tau^2\!+\!r_1)D(\delta_c)+\epsilon\,,$$
$$(r_1\!-\!1)\Phi(\zeta)+(2\tau\!+\!\zeta)\phi(z)+(\tau^2\!+\!r_1)D(\delta_c)\bar\Phi(\zeta)\Big\}.$$

Thus $[(\tau^2\!+\!r_1)D(\delta_c)+1]$ appears both in the first expression and as a multiplier of $\bar\Phi(\zeta)$ in the remainder of the second expression in the max.

To clean the upcoming expressions, note that if we replace the second expression in this max with the bound in which we add the polynomially small ϵ to it, then $r^*_{crit}-1-\epsilon$ becomes independent of ϵ. Accordingly we henceforth make that redefinition of r^*_{crit}. Denoting $\tilde r^*_{crit}=r^*_{crit}-1-\epsilon$ it becomes

$$\max\Big\{(\tau^2\!+\!r_1)D(\delta_c)\,,$$
$$(r_1\!-\!1)\Phi(\zeta)+(2\tau\!+\!\zeta)\phi(z)+(\tau^2\!+\!r_1)D(\delta_c)\bar\Phi(\zeta)\Big\}.$$

Evaluation of the best r^*_{crit} arises in the next subsection from determination of the r_1 that minimizes it.

E. Determination of δ_c:

Here we determine suitable choices of the leveling parameter δ_c. As we know, $\delta_c=0$ corresponds to no-leveling and $\delta_c=snr$ corresponds to the constant power allocation, and both will have their role for very large and very small snr, respectively. Values in between are helpful in conjunction with controlling the rate drop parameter r_{crit}.

Recall the relationship $1+\delta_c=(1+\zeta/\tau)^2/(1+r_1/\tau^2)$, used in analysis of the gap based on $g_{low}(x)$, where ζ is the value of z_x at the upper end point x^* of the interval in which the gap property is invoked. In this subsection we hold ζ fixed and ask for the determination of a suitable choice of δ_c.

In view of the indicated relationship this is equivalent to the determination a choice of r_1. There are choices that arise in obtaining manageable bounds on the rate drop. One is to set $r_1=0$ at which δ_c is near $2\zeta/\tau$, proceeding with a case analysis depending on which of the two terms of r^*_{crit} is largest. In the end this choice permits roughly the right form of bounds, but noticeable improvements in the constants arise with suitable non-zero r_1 in certain regimes.

Secondly, as determined in this section, we can find the r_1 or equivalently $\delta_c=\delta_{match}$ at which the two expressions in the definition of r^*_{crit} match. In some cases this provides the minimum value of r^*_{crit}. Thirdly, keep in mind that we want a small mistake rate δ^*_{mis} as well as a small drop from capacity of the inner code. The use of the overall rate of the composite code provides means to express a combination of δ^*_{mis}, r^*_{crit} and $D(\delta_c)/snr$ to optimize.

In this subsection we address the optimization of δ_c for each ζ and then in the next subsection the choice of nearly best values of ζ. In particular, this analysis provides means to determine regimes for which it is best overall to use δ_{match} or for which it is best to use instead $\delta_c=0$ or $\delta_c=snr$.

For $\zeta>-\tau$, define ζ' by

$$\zeta'=\zeta\,(1\!+\!\zeta/2\tau)$$

for which $(1\!+\!\zeta/\tau)^2=1+2\zeta'/\tau$ and define $\psi=\psi(\zeta)$ by

$$\psi=(2\tau+\zeta)\phi(\zeta)/\Phi(\zeta)$$

and $\gamma=\gamma(\zeta)$ by the small value

$$\gamma=2\zeta'/\tau+(\psi\!-\!1)/\tau^2.$$

***Lemma** 25: Match making.* Given ζ, the choice of $\delta_c=\delta_{match}$ that makes the two expressions in the definition of r^*_{crit} be equal is given by

$$1\!+\!\delta_c=e^{\gamma/(1+\zeta/\tau)^2},$$

at which

$$1\!+\!r_1/\tau^2=(1\!+\!\zeta/\tau)^2e^{-\gamma/(1+\zeta/\tau)^2}.$$

This δ_c is non-negative for ζ such that $\gamma\ge 0$. At this $\delta_c=\delta_{match}$ the value of $\tilde r^*_{crit}=r^*_{crit}-1-\epsilon$ is equal to

$$\tau^2(1\!+\!r_1/\tau^2)D(\delta_c)\;=\;r_1+\psi-1,$$

which yields $\tilde r^*_{crit}/\tau^2$ equal to

$$(1\!+\!\zeta/\tau)^2\big[e^{-\gamma/(1+\zeta/\tau)^2}-1\big]+\gamma,$$

which is less than $\gamma^2/[2(1\!+\!\zeta/\tau)^2]$ for $\gamma>0$. Moreover, the contribution δ^*_{mis} to the mistake rate as in Lemma 17, at this choice of δ_c and corresponding r_1, is equal to

$$\delta^*_{mis}\;=\;\frac{\psi}{2(\tau\!+\!\zeta)^2\,\mathcal{C}}.$$

Remark: Note from the definition of ψ and ζ' that

$$\gamma=\frac{(2+\zeta/\tau)\big(\zeta+\phi(\zeta)/\Phi(\zeta)\big)-1/\tau}{\tau}.$$

Thus γ is near $2\big(\zeta+\phi(\zeta)/\Phi(\zeta)\big)/\tau$.

Using the tail properties of the normal given in the appendix, the expression $\zeta+\phi(\zeta)/\Phi(\zeta)$ is seen to be non-negative and increasing in ζ for all ζ on the line, near to $1/|\zeta|$ for sufficiently negative ζ, and at least ζ for all positive ζ, In particular, γ is found to be non-negative for ζ at least slightly to the right of $-\tau$.

Meanwhile, by such tail properties, $\phi(\zeta)/\Phi(\zeta)$ is near $|\zeta|$ for sufficiently negative ζ, so to keep δ^*_{mis} small we will want to avoid such ζ unless the capacity $\mathcal{C}$ is very large. At $\zeta=0$ the ψ equals $4\tau/\sqrt{2\pi}$ so the δ^*_{mis} there is of order $1/\tau$. When $\mathcal{C}$ is not large we will prefer somewhat positive ζ to produce a small δ^*_{mis} in balance with the rate drop contribution r^*_{crit}/τ^2.

The $\zeta=0$ case is illustrative for the behavior of γ and related quantities. There γ is $(\psi-1)/\tau^2$ equal to $4/\tau\sqrt{2\pi}-1/\tau^2$, with which $\delta_c=e^{\gamma}-1$. Also $r_1/\tau^2=e^{-\gamma}-1$. The $\tilde{r}^*_{crit}/\tau^2=r_1/\tau^2+(\psi-1)/\tau^2$ is then equal to $e^{-\gamma}-1+\gamma$ near to and upper bounded by $\gamma^2/2=(1/2)(\psi-1)^2/\tau^4$ less than $4/\tau^2\pi$. In this $\zeta=0$ case, the slightly positive δ_c, with associated negative r_1, is sufficient to cancel the $(\psi-1)/\tau^2$ part of $\tilde{r}^*_{crit}/\tau^2$, leaving just the small amount bounded by $(1/2)(\psi-1)^2/\tau^4$. With $(\psi-1)/\tau^2$ less than 2, that is a strictly superior value for $\tilde{r}^*_{crit}/\tau^2$ than obtained with $\delta_c=0$ and $\zeta=0$ for which $\tilde{r}^*_{crit}/\tau^2$ is $(\psi-1)/\tau^2$.

Proof of Lemma 25: The $r^*_{crit}-\epsilon$ is the maximum of the two expressions

$$(\tau^2+r_1)D(\delta_c)+1$$

and

$$r_1\Phi(\zeta)+(2\tau+\zeta)\phi(z)+[(\tau^2+r_1)D(\delta_c)+1]\bar{\Phi}(\zeta).$$

Equating these, grouping like terms together using $1-\bar{\Phi}(\zeta)=\Phi(\zeta)$, and then dividing through by $\Phi(\zeta)$ yields

$$(\tau^2+r_1)D(\delta_c)-r_1\;=\;\psi-1.$$

Using τ^2+r_1 equal to $(\tau+\zeta)^2/(1+\delta_c)$ and $[D(\delta_c)-1]/(1+\delta_c)$ equal to $\log(1+\delta_c)-1$ the above equation may be written

$$(\tau+\zeta)^2[\log(1+\delta_c)-1]+\tau^2\;=\;\psi-1.$$

Rearranging, it is

$$\log(1+\delta_c)\;=\;1+\frac{\tau^2+(\psi-1)}{(\tau+\zeta)^2},$$

where the right side may also be written $\gamma/(1+\zeta/\tau)^2$. Exponentiating establishes the solution for $1+\delta_c$, with corresponding r_1 as indicated. Let's call the value that produces this equality $\delta_c=\delta_{match}$. At this solution the value of $\tilde{r}^*_{crit}=r^*_{crit}-1-\epsilon$ satisfies

$$(\tau^2+r_1)D(\delta_c)\;=\;r_1+\psi-1.$$

Likewise, from the identity $(\tau^2+r_1)D(\delta_c)-(r_1-1)=\psi$, multiplying by $\bar{\Phi}(\zeta)$ this establishes that the remainder used in Lemma 16 is in the present case equal to $rem=\psi\,\bar{\Phi}(\zeta)$, while the main part $(2\tau+\zeta)\phi(\zeta)$ is equal to $\psi\,\Phi(\zeta)$. Adding them using $\Phi(\zeta)+\bar{\Phi}(\zeta)=1$ shows that $(2\tau+\zeta)\phi(\zeta)+rem$ is equal to ψ. This is the numerator in the mistake rate expression δ^*_{mis}.

Using the form of r_1 the above expression for $\tilde{r}^*_{crit}/\tau^2$ may also be written

$$(1+\zeta/\tau)^2\big[e^{-\gamma/(1+\zeta/\tau)^2}-1\big]+\gamma.$$

With $y=\gamma/(1+\zeta/\tau)^2$ positive, the expression in the brackets is $e^{-y}-1$ which is less than $y+y^2/2$. Plugging that in, the part linear in y cancels, leaving the claimed bound $\gamma^2/[2(1+\zeta/\tau)^2]$. This completes the proof of Lemma 25.

Lemma 26: *The optimum δ_c quartet.* For each ζ, consider the following minimizations. First, consider the minimization of r^*_{crit} for δ_c in the interval $[0,snr]$. Its minimum occurs at the positive δ_c which is the minimum of the three values δ_{thresh_0}, δ_{match}, and snr, where $\delta_{thresh_0}=\Phi(\zeta)/\bar{\Phi}(\zeta)$.

Second, consider the minimization of

$$(1+D(\delta_c)/snr)(1+r_{crit}/\tau^2)$$

as arises in the denominator of our rate expression. Its minimum for δ_c in $[0,snr)$ occurs at the positive δ_c which is the minimum of the two values δ_{thresh_1} and δ_{match}, where δ_{thresh_1} is $\Phi(\zeta)/[\bar{\Phi}(\zeta)+1/snr]$.

Third, consider the minimization of the following combination of contributions to the inner code rate drop and the simplified mistake rate,

$$\delta^*_{mis,simp}\;+\;(1+D(\delta_c)/snr)(1+r_{crit}/\tau^2)-1,$$

for δ_c in $[0,snr)$. For $\Phi(\zeta)\le 1/(1+2\mathcal{C})$ its minimum occurs at $\delta_c=0$, otherwise it occurs at the positive δ_c which is the minimum of the two values δ_{thresh} and δ_{match} where

$$\delta_{thresh}=\frac{\Phi(\zeta)-\bar{\Phi}(\zeta)/2\mathcal{C}}{1/snr+\bar{\Phi}(\zeta)(1+1/2\mathcal{C})}.$$

The same conclusion holds using $\delta^*_{mis}=\delta^*_{mis,simp}/(1+\zeta/\tau)^2$, replacing the occurrences of $2\mathcal{C}$ in the previous sentence with $2\mathcal{C}(1+\zeta/\tau)^2$. Finally, set

$$\Delta_{\zeta,\delta_c}=\delta^*_{mis}+(1+D(\delta_c)/snr)(1+r_{crit}/\tau^2)-1$$

and extend the minimization to $[0,snr]$ using the previously given specialized values in the $\delta_c=snr$ case. Then for each ζ the minimum Δ_{ζ,δ_c} for δ_c in $[0,snr]$ is equal to the minimum over the four values 0, δ_{thresh}, δ_{match} and snr.

Remark: The Δ_{ζ,δ_c}, when optimized also over ζ, will provide the Δ_{shape} summarized in the introduction. As shown in the next section, motivation for it arises from the total drop rate from capacity of the composition of the sparse superposition code with the outer Reed-Solomon code. For now just think of it as desirable to choose parameters that achieve a good combination of low rate drop and low fraction of section mistakes. As the proof here shows, the proposed combination is convenient for the calculus of this optimization.

Recall for $0\le\delta_c<snr$ that $(1+D(\delta_c)/snr)(1+r_{crit}/\tau^2)$ equals $(1+r_1/\tau^2)D(\delta_c)/snr+1+r^*_{crit}/\tau^2$. In contrast, for $\delta_c=snr$, we set $\delta^*_{mis}=\bar{\Phi}(\zeta)$ and $r_{crit}=\max\{r_{up},0\}$, using the form of r_{up} previously given for this case. These different forms arise because the $g_{low}(x)$ bounds are used for $0\le\delta_c<snr$, whereas $g(x)$ is used directly for $\delta_c=snr$.

Proof of Lemma 26: To determine the δ_c minimizing r^*_{crit}, in the definition of $r^*_{crit} - 1 - \epsilon$ write the first expression $(\tau^2{+}r_1)D(\delta_c)$ in terms of δ_c as

$$(\tau{+}\zeta)^2 \frac{D(\delta_c)}{(1{+}\delta_c)}.$$

Take its derivative with respect to δ_c. The ratio $D(\delta_c)/(1{+}\delta_c)$ has derivative that is equal to $[D'(\delta_c)(1{+}\delta_c) - D(\delta_c)]$ divided by $(1{+}\delta_c)^2$. Now from the form of $D(\delta_c)$, its derivative $D'(\delta_c)$ is $\log(1{+}\delta_c)$, so the expression in brackets simplifies to δ_c, which is non-negative, and multiplying by the positive factor $(\tau{+}\zeta)^2/(1{+}\delta_c)^2$ provides the desired derivative. Thus this first expression is increasing in δ_c, strictly so for $\delta_c > 0$. As for the second expression in the maximum, it is equal to the first expression times $\bar{\Phi}(\zeta)$ plus $r_1 + \psi - 1$ times $\Phi(\zeta)$. So from the relationship of r_1 and δ_c, its derivative is equal to $[\delta_c \bar{\Phi}(\zeta) - \Phi(\zeta)]$ times the same $(\tau{+}\zeta)^2/(1{+}\delta_c)^2$. So the value of the derivative of the first expression is larger than that of the second expression, and accordingly the maximum of the two expressions equals the first expression for $\delta_c \geq \delta_{match}$ and equals the second expression for $\delta_c < \delta_{match}$. The derivative of the second expression, being the multiple of $[\delta_c \bar{\Phi}(\zeta) - \Phi(\zeta)]$ is initially negative so that the expression is initialling decreasing, up to the point $\delta_{thresh_0} = \Phi(\zeta)/\bar{\Phi}(\zeta)$ at which the derivative of this second expression is 0, so the optimizer of r^*_{crit} occurs at the smallest of the three values δ_{match}, δ_{thresh}, and the right end point snr of the interval of consideration.

To minimize $(1{+}D(\delta_c)/snr)(1{+}r_{crit}/\tau^2) - 1$, multiplying through by τ^2, recall that it equals $(\tau^2{+}r_1)D(\delta_c)/snr + r^*_{crit}$ for $0 \leq \delta_c < snr$. Add to the previous derivative values the amount δ_c/snr, which is again multiplied by the same factor $(\tau{+}\zeta)^2/(1{+}\delta_c)^2$. The first expression is still increasing. The second expression, after accounting for that factor, has derivative

$$\delta_c/snr + \delta_c \bar{\Phi}(\zeta) - \Phi(\zeta).$$

It is still initially negative and hits 0 at $\delta_{thresh_1} = \Phi(\zeta)/[\bar{\Phi}(\zeta) + 1/snr]$, which is again the minimizer if it occurs before δ_{match}. Otherwise, if δ_{match} is smaller than δ_{thresh_1} then, since to the right of δ_{match} the maximum equals the increasing first expression, it follows that δ_{match} is the minimizer.

Next determine the minimizer of the criterion that combines the rate drop contribution with the simplified section mistake contribution $\delta^*_{mis,simp}$. Multiplying through by τ^2, we have added the quantity $(\tau^2 + r_1)D(\delta_c) - r_1 + 1$ times $\bar{\Phi}(\zeta)/2\mathcal{C}$ plus the amount $(2\tau + \zeta)\phi(\zeta)/2\mathcal{C}$ not depending on δ_c. So its derivative adds the expression $(\delta_c{+}1)\bar{\Phi}(\zeta)/2\mathcal{C}$ times the same the factor $(\tau{+}\zeta)^2/(1{+}\delta_c)^2$. Thus, when the first part of the max is active, the derivative, after accounting for that factor, is

$$\delta_c + \delta_c/snr + (1{+}\delta_c)\bar{\Phi}(\zeta)/2\mathcal{C},$$

whereas, when the second part of the max is active it is

$$\delta_c\,\bar{\Phi}(\zeta) \,-\, \Phi(\zeta) \,+\, \delta_c\,snr \,+\, (1{+}\delta_c)\bar{\Phi}(\zeta)/2\mathcal{C}.$$

Again the first of these is positive and greater than the second. Where the value δ_c is relative to δ_{match} determines which part of the max is active. For $\delta_c < \delta_{match}$ it is the second. Initially, at $\delta_c = 0$, it is

$$-\Phi(\zeta) + \bar{\Phi}(\zeta)/2\mathcal{C},$$

which is $(1/2\mathcal{C})[1 - \Phi(\zeta)(1{+}2\mathcal{C})]$. If ζ is small enough that $\Phi(\zeta) \leq 1/(1+2\mathcal{C})$, this is at least 0. Then the criterion is increasing to the right of $\delta_c = 0$, whence $\delta_c = 0$ is the minimizer. Else if $\Phi(\zeta) > 1/(1{+}2\mathcal{C})$ then initially, the derivative is negative and the criterion is initially decreasing. Then as before the minimum value is either at δ_{thresh} or at δ_{match} whichever is smallest. Here δ_{thresh} is the point where the function based on the second expression in the maximum has 0 derivative. The same conclusions hold with $\delta^*_{mis} = \delta^*_{mis,simp}/(1{+}\zeta/\tau^2)$ in place of $\delta_{mis,simp}$ except that the denominator $2\mathcal{C}$ is replaced with $2\mathcal{C}(1{+}\zeta/\tau^2)$.

Examining δ_{thresh_1} and δ_{thresh}, it is seen that these are less than snr. Nevertheless, when minimizing over $[0, snr]$, the minimum can arise at snr because of the different form assigned to the expressions in that case. Accordingly the minimum of Δ_{ζ,δ_c} for δ_c in $[0, snr]$ is equal to the minimum over the four values 0, δ_{thresh}, δ_{match} and snr, referred to as the optimum δ_c quartet. This completes the proof of Lemma 26.

Remark: To be explicit as to the form of Δ_{ζ,δ_c} with $\delta_c = snr$, recall that in this case $1{+}r_{up}/\tau^2$ is

$$(1{-}snr\,\bar{\Phi}(\zeta))(1{+}\zeta/\tau)^2/(1{+}snr).$$

Consequently $\Delta_{\zeta,\delta_c} = \delta^*_{mis}{+}(1{+}D(\delta_c)/snr)(1{+}r_{crit}/\tau^2){-}1$, in this $\delta_c = snr$ case, becomes

$$\bar{\Phi}(\zeta) \,+\, \frac{(1{+}D(snr)/snr)}{(1+snr)}(1 + snr\,\bar{\Phi}(\zeta))(1{+}\zeta/\tau)^2 \,-\, 1,$$

when $r_{up} \geq 0$. For $r_{up} < 0$ as is true for sufficiently small contributions from $snr\bar{\Phi}(\zeta)$ and ζ/τ, we simply set $r_{crit} = 0$ to avoid complications from the conditions of Corollary 19. Then Δ_{ζ,δ_c} becomes

$$\bar{\Phi}(\zeta) + D(snr)/snr.$$

F. Inequalities for ψ, γ, and $\tilde{r}^*_{crit}$:

At $\delta_c = \delta_{match}$ we examine $\tilde{r}^*_{crit}$ further. Previously, in Lemma 25 the expression $\tilde{r}^*_{crit}/\tau^2$ is shown to be less than $\gamma^2/[2(1{+}\zeta/\tau)^2]$. Now this bound is refined in the cases of negative and positive ζ. For negative ζ it is shown that $\gamma \leq 2/\tau|\zeta|$ and for positive $|\zeta|$ it is shown that $\tilde{r}^*_{crit}$ is not more than $\max\{2(\zeta')^2, \psi - 1\}$. For sufficiently positive ζ it is not more than $2\zeta^2$.

Recall that γ is less than

$$\frac{(2{+}\zeta/\tau)\big(\zeta + \phi(\zeta)/\Phi(\zeta)\big)}{\tau}$$

and that $\psi = (2\tau + \zeta)\phi(\zeta)/\Phi(\zeta)$.

***Lemma** 27: Inequalities for negative* ζ. For $-\tau < \zeta \leq 0$, the γ is an increasing function less than $\min\{2/|\zeta|, 4/\sqrt{2\pi}\}/\tau$. Likewise the function ψ is less than $2(|\zeta| + 1/|\zeta|)\tau$.

Proof of Lemma 27: For $\zeta \leq 0$, the increasing factor $2{+}\zeta/\tau$ is less than 2 and the factor $\zeta + \phi(\zeta)/\Phi(\zeta)$ is non-negative, increasing, and less than $1/|\zeta|$ by the normal tail inequalities in the appendix. At $\zeta = 0$ this factor is $2/\sqrt{2\pi}$. As for ψ the factor $\phi(\zeta)/\Phi(\zeta)$ is at least $|\zeta|$ and not more than $|\zeta|{+}1/|\zeta|$ for negative ζ again by the normal tail inequalities in the appendix (where improvements are given, especially for $0 \leq |\zeta| \leq 1$). This completes the proof of Lemma 27.

Now we turn attention to non-negative ζ. Three bounds on $\tilde{r}^*_{crit}$ are given. The first based on $\gamma^2/2$ and the other two more exacting to determine the relative effects of $2(\zeta')^2$ and $\psi - 1$.

***Corollary** 28:* For $\zeta \geq 0$ we have $\tilde{r}^*_{crit}/\tau^2 \leq \gamma^2/2$ and

$$\tilde{r}^*_{crit} \;\leq\; 2\big(\zeta + \phi(\zeta)/\Phi(\zeta)\big)^2.$$

Proof of Corollary 28: By Lemma 25, the $\tilde{r}^*_{crit}/\tau^2$ is not more than $\gamma^2/[2(1{+}\zeta/\tau)^2]$. Now γ is not more than

$$\frac{2(1{+}\zeta/2\tau)\big(\zeta + \phi(\zeta)/\Phi(\zeta)\big)}{\tau}$$

Consequently, $\tilde{r}^*_{crit}$ is not more than

$$\frac{2(1{+}\zeta/2\tau)^2\big(\zeta + \phi(\zeta)/\Phi(\zeta)\big)^2}{(1{+}\zeta/\tau)^2}.$$

Using $1{+}\zeta/2\tau$ not more than $1{+}\zeta/\tau$, completes the proof of Lemma 28.

***Lemma** 29: Direct r^*_{crit} bounds.* Let $\tilde{r}^*_{crit} = r^*_{crit} - 1 - \epsilon$ evaluated at δ_{match}. Bounds are provided depending on whether $D(2\zeta'/\tau)$ or $(\psi - 1)/\tau^2$ is larger. In the case $D(2\zeta'/\tau) \geq (\psi - 1)/\tau^2$ the $\tilde{r}^*_{crit}$ satisfies

$$\tilde{r}^*_{crit}/\tau^2 \;\leq\; D(2\zeta'/\tau).$$

In any case, the value of $\tilde{r}^*_{crit}/\tau^2$ may be represented as an average of $D(2\zeta'/\tau)$ and $(\psi{-}1)/\tau^2$, plus a small excess, where the weight assigned to $(\psi{-}1)/\tau^2$ is proportional to the small $2\zeta'/\tau$. Indeed $\tilde{r}^*_{crit}/\tau^2$ equals

$$\frac{D(2\zeta'/\tau) + \frac{(\psi-1)}{\tau^2}2\zeta'/\tau}{1 + 2\zeta'/\tau} + excess$$

where $excess$ is $e^{-v} - (1 - v)$ evaluated at

$$v = \frac{\frac{\psi-1}{\tau^2} - D(2\zeta'/\tau)}{1 + 2\zeta'/\tau}.$$

In the case $(\psi{-}1)/\tau^2 > D(2\zeta'/\tau)$ it satisfies

$$excess \leq \frac{\left[\frac{\psi-1}{\tau^2} - D(2\zeta'/\tau)\right]^2}{2(1 + \zeta/\tau)^4}.$$

Proof of Lemma 29: With the relationship between r_1 and δ_c, recall $(\tau^2{+}r_1)D(\delta_c)$ is increasing in δ_c and hence decreasing in r_1. The r_1 that provides the match makes $(\tau^2{+}r_1)D(\delta_c)$ equal $r_1{+}\psi{-}1$. At $r_1 = 0$, the first is $\tau^2 D(2\zeta'/\tau)$, so if that be larger than $\psi{-}1$ then a positive r_1 is needed to bring it down to the matching value. Then $\tilde{r}^*_{crit}$ is less than $\tau^2 D(2\zeta'/\tau)$. Whereas if $\tau^2 D(2\zeta'/\tau)$ is less than $\psi{-}1$ then $\tilde{r}^*_{crit}$ is greater than $D(2\zeta'/\tau)$, but not by much as we shall see. In any case, write $\tilde{r}^*_{crit}/\tau^2$ as

$$\frac{r_1}{\tau^2} + \frac{\psi - 1}{\tau^2}$$

which by Lemma 25 is

$$\frac{\psi - 1}{\tau^2} \;+\; (1{+}\zeta/\tau)^2 e^{-\gamma/(1+\zeta/\tau)^2} \;-\; 1.$$

Use $\gamma = 2\zeta'/\tau + (\psi - 1)/\tau^2$ and for this proof abbreviate $a = (\psi{-}1)/\tau^2$ and $b = 2\zeta'/\tau$. The exponent $\gamma/(1{+}\zeta/\tau)^2$ is then $(a{+}b)/(1{+}b)$ and the expression for $\tilde{r}^*_{crit}/\tau^2$ becomes

$$a + (1{+}b)e^{-(a+b)/(1+b)} - 1.$$

Add and subtract $D(b)$ in the numerator to write $(a{+}b)/(1{+}b)$ as $(a{-}D(b))/(1{+}b)$ plus $(b{+}D(b))(1{+}b)$, where by the definition of $D(b)$ the latter term is simply $\log(1+b)$ which leads to a cancelation of the $1+b$ outside the exponent. So the above expression becomes

$$a + e^{-(a-D(b))/(1+b)} - 1,$$

which is $a + e^{-v} - 1 = a - v + excess$, where $excess = e^{-v} - (1{-}v)$ and $v = (a{-}D(b))/(1{+}b)$. For $a \geq D(b)$, that is, $v \geq 0$, the excess is less than $v^2/2$, by the second order expansion of e^{-v}, since the second derivative is bounded by 1, which provides the claimed control of the remainder. The $a{-}v$ may be written as $[D(b){+}ba]/(1{+}b)$ the average of $D(b)$ and a with weights $1/(1{+}b)$ and $b/(1{+}b)$, or equivalently as $D(b) + b(a - D(b))/(1{+}b)$. Plugging in the choices of a and b completes the proof of Lemma 29.

An implication when ζ and $\psi - 1$ are positive, is that $\tilde{r}^*_{crit}/\tau^2$ is not more than

$$D(2\zeta'/\tau) + \frac{2\zeta'(\psi - 1)}{\tau^3} + \frac{(\psi - 1)^2}{\tau^4}.$$

This bound, and its sharper form in the above lemma, shows that $\tilde{r}^*_{crit}/\tau^2$ is not much more than $D(2\zeta'/\tau)$, which in turn is less than $2(\zeta')^2/\tau^2$, near $2\zeta^2/\tau^2$.

Another means by which to show that $\tilde{r}^*_{crit}$ is not much more than $2\zeta^2$ uses the upper bound formed by equating the two expressions in r^*_{crit}, with $D(\delta_c)$ replaced by its upper bound $\delta_c^2/2$, where with δ_c expressed as a function of r_1, it leads to a quadratic equation for the solution r_1. It is useful in refining the above, especially in demonstrating that $\tilde{r}^*_{crit}$ is less than $2\zeta^2$ for a range of values of sufficiently positive ζ.

***Corollary** 30: r^*_{crit} bounds based on a quadratic equation.* In the definition of r^*_{crit} consider the value of r_1 that matches upper bounds on the two expressions that arise with $D(\delta_c)$ replaced by $\delta_c^2/2$. Then for every $\zeta \geq 0$, the $\tilde{r}^*_{crit}$ is not more than

$$\frac{1}{1{+}\gamma}\,2(\zeta')^2 \;+\; \frac{\gamma}{1{+}\gamma}\,(\psi - 1).$$

It is less than the value $2(\zeta')^2$ when $\psi{-}1$ is less than that value, and, otherwise, the above expression controls the small amount by which it exceeds $2(\zeta')^2$. The r_1 is given as a solution to a quadratic equation. The sign of r_1 equals the sign of $2(\zeta')^2{+}1{-}\psi$. This r_1 satisfies $r_1 \leq [2(\zeta')^2{+}1{-}\psi]/[1{+}\gamma]$ and when r_1 is negative it satisfies $r_1 \geq [2(\zeta')^2{+}1{-}\psi]/[1{+}2\zeta'/\tau]$. If ζ is such that $\psi \geq 1$, the δ_c is positive. Then $r_1 \leq 2\tau\zeta'$ and $1{+}r_1/\tau^2 \leq 1{+}2\zeta'/\tau$. For every $\zeta \geq 0$, the above bound on $r^*_{crit} - \epsilon$ may be written

$$\frac{2(\zeta')^2 + 1 \,+\, (2\zeta'/\tau)\,\psi \,+\, \psi(\psi{-}1)/\tau^2}{1 + \gamma}.$$

With $\psi \geq 1$ it is further tightly upper bounded by replacing the γ in the denominator with $2\zeta'/\tau$, such that $1{+}\gamma$ is at least $(1{+}\zeta/\tau)^2$. Furthermore, if ζ^2 is at least $\psi{-}1$ and $\zeta \geq 1/(4\tau)$, then

$$\tilde{r}^*_{crit} < 2\zeta^2.$$

The proof of this Corollary 30 is in the appendix.

We remark that later we will choose $\zeta^2 = 2\log\tau/(d\sqrt{2\pi})$ with a value for $d \geq 1/2$, depending on the snr. This arranges $\tau\phi(\zeta) = d$ with which ψ is at least $2d$, so it will be at least 1. Also with d constant, the case that ζ^2 exceeds $\psi{-}1$ is regarded as somewhat typical. Before proceeding further, let's compare the $2\zeta^2$ bound achieved when r_1 is sufficiently positive with the form $2(\zeta')^2$ that holds when $r_1 \geq 0$. The expression $2(\zeta')^2$ equal to $2\zeta^2\big(1+\zeta/(2\tau)\big)^2$, exceeds $2\zeta^2$ by an amount near $2\zeta^3/\tau$, which is of order $(\log\tau)^{1.5}/\tau$. The point of the last claim of the corollary is that this excess amount from ζ' in place of ζ is avoided.

We also take note of the following monotonicity property of the function $\psi(\zeta)$ for $\zeta \geq 0$. It uses the fact that $\tau \geq 1$. Indeed, $\tau \geq \sqrt{2\log B}$ is at least $\sqrt{2\log 2} = 1.18$.

***Lemma** 31: Monotonicity of ψ:* With $\tau \geq 1.0$, the positive function $\psi(z) = (2\tau{+}z)\phi(z)/\Phi(z)$ is strictly decreasing for $z \geq 0$. Its maximum value is $\psi(0) = 4\tau/\sqrt{2\pi} \leq 1.6\,\tau$. Moreover $\gamma = 2\zeta'/\tau + (\psi-1)/\tau^2$ is positive.

Proof of Lemma 31: The function $\psi(z)$ is clearly strictly positive for $z \geq 0$. Its derivative is seen to be

$$\psi'(z) = -[\psi(z) + z(2\tau+z) - 1]\phi(z)/\Phi(z).$$

Note that the function $\tau^2\gamma(z)$ matches the expression in brackets, this derivative equals

$$-\tau^2\gamma(z)\phi(z)/\Phi(z).$$

The $\tau^2\gamma(z)$ is at least $\psi(z) + 2\tau z - 1$, and it remains to show that it is positive for all $z \geq 0$. It is clearly positive for $z \geq 1/2\tau$. For $0 \leq z \leq 1/2\tau$, lower bound it by lower bounding $\psi(z) - 1$ by $2\tau\phi(1/2\tau)/\Phi(1/2\tau) - 1$, which is positive provided $1/2\tau$ is less that the unique point $z = z_{root} > 0$ where $\phi(z) = z\Phi(z)$. Direct evaluation shows that this z is between 0.5 and 0.6. So $\tau \geq 1.0$ suffices for the positivity of $\gamma(z)$ and equivalently the negativity of $\psi'(z)$ for all $z \geq 0$. This completes the proof of Lemma 31.

The monotonicity of $\psi(\zeta)$ is associated with decreasing shortfall δ^* as we increase ζ, though with the cost of increasing r^*_{crit}. Evaluating r^*_{crit} as a function of ζ enables us to control the tradeoff.

Remark 1: For most ζ we use $\Phi(\zeta)$ is not far from 1. Using $\bar\Phi(\zeta) \leq \phi(\zeta)/\zeta$, the expression for $\psi = (2\tau{+}\zeta)\phi(\zeta)/\Phi(\zeta)$ is not more than what one would have with $1-\phi(\zeta)/\zeta$ replacing the $\Phi(\zeta)$ in the denominator. Accordingly for the condition $\psi \leq \zeta^2+1$, rearranging that bound, it is seen to be sufficient that

$$(2\tau + 2\zeta + 1/\zeta)\phi(\zeta) \;\leq\; \zeta^2 + 1.$$

Now with $\zeta = \sqrt{2\log\tau/(d\sqrt{2\pi})}$, the right side is of order $\log\tau$, while the left side is near $2d$. So the condition can be comfortably satisfied.

Remark 2: With $(\psi-1)/\tau$ small compared to ζ, the above proof shows that $r^*_{crit} - 2\zeta^2 - 1$ is bounded by an expression near

$$-[2\zeta/\tau]\big[\zeta^2 + 1 - \psi\big] + \epsilon.$$

So with $\zeta^2 + 1 - \psi$ positive and of the order $\log\tau$, the first part is sufficiently negative that it remains so when we add the polynomially small ϵ. Then we have the even simpler bound, without need for the ϵ,

$$r^*_{crit} \;\leq\; 2\zeta^2 - 1.$$

Remark 3: The rate $R = \mathcal{C}'/(1{+}r/\tau^2)$ has been parameterized by r. As stated in Lemma 24, the relationship between the gap and r, expressed as $gap = (r{-}r_{crit})/[snr(\tau^2{+}r_1)]$, may also be written $r = r_{crit} + snr(\tau^2{+}r_1)gap$. Recall also that we may set $gap = \eta + \bar f + 1/(m-1)$, with $\bar f = mf^*\rho$. In this way, the rate parameter r is determined from the choices of ζ that appear in r_{crit} as well as from the parameters m, $\bar f$ and η that control, respectively, the number of steps, the fractions of false alarms, and the exponent of the error probability.

The importance of ζ in this section is that provides for the evaluation of r_{crit} and through r_1 it controls the location of the upper end of the region in which $g(x) - x$ is shown to exceed a target gap. For any ζ, the above remark conveys the smallest size rate drop parameter r for which that gap is shown to be achieved.

In the rate representation R, draw attention to the product of two of the denominator factors $(1{+}D(\delta_c)/snr)(1{+}r/\tau^2)$. We represent these factors in a way that exhibits the dependence on r^*_{crit} and the gap.

Using r equal to $r_{crit} + snr\,gap\,\tau^2\,(1{+}r_1/\tau^2)$ write the factor $1{+}r/\tau^2$ as $(1{+}r_{crit}/\tau^2)(1 + \xi\,snr\,gap)$ where ξ is the ratio $(1{+}r_1/\tau^2)/(1{+}r_{crit}/\tau^2)$, a value between 0 and 1, typically near 1. Thus the product $(1{+}D(\delta_c)/snr)(1{+}r/\tau^2)$ takes the form

$$\big(1{+}D(\delta_c)/snr\big)\big(1{+}r_{crit}/\tau^2\big)\big(1 + \xi\,snr\,gap\big).$$

Recall that $(1{+}D(\delta_c)/snr)(1{+}r_{crit}/\tau^2)$ is equal to

$$1 + \frac{(1{+}r_1/\tau^2)D(\delta_c)}{snr} + \frac{r^*_{crit}}{\tau^2}$$

which, at $r_1 = r_{1,match}$, is equal to

$$1 + \frac{\tilde{r}^*_{crit}}{snr\,\tau^2} + \frac{\tilde{r}^*_{crit} + 1 + \epsilon}{\tau^2}.$$

So in this way these denominator factors are expressed in terms of the gap and $\tilde{r}^*_{crit}$, where $\tilde{r}^*_{crit}$ is near $2\zeta^2$ by the previous corollary.

We complete this subsection by inquiring whether r_{crit} is positive for relevant ζ. By the definition of r_{crit}, its positivity is equivalent to the positivity of $r^*_{crit} + r_1 D(\delta_c)/snr$ which is not less than $1 + \epsilon + (\tau^2 + r_1)D(\delta_c) + r_1 D(\delta_c)/snr$. The multiplier of $D(\delta_c)$ is $(\tau^2 + r_1)(1 + 1/snr)$ which is positive for $r_1 \geq -\tau^2\,snr/(1{+}snr)$. So we ask whether that be a suitable lower bound on r_1. Recall the relationship between x^* and r_1,

$$1 - x^* = \frac{r - r_1}{snr(\tau^2{+}r_1)} = gap + \frac{r_{crit} - r_1}{snr(\tau^2{+}r_1)}.$$

Recognizing that $r_{crit} - r_1$ equals $r^*_{crit} - r_1$ divided by $1 + D(\delta_c)/snr$, expressing $D(\delta_c)$ in terms of r^*_{crit} and r_1 as above, one can rearrange this relationship to reveal the value of r_1 as a function of $x^* + gap$ and r^*_{crit}. Using $r^*_{crit} > 0$ one finds that the minimal r_1 to achieve positive $x^* + gap$ is indeed greater than $-\tau^2\, snr/(1{+}snr)$.

G. Determination of ζ:

In this subsection we solve, where possible, for the optimal choice of ζ in the expression Δ_{ζ,δ_c} which balances contributions to the rate drop with the quantity δ^*_{mis} related to the mistake rate. As above it is

$$\Delta_{\zeta,\delta_c} = \delta^*_{mis} \; + \; \big(1{+}D(\delta_c)/snr\big)\big(1{+}r_{crit}/\tau^2\big) - 1.$$

We also work with the simplified form $\Delta_{\zeta,\delta_c,simp}$ in which $\delta^*_{mis,simp}$ is used in place of δ^*_{mis}. For $0 \le \delta_c < snr$, this Δ_{ζ,δ_c} coincides, as we saw, with

$$\frac{(2\tau{+}\zeta)\phi(\zeta) + rem}{2\mathcal{C}\,\tau^2(1{+}\zeta/\tau)^2} \; + \; \frac{(1{+}r_1/\tau^2)D(\delta_c)}{snr} \; + \; \frac{r^*_{crit}}{\tau^2},$$

where $rem = \big[(\tau^2{+}r_1)D(\delta_c){-}(r_1{-}1)\big]\bar{\Phi}(\zeta)$. Define $\Delta_{\zeta,\delta_c,simp}$ to be the same but without the $(1{+}\zeta/\tau)^2$ in the denominator of the first part.

Seek to optimize Δ_{ζ,δ_c} or $\Delta_{\zeta,\delta_c,simp}$ over choices of ζ for each of the quartet of choices of δ_c given by 0, δ_{thresh}, δ_{match} and snr. The minimum of Δ_{ζ,δ_c} provides what we denote as Δ_{shape} as summarized in the introduction.

Optimum or near optimum choices for ζ are provided for the cases of δ_c equal to 0, δ_{match}, and snr, respectively. These provide distinct ranges of the signal to noise ratio for which these cases provide the smallest Δ_{ζ,δ_c}. At present we have not been able to determine whether the minimum $\Delta_{\zeta,\delta_{thresh}}$ has a range of signal to noise ratios at which its minimum is superior to what is obtained with the best of the other cases. What we can confirm regarding δ_{thresh} is that for small snr the $\min_\zeta \Delta_{\zeta,\delta_{thresh}}$ requires δ_{thresh} near snr, and that for snr above a particular constant, the minimum $\Delta_{\zeta,\delta_{thresh}}$ matches $\min_\zeta \Delta_{\zeta,0}$ with $\delta_{thresh} = 0$ at the minimizing ζ.

Optimal choices for ζ for the cases of δ_c equal to 0, δ_{match}, and snr, respectively, provide three disjoint intervals R_1, R_2, R_3 of signal to noise ratio. The case of $\delta_c = 0$ provides the optimum for the high end of snr in R_3; the case of $\delta_c = \delta_{match}$ provides our best bounds for the intermediate range R_2; and the case of $\delta_c = snr$ provides the optimum for the low snr range R_1.

Our tactic is consider these choices of δ_c separately, either optimizing over ζ to the extent possible or providing reasonably tight upper bounds on $\min_\zeta \Delta_{\zeta,\delta_c}$, and then inspect the results to see the ranges of snr for which each is best.

Note directly that Δ_{ζ,δ_c} is a decreasing function of snr for the $\delta_c = 0$ and $\delta_c = \delta_{match}$ cases, so $\min_\zeta \Delta_{\zeta,\delta_c}$ will also be decreasing in snr. Likewise for $\Delta_{\zeta,\delta_c,simp}$.

Remember that we are using log base e, so the capacity is measured in *nats*.

***Lemma** 32: Optimization of $\Delta_{\zeta,\delta_c,simp}$ with $\delta_c = 0$.* At $\delta_c = 0$, the $\Delta_{\zeta,0,simp}$ is optimized at the $1/(2\mathcal{C}{+}1)$ quantile of the standard normal distribution

$$\zeta = \zeta_{\mathcal{C}} = \Phi^{-1}(1/(2\mathcal{C}{+}1)).$$

If $\mathcal{C} \ge 1/2$ this $\zeta_{\mathcal{C}}$ is less than or equal to 0 and $\min_\zeta \Delta_{\zeta,0,simp}$ is not more than

$$\frac{(2\mathcal{C}{+}1)\phi(\zeta_{\mathcal{C}})}{\mathcal{C}\,\tau} \; + \; \frac{1}{\tau^2}.$$

Dividing the first term by $(1{+}\zeta_{\mathcal{C}}/\tau)^2$ gives an upper bound on $\min_\zeta \Delta_{\zeta,0}$ valid for $\zeta_{\mathcal{C}} > -\tau$. The bound is decreasing in $\mathcal{C}$ when $\zeta_{\mathcal{C}} > -\tau + 1$. Let $\mathcal{C}$, exponentially large in $\tau^2/2$, be such that $\zeta_{\mathcal{C}_{large}} = -\tau{+}1$. For $\mathcal{C} \ge \mathcal{C}_{large}$, use $\zeta = -\tau{+}1$ in place of $\zeta_{\mathcal{C}}$, then the first term of this bound is exponentially small in $\tau^2/2$ and hence polynomially small in $1/B$.

Thus the $\zeta = \zeta^*_{\mathcal{C}}$ advocated for $\delta_c = 0$ is

$$\zeta^*_{\mathcal{C}} = \max\{\zeta_{\mathcal{C}}, -\tau + 1\}.$$

Examination of the bound shows an implication of this Lemma. When $\mathcal{C}/\sqrt{\log \mathcal{C}}$ is large compared to τ the Δ_{shape} is near $1/\tau^2$. This is clarified in the following corollary which provides slightly more explicit bounds.

***Corollary** 33: Bounding $\min \Delta_{\zeta,0}$ with $\delta_c = 0$.* To present upper bounds on $\min_\zeta \Delta_{\zeta,0,simp}$, the choice $\zeta = 0$ provides

$$\frac{(2\mathcal{C}{+}1)}{\mathcal{C}}\left(\frac{1}{\tau\sqrt{2\pi}} \; + \; \frac{1}{4\tau^2}\right),$$

which also bounds $\min_\zeta \Delta_{\zeta,0}$. Moreover, when $\mathcal{C} \ge 1/2$, the optimum $\zeta_{\mathcal{C}}$ satisfies $|\zeta_{\mathcal{C}}| \le \sqrt{2\log(\mathcal{C}{+}1/2)}$ and provides the following bound, which improves on the $\zeta = 0$ choice when $\mathcal{C} \ge 2.2$,

$$\frac{\xi(|\zeta_{\mathcal{C}}|)}{\mathcal{C}\,\tau} \; + \; \frac{1}{\tau^2},$$

not more than

$$\frac{\xi\big(\sqrt{2\log(\mathcal{C}{+}1/2)}\big)}{\mathcal{C}\,\tau} \; + \; \frac{1}{\tau^2},$$

where $\xi(z)$ equals $z{+}1/z$ for $z \ge 1$ and equals 2 for $0{<}\zeta{<}1$. Dividing the first term by $(1{+}\zeta_{\mathcal{C}}/\tau)^2$, gives an upper bound on $\min_\zeta \Delta_{\zeta,0}$ of

$$\frac{\xi(|\zeta_{\mathcal{C}}|)}{\mathcal{C}\,\tau(1{+}\zeta_{\mathcal{C}}/\tau)^2} \; + \; \frac{1}{\tau^2}.$$

When $B \ge 1{+}snr$, this bound on $\min_\zeta \Delta_{\zeta,0}$ improves on the bound with $\zeta = 0$, for $\mathcal{C} \ge 5.5$. As before, when $\mathcal{C} \ge \mathcal{C}_{large}$ for which $\zeta_{\mathcal{C}_{large}} = -\tau + 1$, we use the bound with $\mathcal{C}_{large}$ in place of $\mathcal{C}$.

The $\min_\zeta \Delta_{\zeta,0,simp}$ bound above is smaller than given below for $\min_\zeta \Delta_{\zeta,\delta_{match},simp}$, when the snr is large enough that an expression of order $\mathcal{C}/(\log \mathcal{C})^{3/2}$ exceeds τ.

There is a role in what follows for the quantity $d = d_{snr} = 2\mathcal{C}/\nu = (1{+}1/snr)\log(1{+}snr)$. It is an increasing function of snr, with value always at least 1.

For $2\mathcal{C}/\nu \ge \tau/\sqrt{2\pi}$ we use non-positive ζ, whereas for $2\mathcal{C}/\nu < \tau/\sqrt{2\pi}$ we use positive ζ. Thus the discriminant of whether we use positive ζ is the ratio $\omega = d/\tau$ and whether it is smaller than $1/\sqrt{2\pi}$. This ratio ω is

$$\omega = \frac{d}{\tau} = \frac{2\mathcal{C}}{\nu\,\tau}.$$

In the next two lemma we use $\delta_c = \delta_{match}$. Using the results of Lemma 25 and $1 + 1/snr = 1/\nu$, the form of $\Delta_{\zeta,\delta_{match}}$ simplifies to

$$\frac{\psi}{2\mathcal{C}\,\tau^2(1+\zeta/\tau)^2} + \frac{1}{\nu}\,\frac{\tilde{r}^*_{crit}}{\tau^2} + \frac{1+\epsilon}{\tau^2}.$$

Recall for negative ζ we have that ψ is near $2\tau|\zeta|$ and $\tilde{r}^*_{crit}$ through $\gamma^2/2$ is near $2/|\zeta|^2$, with associated bounds given in Lemma 27. So it is natural to set a negative ζ that minimizes $-\tau\zeta/\mathcal{C} + 2/(\nu\,\zeta^2\,\tau^2)$ for which the solution is

$$\zeta = -(4\mathcal{C}/\nu\tau)^{1/3} = -(2\omega)^{1/3},$$

which we denote as $\zeta_{1/3}$.

***Lemma** 34: Optimization of Δ_{ζ,δ_c} at $\delta_c = \delta_{match}$: Bounds from non-positive ζ.* The choice of $\zeta = 0$ yields the upper bound on $\min_\zeta \Delta_{\zeta,\delta_{match}}$ of

$$\frac{2}{\mathcal{C}\,\tau\sqrt{2\pi}} + \frac{4}{\nu\,\tau^2\pi} + \frac{1+\epsilon}{\tau^2}.$$

As for negative ζ, the choice $\zeta = \zeta_{1/3} = -(4\mathcal{C}/\nu\tau)^{1/3}$ yields the upper bound on $\min_\zeta \Delta_{\zeta,\delta_{match}}$ of

$$\frac{1}{(1+\zeta_{1/3}/\tau)^2}\left(\frac{2.4}{\nu^{1/3}\mathcal{C}^{2/3}\tau^{4/3}} + \frac{2}{|\zeta_{1/3}|\mathcal{C}\,\tau}\right) + \frac{1{+}\epsilon}{\tau^2}.$$

Amongst the bounds so far with $\zeta \le 0$, the first term controlling δ^*_{mis} is smallest at $\zeta = 0$ where it is $2/[\mathcal{C}\,\tau\sqrt{2\pi}]$. The advantage of going negative is that then the $4/[\nu\,\tau^2\pi]$ term is replaced by terms that are smaller for large $\mathcal{C}$.

Comparison: The two bounds in Lemma 34 may be written as

$$\frac{4}{\nu\,\tau^2}\left[\frac{1}{\sqrt{2\pi}\omega} + \frac{1}{\pi}\right] + \frac{1+\epsilon}{\tau^2}$$

and

$$\frac{4}{\nu\,\tau^2}\left[\frac{1.5}{(2\omega)^{2/3}} + \frac{2}{(2\omega)^{4/3}}\right] + \frac{1+\epsilon}{\tau^2},$$

respectively, neglecting the $(1+\zeta_{1/3}/\tau)$ factor. Numerical comparison of the expressions in the brackets reveals that the former, from $\zeta = 0$, is better for $\omega < 5.37$, while the later from $\zeta = \zeta_{1/3}$ is better for $\omega \ge 5.37$, which is for $|\zeta_{1/3}| \ge 2.2$.

Next compare the leading term of the bound $\zeta_{1/3}$ and $\delta_c = \delta_{match}$ to the corresponding part of the bound using $\zeta_{\mathcal{C}}$ and $\delta_c = 0$. These are, respectively,

$$\frac{2.4}{\nu^{1/3}\mathcal{C}^{2/3}\tau^{4/3}}$$

and

$$\frac{\xi(|\zeta_{\mathcal{C}}|)}{\mathcal{C}\,\tau}.$$

From this comparison the $\delta_c = 0$ solution is seen to be better when

$$\frac{4.5\mathcal{C}}{\nu\,(\xi(|\zeta_{\mathcal{C}}|))^3} > \tau.$$

Modified to take into account the factors $1{+}\zeta_{1/3}/\tau$ and $1{+}\zeta^*_{\mathcal{C}}/\tau$, this condition defines the region R_3 of very large snr for which $\delta_c = 0$ is best. To summarize it corresponds to snr large enough that an expression near $4.5\mathcal{C}/(\log\mathcal{C})^{3/2}$ exceeds τ, or, equivalently, that $\mathcal{C}$ is at least a value of order $\tau(\log\tau)^{3/2}$, near to $(\tau/4.5)(\log(\tau/4.5))^{1.5}$, for sufficient size τ.

Next consider the case of $\omega = d/\tau$ less than $1/\sqrt{2\pi}$ for which we use positive ζ. The function $\phi(\zeta)/\Phi(\zeta)$ is strictly decreasing. From its inverse, let ζ_ω be the unique value at which $\phi(\zeta)/\Phi(\zeta) = 2\omega$. We use it to provide a tight bound on the optimal $\Delta_{\zeta,\delta_{match}}$.

***Lemma** 35: Optimization of Δ_{ζ,δ_c} at $\delta_c = \delta_{match}$: Bounds from positive ζ.* Consider the case that $\tau/\sqrt{2\pi} \ge 2\mathcal{C}/\nu$. Let $\omega = 2\mathcal{C}/\nu\,\tau$. The choice of $\zeta = \zeta_\omega$ yields $\Delta_{\zeta,\delta_{match}}$ not more than

$$\frac{2}{\nu\,\tau^2}\left[2 + \left(\zeta_\omega + 2\omega\right)^2\right] + \frac{1+\epsilon}{\tau^2}.$$

This ζ_ω is not more than $\zeta^*_\omega = \sqrt{2\log\left(1/2 + 1/2\omega\sqrt{2\pi}\right)}$ which is

$$\sqrt{2\log\left(\frac{1}{2} + \frac{\tau\,\nu}{4\mathcal{C}\sqrt{2\pi}}\right)},$$

at which $\Delta_{\zeta,\delta_{match}}$ is not more than

$$\frac{2}{\nu\,\tau^2}\left[2 + \left(\sqrt{2\log\left(\frac{1}{2} + \frac{\tau\,\nu}{4\mathcal{C}\sqrt{2\pi}}\right)} + \frac{4\mathcal{C}}{\tau\,\nu}\right)^2\right] + \frac{1{+}\epsilon}{\tau^2}.$$

For small d/τ the $2\omega = 4\mathcal{C}/\nu\,\tau = 2d/\tau$ term inside the square is negligible compared to the log term. Then the bound is near

$$\frac{4}{\nu\,\tau^2}\left[1 + \log\left(\frac{1}{2} + \frac{\tau}{2d\sqrt{2\pi}}\right)\right] + \frac{1}{\tau^2}.$$

In particular if snr is small the $d = 2\mathcal{C}/\nu$ is near 1 and the bound is near

$$\frac{4}{\nu\,\tau^2}\left[1 + \log\left(\frac{1}{2} + \frac{\tau}{2\sqrt{2\pi}}\right)\right] + \frac{1}{\tau^2}.$$

Finally, consider the case $\delta_c = snr$. The following lemma uses the form of $\Delta_{\zeta,snr}$ given in the remark following Lemma 26.

***Lemma** 36: Optimization of Δ_{ζ,δ_c} at $\delta_c = snr$.* The $\Delta_{\zeta,snr}$ is the maximum of the expressions

$$\bar{\Phi}(\zeta) + \frac{(1{+}D(snr)/snr)}{(1+snr)}(1+snr\,\bar{\Phi}(\zeta))(1{+}\zeta/\tau)^2 - 1$$

and

$$\bar{\Phi}(\zeta) + D(snr)/snr.$$

The first expression in this max is approximately of the form $b\,\bar{\Phi}(\zeta) + 2\zeta/\tau + c$, optimized at

$$\zeta = \sqrt{2\log(\tau(1{+}2\mathcal{C})/2\sqrt{2\pi})},$$

where $b = 1 + 2\mathcal{C}$ and c is equal to the negative value $(1/snr)\log(1{+}snr) - 1$, at which $\bar{\Phi}(\zeta) \le \phi(\zeta) = 2/(\tau b)$. This yields a bound for that expression near

$$\frac{2}{\tau} + \frac{2\sqrt{2\log(\tau(1{+}2\mathcal{C})/2\sqrt{2\pi})}}{\tau} + c,$$

with which we take the maximum of it and

$$\frac{2}{\tau(1{+}2\mathcal{C})} + \frac{D(snr)}{snr}.$$

Recall that $D(snr)/snr \le snr/2$. Because of the $D(snr)/snr$ term the $\Delta_{\zeta,snr}$ is small only when snr is small. In particular $\Delta_{\zeta,snr}$ is less than a constant time $\sqrt{\log\tau}/\tau$ when snr is less than such.

In view of the $\nu = snr/(1{+}snr)$ factor in the denominator of $\Delta_{\zeta,\delta_{match}}$, we see that $\min_\zeta \Delta_{\zeta,snr}$ provides a better bound than $\Delta_{\zeta,\delta_{match}}$ for snr less than a constant times $\sqrt{\log\tau}/\tau$.

Proof of Lemma 32 and its corollary: This Lemma concerns the optimization of ζ in the case $\delta_c = 0$. In this case $1+r_1/\tau^2 = (1{+}\zeta/\tau)^2$, the role of r^*_{crit} is played by r_{up} and the value of $\Delta_{\zeta,0,simp}$ is

$$\frac{1}{2\mathcal{C}}\frac{(2\tau{+}\zeta)\phi(\zeta)+rem}{\tau^2} + \frac{r_1+(2\tau{+}\zeta)\phi(\zeta)+rem}{\tau^2}.$$

Here $rem = -(r_1-1)\bar\Phi(\zeta)$, with $r_1 = 2\zeta\tau+\zeta^2$. Direct evaluation at $\zeta = 0$ gives a bound, at which $r_1 = 0$ and $rem = 1/2$.

Let's optimize $\Delta_{\zeta,0,simp}$ for the choice of ζ. The derivative of $(2\tau{+}\zeta)\phi(\zeta)+rem$ with respect to ζ is seen to simplify to $-2(\tau{+}\zeta)\bar\Phi(\zeta)$. Accordingly, $\Delta_{\zeta,0,simp}$ has derivative

$$2(\tau{+}\zeta)\left(1-(\frac{1}{2\mathcal{C}}+1)\bar\Phi(\zeta)\right),$$

which is 0 at ζ solving $\bar\Phi(\zeta) = 2\mathcal{C}/(2\mathcal{C}{+}1)$, equivalently, $\Phi(\zeta) = 1/(2\mathcal{C}{+}1)$. At this ζ, the quantities multiplying r_1 including the parts from the two occurrences of the remainder remainder are seen to cancel, such that the resulting value of $\Delta_{\zeta,0,simp}$ is

$$\frac{(1{+}1/2\mathcal{C})(2\tau{+}\zeta_{\mathcal{C}})\phi(\zeta_{\mathcal{C}})+1}{\tau^2}.$$

With $2\mathcal{C} > 1$, this $\zeta = \zeta_{\mathcal{C}}$ is negative, so $\Delta_{\zeta,0,simp}$ is not more than

$$\frac{(2\mathcal{C}+1)\phi(\zeta_{\mathcal{C}})}{\mathcal{C}\,\tau}+\frac{1}{\tau^2}.$$

Per the inequality in the appendix for negative ζ, the $\phi(\zeta_{\mathcal{C}})$ is not more than $\xi(|\zeta_{\mathcal{C}}|)\Phi(\zeta_{\mathcal{C}}) = \xi(|\zeta_{\mathcal{C}}|)/(2\mathcal{C}{+}1)$, with $\xi(|\zeta|)$ the nondecreasing function equal to 2 for $|\zeta|\le 1$ and equal to $|\zeta|+1/|\zeta|$ for $|\zeta|$ greater than 1. So at $\zeta = \zeta_{\mathcal{C}}$, the $\Delta_{\zeta,0,simp}$ is not more than

$$\frac{\xi(|\zeta_{\mathcal{C}}|)}{\mathcal{C}\,\tau}+\frac{1}{\tau^2},$$

where from $1/(2\mathcal{C}{+}1) = \Phi(\zeta_{\mathcal{C}}) \le (1/2)e^{-\zeta_{\mathcal{C}}^2/2}$ we have $|\zeta_{\mathcal{C}}| \le \sqrt{2\log(2\mathcal{C}{+}1)/2)}$.

The coefficient $\xi(\sqrt{2\log(2\mathcal{C}{+}1)/2})$ improves on the $(2\mathcal{C}+1)/\sqrt{2\pi}$ from the $\zeta = 0$ case when $(2\mathcal{C}+1)/2$ is less than the value val for which $\xi(\sqrt{2\log val}) = (2/\sqrt{2\pi})val$. Evaluations show val to be between 2.64 and 2.65. So it is an improvement when $2\mathcal{C} \ge 2val-1 = 4.3$, and $\mathcal{C} \ge 2.2$ suffices. The improvement is substantial for large $\mathcal{C}$.

Dividing the first term by $(1{+}\zeta_{\mathcal{C}}/\tau)^2$ produces an upper bound on $\Delta_{\zeta,0}$ when $\zeta_{\mathcal{C}} > -\tau$. Exact minimization of $\Delta_{\zeta,0}$ is possible, though it does not provide an explicit solution. Accordingly we instead use the $\zeta_{\mathcal{C}}$ that optimizes the simpler form and explore the implications of the division by $(1+\zeta_{\mathcal{C}}/\tau)^2$.

Consider determination of conditions on the size $\mathcal{C}$ such that the bound on $\min_\zeta \Delta_{zeta,0}$ is an improvement over the $\zeta = 0$ choice. One can arrange the $|\zeta_{\mathcal{C}}|/\tau$ to be small enough that the factor $(1{+}\zeta_{\mathcal{C}}/\tau)^2$ in the denominator remains sufficiently positive. At $\zeta = \zeta_{\mathcal{C}}$, the bound on $|\zeta_{\mathcal{C}}|$ of $\sqrt{2\log(2\mathcal{C}{+}1)/2)}$ is kept less than $\tau = \sqrt{2\log B}(1{+}\delta_a)$ when B is greater than $\mathcal{C}$, and $|\zeta_{\mathcal{C}}|/\tau$ is kept small if B is sufficiently large compared to $\mathcal{C}$.

In particular, suppose $B \ge 1+snr$, then $\tau^2/4$ is at least $\mathcal{C} = (1/2)\log(1+snr)$, that is, $\tau \ge \sqrt{4\mathcal{C}}$, and $(1{+}\zeta/\tau)$ is greater than $1-\sqrt{(1/2\mathcal{C})\log(2\mathcal{C}{+}1)/2)}$, which is positive for all $\mathcal{C} \ge 1/2$. Then for our non-zero $\zeta_{\mathcal{C}}$ bound on $\Delta_{\zeta,0}$ to provide improvement over the $\zeta = 0$ bound it is sufficient that $\mathcal{C}$ be at least the value $\mathcal{C}_0$ at which $(2\mathcal{C}{+}1)/\sqrt{2\pi}$ equals $\xi(\sqrt{2\log(2\mathcal{C}{+}1)/2})$ divided by $\big[1-2\sqrt{(1/2\mathcal{C})\log((2\mathcal{C}{+}1)/2)}\big]^2$. Numerical evaluation reveals that $\mathcal{C}_0$ is between 5.4 and 5.45.

Next, consider what to do for very large $\mathcal{C}$ for which $\tau+\zeta_{\mathcal{C}}$ is either negative or not sufficiently positive to give an effective bound. This could occur if snr is large compared to B. To overcome this problem, let $\mathcal{C}_{large}$ be the value with $\zeta_{Capacity_{large}} = -\tau+1$. For $\mathcal{C} \ge \mathcal{C}_{large}$, use this $\zeta = \zeta_{\mathcal{C}_{large}}$ in place of $\zeta_{\mathcal{C}}$ so that $\tau{+}\zeta = 1$ stays away from 0. Then upper bound $\Delta_{\zeta,0}$ by replacing the appearance of $\mathcal{C}$ with $\mathcal{C}_{large}$. This $\mathcal{C}_{large}$ has $\sqrt{2\log((2\mathcal{C}_{large}+1)/2)} \ge |\zeta| = \tau-1$ so that

$$2\mathcal{C}_{large}+1 \ge 2e^{(\tau-1)^2/2}.$$

More stringently,

$$\frac{1}{2\mathcal{C}_{large}+1} = \Phi(\zeta) = \Phi(\tau-1) \le \frac{1}{\tau-1}\phi(\tau-1),$$

from which $2\mathcal{C}_{large}{+}1$ is at least $(\tau-1)\sqrt{2\pi}e^{(\tau-1)^2/2}$. Then for $\mathcal{C} \ge \mathcal{C}_{large}$, at $\zeta = \zeta_{\mathcal{C}_{large}}$ the term

$$\frac{\tau\xi(|\zeta|)}{\mathcal{C}\,(\tau+\zeta)^2}$$

is less than

$$\frac{2\tau^2}{(\tau-1)\sqrt{2\pi}e^{(\tau-1)^2/2}-1}$$

which is exponentially small in $\tau^2/2$ and hence of order $1/B$ to within a log factor. Consequently, for such very large $\mathcal{C}$, this term is negligible compared to the $1/\tau^2$.

Finally, consider the matter of the range of $\mathcal{C}$ for which the expression in the first term $(2\mathcal{C}+1)\phi(\zeta_{\mathcal{C}})/[\mathcal{C}\,\tau(1{+}\zeta_{\mathcal{C}}/\tau)^2]$ is decreasing in $\mathcal{C}$ even with the presence of the division by $(1+\zeta_{\mathcal{C}}/\tau)^2$. Taking the derivative of this expression with respect to $\mathcal{C}$, one finds that there is a $\mathcal{C}_{crit}$, with value of $\zeta_{\mathcal{C}_{crit}}$ not much greater than $-\tau$, such that the expression is decreasing for $\mathcal{C}$ up to $\mathcal{C}_{crit}$, after which, for larger $\mathcal{C}$, it becomes preferable to use $\zeta = \zeta_{\mathcal{C}_{crit}}$ in place of $\zeta_{\mathcal{C}}$, though the determination of $\mathcal{C}_{crit}$ is not explicit. Nevertheless, one finds that at $\mathcal{C} = \mathcal{C}_{large}$ where $\zeta_{\mathcal{C}} = -\tau+1$, the derivative of the indicated expression is still negative and hence $\mathcal{C}_{large} \le \mathcal{C}_{crit}$. Thus the obtained bound is monotonically decreasing for $\mathcal{C}$ up to $\mathcal{C}_{large}$, and thereafter the bound for the first term is negligible. This completes the proof of Lemma 32 and its corollary.

Proof of Lemma 34: Recall for negative ζ we have that ψ is bounded by $2\tau[|\zeta|+1/|\zeta|]$. Likewise $\tilde r^*_{crit}/\tau^2$ is bounded by $\gamma^2/[2(1{+}\zeta/\tau)^2]$. Using $\gamma \le 2/|\zeta|\tau$ this yields $\tilde r^*_{crit}/\tau^2$ less than $2/[\zeta^2(\tau+\zeta)^2]$. Plugging in the chosen $\zeta = \zeta_{1/3}$ produces the claimed bound for that case. Likewise directly plugging in $\zeta = 0$ into the terms of $\Delta_{\zeta,\delta_{match}}$ provides a bound for that case. This completes the proof of Lemma 34.

Proof of Lemma 35: As previously developed, at $\delta_c = \delta_{match}$, the form of $\Delta_{\zeta,\delta_{match}}$ simplifies to

$$\frac{\psi}{2\mathcal{C}\,\tau^2(1+\zeta/\tau)^2} + \frac{1}{\nu}\,\frac{\tilde{r}^*_{crit}}{\tau^2} + \frac{1+\epsilon}{\tau^2}.$$

Now by Corollary 28, with $\zeta \geq 0$,

$$\tilde{r}^*_{crit} \,\leq\, 2\big(\zeta + \phi(\zeta)/\Phi(\zeta)\big)^2.$$

Also $\psi(\zeta) = 2\tau(1+\zeta/2\tau)\phi(\zeta)/\Phi(\zeta)$ and the $(1+\zeta/2\tau)$ factor is canceled by the larger $(1+\zeta/\tau)^2$ in the denominator. Accordingly, $\Delta_{\zeta,\delta_{match}}$ has the upper bound

$$\frac{\phi(\zeta)}{\mathcal{C}\,\tau\Phi(\zeta)} + \frac{2}{\nu\,\tau^2}\left(\zeta + \frac{\phi(\zeta)}{\Phi(\zeta)}\right)^2 + \frac{1+\epsilon}{\tau^2}.$$

Plugging in $\zeta = \zeta_\omega$ for which $\phi(\zeta)/\Phi(\zeta) = 2\omega$ produces the claimed bound.

$$\frac{2\omega}{\mathcal{C}\,\tau} + \frac{2}{\nu\,\tau^2}\,(\zeta_\omega + 2\omega)^2 + \frac{1+\epsilon}{\tau^2}.$$

To produce an explicit upper bound on $\Delta_{\zeta,\delta_{match}}$ replace the $\Phi(\zeta)$ in the denominator with its lower bound $1-\sqrt{2\pi}\phi(\zeta)/2$, for $\zeta \geq 0$. This lower bound agrees with $\Phi(\zeta)$ at $\zeta = 0$ and in the limit of large ζ. The resulting upper bound on $\Delta_{\zeta,\delta_{match}}$ is

$$\frac{\phi(\zeta)}{\mathcal{C}\,\tau\big(1-\sqrt{2\pi}\phi(\zeta)/2\big)} + \frac{2}{\nu\,\tau^2}\left(\zeta + \frac{\phi(\zeta)}{\big(1-\sqrt{2\pi}\phi(\zeta)/2\big)}\right)^2$$

plus $(1+\epsilon)/\tau^2$.

The bound on $\phi(\zeta)/\Phi(\zeta)$ of $\phi(\zeta)/\big(1-\sqrt{2\pi}\phi(\zeta)/2\big)$ is found to equal 2ω when $\sqrt{2\pi}\phi(\zeta)$ equals $2/[1+1/\omega\sqrt{2\pi}]$, at which $\zeta = \zeta^*$ is

$$\zeta^* = \sqrt{2\log\left(\frac{1}{2} + \frac{1}{2\omega\sqrt{2\pi}}\right)}.$$

Accordingly, this ζ^* upper bounds ζ_ω and the resulting bound on $\Delta_{\zeta^*,\delta_{match}}$ is

$$\frac{2\omega}{\mathcal{C}\,\tau} + \frac{2}{\nu\,\tau^2}\,(\zeta^* + 2\omega)^2 + \frac{1+\epsilon}{\tau^2}.$$

Using $2\omega = 4\mathcal{C}/\nu\tau$ it is

$$\frac{2}{\nu\,\tau^2}\left[2 + (\zeta + 2\omega)^2\right] + \frac{1+\epsilon}{\tau^2}.$$

This completes the proof of Lemma 35.

To provide further motivation for the choice ζ_ω, the derivative with respect to ζ of the above expression bounding $\Delta_{\zeta,\delta_{match}}$ for $\zeta \geq 0$ is seen, after factoring out $(4/\nu)(\zeta+\phi/\Phi)$, to equal

$$1 - \frac{1}{2\omega}\frac{\phi}{\Phi} - \big(\zeta + \frac{\phi}{\Phi}\big)\frac{\phi}{\Phi},$$

where the last term is negligible if ζ is not small. The first two yield 0 at $\zeta = \zeta_\omega$. Some improvement arises by exact minimization. Set the derivative to 0 including the last term, noting that it takes the form of a quadratic in ϕ/Φ. Then at the minimizer, ϕ/Φ equals $[\sqrt{(\zeta+1/2\omega)^2+4} - (\zeta+1/2\omega)]/2$ which is less than $1/(\zeta+1/2\omega) \leq 2\omega$.

For further understanding of the choice of ζ, note that for ζ not small, $\Phi(\zeta)$ is near 1 and the expression to bound is near $\phi(\zeta)/(\mathcal{C}\,\tau) + 2\zeta^2/\nu\tau^2$, which by analysis of its derivative is seen to be minimized at the positive ζ for which $\phi(\zeta)$ equals $4\mathcal{C}/\nu\tau = 2\omega$. It is $\zeta_1 = \sqrt{2\log 1/(2\omega\sqrt{2\pi})}$. One sees that ζ^* is similar to ζ_1, but has the addition of $1/2$ inside the logarithm, which is advantageous in allowing ω up to $1/\sqrt{2\pi}$. The difference between the use of ζ^* and ζ_1 is negligible when they are large (i.e. when ω is small), nevertheless, numerical evaluation of the resulting bound shows ζ^* to be superior to ζ_1 for all $\omega \leq 1/\sqrt{2\pi}$.

In the next section the rate expression is used to solve for the optimal choices of the remaining parameters.

X. Optimizing Parameters for Rate and Exponent

In this section we determine the parameters that maximize the communication rate for a given error exponent. Moreover, in the small exponent (large L) case, the rate and its closeness to capacity are determined as a function of the section size B and the signal to noise ratio snr.

Recall that the rate of our sparse superposition inner code is

$$R = \frac{(1-h')\mathcal{C}}{(1+\delta_a)^2(1+\delta_{sum}^2)(1+r/\tau^2)},$$

with $(1–h') = (1–h)(1–h_f)$. The inner code makes a weighted fraction of section mistakes bounded by $\delta_m = \delta^* + \eta + \bar{f}$ with high probability, as we have shown previously. If we multiply the weighted fraction by the factor $1/[L\min_\ell \pi_{(\ell)}]$ which equals $\text{fac} = snr\,(1+\delta_{sum}^2)/[2\mathcal{C}(1+\delta_c)]$, then it provides an upper bound on the (unweighted) fraction of mistakes $\delta_{mis} = \text{fac}\,\delta_m$ equal to

$$\delta_{mis} = \text{fac}\,(\delta^* + \eta + \bar{f}).$$

So with the Reed-Solomon outer code of rate $1-\delta_{mis}$, which corrects the remaining fraction of mistakes, the total rate of our code is

$$R_{tot} = \frac{(1-\delta_{mis})(1-h')\,\mathcal{C}}{(1+\delta_{sum}^2)(1+\delta_a)^2(1+r/\tau^2)}.$$

This multiplicative representation is appropriate considering the manner in which the contributions arise. Nevertheless, in choosing the parameters in combination, it is helpful to consider convenient and tight lower bounds on this rate, via an additive expression of rate drop from capacity.

***Lemma** 37: Additive representation of rate drop:* With a non-negative value for r, represented as in Remark 3 above, the rate R_{tot} is at least $(1-\Delta)\mathcal{C}$ with Δ given by

$$\Delta \;=\; \frac{snr\,\delta^*}{(1+\delta_c)2\mathcal{C}} + \frac{r^*_{crit}}{\tau^2} + \frac{(1+r_1/\tau^2)D(\delta_c)}{snr}$$
$$+\,\frac{snr}{2\mathcal{C}}(\eta+\bar{f}) + snr\,gap + h_f + h + 2\delta_a + \frac{2\mathcal{C}}{L\,\nu}.$$

These are called, respectively, the first and second lines of the expression for Δ. The first line of Δ is what we have also denoted in the introduction as Δ_{shape} or in the previous section as Δ_ζ to emphasize its dependence on ζ which determines the values of r_1, δ_c, and δ^*. In contrast the second line of Δ, which we denote Δ_{second}, depends on η, $\bar{f}$, and a. It has the ingredients of Δ_{alarm} and the quantities which determine the error exponent.

Proof of Lemma 37: Consider first the ratio

$$\frac{1-\delta_{mis}}{\big(1+\delta_{sum}^2\big)\big(1+r/\tau^2\big)}.$$

Splitting according to the two terms of the numerator and using the non-negativity of r it is at least

$$\frac{1}{\big(1+\delta_{sum}^2\big)\big(1+r/\tau^2\big)} \;-\; \frac{\delta_{mis}}{1+\delta_{sum}^2}.$$

From the form of fac, the ratio $\delta_{mis}/(1+\delta_{sum}^2)$ subtracted here is equal to

$$\frac{snr}{2\mathcal{C}}\frac{(\delta^*+\eta+\bar{f})}{(1+\delta_c)},$$

where in bounding it further we drop the $(1+\delta_c)$ from the terms involving $\eta+\bar{f}$, but find it useful to retain the term involving δ^*.

Concerning the factors of the first part of the above difference, use $\delta_{sum}^2 \leq D(\delta_c)/snr + 2\mathcal{C}/L\,\nu$ to bound the factor $(1+\delta_{sum}^2)$ by

$$\big(1+D(\delta_c)/snr\big)\big(1+2\mathcal{C}/L\,\nu\big).$$

and use the representation of $\big(1+D(\delta_c)/snr\big)\big(1+r/\tau^2\big)$ developed at the end of the previous section,

$$\left(1+\frac{(1+r_1/\tau^2)D(\delta_c)}{snr}+\frac{r_{crit}^*}{\tau^2}\right)\Big(1+\xi\, snr\, gap\Big)$$

to obtain that the first part of the above difference is at least

$$1\;-\;\left[\frac{r_{crit}^*}{\tau^2}+\frac{(1+r_1/\tau^2)D(\delta_c)}{snr}+snr\,gap+\frac{2\mathcal{C}}{L\,\nu}\right].$$

Proceed in this way, including also the factors $(1-h')$ and $1/(1+\delta_a)$ to produce the indicated bound on the rate drop from capacity. This bound is tight when the individual terms are small, because then the products are negligible in comparison. Here we have used $1/(1+\delta_i) \geq 1-\delta_i$ and $(1-\delta_1)(1-\delta_2)$ exceeds $1-\delta_1-\delta_2$, for non-negative reals δ_i, where the amount by which it exceeds is the product $\delta_1\delta_2$. Likewise inductively products $\prod_i(1-\delta_i)$ exceed $1-\sum_i\delta_i$. This completes the proof of Lemma 37.

This additive form of Δ provides some separation of effects that facilitates joint optimization of the parameters as in the next Lemma. Nevertheless, once the parameters are chosen, it is preferable to reexpress the rate in the original product form because of the slightly larger value it provides.

Let's recall parameters that arise in this rate and how they are interrelated. For the incremental false alarm target use

$$f^* = \frac{1}{\sqrt{2\pi}\sqrt{2\log B}}e^{-a\sqrt{2\log B}},$$

such that

$$\delta_a = \frac{\log 1/[f^*\sqrt{2\pi}\sqrt{2\log B}]}{2\log B}.$$

With a number of steps m at least 2 and with ρ at least 1, the total false alarms are controlled by $\bar{f} = mf^*\rho$ and the exponent associated with failed detections is determined by a positive η. Set h_f equal to $2\,snr\bar{f}$ plus the negligible $\epsilon_3 = 2snr\sqrt{(1+snr)k/L_\pi}+snr/L_\pi$, arising in the determination of the the weights of combination of the test statistic. To control the growth of correct detections set

$$gap = \eta+\bar{f}+1/(m-1).$$

The r_1, r_{crit}^*, δ^* and δ_c are determined as in the preceding section as functions of the positive parameter ζ.

The exponent of the error probability $e^{-L_\pi\mathcal{E}}$ is $\mathcal{E}=\mathcal{E}_\eta$ either given by

$$\mathcal{E}_\eta = 2\eta^2$$

or, if we use the Bernstein bound, by

$$\frac{1}{2}\frac{L}{L_\pi}\,\frac{\eta^2}{V+(1/3)\eta L/L_\pi}$$

where V is the minimum value of the variance function discussed previously. For our power allocation the $L_\pi = 1/\max_\ell\pi_{(\ell)}$ has L/L_π equal to $(2\mathcal{C}/\nu)(1+\delta_{sum}^2)$, which may be replaced by its lower bound $(2\mathcal{C}/\nu)$ yielding

$$\mathcal{E}_\eta = \frac{\eta^2}{V\nu/\mathcal{C}+(2/3)\eta}.$$

In both cases the relationship between $\mathcal{E}$ and η is strictly increasing on $\eta>0$ and invertible, such that for each $\mathcal{E}\geq 0$ there is a unique corresponding $\eta(\mathcal{E})\geq 0$.

Set the Chi-square concentration parameter h so that the exponent $(n-m+1)h_m^2/2$ matches $L_\pi\,\mathcal{E}_\eta$, where h_m equals $(nh-m+1)/(n-m+1)$. Thus $h_m = \sqrt{2\,\mathcal{E}_\eta\,L_\pi/(n-m+1)}$ which means

$$h = (m-1)/n + \sqrt{2\,\mathcal{E}_\eta\,L_\pi(n-m+1)}/n.$$

With $L_\pi \leq (\nu/2\mathcal{C})L$ not more than $(\nu/2)n/\log B$, it yields h not more than $(m-1)/n+h^*$ where

$$h^* = \sqrt{\nu\mathcal{E}_\eta/\log B}.$$

The part $(m-1)/n$ which is $(m-1)\mathcal{C}/L\log B$ is lumped with the above-mentioned remainders $2\mathcal{C}/L\nu$ and ϵ_3, as negligible for large L.

Finally, $\rho>1$ is chosen such that the false alarm exponent $\bar{f}\,\mathcal{D}(\rho)/\rho$ matches $\mathcal{E}_\eta$. The function $\mathcal{D}(\rho)/\rho = \log\rho - 1 + 1/\rho$ is 0 at $\rho=1$ and is an increasing function of $\rho\geq 1$ with unbounded positive range, so it has an inverse function $\rho(\mathcal{E})$ at which we set $\rho=\rho(\mathcal{E}_\eta/\bar{f})$.

It behooves us to pin down as many of these values as we can by exploring the best relationship between rate and error probability achieved by the analysis of our decoder.

We take advantage of the decomposition of Lemma 37.

***Lemma** 38: Optimization of the second line of* Δ. For any given positive η providing the exponent $\mathcal{E}_\eta$ of the error probability, the values of the parameters m, $\bar{f}$, and ρ, are specified to optimize their effect on the communication rate.

The second line Δ_{second} of the total rate drop $(\mathcal{C}–R)/\mathcal{C}$ bound Δ is the sum of three terms

$$\Delta_m \,+\, \Delta_{\bar{f}} \,+\, \Delta_\eta,$$

plus the negligible $\Delta_L = 2\mathcal{C}/(L\nu) + (m–1)\mathcal{C}/(L\log B) + \epsilon_3$. Here

$$\Delta_m \;=\; \frac{snr}{m-1} \,+\, \frac{\log m}{\log B}$$

is optimized at a number of steps m equal to an integer part of $2 + snr \log B$ at which Δ_m is not more than

$$\frac{1}{\log B} \,+\, \frac{\log(2{+}snr\log B)}{\log B}.$$

Likewise $\Delta_{\bar{f}}$ is given by

$$snr(3 + 1/2\mathcal{C})\bar{f} \;-\; \frac{\log\big(\bar{f}\sqrt{2\pi}\sqrt{2\log B}\,\big)}{\log B},$$

optimized at the false alarm level $\bar{f} \,=\, 1/\big[snr(3{+}1/2\mathcal{C})\log B\big]$ at which

$$\Delta_{\bar{f}} \,=\, \frac{1}{\log B} \,+\, \frac{\log\big(snr(3 + 1/2\mathcal{C})\sqrt{\log B}/\sqrt{4\pi}\big)}{\log B}.$$

The Δ_η is given by

$$\Delta_\eta \;=\; \eta\, snr(1{+}1/2\mathcal{C}) \,+\, \frac{\log\rho}{\log B} + h^*$$

evaluated at the optimal $\rho = \rho(\mathcal{E}_\eta/\bar{f})$. It yields Δ_η not more than

$$\eta\, snr(1{+}1/2\mathcal{C}) \,+\, \mathcal{E}_\eta\, snr(3{+}1/2\mathcal{C}) \,+\, 1/\log B \,+\, h^*.$$

Together the optimized $\Delta_m + \Delta_f$ form what is called Δ_{alarm} in the introduction. In the next lemma we use the Δ_η expression, or its inverse, to relate the error exponent to the rate drop.

Proof of Lemma 43: Recall that

$$2\delta_a = \frac{\log\big[\rho\, m/(\bar{f}\sqrt{2\pi}\sqrt{2\log B})\big]}{\log B}.$$

The log of the product is the sum of the logs. Associate the term $\log m/\log B$ with Δ_m and the term $\log\rho/\log B$ with Δ_η and leave the rest of $2\delta_a$ as part of $\Delta_{\bar{f}}$. The rest of the terms of Δ associate in the obvious way. Decomposed in this way, the stated optimizations of Δ_m and Δ_f are straightforward.

For $\Delta_m = snr/(m-1) + (\log m)/(\log B)$ consider it first as a function of real values $m \geq 2$. Its derivative is $-snr/(m-1)^2 + 1/(m\log B)$, which is negative at $m_1 = 1 + snr\log B$, positive at $m_2 = 2 + snr\log B$, and equal to 0 at a point $m_2^* = [m_2 + \sqrt{m_2^2 - 4}]/2$ in between m_1 and m_2. Moreover, the value of Δ_{m_2} is seen to be smaller than the value of m_1. Accordingly, for m in the interval $m_1 < m \leq m_2$, which includes an integer value, the Δ_m remains below what is attained for $m \leq m_1$. Therefore, the minimum among integers occurs at either at the floor $\lfloor 2 + snr\log B\rfloor$ or at the ceiling $\lceil 2 + snr\log B\rceil$ of m_2, whichever produces the smaller Δ_m. [Numerical evaluation confirms that the optimizer tends to coincide with the rounding of m_2^* to the nearest integer, coinciding with a near quadratic shape of Δ_m around m_2^*, by Taylor expansion for m not far from m_2^*.]

When the optimal integer m is less than or equal to $m_2 = 2 + snr\log B$, use that it exceeds m_1 to conclude that $\Delta_m \leq 1/\log B + (\log m_2)/(\log B)$. When the optimal m is a rounding up of m_2, use $snr/(m–1) \leq snr/(1{+}snr\log B)$. Also $\log m$ exceeds $\log m_2$ by the amount $\log(m/m_2) \leq \log(1 + 1/m_2)$ less than $1/(1 + snr\log B)$, to obtain that at the optimal integer, Δ_m remains less than

$$\frac{1}{\log B} + \frac{\log m_2}{\log B}.$$

For $\Delta_{\bar{f}}$ and Δ_η there are two ways to proceed. One is to use the above expression for δ_a, and set $\Delta_{\bar{f}}$ as indicated, which is easily optimized by setting $\bar{f}$ at the value specified.

For Δ_η note that the $\log\rho/\log B$ has numerator $\log\rho$ equal to $1 - 1/\rho + \mathcal{E}_\eta/\bar{f}$ at the optimized ρ, and accordingly we get the claimed upper bound by dropping the subtraction of $1/\rho$. This completes the proof of Lemma 43.

It is noted that in accordance with the inverse function $\rho(\mathcal{E}_\eta/\bar{f})$ there is an indirect dependence of the rate drop on $\bar{f}$ when $\mathcal{E}_\eta > 0$. One can jointly optimize $\Delta_{\bar{f}} + \Delta_\eta$ for $\bar{f}$ for given η, though there is not explicit formula for that solution. The optimization we have claimed is for $\Delta_{\bar{f}}$, which produces a clean expression suitable for use with small positive η.

A closely related presentation is to write

$$2\delta_a = \frac{\log\big[m/(\bar{f}^*\sqrt{2\pi}\sqrt{2\log B})\big]}{\log B}$$

and in other terms involving $\bar{f}$, write it as $\rho\bar{f}^*$. Optimization of

$$snr(3 + 1/2\mathcal{C})\rho\bar{f}^* \;-\; \frac{\log\big(\bar{f}^*\sqrt{2\pi}\sqrt{2\log B}\,\big)}{\log B},$$

occurs at a baseline false alarm level $\bar{f}^*$ that is equal to $1/\big[\rho\, snr(3{+}1/2\mathcal{C})\log B\big]$. These approaches have the baseline level of false alarms (as well as the final value of δ_a) depending on the subsequent choice of ρ.

One has a somewhat cleaner separation in the story, as in the introduction, if $\bar{f}^*$ is set independent of ρ. This is accomplished by a different way of spitting the terms of Δ_{second}. One writes $\bar{f} = \rho\bar{f}^*$ as $\bar{f}^* + (\rho–1)\bar{f}^*$, the baseline value plus the additional amount required for reliability. Then set $\Delta_{\bar{f}^*}$ to equal

$$snr(3 + 1/2\mathcal{C})\bar{f}^* \;-\; \frac{\log\big(\bar{f}^*\sqrt{2\pi}\sqrt{2\log B}\,\big)}{\log B},$$

optimized at $\bar{f}^* = 1/\big[snr(3{+}1/2\mathcal{C})\log B\big]$, which determines a value of δ_a for the rate drop envelope independent of η. In that approach one replaces Δ_η with

$$\eta\, snr(1{+}1/2\mathcal{C}) \,+\, (\rho{-}1)snr(3{+}1/2\mathcal{C}) + h^*,$$

with ρ defined to solve $\bar{f}^*\mathcal{D}(\rho) = \mathcal{E}_\eta$. There is not an explicit solution to the inverse of $\mathcal{D}(\rho)$ at $\mathcal{E}_\eta/\bar{f}^*$. Nevertheless, a satisfactory bound for small η is obtained by replacing $\mathcal{D}(\rho)$ by its lower bound $2(\sqrt{\rho} - 1)^2$, which can be explicitly inverted. Perhaps a downside is that from the form of the $\bar{f}^*$ which minimizes $\Delta_{\bar{f}^*}$ one ends up, multiplying by ρ, with a final $\bar{f}$ larger than before.

With $2(\sqrt{\rho}-1)^2$ replacing $\mathcal{D}(\rho)$, it is matched to $2\eta^2/\bar{f}^*$ by setting $\sqrt{\rho}-1 = \eta/\sqrt{\bar{f}^*}$ and solving for ρ by adding 1 and squaring. The resulting expression used in place of Δ_η is then a quadratic equation in η, for which its root provides means by which to express the relationship between rate drop and error exponent. Then $\rho\bar{f}^*$ is $\big(\sqrt{\bar{f}^*}+\eta\big)^2$.

A twist here, is that in solving for the best $\bar{f}^*$, rather than starting from $\eta = 0$, one may incorporate positive η in the optimization of

$$snr(3+1/2\mathcal{C})\big(\sqrt{\bar{f}^*}+\eta\big)^2 \; - \; \frac{\log\big(\bar{f}^*\sqrt{2\pi}\sqrt{2\log B}\,\big)}{\log B},$$

for which, taking the derivative with respect to $\sqrt{\bar{f}^*}$ and setting it to 0, a solution for this optimization is obtained as the root of a quadratic equation in $\sqrt{\bar{f}^*}$. Upon adding to that the other relevant terms of Δ_{second}, namely $\eta\, snr(1{+}1/2\mathcal{C}) + h^*$, one would have an explicit, albeit complicated, expression remaining in η.

Set $\Delta_B = \Delta(snr, B)$ equal to $\Delta_{shape} + \Delta_m + \Delta_f$ at the above values of ζ, m, $\bar{f}$ (or should we use $\bar{f}^*$?). This Δ_B provides the rate drop envelope as a function only of snr and B. It corresponding to the large L regime in which one may take η to be small. Accordingly, Δ_B provides the boundary of the behavior by evaluating Δ with $\eta = 0$.

The given values of m and $\bar{f}$ optimize Δ_B, and the given ζ provides a tight bound, approximately optimizing the rate drop envelope Δ_B. The associated total rate R_{tot} evaluated at these choices of parameters with $\eta = 0$, denoted $\mathcal{C}_B$, is at least $\mathcal{C}(1-\Delta_B)$. The associated bound on the fraction of mistakes of the inner code is $\delta^*_{mis} = (snr/2\mathcal{C})(\delta^* + \bar{f})$.

Express the Δ_η bound as a strictly increasing function of the error exponent $\mathcal{E}$

$$\eta(\mathcal{E})\, snr(1{+}1/2\mathcal{C}){+}\mathcal{E}(3{+}1/2\mathcal{C}){+}\frac{1-1/\rho(\mathcal{E}/\bar{f})}{\log B}{+}\sqrt{\nu\mathcal{E}/\log B}$$

and let $\mathcal{E}(\Delta)$ denote its inverse for $\Delta \geq 0$, [recognizing also per the statement of the Lemma above the cleaner upper bound dropping the $1/\rho(\mathcal{E}/\bar{f})/\log B$ term]. The part $\eta(\mathcal{E})\, snr/2\mathcal{C}$ within the first term is from the contribution to $2\delta_{mis}$ in the outer code rate. From the rate drop of the superposition inner code, the rest of Δ_η written as a function of $\mathcal{E}$ is denoted $\Delta_{\eta,super}$ and we let $\mathcal{E}_{super}(\Delta)$ denote its inverse function.

For a given total rate $R_{tot} < \mathcal{C}_B$, an associated error exponent $\mathcal{E}$ is

$$\mathcal{E}\big((\mathcal{C}_B - R_{tot})/\mathcal{C}\big),$$

which is the evaluation of that inverse at $(\mathcal{C}_B - R_{tot})/\mathcal{C}$. Alternatively, in place of $\mathcal{C}_B$ we may use its lower bound $\mathcal{C}(1{-}\Delta_B)$ and take the error exponent to be $\mathcal{E}\big(1{-}R_{tot}/\mathcal{C}{-}\Delta_B\big)$. We show either choice provides an error exponent of a code of that specified total rate.

To arrange the constituents of this code, use the inner code mistake rate bound $\delta_{mis} = fac\,(\delta^*{+}\bar{f}{+}\eta(\mathcal{E}))$, and set the inner code rate target $R = R_{tot}/(1-\delta_{mis})$. Accordingly, for any number of sections L, set the codelength n, to be $L\log B/R$ rounded to an integer, so that the inner code rate $L\log B/n$ agrees with the target rate to within a factor of $1 \pm 1/n$, and the total code rate $(1{-}\delta_{mis})R$ agrees with R_{tot} to within the same precision.

***Theorem** 39: Rate and Reliability of the composite code:* As a function of the section size B, let $\mathcal{C}_B$ and its lower bound $\mathcal{C}(1-\Delta_B)$ be the rate envelopes given above, both near the capacity $\mathcal{C}$ for B large. Let a positive $R_{tot} < \mathcal{C}_B$ be given. If $R_{tot} \leq \mathcal{C}(1{-}\Delta_B)$, set the error exponent $\mathcal{E}$ by

$$\mathcal{E}\big(1{-}\Delta_B{-}R_{tot}/\mathcal{C}\big).$$

Alternatively, to arrange the somewhat larger exponent, with η such that $\Delta_\eta = (\mathcal{C}_B - R_{tot})/\mathcal{C}$, suppose that $\Delta_\eta \geq \delta_{mis}$; then set $\mathcal{E} = \mathcal{E}_\eta$, that is, $\mathcal{E} = \mathcal{E}\big((\mathcal{C}_B - R_{tot})/\mathcal{C}\big)$. To allow any $R_{tot} < \mathcal{C}_B$ without further condition, there is a unique $\eta > 0$ such that $\Delta_{\eta,super}\,\mathcal{C} = \mathcal{C}_B/(1{-}\delta^*_{mis}){-}R_{tot}/(1{-}\delta_{mis})$, at which we may set $\mathcal{E} = \mathcal{E}_\eta$. In any of these three cases, for any number of sections L, the code consisting of a sparse superposition code and an outer Reed-Solomon code, having composite rate equal to R_{tot}, to within the indicated precision, has probability of error not more than

$$\kappa e^{-L_\pi \mathcal{E}},$$

which is exponentially small in L_π, near $L\nu/(2\mathcal{C})$, where $\kappa = m(1+snr)^{1/2}B^c + 2m$ is a polynomial in B with $c = snr\,\mathcal{C}$, with number of steps m equal to the integer part of $1 + snr\,\log B$.

Proof of Theorem 39 for rate assumption $R_{tot} < \mathcal{C}(1-\Delta_B)$**:** Set $\eta > 0$ such that $\Delta_\eta = 1-\Delta_B - R_{tot}/\mathcal{C}$. Then the rate R_{tot} is expressed in the form $\mathcal{C}(1-\Delta_B-\Delta_\eta)$. In view of Lemma **??** and the development preceding it, this rate $\mathcal{C}(1{-}\Delta) = \mathcal{C}(1{-}\Delta_B{-}\Delta_\eta)$ is a lower bound on a rate of the established form $(1{-}\delta_{mis})\mathcal{C}'/(1{+}r/\tau^2)$, with parameter values that permit the decoder to be accumulative up to a point x^* with shortfall δ^*, providing a fraction of section mistakes not more than $\delta_{mis} = fac\,(\delta^*{+}\eta{+}\bar{f})$, except in an event of the indicated probability with exponent $\mathcal{E}_h = \mathcal{E}(\Delta_\eta)$. This fraction of mistakes is corrected by the outer code. The probability of error bound from our earlier theorem is

$$m\,e^{-L_\pi\mathcal{E}+m\mathcal{C}} \;+\; 2m\,e^{-L_\pi\mathcal{E}}.$$

With $m \leq 1{+}snr\log B$ it is not more than the given $\kappa e^{-L_\pi\mathcal{E}}$. The other part of the Theorem asserts a similar conclusion but with an improved exponent associated with arranging $\Delta_\eta = (\mathcal{C}_B - R_{tot})/\mathcal{C}$, that is, $R_{tot} = \mathcal{C}_B(1-\Delta_\eta)$. We return to demonstrate that conclusion as a corollary of the next result.

One has the option to state our results in terms of properties of the inner code. At any section size B, recognize that Δ_B above, at the $\eta = 0$ limit, splits into a contribution from $\delta^*_{mis} = (snr/2\mathcal{C})(\bar{f} + \delta^*/(1+\delta_c))$ and the rest which is a bound on the rate drop of the inner superposition code, which we denote Δ^*_{super}, in this small η limit. The rate envelope for such superposition codes is

$$\mathcal{C}^*_{super} = \frac{(1{-}2snr\,\bar{f})\mathcal{C}}{(1{+}D(\delta_c)/snr)[(1{+}\delta_a)^2{+}r/(2\log B)]},$$

evaluated at $\bar{f}$, δ_a, δ_c, r and ζ as specified above, with $\eta = 0$, $h = 0$ and $\rho = 1$, again with a number of steps m equal to the integer part of $1+snr\,\log B$. It has

$$\mathcal{C}^*_{super} \geq \mathcal{C}\,(1 - \Delta^*_{super}).$$

Likewise recall that Δ_η splits into the part $\eta(\mathcal{E})\,snr/2\mathcal{C}$ associated with δ_{mis} and the rest $\Delta_{\eta,super}$ expressed as a function of $\mathcal{E}$, for which $\mathcal{E}_{super}(\Delta)$ is its inverse.

***Theorem** 40: Rate and Reliability of the Sparse Superposition Code:* For any rate $R < \mathcal{C}^*_{super}$, let $\mathcal{E}$ equal

$$\mathcal{E}_{super}(\mathcal{C}^*_{super} - R)/\mathcal{C}).$$

Then for any number of sections L, the rate R sparse superposition code with adaptive successive decoder, makes a fraction of section mistakes less than $\delta^*_{mis} + \eta(\mathcal{E})\,snr/2\mathcal{C}$ except in an event of probability less than $\kappa\,e^{-L_\pi \mathcal{E}}$.

This conclusion about the sparse superposition code would also hold for values of the parameters other than those specified above, producing related tradeoffs between rate and the reliable fraction of section mistakes. Our particular choices of these parameters is specific to the tradeoff that produces the best total rate of the composite code.

Proof of Theorem 40. In view of the preceding analysis, what remains to establish is that the rate

$$\mathcal{C}^*_{super}(1 - \Delta_{\eta,super})$$

is not more than our rate expression

$$\frac{\mathcal{C}(1-h_f)(1-h^*)}{(1+D(\delta_c)/snr)(1+\delta_{a,\rho})^2(1+r_\eta/\tau^2)}$$

where $\Delta_{\eta,super}$ which is

$$\eta\,snr(1+r_1/2\log B) + \mathcal{E}_\eta(3+1/C)(1+1/\log B) + \sqrt{\nu\mathcal{E}/\log B}$$

is at least

$$\eta\,snr(1+r_1/\tau^2) + (\log\rho)/\log B + h^*.$$

with ρ and h^* satisfying the conditions of the Lemma, so that (once we account for the negligible remainder in $1/L$), the indicated reliability holds with this rate. Here we are writing $\delta_{a,\rho} = \delta_a + (\log\rho)/2\log B$ to distinguish the value that occurs with $\rho > 1$ with the value at $\rho = 0$ used in the definition of $\mathcal{C}^*_{super}$. Likewise we are writing r_η/τ^2 for the expression $r/\tau^2 + \eta\,snr(1+r_1/\tau^2)$ to distinguish the value that occurs with $\eta > 0$ with the value of r/τ^2 at $\eta = 0$ used in the definition of $\mathcal{C}^*_{super}$. Factoring out terms in common, what is to be verified is that

$$\frac{1 - \Delta_{\eta,super}}{(1+\delta_a)^2(1+r/\tau^2)}$$

is not more than

$$\frac{(1-h^*)}{(1+\delta_{a,\rho})^2)(1+r_\eta/\tau^2]}.$$

This is seen to be true by cross multiplying, rearranging, expanding the square in $(1+\delta_a + \log\rho/2\log B)^2$, using the lower bound on $\Delta_{\eta,super}$, and comparing term by term for the parts involving h^*, $\log\rho$ and η. This completes the proof of Theorem 40.

Next we prove the rest of Theorem 39, in view of what has been established. For the general rate condition $R_{tot} < \mathcal{C}_B$, for $\eta \geq 0$ the expression

$$\Delta_{\eta,super}\,\mathcal{C} + \frac{R_{tot}}{1 - \delta^*_{mis} - snr\,\eta/2\mathcal{C}}$$

is a strictly increasing function of η in the interval $[0, (2\mathcal{C}/snr)(1-\delta^*_{mis}))$, where the second term in this expression may be interpreted as the rate R of an inner code, with total rate R_{tot}. This function starts at $\eta = 0$ at the value $R_{tot}/(1-\delta^*_{mis})$ which is less than $\mathcal{C}_B/(1-\delta^*_{mis})$ which is $\mathcal{C}^*_{super}$. So there is an η in this interval at which this function hits $\mathcal{C}^*_{super}$. That is $\Delta_{\eta,super}\mathcal{C} + R = \mathcal{C}^*_{super}$, or equivalently, $\Delta_{\eta,super} = (\mathcal{C}^*_{super} - R)/\mathcal{C}$. So Theorem 40 applies with exponent $\mathcal{E}_{super}((\mathcal{C}^*_{super} - R)/\mathcal{C}))$.

Finally, to obtain the exponent $\mathcal{E}((\mathcal{C}_B - R_{tot})/\mathcal{C}))$, let $\Delta_\eta = \mathcal{C}_B - R_{tot}/\mathcal{C}$. Examine the rate

$$\mathcal{C}_B(1 - \Delta_\eta)$$

which is

$$(1-\delta^*_{mis})\mathcal{C}^*_{super}(1 - \Delta_{\eta,super} - \eta snr/2\mathcal{C})$$

and determine whether it is not more than the following composite rate (obtained using the established inner code rate),

$$(1 - \delta^*_{mis} - \eta snr/2\mathcal{C})\mathcal{C}^*_{super}(1 - \Delta_{\eta,super}).$$

These match to first order. Factoring out $\mathcal{C}^*_{super}$ and canceling terms shared in common, the question reduces to whether $-(1-\delta^*_{mis})$ is not more than $-(1-\Delta_{\eta,super})$, that is, whether δ^*_{mis} is not more than $\Delta_{\eta,super}$, or equivalently, whether δ_{mis} is not more than Δ_η, which is the condition assumed in the Theorem for this case. This completes the proof of Theorem 39.

XI. Lower bounds on error exponent:

The second line of the rate drop can be decomposed as

$$\Delta_m + \Delta_{\bar{f}^*} + \Delta_{\eta,\rho},$$

where

$$\Delta_m = \frac{snr}{m-1} + \frac{\log m}{\log B}$$

optimized at a number of steps m equal to to an integer part of $2 + snr\log B$. Further,

$$\Delta_{\bar{f}^*} = \vartheta\bar{f}^* - \frac{\log\left(\bar{f}^*\sqrt{2\pi}\sqrt{2\log B}\right)}{\log B}$$

where $\vartheta = snr(3+1/2\mathcal{C})$. The the optimum value of $\bar{f}^*$ equal to $1/[\vartheta\log B]$ and

$$\Delta_{\eta,\rho} = \eta\vartheta_1 + (\rho-1)/\log B + h.$$

Here $\vartheta_1 = snr(1+1/2\mathcal{C})$.

We have $\Delta_{\eta,\rho}$ is a strictly increasing function of the error exponent $\mathcal{E}$, where

$$\Delta_{\eta,\rho} = \vartheta_1\eta(\mathcal{E}) + (\rho-1)/\log B + h.$$

Let $R_{tot} \leq \mathcal{C}_B$ be given. We need to find the error exponent $\mathcal{E}^* = \mathcal{E}\left((\mathcal{C}_B - R_{tot})/\mathcal{C}\right)$, where $\mathcal{E}$ solves the above equation with $\Delta_{\eta,\rho} = (\mathcal{C}_B - R_{tot})/\mathcal{C}$. That is,

$$\Delta_{\eta,\rho} = \vartheta_1 \eta(\mathcal{E}) + (\rho - 1)/\log B + h,$$

where $\rho = \rho(\mathcal{E}/\bar{f}^*)$.

Now $\rho - 1 = (\sqrt{\rho} - 1)(\sqrt{\rho} + 1)$, which is $(\sqrt{\rho} - 1)^2 + 2(\sqrt{\rho} - 1)$. Correspondingly, using $\mathcal{E} \geq 2\bar{f}^*(\sqrt{\rho} - 1)^2$, we get that $\mathcal{E}/2\bar{f}^* + \sqrt{2\mathcal{E}/\bar{f}^*} \geq \rho - 1$. Further, using $\eta\left(\mathcal{E}\right) = \sqrt{\mathcal{E}/2}$ and $h^* = \sqrt{\nu\mathcal{E}/\log B}$, one gets that

$$\Delta_{\eta,\rho} \leq c_1 \mathcal{E} + c_2 \sqrt{\mathcal{E}},$$

where

$$c_1 = \vartheta/2$$

and

$$c_2 = \left[\frac{\vartheta_1}{\sqrt{2}} + \sqrt{2\vartheta/\log B} + \sqrt{\frac{\nu}{\log B}}\right].$$

Solving the above quadratic in $\sqrt{\mathcal{E}}$ given above, one gets that

$$\mathcal{E} \geq \mathcal{E}_{sol} = \left[\frac{-c_2 + \sqrt{c_2^2 + 4\Delta_{\eta,\rho} c_1}}{2c_1}\right]^2.$$

Let us see what $\mathcal{E}_{sol}$ looks like for $\Delta_{\eta,\rho}$ near 0. Noticing that $\mathcal{E}_{sol}$ has the shape $\Delta_{\eta,\rho}^2$ for $\Delta_{\eta,\rho}$ near 0, we want to find the limit of $\mathcal{E}_{sol}/\Delta_{\eta,\rho}^2$ as $\Delta_{\eta,\rho}$ goes to zero. Using L' Hospital's rule one get that this limiting value is $1/c_2^2$. Correspondingly, using L_π is near $L\nu/2\mathcal{C}$, one gets that the error exponent is near

$$\exp\left\{-L\Delta_{\eta,\rho}^2/\xi_0\right\},$$

for $\Delta_{\eta,\rho}$ near 0, where $\xi_0 = (2\mathcal{C}/\nu)c_2^2$. This quantity behaves like $snr^2\mathcal{C}$ for large snr and has the limiting value of $(1 + 4/\sqrt{\log B})^2/2$ for snr tending to 0.

We now give a simplified expression for $\mathcal{E}_{sol}$. To simplify this, lower bound the function $-a + \sqrt{a^2 + x}$, with $x \geq 0$ with a function of the form $\min\left\{\alpha\sqrt{x}, \beta x\right\}$. It is seen that

$$-a + \sqrt{a^2 + x} \geq \alpha\sqrt{x} \quad \text{for} \quad x \geq \frac{4\alpha^2 a^2}{(1 - \alpha^2)^2}$$

and

$$-a + \sqrt{a^2 + x} \geq \beta x \quad \text{for} \quad x \leq \frac{1 - 2\beta a}{\beta^2}.$$

Clearly, for the above to have any meaning one requires $0 < \alpha < 1$ and $0 < \beta < 1/2a$. Further, it is seen that

$$\begin{aligned}\min\left\{\alpha\sqrt{x}, \beta x\right\} &= \alpha\sqrt{x} \quad \text{for} \quad x \geq (\alpha/\beta)^2 \\ &= \beta x \quad \text{for} \quad x \leq (\alpha/\beta)^2.\end{aligned}$$

Correspondingly, equating $(\alpha/\beta)^2$ with $4\alpha^2 a^2/(1 - \alpha^2)^2$, or equivalently equating $(\alpha/\beta)^2$ with $(1 - 2\beta a)/\beta^2$, we get that $1 - \alpha^2 = 2a\beta$.

We now return to the problem of lower bounding $\mathcal{E}_{sol}$. We take $a = c_2$ and $x = 4\Delta_{\eta,\rho}c_1$. We also take particular choices of β and α to simplify the analysis. We take $\beta = 1/4a$, for which $\alpha = 1/\sqrt{2}$. Then the above gives that

$$\mathcal{E}_{sol} \geq \frac{\left(\min\left\{\alpha\sqrt{4\Delta_{\eta,\rho}c_1},\, \beta 4\Delta_{\eta,\rho}c_1\right\}\right)^2}{4c_1^2}$$

which simplifies to

$$\mathcal{E}_{sol} \geq \min\left\{\Delta_{\eta,\rho}/2c_1,\, \Delta_{\eta,\rho}^2/4c_2^2\right\}.$$

From Theorem 39, one get that the error probability is bounded by

$$\kappa e^{-L_\pi \mathcal{E}_{sol}},$$

which from the above, can also be bounded by the more simplified expression

$$\kappa \exp\left\{-L_\pi \min\left\{\Delta_{\eta,\rho}/2c_1,\, \Delta_{\eta,\rho}^2/4c_2^2\right\}\right\}.$$

We want to express this bound in the form,

$$\kappa \exp\left\{-L \min\left\{\Delta_{\eta,\rho}/\xi_1, \Delta_{\eta,\rho}^2/\xi_2\right\}\right\}$$

for some ξ_1, ξ_2. Using the fact that L_π is near $L\nu/2\mathcal{C}$, one gets that ξ_1 is $(2\mathcal{C}/\nu)(2c_1)$, which gives

$$\xi_1 = (1 + snr)(6\mathcal{C} + 1).$$

We see that ξ_1 goes to 1 as snr tends to zero. Further $\xi_2 = (2\mathcal{C}/\nu)4c_2^2$. which behaves like $4\mathcal{C}\, snr^2$ for large snr. It has the limiting value of $2(1 + 4/\sqrt{\log B})^2$ as snr tends to zero.

Improvement for Rates near Capacity using Bernstein bounds: The improved error bound associated with correct detection is given by

$$\exp\left\{-\frac{\eta^2}{2(V_{tot} + \eta/(3L_\pi))}\right\},$$

where $V_{tot} = V/L$, with $V \leq \tilde{c}_v$, where $\tilde{c}_v = (4\mathcal{C}/\nu^2)(a_1 + a_2/\tau^2)/\tau$. For small η, that is for rates near the rate envelope, the bound behaves like,

$$\exp\left\{-L\frac{\eta^2}{2V}\right\}.$$

Consequently, for such η the exponent is,

$$\mathcal{E} = \frac{1}{d_1}\frac{\eta^2}{2\tilde{c}_v}.$$

Here $d_1 = L_\pi/L$. This corresponds to $\eta = \sqrt{d_2}\sqrt{\mathcal{E}}$, where $d_2 = 2d_1\tilde{c}_v$. Here $\tilde{c}_v = (4\mathcal{C}/\nu^2)(a_1/\tau)$ and that $d_1 = \nu/2\mathcal{C}$ and $\tau \geq \sqrt{2\log B}$, one gets that $d_2 \leq 1.62/\nu\sqrt{\log B}$. Substituting this upper bound for η in the expression for $\Delta_{\eta,\rho}$, we get that

$$\Delta_{\eta,\rho} \leq \tilde{c}_1\mathcal{E} + \tilde{c}_2\sqrt{\mathcal{E}},$$

with $\tilde{c}_1 = \vartheta/2$ and

$$\tilde{c}_2 = \left[\sqrt{d_2}\vartheta_1 + \sqrt{2\vartheta/\log B} + \sqrt{\frac{\nu}{\log B}}\right].$$

Consequently using the same reasoning as above one gets that using the Bernstein bound, for rates close to capacity, the error exponent is like

$$\exp\left\{-L\Delta_{\eta,\rho}^2/\tilde{\xi}_0\right\},$$

for $\Delta_{\eta,\rho}$ near 0, where $\tilde{\xi}_0 = (2\mathcal{C}/\nu)\tilde{c}_2^2$. This quantity behaves like $2d_2 snr^2\mathcal{C}$ for large snr. Further, $2d_2$ is near $3.24/\sqrt{\log B}$ for such snr. Notice now the error exponent is proportional to $L\sqrt{\log B}\Delta_{\eta,\rho}^2$, instead of the $L\Delta_{\eta,\rho}^2$ we had before. We see that for $B > 36300$, the quantity $3.24/\sqrt{\log B}$ is less than one producing a better exponent that before for rates near capacity and for large snr than before.

XII. Optimizing Parameters for Rate and Exponent for No Leveling using the $1-x\nu$ Factor:

From Corollary 20 one gets that

$$GAP = \frac{r - r_{up}}{\nu(\tau^2 + r)}.$$

Simplifying one gets

$$1 + r/\tau^2 = (1 + r_{up}/\tau^2)/(1 - \nu GAP).$$

Recall that the rate of our sparse superposition inner code is

$$R = \frac{(1-h')\mathcal{C}}{(1+\delta_a)^2(1+\delta_{sum}^2)(1+r/\tau^2)}.$$

Here we use the terms involved in the leveling case, even though we are considering the no leveling case here. This will be useful later on when we are generalizing to the case with the leveling. Further, with the Reed-Solomon outer code of rate $1-\delta_{mis}$, which corrects the remaining fraction of mistakes, the total rate of our code is

$$R_{tot} = \frac{(1-\delta_{mis})(1-h')\,\mathcal{C}}{(1+\delta_{sum}^2)(1+\delta_a)^2(1+r/\tau^2)}.$$

which using the above is equal to

$$R_{tot} = \frac{(1-\delta_{mis})(1-h')(1-\nu GAP)\,\mathcal{C}}{(1+\delta_{sum}^2)(1+\delta_a)^2(1+r_{up}/\tau^2)}.$$

***Lemma** 41: Additive representation of rate drop:* With a non-negative value for GAP less than $1/\nu$ the rate R_{tot} is at least $(1-\Delta)\mathcal{C}$ with Δ given by

$$\Delta \;=\; \frac{snr\,\delta^*}{(1+\delta_c)2\mathcal{C}} + \frac{r_{up}}{\tau^2} + \frac{D(\delta_c)}{snr}$$
$$+\frac{snr}{2\mathcal{C}}(\eta + \bar{f}) + \nu\, GAP + h' + 2\delta_a + \frac{2\mathcal{C}}{L\,\nu}.$$

Proof of Lemma 41: Notice that

$$\frac{(1-\delta_{mis})(1-h')(1-\nu GAP)}{(1+\delta_{sum}^2)(1+\delta_a)^2(1+r_{up}/\tau^2)}$$

is at least

$$\frac{(1-h')(1-\nu GAP)}{(1+\delta_{sum}^2)(1+\delta_a)^2(1+r_{up}/\tau^2)}$$

minus $\delta_{mis}/(1+\delta_{sum}^2)$. As before, the ratio $\delta_{mis}/(1+\delta_{sum}^2)$ subtracted here is equal to

$$\frac{snr}{2\mathcal{C}}\frac{(\delta^* + \eta + \bar{f})}{(1+\delta_c)}.$$

Further the first part of the difference is at least

$$1 - h' - \nu GAP - \delta_{sum}^2 - 2\delta_a - r_{up}/\tau^2.$$

Further using $\delta_{sum}^2 \le D(\delta_c)/snr + 2\mathcal{C}/L\nu$ one gets the result. This completes the proof of Lemma 41.

The second line of the rate drop is given by,

$$\frac{snr}{2\mathcal{C}}(\eta(x^*) + \bar{f}) + \nu\, GAP + h' + 2\delta_a + \frac{2\mathcal{C}}{L\,\nu}.$$

where

$$\eta(x^*) = (1 - x^* v)\eta^{std}$$

and

$$GAP = \eta^{std} + \log 1/(1-x^*)/(m-1).$$

Thus νGAP is equal to

$$\nu\eta^{std} + \nu\, c(x^*)/(m-1),$$

where

$$c(x^*) = \log[1/(1-x^*)].$$

Case 1: $h' = h + h_f$ **:** We now optimize the second line of the rate drop when $h' = h + h_f$. We have the following lemma.

***Lemma** 42: Optimization of the second line of* Δ. For any given positive η providing the exponent $\mathcal{E}_\eta$ of the error probability, the values of the parameters m, $\bar{f}^*$ are specified to optimize their effect on the communication rate. The second line Δ_{second} of the total rate drop $(\mathcal{C}-R)/\mathcal{C}$ bound Δ is the sum of three terms

$$\Delta_m \;+\; \Delta_{\bar{f}^*} \;+\; \Delta_{\eta(x^*)},$$

plus the negligible $\Delta_L = 2\mathcal{C}/(L\nu) + (m-1)\mathcal{C}/(L\log B) + \epsilon_3$. Here

$$\Delta_m \;=\; \nu\frac{c(x^*)}{m-1} \;+\; \frac{\log m}{\log B}$$

is optimized at a number of steps m equal to an integer part of $2 + \nu\, c(x^*)\log B$ at which Δ_m is not more than

$$\frac{1}{\log B} + \frac{\log(2 + \;\nu\, c(x^*)\log B)}{\log B}$$

Likewise $\Delta_{\bar{f}^*}$ is given by

$$\vartheta \bar{f}^* \;-\; \frac{\log\left(\bar{f}^*\sqrt{2\pi}\sqrt{2\log B}\right)}{\log B},$$

where $\vartheta = snr(2 + 1/2\mathcal{C})$. The above is optimized at the false alarm level $\bar{f}^* = 1/\left[\vartheta\log B\right]$ at which

$$\Delta_{\bar{f}^*} \;=\; \frac{1}{\log B} \;+\; \frac{\log\left(\vartheta\sqrt{\log B}/\sqrt{4\pi}\right)}{\log B}.$$

The $\Delta_{\eta(x^*)}$ is given by

$$\Delta_{\eta(x^*)} \;=\; \eta^{std}\left[\nu + (1 - x^*\nu)snr/2\mathcal{C}\right] \;+\; (\rho - 1)/\log B + h$$

which is bounded by,

$$\eta^{std}\,\vartheta_1 \;+\; (\rho-1)/\log B + h$$

where $\vartheta_1 = \nu + snr/2\mathcal{C}$.

Remark : Since $1 - x^* = r/(snr\,\tau^2)$ and that $r > r_{up}$, one has that $1 - x^* \ge r_{up}/snr\tau^2$. Correspondingly, $c(x^*)$ is at most $\log(snr) + \log(\tau^2/r_{up})$. The optimum number of steps can be bounded accordingly.

Proof: Club all terms involving the number of steps m to get the expression for Δ_m. It is then seen that optimization of Δ_m give the expression as in the proof statement.

Next, write

$$2\delta_a = \frac{\log\left[m/(\bar{f}^*\sqrt{2\pi}\sqrt{2\log B})\right]}{\log B}.$$

Further write $\bar{f}$ as $(\rho-1)\bar{f}^* + \bar{f}^*$ in terms involving $\bar{f}$. For example $h_f = 2snr\bar{f}$ is written as the sum of $2snr(\rho-1)\bar{f}^*$

plus $2snr\bar f^*$. Now club all terms involving only $\bar f^*$ (that is not $(\rho-1)\bar f^*$) into $\Delta_{\bar f^*}$. We get $\Delta_{\bar f^*}$ to equal

$$\vartheta \bar f^* \;-\; \frac{\log\big(\bar f^*\sqrt{2\pi}\sqrt{2\log B}\,\big)}{\log B},$$

optimized at $\bar f^* = 1/\big[\vartheta\log B\big]$, which determines a value of δ_a for the rate drop envelope independent of η.

The remaining terms are absorbed to give the expression for Δ_η. Thus we get Δ_η is equal to

$$\eta^{std}[\nu{+}(1-x^*\nu)snr/2\mathcal{C}] \;+\; (\rho{-}1)/\log B + h^*.$$

The bound on $\Delta_{\eta(x^*)}$ follows from using $1-x^*\nu\le 1$.

Error exponent: We prefer to use Bernstein bounds for the error bounds associated with correct detection. Recall that that for rates near the rate envelope, that for $\eta(x^*)$ close to 0, the exponent is near

$$\mathcal{E} = \frac{1}{d_1}\frac{(\eta^{std})^2}{2\tilde c_v}.$$

As before, this corresponds to $\eta^{std}=\sqrt{d_2}\sqrt{\mathcal{E}}$, where $d_2 = 2d_1\tilde c_v$. Substituting this upper bound for η^{std} in the expression for $\Delta_{\eta(x^*)}$, we get that

$$\Delta_{\eta,\rho}\le \tilde c_1\mathcal{E}+\tilde c_2\sqrt{\mathcal{E}},$$

with $\tilde c_1=\vartheta/2$ and

$$\tilde c_2 = \left[\sqrt{d_2}\vartheta_1+\sqrt{2\vartheta/\log B}+\sqrt{\frac{\nu}{\log B}}\right].$$

Consequently using the same reasoning as above one gets that using the Bernstein bound, for rates close to capacity, the error exponent is like

$$\exp\left\{-L\Delta_{\eta,\rho}^2/\tilde\xi_0\right\},$$

for $\Delta_{\eta,\rho}$ near 0, where $\tilde\xi_0=(2\mathcal{C}/\nu)\tilde c_2^2$. For small snr, this quantity is near $(\sqrt{d_2}+\sqrt{2/\log B})^2$. This quantity behaves like $d_2 snr^2/2\mathcal{C}$ for large snr. For large snr this is the same as what we got in the previous section.

Case 2 : $1-h' = (1-h)^{m-1}/(1+h)^{m-1}$**.** It is easy to that this implies that $h'\le 2mh$. We give the corresponding Lemma for optimization of rate drop for such h'.

***Lemma** 43: Optimization of the second line of* Δ. For any given positive η providing the exponent $\mathcal{E}_\eta$ of the error probability, the values of the parameters m, $\bar f^*$ are specified to optimize their effect on the communication rate. The second line Δ_{second} of the total rate drop $(\mathcal{C}{-}R)/\mathcal{C}$ bound Δ is the sum of three terms

$$\Delta_m \;+\; \Delta_{\bar f^*} \;+\; \Delta_{\eta(x^*)},$$

plus the negligible $\Delta_L = 2\mathcal{C}/(L\nu)+(m{-}1)\mathcal{C}/(L\log B)+\epsilon_3$. Here

$$\Delta_m \;=\; \nu\frac{c(x^*)}{m-1} \;+\; \frac{\log m}{\log B}$$

is optimized at a number of steps m^* equal to an integer part of $2+\nu\,c(x^*)\log B$ at which Δ_m is not more than

$$\frac{1}{\log B}+\frac{\log(2+\,\nu\,c(x^*)\log B)}{\log B}$$

Likewise $\Delta_{\bar f^*}$ is given by

$$\vartheta \bar f^* \;-\; \frac{\log\big(\bar f^*\sqrt{2\pi}\sqrt{2\log B}\,\big)}{\log B},$$

where $\vartheta = snr/2\mathcal{C}$. The above is optimized at the false alarm level $\bar f^* = 1/\big[\vartheta\log B\big]$ at which

$$\Delta_{\bar f^*} \;=\; \frac{1}{\log B} \;+\; \frac{\log\big(\vartheta\sqrt{\log B}/\sqrt{4\pi}\big)}{\log B}.$$

The $\Delta_{\eta(x^*)}$ is given by

$$\Delta_{\eta(x^*)} \;=\; \eta^{std}\,[\nu{+}(1{-}x^*\nu)snr/2\mathcal{C}] + (\rho{-}1)/\log B{+}2m^*h$$

which is bounded by,

$$\eta^{std}\,\vartheta_1 \;+\; (\rho-1)/\log B + 2m^*h$$

where $\vartheta_1 = \nu{+}snr/2\mathcal{C}$.

Error exponent: Exactly similar to before, we use Bernstein bounds for the correct detection error probabilities to get that,

$$\Delta_{\eta,\rho}\le \tilde c_1\mathcal{E}+\tilde c_2\sqrt{\mathcal{E}},$$

with $\tilde c_1=\vartheta/2$ and

$$\tilde c_2 = \left[\sqrt{d_2}\vartheta_1+\sqrt{2\vartheta/\log B}+2m^*\sqrt{\frac{\nu}{\log B}}\right].$$

Notice that since $m^* = 2+\nu c(x^*)\log B$, one has that

$$\tilde c_2 = \left[\sqrt{d_2}\vartheta_1+\sqrt{\frac{2\vartheta}{\log B}}+4\sqrt{\frac{\nu}{\log B}}+2c(x^*)\nu^{3/2}\sqrt{\log B}\right].$$

As before the error exponent is like

$$\exp\left\{-L\Delta_{\eta,\rho}^2/\tilde\xi_0\right\},$$

for $\Delta_{\eta,\rho}$ near 0, where $\tilde\xi_0=(2\mathcal{C}/\nu)\tilde c_2^2$.

Comparison of envelope and exponent for the two methods with and without factoring $1-x\nu$ **term :** We first concentrate on the envelope, which is given by

$$\frac{1}{\log B}+\frac{m^*}{\log B}+\frac{1}{\log B}\;+\;\frac{\log\big(\vartheta\sqrt{\log B}/\sqrt{4\pi}\big)}{\log B}.$$

For the first method, without factoring out the $1-x\nu$ term, $\vartheta = snr(3{+}1/2\mathcal{C})$ whereas for the second method (Case 2) it is $snr/2\mathcal{C}$. The optimum number of steps m^* is $2{+}snr\log B$ for the first method and is $2+\nu c(x^*)\log B$. Here $c(x^*)$ is $\log(snr)$ plus a term of order $\log\log B$. Correspondingly, the envelope is smaller for the second method.

Next we see that the quantity c_2 is determines the error exponent. The smaller the c_2, the better the exponent. For the first method it is

$$\left[\frac{\vartheta_1}{\sqrt{2}}+\sqrt{2\vartheta/\log B}+\sqrt{\frac{\nu}{\log B}}\right].$$

where $\vartheta_1 = snr(1+1/2\mathcal{C})$. Further ϑ is as given in the previous paragraph. For the second method it is given by

$$\left[\frac{\vartheta_1}{\sqrt{2}}+\sqrt{2\vartheta/\log B}+2m^*\sqrt{\frac{\nu}{\log B}}\right],$$

where $\vartheta_1 = \nu + snr/2\mathcal{C}$. It is seen that for larger snr the latter is less producing a better exponent. To see this, notice that as a function of snr, the first term in c_2, i.e. $\vartheta_1/\sqrt{2}$, behaves like snr for the first case (without factorization of $1 - x\nu$) and is like $snr/2\mathcal{C}$ for the second case. The second term in c_2 is like $\sqrt{snr}$ for the first case and like $\sqrt{snr/2\mathcal{C}}$ for the second. The third is near $\sqrt{1/\log B}$ for the former case and behaves like $\log(snr)$ in the latter case. Consequently, it is the first term in c_2 which determines it behavior for larger snr for both cases. Since ϑ_1 is smaller in the second case, we infer that the second method is better for larger snr.

XIII. Composition with an Outer Code

We use Reed-Solomon (RS) codes ([59], [50]) to correct any remaining mistakes from our adaptive successive decoder. The symbols for the RS code can be associated with that of a Galois field, say consisting of q elements and denoted by $GF(q)$. Here q is typically taken to be of the form of a power of two, say 2^m. Let K_{out}, n_{out} be the message and blocklength respectively for the RS code. Further, if d_{RS} be the minimum distance between the codewords, then an RS code with symbols in $GF(2^m)$ can have the following parameters:

$$n_{out} = 2^m$$
$$n_{out} - K_{out} = d_{RS} - 1$$

Here $n_{out} - K_{out}$ gives the number of parity check symbols added to the message to form the codeword. In what follows we find it convenient to take B to be equal to 2^m so that one can view each symbol in $GF(2^m)$ as giving a number between 1 and B.

We now demonstrate how the RS code can be used as an outer code in conjunction with our inner superposition code, to achieve low block error probability. For simplicity assume that B is a power of 2. First consider the case when L equals B. Taking $m = \log_2 B$, we have that since L is equal to B, the RS codelength becomes L. Thus, one can view each symbol as representing an index specifying the selected term in each of the L sections. The number of input symbols is then $K_{out} = L - d_{RS} + 1$, so setting $\delta = d_{RS}/L$, one sees that the outer rate $R_{out} = K_{out}/n_{out}$, equals $1 - \delta + 1/L$ which is at least $1 - \delta$.

For code composition $K_{out} \log_2 B$ message bits become the K_{out} input symbols to the outer code. The symbols of the outer codeword, having length L, gives the labels of terms sent from each section using our inner superposition with codelength $n = L \log_2 B / R_{inner}$. From the received Y the estimated labels $\hat{j}_i, \hat{j}_2, \ldots \hat{j}_L$ using our adaptive successive decoder can be again thought of as output symbols for our RS codes. If $\hat{\delta}_e$ denotes the section mistake rate, it follows from the distance property of the outer code that if $2\hat{\delta}_e \leq \delta$ then these errors can be corrected. The overall rate R_{comp} is seen to be equal to the product of rates $R_{out} R_{inner}$ which is at least $(1 - \delta) R_{inner}$. Since we arrange for $\hat{\delta}_e$ to be smaller than some δ_{mis} with exponentially small probability, it follows from the above that composition with an outer code allows us to communicate with the same reliability, albeit with a slightly smaller rate given by $(1 - 2\delta_{mis}) R_{inner}$.

The case when $L < B$ can be dealt with by observing ([50], Page 240) that an (n_{out}, K_{out}) RS code as above, can be shortened by length w, where $0 \leq w < K_{out}$, to form an $(n_{out} - w, K_{out} - w)$ code with the same minimum distance d_{RS} as before. This is seen by viewing each codeword as being created by appending $n_{out} - K_{out}$ parity check symbols to the end of the corresponding message string. Then the code formed by considering the set of codewords with the w leading symbols identical to zero has precisely the properties stated above.

With B equal to 2^m as before, we have n_{out} equals B so taking w to be $B - L$ we get an (n'_{out}, K'_{out}) code, with $n'_{out} = L$, $K'_{out} = L - d_{RS} + 1$ and minimum distance d_{RS}. Now since the codelength is L and the symbols of this code are in $GF(B)$ the code composition can be carried out as before.

Appendix I
Distribution of $\mathcal{Z}_{k,j}$

Consider the general $k \geq 2$ case. Focus on the sequence of coefficients

$$\mathcal{Z}_{1,j}, \mathcal{Z}_{2,j}, \ldots, \mathcal{Z}_{k-1,j},\ V_{k,k,j}, V_{k+1,k,j}, \ldots, V_{n,k,j}$$

used to represent X_j for j in J_{k-1} in the basis

$$\frac{G_1}{\|G_1\|}, \frac{G_2}{\|G_2\|}, \ldots, \frac{G_{k-1}}{\|G_{k-1}\|}, \ \xi_{k,k}, \xi_{k+1,k}, \ldots \xi_{n,k},$$

where the $\xi_{i,k}$ for i from k to n are orthonormal vectors in $\mathbb{R}^n$, orthogonal to the $G_1, G_2, \ldots, G_{k-1}$. These are associated with the previously described representation $X_j = \sum_{k'=1}^{k-1} \mathcal{Z}_{k',j}\, G_{k'}/\|G_{k'}\| + V_{k,j}$, except that here $V_{k,j}$ is represented as $\sum_{i=k}^{n} V_{i,k,j}\, \xi_{i,k}$.

Let's prove that conditional on $\mathcal{F}_{k-1}$, the distribution of the $V_{i,k,j}$ is independent across i from k to n, and for each such i the joint distribution of $(V_{i,k,j} : j \in J_{k-1})$ is Normal $N_{J_{k-1}}(0, \Sigma_{k-1})$. The proof is by induction in k. Along the way the conditional distribution properties of G_k, $Z_{k,j}$, and $\mathcal{Z}_{k,j}$ are obtained as consequences. As for $\hat{w}_k$ and δ_k the induction steps provide recursions which permit verification of the stated forms.

The $V_{i,1,j} = X_{i,j}$ are independent standard normals.

To analyze the $k = 2$ case, use the vectors $U_{1,j} = U_j$ that arise in the first step properties in the proof of Lemma 1. There we saw for unit vectors α, that the $U_j^T \alpha$ for $j \in J_1$ have a joint $N_{J_1}(0, \Sigma_1)$ distribution, independent of Y. When represented using the orthonormal basis $Y/\|Y\|, \xi_{2,2}, \ldots, \xi_{n,2}$, the vector U_j has coefficients $Z_j = U_j^T Y/\|Y\|$, and $U_j^T \xi_{2,2}$ through $U_j^T \xi_{n,2}$. Accordingly $X_j = b_{1,j} Y/\sigma + U_j$ has representation in this basis with the same coefficients, except in the direction $Y/\|Y\|$ where Z_j is replaced by $\mathcal{Z}_j = b_{1,j}\|Y\|/\sigma + Z_j$. The joint distribution of $(V_{i,2,j} = U_j^T \xi_{i,2} : j \in J_1)$ is Normal $N_{J_1}(0, \Sigma_1)$, independently for $i = 2$ to n, and independent of $\|Y\|$ and $(Z_j : j \in J_1)$.

Proceed inductively for $k \geq 2$, presuming the stated conditional distribution property of the $V_{i,k,j}$ to be true at k, conduct analysis to demonstrate its validity at $k+1$.

From the representation of $V_{k,j}$ in the basis given above, the G_k has representation in the same basis as $G_{i,k} =$

$\sum_{j \in dec_{k-1}} \sqrt{P_j}\, V_{i,k,j}$ for i from k to n. The coordinates less than k are 0, since the $V_{k,j}$ and G_k are orthogonal to $G_1, \ldots, G_{k-1}$. The value of $\mathcal{Z}_{k,j}$ is $V_{k,j}^T G_k \,/\|G_k\|$ where the inner product (and norm) may be computed in the above basis from sums of products of coefficients for i from k to n.

For the conditional distribution of $G_{i,k}$ given $\mathcal{F}_{k-1}$, independence across i, conditional normality and conditional mean 0 are properties inherited from the corresponding properties of the $V_{i,k,j}$. To obtain the conditional variance of $G_{i,k} = \sum_{j \in dec_{k-1}} \sqrt{P_j}\, V_{i,k,j}$, use the conditional covariance $\Sigma_{k-1} = I - \delta_{k-1}\delta_{k-1}^T$ of $V_{i,k,j}$ for j in J_{k-1}. The identity part contributes $\sum_{j \in dec_{k-1}} P_j$ which is $(\hat{q}_{k-1} + \hat{f}_{k-1})P$; whereas, the $\delta_{k-1}\delta_{k-1}^T$ part, using the presumed form of δ_{k-1}, contributes an amount seen to equal $\nu_{k-1}[\sum_{j \in sent \cap dec_{k-1}} P_j/P]^2\, P$ which is $\nu_{k-1}\hat{q}_{k-1}^2 P$. It follows that the conditional expected square for the coefficients of G_k is

$$\sigma_k^2 \;=\; \big[\, \hat{q}_{k-1} + \hat{f}_{k-1} - \hat{q}_{k-1}^2\, \nu_{k-1} \,\big]\, P.$$

Moreover, conditional on $\mathcal{F}_{k-1}$, the distribution of $\|G_k\|^2 = \sum_{i=k}^n G_{i,k}^2$ is that of $\sigma_k^2\, \mathcal{X}_{n-k+1}^2$, a multiple of a Chi-square with $n-k+1$ degrees of freedom.

Next represent $V_{k,j} = b_{k,j}\, G_k/\sigma_k + U_{k,j}$ using a value of $b_{k,j}$ that follows an update rule that we specify (depending on $\mathcal{F}_{k-1}$). It is represented using $V_{i,k,j} = b_{k,j}\, G_{i,k}/\sigma_k + U_{i,k,j}$ for i from k to n, using the basis built from the $\xi_{i,k}$.

The coefficient $b_{k,j}$ is the value $\mathbb{E}[V_{i,k,j} G_{i,k} | \mathcal{F}_{k-1}]/\sigma_k$. Consider the product $V_{i,k,j}\, G_{i,k}$ in the numerator. Use the representation of $G_{i,k}$ as a sum of the $\sqrt{P_{j'}}\, V_{i,k,j'}$ for $j' \in dec_{k-1}$. Accordingly, the numerator is $\sum_{j' \in dec_{k-1}} \sqrt{P_{j'}}\, \big[1_{j'=j} \,-\, \delta_{k-1,j}\delta_{k-1,j'}\big]$, which simplifies to $\sqrt{P_j}\, \big[1_{j \in dec_{k-1}} \,-\, \nu_{k-1}\hat{q}_{k-1} 1_{j\ sent}\big]$. So for j in $J_k = J_{k-1} - dec_{k-1}$, we have the simplification

$$b_{k,j} = -\, \frac{\hat{q}_{k-1}\, \nu_{k-1} \beta_j}{\sigma_k},$$

for which the product for j, j' in J_k takes the form

$$b_{k,j} b_{k,j'} = \delta_{k-1,j}\delta_{k-1,j'} \frac{\hat{q}_{k-1}\nu_{k-1}}{1 + \hat{f}_{k-1}/\hat{q}_{k-1} - \hat{q}_{k-1}\nu_{k-1}}.$$

Here the ratio simplifies to $\hat{q}_{k-1}^{adj}\nu_{k-1}/(1 - \hat{q}_{k-1}^{adj}\nu_{k-1})$.

Now determine the features of the joint normal distribution of the $U_{i,k,j} = V_{i,k,j} - b_{k,j}\, G_{i,k}/\sigma_k$ for $j \in J_k$, given $\mathcal{F}_{k-1}$. These random variables are conditionally uncorrelated and hence conditionally independent given $\mathcal{F}_{k-1}$ across choices of i, but there is covariance across choices of j for fixed i. This conditional covariance $\mathbb{E}[U_{i,k,j} U_{i,k,j'} | \mathcal{F}_{k-1}]$ by the choice of $b_{k,j}$ reduces to $\mathbb{E}[V_{i,k,j} V_{i,k,j'} | \mathcal{Z}] - b_{k,j} b_{k,j'}$ which, for $j \in J_k$, is $1_{j=j'} - \delta_{k-1,j}\delta_{k-1,j'} - b_{k,j} b_{k,j'}$. That is, for each i, the $(U_{i,k,j} : j \in J_k)$ have the joint $N_{J_k}(0, \Sigma_k)$ distribution, conditional on $\mathcal{F}_{k-1}$, where Σ_k again takes the form $1_{j,j'} - \delta_{k,j}\delta_{k,j'}$ where

$$\delta_{k,j}\delta_{k,j'} \;=\; \delta_{k-1,j}\delta_{k-1,j'} \left\{ 1 \,+\, \frac{\hat{q}_{k-1}^{adj}\, \nu_{k-1}}{1 - \hat{q}_{k-1}^{adj}\nu_{k-1}} \right\},$$

for j, j' now restricted to J_k. The quantity in braces simplifies to $1/(1 - \hat{q}_{k-1}^{adj}\nu_{k-1})$. Correspondingly, the recursive update rule for ν_k is

$$\nu_k \;=\; \frac{\nu_{k-1}}{1 \,-\, \hat{q}_{k-1}^{adj}\, \nu_{k-1}}.$$

Consequently, the joint distribution for $(Z_{k,j} : j \in J_k)$ is determined, conditional on $\mathcal{F}_{k-1}$. It is also the normal $N(0, \Sigma_k)$ distribution and $(Z_{k,j} : j \in J_k)$ is conditionally independent of the coefficients of G_k, given $\mathcal{F}_{k-1}$. After all, the $Z_{k,j} = U_{k,j}^T G_k \,/\|G_k\|$ have this $N_{J_k}(0, \Sigma_k)$ distribution, conditional on G_k and $\mathcal{F}_{k-1}$, but since this distribution does not depend on G_k we have the stated conditional independence.

Now $\mathcal{Z}_{k,j} = X_j^T G_k \,/\|G_k\|$ reduces to $V_{k,j}^T G_k \,/\|G_k\|$ by the orthogonality of the G_1 through G_{k-1} components of X_j with G_k. So using the representation $V_{k,j} = b_{k,j}\, G_k/\sigma_k + U_{k,j}$ one obtains

$$\mathcal{Z}_{k,j} \;=\; b_{k,j}\, \|G_k\|/\sigma_k \,+\, Z_{k,j}.$$

This makes the conditional distribution of the $\mathcal{Z}_{k,j}$, given $\mathcal{F}_{k-1}$, close to but not exactly normally distributed, rather it is a location mixture of normals with distribution of the shift of location determined by the Chi-square distribution of $\mathcal{X}_{n-k+1}^2 = \|G_k\|^2/\sigma_k^2$. Using the form of $b_{k,j}$, for j in J_k, the location shift $b_{k,j}\, \mathcal{X}_{n-k+1}$ may be written

$$-\sqrt{\hat{w}_k\, C_{j,R,B}}\, \big[\, \mathcal{X}_{n-k+1}/\sqrt{n}\, \big]\, 1_{j\ sent},$$

where $\hat{w}_k$ equals $n\, b_{k,j}^2/C_{j,R,B}$. The numerator and denominator has dependence on j through P_j, so canceling the P_j produces a value for $\hat{w}_k$. Indeed, $C_{j,R,B} = (P_j/P)\nu(L/R)\log B$ equals $n(P_j/P)\nu$ and $b_{k,j}^2 = P_j \hat{q}_{k-1}^{adj}\, \nu_{k-1}^2/[1 - \hat{q}_{k-1}^{adj}\nu_{k-1}]$. So this $\hat{w}_k$ may be expressed as

$$\hat{w}_k \;=\; \frac{\nu_{k-1}}{\nu}\; \frac{\hat{q}_{k-1}^{adj}\, \nu_{k-1}}{1 - \hat{q}_{k-1}^{adj}\nu_{k-1}},$$

which, using the update rule for ν_{k-1}, is seen to equal

$$\hat{w}_k = \frac{\nu_{k-1} - \nu_k}{\nu}.$$

Armed with G_k, update the orthonormal basis of $\mathbb{R}^n$ used to represent X_j, $V_{k,j}$ and $U_{k,j}$. From the previous step this basis was $G_1/\|G_1\|, \ldots, G_{k-1}/\|G_{k-1}\|$ along with $\xi_{k,k}, \xi_{k+1,k}, \ldots, \xi_{n,k}$, where only the later are needed for the $V_{k,j}$ and $U_{k,j}$ as their coefficients in the directions $G_1, \ldots G_{k-1}$ are 0.

Now Gram-Schmidt makes an updated orthonormal basis of $\mathbb{R}^n$, retaining the $G_1/\|G_1\|, \ldots, G_{k-1}/\|G_{k-1}\|$, but replacing $\xi_{k,k}, \xi_{k+1,k}, \ldots, \xi_{n,k}$ with $G_k/\|G_k\|$, $\xi_{k+1,k+1}, \ldots, \xi_{n,k+1}$. By the Gram-Schmidt construction process, these vectors $\xi_{i,k+1}$ for i from $k{+}1$ to n are determined from the original basis vectors (columns of the identity) along with the computed random vectors $G_1, \ldots, G_k$ and do not depend on any other random variables in this development.

The coefficients of $U_{k,j}$ in this updated basis are $U_{k,j}^T G_k/\|G_k\|$, $U_{k,j}^T \xi_{k+1,k+1}, \ldots, U_{k,j}^T \xi_{n,k+1}$, which are denoted $U_{k,k+1,j} = Z_{k,j}$ and $U_{k+1,k+1,j}, \ldots, U_{k+1,n,j}$, respectively. Recalling the normal conditional distribution of the $U_{k,j}$, these coefficients $(U_{i,k+1,j} : k \le i \le n,\, j \in J_k)$ are also normally distributed, conditional on $\mathcal{F}_{k-1}$ and G_k, independent across i from k to n (this independence being a consequence of their uncorrelatedness, due to the orthogonality of the $\xi_{i,k+1}$ and the independence of the coefficients $U_{i,k,j}$ across i in the original basis); moreover, as we have seen already for $i = k$, for each i from k to n, the $(U_{i,k+1,j} : j \in J_k)$ inherit a joint normal $N(0, \Sigma_k)$ conditional distribution from the conditional distribution that the $(U_{i,k,j} : j \in J_k)$ have. After all, these coefficients have this conditional distribution, conditioning on the basis vectors and $\mathcal{F}_{k-1}$, and this conditional distribution is the same for all such basis vectors. So, in fact, these $(U_{i,k+1,j} : k \le i \le n,\, j \in J_k)$ are conditionally independent of the G_k given $\mathcal{F}_{k-1}$.

Specializing the conditional distribution conclusion, by separating off the $i = k$ case where the coefficients are $Z_{k,j}$, one has that the $(U_{i,k+1,j} : k+1 \le i \le n,\, j \in J_k)$ have the specified conditional distribution and are conditionally independent of G_k and $(Z_{k,j} : j \in J_k)$ given $\mathcal{F}_{k-1}$. It follows that the conditional distribution of $(U_{i,k+1,j} : k+1 \le i \le n,\, j \in J_k)$ given $\mathcal{F}_k = (\mathcal{F}_{k-1}, \|G_k\|, Z_k)$ is identified. It is normal $N(0, \Sigma_k)$ for each i, independently across i from $k+1$ to n, conditionally given $\mathcal{F}_k$.

Likewise, the vector $V_{k,j} = b_{k,j}\, G_k/\sigma_k + U_{k,j}$ has representation in this updated basis with coefficient $\mathcal{Z}_{k,j}$ in place of $Z_{k,j}$ and with $V_{i,k+1,j} = U_{i,k+1,j}$ for i from $k+1$ to n. So these coefficients $(V_{i,k+1,j} : k+1 \le i \le n,\, j \in J_k)$ have the normal $N(0, \Sigma_k)$ distribution for each i, independently across i from $k+1$ to n, conditionally given $\mathcal{F}_k$.

Thus the induction is established, verifying this conditional distribution property holds for all $k = 1, 2, \ldots, n$. Consequently, the Z_k and $\|G_k\|$ have the claimed conditional distributions.

Finally, repeatedly apply $\nu_{k'}/\nu_{k'-1} = 1/(1 - \hat{q}^{adj}_{k'-1}\, \nu_{k'-1})$, for k' from k to 2, each time substituting the required expression on the right and simplifying to obtain

$$\frac{\nu_k}{\nu_{k-1}} \;=\; \frac{1 \,-\, (\hat{q}^{adj}_1 + \ldots + \hat{q}^{adj}_{k-2})\; \nu}{1 - (\hat{q}^{adj}_1 + \ldots + \hat{q}^{adj}_{k-2} + \hat{q}^{adj}_{k-1})\, \nu}.$$

This yields $\nu_k = \nu \hat{s}_k$, which, when plugged into the expressions for $\hat{w}_k$, establishes the claims. The proof of Lemma 2 is complete.

Appendix II
The Method of Nearby Measures

Recall that the Rènyi relative entropy of order $\alpha > 1$ (also known as the α divergence) of two probability measures $\mathbb{P}$ and $\mathbb{Q}$ with density functions $p(Z)$ and $q(Z)$ for a random vector Z is given by

$$D_\alpha(\mathbb{P}\|\mathbb{Q}) = \frac{1}{\alpha - 1} \log \mathbb{E}_{\mathbb{Q}}[(p(Z)/q(Z))^\alpha].$$

Its limit for large α is $D_\infty(\mathbb{P}\|\mathbb{Q}) = \log \|p/q\|_\infty$.

Lemma 44: Let $\mathbb{P}$ and $\mathbb{Q}$ be a pair of probability measures with finite $D_\alpha(\mathbb{P}\|\mathbb{Q})$. For any event A, and $\alpha > 1$,

$$\mathbb{P}[A] \;\le\; \left[\mathbb{Q}[A] e^{D_\alpha(\mathbb{P}\|\mathbb{Q})}\right]^{(\alpha-1)/\alpha}.$$

If $D_\alpha(\mathbb{P}\|\mathbb{Q}) \le c_0$ for all α, then the following bound holds, taking the limit of large α,

$$\mathbb{P}[A] \;\le\; \mathbb{Q}[A] e^{c_0}.$$

In this case the density ratio $p(Z)/q(Z)$ is uniformly bounded by e^{c_0}.

Proof of Lemma 44: For convex f, as in Csiszar's f-divergence inequality, from Jensen's inequality applied to the decomposition of $\mathbb{E}_{\mathbb{Q}}[f(p(Z)/q(Z))]$ using the distributions conditional on A and its complement,

$$\mathbb{Q}A\, f(\mathbb{P}A/\mathbb{Q}A) \,+\, \mathbb{Q}A^c\, f(\mathbb{P}A^c/\mathbb{Q}A^c) \;\le\; \mathbb{E}_{\mathbb{Q}} f(p(Z)/q(Z)).$$

Using in particular $f(r) = r^\alpha$ and throwing out the non-negative A^c part, yields

$$(\mathbb{P}A)^\alpha \;\le\; (\mathbb{Q}A)^{\alpha-1} \mathbb{E}_{\mathbb{Q}}[(p(Z)/q(Z))^\alpha].$$

It is also seen as Holder's inequality applied to $\int q(p/q) 1_A$. Taking the α root produces the stated inequality.

Lemma 45: Let $\mathbb{P}_Z$ be the joint normal $N(0, \Sigma)$ distribution, with $\Sigma = I - bb^T$ where $\|b\|^2 = \nu < 1$. Likewise, let $\mathbb{Q}_Z$ be the distribution that makes the Z_j independent standard normal. Then the Rènyi divergence is bounded. Indeed, for all $1 \le \alpha \le \infty$,

$$D_\alpha(\mathbb{P}_Z\|\mathbb{Q}_Z) \;\le\; c_0.$$

where $c_0 = -(1/2) \log[1 - \nu]$. With $\nu = P/(\sigma^2 + P)$, this constant is $c_0 = (1/2) \log[1 + P/\sigma^2]$.

Proof of Lemma 45: Direct evaluation of the α divergence between $N(0, \Sigma)$ and $N(0, I)$ reveals the value

$$D_\alpha \;=\; -\frac{1}{2} \log |\Sigma| \;-\; \frac{1}{2(\alpha - 1)} \log |\alpha I - (\alpha - 1)\Sigma|$$

Expressing $\Sigma = I - \Delta$, it simplifies to

$$-\frac{1}{2} \log |I - \Delta| \;-\; \frac{1}{2(\alpha - 1)} \log |I + (\alpha - 1)\Delta|$$

The matrix Δ is equal to bb^T, with b as previously specified with $\|b\|^2 = \nu$. The two matrices $I - \Delta$ and $I + (\alpha - 1)\Delta$ each take the form $I + \gamma bb^T$, with γ equal to -1 and $(\alpha - 1)$ respectively.

The form $I + \gamma bb^T$ is readily seen to have one eigenvalue of $1 + \gamma\nu$ corresponding to an eigenvector $b/\|b\|$ and $L - 1$ eigenvalues equal to 1 corresponding to eigenvectors orthogonal to the vector b. The log determinant is the sum of the logs of the eigenvalues, and so, in the present context, the log determinants arise exclusively from the one eigenvalue not equal to 1. This provides evaluation of D_α to be

$$-\frac{1}{2} \log[1 - \nu] \;-\; \frac{1}{2(\alpha - 1)} \log[1 + (\alpha - 1)\nu],$$

where an upper bound is obtained by tossing the second term which is negative.

We see that $\max_Z p(Z)/q(Z)$ is finite and equals $[1/(1 - \nu)]^{1/2}$. Indeed, from the densities $N(0, I - bb^T)$ and $N(0, I)$ this claim can be established, noting after orthogonal transformation that these measures are only different in one variable, which is either $N(0, 1-\nu)$ or $N(0, 1)$, for which the maximum ratio of the densities occurs at the origin and is simply the ratio of the normalizing constants. This completes the proof of Lemma 45.

With $\nu = P/(\sigma^2+P)$ this limit $-(1/2)\log[1-\nu]$ which we have denoted as c_0 is the same as $(1/2)\log[1+P/\sigma^2]$. That it is the same as the capacity $\mathcal{C}$ appears to be coincidental, as we do not have any direct communication rate interpretation of the operation of taking the log of the L_∞ norm of the ratio of the densities that arise here.

Proof of Lemma 3: We are to show that for events A determined by $\mathcal{F}_k$ the probability $\mathbb{P}[A]$ is not more than $\mathbb{Q}[A]e^{kc_0}$. Write the probability as an iterated expectation conditioning on $\mathcal{F}_{k-1}$. That is, $\mathbb{P}[A] = \mathbb{E}\left[\mathbb{P}[A|\mathcal{F}_{k-1}]\right]$. To determine membership in A, conditional on $\mathcal{F}_{k-1}$, we only need $Z_{k,J_k} = (Z_{k,j} : j \in J_k)$ where J_k is determined by $\mathcal{F}_{k-1}$. Thus

$$\mathbb{P}[A] = \mathbb{E}_{\mathbb{P}}\left[\mathbb{P}_{\mathcal{X}^2_{n-k+1}, Z_{k,J_k}|\mathcal{F}_{k-1}}[A]\right],$$

where we use the subscript on the outer expectation to denote that it is with respect to $\mathbb{P}$ and the subscripts on the inner conditional probability to indicate the relevant variables. For this inner probability switch to the nearby measure $\mathbb{Q}_{\mathcal{X}_{n-k+1}, Z_{k,J_k}|\mathcal{F}_{k-1}}$. These conditional measures agree concerning the distribution of the independent $\mathcal{X}^2_{n-k+1}$, so the α relative entropy between them arises only from the normal distributions of the Z_{k,J_k} given $\mathcal{F}_{k-1}$. This α relative entropy is bounded by c_0.

To see this, recall that from Lemma 2 that $\mathbb{P}_{Z_{k,J_k}|\mathcal{F}_{k-1}}$ is $N_{J_k}(0, \Sigma_k)$ with $\Sigma_k = I - \delta_k\delta_k^T$. Now

$$||\delta_k||^2 = \nu_k \sum_{j \in sent \cap J_k} P_j/P$$

which is $(1 - (\hat{q}_1 + \ldots + \hat{q}_{k-1}))\nu_k$. Noting that $\nu_k = \hat{s}_k\nu$ and $\hat{s}_k(1 - (\hat{q}_1 + \ldots + \hat{q}_{k-1}))$ is at most 1, we get that $||\delta_k||^2 \leq \nu$. Thus from Lemma 45, for all $\alpha \geq 1$, the α relative entropy between $\mathbb{P}_{Z_{k,J_k}|\mathcal{F}_{k-1}}$ and the corresponding $\mathbb{Q}$ conditional distribution is at most c_0.

So with the switch of conditional distribution we obtain a bound with a multiplicative factor of e^{c_0}. The bound on the inner expectation is then a function of $\mathcal{F}_{k-1}$, so the conclusion follows by induction. This completes the proof of Lemma 3.

Appendix III
Proof of Lemmas on the Progress of $q_{1,k}$

Proof of Lemma 6: Consider any step k with $q_{1,k-1} - f_{1,k-1} \leq x^*$. We have that $x = q^{adj}_{1,k-1}$ is at least $\tilde{x} = q_{1,k-1} - f_{1,k-1}$, where these are initialized to be 0 when $k = 1$. Consider $q_{1,k} = g_L(x) - \eta_k$ which is at least $g_L(\tilde{x}) - \eta_k$, since the function g_L is increasing. By the gap property, it is at least $\tilde{x} + gap(\tilde{x}) - \eta_k$, which in turn is at least $q_{1,k-1} - \bar{f}(x) + gap(x) - \eta(x)$, which is at least $q_{1,k-1} + gap'$.

The increase $q_{1,k} - q_{1,k-1}$ is at least gap' each such step, so the number of of such steps $m-1$ is not more than $1/gap'$. At the final step m, the $\tilde{x} = q_{1,m-1} - f_{1,m-1}$ exceeds x^* so $q_{1,m}$ is at least $g_L(x^*) - \eta_m$ which is $1 - \delta^* - \eta_m$. This completes the proof of Lemma 6.

Proof of Lemma 5: With a constant gap bound, the claim when $f_{1,k} \leq \bar{f}$ follows from the above, specializing $\bar{f}$ and η to be constant. As for the claim when $f_{1,k} = kf$, it is actually covered by the case that $f_{1,k} \leq \bar{f}$, in view of the choice that $f \leq \bar{f}/m^*$. This completes the proof of Lemma 5.

Appendix IV
Accumulative g in the Large Code Limit

Our primary work concerns finite L and B regimes. Nevertheless, in providing intuition concerning the power allocation and the behavior of the update function, it is sensible to briefly discuss the matter via a limiting arguments for large B. The power allocation should be proportional to $e^{-2\mathcal{C}\ell/L}$ for the limit of the update function to be accumulative at rates up to capacity.

The fixed L and large B extreme: One justification of such power allocation aries from the setting in which the decoding can be done successively without any adaptation. This comes from consideration of fixed L and exponentially large $B = 2^{nR/L}$. With nR/L very large, practicality would be lost, but in theory it would permit reliable decoding one section at a time, by what is called rate-splitting and successive-decoding, as previously cited. With total sum rate R arranged to be near the capacity $\mathcal{C} = (1/2)\log(1 + P/\sigma^2)$, the choice of the power allocations proportional to $e^{-2\ell\mathcal{C}/L}$ simply arise as the choice that makes the incremental decoding capacities be $\mathcal{C}/L$, commensurate with the choice of sections of equal size near $2^{n\mathcal{C}/L}$.

To express what is meant by an incremental decoding capacity $\mathcal{C}_\ell$ for the decoding of section ℓ, for any power allocation, the power of the as yet undecoded sections $P_{\ell+1,L} = P_{(\ell+1)} + \ldots + P_{(L)}$ adds to the noise variance, to express $\mathcal{C}_\ell = (1/2)\log(1 + P_{(\ell)}/(\sigma^2 + P_{\ell+1,L}))$. These incremental capacities sum to the total capacity $(1/2)\log(1+P/\sigma^2)$, afterall, the incremental capacities are seen to be the difference of the values of $(1/2)\log(\sigma^2 + P_{\ell,L})$ at ℓ and $\ell+1$, so they sum to the difference of $(1/2)\log(\sigma^2 + P)$ and $(1/2)\log(\sigma^2)$.

The power allocation proportional to the mild exponential decay, $e^{-2\ell\mathcal{C}/L}$, is indeed the unique choice that makes these incremental capacities be equal across the sections.

Now since $\sum_\ell \mathcal{C}_\ell = \mathcal{C}$ for any power allocation with the specified total power, it follows that $\min_\ell \mathcal{C}_\ell \leq \mathcal{C}/L$, with strict inequality if the power allocation differs from the specified choice. Using the constant rate split R/L, reliability of successive decoding of the sections would require it to be not more than the minimum of the $\mathcal{C}_\ell$. So with successive decoding and a constant rate split, to permit R up to $\mathcal{C}$ requires that the power allocation be as specified.

It is instructive to deduce the update function for this case. In this extreme, the number of steps would be L and there would be a linear increase of $1/L$ in the (unweighted) fraction decoded each step. As for the weighted fraction, with $\pi_{(\ell)} = e^{-2\mathcal{C}(\ell-1)/L}(1 - e^{-2\mathcal{C}/L})/(1 - e^{-2\mathcal{C}})$, the π size of the previously decoded set $\{1, 2, \ldots, \ell-1\}$ is $x = (1 - e^{-2\mathcal{C}(\ell-1)/L})/(1 - e^{-2\mathcal{C}})$ which is increased to

$g_L(x) = x + \pi_{(\ell)}$. The expression for x may be reexpressed as $1/\nu - \pi_{(\ell)}/(1-e^{-2\mathcal{C}/L})$, so this $\pi_{(\ell)}$ may be expressed as $(1/\nu - x)(1-e^{-2\mathcal{C}/L})$. Consequently, in this extreme case of exponentially large B one has

$$g_L(x) = x + (1-x\nu)(1-e^{-2\mathcal{C}/L})/\nu.$$

It has positive gap $g_L(x) - x$ near $(1-x\nu)2\mathcal{C}/(\nu L)$. This expression shows how close the update function $g_L(x)$ is to x for the capacity achieving update function in the large B limit, and the power allocation that achieves it is identified.

In contrast to this fixed L, exponentially large B setting, with one section decoded each step, the adaptive decoder is built for large L, arranged to be within a log-factor of n, allowing more moderate section sizes B, to achieve a feasible size dictionary. Doing so entails less separation between the distribution of statistics for term selection, necessitating the adaptive selection, while retaining exponentially small error probability (now in L rather than in n). With the number of steps m of order $\log B$ which remains much smaller than L, the update function $g_L(x)$ is seen to take a similar form, but with a gap of order $1/m$ rather than $1/L$.

Integral characterization: As we have seen for specific power allocations, an integral approximation to $g_L(x)$ may be used. This may be used for a calculus examination of conditions for accumulation in the large code limit. Consider power allocations sorted to be decreasing in ℓ/L. In particular, suppose the power allocation weights $\pi_{(\ell)} = P_{(\ell)}/P$ are expressible as proportional to a decreasing differentiable function $u(t)$ for t in $[0,1]$, evaluated at $t = \ell/L$. That is

$$\pi_{(\ell)} = (1/L)u(t)/U_L$$

where the normalization constant $U_L = (1/L)\sum_{\ell=1}^{L} u(\ell/L)$ is approximated by the integral $U(1) = \int_0^1 u(t)dt$.

Ignoring the effects of the small a and small h, the $g_L(x)$ becomes

$$\sum_{\ell} \pi_\ell \, \Phi\left((\sqrt{\frac{L\pi_\ell \nu/2R}{1-x\nu}} - 1)\tau \right)$$

where, with $t = \ell/L$, the normal probabilities in this sum become

$$\Phi\left((\sqrt{\frac{u(t)\nu/[2RU(1)]}{1-x\nu}} - 1)\tau \right).$$

The sum is approximated by the corresponding integral of these with respect to the density $u(t)/U(1)$ on $[0,1]$ of the power allocation measure π with cumulative distribution $U(t)/U(1)$ where $U(t) = \int_0^t u(\tilde{t})d\tilde{t}$.

Some analysis for large τ proceeds as follows. These probabilities at t approaches 1 or 0, respectively, according to whether $u(t)\nu/[2RU(1)]$ is greater than or less than $1 - x\nu$, which means that it is 1 for t less than a value t_x at which

$$u(t_x)\nu/[2RU(1)] = 1 - x\nu.$$

This large τ limit is thus called the saturated case, in which the detection probabilities are 1 or 0, with the cut-off between 1 and 0 occurring at the point t_x.

This t_x is increasing in x in an interval $[0, x_1]$ contained in $[0,1]$, where x_1 is the point where $t_x = 1$. To get started with t_0 non-negative, it is required that $u(0)\nu/[2U(1)]$ be at least R. To end with the value $t_x = 1$ at $x_1 \le 1$, it is required that $u(1)\nu/[(1-\nu)2U(1)]$ be at least R.

In accordance with this 1 and 0 characterization of the probabilities in the large τ and L limit, the $g_L(x)$ approaches the π measure of the interval up to t_x, that is, the update function becomes

$$g(x) = U(t_x)/U(1)$$

on $[0, x_1]$.

It is only in this section of the appendix that we use $g(x)$ to denote this large τ and large L limit. In the rest of the paper $g(x)$ denotes the integral approximation in which explicit dependence on τ is retained.

From the equations characterizing t_x and $g(x)$ here, their derivatives with respect to x satisfy $t'_x u'(t_x) = -2RU(1)$ and $g'(x) = t'_x u(t_x)/U(1)$, so that $u'(t_x)/u(t_x) = -2R/g'(x)$ where $t'_x = dt_x/dx$.

In particular, for the asymptotic update function $g(x)$ to track x requires that $g'(x) = 1$ and consequently $u'(t)/u(t) = -2R$, which means that $u(t)$ is proportional to e^{-2Rt}. With R approaching $\mathcal{C}$ this gives additional motivation for the choice of variable power assignment with $u(t) = e^{-2\mathcal{C}t}$. It has $U(1) = (1-e^{-2\mathcal{C}})/(2\mathcal{C})$ equal to $\nu/(2\mathcal{C})$, with which $u(0)\nu/[2U(1)] = \mathcal{C}$ is indeed at least R.

An interesting aspect of this saturated setting is a rejuvenation property analogous to that already seen in the above discussion in the finite L case. Namely, for any $0 < a < 1$, once the π measure of the set decoded hits $x = a$ corresponding to a portion decoded of t_a, what remains is an analogous problem decoding a portion $1 - t_a$ with the remaining portion of the power $(1-a)P$.

To attempt to make $g(x)$ track a higher trajectory equal to $(R/\gamma)x$ for an initial interval of values of x, for some positive $\gamma < R$, we could set $u'(t)/u(t) = -2\gamma$ as satisfied by the power assignment with $u(t)$ proportional to $e^{-2\gamma t}$. Denote $\mathcal{C}_\gamma = \gamma(1-e^{-2\mathcal{C}})/(1-e^{-2\gamma})$ which is the value of $u(0)\nu/[2U(1)]$ in this setting. It is between γ and $\mathcal{C}$. The initialization requirement that $u(0)\nu/[2U(1)]$ be at least R becomes the limitation $R \le \mathcal{C}_\gamma$. So such $\gamma < \mathcal{C}$ keeps the rate less than capacity.

It is noted if γ is near $\mathcal{C}$ reasonable rates can be achieved by straddling R between γ and $\mathcal{C}_\gamma$.

Presumably, as holds true in the finite L setting, in this limiting L setting, $u(t)$ proportional to $e^{-2\mathcal{C}t}$ is the unique choice of power allocation density for which the $g(x)$ is accumulative for rates up to $\mathcal{C}$. We shall not demonstrate that here, but do take note that if the differentiable function $g(x) \ge x$ does not match x throughout $[0,1]$, then there is a first $x_0 \ge 0$ at which $g'(x_0) > 1$, from which there is a small interval of values of x starting at x_0 and corresponding small interval of values of t within which by Taylor expansion $u(t)$ is near $u(t_{x_0})e^{-2\gamma t}$ with a $\gamma < \mathcal{C}$ and $g(x)$ is strictly greater than x. Using the rejuvenation property it should then be possible to split the analysis into intervals in which the use of such power allocation restricts the achievable rate to less than capacity.

For our finite L and B demonstration of the accumulative property of the proposed power allocations, the analysis is similar. However, instead of a sharp transition from 1 to 0 for the detection probabilities, there is quantified by $\Phi(\mu_j(x) - \tau)$ a more gradual transition extending over a significant portion of the sections. This gradual detection quantification provids the overlap associated with building up the decoding adaptively.

Appendix V
The gap has not more than one oscillation

Proof of Lemma 21: In the same manner as the derivative result for $g_{num}(x)$, the $g_{low}(x)$ has derivative with respect to x given by the following function, evaluated at $z = z_x$,

$$\left\{\frac{\tau\Delta_c}{2}\left(1+\frac{z}{\tau}\right)^3\phi(z) \,+\, \int_z^\infty \left(1+t/\tau\right)^2\phi(t)dt\right\}\frac{R}{\mathcal{C}'}.$$

Subtracting $1{+}D(\delta_c)/snr$ from it gives the function $der(z)$, which at $z = z_x$ is the derivative with respect to x of $G(z_x) = g_{low}(x) - x - xD(\delta_c)/snr$. The mapping from x to z_x is strictly increasing, so the sign of $der(z)$ provides the direction of movement of either $G(z)$ or of $G(z_x)$.

Consider the behavior of $der(z)$ for $z \geq -\tau$ which includes $[z_0, z_1]$. At $z = -\tau$ the first term vanishes and the integral is not more than $1 + 1/\tau^2$, so under the stated condition on R, the $der(z)$ starts out negative at $z = -\tau$. Likewise note that $der(z)$ is ultimately negative for large z since it approaches $-(1{+}D(\delta_c)/snr)$. Let's see whether $der(z)$ goes up anywhere to the right of $-\tau$. Taking its derivative with respect to z, we obtain

$$der'(z) \,=\, \Big\{ -\frac{\tau\Delta_c}{2}\big(1{+}z/\tau\big)^3 z\phi(z) \,+\, \frac{3\Delta_c}{2}\big(1{+}z/\tau\big)^2\phi(z) \\ -\big(1{+}z/\tau\big)^2\phi(z)\Big\}\,\frac{R}{\mathcal{C}'}.$$

The interpretation of $der'(z)$ is that since $der(z_x)$ is the first derivative of $G(z_x)$, it follows that $z_x'\, der'(z_x)$ is the second derivative, where z_x' as determined in the proof of Corollary 13 is strictly positive for $z > -\tau$. Thus the sign of the second derivative of the lower bound on the gap is determined by the sign of $der'(z)$.

Factoring out the positive $(1{+}z/\tau)^2\phi(z)R/\mathcal{C}'$ for $z > -\tau$, the sign of $der'(z)$ is determined by the quadratic expression

$$-(\tau\Delta_c/2)\big(1{+}z/\tau\big)z \,+\, 3\Delta_c/2 \,-\, 1,$$

which has value $3\Delta_c/2 - 1$ at $z = -\tau$ and at $z = 0$. The discriminant of whether there are any roots to this quadratic yielding $der'(z) = 0$ is given by $(\tau\Delta_c)^2/4 {-} 2\Delta_c(1{-}3\Delta_c/2)$. Its positivity is determined by whether $\tau^2\Delta_c/4 > 2 - 3\Delta_c$, that is, whether $\Delta_c > 2/(\tau^2/4 + 3)$. If $\Delta_c \leq 2/(\tau^2/4 + 3)$ which is less than $2/3$, then $der'(z)$, which in that case starts out negative at $z = -\tau$, never hits 0, so it stays negative for $z \geq -\tau$, so $der(z)$ never goes up to the right of $-\tau$ and $G(z)$ remains a decreasing function. In that decreasing case we may take $z_G = z_{max} = -\tau$.

If $\Delta_c > 2/(\tau^2/4 + 3)$, then by the quadratic formula there is an interval of values of z between the pair of points $-\tau/2 \pm \sqrt{\tau^2/4 - (2/\Delta_c)(1 - 3\Delta/2)}$ within which $der'(z)$ is positive, and within the associated interval of values of x the $G(z_x)$ is convex in x. Outside of that interval we have concavity of $G(z_x)$. So then either $der(z)$ remains negative, so that $G(z)$ is decreasing for $z \geq -\tau$, or there is a root $z_{crit} > -\tau$ where $der(z)$ first hits 0 and $der'(z) > 0$, i.e. that root, if there is one, is in this interval. Suppose there is such a root. Then from the behavior of $der'(z)$ as a positive multiple of a quadratic with two zero crossings, the function $G(z)$ experiences an oscillation.

Indeed, until that point z_{crit}, the $der(z)$ is negative so $G(z)$ is decreasing. After that root, the $der(z)$ is increasing between z_{crit} and z_{right}, the right end of the above interval, so $der(z)$ is positive and $G(z)$ is increasing between those points as well. Now consider $z \geq z_{right}$, where $der'(z) \leq 0$, strictly so for $z > z_{right}$. At z_{right} the $der(z)$ is strictly positive (in fact maximal) and ultimately for large z the $der(z)$ is negative, so for $z > z_{right}$ the $G(z)$ rises further until a point $z = z_{max}$ where $der(z) = 0$. To the right of that point since $der'(z) < 0$, the $der(z)$ stays negative and $G(z)$ is decreasing. Thus $der(z)$ is identified as having two roots z_{crit} and z_{max}, and $G(z)$ is unimodal to the right of z_{crit}.

To determine the value of $der(z)$ at $z = 0$, evaluate the integral $\int_z^\infty (1{+}t/\nu)^2\pi(t)dt$. In the same manner as in the preceding subsection, it is $(1{+}1/\tau^2)\bar{\Phi}(z) + (2\tau{+}z)\phi(z)/\tau^2$. Thus $der(z)$ is

$$\frac{R}{\mathcal{C}'}\left\{\frac{\tau\Delta_c}{2}\left(1{+}\frac{z}{\tau}\right)^3\phi(z) + \frac{2\tau{+}z}{\tau^2}\phi(z) + \big(1 + \frac{1}{\tau^2}\big)\bar{\Phi}(z)\right\} \\ -\left(1 + \frac{D(\delta_c)}{snr}\right).$$

At $z = 0$ it is

$$\frac{R}{\mathcal{C}'}\left\{\left(\frac{\tau\Delta_c}{2} + \frac{2}{\tau}\right)\frac{1}{\sqrt{2\pi}} + \left(1 + \frac{1}{\tau^2}\right)/2\right\} - \big[1{+}D(\delta_c)/snr\big].$$

It is non-negative if $\tau\Delta_c/(2\sqrt{2\pi}\,)$ exceeds

$$\big[1 + D(\delta_c)/snr\big]\,\mathcal{C}'/R \,-\, (1{+}1/\tau^2)/2 \,-\, 2/(\tau\sqrt{2\pi}\,)$$

which using $\mathcal{C}'/R = 1{+}r/\tau^2$ is

$$1/2 + \frac{(r{-}1/2)}{\tau^2} + \frac{D(\delta_c)}{snr}(1{+}r/\tau^2) - 2/(\tau\sqrt{2\pi}\,).$$

It is this expression which we call *half* for it tends to be not much more than 1/2. For instance, if $D(\delta_c)/snr \leq 1/2$ and $(3/2)r \leq (2/\sqrt{2\pi})\tau$, then this expression is not more than $1 - 1/(2\tau^2)$ which is less than 1.

So then $der(z)$ is non-negative at $z = 0$ if

$$\Delta_c \geq \frac{2\sqrt{2\pi}\,\textit{half}}{\tau}.$$

Non-negativity of $der(0)$ implies that the critical value of the function G satisfies $z_G \leq 0$.

Suppose on the other hand that $der(0) < 0$. Then $\Delta_c < 2\sqrt{2\pi}\,\textit{half}/\tau$, which is less than $2/3$ when τ is at least $3\sqrt{2\pi}\,\textit{half}$. Using the condition $\Delta_c \leq 2/3$, the $der'(z) < 0$ for $z > 0$. It follows that $G(z)$ is decreasing for $z > 0$, and both z_G and z_{max} are non-positive.

Next consider the behavior of the function $A(z)$, for which we show that it too has at most one oscillation. Differentiating and collecting terms obtain that $A'(z)$ is

$$A'(z) = -2(1-\Delta_c)(z{+}\tau)\Phi(z) + \Delta_c(z{+}\tau)^2\phi(z).$$

Consider values of z in $I_\tau = (-\tau, \infty)$ to the right of $-\tau$. Factoring out $2(z{+}\tau)$, the sign behavior of $A'(z)$ is determined by the function

$$M(z) \,=\, -\,(1{-}\Delta_c)\,\Phi(z) \,+\, (\Delta_c/2)\,(z{+}\tau)\,\phi(z)\,.$$

This function $M(z)$ is negative for large z as it converges to $-2(1{-}\Delta_c)$. Thus $A(z)$ is decreasing for large z. At $z = -\tau$ the sign of $M(z)$ is determined by whether $\Delta_c{<}1$, if so then $M(z)$ starts out negative, so then $A(z)$ is initially decreasing, whereas in the unusual case of $\Delta_c \geq 1$, the $A(z)$ is initially increasing and we set $z_A = -\tau$. Consider the derivative of $M(z)$ given by

$$M'(z) = -\big[1{-}3\Delta_c/2 \,+\, (\Delta_c/2)\,z(z{+}\tau)\big]\,\phi(z).$$

The expression in brackets is the same quadratic function of z considered above. It is centered and extremal at $z_{cent} = -\tau/2$. This quadratic attains the value 0 only if Δ_c is at least $\Delta_c^* = 2/(\tau^2/4+3)$.

For $\Delta_c{<}\Delta_c^*$, which is less than 1, the $M'(z)$ stays negative and consequently $M(z)$ is decreasing, so $M(z)$ and $A'(z)$ remains negative for $z > -\tau$. Then $A(z)$ is decreasing in I_τ (which actually implies the monotonicity of $G(z)$ under the same condition on Δ_c).

For $\Delta_c \geq \Delta_c^*$, for which the function $M'(z)$ does cross 0, this $M'(z)$ is positive in the interval of values of z centered at $z_{cent} = -\tau/2$ and heading up to the point z_{right} previously discussed. In this interval including $[-\tau/2, z_{right}]$ the function $M(z)$ is increasing.

Let's see whether $M(z)$ is positive, at or to the left of z_{cent}. For $\Delta_c > 1$ that positivity already occurred at and just to the right of $-\tau$. For $\Delta_c \leq 1$, use the inequality $\Phi(z) \leq \phi(z)/(-z)$ for $z < 0$. This lower bound is sufficient to demonstrate positivity in an interval of values of z centered at the same point $z_{cent} = -\tau/2$, provided $\Delta_c\tau^2/4$ is at least $2(1{-}\Delta_c)$, that is, Δ_c at least $\Delta_c^{**} = 2/(\tau^2/4+2)$. Then z_A is not more than the left end of this interval, which is less than $-\tau/2$. For $\Delta_c \geq \Delta_c^{**}$, this interval is where the same quadratic $z(z{+}\tau)$ is less than $-2(1{-}\Delta_c)/\Delta_c$. Then the $M(z)$ is positive at $-\tau/2$ and furthermore increasing from there up to z_{right}, while, further to the right it is decreasing and ultimately negative. It follows that such $M(z)$ has only one root to the right of $-\tau/2$. The $A'(z)$ inherits the same sign and root characteristics as $M(z)$, so $A(z)$ is unimodal to the right of $-\tau/2$.

If Δ_c is between Δ_c^* and Δ_c^{**}, the lower bound we have invoked is insufficient to determine the precise conditions of positivity of $M(z)$ at z_{cent}, so we resort in this case to the milder conclusion, from the negativity of $M'(z)$ to the right of z_{right}, that $M(z)$ is decreasing there and hence it and $A'(z)$ has at most one root to the right of that point, so $A(z)$ is unimodal there. Being less than Δ_c^{**}, the value of Δ_c is small enough that $2/\Delta_c > \tau^2/4+2$, and hence z_{right} is not more than $[-\tau+\sqrt{4}]/2$ which is $-\tau/2+1$.

This completes the proof of Lemma 21.

We remark concerning $G(z)$ that one can pin down down the location of z_G further. Under conditions on Δ_c, it is near to and not more that a value near

$$-\sqrt{2\log\left(\frac{1}{2\pi}\,\frac{\tau\Delta_c/2}{D(\delta_c)/snr{+}(r{-}1)/\tau^2}\right)},$$

provided the argument of the logarithm is of a sufficient size. As we have said, precise knowledge of the value of z_G is not essential because the shape properties allow us to take advantage of the tight lower bounds on $A(z)$ for negative z.

Appendix VI
The gap in the constant power case

Proof of Corollary 19. We are to show under the stated conditions that $g(x) - x$ is smallest in $[0, x^*]$ at $x = x^*$, when the power allocation is constant. For x in $[0,1]$ the function z_x is one to one. In this $u_{cut} = 1$ case, it is equal to $z_x = [\sqrt{(1{+}r/\tau^2)/(1-x\nu)} - 1]\tau$. It starts at $x = 0$ with z_0 and at $x = x^*$ it is ζ. Note that $(1{+}z_0/\tau)^2 = 1{+}r/\tau^2$. If $r \geq 0$ the $z_0 \geq 0$, while, in any case, for $r > -\tau^2$ the z_0 at least exceeds $-\tau$. Invert the formula for $z = z_x$ to express x in terms of z. Using $g(x) = \Phi(z)$ and subtracting the expression for x, we want the minimum of the function

$$G(z) = \Phi(z) - \frac{1}{\nu}\left(1 - \frac{(1{+}r/\tau^2)}{(1{+}z/\tau)^2}\right).$$

Its value at z_0 is $G(z_0) = \Phi(z_0)$. Consider the minimization of $G(z)$ for $z_0 \leq z \leq \zeta$, but take advantage, when it is helpful, of properties for all $z > -\tau$. The first derivative is

$$\phi(z) - \frac{2}{\nu\,\tau}\frac{(1{+}r/\tau^2)}{(1{+}z/\tau)^3},$$

ultimately negative for very large z. This function has 0, 1, or 2 roots to the right of $-\tau$. Indeed, to be zero it means that z solves

$$z^2 - 6\log(1{+}z/\tau) = 2\log(\nu\tau/c)$$

where $c = 2(1+r/\tau^2)\sqrt{2\pi}$. The function on the left side $v(z) = z^2 - 6\log(1{+}z/\tau)$ is convex, with a value of 0 and a negative slope at $z = 0$ and it grows without bound for large z. This function reaches its minimum value (lets call it $val < 0$) at a point $z = z_{crit} > 0$, which solves $2z - 6/(\tau+z) = 0$, given by $z_{crit} = (\tau/2)\big[\sqrt{1{+}12/\tau^2} - 1\big]$ not more than $3/\tau$.

When $val > 2\log(\nu\,\tau/c)$ there are no roots, so $G(z)$ is decreasing for $z > -\tau$ and has its minimum on $[0,\zeta]$ at $z = \zeta$.

When $2\log(\nu\tau/c)$ is positive (that is, when $\nu\tau > c$, which is the condition stated in the corollary), it exceeds the value of the expression on the left at $z = 0$, and G is increasing there. So from the indicated shape of the function $v(z)$, there is one root to the right of 0, which must be a maximizer of $G(z)$, since $G(z)$ is eventually decreasing. So then $G(z)$ is unimodal for positive z and so if $z_0 \geq 0$ its minimum in $[z_0, \zeta]$ is at either $z = z_0$ or $z = \zeta$ and this minimum is at least $\min\{G(0), G(\zeta)\}$. The value at $z = 0$ is $G(0) = 1/2$. So, with $(r - r_{up})/[snr(\tau^2{+}r_1)]$ less than $1/2$, the minimum for

$z \geq 0$ occurs at $z = \zeta$, which demonstrates the first conclusion of Corollary 19.

If r is negative then $z_0 < 0$, and we consider the shape of $G(z)$ for negative z. Again with the assumption that $2\log(\nu\tau/c)$ is positive, the function $G(z)$ for $z \geq -\tau$ is seen to have a minimizer at a negative $z = z_{min}$ solving $z^2 = 2\log(\nu\tau/c) - 6\log(1{+}z/\tau)$, where $G'(z) = 0$, and $G(z)$ is increasing between z_{min} and 0. We inquire as to whether $G(z)$ is increasing at z_0. If it is, then $z_0 \geq z_{min}$ and $G(z)$ is unimodal to the right of z_0. The value of the derivative there is $\phi(z_0) - \frac{2}{\nu\,(\tau+z_0)}$, which is positive if

$$|z_0| \leq \sqrt{2\log\left(\nu(\tau{+}z_0)/2\sqrt{2\pi}\right)}.$$

As we shall see momentarily, z_0 is between r/τ and $r/2\tau$, so this positive derivative condition is implied by

$$r/\tau \geq -\sqrt{2\log\left(\nu(\tau{+}r/\tau)/2\sqrt{2\pi}\right)}.$$

Then $G(z)$ is unimodal to the right of z_0 and has minimum equal to $\min\{G(z_0), G(\zeta)\}$.

From the relationship $(1{+}z_0/\tau)^2 = 1{+}r/\tau^2$, with $-\tau < z_0 \leq 0$, one finds that $r = z_0(2\tau + z_0)$, so it follows that $z_0 = r/(2\tau + z_0)$ is between r/τ and $r/2\tau$.

Lower bound $G(z_0) = \Phi(z_0)$ for $z_0 \leq 0$ by the tangent line $(1/2){+}z_0\phi(0)$, which is at least $(1/2){+}r/(\tau\sqrt{2\pi})$. Thus when r is such that the positive derivative condition holds, we have the gap lower bound allowing $r_{up} < r \leq 0$ which is

$$\min\left\{1/2 + r/(\tau\sqrt{2\pi})\,,\,(r - r_{up})/[snr\,(\tau^2{+}r_1)]\right\}.$$

This completes the proof of Corollary 19.

Next we ask whether a useful bound might be available if $G(z)$ is not increasing at this $z_0 \leq 0$. Then $z_0 \leq z_{min}$, and the minimum of $G(z)$ in $[z_0, \zeta]$ is either at z_{min} or at ζ. The $G(z)$ is

$$\Phi(z) + \frac{(1{+}z_0/\tau)^2 - (1{+}z/\tau)^2}{(1{+}z/\tau)^2}.$$

Now since z_{min} is the negative solution to $z^2 = 2\log(\nu\tau/c) - 6\log(1{+}z/\tau)$, it follows that there z_{min} is near $-\sqrt{2\log(\nu\tau/c)}$. From the difference of squares, the second part of $G(z_{min})$ is near $2(z_0 - z_{min})/\tau$ which is negative. So for $G(z_{min})$ to be positive the $\Phi(z_{min})$ would need to overcome that term. Now $\Phi(z_{min})$ is near $\phi(z_{min})/|z_{min}|$, and $G'(z) = 0$ at z_{min} means that $\phi(z_{min})$ equals the value $(2/\nu\,\tau)(1+z_0/\tau)^2/(1+z_{min}/\tau)^3$. Accordingly, $G(z_{min})$ is near

$$\frac{2(1{+}z_0/\tau)^2}{\nu\,\tau\sqrt{2\log(\nu\,\tau/c)}} + \frac{2(z_0 + \sqrt{2\log(\nu\tau/c)})}{\tau}.$$

The implication is that by choice of r one can not push z_0 much to the left of $-\sqrt{2\log(\nu\tau/c)}$ without losing positivity of $G(z)$.

Next we examine when r_{up} is negative, whether r arbitrarily close to r_{up} can satisfy the conditions. That would require the r_{up}/τ to be greater than $-\sqrt{2\pi}/2$ and greater than

$$-\sqrt{2\log\left(\nu\tau(1 + r_{up}/\tau^2)/2\sqrt{2\pi}\right)}.$$

However, in view of the formula for r_{up}, it is near $[1/(1{+}snr) - 1]\tau^2 = -\nu\tau^2$ when $snr\,\bar{\Phi}(\zeta)$ and ζ/τ are small. Consequently, r_{up}/τ is near $-\nu\tau$. So if $\nu\tau$ is greater than a constant near $\sqrt{2\pi}/2$ then the first of these conditions on r_{up}/τ is not satisfied. Also with this r_{up}/τ near $-\nu\tau$ the argument of the logarithm becomes $\nu(1{-}\nu)\tau/2\sqrt{2\pi}$, needed to be greater than 1. So if $\nu\tau$ is less than a constant near $\sqrt{2\pi}/2$ then this argument of the logarithm is strictly less than 1. Thus the conditions for allowance of such negative r so close to r_{up} are vacuous. It is not possible to use an r so close to r_{up} when it is negative.

If when r_{up}/τ is negative, near $-\nu\tau$, we try instead to have $r/\tau = -\alpha\sqrt{2\pi}/2$ with $0 \leq \alpha < 1$, then the first expression in the minimum becomes $(1 - \alpha)/2$, the second expression becomes $r - r_{up}/[\nu(\tau{+}\zeta)^2]$ near $1 + r/[\nu\,\tau^2]$ equal to $1 - \alpha\sqrt{2\pi}/(2\nu\tau)$, and the additional condition becomes

$$\alpha\sqrt{2\pi}/2 \leq \sqrt{2\log\left(\nu(\frac{\tau}{2\sqrt{2\pi}} - \alpha/4)\right)}.$$

Which is acceptable with $\nu\tau$ at least a little more than $2\sqrt{2\pi}e^{\pi/4}$. So in this way the $1{+}r/\tau^2$ factor becomes at best near $1\sqrt{2\pi}/2\tau$. That is indeed a nice improvement factor in the rate, though not as ambitious as the unobtainable $1{+}r_{up}/\tau^2$ near $1 - \nu$.

A particular negative r of interest would be one that makes $(1{+}D(snr)/snr)(1{+}r/\tau^2) = 1$, for then even with constant power it would provide no rate drop from capacity. With this choice $1{+}r/\tau^2 = 1/(1{+}D(snr)/snr)$, the

$$r/\tau = \frac{-\tau D(snr)/snr}{1 + D(snr)/snr}.$$

That a multiple of $-\tau$, where the multiple is near $snr/2$ when snr is small. For $G(z)$ to be increasing at the corresponding z_0, we would want the magnitude $-r/\tau$ to be less than $\sqrt{2\log\left(\nu, \tau(1{+}r/\tau^2)/2\sqrt{2\pi}\right)}$, where the $\nu(1 + r/tau^2)/2$ may be expressed as a function of snr, and is also near $snr/2$ when snr is small. But that would mean that $b = \tau snr/2$ is a value where $b^2 \leq 2\log(b/\sqrt{2\pi}$, which a little calculus shows is not possible. Likewise, the above development of the case that z_0 is to the left of z_{min}, shows that we can not allow $-r/\tau$ to be much greater than the same value.

Omit these next three paragraphs which say more about the positive r case. To drop the stated requirement on $2\log(\tau/c)$ consider the remaining (perhaps unusual) case that $0 > 2\log(\tau/c) > val$. Then there are two positive roots, one to the left of z_{crit} and one to the right of z_{crit}. The root to the left of z_{crit} provides a candidate minimizer of $G(z)$, while to the right of that point the function is unimodal. So if z_0 is at least z_{crit}, for which it would be sufficient that r be at least 6, then the unimodality in $[z_0, \zeta]$ remains, with the gap minimum at $z = \zeta$, if also r is not more than $r_{up} + snr(\tau^2 + r_1)/2$ as assumed.

In this case that $0 > 2\log(\tau/c) > val$, if $z_0 < z_{crit}$, the minimum $G(z)$ in $[z_0, \zeta]$ occurs at either $z = z_{crit}$ or at $z = \zeta$. Then for the gap we have $\min\{G(z_{crit}), G(\zeta)\}$ at least

$$\min\left\{\Phi(z_{crit}) + \frac{1}{\nu}\left[\frac{(1 + r/\tau^2)}{(1 + z_{crit}/\tau)^2} - 1\right], \frac{r - r_{up}}{snr(\tau^2 + r_1)}\right\},$$

from which, recognizing that both expressions are linear in r, one can solve a linear equation for the maximum $r-r_{up}$ for which the expression on the right is the minimizer. Alternatively, for a simpler expression of a sufficient condition, note that the $1+r/\tau^2$ may be written $(1+z_0/\tau)^2$. The quantity in brackets is negative for $0 \le z_0 < z_{crit}$, but at least $-2z_{crit}/\tau$. So using $z_{crit} \le 3/\tau$, the gap in this case is at least

$$\min\left\{\Phi(3/\tau) - \frac{6}{\nu\,\tau^2}\,,\, \frac{r-r_{up}}{snr(\tau^2+r_1)}\right\}.$$

The lower bound on $G(z_{crit})$ in this case is near $\Phi(0) = 1/2$ for large τ^2 and it is positive when $\nu\,\tau^2 > 12$. So then the gap is at least $(r-r_{up})/[snr(\tau^2+r_1)]$ for $r-r_{up}$ not more than $snr(\tau^2+r_1)\big[1/2 - 6/(\nu\,\tau^2)\big]$.

Comparing the conditions $\tau > c = 2(1+r/\tau^2)\sqrt{2\pi}/\nu$ and $\nu\,\tau^2 > 12$, using $\nu = snr/(1+snr)$, one sees that the former requires snr to exceed an expression of order $1/\tau$ whereas the latter allows snr of order $1/\tau^2$. Moreover, the latter condition is less restrictive when $\tau > 6/\sqrt{2\pi}$. Also, if $r \ge 6$ then neither of those conditions are required and snr may be arbitrarily small.

Appendix VII
The r^*_{crit} bound based on a quadratic equation for r_1

Proof of Corollary 30: Recall that $D(\delta_c) \le \delta_c^2/2$ and $1+\delta_c = (1+\zeta/\tau)^2/(1+r_1/\tau^2)$, so with $\zeta' = \zeta(1+\zeta/(2\tau))$, we have

$$\delta_c = \frac{2\big(\zeta' - r_1/(2\tau)\big)}{\tau(1+r_1/\tau^2)},$$

so that $(\tau^2+r_1)D(\delta_c)$ is not more than

$$\frac{2\big(\zeta'-r_1/2\tau\big)^2}{1+r_1/\tau^2}.$$

Thus for $r^*_{crit}-\epsilon$ we investigate the optimization of the bound

$$\max\Big\{\frac{2\big(\zeta'-r_1/2\tau\big)^2}{1+r_1/\tau^2}+1\,,$$

$$r_1\Phi(\zeta)+(2\tau{+}\zeta)\phi(\zeta)+\Big[\frac{2\big(\zeta'-r_1/2\tau\big)^2}{1+r_1/\tau^2}+1\Big]\bar\Phi(\zeta)\Big\}.$$

Plugging in $r_1 = 0$ produces a bound for that case. To improve it we allow non-zero r_1. First let's analyze the requirement for its optimization. Observe that $2\big(\zeta'{-}r_1/2\tau\big)^2/(1{+}r_1/\tau^2)$ is decreasing as a function of r_1 between $-\tau^2$ and $2\tau\zeta'$. Its derivative with respect to r_1 is equal to $-(2/\tau)\big(\zeta' - r_1/2\tau\big)/(1{+}r_1/\tau^2)$ plus $-(1/\tau^2)2\big(\zeta'{-}r_1/2\tau\big)^2/(1{+}r_1/\tau^2)^2$, which is indeed negative, together equal to $-\delta_c[1+1/(\tau(1+r_1/\tau^2)^2)]$. Also, the second expression in the maximum is the increasing function $r_1\Phi(\zeta)+(2\tau+\zeta)\phi(\zeta)$ plus the additional amount less than the first expression in the max by the factor $\bar\Phi(\zeta) < 1$. Its derivative is $\Phi(\zeta)$ plus $\bar\Phi(\zeta)$ times the derivative of the first expression. Accordingly, in the same manner as the proof of the previous lemma, the minimum of the r^*_{crit} bound is either at the r_1 providing a match or at the point at which the derivative of the second expression is 0. Here we analyze the bound obtained by choosing the value of r_1 which make the two expressions equal. These derivative properties reveal there will be at most one root between $-\tau^2$ and $2\tau\zeta'$.

Equating the two expressions, grouping together like terms using $\bar\Phi(\zeta) = 1-\Phi(\zeta)$, and then dividing through by $\Phi(\zeta)$, the equation may be simplified to

$$\frac{2(\zeta'-r_1/2\tau)^2}{1+r_1/\tau^2} = r_1 - 1 + (2\tau{+}\zeta)\phi(\zeta)/\Phi(\zeta).$$

Evaluating either side of this equation at the solution r_1 produces our $r^*_{crit}-1-\epsilon$ bound. The symbol ψ is used to abbreviate the expression $(2\tau{+}\zeta)\phi(\zeta)/\Phi(\zeta)$. Positivity of the expression on the right requires $r_1 \ge -(\psi{-}1)$. As for positivity of the expression on the left, the denominator $1{+}r_1/\tau^2$ will then be at least $1{-}(\psi{-}1)/\tau^2$ which is at least $1{+}1/\tau^2{-}1.6/\tau$, near 1 for reasonable size τ. In any case, this lower bound is $\tau + 1/\tau - 1.6$ dividing by τ, which is positive since $\tau+1/\tau$ stays at least 2.

Multiply through by $1+r_1/\tau^2$, expand the square, and simplify. This resulting equation is seen to be quadratic in r_1 and may be written

$$\big[1{+}\gamma\big]\,r_1 = \big[2(\zeta')^2{+}1{-}\psi\big] - r_1^2/(2\tau^2),$$

where $\gamma = 2\zeta'/\tau + (\psi{-}1)/\tau^2$. One may use the quadratic formula to provide the solution

$$r_1 = \tau^2\left[\sqrt{(1{+}\gamma)^2+(2/\tau^2)\mathit{diff}} - (1{+}\gamma)\right].$$

where

$$\mathit{diff} = 2(\zeta')^2{-}(\psi{-}1).$$

The expression in the square root may be regrouped to show that it equals $(1{+}2\zeta'/\tau)^2+\gamma^2$, exceeding $(1{+}\zeta'/\tau)^2$. This shows that r_1 exceeds $\tau^2\big[(1{+}2\zeta'/\tau)-(1{+}\gamma)\big]$ which equals $-(\psi{-}1)$, in accordance with the requirement of positivity of $r_1+\psi-1$, and thence of $1+r_1/\tau^2$, thereby verifying that among the two roots to the quadratic we have identified the appropriate one. At this optimal r_1 the minimized r^*_{crit} is

$$r^*_{crit} = r_1 + \psi + \epsilon.$$

Write the solution for r_1 as

$$r_1 = \tau^2(1{+}\gamma)\left[\sqrt{1+(2/\tau^2)\mathit{diff}/(1{+}\gamma)^2} - 1\right].$$

Which shows that r_1 has sign matching the sign of *diff*. Using $\sqrt{1{+}2d} \le 1{+}d$ the formula yields the upper bound

$$r_1 \le r_1^* = \mathit{diff}/(1{+}\gamma).$$

This bound also arises from the quadratic equation preceeding the formula, since dividing it by $[1{+}\gamma]$ yields r_1 equal to an expression that provides the same upper bound by dropping the r_1^2/τ^2 term. Likewise $\sqrt{1{+}2d} \ge 1{+}d/(d{+}1)$, valid for $d \ge -1/2$, yields $r_1 \ge \mathit{diff}/[1{+}\gamma{+}\mathit{diff}/(\tau^2(1{+}\gamma))]$, where for $\mathit{diff} < 0$ the denominator is at least $(1{+}\zeta/\tau)^2$.

Using $r_1 \le r_1^*$, the $r^*_{crit}-1-\epsilon$ bound, equal to $r_1+\psi{-}1$, is not more than $r_1^*+\psi{-}1$, which may be written as

$$\frac{1}{1{+}\gamma}\big[2(\zeta')^2\big] + \frac{\gamma}{1{+}\gamma}\big[\psi{-}1\big],$$

an average of $2(\zeta')^2$ and $\psi{-}1$. Equivalently, it may be written

$$2(\zeta')^2 - \frac{\gamma}{1{+}\gamma}\big[2(\zeta')^2 + 1{-}\psi\big],$$

which is $2(\zeta')^2 - \gamma r_1^*$. Furthermore, since $(\tau^2{+}r_1)D(\delta_c)$ has the bound made to match $r_1 + \psi{-}1$, it is bounded by these same expressions.

Likewise, from the identity

$$\frac{2(\zeta'{-}r_1/2\tau)^2}{1{+}r_1/\tau^2} - (r_1 - 1) \,=\, \psi,$$

multiplying by $\bar\Phi(\zeta)$ this bounds the remainder used in Lemma 16 by $rem \le \psi\,\bar\Phi(\zeta)$. Accordingly, concerning the expression $(2\tau{+}\zeta)\phi(\zeta){+}rem$ equal to $\psi\,\Phi(\zeta){+}rem$ which arises as the numerator in the shortfall bound δ^*, using $\Phi(\zeta) + \bar\Phi(\zeta) = 1$, it is seen to be not more than ψ.

Next we check that δ_c is positive. From the above expression for δ_c that is equivalent to the positivity of $\zeta'{-}r_1/(2\tau)$, which is at least $\zeta' - r_1^*/(2\tau)$, which, using the form of r_1^* and regrouping, is seen to equal

$$\frac{\zeta'}{1{+}\gamma} + \frac{(\psi - 1)(1{+}2\zeta'/\tau)}{2\tau(1{+}\gamma)}$$

which is positive when ζ is such that $\psi \ge 1$.

This positivity of δ_c corresponds to $r_1/\tau < 2\zeta'$ and $1{+}r_1/\tau^2$ less than $1{+}2\zeta'/\tau$. For a lower bound on $1{+}r_1/\tau^2$ we may use $1 - (\psi{-}1)/\tau^2$ as previously discussed.

As for an upper bound on δ_c, of course $\delta_c \le 2\zeta'/\tau$ when $r_1 \ge 0$. Otherwise, from $r_1 \ge -(\psi{-}1)$ we have

$$\delta_c \,=\, \frac{2\zeta'{-}r_1/\tau}{\tau\,(1{+}r_1/\tau^2)} \,\le\, \frac{2\zeta'/\tau + (\psi{-}1)/\tau^2}{1 - (\psi{-}1)/\tau^2},$$

where the numerator of the right side matches γ. Using the upper bound on ψ it gives

$$\delta_c \le \frac{2\zeta' + 1.6}{\tau + 1/\tau - 1.6}.$$

To show a little more, it is not necessary to constrain $\psi \ge 1$ for the positivity of δ_c. Having $\tau \ge 1.4$ as true with $B > 2$ is sufficient for this positivity. Indeed factoring out $1/(1{+}\gamma)$, the above expression is at least $\zeta'(1{-}1/\tau^2) - 1/(2\tau){+}\psi/(2\tau)$. Using $\psi \ge 4\tau\phi(\zeta)$, it is at least $\zeta(1{-}1/\tau^2) - 1/(2\tau){+}2\phi(\zeta)$. At $\zeta = 0$ it is $2/\sqrt{2\pi} - 1/(2\tau)$ which is positive. Moreover, it goes up from there, since its derivative $1{-}1/\tau^2 - 2\zeta\phi(\zeta)$ is at its minimum at $\zeta = 1$, where it equals $1{-}1/\tau^2{-}2/\sqrt{2\pi e}$, positive for τ^2 at least $1/[1 - 2/\sqrt{2\pi e}]$, which is not more than 1.94. So $\tau \ge 1.40 \ge \sqrt{1.94}$ is sufficient for the δ_c to be positive.

The remainder of this proof is to established the simplified bound when ζ^2 exceeds $\psi - 1$. Use $\zeta' = \zeta(1 + \zeta/2\tau)$. Subtracting $2\zeta^2$ from our bound, what remains is

$$2\zeta^2\big[\big(1 + \zeta/2\tau\big)^2 - 1\big] - \frac{\gamma}{1+\gamma}\big[2(\zeta')^2 + 1{-}\psi\big].$$

We want this to be negative, so that $r_{crit}^* - 1 - \epsilon - 2\zeta^2 \le 0$. Multiplying through by $1 + \gamma$, the expression desired to be negative becomes

$$(1+\gamma)2\zeta^2\big[\big(1 + \zeta/2\tau\big)^2 - 1\big] - \gamma\big[2(\zeta')^2 + 1{-}\psi\big].$$

Using $\gamma = 2\zeta'/\tau + (\psi{-}1)/\tau^2$, this is a quadratic in ψ so it is possible to specify by quadratic formula the condition for negativity. To give a simplified sufficient condition, split the bound according to the two parts of γ, as a sum of the following two expressions. The main part is

$$(A) = (1{+}2\zeta'/\tau)2\zeta^2\big[\big(1{+}\zeta/2\tau\big)^2{-}1\big] - (2\zeta'/\tau)\big[2(\zeta')^2{+}1{-}\psi\big]$$

and the other is $(\psi - 1)/\tau^2$ times the expression

$$(B) = 2\zeta^2\big[\big(1 + \zeta/2\tau\big)^2 - 1\big] - \big[2(\zeta')^2 + 1{-}\psi\big].$$

We show that both (A) and (B) are negative when $\zeta^2 + 1$ exceeds ψ. First concerning (A), expanding $(1{+}\,\zeta/2\tau)^2 - 1$, rearranging, and simplifying, yields

$$(A) = -(2\zeta/\tau)(1 + \zeta/2\tau)\big[\zeta^2 + 1{-}\psi\big] - \zeta^4/2\tau^2$$

Likewise for the expression (B) it simplifies to

$$(B) = -\big[\zeta^2 + 1{-}\psi\big] - \zeta^2.$$

Both of these are negative when $\zeta^2{+}1$ exceeds ψ. That suffices for $(A) + (B)(\psi - 1)/\tau^2$ negative, assuming also $\psi{-}1 \ge 0$. Accordingly $r_{crit}^* - 2\zeta^2 - 1 - \epsilon$ is negative. This completes the proof of Corollary 30.

From this development, when ζ^2 exceeds $\psi - 1$, not only is $r_{crit}^* < 2\zeta^2 + 1 + \epsilon$, but also, the value of

$$\frac{2(\zeta' - r_1/2\tau)^2}{1 + r_1/2\tau^2} = r_1 - 1 + \psi$$

is also less than $2\zeta^2$. Consequently, we have the additional controls that

$$r_1 < 2\zeta^2 + 1{-}\psi$$

and that

$$(\tau^2 + r_1)D(\delta_c) < 2\zeta^2.$$

Appendix VIII
The variance of $\sum_{j\;sent} \pi_j 1_{H_{\lambda,k,j}}$

The variance of this weighted sum of Bernoulli that we wish to control is $V/L = \sum_{j\;sent} \pi_j^2 \Phi(\mu_{k,j})\bar\Phi(\mu_{k,j})$ with $\mu_{k,j} = \text{shift}_{k,j} - \tau$. The $\text{shift}_{k,j}$ may be written as $\sqrt{c_k\pi_j}\tau$, where $c_k = \nu L(1{-}h')/(2R(1{-}x\nu)(1{+}\delta_a)^2)$ evaluated at $x = q_{1,k-1}^{adj}$. Thus

$$V/L = \sum_{ell} \pi_{(\ell)}^2 \Phi\bar\Phi((\sqrt{c_k\pi_{(\ell)}} - 1)\tau)$$

where $\Phi\bar\Phi(z)$ is the function formed by the product $\Phi(z)\bar\Phi(z)$.

In the no-leveling $(c{=}0)$ case $\pi_{(\ell)} = e^{-2\mathcal{C}(\ell-1)/L}2\tilde{\mathcal{C}}/(\nu L)$ and $c_k\pi_{(\ell)} = u_\ell R'/(R(1{-}x\nu))$ with $R' = \tilde{\mathcal{C}}(1{-}h')/(1{+}\delta_a)^2$, where $u_\ell = e^{-2\mathcal{C}(\ell-1)/L}$.

With a quantifiably small error as before we may replace the sum over the grid of values of $t = \ell/L$ in $[0, 1]$ with the integral over this interval, yielding the value

$$V = (2\tilde{\mathcal{C}}/\nu)^2 \int_0^1 e^{-4\mathcal{C}t}\Phi\bar\Phi\left(\left(\sqrt{\frac{e^{-2\mathcal{C}t}\mathcal{C}'/R}{1 - x\nu}} - 1\right)\tau\right)dt.$$

If we change variables to $\tilde u = e^{-\mathcal{C}t}$ it is expressed as

$$V = \frac{(2\tilde{\mathcal{C}}/\nu)^2}{\mathcal{C}} \int_{e^{-\mathcal{C}}}^1 \tilde u^3\Phi\bar\Phi\left(\left(\tilde u\sqrt{\frac{R'/R}{1 - x\nu}} - 1\right)\tau\right)d\tilde u.$$

To upper bound it we replace the $\tilde{u}^3$ factor with 1 and change variables further to

$$z = \left(\tilde{u}\sqrt{\frac{R'/R}{1-x\nu}} - 1\right)\tau.$$

Thereby we obtain an upper bound and V of

$$\frac{\sqrt{1-x\nu}}{\tau\sqrt{R'/R}}\frac{(2\tilde{\mathcal{C}}/\nu)^2}{\mathcal{C}}\int \Phi\bar{\Phi}(z)dz.$$

Now $\Phi\bar{\Phi}(z)$ has the upper bound $(1/4)e^{-z^2/2}$, which is $\sqrt{2\pi}\phi(z)/4$, which when integrated on the line yields

$$V \le \frac{\sqrt{1-x\nu}}{\tau\sqrt{\mathcal{C}'/R}}\frac{(\tilde{\mathcal{C}}/\nu)^2}{\mathcal{C}}\sqrt{2\pi}.$$

When $R \le R'$, then using $\tilde{\mathcal{C}} \le \mathcal{C}$ and $x \le 1$, it yields

$$V \le \frac{\sqrt{2\pi}\mathcal{C}}{\nu^2\tau}.$$

This provides the desired upper bound on the variance.

APPENDIX IX
SLIGHT IMPROVEMENT TO THE VARIANCE OF $\sum_{j\ sent}\pi_j 1_{H_{\lambda,k,j}}$

The variance of this weighted sum of Bernoulli that we wish to control is $V/L = \sum_{j\ sent}\pi_j^2\Phi(\mu_{k,j})\bar{\Phi}(\mu_{k,j})$ with $\mu_{k,j} = \text{shift}_{k,j} - \tau$. The $\text{shift}_{k,j}$ may be written as $\sqrt{c_k\pi_j}\tau$, where $c_k = \nu L(1-h')/(2R(1-x\nu)(1+\delta_a)^2)$ evaluated at $x = q_{1,k-1}^{adj}$. Thus

$$V/L = \sum_{ell}\pi_{(\ell)}^2\Phi\bar{\Phi}((\sqrt{c_k\pi_{(\ell)}} - 1)\tau)$$

where $\Phi\bar{\Phi}(z)$ is the function formed by the product $\Phi(z)\bar{\Phi}(z)$.

In the no-leveling ($c{=}0$) case $\pi_{(\ell)} = e^{-2\mathcal{C}(\ell-1)/L}2\tilde{\mathcal{C}}/(\nu L)$ and $c_k\pi_{(\ell)} = u_\ell R'/(R(1-x\nu))$ with $R' = \tilde{\mathcal{C}}(1-h')/(1+\delta_a)^2$, where $u_\ell = e^{-2\mathcal{C}(\ell-1)/L}$.

With a quantifiable small error as before we may replace the sum over the grid of values of $t = \ell/L$ in $[0,1]$ with the integral over this interval, yielding the value

$$V = (2\tilde{\mathcal{C}}/\nu)^2\int_0^1 e^{-4\mathcal{C}t}\Phi\bar{\Phi}\left(\left(\sqrt{\frac{e^{-2\mathcal{C}t}\mathcal{C}'/R}{1-x\nu}} - 1\right)\tau\right)dt.$$

If we change variables to $\tilde{u} = e^{-\mathcal{C}t}$ it is expressed as

$$V = \frac{(2\tilde{\mathcal{C}}/\nu)^2}{\mathcal{C}}\int_{e^{-\mathcal{C}}}^1 \tilde{u}^3\Phi\bar{\Phi}\left(\left(\tilde{u}\sqrt{\frac{R'/R}{1-x\nu}} - 1\right)\tau\right)d\tilde{u}.$$

To upper bound the above we change variables further to

$$z = \left(\tilde{u}\sqrt{\frac{R'/R}{1-x\nu}} - 1\right)\tau.$$

Thereby we obtain an upper bound and V of

$$\frac{(1-x\nu)^2}{\tau(R'/R)^2}\frac{(2\tilde{\mathcal{C}}/\nu)^2}{\mathcal{C}}\int_{(e^{-\mathcal{C}}c_0-1)\tau}^{(c_0-1)\tau}(1+z/\tau)^3\,\Phi\bar{\Phi}(z)dz,$$

where $c_0 = \sqrt{R'/R(1-x\nu)}$. Now notice that $(e^{-\mathcal{C}}c_0-1)\tau$ is at least $-\tau$, making $1+z/\tau \ge 0$ on the interval of integration. Accordingly, the above integral is can be bounded from above by,

$$\int_{z\ge-\tau}(1+z/\tau)^3\,\Phi\bar{\Phi}(z)dz.$$

Further, the integral of $(1+z/\tau)^3\,\Phi\bar{\Phi}(z)$ for $z \le -\tau$ is a negligible term that is polynomially small in $1/B$. We ignore that term in the rest of the analysis. Correspondingly we need to bound the integral,

$$\frac{(1-x\nu)^2}{\tau(R'/R)^2}\frac{(2\tilde{\mathcal{C}}/\nu)^2}{\mathcal{C}}\int(1+z/\tau)^3\,\Phi\bar{\Phi}(z)dz.$$

Noticing that $\Phi\bar{\Phi}(z)$ is a symmetric function, the terms that are involve z and z^3 after the expansion of $(1+z/\tau)^3$ above vanish upon integrating. Consequently, we only need to bound the integral of $\Phi\bar{\Phi}(z)$ and $z^2\Phi\bar{\Phi}(z)$. Doing this numerically we see that the integral of the former is bounded by $a_1 = 0.57$ and that of the latter is bounded by $a_2 = 0.48$.

So ignoring the polynomially small term, the variance can be bounded by

$$V \le \frac{(1-x\nu)^2}{(R'/R)^2}\frac{(2\tilde{\mathcal{C}}/\nu)^2}{\mathcal{C}\tau}(a_1 + a_2/\tau^2).$$

which is less than,

$$(1-x\nu)^2\frac{(4\mathcal{C}/\nu^2)}{\tau}(a_1 + a_2/\tau^2).$$

We bound the above quantity by $(4\mathcal{C}/\nu^2)(a_1 + a_2/\tau^2)/\tau$. Let's ignore the a_2/τ^2 term since this of smaller order. Then the variance can be bounded by $1.62/\sqrt{\log B}$, where we use that $\tau \ge \sqrt{2\log B}$ and that $4a_1/\sqrt{2}$ is less than 1.62.

APPENDIX X
NORMAL TAILS

Let Z be a standard normal random variable and let $\phi(z)$ be its probability density function, $\Phi(z)$ be its cumulative distribution and $\bar{\Phi}(z) = 1 - \Phi(z)$ be its upper tail probability for $z > 0$. Here we collect some properties of this probability, beginning with a conclusion from Feller. Most familiar is his bound $\bar{\Phi}(z) \le (1/z)\phi(z)$ which may be stated as $\phi(z)/\bar{\Phi}(z)$ being at least z. His lower bound $\bar{\Phi}(z) \ge (1/z - 1/z^3)\phi(z)$ has certain natural improvements, which we express through upper bounds on $\phi(z)/\bar{\Phi}(z)$ showing how close it is to z.

Lemma 46: For positive z the upper tail probability $\mathbb{P}\{Z > z\} = \bar{\Phi}(z)$ satisfies $\bar{\Phi}(z) \le (\sqrt{2\pi}/2)\phi(z)$ and satisfies the Feller expansion

$$\bar{\Phi}(z) \sim \phi(z)\left(\frac{1}{z} - \frac{1}{z^3} + \frac{3}{z^5} - \frac{3\cdot 5}{z^7} + \dots\right),$$

with terms of alternating sign, where terminating with any term of positive sign produces an upper bound and terminating with any term of negative sign produces a lower bound. Furthermore, for $z > 0$ the ratio $\phi(z)/\bar{\Phi}(z)$ is increasing and is less than $z + 1/z$. Further improved bounds are that it is less than $\xi(z)$ equal to 2 for $0 \le z \le 1$ and equal to $z + 1/z$ for $z \ge 1$, and, slightly better, $\phi(z)/\bar{\Phi}(z)$ is less than $[z + \sqrt{z^2+4}]/2$. Moreover, the positive $\phi(z)/\bar{\Phi}(z) - z$ is a decreasing function of z.

Proof of Lemma 46: The expansion is from Feller [32], Chapter VII, where we note in particular that the first order upper bound $\bar{\Phi}(z) < (1/z)\phi(z)$ is obtained from $\phi'(t) = -t\phi(t)$ by noting that $z\bar{\Phi}(z) = z\int_z^\infty \phi(t)dt$ is less than $\int_z^\infty t\phi(t)dt = \phi(z)$. Thus the ratio $\phi(z)/\Phi(z)$ exceeds z. It follows that the derivative of the ratio $\phi(z)/\bar{\Phi}(z)$ which is $[\phi(z)/\bar{\Phi}(z) - z]\phi(z)/\bar{\Phi}(z)$ is positive, so this ratio is increasing and at least its value at $z = 0$, which is $2/\sqrt{2\pi}$.

Now for any positive c consider the positive integral $\int_z^\infty (t/c - 1)^2\phi(t)dt$. By expanding the square and using that $(t^2 - 1)\phi(t)$ is the derivative of $-t\phi(t)$ on sees that this integral is $(1 + 1/c^2)\bar{\Phi}(z) - (2/c - z/c^2)\phi(z)$. Multiplying through by c^2, and assuming $2c > z$, its positivity gives the family of bounds

$$\phi(z)/\bar{\Phi}(z) \le \frac{c^2 + 1}{2c - z}.$$

Evaluating it at $c = z$ gives the upper bound on the ratio of $(z^2 + 1)/z = z + 1/z$. Note that since $z/(z^2 + 1)$ equals $1/z - 1/[z(z^2 + 1)]$ it improves on $1/z - 1/z^3$ for all $z \ge 0$. Since $\phi(z)/\bar{\Phi}(z)$ is increasing we can replace the upper bound $z + 1/z$ with its lower increasing envelope, which is the claimed bound $\xi(z)$, noting that $z + 1/z$ takes its minimum value of 2 at $z = 1$ and is increasing thereafter. For further improvement note that $\phi(z)/\bar{\Phi}(z)$ equals a value not more than 1.53 at $z = 1$, so the bound 2 for $0 \le z \le 1$ may be replaced by 1.53.

Next let's determine the best bound of the above form by optimizing the choice of c. The derivative of the bound is the ratio of $2c(2c - z) - 2(c^2 + 1)$ and $(2c - z)^2$ and the c that sets it to 0 solves $c^2 - zc - 1 = 0$ for which $c = [z + \sqrt{z^2 + 4}]/2$, and the above bound is then equal to this c.

As for the monotonicity of $\phi(z)/\bar{\Phi}(z) - z$, its derivative is $(\phi/\bar{\Phi})^2 - z(\phi/\bar{\Phi}) - 1$ which is a quadratic in the positive quantity $\phi/\bar{\Phi}$, abbreviating $\phi(z)/\Phi(z)$. Hence by inspecting the quadratic formula, this derivative is negative if $\phi/\bar{\Phi}$ is less than or equal to $[z + \sqrt{z^2 + 4}]/2$, which it is by the above bound. This completes the proof of Lemma 46.

We remark that $\log \phi(z)/\bar{\Phi}(z)$ has first derivative $\phi(z)/\bar{\Phi}(z) - z$ equal to the quantity studied in this lemma and second derivative found above to be negative. So the fact that $\phi(z)/\bar{\Phi}(z) - z$ is decreasing is equivalent to the normal hazard function $\phi(z)/\bar{\Phi}(z)$ being log-concave.

Appendix XI
Tails for weighted Bernoulli sums

Lemma 47: Let W_j, $1 \le j \le N$ be N independent Bernoulli(r_j) random variables. Furthermore, let α_j, $1 \le j \le K$ be non-negative weights that sum to 1 and let $N_\alpha = 1/\max_j \alpha_j$. Then the weighted sum $\hat{r} = \sum_j \alpha_j W_j$ which has mean given by $r^* = \sum_j \alpha_j r_j$, satisfies the following large deviation inequalities. For any r with $0 < r < r^*$,

$$P(\hat{r} < r) \le \exp\{-N_\alpha D(r\|r^*)\}$$

and for any $\tilde{r}$ with $r^* < \tilde{r} < 1$,

$$P(\hat{r} > \tilde{r}) \le \exp\{-N_\alpha D(\tilde{r}\|r^*)\}$$

where $D(r\|r^*)$ denotes the relative entropy between Bernoulli random variables of success parameters r and r^*.

Proof of Lemma 47: Let's prove the first part. The proof of the second part is similar.

Denote the event

$$\mathcal{A} = \{\underline{W} : \sum_j \alpha_j W_j \le r\}$$

with $\underline{W}$ denoting the N-vector of W_j's. Proceeding as in Csiszar [24] we have that

$$\begin{aligned} P(\mathcal{A}) &= \exp\{-D(P_{\underline{W}|\mathcal{A}} \| P_{\underline{W}})\} \\ &\le \exp\Big\{-\sum_j D(P_{W_j|\mathcal{A}} \| P_{W_j})\Big\} \end{aligned}$$

Here $P_{\underline{W}|\mathcal{A}}$ denotes the conditional distribution of the vector $\underline{W}$ conditional on the event $\mathcal{A}$ and $P_{W_j|\mathcal{A}}$ denotes the associated marginal distribution of W_j conditioned on $\mathcal{A}$. Now

$$\sum_j D(P_{W_j|\mathcal{A}} \| P_{W_j}) \ge N_\alpha \sum_j \alpha_j D(P_{W_j|\mathcal{A}} \| P_{W_j}).$$

Furthermore, the convexity of the relative entropy implies that

$$\sum_j \alpha_j D(P_{W_j|\mathcal{A}} \parallel P_{W_j}) \ge D\Big(\sum_j \alpha_j P_{W_j|\mathcal{A}} \parallel \sum_j \alpha_j P_{W_j}\Big).$$

The sums on the right denote α mixtures of distributions $P_{W_j|\mathcal{A}}$ and P_{W_j}, respectively, which are distributions on $\{0, 1\}$, and hence these mixtures are also distributions on $\{0, 1\}$. In particular, $\sum_j \alpha_j P_{W_j}$ is the Bernoulli(r^*) distribution and $\sum_j \alpha_j P_{W_j|\mathcal{A}}$ is the Bernoulli(r_e) distribution where

$$r_e = \mathrm{E}\big[\sum_j \alpha_j W_j \,|\, \mathcal{A}\big] = \mathrm{E}\big[\hat{r} \,|\, \mathcal{A}\big].$$

But in the event $\mathcal{A}$ we have $\hat{r} \le r$ so it follows that $r_e \le r$. As $r < r^*$ this yields $D(r_e \parallel r^*) \ge D(r \parallel r^*)$. This completes the proof of Lemma 47.

Appendix XII
Lower Bounds on D

Lemma 48: For $p \ge p^*$, the relative entropy between Bernoulli(p) and Bernoulli(p^*) distributions has the succession of lower bounds

$$D_{Ber}(p\|p^*) \ge D_{Poi}(p\|p^*) \ge 2\big(\sqrt{p} - \sqrt{p^*}\big)^2 \ge \frac{(p - p^*)^2}{2p}$$

where $D_{Poi}(p\|p^*) = p \log p/p^* + p^* - p$ is also recognizable as the relative entropy between Poisson distributions of mean p and p^* respectively.

Remark a: There are analogous statements for pairs of probability distributions P and P^* on a measurable space $\mathcal{X}$ with densities $p(x)$ and $p^*(x)$ with respect to a dominating measure μ. The relative entropy $D(P\|P^*)$ which is $\int p(x) \log p(x)/p^*(x)\mu(dx)$ may be written as the integral of the non-negative integrand $p(x) \log p(x)/p^*(x) + p^*(x) - p(x)$, which exceeds $(1/2)\big(p(x) - p^*(x)\big)^2/\max\{p(x), p^*(x)\}$. It is familiar that $D(P\|P^*)$ exceeds the squared Hellinger distance

$H^2(P,P^*) = \int \big(\sqrt{p(x)} - \sqrt{p^*(x)}\,\big)^2 \mu(dx)$. That fact arises for instance via Jensen's inequality, from which D exceeds $2\log 1/(1-(1/2)H^2)$ which in turn is at least H^2. However, we have not been able to get $D \geq 2H^2$ in general. The above lemma with the factor of 2 is for $p > p^*$, not for general integrands.

Remark b: When $\hat{p}$ is the relative frequency of occurrence in N independent Bernoulli trials it has the bound $P\{\hat{p} > p\} \leq e^{-ND_{Ber}(p\|p^*)}$ on the upper tail of the Binomial distribution of $N\hat{p}$ for $p > p^*$. In accordance with the Poisson interpretation of the lower bound on the exponent, one sees that this upper tail of the Binomial is in turn bounded by the corresponding large deviation expression that would hold if the random variables were Poisson.

Proof of Lemma 48: The Bernoulli relative entropy may be expressed as the sum of two positive terms, one of which is $p\log p/p^* + p^* - p$, and the other is the corresponding term with $1-p$ and $1-p^*$ in place of p and p^*, so this demonstrates the first inequality. Now suppose $p > p^*$. Write $p\log p/p^* + p^* - p$ as $p^* F(s)$ where $F(s) = 2s^2\log s + 1 - s^2$ with $s^2 = p/p^*$ which is at least 1. This function F and its first derivative $F'(s) = 4s\log s$ have value equal to 0 at $s = 1$, and its second derivative $F''(s) = 4 + 4\log s$ is at least 4 for $s \geq 1$. So by second order Taylor expansion $F(s) \geq 2(s-1)^2$ for $s \geq 1$. Thus $p\log p/p^* + p^* - p$ is at least $2\big(\sqrt{p} - \sqrt{p^*}\,\big)^2$. Furthermore $2(s-1)^2 \geq (s^2-1)^2/(2s^2)$ as, taking the square root of both sides, it is seen to be equivalent to $2(s-1) \geq s^2 - 1$, which, factoring out $s-1$ from both sides, is seen to hold for $s \geq 1$. From this we have the final lower bound $(p-p^*)^2/(2p)$.

Acknowledgment

We thank Dan Spielman, Imre Teletar, Martin Wainwright, Sergio Verdú, Cong Huang, Xi Luo, Edmund Yeh, John Hartigan, Mokshay Madiman, Harrison Zhou and Joseph Chang for helpful conversations. We especially thank David Smalling who completed a number of simulations of earlier incarnations of the decoding algorithm for his Yale applied math senior project in spring term of 2009 and Yale statistics masters student Creighton Hauikulani who took the simulations further in 2009 and 2010. Their work was instrumental to us in recognizing that without modification, direct convex projection methods have a rate threshold substantially below capacity when the signal-to-noise ratio is high.

We especially thank anonymous reviewers of the companion paper [10] who generously provided helpful review of aspects of the sparse signal recovery literature, much of which we have incorporated here.

References

[1] A. Abbe and A.R. Barron, "Polar codes for the AWGN," *Proc. IEEE Intern. Symp. Inform. Theory*, St. Petersburg, Russia, August 2011.

[2] E. Abbe and E. Telatar, *Polar codes for the m-user MAC,* in Proc. 2010 International Zurich Seminar on Communications, Zurich, 2010. Available at arXiv:1002.0777v2.

[3] Y. Altuğ and A. Wagner, "Moderate deviation analysis of channel coding," *Proc. IEEE Intern. Symp. Inform. Theory*, Austin, TX, June 13-18, 2010.

[4] E. Arikan, "Channel polarization," *IEEE Trans. Inform. Theory*, Vol.55, 2009.

[5] E. Arikan and E. Telatar, "On the rate of channel polarization," *Proc. IEEE Internat. Symp. Inform. Theory*, Seoul, July 2009.

[6] A. Barg and G. Zémor, "Error exponents of expander codes," *IEEE Trans. Inform. Theory*, vol.48, pp.1725-1729, June 2002.

[7] A. Barg and G. Zémor, "Error exponents of expander codes under linear-complexity decoding," *SIAM J. Discrete Math*, vol.17, no.3, pp.426-445, 2004.

[8] A.R. Barron, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Trans. Inform. Theory*, vol.39, pp.930-944, 1993.

[9] A. Barron, A. Cohen, W. Dahmen and R. Devore, "Approximation and learning by greedy algorithms," *Annals of Statistics*, vol.36 ,no.1, pp.64-94, 2007.

[10] A.R. Barron A. Joseph, "Least squares superposition coding of moderate dictionary size, reliable at rates up to channel capacity," Full manuscript submitted to the *IEEE Trans. Inform. Theory*. Conference version in *Proc. IEEE Intern. Symp. Inform. Theory*, Austin, TX, June 13-18, 2010.

[11] A.R. Barron and A. Joseph, "Toward fast reliable communication at rates near capacity with Gaussian noise, *Proc. IEEE Intern. Symp. Inform. Theory*, Austin, TX, June 13-18, 2010.

[12] Y. Benjimini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing, *J. Roy. Statist. Soc. Ser. B*, vol57,pp.289-300, 1995.

[13] C. Berrou, A. Glavieux and P. Thitimajshima, "Near Shannon limit error correcting coding and decoding: Turbo coes," *Proc. IEEE Int. Conf. Communications*, Geneva, May 1993, pp.1064-1070.

[14] R. Calderbank, Howard, and S. Jafarpour, "Construction of a large class of deterministic sensing matrices that satisfy a statistical isometry property *IEEE Journal of Selected Topics in Signal Processing: Special Issue on Compressed Sensing*, 2009.

[15] E. Candès and Y. Plan, "Near-ideal model selection by ℓ_1 minimization," *Annals of Statistics*, 2009.

[16] E. Candés and T. Tao, "Near-optimal signal recovery from random projections: Universal Encoding Strategies," *IEEE Trans. Inform. Theory*, vol.52, no.12, pp.5406-5425, Dec. 2006.

[17] E. Candes and T. Tao, "Decoding by linear programming," *IEEE Trans. Inform. Theory*, vol.51, 2005.

[18] J. Cao and E.M. Yeh, "Asymptotically optimal multiple-access communication via distributed rate splitting," *IEEE Trans. Inform. Theory*, vol.53, pp.304-319, Jan.2007.

[19] S.S. Chen, D. Donoho, and M.A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Schientific Computing*, vol.20, pp.31-61, 1999.

[20] J.H. Conway and N.J.A. Sloane, *Sphere Packings, Lattices and Groups*, New York, Springer-Verlag, 1988.

[21] D.J. Costello, Jr. and G.D. Forney, Jr. "Channel coding: The road to channel capacity," *Proceedings of the IEEE.* 2007.

[22] T.M. Cover, "Broadcast channels," *IEEE Trans. Inform. Theory*, vol.18, pp.2-14, 1972.

[23] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, New York, Wiley-Interscience, 2006.

[24] I. Csiszár, "Sanov properties, generalized I-projection, and a conditional limit theorem. *Annals of Probability*, vol.12, pp.768-793,1984.

[25] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, New York, Academic Press, 1980.

[26] D.L. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol.52, no.4, pp.1289-1306, April 2006.

[27] D.L Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Trans. Inform. Theory*, vol.47, pp.2845-2862, Nov. 2001.

[28] D.L. Donoho, M. Elad, and V.M. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Inform. Theory*, vol.52, no.1, pp.6-18, Jan. 2006.

[29] , "Multiple regression analysis," In A. Ralston and H.S. Wilf (editors) *Mathematical Methods for Digital Computers*, Wiley, New York, 1960.

[30] A. El. Gamal and T.M. Cover, "Multiple user information theory, *Proc. IEEE* vol.68, pp.1466-1483, 1980.

[31] J. Fan and J. Lv, "A selective overview of variable selection in high dimensional feature space," *Statistica Sinica*, Vol.20, 2010.

[32] W. Feller, *An Introduction to Probability Theory and Its Applications* Vol.I, New York, John Wiley and Sons, Third Edition, 1968.

[33] A.K. Fletcher, S. Rangan, and V.K. Goyal, "Necessary and sufficient conditions for sparsity pattern recovery," *IEEE Trans. Inform. Theory*, vol.55, no.12, pp.5758-5773.

[34] A.K. Fletcher, S. Rangan, and V.K. Goyal, "On-Off Random Access Channels: A Compressed Sensing Framwork." ArXiv:0903.1022v2, 2006.

[35] G.D. Forney, Jr. *Concatenated Codes*, Research Monograph No. 37, Cambridge, Massachusetts, M.I.T. Press. 1966.

[36] G.D. Forney, Jr., G. Ungerboeck, "Modulation and Coding for Linear Gaussian Channels," *IEEE Trans. Inform. Theory*, vol.44, no.6, pp.2384-2415, Oct. 1998.

[37] J.J. Fuchs, "Recovery of exact sparse representations in the presence of bounded noise," *IEEE Trans. Inform. Theory*, vol.51, no.10, Oct. 2005.

[38] R.G. Gallager, *Low Density Parity-Check Codes*, Cambridge, Massachusetts, M.I.T. Press. 1963.

[39] R.G. Gallager, *Information Theory and Reliable Communication*, New York, John Wiley and Sons, 1968.

[40] S. Gurevich and R. Hadani, "The statistical restricted isometry property and the Wigner semicircle distribution of incoherent dictionaries," ArXiv:0812.2602.

[41] A.C. Gilbert and J.A. Tropp, "Applications of sparse approximation in communications," *Proc. IEEE Internat. Symp. Inform. Theory*, pp.1000-1004, Adelaide, Sept. 2005.

[42] J. Hagenauer, E. Offer, and L. Papke, "Iterative decoding of binary block and convolutional codes," *IEEE Trans. Inform. Theory*, vol.42, no.2, pp.429-445, Mar.1996.

[43] , C. He, M. Lentmaier, D.J. Costello, and K.Sh. Zigangirov, "Joint permutor analysis and design for multiple Turbo codes," vol.52, no.9, Sept.2006.

[44] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *J. American Statist. Assoc.*, pp.13-30, March, 1963.

[45] C. Huang, A.R. Barron, and G.H.L. Cheang, "Risk of penalized least squares, greedy selection and ℓ_1-penalization for flexible function libraries" Preprint, 2008.

[46] S. Jafarpour, W. Xu, B. Hassibi, R. Calderbank, "Efficient compressed sensing using high-quality expander graphs," *IEEE Trans. Information Theory* Vol.55, 2009.

[47] L. Jones, "A simple lemma for optimization in a Hilbert space, with application to projection pursuit and neural net training," *Annals of Statistics*, vol.20, pp.608-613, 1992.

[48] A.M. Kakhaki, H.K. Abadi,P. Pad,H. Saeedi,K. Alishahi and F. Marvasti, "Capacity achieving random sparse linear codes," arXiv:1102.4099v1, Feb. 2011.

[49] W.S. Lee, P. Bartlett, B. Williamson, *IEEE Trans. Inform. Theory*, vol.42, pp.2118-2132, 1996.

[50] S. Lin, D. Costello *Error Control Coding*, Pearson, Prentice Hall, 2004.

[51] M.G. Luby, M. Mitzenmacher, M.A. Shokrollahi, and D.A. Spielman, "Efficient erasure correcting codes." *IEEE Trans. Inform. Theory*, vol.47, pp.569-584, Feb.2001.

[52] M.G. Luby, M. Mitzenmacher, M.A. Shokrollahi, and D.A. Spielman, "Improved low-density parity-check codes using irregular graphs." *IEEE Trans. Inform. Theory*, vol.47, pp.585-598, Feb.2001.

[53] F.J. MacWilliams and N.J.A. Sloane, *The Theory of Error-Correcting Codes*, Amsterdam and New York, North-Holland Publishing Co.

[54] S. Mallat and Z. Zhang, "Matching pursuit with time-frequency dictionaries. *IEEE Trans. Signal Proc.*, vol.41, pp.3397-3415, 1993.

[55] M. Malloy and R. Nowak, "Sequential analysis in high-dimensional multiple testing and sparse recovery," Department of Electrical Engineering Technical Report, University of Wisconsin, 2011.

[56] Y.C. Pati, R. Rezaiifar and P.S. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," *Proc. 27th Ann. Asilomar Conf. Signals, Systems, and Computers* Nov.1993.

[57] L. Perez, J. Seghers, and D.J. Costello Jr., "A distance spectrum interpretation of turbo codes," *IEEE Trans. Inform. Theory*, vol.42, no.6, pp.1698-1709, Nov.1996.

[58] Y. Polyanskiy, H.V. Poor and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inform. Theory*, May 2010.

[59] I.S. Reed and G. Solomon, "Polynomial codes over certain finite fields." *J. SIAM*, vol.8, pp.300-304, June 1960.

[60] G. Reeves and M. Gastpar, "Approximate sparsity pattern recovery: information-theoretic lower bounds," arXiv:1002.4458v2, March 2010.

[61] T. Richardson and R. Urbanke, *Modern Coding Theory*, Cambridge Univ. Press, 2008.

[62] B. Rimoldi and R. Urbanke, "A rate-splitting approach to the Gaussian multiple-access channel capacity." *IEEE Trans. Inform. Theory*, vol.47, pp.364-375, Mar. 2001.

[63] C.E. Shannon, "A mathematical theory of communication." *Bell Syst. Tech. J.*, vol.27, pp.379-423 and 623-656, 1948.

[64] M. Sipser and D.A. Spielman, "Expander codes," *IEEE Trans. Inform. Theory*, vol.42, pp.1710-1722, Nov. 1996.

[65] V.N. Temlyakov and P. Zheltov, "On performance of greedy algorithms," Submitted to *J. Approximation Theory*, 2010.

[66] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *J. Roy. Statist. Soc. Ser. B*, Vol.58, no.1, pp.267-288.

[67] J.A. Tropp, "Just relax: convex programming methods for identifying sparse signals in noise," *IEEE Trans. Inform. Theory*, vol.52, no.3, pp. 1030-1051, Mar. 2006. (Correction note, vol.55, pp.917-918, Feb.2009).

[68] J.A. Tropp, "On the conditioning of random subdictionaries," *Appl. Comput. Harmonic Anal.*, vol.25, pp.1-24, 2008.

[69] J.A. Tropp, "Norms of random submatrices and sparse approximation," *C. R. Acad. Sci. Paris, Ser. I*, vol.346, pp.1271-1274, 2008.

[70] J.A. Tropp and A.C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inform. Theory*, vol.53, pp.4655-4666, Dec.2007.

[71] S. Verdú, *Multi-User Detection*, Cambridge University Press, 1998.

[72] M.J. Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso)." *IEEE Trans. Inform. Theory*, vol. 55, no. 5, pp. 2183-2202, May 2009.

[73] M.J. Wainwright, "Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting," *IEEE Trans. Inform. Theory*, vol.55, no.12, pp.5728-5741, December 2009.

[74] Y. Wu and S. Verdú, "The impact of constellation cardinality on Gaussian channel capacity," *Proc. Forty-Eighth Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, Sept. 2010.

[75] G. Zémor, "On expander codes," *IEEE Trans. Inform. Theory (Special Issue on Codes on Graphs and Iterative Algorithms*, vol.47,pp.835-837, Feb. 2001.

[76] T. Zhang, "Adaptive forward-backward greedy algorithm for learning sparse representations," *IEEE Trans. Inform. Theory*, To appear, Vol.57, 2011.

[77] T. Zhang, "Sparse recovery with orthogonal matching pursuit," *IEEE Trans. Inform. Theory*, To appear, Vol.57, 2011.

[78] T. Zhang, "Multistage convex relaxation for feature selection," Department of Statistics Technical Report, Rutgers University, 2011.