Achieving Information-Theoretic Limits with High-Dimensional Regression

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Antony Joseph

Dissertation Director: Andrew Barron

June 6, 2012

Abstract

Achieving Information-Theoretic Limits with High-Dimensional Regression

Antony Joseph

2012

We focus on variable selection in the high-dimensional regression setting with random Gaussian designs. More specifically, we consider the linear model,

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

where the response $Y \in \mathbb{R}^n$, the design matrix $X \in \mathbb{R}^{n \times N}$, the coefficient vector $\beta \in \mathbb{R}^N$, and noise $\epsilon \in \mathbb{R}^n$. We deal with the high-dimensional setting, where the dimension N is typically much greater than the sample size n. We also assume that the coefficient vector β is sparse, that is, it has L non-zeroes, where L is much smaller than the dimension N. We apply the above setup to the problem of reliable communication through a noisy channel.

Two estimation procedures are analyzed. The first, which we call the *Least Squares decoder*, involves an exhaustive search over all allowed coefficient vectors and selecting the β which minimizes $\|Y - X\beta\|^2$, or the distance of the response Y from the fitted value $X\beta$. Although it can be shown that this is the optimal estimator for the problem, it is not computationally feasible. The second procedure is a computationally feasible algorithm, which is similar in spirit to iterative algorithms such as forward stepwise regression. We call this the *Iterative Thresholding Algorithm*.

For both procedures we demonstrate that the sample size required, in certain regimes, is optimal when compared to information-theoretic limits. In other words, we show that the sample size in relation to the sparsity and dimension, is correct not just up to orders of magnitude, but also up to the constant. Translated to the communication setting, this provides theoretically provable, low computational complexity communication systems based on our statistical framework.

Acknowledgements

I am indebted to all faculty associated with the Statistics department at Yale for their help and support during my five years as a Ph.D. student. The knowledge and advice they provided have helped me immeasurably. I would also like to thank Joann and Ashley for their kindness and help in so many matters.

I am extremely grateful to my advisor Professor Andrew Barron. Not only has he encouraged and guided me in working on a very challenging problem, he has on several occasions gone out of his way to help me in other professional matters.

I would also like to thank Professor David Pollard for teaching advanced courses in probability, such as Asymptotics, which have proven to be invaluable to my research.

The relaxed weekly seminars of the Yale Probabilistic Networks Group not only introduced me to some exciting results, but also gave me an opportunity to improve my presentation skills.

The numerous friends I made have also helped make my stint at Yale a thoroughly enjoyable experience, and I wish to thank them too. There are also so many things in the university campus that I cherish – the libraries, coffee shops, restaurants, Payne Whitney gym, bike trails, tennis courts, East Rock, Helen Hadley Hall, to name a few.

Dedicated to my parents

Contents

List of Figures

Chapter 1

Introduction

The high-dimensional linear regression model has been of immense interest in recent times. The model may be expressed as

$$Y = X\beta + \epsilon, \tag{1.1}$$

where Y is an $n \times 1$ response vector, X an $n \times N$ design matrix, β an $N \times 1$ coefficient vector, and ϵ an $n \times 1$ noise vector.

As opposed to the classical regression model, in the high-dimensional case, the dimension N need not be small – in fact, typically it is much larger – compared to the sample size n. Under the above scenario, accurate statistical estimation of the coefficient vector β is not possible unless there is some structure imposed on it. The most common assumption on β is the *sparsity* assumption, which, in its simplest form, states that the size of the set of non-zeroes of β, or the support of β, is small compared to N. We denote as L the number of non-zeroes of β. An aim, under the above setup, is the prediction of the response variable Y, or in other words, estimation of $X\beta$.

An equally important task is the estimation of the coefficient vector β. An estimator $\hat{\beta}$ is evaluated in terms of closeness to β with respect to some metric, for example the ℓ_2 norm. When β has a small support, another test of goodness of the estimator $\hat{\beta}$ is to see how well one can recover the support of β from that $\hat{\beta}$.

The above setup for recovering sparse signals has found application in many modern day problems. For example, in genomics, people are interested in identifying positions in the human genome responsible for causing a particular disease. When framed as a high-dimensional regression model, each row of the X matrix stores genetic data for a particular person. More specifically, each entry in

the row vector corresponds to data at a particular position of the person's genome. The corresponding entry of the Y vector gives the disease status of that person. This problem is high-dimensional in nature since the number of positions in the genome that are sampled (which corresponds to the dimension N) is typically in the millions, whereas, the number of people involved in the study (the sample size n) is much smaller. Sparsity appears from assuming that only a few positions in the genome play a role for causing the disease in the population.

Other examples where the high-dimensional regression framework has found applicability include graphical model selection (Meinshausen and Buhlmann [27]), compressed sensing (Donoho [16], Candes and Tao [11]), and computer vision (Wright et al. [40]).

In this document we discuss and analyze a novel use of the above model (see also Barron and Joseph [5], Barron and Joseph [6] for earlier versions) for the age-old problem of communicating reliably through a noisy medium. Our theoretical analysis demonstrates how information-theoretic limits for communication, discovered by Shannon [32], can be attained in the high-dimensional regression framework. These limits translate into relationships between sample size n, dimension N, sparsity L, and signal-to-noise ratio, for support recovery in regression. As a consequence, our analysis also contributes to a greater understanding of the statistical problem of estimation in the high-dimensional regression framework by identifying regimes where the information-theoretic limits can be attained.

In the communication context, our work attempts to provide a theoretically provable and practical solution to the Shannon coding problem. This is important since even though Shannon theory identifies key information-theoretic limits for reliable communication, it does not provide practical solutions. Modern day communication schemes have been demonstrated to have good empirical performance, however, a theoretical understanding of these schemes are limited only to certain special cases.

The document is organized as follows. In Chapter 2, we describe the communication problem, along with the associated information-theoretic limits. We also describe its formulation as a regression problem. In Chapters 3 and 4 we analyze two decoding schemes and demonstrate that these schemes do attain information-theoretic limits. In Chapter 3 we analyze the *Least Squares* decoding scheme, which is the optimal scheme for this problem. However, this scheme is not computationally feasible. In Chapter 4 a practical *Iterative Thresholding* algorithm, which is similar in spirit to greedy algorithms such as forward-stepwise regression [22],[7], [26], [28], is analyzed.

Chapter 2

The communication problem

2.1 Introduction

The goal in communication is to send information reliably through a noisy medium, also known as a *channel*. The framework of modern day communication schemes remains true to the scheme laid down by Shannon [32] . His path breaking work was instrumental in giving rise to the era of digital communication. For example, in communication using telephones, digitizing involves converting sound waves for a particular time interval into digital information in the form of binary strings of a particular length. For transmission through the channel, the sender encodes these binary strings into codewords. At the other end of the channel, the receiver gets a noisy version of the codeword. His goal is to decode the codeword sent, or equivalently the associated binary string, from knowledge of the dictionary of codewords and the channel model. The decoded binary string is reconverted back to a sound wave for the receiver to hear.

Here, we are only concerned with the part of the communication process involving mapping of the binary information as real-valued codewords and the recovery of the correct codeword from noisy versions of it. An important ingredient in this is the modeling of the noisy medium or the channel. Shannon's theoretical analysis identified key information-theoretic limits, depending on the channel model, for reliable communication. Below, in section 2.2 , we describe in greater detail communication using a popular channel known as the Gaussian channel along with the associated information-theoretic limits.

Even though Shannon's theory was instrumental in quantifying the efficiency of a coding scheme,

it does not provide practical schemes for encoding and decoding. Modern day practical schemes, for example using LDPC codes and Turbo codes [18], [9], have been empirically demonstrated to perform well when compared to these limits. However, a good theoretical understanding of these schemes has been lacking, even to this day. In section 2.3 we discuss an entirely different approach, based on the high-dimensional regression framework, for the construction of practical coding schemes. Chapters 3 and 4 deal with the theoretical analysis of the scheme.

2.2 Communication with the Gaussian channel

A Gaussian channel takes inputs $c \in \mathbb{R}$ and outputs $Y \in \mathbb{R}$, where

$$Y = c + \epsilon, \qquad \epsilon \sim N(0, \sigma^2). \tag{2.1}$$

Here $\sigma^2 > 0$ represents the noise variance in the channel. We now mention in greater detail communication using the Gaussian channel. The goal is to send any one of a set of messages reliably through the channel. The set of messages comprises of binary strings of a particular length K, allowing for a total for 2^K possible choices of messages.

Prior to transmission, an encoder is used to map each input bit string $u = (u_1, u_2 \ldots, u_K)$, where each $u_i \in \{0, 1\}$, into a length n vector of real numbers $(x_1, x_2, \ldots, x_n)$, known as a *codeword*, which may satisfy,

$$\frac{1}{n} \sum_{i=1}^{n} x_i^2 \leq P. \tag{2.2}$$

The positive quantity P is called the *power*. This terminology arises from the fact that the above constraint on the average ℓ_2 norm of each codeword translates to a constraint on the energy allocated per transmission.

Without loss of generality, we denote the set of input bit strings as $\mathcal{A} = \{1, 2 \ldots, 2^K\}$, with the understanding that each element of this set corresponds to a binary vector of length K. For any $j \in \mathcal{A}$, denote the corresponding codeword in $\mathbb{R}^n$ as X_j. From the power constraint one has $\|X_j\|^2/n \leq P$, for $j = 1, 2 \ldots, 2^K$, where $\|.\|$ denotes the ℓ_2- norm. Denote as X the *codebook*, which is the $n \times N$ matrix comprising of the codewords. In other words,

$$X_{n \times N} = [X_1 : X_2 : \ldots : X_{2^K}].$$

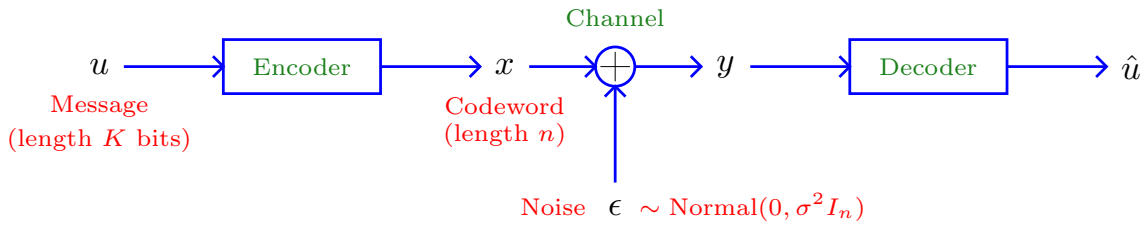

Figure 2.1: The general framework of communication using a Gaussian channel.

For transmission, if the sender wants to send an input bit string $j \in \mathcal{A}$ through the channel he does this by selecting the corresponding codeword X_j and transmits it through the channel. In a broad sense, one may assume that this involves sending each entry of X_j sequentially through the channel defined in (2.1). Thus, the sender makes a total of n transmissions.

The receiver gets $Y \in \mathbb{R}^n$, where

$$Y = X_j + \epsilon \qquad \text{where } \epsilon \sim N(0, \sigma^2 I_n).$$

The receiver's goal is to detect the j sent from the knowledge of the received string Y and the dictionary matrix X. The receiver is also aware of the power P and the noise variance σ^2. A schematic rendering of the setup is shown in figure 2.1.

A quantity of interest is the *rate* R of transmission, given by

$$R = \frac{K}{n}. \tag{2.3}$$

The above gives the ratio of the length of the input string K to the number of transmissions n. Recalling that $K = \log_2 |\mathcal{A}|$, where $|\mathcal{A}|$ is the cardinality of $\mathcal{A}$, another equivalent way of expressing the rate is

$$R = \frac{\log |\mathcal{A}|}{n}. \tag{2.4}$$

Notice that we have suppressed the fact that the logarithm in the above expression is of base 2. The choice of base determines the unit in which the rate is expressed. If logarithm base is 2 then the unit is *bits*, whereas it is *nats* if the base is e.

Ideally, for a given n, one would like to have the message set $\mathcal{A}$ as large as possible. However,

the power constraint restricts the rate R from being arbitrarily large. We make this rigorous in the sequel.

For a given dictionary matrix X, assume that from the received Y the receiver makes an estimate $\hat{j} = \hat{j}(Y, X)$ of the input string j. One way of measuring the performance of the coding scheme is the average probability of error given by,

$$p_{err,X} = \frac{1}{|\mathcal{A}|} \sum_{j \in \mathcal{A}} \mathbb{P}_j(\hat{j} \neq j \,|X), \tag{2.5}$$

where, for the given X matrix, $\mathbb{P}_j(.|X)$ gives the probability distribution of Y if j was sent.

From the following lemma, see for example Cover et al. [15], one infers that the quantity

$$\mathcal{C} = \frac{1}{2} \log(1 + P/\sigma^2), \tag{2.6}$$

also called the *capacity* of the Gaussian channel, is an upper bound on the rate R for reliable communication.

Theorem 1 (converse to Shannon coding theorem)**.** *For any rate $R > 0$,*

$$p_{err,\,X} \geq 1 - \frac{\mathcal{C}}{R} - \frac{1}{nR}. \tag{2.7}$$

From the above one sees that if $R > \mathcal{C}$ then the probability of error is asymptotically, for large values of codelength n, bounded away from 0. However, since it is only a converse result, it doesn't guarantee the existence of codes, for any rate $R < \mathcal{C}$, for which one can make $p_{err,X}$ arbitrarily small.

Notice that, for a given codebook X, the optimal decoding rule, which is the scheme that produces an estimate $\hat{j}$ that minimizes $p_{err,X}$ among all decoders, is the maximum likelihood scheme given by,

$$\hat{j} = \arg\min_{j \in \mathcal{A}} \|Y - X_j\|^2,$$

where $\|.\|$ denotes the usual ℓ_2-norm. Indeed, with the prior on j to be uniform and the likelihood for Y for the given X and j being Gaussian, this is the choice that minimizes the posterior probability of error given Y and X.

To prove the existence of codes achieving rates R arbitrarily close to capacity $\mathcal{C}$, Shannon devised

an random encoding and decoding argument. He took the entries of X to be *i.i.d.* $N(0, P)$. An input string $j \in \mathcal{A}$ is then mapped to the corresponding column of this randomly selected dictionary. Notice that the expectation $E\|X_j\|^2$ is nP, for $j = 1, \ldots, \|\mathcal{A}|$. Thus the power constraint is satisfied in an expected sense rather than the strict sense in (2.2). Using the above random design, Shannon was able to prove that for the maximum likelihood rule, the error probability

$$p_{err} = E(p_{err,X}), \tag{2.8}$$

where the outside expectation is with respect to the distribution of the X matrix, goes to zero exponentially fast (in n) for any fixed $R < C$. Demonstration of small p_{err} implies the existence of a codebook X whose corresponding error probability $p_{err,X}$ is small.

In order to make the individual codewords satisfy the power constraint (2.2), an expurgation argument can be used (see for example Cover et al. [15]), whereby one removes codewords violating (2.2).

Since for the code construction that we discuss in section 2.3 we choose a random codebook, from hereon we control the power only by requiring that $E\|X_j\|^2/n \leq P$, for $j = 1, \ldots, |\mathcal{A}|$. Here the expectation is with respect to the distribution of the codebook X. Assumption that the power constraint be satisfied in this expected sense does not change the information-theoretic limits for reliable communication.

Since the work of Shannon, there have been works characterizing the optimal form of the exponent for the error probability p_{err}, see for example Gallager [19] and Polyanskiy et al. [29]. We summarize these results in the following theorem.

Theorem 2 (Gallager [19], Polyanskiy et al. [29])**.** *For any $R < \mathcal{C}$, one has*

$$p_{err} \leq e^{-n\,\kappa\,\min\{\Delta, \Delta^2\}},$$

where κ is a positive constant and $\Delta = \mathcal{C} - R$. Here p_{err} is calculated using the maximum likelihood rule.

It is still an area of active research to give better characterizations of the optimal error exponent. The above theorem was stated since it serves as a benchmark against which we can compare the error exponents obtained from our regression coding scheme.

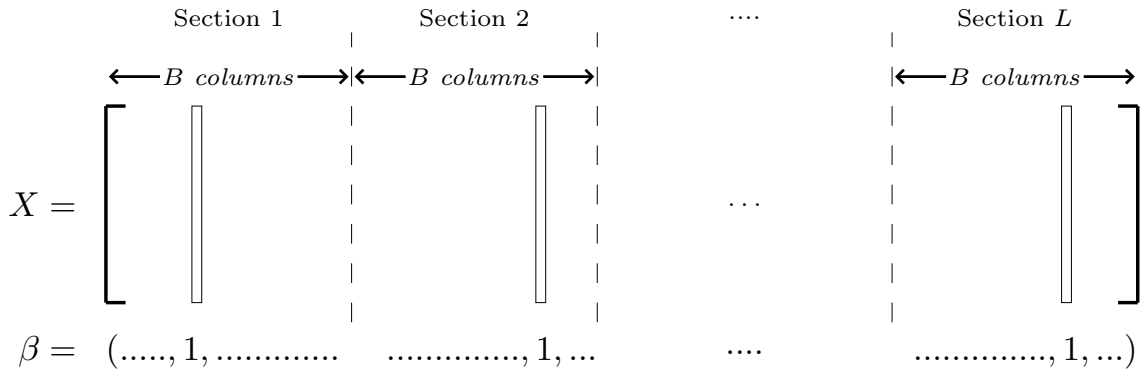

Figure 2.2: Schematic rendering of the dictionary matrix X. The vertical bars in the X matrix indicate the selected columns from a section.

Before discussing the framework for the regression based codes, we mention that Shannon's method of code construction is not a practical approach. The reason for this is that the size of the codebook X is exponential in n, making it too large for storage purposes. To see this, notice that from relation (2.3), the number of columns of the X matrix is 2^{nR}. Correspondingly, for any fixed rate $R > 0$, the size of X, which is $n \times 2^{nR}$ is impractically large.

The above problem arises since each column of X corresponds to a codeword, and, for a fixed rate R, the number of codewords is exponential in n. Our regression codes rectify this by defining codewords to be sparse linear combinations (or superpositions) of elements of a much smaller *dictionary matrix*. Indeed, in chapters 3 and 4, we show that using a dictionary with number of columns that is only polynomial in the number of rows n (or the codelength), one can communicate at rates up to capacity.

2.3 The regression formulation

We now describe the superposition coding scheme. The story begins with the dictionary (design matrix) $X \in \mathbb{R}^{n \times N}$, with columns $X_j \in \mathbb{R}^n$ for $j = 1, 2, \ldots, N$. We further assume that $N = LB$, with L and B being positive integers, and partition the dictionary into L sections, each of size B as depicted in figure 2.2.

The codewords takes the form of particular linear combinations of subsets of columns of the dictionary. Specifically, each codeword is of the form $X\beta$, where $\beta \in \mathbb{R}^N$ belongs to a set $\mathcal{B}$ given

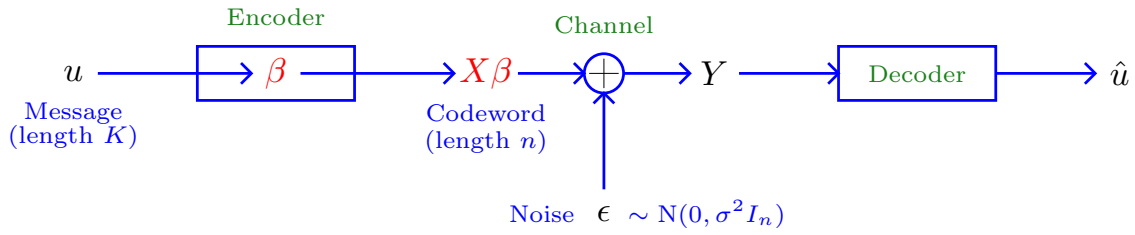

Figure 2.3: The framework for communication using regression.

by

$$\mathcal{B} = \{\beta \in \{0, 1\}^N : \beta_j \text{ has one 1 in each section}\}.$$

Notice that $|\mathcal{B}| = B^L$ For $\beta \in \mathcal{B}$, the codeword $X\beta$ is now a superposition of L columns of X, with exactly one column selected from each section. The quantity L can also be viewed as the sparsity parameter discussed in chapter 1. The received vector is then in accordance with the statistical linear model

$$Y = X\beta + \epsilon \tag{2.9}$$

where ϵ is the noise vector distributed Normal$(0, \sigma^2 I)$.

The entries of X are drawn independently from a normal distribution with mean zero and variance P/L. With this distribution one has that for each $\beta \in \mathcal{B}$, the expected codeword power, given by $\mathrm{E}\|X\beta\|^2/n$, is equal to P. Thus power is controlled in an expected sense rather than the traditional sense of requiring that for each codeword $X\beta$, to have $\|X\beta\|^2/n \leq P$. In section 2.8, we discuss the implications of our power control on the power of individual codewords.

The diagram showing the framework for communication using regression is shown in figure 2.3.

Notice that the above setup reduces to the Shannon's random coding scheme setup, discussed in section 2.2, if we take $L = 1$. However, as mentioned earlier, this forced N to be exponential in the codelength n (or the number of rows of the X matrix) in order to communicate at positive rates. By allowing the sparsity L (or the number of sections) to grow with n, we demonstrate that one can arrange for the dimension N of the dictionary to be only polynomial in n and still communicate at rate arbitrarily close to capacity.

A final piece in the puzzle is the mapping of input bit strings to coefficient vectors β. In other words, we need to find a bijection from the set $\mathcal{A} = \{1, \ldots, 2^K\}$ of input bit strings to the set $\mathcal{B}$

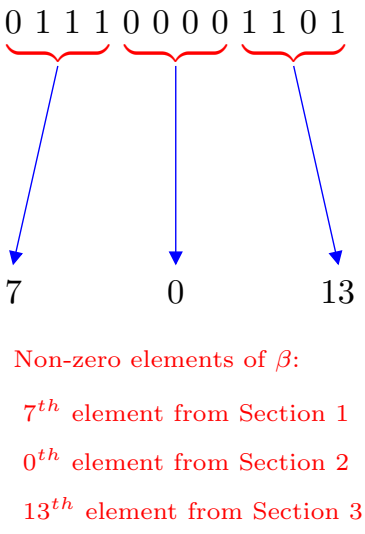

Figure 2.4: Diagram representing mapping of a binary string to a coefficient vector β in $\mathcal{B}$. Assume that $u = (0\ 1\ 1\ 1\ 0\ 0\ 0\ 0\ 1\ 1\ 0\ 1)$ has length $K = 12$. Take $L = 3$ and $\log_2 B = 4$, so that $K = L \log_2 B$. The string splits give the binary address of the index where β is non-zero in a section. Here we assume that the indices are numbered 0 through $B - 1$.

of allowed coefficient vectors. Once this mapping is defined, one can encode the bit string as the corresponding $X\beta$. For Shannon's coding scheme this encoding was trivial – each input bit string was encoded as the corresponding column of the X matrix. Luckily, our partitioned setup allows us to also define the encoding without any hassle.

For convenience, it is assumed that the section size B is a power of 2. Also assume that the input bit string are of length $K = L \log_2 M$. Split this string into L substrings of size $\log_2 B$. The encoder maps u to β simply by interpreting each substring of u as giving the index of which coordinate of β is non-zero in the corresponding section. That is, each substring is the binary representation of the corresponding index. For an input string u, we find it sometimes convenient to write the corresponding β as $\beta(u)$. A diagram providing an example of the mapping $u \to \beta(u)$ is shown in figure 2.4.

As we have said, the rate of the code is $R = K/n$ input bits per channel uses and we arrange

for R arbitrarily close to C. For our code, this rate is

$$R = (L \log B)/n.$$

For specified rate R, the codelength $n = (L/R) \log B$. We take the section size B to be related to the number of sections L by an expression of polynomial size. Consequently, the length n and the number of terms L agree to within a log factor.

Control of the dictionary size is critical to computationally advantageous coding and decoding. If the number of sections L were fixed, then X has size $N = L2^{nR/L}$ that is exponential in n, making its direct use impractical. Instead, with L agreeing with n to within a log-factor, the dictionary size is more manageable. In this setting, we construct reliable, high-rate codes with codewords corresponding to linear combinations of subsets of terms in moderate size dictionaries.

The idea of superposition codes for Gaussian channels began with Cover [14] in the context of determination of the capacity region of certain multiple user channels. There L represents the number of messages decoded and a selected column represents the codeword for a message. Codes for the Gaussian channel based on sparse linear combinations have been proposed in the compressed sensing community by Tropp [36]. However, as he discusses, the rate achieved there is not up to capacity. Relationship of our work to that in these communities will is discussed in further detail in section 3.3.

2.4 Capacity achieving decoders

Two decoders are analyze in Chapters 3 and 4. In Chapter 3 we analyze the maximum likelihood decoder, given by,

$$\hat{\beta} = \arg\min_{\beta \in \mathcal{B}} \|Y - X\beta\|^2.$$

This is the optimal decoder since, for any given X matrix, it minimizes the average probability of error,

$$p_{err,X}(\tilde{\beta}) = \frac{1}{|\mathcal{B}|} \sum_{\beta \in \mathcal{B}} \mathbb{P}_\beta(\tilde{\beta} \neq \beta | X)$$

over all possible choices of decoders $\tilde{\beta}$. Here $\mathbb{P}_\beta(|X)$ gives the distribution of Y, assuming β was sent.

Although the above estimator is optimal, it is not computationally feasible since its evaluation

entails a search over all the B^L (or 2^{nR}) vectors in $\mathcal{B}$. This is impractical since the size of the search set is exponential in n. Accordingly, in Chapter 4 we pursue the problem of achieving capacity using computationally feasible schemes. The scheme we analyze identifies correct terms through an *iterative thresholding* algorithm. This is similar in spirit to iterative decoding techniques such as forward stepwise regression [22], [7] and orthogonal matching pursuit [26][28].

2.5 Variants of the regression scheme

We also call our scheme based on the regression model, the *sparse superposition* coding scheme. To distinguish it from other sparse superposition codes, the code analyzed here may be called a *partitioned superposition* code. The motivations for introducing the partitioning versus arbitrary subsets, in the superposition coding scheme, are the ease in mapping the input bit string to the coefficient vector and the ease in composition with the outer Reed-Solomon code. Natural variants of the schemes are *subset superposition* coding, where one arranges for a number L of the coordinates to be non-zero and taking the value 1, with the message conveyed by the choice of subset. With somewhat greater freedom one may have *signed superposition* coding, where one arranges the non-zero coefficients to be $+1$ or -1. Then the message is conveyed by the sequence of signs as well as the choice of subset. In both cases if one takes the elements of X to be i.i.d $N(0, P/L)$ as before, then the expected power of each codeword is P. The signed superposition coding scheme has been proposed in Tropp [36], Gilbert and Tropp [20].

As mentioned earlier, superposition codes began with Cover [14] for multi-user channels in the context of determination of the capacity region of Gaussian broadcast channels. There the number of users corresponds to L. The codewords for user ℓ, for $\ell = 1, \ldots, L$, corresponds to the columns in section ℓ. In that setting what is sent is the sum of codewords, one from each user. With L fixed, $B = 2^{nR/L}$ is exponential in L. Here for the single user channel, by allowing L to be of the same order as n, to within a log factor, we make it possible to achieve rates close to capacity with polynomial size dictionaries.

2.6 Block, bit, and section error probabilities

Hereon we focus on the *partitioned superposition* coding scheme discussed in section 2.3. Recall that we denote $\mathcal{A} = \{1, \ldots, 2^K\}$ as the set of input bit strings, and $\mathcal{B}$ as the corresponding set of

coefficient vectors. Using $K = L \log_2 B$, one sees that $|\mathcal{A}| = |\mathcal{B}| = M^L$.

Ideally one would like to produce a bound on the error probability $p_{err,X}$, given in (2.5), conditioned on the given X matrix. However, since bounding the above is difficult, we follow Shannon theory tradition and try to bound the simple quantity p_{err}, given in (2.8), obtained after averaging over the distribution of the X matrix.

There are two equivalent ways of expressing p_{err}. The first way, as in (2.8), is to write p_{err} as the average probability of the error event that $\hat{j}$, the estimate of the input string, is not equal to the input string j. This may also be written as,

$$p_{err} = \frac{1}{|\mathcal{A}|} \sum_{j \in \mathcal{A}} \mathbb{P}_j(\hat{j} \neq j), \tag{2.10}$$

where $\mathbb{P}_j(.) = E\mathbb{P}_j(.|X)$, with $\mathbb{P}_j(.|X)$ given by (2.5). Another equivalent representation of p_{err} is seen by noting that each input bit string corresponds to a coefficient vector β in $\mathcal{B}$, and vice versa. Correspondingly,

$$p_{err} = \frac{1}{|\mathcal{B}|} \sum_{\beta \in \mathcal{B}} \mathbb{P}_\beta(\hat{\beta} \neq \beta), \tag{2.11}$$

where $\mathbb{P}_\beta(.)$ is interpreted analogously.

The quantity p_{err} is called the *block error probability*. Theorem 2 stated bounds on this probability for the Shannon random coding scheme. Ideally, one would like to get similar exponentially small error probabilities for our regression scheme. However, we are fairly convinced this is not possible. The reason for this is the relatively small magnitude, of order $1/\sqrt{L}$, of the non-zeroes of β. As a result, it is not possible for any algorithm, practical or otherwise, to distinguish between a β and β' belonging in $\mathcal{B}$, when they differ in only a few sections. Accordingly, instead of controlling the block error probability we control a less stringent *bit error* probability, which we now describe.

Define the *bit error rate* $d(u, \hat{u})$, between the input string u and its estimate $\hat{u}$, as

$$d(u, \hat{u}) = \frac{1}{K} \sum_{i=1}^{K} I_{\{u_i \neq \hat{u}_i\}}, \tag{2.12}$$

which is the fractions of positions where u and $\hat{u}$ differ. Denoting j, $\hat{j}$ as the elements in $\mathcal{A}$ corresponding to u, $\hat{u}$ respectively, with a slight abuse of notation, we also denote the bit error rate as $d(j, \hat{j})$.

For a given fraction of mistakes α_0, the corresponding *bit error probability* is given by,

$$p_{err,\alpha_0} = \frac{1}{|\mathcal{A}|} \sum_{j \in \mathcal{A}} \mathbb{P}_j(d(j, \hat{j}) > \alpha_0). \tag{2.13}$$

Notice that $p_{err,0}$ is simple the block error probability. A small bit error probability ensures that, with high probability, the estimated string $\hat{u}$ does not differ from u in too many positions. More precisely, with high probability, the strings are different in at most α_0 fraction of positions.

Instead of directly showing that the probability in (2.13) is small, we find it more convenient to bound the probability of high *section error rate*, which is the fraction of sections where mistakes are made. In particular, denoting *mistakes* as the number of sections where our estimate $\hat{\beta}$ differs from the true β, let

$$\mathcal{E}_{\alpha_0} = \{mistakes \geq \alpha_0 L\}. \tag{2.14}$$

In our analysis we bound the probability that the section error rate is greater that α_0, given by,

$$\bar{\mathbb{P}}(\mathcal{E}_{\alpha_0}) = \frac{1}{B^L} \sum_{\beta \in \mathcal{B}} \mathbb{P}_\beta(\mathcal{E}_{\alpha_0}). \tag{2.15}$$

We call the above the *section error probability*.

Notice that if β and $\hat{\beta}$ differ in α fraction of sections then the corresponding u and $\hat{u}$ differ in *at most* α fraction of positions. Accordingly,

$$\{d(u, \hat{u}) > \alpha_0\} \subseteq \mathcal{E}_{\alpha_0}.$$

Correspondingly,

$$p_{err,\alpha_0} \leq \bar{\mathbb{P}}[\mathcal{E}_{\alpha_0}].$$

Thus a bound on the section error probability also translates to a bound on bit error probability.

2.7 Control on block error from section error

Assume that we are able to demonstrate that the section error probability $\bar{\mathbb{P}}[\mathcal{E}_{\alpha_0}]$ is exponentially small. Here, we devise a scheme that has low block error probability as well. At a high-level, the idea is that instead of encoding the input string u (of length K) directly as the coefficient vector

β, one encodes it first as a string whose elements belong to Galois field of B elements. Each such encoded string may be viewed as binary string $\tilde{u}$, of length $\tilde{K} > K$. This encoding, between bit strings of length K and those of length $\tilde{K}$, satisfies the following:

For any two distinct bit strings of length K, say u, u', the corresponding encoded strings $\tilde{u}$, $\tilde{u}'$ of length $\tilde{K}$, satisfy,

$$d(\tilde{u}, \tilde{u}') \geq 2\alpha_0.$$

In other words, any two distinct encoded strings differ in at least $2\alpha_0 \tilde{K}$ positions. We employ Reed-Solomon (RS) codes ([30], [25]) to do this. The encoded string $\tilde{u}$ is then mapped to the coefficient vector $\beta = \beta(\tilde{u})$ in the usual manner. A more detailed description of Reed-Solomon codes as well as the encoding of strings $\tilde{u}$ to coefficients vectors $\beta(\tilde{u})$ is given below.

It turns out that any two distinct coefficient vectors, encoded in such a manner, differ in at least $2\alpha_0$ fraction of sections. In other words, if

$$\tilde{\mathcal{B}} = \{\beta(\tilde{u}) : \tilde{u} \text{ is a Reed-Solomon encoding of a bit string } u \text{ of length } K\},$$

then any two distinct β and β' in $\tilde{\mathcal{B}}$ differ in at least $2\alpha_0$ fraction of sections.

We now describe how we obtain low block error probability from a demonstration that the section error probability is small. Notice that since for a given α_0, the section error probability $\bar{\mathbb{P}}[\mathcal{E}_{\alpha_0}]$ is small, the estimate $\hat{\beta}$ differs from the true β, now in the set $\tilde{\mathcal{B}}$ instead of $\mathcal{B}$, in at most α_0 sections, with high probability. We remind that the estimate $\hat{\beta}$ need not belong to $\tilde{\mathcal{B}}$. Clearly, one can identify β exactly from the estimate $\hat{\beta}$ by selecting the unique β in $\tilde{\mathcal{B}}$ for which the number of sections in which β and $\hat{\beta}$ differ is the least. This follows from the fact that any two distinct β and β' in $\tilde{\mathcal{B}}$ differ in at least $2\alpha_0$ fraction of sections. Also notice that the bound on the block error probability, when the coefficients lie in $\tilde{\mathcal{B}}$, obtained in this fashion is the same as the bound on the section error probability $\bar{\mathbb{P}}[\mathcal{E}_{\alpha_0}]$.

For any fixed α_0, the effective communication rate is reduced as a result of the above approach. The larger the α_0, the greater is the reduction. To quantify the effective communication rate, we now describe in greater detail encoding using the above scheme.

We call the Reed-Solomon (RS) code the *outer code* in our scheme. The superposition code corresponds to the *inner code*. The symbols for the RS code come from a Galois field consisting of q elements denoted by $GF(q)$, with q typically taken to be of the form 2^m. If K_{out}, n_{out} represent

message and codeword lengths respectively, then an RS code with symbols in $GF(2^m)$ and minimum distance between codewords given by d_{RS} can have the following parameters:

$$n_{out} = 2^m \tag{2.16}$$

$$n_{out} - K_{out} = d_{RS} - 1 \tag{2.17}$$

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 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 in each of the L sections. The number of input symbols is then $K_{out} = L - d_{RS} + 1$, so setting $2\alpha_0 = d_{RS}/L$, one sees that the outer rate R_{out}, equals $1 - 2\alpha_0 + 1/L$ which is at least $1 - 2\alpha_0$.

For code composition $K_{out} \log_2 B$ message bits become the K_{out} input symbols to the outer code. The symbols $j_1, j_2, \ldots, j_{n_{out}}$ of the outer codeword, having length $n_{out} = L$, with each term in the codeword coming from a symbol set consisting of B values. Thus the codeword can be described by $\tilde{K} = n_{out} \log_2 B$ message bits. The rate of the code is $R_{out} = K/\tilde{K}$.

These symbols of the outer code gives the labels of terms sent from each section using our inner superposition with codelength $n = n_{inner} = L(\log_2 B)/R_{inner}$. From the received Y the estimated labels $\hat{j}_i, \hat{j}_2, \ldots \hat{j}_L$ using our decoder can be again thought of as output symbols for our RS codes. If $\hat{\alpha}_0 = mistakes/L$ denotes the section mistake rate, it follows from the distance property of the outer code that if $\hat{\alpha}_0 \leq \alpha_0$, then these errors can be corrected.

The overall rate $R = R_{comp}$ is seen to be equal to the product of rates $R_{out} R_{inner}$. To see note that $R = K/n$. Using $K = K_{out} \log_2 B$ and $n = n_{inner}$, one has

$$\begin{aligned} R &= \frac{K}{n_{out} \log_2 B} \quad \frac{n_{out} \log_2 M}{n} \\ &= \frac{K}{\tilde{K}} \quad \frac{n_{out} \log_2 M}{n} \\ &= R_{out} \quad \frac{n_{out} \log_2 M}{n} \end{aligned}$$

Now use the fact that $n_{out} = L$, and $n = L(\log_2 B)/R_{inner}$ from which one gets,

$$R = R_{out}\, R_{inner}.$$

Since we arrange for $\hat{\alpha}_0$ to be smaller than some α_0 with exponentially small probability ϵ, 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 $(1 - 2\alpha_0)R_{inner}$.

The case when $L < B$ can be dealt with by observing ([25], 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 easily 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 outer codelength is L and symbols of this code are in $GF(B)$, the code composition can be carried out as before.

The above gives us a scheme with low block error probability from a scheme that has a low section error probability, with the same bound on error probability. The effective rate of communication $(1 - 2\alpha_0)R_{inner}$ depends on the section mistake rate as well as the rate of the inner superposition. Correspondingly, from hereon our results will mainly focus on getting good controls on the section error probability.

2.8 Control of power

Recall that our codewords take the form $X\beta$, with β in $\mathcal{B}$. For an $x \in \mathbb{R}^n$, denote as $|x|^2 = \|x\|^2/n$ the normalized sum of squares. Here we investigate the distribution of codeword powers, which is the distribution of $|X\beta|^2$ when β varies in $\mathcal{B}$, from our power control given by $E|X\beta|^2 = P$, for each $\beta \in \mathcal{B}$. More specifically, we investigate the average codeword power,

$$P_{avg} = \frac{1}{B^L} \sum_{\beta \in \mathcal{B}} |X\beta|^2 \tag{2.18}$$

and the worst case codeword power given by,

$$P_{max} = \max_{\beta \in \mathcal{B}} |X\beta|^2 \tag{2.19}$$

We first analyze P_{avg}. With the non-zero indices of β in a section chosen uniformly from the B possible choices, P_{avg} can we viewed as the expectation of $|X\beta|^2$ with respect to this distribution of β. Correspondingly, using the rule that an expected square is the square of the expectation plus the variance, one gets that,

$$P_{avg} = \sum_{\ell=1}^{L} \sum_{j \in section\ \ell} \frac{|X_j - \bar{X}_\ell|^2}{B} + |\sum_{\ell=1}^{L} \bar{X}_\ell|^2, \tag{2.20}$$

where $\bar{X}_\ell$ is the average of the columns in section ℓ. Using the independence of $\bar{X}_\ell$ and $(X_j - \bar{X}_\ell : j \in section\ \ell)$, we have that the first term in the above expression is $P/(LBn)$ times a Chi-square random variable with $nL(B-1)$ degrees of freedom, and the second term is $P/(nB)$ times an independent Chi-square random variable with n degrees of freedom.

We use the following inequality for the concentration of Chi-squares random variables, see for example Donoho [17]. For any $h > 0$,

$$\mathbb{P}(\mathcal{X}_n^2 \geq n(1+h)^2) \leq e^{-nh^2/2}. \tag{2.21}$$

Here $\mathcal{X}_n^2$ denotes a Chi-squared random variable with n degrees of freedom.

Using the above we have

$$\mathbb{P}(P_{avg} > P(1+h)^2) \leq 2e^{-nh^2/2}. \tag{2.22}$$

To see this, denote the first and second terms in (2.20) as W_1 and W_2 respectively. Using (2.21) one has

$$\mathbb{P}\left(W_1 > P\frac{B-1}{B}(1+h)^2\right) \leq \exp\{-nL(B-1)h^2/2\},$$

the right side of which is bounded by $e^{-nh^2/2}$. Also,

$$\mathbb{P}\left(W_2 > (P/B)(1+h)^2\right) \leq \exp\{-nh^2/2\}.$$

Correspondingly, (2.22) follows from using a union bound, along with $P_{avg} = W_1 + W_2$.

Taking $h_\epsilon = \sqrt{(2/n)\log(2/\epsilon)}$, from (2.22) one sees that P_{avg} is less that $P(1+h_\epsilon)^2$ with probability at least $1-\epsilon$. Thus for fixed error probability ϵ, and for large n, the average power P_{avg} is not much larger than P with high probability .

Next, we investigate P_{max}. The simplest distribution bound is to note that for each β, the codeword $X\beta$ is distributed as a random vector with independent $N(0,P)$ coordinates. Accordingly, $|X\beta|^2$ is P/n times a Chi-square n random vector. There are e^{nR} such codewords, with the rate written in nats. Using (2.21) and a union bound, one gets

$$\mathbb{P}(P_{max} > P(1+h)^2) \leq e^{nR} e^{-nh^2/2}.$$

Correspondingly, taking $h = \tilde{h}_\epsilon$, where

$$\tilde{h}_\epsilon = \sqrt{2}\sqrt{R + (1/n)\log(1/\epsilon)},$$

we sees that P_{max} is bounded by $P(1+\tilde{h}_\epsilon)^2$, except on a set of probability at most ϵ.

Notice that unlike h_ϵ, the quantity $\tilde{h}_\epsilon$ does not become small for large n. According to this characterization, one does not have the norms $|X\beta|^2$ being uniformly close to their expectation from our expected power control.

Chapter 3

The Least Squares decoder

3.1 Introduction

Here we analyze the maximum likelihood decoder. This decoder is the same as that which chooses the β that maximizes the posterior probability when the prior distribution is uniform over $\mathcal{B}$. The decoder is given by,

$$\hat{\beta} = \arg\min_{\beta \in \mathcal{B}} \|Y - X\beta\|^2 \tag{3.1}$$

where $\|.\|$ denotes the euclidean norm. Here we implicitly assume that if the minimization has a non-unique solution, one may take $\hat{\beta}$ to be any value in the solution set. Since the above is a least squares minimization problem over coefficient vectors in $\mathcal{B}$, we also call this the *least squares decoder*. Although the above decoder is not a computationally feasible scheme, the result is significant since we show that one can achieve rates up to capacity with a codebook that has a compact representation in the form of the dictionary X. Recall that the entries of X are drawn i.i.d $N(0, P/L)$.

3.2 Main result

We now describe our main result concerning the performance of the least squares decoder. We show that if $B = L^a$, for any a exceeding a particular positive function of the signal-to-noise ratio v, then rates arbitrarily close to capacity can be achieved. This function is near $16/v^2$ for small v and near 1 for large v. Consequently, the dictionary has size $N = L^{a+1}$ that is polynomial in L. This required section size does not depend on the gap $C - R$ and thus the dictionary has a compact

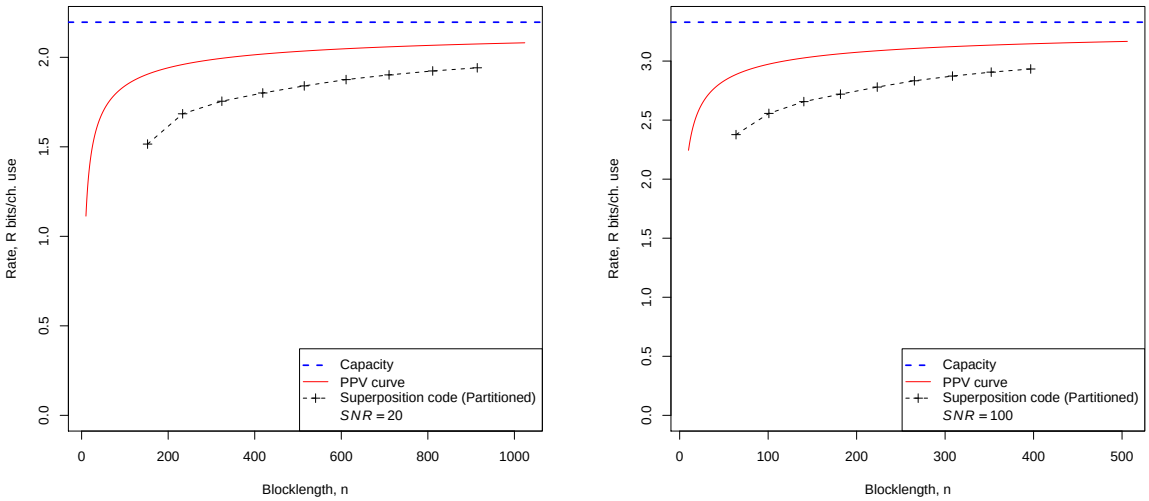

Figure 3.1: Plots of comparison between achievable rates using our scheme and the theoretical best possible rates for *block error* probability of 10^{-4} and signal-to-noise ratio (v) values of 20 and 100. The curves for our partitioned superposition code were evaluated at points with number of sections L ranging from 20 to 100 in steps of 10, with corresponding B values taken to be L^{a_v}, where a_v is as given in Lemma 7, equations (3.32), (3.33) later on. For the v values of 20 and 100 shown above, a_v is around 2.6 and 1.6 respectively. For details on computations please refer to appendix 3.D.

representation irrespective of the closeness of R to C.

We restate the relevant notation introduced in section 2.6 of chapter 1. For $\beta \in \mathcal{B}$, let $\mathbb{P}_\beta(.)$ denote the joint distribution of Y, X given β. Further, let *mistakes* denote the number of mistakes made by the least squares decoder, that is, the number of sections in which the position of the non-zero term in $\hat{\beta}$ is different from that in the true β. Denote the error event

$$\mathcal{E}_{\alpha_0} = \{mistakes \geq \alpha_0 L\} \tag{3.2}$$

that the decoder makes mistakes in at least α_0 fraction of sections. Assuming that β is drawn from a uniform distribution over all B^L elements from $\mathcal{B}$, the average probability of error conditional on X is given by,

$$\bar{\mathbb{P}}\left[\mathcal{E}_{\alpha_0}|X\right] = \frac{1}{B^L}\sum_{\beta\in\mathcal{B}} \mathbb{P}_\beta[\mathcal{E}_{\alpha_0}|X].$$

Deriving bounds for the above is not easy. We follow the information theory tradition and bound

the average of the above over the distribution of X, given by,

$$\bar{\mathbb{P}}[\mathcal{E}_{\alpha_0}] = \mathbb{E}_X \bar{\mathbb{P}}[\mathcal{E}_{\alpha_0}|X]. \tag{3.3}$$

For positive x, let $g(x) = \sqrt{1+4x^2} - 1$. Further, for $R \leq C$, let

$$w_v = \frac{v}{[4(1+v)^2]\sqrt{1+(1/4)v^3/(1+v)}} \tag{3.4}$$

A positive expression $a_{v,L}$ possessing properties explained in section 3.6, lemma 7 is used. For large L it is near a function a_v near $16/v^2$ for small v and near 1 for large v. Our main result is the following.

Proposition 3. *Assume $B = L^a$, where $a \geq a_{v,L}$, and rate R is less than capacity C. Let α_0 represents the fraction of section mistakes made by the least squares decoder. Then,*

$$\bar{\mathbb{P}}[\mathcal{E}_{\alpha_0}] = e^{-nE(\alpha_0,R)}$$

with $E(\alpha_0, R) \geq h(\alpha_0, C-R) - (\log 2L)/n$, where

$$h(\alpha, \Delta) = \min\left\{\alpha w_v \Delta,\ \frac{1}{4} g\left(\frac{\Delta}{2\sqrt{v}}\right)\right\} \tag{3.5}$$

is evaluated at $\alpha = \alpha_0$ and $\Delta = C - R$.

Proposition 3 is proved in Section 3.7.

Remark: It is shown in appendix 3.C that the exponent $E(\alpha_0, R)$ can be improved by replacing $h(\alpha_0, C-R)$ with $\tilde{h}(\alpha_0, C-R)$ where,

$$\tilde{h}(\alpha, \Delta) = \min\left\{c_{\alpha,v}\,\alpha,\ \frac{1}{4} g\left(\frac{\Delta}{2\sqrt{v}}\right)\right\}.$$

Here $c_{\alpha,v}$ is a positive function of α and v, which for given v is near $\tau_v \tilde{w}_v/4$ for small α, where τ_v, $\tilde{w}_v$ are positive expressions as in (3.42), (3.43) later on.

Let $g^*(x) = \min\{\sqrt{2}x, x^2\}$. Then it is not hard to see that

$$g(x) \geq g^*(x) \quad \text{for all} \quad x \geq 0. \tag{3.6}$$

Accordingly, the function $g(.)$, appearing in the lower bound (3.5), may be replaced by $g^*(.)$, revealing that the exponent is, up to a constant, of the form $\min\{\alpha_0\Delta, \Delta^2\}$, where $\Delta = C - R$. With the improved bound in appendix 3.C, it is of the form $\min\{\alpha_0, \Delta^2\}$.

Further, using the technique of composition with an outer code described in section 2.7, we state the proposition for exponentially small block error probability for any $R < \mathcal{C}$.

Proposition 4. *For given positive error probability ϵ and fraction of mistakes α_0, let R be the rate for which the partitioned superposition code with L sections has*

$$\bar{\mathbb{P}}\{\mathcal{E}_{\alpha_0}\} \leq \epsilon.$$

Then through concatenation with an outer Reed-Solomon code, one obtains a code with rate $(1 - 2\alpha_0)R$ and block error probability less than or equal to ϵ.

The proof of the above is immediate from the description of the scheme involving composition with the outer Reed-Solomon code described in section 2.7.

Particular interest is given to the case that the rate R is made to approach the capacity C. Arrange $R = C - \Delta_n$ and $\alpha_0 = \Delta_n$. One may let the rate gap Δ_n tend to zero (e.g. at a $1/\log n$ rate or any polynomial rate not faster than $1/\sqrt{n}$), then the overall rate $R_{tot} = (1 - 2\alpha_0)(C - \Delta_n)$ continues to have drop from capacity of order Δ_n, with the composite code having block error probability of order

$$\exp\{-nc\,\Delta_n^2\}.$$

The exponent above, of order $(C-R)^2$ for R near C, is in agreement with the form of the optimal reliability bounds as in [19], [29], though here our constant c is not demonstrated to be optimal.

In Figure 3.1 we plot curves of achievable rates using our scheme for block error probability fixed at 10^{-4} and signal to noise ratios of 20 and 100. We also compare this to a rate curve given in Polyanskiy, Poor and Verdu [29] (PPV curve), where it is demonstrated that for a Gaussian channel with signal to noise ratio v, the block error probability ϵ, codelength n and rate R with an optimal

code can be well approximated by the following relation,

$$R \approx C - \sqrt{\frac{V}{n}} Q^{-1}(\epsilon) + \frac{1}{2}\frac{\log n}{n} \tag{3.7}$$

where $V = (v/2)(v+2)\log^2 e/(v+1)^2$ is the channel dispersion and Q^{-1} is the inverse normal cdf.

For the superposition code curve, the y-axis gives the highest R_{comp} for which the error probability stays below 10^{-4}. We see for the given v and block error probability values, the achievable rates using our scheme are reasonably close to the theoretically best scheme. Note that the PPV curve was computed with an approach that uses a codebook of size that is exponential in blocklength, whereas our dictionary, of size LB, is of considerably smaller size.

Section 3.4 contains brief preliminaries. Section 3.5 provides core lemmas on the reliability of least squares for our superposition codes. Section 3.6 analyzes the matter of section size sufficient for reliability. In section 3.7 we give proofs propositions 3 and in section 3.8 we discuss how the our results can be adapted for an approximate form of the least squares decoder. The appendix collects some auxiliary matters.

3.3 Related work on sparse signal recovery

While reviewing works on sparse signal recovery and compressed sensing, we adhere to our notation that we have a linear model of the form

$$Y = X\beta + \epsilon$$

where $X \in \mathbb{R}^{n \times N}$ is a deterministic or random matrix and $\beta \in \mathbb{R}^N$ has exactly L non-zero values. The quantities n, N, L and β will be called parameters for the model. In our description below we denote as *const* some positive constant whose value will change from time to time.

The conclusions here complement recent work on sparse signal recovery in the linear model setup as we now discuss. In a broad sense, these works analyze for various schemes (practical or otherwise), conditions on the parameters so that certain reliability requirements are satisfied with high probability. Closely related to our work is the requirement that only the indices corresponding to the non-zero elements of β, that is the support of β, be recovered exactly or almost exactly. The

most typical assumption on β is that it belongs to a set

$$\mathcal{B}' = \{\beta : \beta \text{ has } L \text{ non-zeroes with magnitude at least } \beta_{min}\}, \tag{3.8}$$

where β_{min} is a positive quantity. Notice that unlike ours, the values of the non-zeroes are unknown in these works. The only assumption they make is that the non-zeroes are least a positive value β_{min}. This added flexibility in coefficient vectors proves to be a major hinderance in establishing precise statements regarding achieved rate in these works.

Also note that for our scheme $n = (L/R)\log B$, which using $L = B^a$ and $N = LB$, one gets that $n = [a/R(a+1)]\, L \log N$ is sufficient for subset recovery. Correspondingly, n needs to be of the order of $L \log B$ for communication at positive rates.

In this document, in order to achieve rates arbitrarily close capacity we require $N = L^{a+1}$, with precise values of a specified later on, putting us in the sub-linear sparsity regime, that is $L/N \to 0$ as L, $N \to \infty$. Also, if we change the scale and take the elements of the X matrix as i.i.d standard normal, the non-zero values of β assume the value $\sqrt{P/L}$. Accordingly, although most the claims in this area are for more general sparsity regimes and values of β, the results most relevant to us are those for the sub-linear sparsity regime and when the non-zero β_j's are at least $const/\sqrt{L}$.

We mention there are works on computationally feasible algorithms such as Lasso and Orthogonal Matching Pursuit which, when applied to our setting, do demonstrate that communication at positive rates is indeed possible using these codes. These are reviewed in section 4.6. In this section we focus on works closely related to our Least Squares decoding and also converse results for support recovery.

Support recovery of a least squares decoder is analyzed by Wainwright [38], and Akçakaya and Tarokh [1] for Gaussian X matrices, where [1] also addresses the issue of partial support recovery. One can infer from these that communication at positive rates is possible using random designs. However, since the values of the non-zeroes are unknown, the least square decoder considered in these works involves not only an exhaustive search over all subsets of size L but also, for a selected subset, one needs to estimate the coefficient vector for the regression of Y on the selected subset. Since the signal recovery purpose is somewhat different here from our communications purpose, in that the work typically does not constrain the non-zero coefficients to the same value, the resulting freedom in their values lead to order of magnitude conclusions that obstruct interpretation in terms of exact rate.

There are also results giving converse results, for exact support recovery [38], and for partial support recovery [1]. These results, which are of the same flavor as theorem 1, gives lower bounds on the sample size n for recovery of the support. Both these agree in terms of order of magnitude, requiring an order of $L \log N$ for the regime we deal with. In Reeves and Gastpar [31] it is shown that in the linear sparsity regime, that is, when L is of the same order as N, one requires $n \geq const N$ for reliable recovery of the support. An implication of this is that the sub-linear sparsity regime is necessary for communication at positive rates.

Consequently one can infer, from some of the aforementioned works, that communication at positive rates is possible with sparse superposition codes. Section 4.6 gives details on the performance of practical schemes. We add to the existing literature by showing that one can achieve any rate up to capacity in certain sparsity regimes with a compact dictionary, albeit for a computationally infeasible scheme. Further we demonstrate that the error exponents are of the optimal form.

3.4 Preliminaries

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 $a \cdot b = (1/n)\sum_{i=1}^n a_i b_i$ be the associated inner product. It is a matter of taste, but we find it slightly more convenient to work henceforth with the norm $|a|$ rather than $\|a\|$.

Concerning the base of the logarithm (log) and associated exponential (exp), base 2 is most suitable for interpretation and base e most suitable for the calculus. For instance, the rate $R = (L \log B)/n$ is measured in bits if the log is base 2 and nats if the log is base e. Typically, conclusions are stated in a manner that can be interpreted to be invariant to the choice of base, and base e is used for convenience in the derivations.

We make repeated use of the following moment generating function and its associated Cramer-Chernoff large deviation exponent in constructing bounds on error probabilities. If Z and $\tilde{Z}$ are normal with means equal to 0, variances equal to 1, and correlation coefficient ρ, then

$$\mathrm{E}(e^{(\lambda/2)(Z^2-\tilde{Z}^2)}) = 1/[1-\lambda^2(1-\rho^2)]^{1/2} \tag{3.9}$$

when $\lambda^2 < 1/(1-\rho^2)$ and infinity otherwise. So, taking the logarithm, the associated cumulant generating function of $(1/2)(Z^2 - \tilde{Z}^2)$ is $-(1/2)\log(1-\lambda^2(1-\rho^2))$, with the understanding that

the minus log is replaced by infinity when λ^2 is at least $1/(1-\rho^2)$. For positive Δ we define the quantity $D = D(\Delta, 1-\rho^2)$ given by

$$D = \max_{\lambda \geq 0} \big\{ \lambda\Delta + (1/2)\log(1 - \lambda^2(1-\rho^2)) \big\}. \tag{3.10}$$

The expression corresponding to D but with the maximum restricted to $0 \leq \lambda \leq 1$ is denoted $D_1 = D_1(\Delta, 1-\rho^2)$, that is,

$$D_1 = \max_{0 \leq \lambda \leq 1} \big\{ \lambda\Delta + (1/2)\log(1 - \lambda^2(1-\rho^2)) \big\}. \tag{3.11}$$

When the optimal λ is strictly less than 1, the value of D_1 matches D as given above.

The $\lambda = 1$ case occurs when $1 + 4\Delta^2/(1-\rho^2) \geq (1+2\Delta)^2$, or equivalently $\Delta \geq (1-\rho^2)/\rho^2$. Then the exponent is $D_1 = \Delta + (1/2)\log\rho^2$, which is as least $\Delta - (1/2)\log(1+\Delta)$. Consequently, in this regime D_1 is between $\Delta/2$ and Δ. The special case $\rho^2 = 1$ is included with $D_1 = \Delta$.

There is a role for the function

$$C_\alpha = \frac{1}{2}\log(1 + \alpha v) \tag{3.12}$$

for $0 \leq \alpha \leq 1$, where $v = P/\sigma^2$ is the signal-to-noise ratio and $C_1 = C = (1/2)\log(1+v)$ is the channel capacity. We note that $C_\alpha - \alpha C$ is a non-negative concave function equal to 0 when α is 0 or 1 and strictly positive in between. The quantity $C_\alpha - \alpha R$ is larger by the additional amount $\alpha(C - R)$, positive when the rate R is less than the Shannon capacity C.

Remark on average codeword power: The average codeword power $B^{-L}\sum_{\beta \in \mathcal{B}} |X\beta|^2$ has expectation with respect to X that matches $\mathrm{E}|X\beta|^2 = P$, for all $\beta \in \mathcal{B}$. The distribution of the average codeword power is tightly concentrated around P as explained in the appendix of [4], and will not be explored further here.

3.5 Performance of Least Squares

In this section we examine the performance of the least squares decoder (3.1) in terms of rate and reliability. For $\beta \in \mathcal{B}$, let $S(\beta) = \{j : \beta_j = 1\}$ denote the set of indices for which β is non-zero. Further, let

$$\mathcal{A} = \{S(\beta) : \beta \in \mathcal{B}\} \tag{3.13}$$

denote the set of allowed subset of terms. It corresponds to the B^L subsets of $\{1, \ldots, N\}$ of size L and comprising of exactly one term from each section.

Recall that we are interested in bounding $\bar{\mathbb{P}}[\mathcal{E}_{\alpha_0}]$ given in (3.3). By symmetry,

$$\bar{\mathbb{P}}[\mathcal{E}_{\alpha_0}] = \mathbb{P}_\beta[\mathcal{E}_{\alpha_0}] \quad \text{for all} \quad \beta \in \mathcal{B},$$

where $\mathbb{P}_\beta[\mathcal{E}_{\alpha_0}] = \mathrm{E}_X \mathbb{P}_\beta[\mathcal{E}_{\alpha_0}|X]$. Correspondingly, for fixed $\beta^* \in \mathcal{B}$, we proceed to obtain bounds for $\mathbb{P}_{\beta^*}[\mathcal{E}_{\alpha_0}]$. Let $S^* = S(\beta^*)$. Further, let $\hat{\beta}$ be the least squares solution (3.1) and $\hat{S} = S(\hat{\beta})$. Notice that $mistakes = card(\hat{S} - S^*)$, which is also the number of sections incorrectly decoded.

For $\ell \in \{1, 2 \ldots, L\}$, let $E_\ell = \{mistakes = \ell\}$ be the event that there are exactly ℓ mistakes. Now $\mathcal{E}_{\alpha_0}$ can be expressed as a disjoint union of E_ℓ, for $\ell \geq \alpha_0 L$. Correspondingly,

$$\mathbb{P}_{\beta^*}[\mathcal{E}_{\alpha_0}] = \sum_{\ell \geq \alpha_0 L} \mathbb{P}_{\beta^*}[E_\ell]. \tag{3.14}$$

In the next two lemmas we give bounds for $\mathbb{P}_{\beta^*}(E_\ell)$ for $\ell = 1, \ldots, L$.

Lemma 5. *Set $\alpha = \ell/L$ for an $\ell \in \{1, 2, \ldots, L\}$. The probability $\mathbb{P}_{\beta^*}(E_\ell)$ can be bounded by $err_1(\alpha)$, where*

$$err_1(\alpha) = \binom{L}{\alpha L} \exp\left\{-nD_1(\Delta_\alpha, 1 - \rho_\alpha^2)\right\}, \tag{3.15}$$

where $\Delta_\alpha = C_\alpha - \alpha R$ and $1 - \rho_\alpha^2 = \alpha v/(1 + \alpha v)$. Here v is the signal-to-noise ratio.

Remark: Notice that $err_1(\alpha)$ depends also on L, n and v. Whether $err_1(\alpha)$ is exponentially small depends on the relative size of the combinatorial term $\binom{L}{\alpha L}$ and the exponential term in n and α.

Proof of Lemma 5: For the occurrence of E_ℓ, there must be an $S \in \mathcal{A}$ which differs from the subset S^* sent in an amount $card(S - S^*) = card(S^* - S) = \ell$ and which has $|Y - X_S|^2 \leq |Y - X_{S^*}|^2$, or equivalently has $T(S) \leq 0$, where

$$T(S) = \frac{1}{2}\left[\frac{|Y - X_S|^2}{\sigma^2} - \frac{|Y - X_{S^*}|^2}{\sigma^2}\right]. \tag{3.16}$$

The analysis proceeds by considering an arbitrary such S, bounding the probability that $T(S) \leq 0$, and then using an appropriately designed union bound to put such probabilities together. Notice that the subsets S and S^* have an intersection $S_1 = S \cap S^*$ of size $L - \ell$ and difference $S_2 = S - S_1$

of size $\ell = \alpha L$.

Let $p(Y, X)$ denote the joint density of Y and X when S^* is sent. Further, let $X_{S_1} = \sum_{j \in S_1} X_j$. The actual density of Y given X_{S_1}, denoted by $p(Y|X_{S_1})$, has mean X_{S_1} and variance $(\sigma^2 + \alpha P)I$. Further, there is conditional independence of Y and X_{S_2} given X_{S_1}.

Next consider the alternative hypothesis that S was sent and let $p_h(Y, X)$ denote the corresponding joint density under this hypothesis. The conditional density for Y given X_{S_1} and X_{S_2}, denoted by $p_h(Y|X_{S_1}, X_{S_2})$, is now Normal$(X_S, \sigma^2 I)$. With respect to this alternative hypothesis, the conditional distribution for Y given X_{S_1} remains Normal$(X_{S_1}, (\sigma^2 + \alpha P)I)$. That is, $p_h(Y|X_{S_1}) = p(Y|X_{S_1})$.

We decompose the test statistic $T(S)$ in (3.16) as $T_1 + T_2$, where

$$T_1 = \frac{1}{2}\left[\frac{|Y - X_{S_1}|^2}{\sigma^2 + \alpha P} - \frac{|Y - X_{S^*}|^2}{\sigma^2}\right] \tag{3.17}$$

and

$$T_2 = \frac{1}{2}\left[\frac{|Y - X_S|^2}{\sigma^2} - \frac{|Y - X_{S_1}|^2}{\sigma^2 + \alpha P}\right]. \tag{3.18}$$

Note that $T_1 = T_1(S_1)$ depends only on terms in S^*, whereas $T_2 = T_2(S)$ depends also on the part of S not in S^*.

Concerning T_2, note that we may express it as

$$T_2(S) = \frac{1}{n} \log \frac{p(Y|X_{S_1})}{p_h(Y|X_S)} + C_\alpha, \tag{3.19}$$

where

$$C_\alpha = \frac{1}{2} \log\left(1 + \alpha \frac{P}{\sigma^2}\right)$$

is the adjustment by the logarithm of the ratio of the normalizing constants of these densities. Using Bayes rule notice that

$$\frac{p_h(X_{S_2}|Y, X_{S_1})}{p(X_{S_2})} = \frac{p_h(Y|X_{S_1}, X_{S_2})}{p(Y|X_{S_1})}.$$

Correspondingly, one gets from (3.19) that

$$T_2(S) = \frac{1}{n} \log \frac{p(X_{S_2})}{p_h(X_{S_2}|Y, X_{S_1})} + C_\alpha \tag{3.20}$$

We are examining the event E_ℓ that there is an $S \in \mathcal{A}$, with $card(S - S^*) = \ell$ and $T(S) \leq 0$. For positive λ the indicator of this event satisfies

$$1_{E_\ell} \leq \sum_{S_1} \left(\sum_{S_2} e^{-nT(S)} \right)^\lambda ,$$

where $S_1 = S \cap S^*$ is of size $L - \ell$ and $S_2 = S - S_1$ of size ℓ. The above follows since if there is such an S with $T(S) \leq 0$, then indeed that contributes a term on the right side of value at least 1. Here the outer sum is over $S_1 \subset S^*$. For each such S_1, for the inner sum, we have ℓ sections in each of which, to comprise S_2, there is a term selected from among $B - 1$ choices other than the one prescribed by S^*.

To bound the probability of E_ℓ, take the expectation of both sides, bring the expectation on the right inside the outer sum, and write it as the iterated expectation, where on the inside condition on Y, X_{S_1} and X_{S^*} to pull out the factor involving T_1, to get that $\mathbb{P}_{\beta^*}[E_\ell]$ is not more than

$$\sum_{S_1} \mathrm{E} e^{-n\lambda T_1(S_1)} \mathrm{E}_{X_{S_2}|Y,X_{S_1},X_{S^*}} \left(\sum_{S_2} e^{-nT_2(S)} \right)^\lambda .$$

Notice that $p(X_{S_2}|Y, X_{S_1}, X_{S^*}) = p(X_{S_2})$, that is X_{S_2} is independent of Y, X_{S_1} and X_{S^*}. Correspondingly, the inner expectation may be expressed as $\mathrm{E}_{X_{S_2}}(.)$. Further, we arrange for λ to be not more than 1. Then by Jensen's inequality, the expectation $\mathrm{E}_{X_{S_2}}(.)$ may be brought inside the λ power and inside the inner sum, yielding

$$\mathbb{P}_{\beta^*}[E_\ell] \leq \sum_{S_1} \mathrm{E} e^{-n\lambda T_1(S_1)} \left(\sum_{S_2} \mathrm{E}_{X_{S_2}} e^{-nT_2(S)} \right)^\lambda . \tag{3.21}$$

Recall that

$$e^{-nT_2(S)} = \frac{p_h(X_{S_2}|Y, X_{S_1})}{p(X_{S_2})} e^{-nC_\alpha}$$

from (3.20). Consequently, one has

$$\mathrm{E}_{X_{S_2}} e^{-nT_2(S)} = \mathrm{E}_{X_{S_2}|Y,X_{S_1}} e^{-nC_\alpha}$$

which is equal to e^{-nC_α}. The sum over S_2 entails less than $B^\ell = e^{nR\alpha}$, where $\alpha = \ell/L$, choices so

the bound (3.21) becomes

$$\mathbb{P}_{\beta^*}[E_\ell] \leq \sum_{S_1} \mathrm{E} e^{-n\lambda T_1(S_1)} e^{-n\lambda[C_\alpha - \alpha R]}. \tag{3.22}$$

The sum over S_1 in the above expression is over $\binom{L}{\alpha L}$ terms. Further, $nT_1(S_1)$ is a sum of n independent mean-zero random variables each of which is the difference of squares of normals for which the squared correlation is $\rho_\alpha^2 = 1/(1+\alpha v)$. So using (3.9), the expectation $\mathrm{E}e^{-n\lambda T_1(S_1)}$ is found to be equal to $[1/[1-\lambda^2\alpha v/(1+\alpha v)]]^{n/2}$. When plugged in above and optimized over λ in $[0,1]$, one gets from the expression of D_1 given in (3.11) that the expectation in the right side of (3.22) is equal to $e^{-nD_1(\Delta_\alpha, 1-\rho_\alpha^2)}$. This completes the proof of the lemma.

Remark: A natural question to ask is why we didn't use the simpler union bound for $\mathbb{P}_{\beta^*}(E_\ell)$ given by,

$$\binom{L}{\ell} B^\ell \, \mathbb{P}_{\beta^*}[T(S) \leq 0],$$

where $S \in \mathcal{A}$, is any set with $card(S - S^*) = \ell$. One could then use a Chernoff bound for the term $\mathbb{P}_{\beta^*}[T(S) \leq 0]$. Indeed, this is what we tried initially; however, due to the presence of the two combinatorial terms, we were unable to make the above go to zero, with large n, for all rates less than capacity. In our proof above, by introducing the λ term in the exponent, we were able to reduce the B^ℓ term to $B^{\lambda\ell}$. Optimizing over λ revealed the best bound using this method. Somewhat similar analysis has been done before to obtain error exponent for the standard channel coding problem, for example in [19].

A difficulty with the Lemma 5 bound is that for α near 1 and for R correspondingly close to C, in the key quantity $\Delta_\alpha^2/(1-\rho_\alpha^2)$, the order of Δ_α^2 is $(1-\alpha)^2$, which is too close to zero to cancel the effect of the combinatorial coefficient $\binom{L}{\alpha L}$.

The following lemma refines the analysis of Lemma 5, obtaining the same exponent with an improved correlation coefficient. The denominator of $\Delta_\alpha^2/(1-\rho_\alpha^2)$ now becomes $\alpha(1-\alpha)/(1+\alpha v)$. This is an improvement due to the presence of the factor $(1-\alpha)$ allowing the conclusion to be useful also for α near 1. The price we pay is the presence of an additional term in the bound.

Lemma 6. *Let a positive integer $\ell \leq L$ be given and let $\alpha = \ell/L$. Then $\mathbb{P}_{\beta^*}[E_\ell]$ is bounded by the*

minimum for t_α in the interval $[0, C_\alpha - \alpha R]$ of $err_2(\alpha, t_\alpha)$, where

$$err_2(\alpha, t_\alpha) = \binom{L}{L\alpha} \exp\left\{-nD_1(\triangle_\alpha, 1-\rho_\alpha^2)\right\} + \exp\left\{-nD(t_\alpha, \alpha^2 v/(1+\alpha^2 v)]\right\}, \tag{3.23}$$

where here the quantities $\triangle_\alpha = C_\alpha - \alpha R - t_\alpha$ and $1-\rho_\alpha^2 = \alpha(1-\alpha)v/(1+\alpha v)$.

Proof of Lemma 6: Split the test statistic $T(S) = \tilde{T}(S) + T^*$ where

$$\tilde{T}(S) = \frac{1}{2}\left[\frac{|Y - X_S|^2}{\sigma^2} - \frac{|Y - (1-\alpha)X_{S^*}|^2}{\sigma^2 + \alpha^2 P}\right]$$

and

$$T^* = \frac{1}{2}\left[\frac{|Y - (1-\alpha)X_{S^*}|^2}{\sigma^2 + \alpha^2 P} - \frac{|Y - X_{S^*}|^2}{\sigma^2}\right]$$

Take positive $\tilde{t} = t_\alpha$ and negative $t^* = -t_\alpha$. Then $E_\ell \subseteq \tilde{E}_\ell \cup E_\ell^*$, with $\tilde{E}_\ell$ being the event that there is an $S \in \mathcal{A}$, with $card(S - S^*) = \ell$ and $\tilde{T}(S) \leq \tilde{t}$. Similarly E_ℓ^* is the corresponding event that $T^* \leq t^*$. The part T^* has no dependence on S so its treatment is more simple. It is a mean zero average of differences of squared normal random variables, with squared correlation $1/(1+\alpha^2 v)$. So using its moment generating function, $\mathbb{P}_{\beta^*}[E_\ell^*]$ is exponentially small, bounded by the second of the two expressions in (3.23).

Concerning $\mathbb{P}_{\beta^*}[\tilde{E}_\ell]$, its analysis is much the same as for Lemma 5. We again decompose $\tilde{T}(S)$ as the sum $\tilde{T}_1(S_1) + \tilde{T}_2(S)$, where $\tilde{T}_2(S) = T_2(S)$ is the same as before. The difference is that in forming $\tilde{T}_1(S_1)$ we subtract $\frac{|Y-(1-\alpha)X_{S^*}|^2}{\sigma^2+\alpha^2 P}$ rather than $\frac{|Y-X_{S^*}|^2}{\sigma^2}$. Consequently,

$$\tilde{T}_1(S_1) = \frac{1}{2}\left[\frac{|Y - X_{S_1}|^2}{\sigma^2 + \alpha P} - \frac{|Y - (1-\alpha)X_{S^*}|^2}{\sigma^2 + \alpha^2 P}\right],$$

which again involves a difference of squares of standardized normals. But here the coefficient $(1-\alpha)$ multiplying X_{S^*} is such that we have maximized the correlations between the $Y - X_{S_1}$ and $Y - (1-\alpha)X_{S^*}$. Consequently, we have reduced the spread of the distribution of the differences of squares of their standardizations as quantified by the cumulant generating function. One finds that the squared correlation coefficient is $\rho_\alpha^2 = (1+\alpha^2 v)/(1+\alpha v)$ for which $1-\rho_\alpha^2 = \alpha(1-\alpha)v/(1+\alpha v)$. Accordingly we have that the moment generating function is $Ee^{-n\lambda\tilde{T}(S_1)} = \exp\{-(n/2)\log[1 - \lambda^2(1-\rho_\alpha^2)]\}$ which gives rise to the bound appearing as the first of the two expressions in (3.23).

This completes the proof of Lemma 6.

From Lemma 6, one gets that $\mathbb{P}_{\beta^*}[E_\ell] \leq err_2(\alpha)$, where

$$err_2(\alpha) = \min_{t_\alpha \in [0, C_\alpha - \alpha R]} err_2(\alpha, t_\alpha).$$

Consequently, from Lemmas 5 and 6, along with (3.14), one gets that $\mathbb{P}_{\beta^*}[\mathcal{E}_{\alpha_0}] \leq err_{tot}(\alpha_0)$, where,

$$err_{tot}(\alpha_0) = \sum_{\ell \geq \alpha_0 L} \min\left\{err_1(\ell/L),\, err_2(\ell/L)\right\}. \tag{3.24}$$

This is the bound we use to numerically compute the rate curve in Figure 3.1. Accordingly, the error exponent $E(\alpha_0, R)$ of Proposition 3 satisfies,

$$E(\alpha_0, R) \geq -\frac{1}{n}\log(err_{tot}(\alpha_0)). \tag{3.25}$$

Our task will be to give simplified lower bounds for the right side of (3.25) for all $R < C$. In the next section we characterize the section size required to achieve rates up to capacity. In Section 3.7 we prove Proposition 3. We also remark that in Appendix 3.F we discuss how the bounds of the above two lemmas may be modified to deal with the subset superposition coding scheme described in Subsection 3.F.

Since the bounds of lemma 6 are better than those in lemma 5 for α values near 1, for simplicity we only use the bounds from lemma 6 in characterizing the error exponents. Correspondingly, from hereon we take

$$\triangle_\alpha = C_\alpha - \alpha R - t_\alpha, \quad 1 - \rho_\alpha^2 = \frac{\alpha(1-\alpha)v}{1+\alpha v} \tag{3.26}$$

as in lemma 6.

3.6 Sufficient Section Size

We call $a = (\log B)/(\log L)$ the *section size rate*, that is, the bits required to describe the member of a section relative to the bits required to describe which section. It is invariant to the base of the log. Equivalently we have B and L related by $B = L^a$. Note that the size of a controls the polynomial size of the dictionary $N = LB = L^{a+1}$.

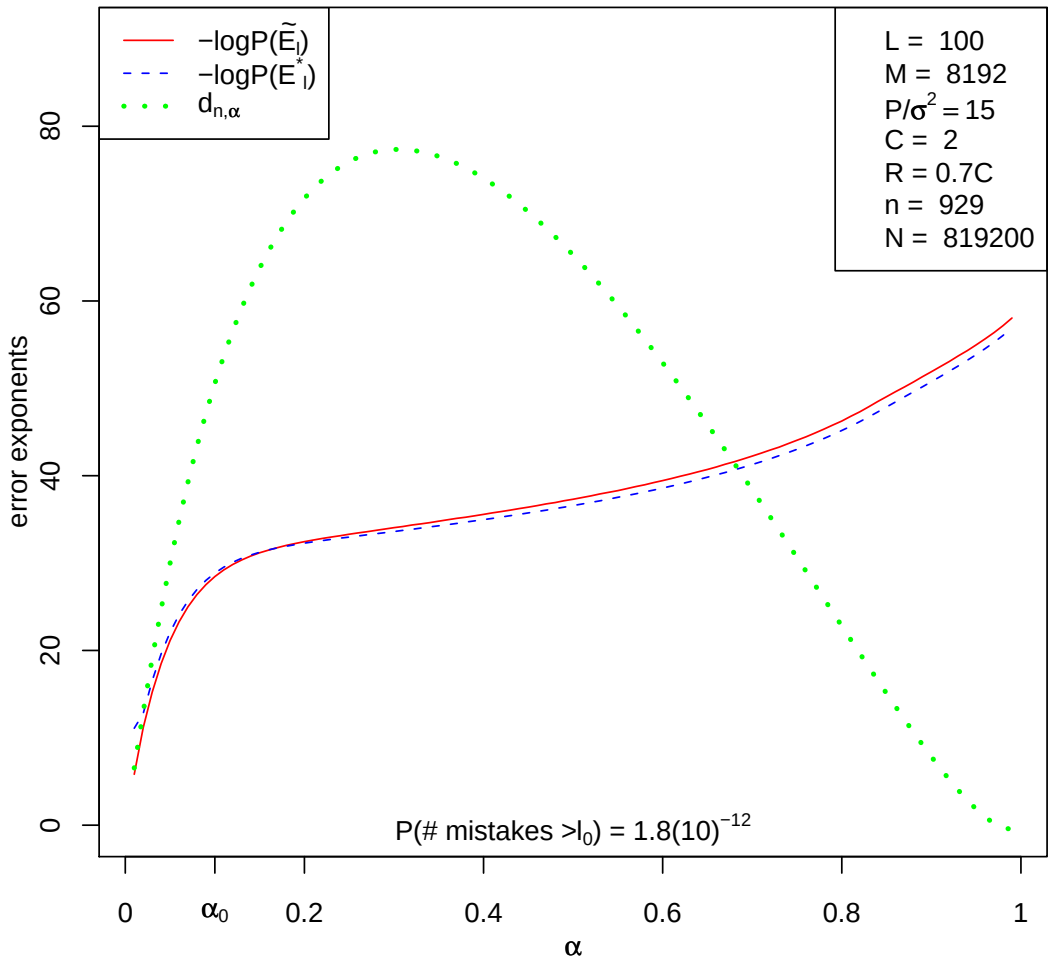

Figure 3.2: Exponents of contributions to the error probability as functions of $\alpha = \ell/L$ using exact least squares, i.e., $t = 0$, with $L = 100, B = 2^{13}$, signal-to-noise ratio $v = 15$, and rate 70% of capacity. The red and blue curves are the $-\log \mathbb{P}[\tilde{E}_\ell]$ and $-\log \mathbb{P}[E^*_\ell]$ bounds, using the natural logarithm, from the two terms in lemma 6 with optimized t_α. The dotted green curve is $d_{n,\alpha}$ (3.27). With $\alpha_0 = 0.1$, the total probability of at least that fraction of mistakes is bounded by $1.8(10)^{-12}$.

The codelength may be written as

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

We do not want a requirement on the section sizes with a of order $1/(C-R)$ for then the complexity would grow exponentially with this inverse of the gap from capacity. So instead we decompose $\triangle_\alpha = \tilde{\triangle}_\alpha + \alpha(C-R) - t_\alpha$ where $\tilde{\triangle}_\alpha = C_\alpha - \alpha C$. We investigate in this section the use of $\tilde{\triangle}_\alpha$ to cancel out the combinatorial coefficient $\binom{L}{\alpha L}$ appearing in the first term in (3.23). In subsequent sections, excess in $\tilde{\triangle}_\alpha$, beyond that needed to cancel the combinatorial coefficient, plus $\alpha(C-R) - t_\alpha$ are used to produce exponentially small error probability.

Define $D_{\alpha,v} = D_1(\triangle_\alpha, 1-\rho_\alpha^2)$ and $\tilde{D}_{\alpha,v} = D_1(\tilde{\Delta}_\alpha, 1-\rho_\alpha^2)$. Now $D_1(\Delta, 1-\rho^2)$ is increasing as a function of Δ, so $D_{\alpha,v}$ is greater than $\tilde{D}_{\alpha,v}$ whenever $\triangle_\alpha > \tilde{\triangle}_\alpha$. Accordingly, we decompose the exponent $D_{\alpha,v}$ as the sum of two components, namely, $\tilde{D}_{\alpha,v}$ and the difference $D_{\alpha,v} - \tilde{D}_{\alpha,v}$.

We then ask whether the first part of the exponent denoted $\tilde{D}_{\alpha,v}$ is sufficient to cancel out the effect of the log combinatorial coefficient $\log \binom{L}{L\alpha}$. That is, we want to arrange for the nonnegativity of the difference

$$d_{n,\alpha} = n\tilde{D}_{\alpha,v} - \log \binom{L}{L\alpha}. \tag{3.27}$$

Consequently, using $n = (aL \log L)/R$, one finds that for sufficiently large a depending on v, the difference $d_{n,\alpha}$ is nonnegative uniformly for the permitted α in $[0,1]$. The smallest such section size rate is

$$a_{v,L} = \max_\alpha \frac{R \log \binom{L}{L\alpha}}{\tilde{D}_{\alpha,v}\, L \log L}, \tag{3.28}$$

where the maximum is for α in $\{1/L, 2/L, \ldots, 1-1/L\}$. This definition is invariant to the choice of base of the logarithm, assuming that the same base is used for the communication rate R and for the $C_\alpha - \alpha C$ that arises in the definition of $\tilde{D}_{\alpha,v}$.

In the above ratio the numerator and denominator are both 0 at $\alpha = 0$ and $\alpha = 1$ (yielding $d_{n,\alpha} = 0$ at the ends). Accordingly, we have excluded 0 and 1 from the definition of $a_{v,L}$ for finite L. Nevertheless, limiting ratios arise at these ends.

We give bounds for $a_{v,L}$ and show that the value of $a_{v,L}$ is fairly insensitive to the value of L, with the maximum over the whole range being close to a limit a_v which is characterized by values in the vicinity of $\alpha = 1$.

Let v^* near 15.8 be the solution to $(1+v^*)\log(1+v^*) = 3v^* \log e$.

Lemma 7. *The quantity $a_{v,L}$ has the following properties,*

(a) For $L > 2$,

$$a_{v,L} \leq \frac{64R}{(1-\delta_L)}(1+v)^4/v^3 \tag{3.29}$$

where $\delta_L = \log 2/\log L$.

(b) The limit for large L of $a_{v,L}$ is a continuous function a_v which is given, for $0 < v < v^*$, by

$$\frac{8Rv(1+v)\log e}{[(1+v)\log(1+v) - v\log e]^2} \tag{3.30}$$

and for $v \geq v^*$ by

$$\frac{2R(1+v)}{[(1+v)\log(1+v) - 2v\log e]}. \tag{3.31}$$

(c) For all $R \leq C$ and using log base e, the a_v above is bounded by,

$$\frac{4v(1+v)\log(1+v)}{[(1+v)\log(1+v) - v]^2} \tag{3.32}$$

in the case $0<v<v^*$, which is approximately $16/v^2$ for small positive v; whereas, in the case $v \geq v^*$ it is bounded by

$$\frac{(1+v)\log(1+v)}{(1+v)\log(1+v) - 2v} \tag{3.33}$$

which asymptotes to the value 1 for large v.

The proof of the above lemma is routine. For convenience it is given in Appendix 3.B.

While a_v is undesirably large for small v, we have reasonable values for moderately large v. In particular, a_v equals 5.0 and 3, respectively, at $v = 7$ and $v^* = 15.8$, and it is near 1 for large v.

Numerically, it is of interest to ascertain the minimal section size rate $a_{v,L,\epsilon,\alpha_0}$, for a specified L such as $L = 64$, for R chosen to be a given high fraction of C, say $R = 0.8C$, for α_0 at a fixed small target fraction of mistakes, say $\alpha_0 = 0.1$, and for ϵ to be a small target probability, so as to obtain $err_{tot}(\alpha_0) \leq \epsilon$. Here $err_{tot}(\alpha_0)$ as in (3.24). This is illustrated in Figure 3.3 plotting the minimal section size rate as a function of v for $\epsilon = e^{-10}$. With such R moderately less than C, we observe substantial reduction in the required section size rate.

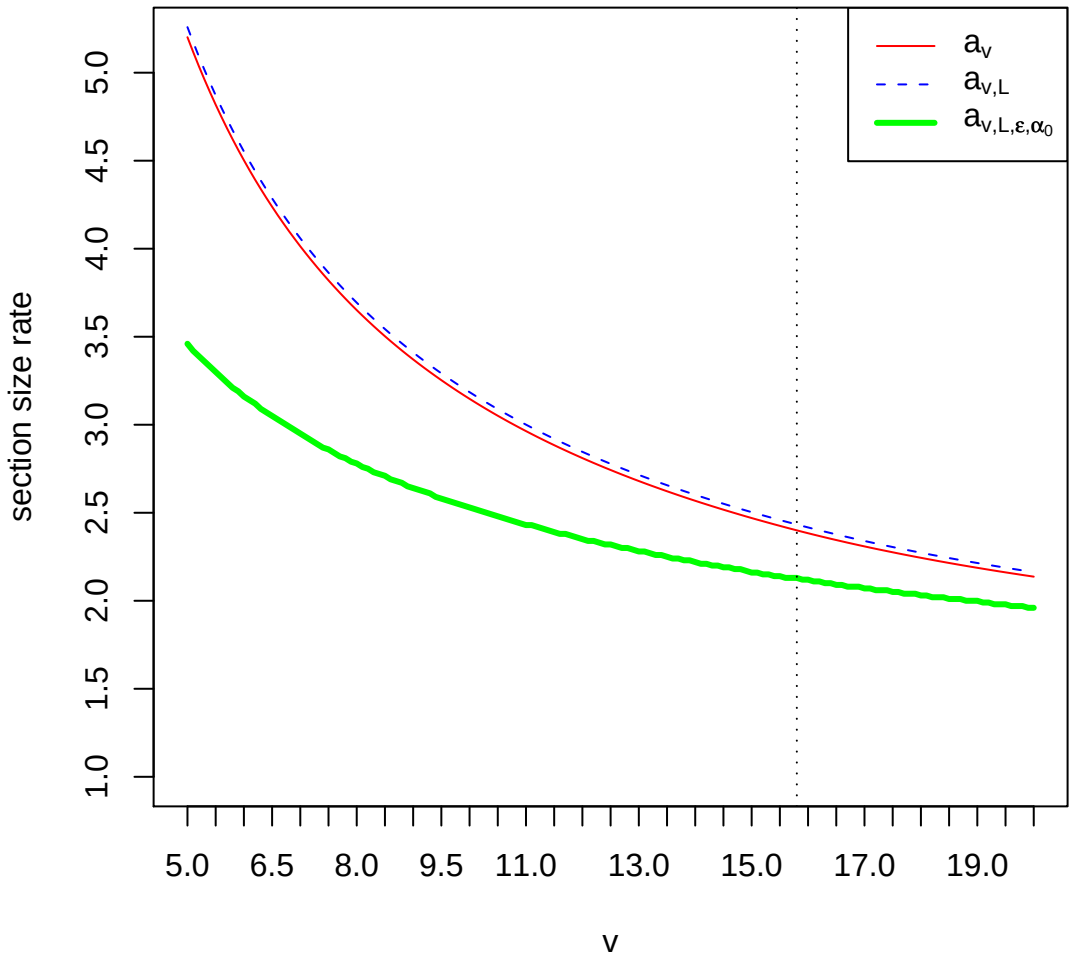

Figure 3.3: Sufficient section size rate a as a function of the signal-to-noise ratio v. The dashed curve shows $a_{v,L}$ at $L = 64$. Just below it the thin solid curve is the limit for large L. For section size $B \geq L^a$ the error probabilities are exponentially small for all $R < C$ and any $\alpha_0 > 0$. The bottom curve shows the minimal section size rate for the bound on the error probability contributions to be less than e^{-10}, with $R = 0.8C$ and $\alpha_0 = 0.1$ at $L = 64$.

3.7 Proof of Proposition 3

In this section we put the above conclusions together to prove proposition 3 demonstrating the reliability of approximate least squares. The following lemma will be useful in proving the lower bound for the error exponent in proposition 3. Let $g(x) = \sqrt{1+4x^2} - 1$ as before.

Lemma 8. *The following bounds hold.*

(a) For positive Δ and correlation $\rho \in (0,1)$, let $q = \Delta/\sqrt{1-\rho^2}$. Then,

$$D(\Delta, 1-\rho^2) \geq g(q)/4 \tag{3.34}$$

and

$$D_1(\Delta, 1-\rho^2) \geq \min\{g(q)/4,\, \Delta/2\}\,. \tag{3.35}$$

(b) For $\alpha \in [0,1]$, let $\tilde{\Delta}_\alpha = C_\alpha - \alpha C$. Then

$$\frac{v^2}{2(1+v)}\alpha(1-\alpha) \geq \tilde{\Delta}_\alpha \geq \frac{v^2}{4(1+v)^2}\alpha(1-\alpha) \tag{3.36}$$

For convenience we put its proof in appendix 3.A.

We now prove Proposition 3. Consider the exponent $D_{\alpha,v} = D_1(\Delta_\alpha, 1-\rho_\alpha^2)$ appearing in the error bound (3.23). Now $D_1(\Delta, 1-\rho^2)$ has a nondecreasing derivative with respect to Δ. So $D_{\alpha,v} = D_1(\triangle_\alpha, 1-\rho_\alpha^2)$ is greater than $\tilde{D}_{\alpha,v} = D_1(\tilde{\Delta}_\alpha, 1-\rho_\alpha^2)$. Consequently, it lies above the tangent line (the first order Taylor expansion) at $\tilde{\Delta}_\alpha$, that is,

$$D_{\alpha,v} \;\geq\; \tilde{D}_{\alpha,v} \;+\; (\Delta_\alpha - \tilde{\Delta}_\alpha)\; D', \tag{3.37}$$

where $D' = D_1'(\Delta)$ is the derivative of $D_1(\Delta) = D_1(\Delta, 1-\rho_\alpha^2)$ with respect to Δ, which is here evaluated at $\tilde{\Delta}_\alpha$. In detail, the derivative $D_1'(\Delta)$ is seen to equal

$$\frac{1}{1+\sqrt{1+4\Delta^2/(1-\rho_\alpha^2)}}\,\frac{2\Delta}{1-\rho_\alpha^2} \tag{3.38}$$

when $\Delta < (1-\rho_\alpha^2)/\rho_\alpha^2$, and this derivative is equal to 1 otherwise. [The latter case with derivative equal to 1 includes the situations $\alpha = 0$ and $\alpha = 1$ where $1-\rho_\alpha^2 = 0$ with $D_1 = \Delta$; all other α have $1-\rho_\alpha^2 > 0$.]

We now lower bound the derivative $D' = D_1'(\Delta)$ evaluated at $\Delta = \tilde{\Delta}_\alpha$. Using the upper bound on $\tilde{\Delta}_\alpha$ given in (3.36) and the form of $1-\rho_\alpha^2$, one gets that $4\tilde{\Delta}_\alpha^2/(1-\rho_\alpha^2)$ is bounded by $[v^3(1+\alpha v)/(1+v)^2]\alpha(1-\alpha)$, which using $1+\alpha v \le 1+v$ and $\alpha(1-\alpha) \le 1/4$, one gets that

$$4\tilde{\Delta}_\alpha^2/(1-\rho_\alpha^2) \le v^3/[4(1+v)].$$

Further using the lower bound in (3.36), one has $2\tilde{\Delta}_\alpha/(1-\rho_\alpha^2)$ is at least $(1/2)v/(1+v)^2$, where we make use of $1+\alpha v \ge 1$. Correspondingly,

$$D_1'(\tilde{\Delta}_\alpha) \ge \frac{v}{[2(1+v)^2]\sqrt{1+(1/4)v^3/(1+v)}}, \tag{3.39}$$

the right side of which is $2w_v$, where w_v is as in (3.4).

Now we are in position to apply lemma 6 and lemma 7. If the section size rate a is at least $a_{v,L}$ we have that $n\tilde{D}_{\alpha,v}$ cancels the combinatorial coefficient $\binom{L}{\alpha L}$ and hence the first term in the $\mathbb{P}_{\beta^*}[E_\ell]$ bound (3.23) (the part controlling $\mathbb{P}_{\beta^*}[\tilde{E}_\ell]$) is not more than

$$\exp\{-n[\Delta_\alpha - \tilde{\Delta}_\alpha]\, D'\},$$

where $\alpha = \ell/L$. Using $\Delta_\alpha = C_\alpha - \alpha R - t_\alpha$ and $\tilde{\Delta}_\alpha = C_\alpha - \alpha C$ and (3.39), yields $\mathbb{P}_{\beta^*}[E_\ell]$ not more than the sum of

$$\exp\{-2w_v n[\alpha(C-R) - t_\alpha]\}$$

and

$$\exp\{-nD(t_\alpha, \alpha^2 v/(1+\alpha^2 v))\},$$

for any choice of $t_\alpha \in [0, \alpha(C-R)]$. For convenience we take t_α to be $\alpha(C-R)/2$. In this case the first part of the above sum is $\exp\{-nw_v\alpha(C-R)\}$.

Now use (3.34) to get that $D(t_\alpha, \alpha^2 v/(1+\alpha^2 v))$ is at least $g(q)/4$, where $q = (C-R)\sqrt{1+\alpha^2 v}/(2\sqrt{v})$. Correspondingly, using $(1+\alpha^2 v) \ge 1$, one gets that $q \ge (C-R)/(2\sqrt{v})$. Accordingly, $D(t_\alpha, \alpha^2 v/(1+\alpha^2 v))$ is at least $g((C-R)/(2\sqrt{v}))/4$.

It follows from the above that

$$\mathbb{P}_{\beta^*}[E_\ell] \le 2e^{-n\min\left\{\alpha w_v \Delta,\, \frac{1}{4} g\left(\frac{\Delta}{2\sqrt{v}}\right)\right\}},$$

where $\alpha = \ell/L$, $\Delta = C - R$. Consequently, summing over all $\ell \geq \alpha_0 L$, for which $\alpha \geq \alpha_0$, one gets

$$\mathbb{P}_{\beta^*}[\mathcal{E}_{\alpha_0}] \leq 2Le^{-n \min\left\{\alpha_0 w_v \Delta,\, \frac{1}{4} g\left(\frac{\Delta}{2\sqrt{v}}\right)\right\}}.$$

The exponent in the right side of the above is $h(\alpha_0, \Delta) - (\log 2L)/n$. Now use the $\mathbb{P}_{\beta^*}[\mathcal{E}_{\alpha_0}] = \bar{\mathbb{P}}[\mathcal{E}_{\alpha_0}]$ to complete the proof of proposition 3.

Remarks: The form given for the exponential bound is meant only to reveal the general character of what is available. A compromise was made, by introduction of an inequality (the tangent bound on the exponent) to proceed most simply to this demonstration. Now understanding that it is exponentially small, our best evaluation avoids this compromise and proceeds directly, using the bound (3.24), as it provides substantial numerical improvement.

3.8 Generalization to approximate least squares

In conclusion, we remark that our results are equally valid for an *approximate least squares decoder*, which for some non-negative δ_0, chooses a $\hat{\beta} \in \mathcal{B}$ satisfying,

$$|Y - X\hat{\beta}|^2 \leq |Y - X\beta^*|^2 + \delta_0, \tag{3.40}$$

where β^* is what is sent. Since the above is less restrictive than (3.1), it may be possible to find a computationally feasible algorithm for it. Indeed, we show in Appendix 3.E that any computationally feasible algorithm, if it be an accurate decoder then it must be an approximate least squares decoder for some small δ_0.

We now describe how our error probability bounds can be generalized to incorporate (3.40). We note that (3.40) is equivalent to finding an $\hat{S} \in \mathcal{A}$, so that $T(\hat{S}) \leq t$, with $t = \delta_0/(2\sigma^2)$, where $T(S)$ is as in (3.16). We find that the expression for $err_1(\alpha)$ in lemma 5 holds for approximate least squares decoders with $t \leq C_\alpha - \alpha R$, if we replace $\triangle_\alpha$ by $\triangle_\alpha = C_\alpha - \alpha R - t$. Further, the expression for $err_2(\alpha, t_\alpha)$ of lemma 6 is also true for $t \leq C_\alpha - \alpha R$, if one replaces the t_α appearing in the second term of the bound by $t_\alpha - t$. Accordingly, for such approximate decoders, with $t \leq C_\alpha - \alpha R$,

the bound corresponding to lemma 6 becomes,

$$err_2(\alpha, t_\alpha) = \binom{L}{L\alpha} \exp\left\{-nD_1(\triangle_\alpha, 1-\rho_\alpha^2)\right\} + \exp\left\{-nD(t_\alpha - t, \alpha^2 v/(1+\alpha^2 v)]\right\},$$

where $\triangle_\alpha = C_\alpha - \alpha R - t_\alpha$ and $1-\rho_\alpha^2 = \alpha(1-\alpha)v/(1+\alpha v)$ is as in lemma 6.

The analysis of this decoder is quite similar to that of (3.1). Interested readers may refer to the document [4] for a more general analysis incorporating (3.40).

3.A Proof of Lemma 8

We first prove part (a). Write $D(\Delta, 1-\rho^2)$ explicitly as an increasing function of the ratio $q = \Delta/\sqrt{1-\rho^2}$. Working with logarithm base e, the derivative with respect to λ of the expression being maximized yields a quadratic equation which can be solved for the optimal

$$\lambda^* = \frac{1}{2\Delta}\big(\sqrt{1+4\Delta^2/(1-\rho^2)} - 1\big).$$

Using this λ^* we get that $D = (1/2)\big(\gamma - \log(1+\gamma/2)\big)$, which is at least $\gamma/4$. Here $\gamma = \sqrt{1+4q^2} - 1$, with $q = \Delta/\sqrt{1-\rho^2}$. Correspondingly, $D(\Delta, 1-\rho^2) \geq g(q)/4$. This proves (3.34).

For the lower bound on $D_1 = D_1(\Delta, 1-\rho^2)$, recall that the case $\lambda = 1$ case occurs when $1 + 4\Delta^2/(1-\rho^2) \geq (1+2\Delta)^2$, in which case D_1 is at least $\Delta - (1/2)\log(1+\Delta)$. Using $\Delta - (1/2)\log(1+\Delta) \geq \Delta/2$ proves (3.35).

Next we prove part (b). Notice the $\tilde{\Delta}_\alpha$ has second derivative $-(1/2)v^2/(1+\alpha v)^2$. It follows that $\tilde{\triangle}_\alpha \geq (1/4)\alpha(1-\alpha)v^2/(1+v)^2$, since the difference of the two sides has negative second derivative, so it is concave and equals 0 at $\alpha=0$ and $\alpha=1$.

For the upper bound, notice that the derivative of Δ_α is v_1 at $\alpha = 0$ and $-v_2$ at $\alpha = 1$, where $v_1 = v/2 - C$ and $v_2 = C - v/[2(1+v)]$. Correspondingly, $\tilde{\Delta}_\alpha$ is bounded from above by the minimum of $v_1\alpha$ and $v_2(1-\alpha)$. Now it is not hard to see that

$$\min\{v_1\alpha,\, v_2(1-\alpha)\} \leq \alpha(1-\alpha)(v_1+v_2).$$

Correspondingly, we get the upper bound in (3.36).

3.B Proof of Lemma 7

We first prove part (a). Define $q = \tilde{\Delta}_\alpha/\sqrt{1-\rho_\alpha^2}$, which, using the lower bound on $\tilde{\Delta}_\alpha$ given in lemma 8 (b) and $1-\rho_\alpha^2 = \alpha(1-\alpha)v/(1+\alpha v)$, is at least $(1/4)\sqrt{\alpha(1-\alpha)}v^{3/2}(1+\alpha v)^{1/2}/(1+v)^2$. Consequently, q is at least $(1/4)\sqrt{\alpha(1-\alpha)}v^{3/2}/(1+v)^2$ using $1+\alpha v \geq 1$. Correspondingly, using (3.35) and the lower bound (3.6), one gets that $\tilde{D}_{\alpha,v} = D_1(\tilde{\Delta}_\alpha,\, 1-\rho_\alpha^2)$ is at least

$$\min\left\{\frac{1}{8\sqrt{2}}\frac{\sqrt{\alpha(1-\alpha)}v^{3/2}}{(1+v)^2},\ \frac{\alpha(1-\alpha)v^3}{64(1+v)^4},\ \frac{1}{8}\frac{\alpha(1-\alpha)v^2}{(1+v)^2}\right\},$$

which is equal to $v^{3/2}/[8(1+v)^2]$ times

$$\min\left\{\frac{\sqrt{\alpha(1-\alpha)}}{\sqrt{2}},\ \frac{\alpha(1-\alpha)v^{3/2}}{8(1+v)^2},\ \alpha(1-\alpha)\sqrt{v}\right\}.$$

Further $\log\binom{L}{L\alpha}$ can be bounded by $\min(\alpha, 1-\alpha)L\log L$ and $L\log 2$. Therefore, it is at most $\alpha(1-\alpha)(L\log L)/(1-\delta_L)$, where $\delta_L = (\log 2)/\log L$. Using this, the lower bound on $\tilde{D}_{\alpha,v}$ and the form of $a_{v,L}$ given in (3.28), one gets that $a_{v,L}$ can be bounded by $8R(1+v)^2/[(1-\delta_L)v^{3/2}]$ times

$$\max\left\{\sqrt{2\alpha(1-\alpha)},\ \frac{8(1+v)^2}{v^{3/2}},\ 1/\sqrt{v}\right\}.$$

Now use $\alpha(1-\alpha) \leq 1/4$ to get that,

$$a_{v,L} \leq \frac{8R(1+v)^2}{(1-\delta_L)v^{3/2}}\max\left\{1/\sqrt{2},\ \frac{8(1+v)^2}{v^{3/2}},\ 1/\sqrt{v}\right\}.$$

Now observe that the second term in the maximum above dominates the other two terms for all v. This completes the proof of part (a).

Next we prove part (b). For α in $(0,1)$ we use $\log\binom{L}{L\alpha} \leq L\log 2$ and the strict positivity of $\tilde{D}_{\alpha,v}$ to see that the ratio in the definition of $a_{v,L}$ tends to zero uniformly within compact sets interior to $(0,1)$. So the limit a_v is determined by the maximum of the limits of the ratios at the two ends. In the vicinity of the left and right ends we replace $\log\binom{L}{L\alpha}$ by the continuous upper bounds $\alpha L\log L$ and $(1-\alpha)L\log L$, respectively, which are tight at $\alpha = 1/L$ and $1-\alpha = 1/L$, respectively. Then in accordance with L'Hôpital's rule, the limit of the ratios equals the ratios of the derivatives at $\alpha = 0$

and $\alpha=1$, respectively. Accordingly,

$$a_v = \max\left\{\frac{R}{\tilde{D}'_{0,v}}, \frac{-R}{\tilde{D}'_{1,v}}\right\}, \tag{3.41}$$

where $\tilde{D}'_{0,v}$ and $\tilde{D}'_{1,v}$ are the derivatives of $\tilde{D}_{\alpha,v}$ with respect to α evaluated at $\alpha=0$ and $\alpha=1$, respectively.

To determine the behavior of $\tilde{D}_\alpha = \tilde{D}_{\alpha,v}$ in the vicinity of 0 and 1 we first need to determine whether the optimal λ in its definition is strictly less than 1 or equal to 1. From section 3.4, the case $\lambda<1$ occurs if and only if $\tilde{\Delta}_\alpha < (1-\rho_\alpha^2)/\rho_\alpha^2$. The right side of this is $\alpha(1-\alpha)v/(1+\alpha^2 v)$. So it is equivalent to determine whether the ratio

$$\frac{(C_\alpha - \alpha C)(1+\alpha^2 v)}{\alpha(1-\alpha)v}$$

is less than 1 for α in the vicinity of 0 and 1. Using L'Hôpital's rule it suffices to determine whether the ratio of derivatives is less than 1 when evaluated at 0 and 1. At $\alpha=0$ it is $(1/2)[v-\log(1+v)]/v$ which is not more than $1/2$ (certainly less than 1) for all positive v; whereas, at $\alpha=1$ the ratio of derivatives is $(1/2)[(1+v)\log(1+v)-v]/v$ which is less than 1 if and only if $v<v^*$. In other words, at $\alpha=0$ the optimum λ is less than one for all v, whereas at $\alpha=1$ it is less than one if and only if $v<v^*$.

For the cases in which the optimal $\lambda<1$, we need to determine the derivative of $\tilde{D}_\alpha$ at $\alpha=0$ and $\alpha=1$. Recall that $\tilde{D}_\alpha$ is the composition of the functions $(1/2)(\gamma-\log(1+\gamma/2))$ and $\gamma=\sqrt{1+q}-1$ and $q=q_\alpha=4\tilde{\Delta}_\alpha^2/(1-\rho_\alpha^2)$. Also recall that the limit of q_α, as α tends to 0 or 1, is zero.

Use chain rule for finding the derivative of $\tilde{D}_\alpha$, taking the products of the associated derivatives. The first of these functions has derivative $(1/2)(1-1/(2+\gamma))$ which is $1/4$ at $\gamma=0$, the second of these has derivative $1/(2\sqrt{1+u})$ which is $1/2$ at $u=0$, and the third of these functions is

$$u_\alpha = \frac{\big(\log(1+\alpha v) - \alpha\log(1+v)\big)^2}{\alpha(1-\alpha)v/(1+\alpha v)}$$

which has derivative that evaluates to $(v-\log(1+v))^2/v$ at $\alpha=0$ and evaluates to $-[(1+v)\log(1+v)-v]^2/[v(1+v)]$ at $\alpha=1$. Corresponding, for $\alpha=0$, the derivative of $\tilde{D}_\alpha$ is $(v-\log(1+v))^2/(8v)$ for all v; whereas for $\alpha=1$, its derivative is $-[(1+v)\log(1+v)-v]^2/[8v(1+v)]$ for $v<v^*$.

For $v<v^*$, the magnitude of the derivative of $\tilde{D}_\alpha$ at 1 is smaller than at 0. Indeed, taking

square roots this is the same as the claim that $(1+v)\log(1+v)-v < \sqrt{1+v}(v-\log(1+v))$. Replacing $s=\sqrt{1+v}$ and rearranging, it reduces to $s\log s < (s^2-1)/2$, which is true for $s>1$ since the two sides match at $s=1$ and have derivatives $1+\log s < s$. Thus the limiting value for α near 1 is what matters for the maximum. This produces the claimed form of a_v for $v < v^*$.

In contrast for $v > v^*$, the optimal λ equal 1 for α in the vicinity of 1. In this case we use $\tilde{D}_\alpha = \tilde{\Delta}_\alpha + (1/2)\log\rho_\alpha^2$ which has derivative equal to $-(1/2)[(1+v)\log(1+v)-2v]/(1+v)$ at $\alpha=1$, which is again smaller in magnitude than the derivative at $\alpha=0$, producing the claimed form for a_v for $v > v^*$.

At $v = v^*$ we equate $(1+v)\log(1+v) = 3v$ and see that both of the expressions for the magnitude of the derivative at 1 agree with each other (both reducing to $v^*/(2(1+v^*))$) so the argument extends to this case, and the expression for a_v is continuous in v.

Part (c) is proved by using $R \le (1/2)\log(1+v)$ and simplifying the resulting expression. This completes the proof of Lemma 7.

3.C Improvement in form of exponent

The following improvement in the form of the exponent in Proposition 3 can be obtained.

Theorem 9. *Assume $B = L^a$, where $a \ge a_{v,L}$, and rate R is less than capacity C. For the least squares decoder*

$$\bar{\mathbb{P}}[\mathcal{E}_{\alpha_0}] \le 2Le^{-n\tilde{h}(\alpha_0, C-R)}.$$

Here

$$\tilde{h}(\alpha_0, C-R) = \min\left\{c_{\alpha_0,v}\,\alpha_0,\ \frac{1}{4}g\left(\frac{C-R}{2\sqrt{v}}\right)\right\},$$

where $c_{\alpha,v}$ is positive and tends to $\tau_v\tilde{w}_v/4$ as α tends to 0. Here

$$\tau_v = (1/2)[v-\log(1+v)] - [2vR/a]^{1/2} \tag{3.42}$$

and

$$\tilde{w}_v = (v/2 - C)/(2v). \tag{3.43}$$

Proof of Theorem 9: Here we determine the minimum value of Δ for which the combinatorial term $\binom{L}{L\alpha}$ is canceled, and we characterize the amount beyond that minimum which makes the error

probability exponentially small. Arrange $\Delta_\alpha^{\min}$ to be the solution to the equation

$$nD_1(\Delta_\alpha^{\min}, 1-\rho_\alpha^2) = \log \binom{L}{L\alpha}.$$

To see its characteristics, let $\Delta_\alpha^{target} = (1-\rho_\alpha^2)^{1/2} G(r_\alpha)$ at

$$r_\alpha = \frac{1}{n} \log \binom{L}{L\alpha},$$

using log base e. Here $G(r)$ is the inverse of the function $D(\delta, 1)$ which is the composition of the increasing functions $(1/2)[\gamma - \log(1+\gamma/2)]$ and $\gamma = \sqrt{1+4\delta^2} - 1$ previously discussed in Section 3.4. This $G(r)$ is near $\sqrt{2r}$ for small r. When $G(r) < (1-\rho_\alpha^2)^{1/2}/\rho_\alpha^2$ the condition $\lambda < 1$ is satisfied and $\Delta_\alpha^{\min} = \Delta_\alpha^{target}$ indeed solves the above equation; otherwise $\Delta_\alpha^{\min} = r_\alpha - (1/2)\log \rho_\alpha^2$ provides the solution.

Now $r_\alpha = (R/a)(\log \binom{L}{\alpha L})/(L \log L)$, which from before can be bounded by $(R/a)\alpha(1-\alpha)/(1-\delta_L)$. Also $1-\rho_\alpha^2 = \alpha(1-\alpha)v/(1+\alpha v)$. Consequently, $\Delta_\alpha^{\min}$ is small for large L; moreover, for α near 0 and 1, it is of order α and $1-\alpha$, respectively, and via the indicated bounds, derivatives at 0 and 1 can be explicitly determined.

The analysis in Lemma 7 may be interpreted as determining section size rates a such that the differentiable upper bounds on $\Delta_\alpha^{\min}$ are less than or equal to $\tilde{\Delta}_\alpha = C_\alpha - \alpha C$ for $0 \le \alpha \le 1$, where, noting that these quantities are 0 at the endpoints of the interval, the critical section size rate is determined by matching the slopes at $\alpha = 1$. At the other end of the interval, the bound on the difference $\tilde{\Delta}_\alpha - \Delta_\alpha^{\min}$ has a strictly positive slope for $a \ge a_v$ at $\alpha = 0$, given by τ_v as in (3.42). The positivity of τ_v follows from recalling that $a_v > R/\tilde{D}'_{0,v}$, since the second term in (3.41) always turns out to be the greater one. Consequently, one may take $\tilde{\Delta}_\alpha - \Delta_\alpha^{\min} = \tau_{\alpha,v}\,\alpha$ for some positive $\tau_{\alpha,v}$, where $\tau_{\alpha,v}$ tends to τ_v as α tends to 0.

Recall that $\Delta_\alpha = C_\alpha - \alpha R - t_\alpha$. Express Δ_α as the sum of $\Delta_\alpha^{\min}$, needed to cancel the combinatorial coefficient, and $\Delta_\alpha^{extra} = C_\alpha - \alpha R - \Delta_\alpha^{\min} - t_\alpha$, which is positive. This Δ_α^{extra} arises in establishing that the main term in the probability bound is exponentially small. It decomposes as $\Delta_\alpha^{extra} = \alpha(C-R) + (\tilde{\Delta}_\alpha - \Delta_\alpha^{\min}) - t_\alpha$. Arrange t_α to be $(1/2)[\alpha(C-R) + \tilde{\Delta}_\alpha - \Delta_\alpha^{\min}]$ so that $\Delta_\alpha^{extra} = (1/2)[\alpha(C-R) + \tau_{\alpha,v}\alpha]$.

Consider the exponent $D_{\alpha,v} = D_1(\Delta_\alpha, 1-\rho_\alpha^2)$ as given in lemma 6. We take a reference Δ_α^{ref} for which $\Delta_\alpha > \Delta_\alpha^{ref}$ and for which Δ_α^{ref} is at least $\Delta_\alpha^{\min}$ and at least a multiple of $\tilde{\Delta}_\alpha$. For

convenience,we set $\Delta_\alpha^{ref} = (1/2)[\Delta_\alpha + \Delta_\alpha^{min}]$ to be half way between $\Delta_\alpha^{\min}$ and Δ_α. Recall that $D_1(\Delta, 1-\rho^2)$ has a nondecreasing derivative with respect to Δ. So $D_{\alpha,v} = D_1(\triangle_\alpha, 1-\rho_\alpha^2)$ is greater than $D_{\alpha,v}^{ref} = D_1(\Delta_\alpha^{ref}, 1-\rho_\alpha^2)$. Consequently, it lies above the tangent line at Δ_α^{ref}, that is,

$$D_{\alpha,v} \ge D_{\alpha,v}^{ref} + (\Delta_\alpha - \Delta_\alpha^{ref})\, D',$$

where as before $D' = D_1'(\Delta)$ is the derivative of $D_1(\Delta) = D_1(\Delta, 1-\rho_\alpha^2)$ with respect to Δ, which is here evaluated at Δ_α^{ref}. Its expression is as in (3.38).

We wish to examine the behavior of $D_1'(\Delta_\alpha^{ref})$ for α near 0. For this, we first lower bound the derivative $D_1'(\Delta_\alpha^{ref})$. Since this derivative is non-decreasing it is at least as large as the value at $\Delta = (1/4)\tilde{\Delta}_\alpha$. Now recall that $\tilde{\Delta}_\alpha^2/(1-\rho_\alpha^2)$ has a limit 0 as α tends to 0. Further, $\tilde{\Delta}_\alpha/(1-\rho_\alpha^2)$ has limit $(v/2-C)/v$ as α tends to 0. Consequently, from (3.38), at $\Delta = (1/4)\tilde{\Delta}_\alpha$, we have $D_1'(\Delta)$ tends to $\tilde{w}_v$, given by (3.43), as α tends to 0. Consequently, $D_1'(\Delta_\alpha^{ref}) \ge \tilde{w}_{\alpha,v}$, where $\tilde{w}_{\alpha,v}$ is positive and tends to $\tilde{w}_v$ as α goes to 0.

Next examine $D_{\alpha,v}^{ref}$. Since Δ_α^{ref} is at least $\Delta_\alpha^{\min}$, it follows that $D_{\alpha,v}^{ref}$ is at least $D_{\alpha,v}^{\min} = D(\Delta_\alpha^{\min}, 1-\rho_\alpha^2)$. Consequently, as in the proof of proposition 3, if the section size rate a is at least $a_{v,L}$ then the $\mathbb{P}_{\beta^*}[E_\ell]$ bound (3.23) is not more than the sum of

$$\exp\{-n[\Delta_\alpha - \Delta_\alpha^{ref}]\, D'\},$$

and

$$\exp\{-nD(t_\alpha, \alpha^2 v/(1+\alpha^2 v))\}.$$

Using Δ_α^{ref} half way between Δ_α^{min} and Δ_α, the first part of the bound is at most

$$\exp\{-n(1/4)[\alpha(C-R) + \tau_{\alpha,v}\alpha]\, \tilde{w}_{\alpha,v}\}.$$

This bound is superior to the previous one, when R closely matches C, because of the addition of the non-negative $\tau_{\alpha,v}\alpha$ term. The second part of the bound can be dealt with as in proposition 3. Accordingly, we have proved that

$$\bar{\mathbb{P}}[\mathcal{E}_{\alpha_0}] \le 2Le^{-n\min\left\{c_{\alpha_0,v}\,\alpha_0,\, \frac{1}{4}g\left(\frac{C-R}{2\sqrt{v}}\right)\right\}},$$

where $c_{\alpha,v} = \tau_{\alpha,v}\tilde{w}_{\alpha,v}/4$ for small α. It tends to $\tau_v\tilde{w}_v/4$ as α tends to 0. This completes the proof of theorem 9.

3.D Computations

We describe how the rate curves in figure 3.1 were computed. The block error probability ϵ was fixed at 10^{-4} and the signal-to-ratio v was taken to be 20 and 100. The PPV curve was curve was computed using right side of (3.7) for the given ϵ and v. The maximum achievable (composite) rate for the superposition code was calculated in the following manner. The number of sections, L ranged from 20 to 100 in steps of 10, with the corresponding section size B taken to be L^{a_v}, where a_v as in (3.32), (3.33).

For given ϵ and values of v, L and B, the inner coder rate R_{inner} was decreased from $.99C$ to $.05C$ in decrements of $.001C$. For a given R_{inner}, the minimum section mistake rate $\alpha(R_{inner})$ so that the error probability, computed using bounds (3.24), is at most ϵ was computed. The corresponding composite rate is taken to be

$$R_{comp}(R_{inner}) = (1 - 2\alpha(R_{inner}))R_{inner}.$$

The maximum of the composite rates $R_{comp}(R_{inner})$, when R_{inner} ranged from $.99C$ to $.05C$ in decrements of $.001C$, is the reported maximum achievable rate for the superposition code for the given values of ϵ, v, L and B.

3.E Accurate decoder $\Rightarrow$ approximate least squares

In Lemma 11 below we show that any decoder is an approximate least squares decoder. More specifically, we show that if the fraction of mistakes α made by a decoder is small, the distance of the estimated fit $X\hat{\beta}$ from Y cannot be much greater than distance of the codeword sent, that is $X\beta^*$, from Y. To prove this we require the following lemma, which is a consequence of the restricted isometry property [**?**], [**?**] for Gaussian random matrices. We recall that the entries of our X matrix are i.i.d $N(0, P/L)$.

Lemma 10. *Let $R < C$ and $n = (L \log B)/R$. Then the following holds except on a set with*

probability at most $e^{-n(C-R)}$*:*

$$|X\beta - X\beta'| \leq c_{rip}\frac{\|\beta - \beta'\|}{\sqrt{L}} \quad \textit{for all} \quad \beta,\, \beta' \in \mathcal{B}. \tag{3.44}$$

where $c_{rip} = \sqrt{P}(1 + \sqrt{C/\log B} + \sqrt{2C})$ *is related to the restricted isometry property constant.*

Proof: Statement (3.44) is equivalent to giving uniform bounds on the maximum singular value of the matrices $W_S = X_S/\sqrt{n}$, for all $S \in \mathcal{A}$, where $\mathcal{A}$ is as in (3.13). For $S \in \mathcal{A}$, let $\lambda(W_S)$ denote the maximum singular value of W_S. We use a result in Szarek [33] (see also [12]), giving tail bounds for the maximum singular value for Gaussian matrices, from which one gets that for positive r,

$$\mathbb{P}[\lambda(W_S) > 1 + \sqrt{L/n} + r] \leq e^{-nr^2/2}.$$

Accordingly, choose $r = \sqrt{2C}$ and use $\sqrt{L/n} \leq \sqrt{C/\log B}$, to get that $\lambda(W_S) \leq c_{rip}$, except on a set with probability e^{-nC}.

We need $\lambda(W_S) \leq c_{rip}$ to hold uniformly for all B^L sets $S \in \mathcal{A}$, with high probability. Correspondingly, using $B^L = e^{nR}$, using a union bound one gets that the probability of the event

$$\lambda(W_S) \leq c_{rip} \quad \text{for all} \quad S \in \mathcal{A}$$

is at least $1 - e^{-n(C-R)}$. This completes the proof of the lemma.

If $\epsilon \sim N_n(0, \sigma^2)$, then from standard results on the tail bounds of chi-square random variables, one has that

$$\mathbb{P}[|\epsilon| > 2\sigma] \leq e^{-n/2}. \tag{3.45}$$

Lemma 10 and (3.45) gives us the following.

Lemma 11. *Assume that a decoder for the superposition code, operating at rate* $R < C$*, makes at most* α *section of mistakes. Denote as* $\hat{\beta}$ *the estimate of the true* β^* *outputted by the decoder. Then, with probability at least* $1 - 2e^{-n\min\{(C-R),1/2\}}$*, the estimate* $\hat{\beta}$ *satisfies,*

$$|Y - X\hat{\beta}|^2 \leq |Y - X\beta^*|^2 + \delta_0,$$

with $\delta_0 = c_{rip}\left(4\sqrt{2}\sigma + 2c_{rip}\sqrt{\alpha}\right)\sqrt{\alpha}$*. In other words, with high probability,* $\hat{\beta}$ *is the solution of an*

approximate least squares decoder (3.40) with the given δ_0.

Proof: We need to show that $|Y - X\hat{\beta}|^2$ cannot be much greater than $|Y - X\beta^*|^2$. Notice that,

$$\begin{aligned} |Y - X\hat{\beta}|^2 &\leq \Big(|Y - X\beta^*| + |X\hat{\beta} - X\beta^*|\Big)^2 \\ &= |Y - X\beta^*|^2 + 2|\epsilon||X\hat{\beta} - X\beta^*| \\ &\quad + |X\hat{\beta} - X\beta^*|^2 \end{aligned} \tag{3.46}$$

where for (3.46) we use the fact that the noise $\epsilon = Y - X\beta^*$. Now, $\|\hat{\beta} - \beta^*\|^2/L \leq 2\alpha$, since the decoder makes at most α mistakes. Accordingly, using lemma 10 and (3.45), one gets that with probability at least $1 - 2e^{-n\min\{(C-R),1/2\}}$, one gets that $|X\hat{\beta} - X\beta^*| \leq c_{rip}\sqrt{2}\sqrt{\alpha}$ and $|\epsilon| \leq 2\sigma$. Consequently, from (3.46), one gets

$$|Y - X\hat{\beta}|^2 \leq |Y - X\beta^*|^2 + \delta_0,$$

with probability at least $1 - 2e^{-n\min\{(C-R),1/2\}}$, where $\delta_0 = c_{rip}\left(4\sqrt{2}\sigma + 2c_{rip}\sqrt{\alpha}\right)\sqrt{\alpha}$.

3.F Error bounds for subset superposition codes

The method of analysis also allows consideration of *subset superposition coding* described in section 2.5. For, in this case all $\binom{N}{L}$ subsets of size L correspond to codewords, so with the rate in nats we have $e^{nR} = \binom{N}{L}$. The analysis proceeds in the same manner, with the same number $\binom{L}{L-\ell}$ of choices of sets $S_1 = S \cap S^*$ where S and S^* agree on $L - \ell$ terms, but now with $\binom{N-L}{\ell}$ choices of sets $S_2 = S - S^*$ of size ℓ where they disagree. We obtain the same bounds as above except that where we have $B^\ell = e^{n\alpha R}$, with the exponent αR, it is replaced by $\binom{N-L}{\ell} = e^{nR(\alpha)}$, with the exponent $R(\alpha)$ defined by $R(\alpha) = R\log\binom{N-L}{\alpha L}/\log\binom{N}{L}$.

Correspondingly, for subset superposition coding, the probability $P_{\beta^*}[E_\ell]$ is bounded by the minimum of the same expressions given in Lemma 5 and Lemma 6, except that the term αR appearing in these expression is replaced by the quantity $R(\alpha)$ defined above. We haven't investigated in greater detail for whether there is reliability for any rate below capacity for these codes.

Chapter 4

Decoding using the Iterative Algorithm

4.1 Introduction

Here we discuss a computationally feasible decoder and provide theoretical analysis for it. Since the decoding is done iteratively, involving multiple steps, with decisions during a particular step based on whether certain statistics exceed a threshold value, we call our algorithm the *iterative thresholding* algorithm.

As mentioned earlier, in order to demonstrate that rates up to capacity can be achieved using our feasible algorithm, we make some minor modifications to the code construction. The difference we introduce is that we allow the non-zero weights of β to vary across sections. This is explained in greater detail below.

Instead of drawing the entries of X to be i.i.d. $N(0, P/L)$, as in Chapter 3, one may, through a change of scale, assume that the entries are i.i.d $N(0, 1)$. We take the non-zeroes of β to be $\sqrt{P_{(\ell)}}$ is section ℓ, for $\ell = 1, \ldots, L$. Here the $P_{(\ell)}$'s are positive and satisfy,

$$\sum_{\ell=1}^{L} P_{(\ell)} = P.$$

The above power allocation implies that $E\|X\beta\|^2/n = P$, ensuring that our power control is satisfied. A schematic rendering of the setup is shown in figure 4.1.

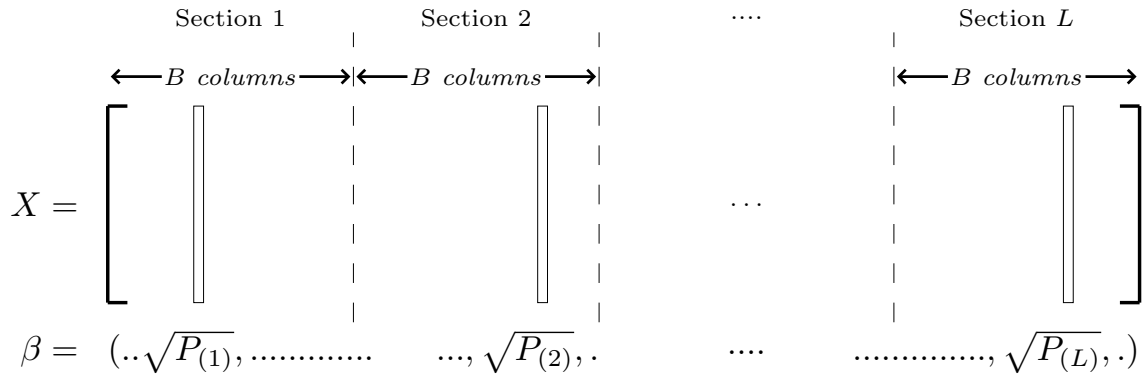

Figure 4.1: Schematic rendering of the dictionary matrix X and coefficient vector β as used here for the analysis of the iterative algorithm. The entries of X are i.i.d. $N(0, 1)$.

Also, notice that if $P_{(\ell)} = P/L$, for each ℓ, then the above code construction is the same as that studied in Chapter 3. As will be seen, choice of such equal weights will only allow us to achieve rates up to a threshold rate R_0, where,

$$R_0 = \frac{1}{2}\frac{P}{P+\sigma^2}$$

using our feasible algorithm.

To achieve rates arbitrarily close to capacity we choose power allocations $P_{(\ell)}$ proportional to $e^{-2\mathcal{C}(\ell-1)/L}$. Concerning the advantages of variable power, which allows our scheme to achieve rates near capacity, the idea is that power allocations proportional to $e^{-2\mathcal{C}(\ell-1)/L}$ give some favoring to the decoding of higher power sections among those that remain each step. In other words, it gives more statistical power to our iterative statistics to successively detect higher power sections among those that remain.

These power allocations also arise in successive decoding and rate splitting for multi-user communication by Cover [14]. As mentioned earlier, in this setup the number of users is L and the columns of a particular section corresponds to codewords for a particular user. Each user's message corresponds to one column from his section. The received string is simply the sum of the codewords sent by individual users – the goal being to detect each user's codeword. In successive decoding, the codewords are detected in succession, starting with the first user. In particular, the first user's codeword is decoded in the beginning, ignoring the contributions from the other users. The effective

signal-to-noise ratio is thus,

$$\frac{P_{(1)}}{P_{(2)} + \ldots + P_{(L)} + \sigma^2}.$$

After the first user's message is decoded, it is subtracted from the response Y and the second user's codeword is detected, again by ignoring effects of the other remaining users. The effective signal-to-noise ratio, assuming that the first step was done correctly, now becomes,

$$\frac{P_{(2)}}{P_{(3)} + \ldots + P_{(L)} + \sigma^2}.$$

The process is continued until all the messages are detected.

The power allocation $P_{(\ell)} \propto e^{-2\mathcal{C}(\ell-1)/L}$ corresponds to the unique allocation of powers for each user so that

$$\frac{P_{(\ell)}}{P_{(\ell+1)} + \ldots + P_{(L)} + \sigma^2}$$

is the same for each $\ell = 1, \ldots, L$. Thus the above power allocation allows each user's codeword to be detected with the same reliability using successive decoding.

As mentioned earlier, in [14] it is assumed that the number of users were fixed. Consequently, for each user to communicate at positive rates, the number of columns in a section (the section size) needed to be exponential in the sample size n. We overcome this problem in the single user setup by assuming that L, which corresponded to the number of users in the multi-user setup, is large. By making the communication rate negligible in each section, it allows for a section size that is only polynomial in n. The rate of communication is the sum of the rates for each section, which, because L is large, is not negligible. It can be shown that direct use of successive decoding to this setup will not result in exponentially small error probabilities. It is for this reason that we propose an alternative technique, in which multiple sections are detected during a step.

For rates near capacity, it helpful to use a modified power allocation, with power

$$P_{(\ell)} \propto \max\{e^{-2\mathcal{C}\frac{\ell-1}{L}}, u_{cut}\},$$

where $u_{cut} = e^{-2\mathcal{C}}\big(1 + \delta_c\big)$ with $\delta_c = c/\sqrt{2\log B}$, with a non-negative value of 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.

4.2 Intuition behind the algorithm

From the received Y and knowledge of the dictionary, we decode which terms were sent by an iterative algorithm. We now describe this algorithm.

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

$$\mathcal{H}_j = \mathcal{H}_{1,j} = \{X_j \cdot Y \geq T\}.$$

In terms of a normalized version, $\mathcal{Z}_{1,j} = X_j^T Y / \|Y\|$, of the above test statistic, this first step test is the same as comparing $\mathcal{Z}_{1,j}$ to a threshold τ.

Denote as

$$sent = \{j : \beta_j \neq 0\} \qquad \text{and} \qquad others = \{j : \beta_j = 0\}.$$

The set *sent* consists of one columns from each section.

As we shall see below, the distribution of $\mathcal{Z}_{1,j}$ is quite similar to that of a location shifted normal, where the shift is 0 for any j in *others* and is a positive quantity for j in *sent*. This positive shift for j in *sent* is seen to depend on $P_{(\ell)}$, the signal-strength in section ℓ; The larger the $P_{(\ell)}$, the more is this location shift.

The above gives us a means to identify at least some of the correct terms (that is those in *sent*), in the first step. Let $\mathcal{H}_{1,j} = \{\mathcal{Z}_{1,j} \geq \tau\}$. The threshold is chosen to be

$$\tau = \sqrt{2 \log B} + a. \tag{4.1}$$

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 will be above threshold and hence will be correctly decoded on this first step. Take $dec_1 = \{j \in J : 1_{\mathcal{H}_{1,j}} = 1\}$ as the set of terms detected in the first step. Recall that this is also the set of terms with the test statistic above threshold.

Denoting $P_j = P_{(\ell)}$ if j is in section ℓ, the output of the first step consists of the set of decoded terms dec_1 and the vector

$$F_1 = \sum_{j \in J_1} \sqrt{P_j}\, X_j 1_{\mathcal{H}_{1,j}}$$

which forms the first part of the fit. Notice that F_1 may also be expressed as $\sum_{\ell=1}^{L} \sum_{j \in section\ell} \sqrt{P_{(\ell)}} X_j$. 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 very natural way to conduct subsequent steps is 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 $\mathcal{Z}_{2,j}^{res} = X_j^T R_2 / \|R_2\|$ which may be compared to the same threshold $\tau = \sqrt{2 \log B} + a$. Then dec_2, the set for decoded terms for the second step, could be chosen in a manner similar to that in the first step. In other words, defining $\mathcal{H}_{2,j}^{res} = \{\mathcal{Z}_{2,j}^{res} \geq \tau\}$, we take $dec_2 = \{j \in J_2 : \mathcal{H}_{2,j}^{res} = 1\}$. From the set dec_2, one could the 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 second step. For any step k, we are only interested in

$$J_k = J - dec_1 \cup dec_2 \ldots \cup dec_{k-1},$$

that is, terms not decoded previously. One first computes the residual vector $R_k = Y - (F_1 + \ldots + F_{k-1})$. Accordingly, for terms in J_k, we get dec_k as the set of terms for which $\mathcal{Z}_{k,j}^{res} = X_j^T R_k / \|R_k\|$ is above τ.

The decoding stops when the size of the cardinality of the set of all decoded term becomes L or is there are no terms above threshold in a particular step.

The algorithm we analyze, although similar in spirit, is a slight modification of the above algorithm. The modifications are made so as to help characterize the distributions of $\mathcal{Z}_{k,j}^{res}$, for $k \geq 2$. These are described in the section below.

4.3 Modifications to the above algorithm

The algorithm we analyze is a modification of the above algorithm. The reason for the modification is purely for analysis purposes. The main reason for the modification is due to the difficulty in analyzing the statistics $\mathcal{Z}_{k,j}^{res}$, for $j \in J_k$ and for steps $k \geq 2$.

The distribution of the statistic $\mathcal{Z}_{1,j}$, used in the first step, is easy, as will be seen below. This

is because of the fact that the random variables

$$\{X_j,\, j \in J\} \quad \text{and} \quad Y$$

are jointly multivariate normal. However, this fails to hold for the random variables,

$$\{X_j,\, j \in J_k\} \quad \text{and} \quad R_k$$

used in forming $\mathcal{Z}_{k,j}^{res}$.

It is not hard to see why this joint Gaussianity fails. Recall that R_k may be expressed as,

$$R_k = Y - \sum_{j \in dec_{1,k-1}} \sqrt{P_j} X_j.$$

Correspondingly, since the event $dec_{1,k-1}$ is not independent of the X_j's, the quantities R_k, for $k \geq 2$, are no longer normal random vectors. It is for this reason the we introduce the following two modifications.

4.3.1 The first modification : Using a combined statistic

We overcome the above difficulty in the following manner. Recall that each

$$R_k = Y - F_1 + \ldots - F_{k-1}, \tag{4.2}$$

is a sum of Y and $-F_1, \ldots, -F_{k-1}$. Let $G_1 = Y$ and denote G_k, for $k \geq 2$, as the part of $-F_k$ that is orthogonal to the previous G_k's. In other words, perform Grahm-Schmidt orthogonalization on the vectors $Y, -F_1, \ldots, -F_k$, to get $G_{k'}$, with $k' = 1, \ldots, k$. Then, from (4.2), each R_k may be represented as

$$\frac{R_k}{\|R_k\|} = weight_1 \frac{G_1}{\|G_1\|} + weight_2 \frac{G_2}{\|G_2\|} + \ldots + weight_k \frac{G_k}{\|G_k\|},$$

for some weights, denoted by $weight_{k'} = weight_{k',k}$, for $k' = 1, \ldots, k$. More specifically,

$$weight_{k'} = \frac{R_k^{\mathrm{T}} G_{k'}}{\|R_k\| \|G_{k'}\|}.$$

It is not hard to see that one must have,

$$weight_1^2 + \ldots + weight_k^2 = 1.$$

Correspondingly, the statistic $\mathcal{Z}_{k,j}^{res} = X_j^{\mathrm{T}} R_k / \|R_k\|$, which we want to use for k th step detection, may be expressed as,

$$\mathcal{Z}_{k,j}^{res} = weight_1\, \mathcal{Z}_{1,j} + weight_2\, \mathcal{Z}_{2,j} + \ldots + weight_{k-1}\, \mathcal{Z}_{k,j},$$

where,

$$\mathcal{Z}_{k,j} = X_j^T G_k / \|G_k\|. \tag{4.3}$$

Instead of using the statistic $\mathcal{Z}_{k,j}^{res}$, for $k \geq 2$, we find it more convenient to use the statistic of the form,

$$\mathcal{Z}_{k,j}^{comb} = \lambda_{1,k}\, \mathcal{Z}_{1,j} + \lambda_{2,k}\, \mathcal{Z}_{2,j} + \ldots + \lambda_{k,k}\, \mathcal{Z}_{k,j}, \tag{4.4}$$

where $\lambda_{k',k}$, for $k' = 1, \ldots, k$ are positive deterministic quantities satisfying,

$$\sum_{k'=1}^{k} \lambda_{k',k}^2 = 1.$$

For convenience, unless there is some ambiguity, we suppress the dependence on k and denote $\lambda_{k',k}$ as simply $\lambda_{k'}$. Essentially, we choose λ_1 so that it is a deterministic proxy for $weight_1$ given above. Similarly, $\lambda_{k'}$ is a proxy for $weight_{k'}$ for $k' \geq 2$. The important modification we make, of replacing the random $weight_k$'s by deterministic counterparts, enables us to give an explicit characterization of the distribution of the statistic $\mathcal{Z}_{k,j}^{comb}$, which we use as a proxy for $\mathcal{Z}_{k,j}^{res}$ for detection of additional terms in successive iterations.

We now describe the algorithm after incorporating the above modification. For the time-being assume that for each k we have a vector of deterministic weights,

$$(\lambda_{k',k} : k' = 1, \ldots, k),$$

satisfying $\sum_{k'=1}^{k} \lambda_{k'}^2 = 1$, where recall that for convenience we denote $\lambda_{k',k}$ as $\lambda_{k'}$. Recall $G_1 = Y$.

For step $k = 1$, do the following

- For $j \in J$, compute

$$\mathcal{Z}_{1,j} = X_j^{\mathrm{T}} G_1 / \|G_1\|.$$

  To provide consistency with the notation used below, we also denote $\mathcal{Z}_{1,j}$ as $\mathcal{Z}_{1,j}^{comb}$.

- Update

$$dec_1 = \{j \in J : \mathcal{Z}_{1,j}^{comb} \geq \tau\}, \tag{4.5}$$

  which corresponds to the set of decoded terms for the first step. Also let $dec_{1,1} = dec_1$. Update

$$F_1 = \sum_{j \in dec_1} \sqrt{P_j} X_j.$$

This completes the actions of the first step. Next, perform the following steps for $k \geq 2$, with the number od steps k to be at most a pre-define value m.

- Define G_k as the part of $-F_{k-1}$ orthogonal to $G_1, \ldots, G_{k-1}$.

- For $j \in J_k = J - dec_{1,k-1}$, calculate

$$\mathcal{Z}_{k,j} = \frac{X_j^{\mathrm{T}} G_k}{\|G_k\|} \tag{4.6}$$

- Next for $j \in J_k$, compute the combined statistic using the above $\mathcal{Z}_{k,j}$ and $\mathcal{Z}_{k',j}$, $0 \leq k' \leq k-1$, given by,

$$\mathcal{Z}_{k,j}^{comb} = \lambda_1 \mathcal{Z}_{1,j} + \lambda_2 \mathcal{Z}_{2,j} + \ldots + \lambda_k \mathcal{Z}_{k,j},$$

  where the weights $\lambda_{k'} = \lambda_{k,k'}$, which we specify later, are positive and have sum of squares equal to 1.

- Update

$$dec_k = \{j \in J_k : \mathcal{Z}_{k,j}^{comb} \geq \tau\}, \tag{4.7}$$

  which corresponds to the set of decoded terms for the k th step. Also let $dec_{1,k} = dec_{1,k-1} \cup dec_k$, which is the set of terms detected after k steps.

- This completes the k th step. Stop if either L terms have been decoded, or if no terms are above threshold, or if $k = m$. Otherwise increase k by 1 and repeat.

Part of what makes the above work is our ability to assign deterministic weights ($\lambda_{k,k'} : k' = 1, \ldots, k$), for each step $k = 1, \ldots, m$. To be able to do so, we need good control on the (weigthed) sizes of the set of decoded terms $dec_{1,k}$ after step k, for each k. In particular, defining for each j, the quantity $\pi_j = P_j/P$, we define the size of the set dec_k as $size_k$, where

$$size_k = \sum_{j \in dec_k} \pi_j.$$

Notice that $size_k$, for each k, is a random quantity which depends on the number of correct detections and false alarms in each step. As we shall see, we need to provide good upper and lower bounds for the $size_1, \ldots, size_{k-1}$ that are satisfied with high probability, to be able to provide deterministic weights of combination, $\lambda_{k',k}$, for $k' = 1, \ldots, k$, for the kth step.

It turns out that the existing algorithm does not provide the means to give good controls on the $size_k$'s. To be able to do so, we need to further modify our algorithm.

4.3.2 The second modification : Pacing the steps

As mentioned above, we need to get good controls on the quantity $size_k$, for each k, where $size_k$ is defined as above. For this we need to modify the algorithm even further.

Let $thresh_1 = \{j \in J : \mathcal{Z}_{1,j} \geq \tau\}$ be the set of terms with the test statistic above threshold. We restrict decoding on the first step to terms in $thresh_1$ so as to avoid false alarms. For the algorithm described above, $thresh_1$ was equal to dec_1. Instead of taking dec_1 equal to $thresh_1$, we take dec_1 as a subset of $thresh_1$ satisfying the following condition:

Let $pace_1 > 0$, be a fixed value. Considering terms in J in order of decreasing $\mathcal{Z}_{1,j}$, we include in dec_1 as many as we can so that

$$size_1 = \sum_{j \in dec_1} \pi_1 \tag{4.8}$$

no more than $pace_1$.

Similarly, for steps $k \geq 2$, values $pace_k > 0$ are specified. Define:

$$thresh_k = \{j \in J_k : \mathcal{Z}_{k,j}^{comb} \geq \tau\}.$$

Then, instead of taking $dec_k = thresh_k$, we take dec_k to be a subset of $thresh_k$. As in the first step, this subset is chosen by considering the terms in J_k in the order of decreasing $\mathcal{Z}_{k,j}$ values,

and including in dec_k as many as we can so that

$$size_k = \sum_{j \in dec_k} \pi_j \tag{4.9}$$

is not more than $pace_k$.

The values for $pace_k$, for $k \geq 1$, will be specified shortly. It will also be seen that this simple modification allows us to get good controls on the $size_{k'}$'s, which in turns allows us to characterize the distribution of $\mathcal{Z}_{k,j}^{comb}$ for subsequent steps.

4.4 Performance of the algorithm

We allow not only for fixed rates $R < \mathcal{C}$, but also for rates for which the gap from capacity is of the order of a polynomial in $1/\log B$.

Recall that the power allocation $P_{(\ell)}$ we consider are of the form,

$$P_{(\ell)} \propto \max \left\{ e^{-2\mathcal{C}(\ell-1)/L}, \, u_{cut} \right\}$$

where $u_{cut} = e^{-2\mathcal{C}}(1+\delta_c)$, with $\delta_c = c/\sqrt{2 \log B}$ for a non-negative c. There are two cases analyzed for our adaptive successive decoder, depending on the presence of the slight leveling of the power allocation.

The first case has no leveling ($c = 0$) and a number of steps m of order $\sqrt{\log B}$. The rate R in nats is expressed as $\mathcal{C}/(1+\delta_b)^2$, where $\delta_b = b/\sqrt{2 \log B}$, with a minimal permitted b given by $b^* = a + 1/\sqrt{2\pi}$, where a is the positive value used in the threshold τ given by (4.1). Here we are ignoring terms given later in the analysis that are polynomially small in $1/L$ and negligible in comparison to what we state here provided the signal-to-noise ratio is not too small. In this case we have rate drop from capacity not smaller than a multiple of $1/\sqrt{\log B}$. The parameter a is free to be any positive value, though typically a between 0.5 and 0.8, gives acceptable performance tradeoffs for reasonable size codes. For large B, we advocate $a = (3/2) \log\log B/\sqrt{2 \log B}$. The probability of a fraction of mistakes more than a value of order $1/\sqrt{\log B}$ is shown to be exponentially small in $L/\sqrt{\log B}$.

The second case, with slight leveling of the power using a positive c, is shown to allow higher

rate for large B. We allow rate R up to $\mathcal{C}^*$, where $\mathcal{C}^*$ can be written as

$$\mathcal{C}^* = \frac{\mathcal{C}}{1 + drop^*}.$$

Here $drop^*$ is a positive quantity given explicitly later in this paper, for which we mention here two approximations.

1. When snr is large compared to $\log \log B$, it can be approximated by,

$$drop^* \approx \frac{5 \log \log B + 8\mathcal{C} + 8.23}{2 \log B}.$$

2. For snr near 1,

$$drop^* \approx \frac{7 \log \log B + 8.71}{2 \log B}$$

This $\mathcal{C}^*$ is within order $\log \log B / \log B$ of capacity and tends to $\mathcal{C}$ for large B. In summary form, the following expresses the main result of the performance of the iterative algorithm.

Proposition 12. *For any inner code rate $R < \mathcal{C}^*$, express it in the form*

$$R = \frac{\mathcal{C}^*}{1 + \kappa / \log B}, \tag{4.10}$$

with $\kappa \geq 0$. Then, for the partitioned superposition code,

I) Our iterative thresholding algorithm admits fraction of section mistakes less than

$$\delta_{mis} = \frac{3\kappa + 4r_1 + 4}{8\mathcal{C} \log B} \tag{4.11}$$

except in a set of probability not more than

$$p_e = \kappa_1 e^{-\kappa_2 L \min\left\{\kappa_3 \Delta^*, \kappa_4 (\Delta^*)^2\right\}},$$

where

$$\Delta^* = (\mathcal{C}^* - R)/\mathcal{C}^*.$$

Here r_1 is near $\log \log B + 1.38$ and κ_1 is a constant to be specified later that is only polynomial in B. See subsection 4.18.5 for details. Also, κ_2, κ_3 and κ_4 are constants that depend on the

snr.

II) After composition with an outer Reed Solomon code the decoder admits block error probability less than p_e, with the composite rate being $R_{total} = (1 - 2\delta_{mis})\, R$.

The proof of the above proposition is given in subsection 4.18.5.

Remarks:

a) The constants κ_2, κ_3 and κ_4 can be specified as follows. Define $\nu = snr/(1 + snr)$. We have that κ_2 is near $\nu/(2\mathcal{C})$. Further

$$\kappa_3 = \min \left\{ 1/(2snr^2),\ \frac{\log B}{16snr},\ 1/(16\mathcal{C}^*) \right\}$$

and

$$\kappa_4 = 1/(4snr).$$

b) If κ is a constant or smaller order, then Δ^* is of order $1/\log B$ below capacity. Further, ignoring $\log \log B$ factors, $\Delta = (\mathcal{C} - R)/\mathcal{C}$ is also $1/\log B$ below capacity.

c) If κ is small compared to $\log B$, then the R_{total} is close to R, with $\mathcal{C}^* - R$ and $\mathcal{C}^* - R_{total}$ of the same order. For such κ the exponent might as well be expressed with the $\mathcal{C}^* - R$ replaced by $\mathcal{C}^* - R_{total}$.

d) In the large *snr* regime, the presence of the $8\mathcal{C}$ term in the approximation for $\mathcal{C}^*$ reveals a need for $\log B$ to be large compared to the capacity $\mathcal{C}$ to achieve a rate that is a high fraction of capacity.

4.5 Comparison with Least Squares estimator

Here we compare the rate achieved here by our practical decoder with what is achieved with the theoretically optimal, but possibly impractical, least squares decoding of these sparse superposition codes shown in Chapter 3.

Let $\Delta = (\mathcal{C} - R)/\mathcal{C}$ be the rate drop from capacity, with R not more than $\mathcal{C}$. The rate drop Δ takes values between 0 and 1. With power allocated equally across sections, that is with $P_{(\ell)} = P/L$, it was shown in Chapter 3 that from any $\delta_{mis} \in [0, 1)$, 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 positive rate drop Δ and any size n. This bound is better than that obtained for our practical decoder in its freedom of any choice of mistake fraction, rate drop and size of the dictionary matrix X.

Here, as mentioned in Remark (b), we allow for rate drop Δ to be of order $1/\log B$. Further, from the expression (4.11), we have δ_{mis} is of order $1/\log B$, when κ is taken to be of $O(1)$ and ignoring $\log\log B$ factors. Consequently, we compare the error exponents obtained here with that of the least squares estimator of Chapter 3, when both Δ and δ_{mis} are of order $1/\log B$.

Using the expression given above for the least squares decoder one sees that the exponent is of order $n/(\log B)^2$, or equivalently $L/\log B$, using $n = (L \log B)/R$. For our decoder, neglecting loglog factors, the error probability bound is seen to be exponentially small in $L/(\log B)^2$ using the expression given in proposition 12. This bound is within a $(\log B)$ factor of what we obtained for the optimal least squares decoding of sparse superposition codes.

4.6 Further relationships to sparse signal recovery

Here we comment further on the relationships to high-dimensional regression. As mentioned in the introduction, a very common assumption is that the coefficient in *sparse*, meaning that it has only a few, say L, non-zeroes, with L typically much smaller than the dimension N. Note, unlike our communication setting, it is not assumed the the magnitude of the non-zeroes be known. Most relevant to our setting are works on *support recovery*, or the recovery of the non-zeroes of β. As mentioned earlier in section 3.3, β is typically allowed to belong to the set $\mathcal{B}'$ of all coefficient vectors, with L non-zeroes, with the magnitude of the non-zeroes being at least a certain positive value, say β_{min}.

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 [13], or the Lasso [34]. 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. Such works on sparse approximation and term selection concerns a class of iterative procedures which may be called relaxed greedy algorithms (including orthogonal matching pursuit or OMP) as studied in [22], [2], [28], [24], [7], [21], [37], [42], [23]. 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. Here by relaxed it is taken to mean that in updating the fit based on the newly selected term and the terms selected in previous steps, the contribution of terms selected previously are down-weighted. These procedures solves, to within specified precision, for the least squares convex projection onto the convex hull of a given set of vectors. A variant of it can also solve for the ℓ_1-penalized least squares solution to within a given precision, as shown in Huang et al. [21].

Results on support recovery can broadly be divided into two categories. The first involves giving, for a given X matrix, uniform guarantees for support recovery. In other words, it guarantees, for any β in the allowed set of coefficient vectors, that the probability of recovery is high. More specifically, denoting as S the support β, and as $\hat{S}$ its estimate obtained using a certain procedure, interest is mainly on conditions on X so that

$$P_{err,\,X} = \sup_{\beta \in \mathcal{B}'} \mathbb{P}_\beta \left(\mathcal{E} | X \right) \tag{4.12}$$

is small. Here $\mathcal{E}$ is the event that $\hat{S}$ is not equal to S. Observe that if $P_{err,X}$ is small, it gives strong guarantees on support recovery, since it ensures that any $\beta \in \mathcal{B}'$ can be recovered with high probability.

The second category of research involves results where the probability of recovery is obtained after certain averaging, where the averaging is over a distribution of X matrix. In particular, one seeks to make the quantity

$$P_{err} = \sup_{\beta \in \mathcal{B}'} \mathbb{P}_\beta(\mathcal{E}) \tag{4.13}$$

small. Here $\mathbb{P}_\beta(\mathcal{E}) = E_X \mathbb{P}_\beta(\mathcal{E}|X)$, where the expectation on the right is over the distribution of X. Notice that if (4.12) is small, it implies that (4.13) is small. Thus the second category of research provides a somewhat weaker characterization of the error probability. We describe results on both approaches in the sequel.

As mentioned in the previous paragraph, the first approach involves results for a given X matrix, satisfying certain conditions, high probability statements for the recovery of the non-zeroes of β. A common condition on the X matrix is the *mutual incoherence condition*, which assumes that the correlation between any two distinct columns be small. In particular, assuming that $\|X_j\|^2 = n$, for each $j = 1, \ldots, N$, it is assumed that,

$$\gamma(X) = \frac{1}{n} \max_{j \neq j'} \left| X_j^{\mathrm{T}} X_{j'} \right| \quad \text{is } \mathrm{O}(1/L). \tag{4.14}$$

Another related criterion is the *irrepresentable criterion* [35], [43], which assumes, for all subsets T of size L, that

$$\|(X_T^{\mathrm{T}} X_T)^{-1} X_T^{\mathrm{T}} X_j\|_1 < 1, \quad \text{for all} \quad j \in T^c. \tag{4.15}$$

Here $\|.\|_1$ denotes the ℓ_1 norm.

It can be shown, see for example [43], that conditions similar to that above are indeed necessary as well for support recovery. The above conditions are too stringent for our purpose of communicating at rate up to capacity. Indeed, for the i.i.d $N(0,1)$ designs that we consider, one requires n to be $\Omega(L^2 \log B)$ for the above condition to be satisfied. In other words, the rate R is of order $1/L$, which goes to 0 for large L.

As mentioned in Chapter 2, the idea of adapting techniques in high-dimensional to solve the communication problem began with Tropp [36], where he proposed using the signed superposition coding scheme discussed in section 2.5. However, since he used a condition similar to the irrepresentable condition discussed above, his results do not demonstrate communication at positive rates, let alone rates up to capacity.

We also remark that conditions (4.14) and (4.15) are required by algorithms such as Lasso

and Orthogonal Matching Pursuit for providing uniform guarantees on support recovery. However, there are algorithms which provided guarantees with much weaker conditions on X. Examples include the iterative forward-backward algorithm [42] and least squares minimization using concave penalties [41]. Even though these results, when translated to our setting, do imply communication at positive rates is possible, a demonstration that rates up to capacity can be achieved has been lacking.

The second approach is to assign a distribution for the X-matrix and analyze performance after averaging over this distribution. Wainwright [39] considers X matrices with rows i.i.d. $N(0, \Sigma)$, where Σ satisfies certain conditions, and shows that recovery is possible with the Lasso with n that is $\Omega(L \log B)$. In particular his results hold for the i.i.d. Gaussian ensembles that we consider here. Analogous results for the OMP was shown by Joseph [23]. Another result in the same spirit of average case analysis is done by Candès and Plan [10] for the Lasso, where the authors assign a prior distribution to β and study the performance after averaging over this distribution. The X matrix is assumed to satisfy a weaker form of the incoherence condition that holds with high-probability for i.i.d Gaussian designs, with n again of the right order.

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 coefficients vectors, such as that described above, and still recover their support and estimate the coefficient values. The main distinction from us being that our coefficient vectors belong to a finite set, of B^L elements, whereas in the above literature the class of coefficients vectors is almost always infinite. This additional flexibility is one of the reasons why an exact characterization of achieved rate has not been done in these works.

Another point of distinction is that majority of these works focus on exact recovery of the support of the true of coefficient vector β. As mentioned before, as our non-zeroes are quite small (of the order of $1/\sqrt{L}$), we are certain one cannot get exponentially small error probabilities for exact support recovery. Correspondingly, it is essential to relax the stipulation of exact support recovery and allow for a certain small fraction of mistakes (both false alarms and failed detection). To the best of our knowledge, there is still a need in the sparse signal recovery literature to provide proper controls on these mistakes rates to get significantly lower error probabilities.

4.7 Weighted measures of correct detections and false alarms

The following measures of performance of a step of the algorithm are important in characterizing the distribution of the statistics $\mathcal{Z}_{k,j}^{comb}$ for subsequent steps.

Let $\pi_j = P_j/P$, which sums to 1 across j in *sent*, and sums 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

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

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 $size_k = \hat{q}_k + \hat{f}_k$ 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

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

When $dec_k = thresh_k$, this total 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, 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 $\hat{f}_k^{tot} = \sum_{j \in \cap dec_{1,k}} \pi_j$ 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

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

As we will see in section 4.9, the following measure of correct detections in step, adjusted for

false alarms, plays an important role in characterizing the distributions of the statistics involved in an iteration.

$$\hat{q}_k^{adj} = \frac{\hat{q}_k}{1 + \hat{f}_k/\hat{q}_k}. \tag{4.16}$$

As we mentioned earlier, the distribution of $\mathcal{Z}_{1,j}$ is easy to characterize. Correspondingly, we do this separately in the next section. In section 4.9 we provide the analysis the the distribution of $\mathcal{Z}_{k,j}^{comb}$, for $k \geq 2$.

4.8 Analysis of the first step

In lemma 13 below we derive the distributional properties of $(\mathcal{Z}_{1,j} : j \in J)$. Lemma 14 in the next subsection characterizes the distribution of $(\mathcal{Z}_{k,j} : j \in J_k)$ for steps $k \geq 2$.

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 $C_{j,R} = \pi_j \, L\nu/(2R)$, where $\pi_j = P_j/P$ and $\nu = \nu_1 = P/(\sigma^2+P)$. Likewise define

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

Note also that it simplifies to

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

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 $C_{j,R,B} = C_{R,B} := (R_0/R)\, 2 \log B$ are equal for all j. This $C_{j,R}$ equals 1 when the rate R equals R_0 and then $C_{j,R,B} = 2 \log B$.

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. Let

$$\tilde{\mathcal{C}} = (L/2)[1 - e^{-2\mathcal{C}/L}], \tag{4.17}$$

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

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

Note, the $C_{j,R}$ simplifies to the value $e^{-2\mathcal{C}(\ell-1)/L}$ when the rate is $R = \tilde{\mathcal{C}}$.

We now are in a position to give the lemma for the distribution of $\mathcal{Z}_{1,j}$, for $j \in J$. The lemma below show that each $\mathcal{Z}_{1,j}$ is distribution as a shifted normal, where the shift is positive for any j in *sent* and is 0 for j in *others*.

Lemma 13. *For each $j \in J$, the statistic $\mathcal{Z}_{1,j}$, conditional on Y, 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 $N(0, \Sigma_1)$. Also,

$$\mathcal{X}_n^2 = \frac{\|Y\|^2}{\sigma_Y^2}$$

is a Chi-square n random variable that is independent of Z_1. Here $\sigma_Y = \sqrt{P + \sigma^2}$ is the standard deviation of each coordinate of Y.

Further, the covariance matrix Σ_1 can be expressed as $\Sigma_1 = I - \beta\beta^T/\sigma_Y^2$.

Proof. 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

The conditional distribution of each X_j given Y may be expressed as,

$$X_j = \beta_j\, Y/\sigma_Y^2 + U_j,$$

where $U_j = U_{1,j}$ is a vector in $\mathbb{R}^N$ having a multivariate normal distribution. Denote $b = \beta/\sigma_Y$. It is seen that

$$U_j \sim N_n\left(0, (1 - b_j^2)I\right),$$

where b_j is the j th coordinate of b.

Further, letting $U = [U_1 : \ldots : U_N]$, it follows from the fact that the rows of $A = [X : \epsilon]$ are i.i.d., that the rows of the matrix U are i.i.d.

Further, for row i of U, the random variables $U_{i,j}$ and $U_{i,j'}$ have mean zero and expected product

$$1_{\{j=j'\}} - b_j b_{j'}.$$

In general, the covariances $(\mathrm{E}[U_{i,j}U_{i,j'}] : j, j' \in J)$ organize into a matrix

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

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 $Z_{1,j} = U_j^T Y / \|Y\|$, also denoted $Z_{1,j}$ when making explicit that it is for the first step. Conditional on Y, one has that jointly across j, these $Z_{1,j}$ have the normal $N(0, \Sigma)$ distribution. Correspondingly, $(Z_{1,j} : j \in J)$ is independent of Y, and has a $N(0, \Sigma)$ distribution.

Where this gets us is revealed via the representation of the inner product $X_j^T Y$, conditional on Y, as

$$X_j^T Y \;=\; \beta_j \, \frac{\|Y\|^2}{\sigma_Y^2} \;+\; \|Y\| \, Z_{1,j}.$$

Correspondingly,

$$\mathcal{Z}_{1,j} = \; \beta_j \, \frac{\|Y\|}{\sigma_Y^2} \; + Z_{1,j}.$$

The proof is completed by noticing that for $j \in sent$, one has $\sqrt{C_{j,R,B}} = \beta_j \sqrt{n}/\sigma_Y$. □

Notice that for the constant power allocation case, $b_{1,j} = \sqrt{\nu/L}\, 1_{j\ sent}$, leading to

$$\mathcal{Z}_{1,j} \;=\; \sqrt{C_{R,B}} \, \frac{\|Y\|}{\sqrt{n}\sigma_Y} \, 1_{j\ sent} \;+\; Z_j,$$

where recall that $C_{R,B} = 2(R_0/R) \log B$.

4.9 Analysis of steps $k \geq 2$

We need the characterize the distribution of the statistic $\mathcal{Z}_{k,j}^{comb}$, $j \in J_k$, used the decoding additional terms in the k steps.

The statistic $\mathcal{Z}_{k,j}^{comb}$, $j \in J_k$, can be expressed more clearly in the following manner. For each $k \geq 1$, denote,

$$\mathcal{Z}_k = X^{\mathrm{T}} \frac{G_k}{\|G_k\|}.$$

Further, define

$$\mathcal{Z}_{1,k} = [\mathcal{Z}_1 : \mathcal{Z}_2 : \ldots : \mathcal{Z}_k]$$

and let $\Lambda_k = (\lambda_{k,1}, \lambda_{k,2}, \ldots, \lambda_{k,k})^{\mathrm{T}}$ be the deterministic vector of weights of combinations used for the statistics $\mathcal{Z}_{k,j}^{comb}$. Then $\mathcal{Z}_{k,j}^{comb}$ is simply the j th element of the vector

$$\mathcal{Z}_k^{comb} = \mathcal{Z}_{1,k}\Lambda_k.$$

We remind that for step k we are only interested in elements $j \in J_k$, that is those that were not decoded in previous steps.

Below we characterize the distribution of $\mathcal{Z}_k^{comb}$ conditioned on the what occurred on previous steps in the algorithm. More explicitly, we define $\mathcal{F}_{k-1}$ as

$$\mathcal{F}_{k-1} = \sigma\{G_1, G_2, \ldots, G_{k-1}, \mathcal{Z}_1, \ldots, \mathcal{Z}_{k-1}\}, \tag{4.18}$$

the sigma-field generated by the random variables, $G_1, G_2, \ldots, G_{k-1}$, as well as the statistics $\mathcal{Z}_1, \ldots, \mathcal{Z}_{k-1}$. This sigma field represents the events taking place up to step $k-1$. Notice that from the knowledge of $\mathcal{Z}_{k'}$, for $k' = 1, ..., k-1$, one can compute $\mathcal{Z}_{k'}^{comb}$, for $k' < k$. Correspondingly, the set of decoded terms $dec_{k'}$, till step $k-1$, is completely specified from knowledge of $\mathcal{F}_{k-1}$.

Next, note that in $\mathcal{Z}_{1,k}$, only the vector $\mathcal{Z}_k$ does not belong to $\mathcal{F}_{k-1}$. Correspondingly, the conditional distribution of $\mathcal{Z}_k^{comb}$, given $\mathcal{F}_{k-1}$, is described completely by finding the distribution of $\mathcal{Z}_k$ given $\mathcal{F}_{k-1}$. Accordingly, we only need to characterize the conditional distribution of $\mathcal{Z}_k$ given $\mathcal{F}_{k-1}$.

Next, we state a lemma showing that for $k \geq 2$, the distribution of $\mathcal{Z}_{k,j}$, with $j \in J_k$, can also be expressed in a manner similar to that of $\mathcal{Z}_{1,j}$. In particular, $\mathcal{Z}_{k,j}$ can be expressed as a normal random variable $Z_{k,j}$ plus a location shift depending on whether j is in *sent* or not.

Notice that we maintain the pattern used in lemma 13 and use $\mathcal{Z}_{k,j}$ to denote the test statistics that incorporate the shift for j in *sent* and standard font $Z_{k,j}$ to denote their counterpart mean zero normal random variables before the shift.

Initializing with the distribution of $\mathcal{Z}_1$ derived in lemma 13, we provide the conditional distributions $\mathcal{Z}_{k,J_k} = (\mathcal{Z}_{k,j} : j \in J_k)$, for $k = 2, \ldots, n$. As in the first step, we show that the distribution of $\mathcal{Z}_{k,J_k}$ can be expressed as the sum of a mean vector and a multivariate normal noise vector

$Z_{k,J_k} = (Z_{k,j} : j \in J_k)$. 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 14. *For each $k \geq 2$, the conditional distribution of $\mathcal{Z}_{k,j}$, $j \in J_k$, given $\mathcal{F}_{k-1}$ has the representation*

$$\sqrt{\hat{w}_k \, C_{j,R,B}} \, (\mathcal{X}_{d_k}/\sqrt{n}) \, 1_{\{j \, \in \, sent\}} \; + \; Z_{k,j}. \tag{4.19}$$

Recall that $C_{j,R,B} = n\pi_j\nu$. Further, $\hat{w}_k = \hat{s}_k - \hat{s}_{k-1}$, which are increments of a series with total

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

where

$$\hat{q}_k^{adj,tot} = \hat{q}_1^{adj} + \ldots + \hat{q}_k^{adj}. \tag{4.20}$$

Here $\hat{s}_1 = 1$. Also, $\sigma_k^2 = \hat{s}_{k-1}/\hat{s}_k$. The quantities $\hat{q}_k^{adj}$ is given by (4.16).

The conditional distribution $\mathbb{P}_{Z_{k,J_k}|\mathcal{F}_{k-1}}$ is normal $N_{J_k}(0, \Sigma_k)$, where the covariance Σ_k has the representation

$$\Sigma_k = I - \delta_k \delta_k^T / P.$$

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

The distribution of $\mathcal{X}_{d_k}^2 = \|G_k\|^2/\sigma_k^2$, given $\mathcal{F}_{k-1}$, is chi-square with $d_k = n - k + 1$ degrees of freedom, and further, it is independent $Z_{k,J_k} = (Z_{k,j} : j \in J_k)$.

The proof of the above lemma is is considerably more involved. It is given in section 4.19. From the above lemma one gets that $\mathcal{Z}_{k,j}$ is the sum of two terms - the 'shift' term and the 'noise' term $Z_{k,j}$. The lemma also provided that the noise term is normal with a certain covariance matrix Σ_k.

In the next section, we demonstrate that $Z_{k,j}$, for $j \in J_k$, are very close to being independent and identically distributed (i.i.d.).

4.10 The nearby distribution

Two probability measures $\mathbb{Q}$ and $\mathbb{P}$ are specified. Here $\mathbb{Q}$ is the approximating distribution. It makes all the $Z_{k',j}$, for $j \in J$, for $k' = 1, 2, \ldots, k$ independent standard normal, and like $\mathbb{P}$, the measure $\mathbb{Q}$ makes the $\mathcal{X}^2_{n-k'+1} = \|G_{k'}\|^2/\sigma^2_{k'}$ independent Chi-square$(n-k'+1)$ random variables.

Fill out of specification of the distribution assigned by $\mathbb{P}$, via a sequence of conditionals $\mathbb{P}_{Z_{k,J}|\mathcal{F}_{k-1}}$ for $Z_{k,J} = (Z_{k,j} : j \in J)$, which is for all j in J, not just for j in J_k. 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 $J - J_k$, given $\mathcal{F}_{k-1}$, we conveniently arrange them to have the same independent standard normal as is used by $\mathbb{Q}$. This definition is contrary to the definition above of $Z_{k,j}$ for j with $1_{H_{k',j}} = 1$ for earlier $k' < k$, but it is a simpler extension of the conditional distribution that shares the same marginalization to the true distribution of $(Z_{k,j} : j \in J_k)$ given $\mathcal{F}_{k-1}$.

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 15. *For any event A determined by $\mathcal{F}_k$,*

$$\mathbb{P}[A] \leq \mathbb{Q}[A] e^{k c_0},$$

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 $\mathcal{F}_k$.

For ease of exposition we relegate the proof to appendix 4.A.

4.11 Simple device in bounding detections and false alarms

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

$$\hat{q}^{tot}_k \geq \sum_{j\ sent} \pi_j 1_{\mathcal{H}_{k,j}} \tag{4.21}$$

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 $\mathcal{H}_{k',j}$.

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, one sees that it is at most

$$\hat{f}_k \leq \sum_{j \in other} \pi_j 1_{\mathcal{H}_{k,j}}. \tag{4.22}$$

The point of these simple inequalities is to 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 J_k.

The above gives us a mean to provide deterministic lower bounds for $\hat{q}_k^{tot}$ and upper bounds for $\hat{f}_k$ that are satisfied on sets with high probability. We call the lower bound on $\hat{q}_k^{tot}$ as $q_{1,k}$ and the upper bound on $\hat{f}_k$ as f_k. However, to proceed with our analysis, we also need to find good lower bounds on $\hat{q}_k$, the weighed proportion of correct detections on step k. This is where we make use of the pacing approach described in subsection 4.3.2.

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 $size_k$ is near a deterministic quantity which we call $pace_k$. Denote

$$size_{1,k} = \sum_{j \in dec_{1,k}} \pi_j = size_1 + \ldots + size_k.$$

The above will yield a sum $size_{1,k}$ bounded by $\sum_{k'=1}^{k} pace_{k'}$, which we arrange to match $q_{1,k}$.

In particular, setting $pace_k = q_{1,k} - q_{1,k-1}$, 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 $size_{1,k-1}$, one continues to add terms to dec_k, if possible, until their sum satisfies the following requirement,

$$q_{1,k} - 1/L_\pi < size_{1,k} \leq q_{1,k}, \tag{4.23}$$

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 $thresh_k$ might not be large enough to arrange for $size_k$ satisfying the above requirement. Nevertheless, it is satisfied, provided

$$size_{1,k-1} + \sum_{j \in thresh_k} \pi_j \geq q_{1,k}$$

or equivalently,

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

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 $\sum_{j \in sent} \pi_j 1_{\mathcal{H}_{k,j}}$. As mentioned earlier, our analysis demonstrates, for each k, that the inequality

$$\sum_{j \in sent} \pi_j 1_{\mathcal{H}_{k,j}} > q_{1,k}$$

holds with high probability. Accordingly, the above requirement is satisfied for each step, with high probability, and thence $size_k$ matches $pace_k$ to within $1/L_\pi$.

We now show that one can get a deterministic lower bound for $\hat{q}_k$ using the above approach. In particular, notice that from (4.23), one has,

$$q_{1,k} - 1/L_\pi < size_{1,k} \leq q_{1,k},$$
$$q_{1,k-1} - 1/L_\pi < size_{1,k-1} \leq q_{1,k-1}$$

Correspondingly, from the above one has,

$$size_k = \hat{q}_k + \hat{f}_k \geq q_{1,k} - q_{1,k-1} - 1/L_\pi$$

or re-arranging,

$$\hat{q}_k \geq q_{1,k} - q_{1,k-1} - 1/L_\pi - \hat{f}_k$$

Let f_k be a quantity so that $\hat{f}_k \leq f_k$ with high probability. Define

$$q_k = q_{1,k} - q_{1,k-1} - 1/L_\pi - f_k. \tag{4.24}$$

Then one has $\hat{q}_k$ is at least q_k with high probability.

4.12 Separation analysis

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 14 is through the sum

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

of the adjusted values $\hat{q}_k^{adj} = \hat{q}_k/(1 + \hat{f}_k/\hat{q}_k)$. We will show that each $\hat{q}_k$ is at least q_k, and each $\hat{f}_k$ is at most f_k, with high probability. Accordingly, for each k, one has $\hat{q}_k^{adj} \geq q_k^{adj}$, where

$$q_k^{adj} = q_k/(1 + f_k/q_k).$$

Also recall that from lemma 14, that,

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

From the above, $\hat{w}_k$ is at least w_k^*. which is given by,

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

where for each k, we take $q_k^{adj,tot} = q_1^{adj} + \ldots + q_k^{adj}$. Using this w_k^*, we define the corresponding vector of weight λ^*, used in forming the statistics $\mathcal{Z}_{k,j}^{comb}$, to have coordinates $\lambda_{k',k}^* = w_{k'}^*/[1 + w_2^* + \ldots + w_k^*]$, for $k' = 1$ to k.

Given that the algorithm has run for $k-1$ steps, we now proceed to describe how we calculate the bounds on the total detections and false alarms after k steps. Define the exception sets

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

Also denote the set

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

For convenience we suppress the dependence on k in these sets. From the argument given in the previous section, outside of A_q, we have that $\hat{q}_{k'}^{tot} \geq q_{1,k'}$ for each $1 \leq k' < k$, ensuring that for each such k' one can get decoding sets $dec_{k'}$ such that the corresponding $size_{1,k'}$ is at most $1/L_\pi$ below $q_{1,k'}$. Also, notice that outside of $A_q \cup A_f$ one has $\hat{q}_{k'} \geq q_{k'}$, where $q_{k'}$ as in (4.24).

Define the additional exception event

$$A_h = \cup_{k'=1}^{k} \{\mathcal{X}_{d_k}^2 \leq 1-h\},$$

where the term $\mathcal{X}_{d_k}$ is as given in lemma 14. Recall that

$$\mathcal{Z}_{1,j} = \sqrt{C_{j,R,B}}\, \mathcal{X}_n\, 1_{j\ sent} + Z_{1,j},$$

and, for $k' \geq 2$,

$$\mathcal{Z}_{k',j} = \sqrt{\hat{w}_{k'} C_{j,R,B}}\, \mathcal{X}_{d_k}\, 1_{j\ sent} + Z_{k',j}.$$

We recall that for j in $J_{k'}$ these $Z_{k',j}$ are determined as in a previous section by the data X,Y and the sequence of previous test outcomes, whereas for j outside $J_{k'}$, the $Z_{k',j}$ are auxiliary independent standard normals used in the analysis.

Define

$$C_{j,R,B,h} = C_{j,R,B}(1-h). \tag{4.25}$$

For $j \in J \cap J_k$, we have that $\mathcal{Z}_{k,j} = Z_{k,j}$. Further, except in $A = A_q \cup A_f \cup A_h$, we have

$$\mathcal{Z}_{1,j} \geq \sqrt{C_{j,R,B,h}}\ +\ Z_{1,j},$$

and for $k' \geq 2$,

$$\mathcal{Z}_{k',j} \geq \sqrt{\hat{w}_{k'}}\, \sqrt{C_{j,R,B,h}}\ +\ Z_{k',j}.$$

Recall that,

$$\mathcal{Z}_{k,j}^{comb} = \lambda_1^*\, \mathcal{Z}_{1,j}\ +\ \lambda_2^*\, \mathcal{Z}_{2,j}\ +\ \ldots\ +\ \lambda_k^*\, \mathcal{Z}_{k,j}.$$

Again we recognize for j in J_k this is the test statistic solely determined by the data X,Y via the sequence of previous test outcomes, whereas for j outside J_k this random variable $\mathcal{Z}_{k,j}$ is constructed in part from the auxiliary random variables used only in the analysis.

For j in *other* this $\mathcal{Z}_{k,j}^{comb}$ equals $Z_{k,j}^{comb}$ and for j in *sent*, when outside the exception set, this combination exceeds

$$[\lambda_1^* + \lambda_2^*\, \hat{w}_2 + \ldots + \lambda_k^*\, \hat{w}_k]\, \sqrt{C_{j,R,B,h}}\ 1_{j\ sent}\ +\ Z_{k,j}^{comb}$$

which is at least

$$\sqrt{\frac{C_{j,R,B,h}}{1 - q_{1,k-1}^{adj,tot}\,\nu}} \;+\; Z_{k,j}^{comb},$$

using $\hat{w}_k \geq w_k^*$.

Here for $j \in J_k$,

$$Z_{k,j}^{comb} = \lambda_1^* Z_{1,j} + \lambda_2^* Z_{2,j} + \ldots + \lambda_k^* Z_{k,j}.$$

Summarizing,

$$\mathcal{Z}_{k,j}^{comb} = Z_{k,j}^{comb} \qquad \text{for } j \in others$$

and, outside the exception set $A = A_q \cup A_f \cup A_h$,

$$\mathcal{Z}_{k,j}^{comb} \geq \text{shift}_{k,j} \;+\; Z_{k,j}^{comb}, \qquad \text{for } j \in sent,$$

where

$$\text{shift}_{k,j} = \sqrt{\frac{C_{j,R,B,h}}{1 - x_{k-1}\,\nu}}$$

with $x_0 = 0$ and $x_{k-1} = q_{1,k-1}^{adj,tot}$, for $k \geq 2$.

Since the x_k's are increasing, the $\text{shift}_{k,j}$'s increase with k. It is this increase in the mean shifts that helps in additional detections.

Set $H_{k,j}$ be the event,

$$H_{k,j} = \left\{\text{shift}_{k,j} 1_{\{j \in sent\}} \;+\; Z_{k,j} \geq \tau\right\} \tag{4.26}$$

Notice that

$$H_{k,j} = \mathcal{H}_{k,j} \qquad \text{for } j \in others.$$

Further, outside the set A, define above, one has

$$H_{k,j} \subseteq \mathcal{H}_{k,j} \qquad \text{for } j \in sent.$$

4.13 Target False Alarm Rates

A target false alarm rate for step k arises as a bound f_k^* on the expected value of $\sum_{j\ other} \pi_j 1_{H_{k,j}}$. This expected value is $(B-1)\bar{\Phi}(\tau_k)$, where $\bar{\Phi}(\tau_k)$ is the upper tail probability with which a standard

normal exceeds the threshold $\tau_k = \sqrt{2 \log B} + a_k$. If a_k is equal to the same value a for all k, this makes $\tau_k = \tau = \sqrt{2 \log B} + a$. Then $f_k^* = f^*$ where we set

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

That this is a bound on the target false alarm rate arises from the fact that $\bar{\Phi}(x) \le \phi(x)/x$ for positive x, with ϕ being the standard normal density. Likewise set values $f_k = f > f^*$. We express $f = \rho f^*$ with a constant factor $\rho > 1$.

By a union bound the total false alarm rate target at step k is $f_{1,k}^* = k f^*$. A corresponding upper bound would be $f_{1,k} = kf$ at step k. Another choice that will be seen to have some advantages of improved exponent is to use the same upper bound $f_{1,k} = mf$ for the partial totals as used for the final total at step m.

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 $f_{1,m}^*$ to be bounded by a multiple of $1/\log B$. Moreover, the number of steps m will be of order $\log B$. So with $f_{1,m}^* = mf^*$ 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 $(3/2)(\log \log B)/\sqrt{2 \log B}$.

4.14 Target Total Detection Rate

Per the preceding subsection, we are arranging an increasing sequence of likely lower bounds $\text{shift}_{k,j}$, on the shift of our combined test statistic for the terms j in *sent* that facilitates decoding on step k. Correspondingly, set $\mu_{k,j} = \text{shift}_{k,j} - \tau$. Except in $A_q \cup A_f \cup A_h$, the event $\mathcal{H}_{k,j}$ contains $H_{k,j}$.

For each k, set

$$q_{1,k}^* = \sum_{j \; sent} \pi_j \bar{\Phi}(-\mu_{k,j}).$$

That is, $q_{1,k}^*$ is the expectation, with respect to $\mathbb{Q}$, of

$$\hat{q}_{1,k} = \sum_{j \; sent} \pi_j 1_{H_{k,j}}. \tag{4.27}$$

That is, it is the expectation that would arise if the $Z_{k,j}$ were replaced by standard normals. It is this specification of $q_{1,k}^*$ that finally enables us to define the quantity $q_{1,k}$, introduced in section

4.11, that plays such an important role in the analysis. Arrange $q_{1,k}$ to be a value slightly less than $q^*_{1,k}$. Let $\eta_k = q^*_{1,k} - q_{1,k}$. For instance, we may set $q_{1,k} = q^*_{1,k} - \eta$, with $\eta_k = \eta$. As developed in the next subsection the value of η_k is arranged to be smaller than the increase in $q^*_{1,k}$ on step k.

This specification of $q^*_{1,k}$ and the related $q_{1,k}$ is a recursive definition, depending on the values of $q_{1,k-1}$ and $f_{1,k-1}$ from the previous step. Recall that the value of $\mu_{k,j}$ is obtained by plugging this choice of $x_{k-1} = q^{adj,tot}_{k-1}$ in the function

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

Then $q^*_{1,k}$ equals the function

$$g(x) = \sum_{j\ sent} \pi_j \bar{\Phi}(-\mu_j(x)) \tag{4.28}$$

evaluated at x_{k-1}.

For instance, in the constant power allocation case, $C_{j,R,B,h} = C_{R,B,h}$ is equal to $(R_0(1-h)/R)\,2\log B$, the same for all j in *sent*. So all such j experience the same level of likely shift $\text{shift}_{k,j} = \text{shift}_k = \sqrt{s_k}\,\sqrt{C_{R,B,h}}$. In this case $\mu_{k,j} = \mu_k = \text{shift}_k - \tau_k$ is the same value for each j in *sent*, which may also be written as $\mu(x)$ evaluated at x_{k-1}, with $\mu(x) = \sqrt{1/(1-x\nu)}\,\sqrt{C_{R,B,h}} - \tau$. Then the target fraction decoded on step k is

$$q^*_{1,k} = \bar{\Phi}(-\mu_k).$$

It obeys the recursion $q^*_{1,k} = g(x)$ evaluated at x_{k-1}, here with $g(x) = \bar{\Phi}(-\mu(x))$.

Where this will get us is demonstration that $q_{1,k}$ is a likely lower bound on $\hat{q}_{1,k} = \sum_{j\ sent} \pi_j 1_{H_{k,j}}$, which, as we have said, is a likely lower bound on the total fraction of correct detections using (4.21). If, for a suitable number of steps, we have arranged sufficient growth in $q^{adj,tot}_{k-1}$, then $q_{1,k}$ will be near 1 at the final k.

4.15 Building Up the Total Detection Rate

Let's demonstrate here how the likely total correct detection rate $q_{1,k}$ builds up to a value near 1, followed by the corresponding conclusion of reliability. Here we define the notion of correct detection being *accumulative*. This notion holds for the power allocations we study. In this section we illustrate the matter in the case of constant power allocation and R at most $R_0(1-h)$. Thereafter,

we handle variable power allocation and rates R up to the capacity.

Recall that for our iterative algorithm, from the function $g(.)$, given by (4.28), For our adaptive successive decoder there is a function $g(x)$ for $0 \leq x \leq 1$ such that for each step k,

$$q^*_{1,k} = g(q^{adj,tot}_{k-1}),$$

with which we then update the new $q_{1,k}$ by choosing it to be slightly less than $q^*_{1,k}$. That can be done by setting a small constant η for which $q_{1,k} = q^*_{1,k} - \eta$. Slightly better alternative choices motivated by the reliability bounds are to arrange $\sqrt{1-q_{1,k}} = \sqrt{1-q^*_{1,k}} + \eta$ or $D(q_{1,k} \| q^*_{1,k}) = \eta^2$ where D is the relative entropy between Bernoulli random variables of the indicated success probabilities.

Let x_r be any given value between 0 and 1, preferably near 1.

Definition: A function $g(x)$ is said to be *accumulative* for $0 \leq x \leq x_r$ with a positive *gap*, if

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

for all $0 \leq x \leq x_r$. Moreover, an adaptive successive decoder is *accumulative* with a given rate and power allocation if the function $g(x)$ used to update its likely correct detection rate satisfies this property for given x_r and positive *gap*.

To detail the progression of the $q_{1,k}$ consider the following Lemma.

Lemma 16. *Suppose $g(x)$ is accumulative on $[0, x_r]$ with a positive gap. Choose small positive η and $f > f^*$. Arrange $gap - \eta$ to be positive and for $4f\, x_r \leq (gap - \eta)^2$. Arrange $q_{1,k} = q^*_{1,k} - \eta$. Then the increase $q_{1,k} - q_{1,k-1}$ on each step for which $q^{adj,tot}_{k-1} \leq x_r$ is at least Λ, where Λ satisfies the equation*

$$\Lambda = (gap - \eta) - x_r\, f/\Lambda,$$

quadratic in Λ, for which the solution

$$\Lambda = (gap - \eta)\{1 + \left(1 - 4\, x_r\, f/(gap - \eta)^2\right)^{1/2}\}/2,$$

has its value between $(gap - \eta)/2$ and $(gap - \eta)$. A slightly larger Λ solving $\Lambda = (gap - \eta) - x_r\, f/(f + \Lambda)$ also satisfies $q_{1,k} - q_{1,k-1} \geq \Lambda$. In either case, the number of steps m required such that on step

$m-1$, *the* $q^{adj}_{1,m-1}$ *first exceeds* x_r, *is bounded by* $m \le 1/\Lambda$ *steps. At the final step* $q_{1,m}$ *exceeds* $g(x_r)-\eta$, *with* $g(x_r)-\eta$ *being at least* $x_r + (gap - \eta)$.

The proof of lemma 16 is given in appendix 4.B.

As for the matter of the choice of $f_{1,k}$, though it may seem wise to set it to kf, one finds that from small k, e.g. $k = 1$, a term in the probability bound has exponent of LfD, where D will be given as a function of ρ. With f of order $1/(\log B)^2$, such an exponent is not as large as desired. Instead fixing $f_{1,k} = mf$ for $k = 1, 2, \ldots, m$, allows the exponent to be at least $LmfD$ for all k. So that is better by a factor of $\log B$ in the exponent.

4.16 Reliability of the Adaptive Successive Decoder:

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.

Theorem 17. *Reliable communication by sparse superposition codes with adaptive successive decoding. With false alarm rate targets* $f_k > f^*_k$ *and update function* g_L, *set recursively the detection rate targets* $q_{1,k} = g_L(q^{adj,tot}_{k-1}) - \eta_k$, *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:*

$$\sum_{k=1}^{m} \Big[e^{-L_\pi D(q_{1,k} \| q^*_{1,k}) + c_0 k} \Big]$$

$$+ \sum_{k=1}^{m} \Big[e^{-L_\pi (B-1) D(p_k \| p^*_k)} \Big]$$

$$+ \sum_{k=1}^{m} e^{-(n-k+1) D_{h_k}},$$

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 $\{\|G\|_k^2/\sigma_k^2 \le n(1\text{–}h)\}$, *on steps* 1 *to* m. *Here* $L_\pi = 1/\max_j \pi_j$. *The* p_k, p_k^* *equal the corresponding* f_k, f_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.

Corollary 18. *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* $f_k = \bar{f}$ *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,*

$$me^{-2L_\pi \eta^2 + mc_0} + me^{-L_\pi \bar{f} \mathcal{D}(\rho)/\rho}$$

$$+me^{-(n-m+1)h_m^2/2}.$$

Proof of theorem 17 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 $H_{k,j}$, 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 $\hat{f}_k = \sum_{j \in other \cap J_k} \pi_j 1_{\mathcal{H}_{k,j}}$. Recall from (4.22) that $\hat{f}_k$ is bounded by $\sum_{j \in other} \pi_j 1_{\mathcal{H}_{k,j}}$.

Recall, as previously discussed, for all such j in *other*, the event $H_{k,j}$ is the event that $Z_{k,j}^{comb}$ exceeds τ, where the $Z_{k,j}^{comb}$ are standard normal random variables, independent across j in *other*. So the events $H_{k,j}$ are independent and equiprobable across such j, for each k. Let p_k^* be its probability or an upper bound on it, and let $p_k > p_k^*$. Then $A_{f,k} = \{\hat{f}_k \ge f_k\}$ is the same as $\{\hat{p}_k \ge p_k\}$ where

$$\hat{p}_k = \frac{1}{B-1} \sum_{j \in other} \pi_j \, 1_{H_{k,j}}.$$

Moreover, by lemma 29 in the appendix 4.D, the probability of the event $\{\hat{p}_k \ge p_k\}$ is less than $e^{-L_\pi (B-1) D(p_k \| p_k^*)}$. Therefore, the probability of $\{\hat{f}_k \ge f_k\}$ is less than

$$e^{-L_\pi (B-1) D(p_k \| p_k^*)}.$$

Likewise, we investigate the weighted proportion of correct decodings $\hat{q}_m^{tot}$ and the associated values $\hat{q}_{1,k} = \sum_{j\ sent} \pi_j 1_{H_{k,j}}$ which we compare 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 $A_{q,k} = \{\hat{q}_{1,k} \leq q_{1,k}\}$, $A_{f,k} = \{\hat{f}_k \geq f_k\}$ and $A_{h,k} = \{\mathcal{X}^2_{d_k}/n < 1-h\}$. 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, for $j \in sent$, the statistic $\mathcal{Z}^{comb}_{k,j}$ exceeds,

$$\sqrt{s_k}\,\sqrt{C_{j,R,B,h}}\ +\ Z^{comb}_{k,j},$$

where $s_k = 1/[1 - q^{adj,tot}_{k-1}\nu]$, Correspondingly, $A_{q,k}$ is contained in

$$\{\hat{q}_{1,k} < q_{1,k}\}$$

where

$$\hat{q}_{1,k} = \sum_{j\ sent} \pi_j 1_{\{Z^{comb}_{k,j} \geq a_{k,j}\}}.$$

Here $a_{k,j} = \tau - \sqrt{s_k}\,\sqrt{C_{j,R,B,h}}$. With respect to $\mathbb{Q}$, these $Z^{comb}_{k,j}$ are standard normal, independent across j, so the Bernoulli random variables $1_{\{Z^{comb}_{k,j} \geq a_{k,j}\}}$ have success probability $\bar{\Phi}(a_{k,j})$ and accordingly, with respect to $\mathbb{Q}$, the $\hat{q}_{1,k}$ has expectation $q^*_{1,k} = \sum_{j\ sent} \pi_j \bar{\Phi}(a_{k,j})$. Thus, again by Lemma 29 in the appendix the probability of

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

is not more than

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

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

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

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 $\sum_{j\ sent} \pi_j 1_{\mathcal{H}_{m,j}}$, by equation (4.21). Outside the exception sets A_h, A_f and $A_{q,m-1}$, it is at least $\hat{q}_{1,m} = \sum_{j\ sent} \pi_j 1_{H_{m,j}}$. With the indicated intersections, the above event is contained in $A_{q,m} = \{\hat{q}_{1,m} \leq q_{1,m}\}$, 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 differs from $q^*_{1,k}$ by the amount η.

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 appendix 4.E. 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 ρ.

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 $\bar{f}\,\mathcal{D}(\rho)/\rho$.

□

4.17 Computational Illustrations

We illustrate in two ways the performance of our algorithm. First, for fixed values L, B, snr and rates below capacity we evaluate detection rate as well as probability of exception set $P_{\mathcal{E}}$ using the theoretical bounds given in theorem 17. Plots demonstrating the progression of our algorithm are also shown. These highlight the crucial role of the function g_L in achieving high reliability.

Figure 4.2 presents the results of computation using the reliability bounds of theorem 17 for fixed L and B and various choices of snr and rates below capacity. The dots in these figures denotes $q_{1,k}$, for each k.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. The probability of error $P_{\mathcal{E}}$ is set to be near 10^{-3}. For each snr value the maximum rate, over a grid of values, for which the error probability is less than $P_{\mathcal{E}}$ is determined. With $snr = 1$ (Fig 4.2), this rate R is 0.3 bits which is 59% of capacity. When snr is 7 and 15 (Fig 4.2) , these rates correspond to 49% and 42% of their corresponding capacities.

For the above computations we choose power allocations proportional to $e^{-2\gamma l/L}(1+\delta_c)$, 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,

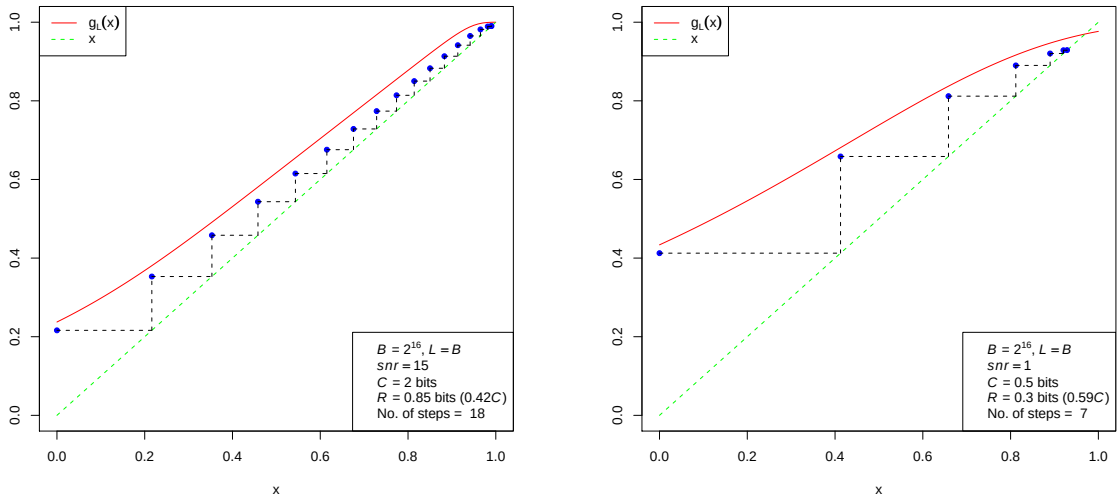

Figure 4.2: Plots demonstrating progression of our algorithm. (Plot on left) $snr = 15$. The weighted (unweighted) detection rate is 0.995 (0.985) for a failed detection rate of 0.014 and the false alarm rate is 0.005. (Plot on right) $snr = 1$. The detection rate (both weighted and un-weighted) is 0.944 and the false alarm and failed detection rates are 0.016 and 0.055 respectively.

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 Proposition 12 and in subsection 4.18.4 later on, γ is constrained to be equal to $\mathcal{C}$.

Figure 4.3 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 error 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 error 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 is quite close to the detailed envelope curve, showing that our theoretical bounds are quite conservative.

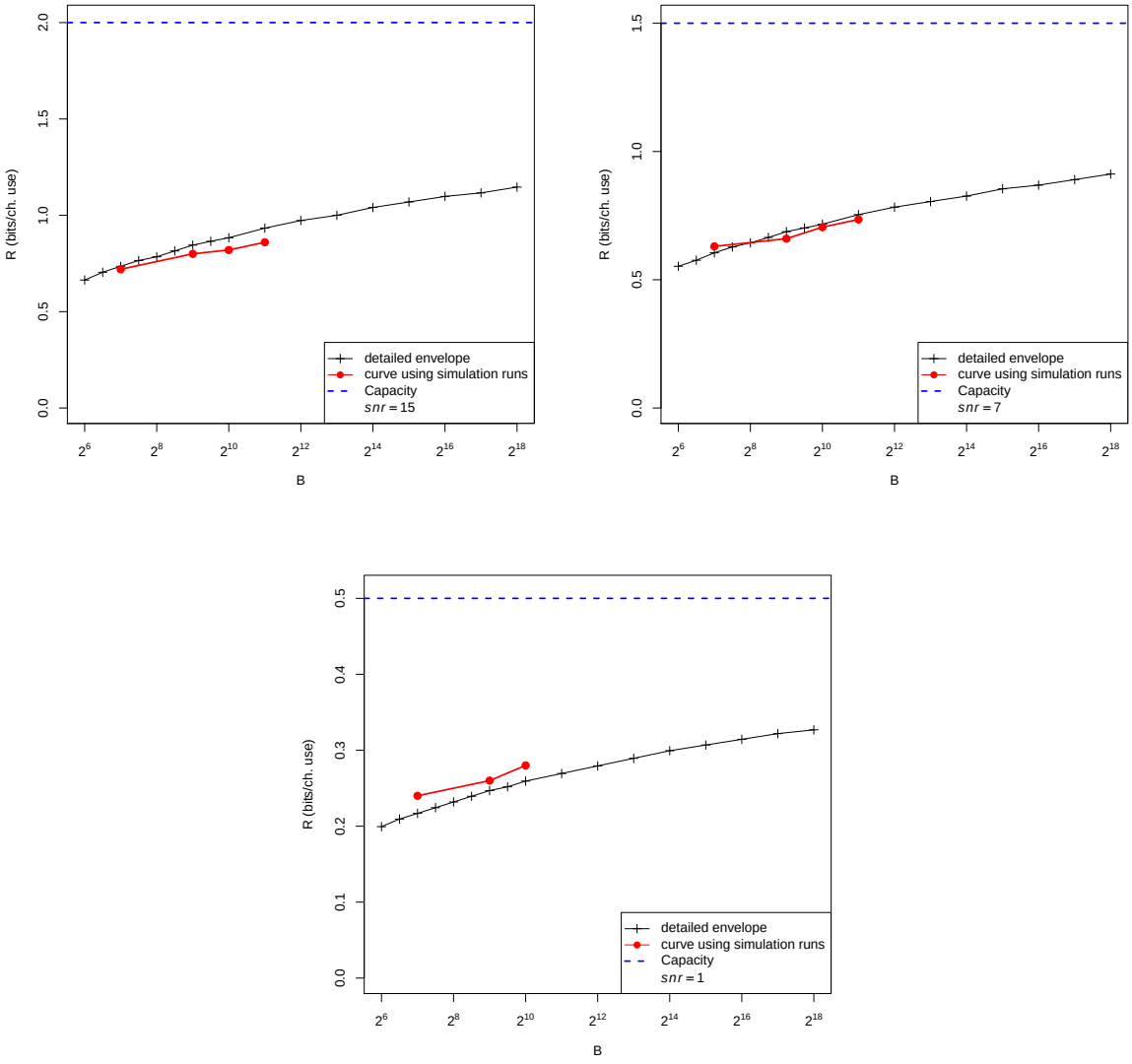

Figure 4.3: Plots of achievable rates as a function of B for snr values of 15, 7 and 1. Section error rate is controlled to be between 9 and 10%. For the curve using simulation runs the error probability of making more than 10% section mistakes is taken to be 10^{-3}.

4.18 Achievable Rates approaching Capacity

We demonstrate analytically that rates R moderately close to $\mathcal{C}$ are attainable by showing that the function $g_L(x)$ providing the updates for the fraction of correctly detected terms is indeed accumulative is indeed accumulative for suitable x_r and *gap*. Then the reliability of the decoder can be established via theorem 17. In particular, the matter of normalization of the weights $\pi_{(\ell)}$ is developed in subsection 4.18.1. An integral approximation $g(x)$ to the sum $g_L(x)$ is provided in subsection 4.18.2 and in subsection 4.18.3 we show that it is accumulative. Subsection 4.18.4 addresses the issue of optimization of parameters that arise in specifying the code. In subsection 4.18.5, we give the proof of proposition 12.

Let $g_L(x)$ for $0 \le x \le 1$ be the function given by

$$g_L(x) = \sum_{j\ sent} \pi_j\, \Phi(\mu(x, C_{j,R,h})),$$

where $\mu(x,u)$ for $0 \le x \le 1$ and $u \ge 0$ is the functions given by

$$\mu(x,u) \;=\; \big(\sqrt{u/(1-x\nu)} - 1\big)\sqrt{2\log B} \;-\; a.$$

Then $\mu_{k,j} = \mu(x, C_{j,R,h})$ where $C_{j,R,h} = C_{j,R}(1-h)$.

The weight in section ℓ is also denoted as π_ℓ. At slight risk of abuse of notation, it is also convenient to denote $C_{\ell,R} = \pi_{(\ell)}\, L\nu/(2R)$ and likewise $C_{\ell,R,h} = C_{\ell,R}(1-h)$, so that

$$g_L(x) = \sum_{\ell=1}^{L} \pi_{(\ell)}\, \Phi(\mu(x, C_{\ell,R,h})). \tag{4.29}$$

4.18.1 Variable power allocations

We consider two closely related schemes for allocating the power. First suppose $P_{(\ell)}$ is proportional to $e^{-2\mathcal{C}\ell/L}$. Then the weight for section ℓ is $\pi_{(\ell)}$ given by $P_{(\ell)}/P$. In this case recall that $C_{\ell,R} = \pi_{(\ell)}\, L\nu/(2R)$, from investigation of the normalizing constant, 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, and $\tilde{\mathcal{C}}$ is as in (4.17). The presence of the factor $\tilde{\mathcal{C}}/R$, arranged to be at least slightly greater than 1, has a role in compensating for the presence of the $-a$ in $\mu(x, C_{\ell,R,h})$. If R is too close to $\tilde{\mathcal{C}}$, the a, though helpful for the false alarm control, is a bias that presents an obstacle to keeping $g_L(x)$ above x. The difficulty arises especially for x near 1.

Therefore we modify the power allocation so that $u_\ell = \exp\{-2\mathcal{C}\frac{\ell-1}{L}\}$ for most ℓ/L and for large ℓ/L it is leveled to be not less than a value $u_{cut} = e^{-2\mathcal{C}}(1+\delta_c)$. Here δ_c will take the form $\delta_c = c/\sqrt{2\log B}$ with a value of c to be set below. Then let

$$\pi_\ell \propto \max\{u_\ell, u_{cut}\}$$

The idea is that by leveling the height to a slightly larger value for ℓ/L near 1, we can help overcome the bias from a.

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 $sum = \nu/(2\tilde{\mathcal{C}})$ as previously seen, where $\nu = 1-e^{-2\mathcal{C}}$. If $c > 0$ and $u_{cut} < 1$, it is the sum of two parts, depending on whether the quantity $e^{-2\mathcal{C}(\ell-1)/L}$ 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 $snr = \nu/(1-\nu)$ and $D(\delta) = (1+\delta)\log(1+\delta) - \delta$ is not more than $\delta^2/2$. Accordingly sum

may be expressed as

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

where

$$\delta^2_{sum} = D(\delta_c)/snr + 2\theta\mathcal{C}/(L\nu), \tag{4.30}$$

which is not more than $\delta_c^2/(2snr) + 2\mathcal{C}/(L\nu)$. Thus

$$\pi_{(\ell)} = \frac{\max\{u_\ell, u_{cut}\}}{L\, sum} = \frac{2\mathcal{C}}{L\nu}\frac{\max\{u_\ell, u_{cut}\}}{1+\delta^2_{sum}}. \tag{4.31}$$

In this case $C_{\ell,R,h} = C_{\ell,R}(1-h)$ satisfies

$$C_{\ell,R,h} = \max\{u_\ell, u_{cut}\}\, \mathcal{C}'/R \tag{4.32}$$

where

$$\mathcal{C}' \;=\; \tilde{\mathcal{C}}\frac{1-h}{1+\delta^2_{sum}}.$$

As we have seen if $c = 0$ the $C_{\ell,R,h} = u_\ell\, \mathcal{C}'/R$ takes a similar form but without the max, where for $c = 0$ the $\mathcal{C}' = \tilde{\mathcal{C}}\,(1-h)$.

4.18.2 Formulation and evaluation of the integral $g(x)$

From the expression of $g_L(x)$ given in (4.29) and using (4.31) and (4.32), one gets that

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

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 $g_{num}(x)/(1+\delta^2_{sum})$ where the numerator is

$$g_{num}(x) = \frac{2\mathcal{C}}{\nu}\int_0^1 \max\{e^{-2\mathcal{C}t}, u_{cut}\}\Phi\big(\mu(x, \max\{e^{-2\mathcal{C}t}, u_{cut}\}\mathcal{C}'/R)\big)dt. \tag{4.33}$$

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

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

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 $g_{num}(x)$. Let $z_x^{low} = \mu(x, u_{cut}\,\mathcal{C}'/R)$ and $z_x^{max} = \mu(x, \mathcal{C}'/R)$ and let $\delta_a = a/\sqrt{2\log B}$. For emphasis we write out that $z_x = z_x^{low}$ takes the form

$$z_x = \left[\frac{\sqrt{u_{cut}\mathcal{C}'/R}}{\sqrt{1\,-\,x\nu}} - (1+\delta_a)\right]\sqrt{2\log B}.$$

Notice that $g_{num}(x)$ in (4.33) is equal to

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

plus the function

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

The above function can also be written as $[\delta_c + D(\delta_c)]\Phi(z_x)(1-\nu)/\nu$.

Change the variable of integration from t to $u = e^{-2\mathcal{C}t}$, to see that

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

Now since

$$\Phi(\mu) = \int 1_{\{z\le\mu\}}\,\phi(z)\,dz,$$

it follows that

$$g_{num,1}(x) = \int\int 1_{\left\{u_{cut}\le u\le 1\right\}} 1_{\left\{z\le\mu(x,u\,\mathcal{C}'/R)\right\}}\,\phi(z)\,dz\,du/\nu,. \tag{4.34}$$

In (4.34), the inequality

$$z \le \mu(x, u\,\mathcal{C}'/R)$$

is the same as

$$\sqrt{u}\ \ge\ \sqrt{u_x R/\mathcal{C}'}\,\big(1 + (z+a)/\sqrt{2\log B}\,\big),$$

provided $z_x^{low} \le z \le z_x^{max}$. Here $u_x = 1 - x\nu$. Thereby for all z the length of this interval of values of u can be written as

$$\left[1 - \max\left\{u_x\frac{R}{\mathcal{C}'}\Big(1+\frac{z+a}{\sqrt{2\log B}}\Big)_+^2, u_{cut}\right\}\right]_+.$$

Thus

$$g_{num}(x) = [\delta_c + D(\delta_c)]\Phi(z_x)\frac{1-\nu}{\nu} + \frac{1}{\nu}\int \left[1 - \max\left\{u_x \frac{R}{\mathcal{C}'}\Big(1 + \frac{z+a}{\sqrt{2\log B}}\Big)_+^2, u_{cut}\right\}\right]_+ \phi(z)dz \tag{4.35}$$

where $z_x = z_x^{low}$.

Lemma 19. Derivative evaluation. *The derivative $g'_{num}(x)$ is equal to*

$$\frac{z'_x}{snr}[\delta_c + D(\delta_c)]\phi(z_x) + \int_{z_x^{low}}^{z_x^{max}} \frac{R}{\mathcal{C}'}\big(1+\delta_a+\delta_z\big)^2\phi(z)dz,$$

with

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

where $\delta_z = z\delta$ with $\delta = 1/\sqrt{2\log B}$.

In particular if $c = 0$ the derivative $g'(x)$ may be expressed as

$$\frac{R}{\mathcal{C}'}\int_{z_x^{low}}^{z_x^{max}} \big(1+\delta_a+\delta_z\big)^2\phi(z)dz,$$

and then, if also $R = \mathcal{C}'/[(1+\delta_a)^2(1+r/\log B)]$ with $r \geq 1/[2(1+\delta_a)^2]$, the difference $g(x) - x$ is a decreasing function of x.

Proof. The integrand in the expression for g_{num} is continuous and piecewise differentiable in x, and the integral of its derivative is expressed above. It is in agreement with the derivative with respect to x of the integral $g_{num}(x)$.

In the $c = 0$ case this derivative specializes to

$$g'(x) = \frac{R}{\mathcal{C}'}\int_{z_x^{low}}^{z_x^{max}} \big(1+\delta_a+\delta_z\big)^2\phi(z)dz,$$

which is less than

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

which by the choice of R is less than 1 for $r \geq 1/[2(1+\delta_a)^2]$. Then $g(x) - x$ is decreasing as it has a negative derivative. □

Corollary 20. A lower bound. *The $g_{num}(x)$ is at least*

$$g_{low}(x) = [\delta_c + D(\delta_c)]\Phi(z_x)\frac{1-\nu}{\nu} +$$

$$\frac{1}{\nu}\int_{z_x^{low}}^{\infty}\Big[1-(R/\mathcal{C}')u_x\left(1+(z+a)/\sqrt{2\log B}\right)^2\Big]\,\phi(z)dz.$$

This $g_{low}(x)$ is equal to

$$\Big[1+\frac{1-\nu}{\nu}D(\delta_c)\Big]\Phi(z_x) + \Big[x+\delta_R\frac{u_x}{\nu}\Big]\left[1-\Phi(z_x)\right]$$

$$-\,2(1+\delta_a)\frac{R}{\mathcal{C}'}\frac{u_x}{\nu}\,\frac{\phi(z_x)}{\sqrt{2\,\log B}} - \frac{R}{\mathcal{C}'}\frac{u_x}{\nu}\,\frac{z_x\phi(z_x)}{2\log B}.$$

where

$$\delta_R = 1-\frac{R}{\mathcal{C}'}\left[(1+\delta_a)^2+1/(2\log\,B)\right].$$

Moreover, this $g_{low}(x)$ has the analogous integral characterization as in (4.35)*, but with removal of the outer positive part restriction, and accordingly $g_{low}(x)$ has derivative $g'_{low}(x)$ given by*

$$\frac{z'_x}{snr}[\delta_c + D(\delta_c)]\phi(z_x) + \frac{R}{\mathcal{C}'}\int_{z_x}^{\infty}\big(1+\delta_a+\delta_z\big)^2\phi(z)dz,$$

where the first part vanishes when $c=0$.

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}. The evaluation of $g_{low}(x)$ is fairly straightforward after using $z\phi(z) = -\phi'(z)$ and $z^2\phi(z) = \phi(z) - (z\phi(z))'$. Also use that $\Phi(z)$ tends to 1, while $\phi(z)$ and $z\phi(z)$ tend to 0 as $z\to\infty$. This completes the proof of Corollary 20. □

Remark: What we gain with this lower bound is simplification because the result depends on x only through $z_x = z_x^{low}$.

With the rate R taken to be not more than $\mathcal{C}'/(1+\delta_a)^2$, we write it as

$$R = \frac{\mathcal{C}'}{(1+\delta_b)^2} = \frac{\mathcal{C}'}{(1+\delta_a)^2(1+r/\log B)} \tag{4.36}$$

where $\delta_b = b/\sqrt{2\log B}$.

4.18.3 Showing $g(x)$ is greater than x

This section shows that $g_L(x)$ is accumulative, that is, it is at least x for most of the interval from 0 to 1, using the lower bound which is here explored further.

The preceding subsection established that $g_L(x)$ is at least $(integ/sum)g(x)$, where

$$\frac{integ}{sum} = \frac{1+D(\delta_c)/snr}{1+\delta_{sum}^2} = 1 - \frac{2\theta\mathcal{C}/(L\nu)}{1+\delta_{sum}^2},$$

which is at least $1 - 2\mathcal{C}/(L\nu)$. It follows that $g_L(x) - x$ is at least

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

So to establish positivity of $g_L(x) - x$ it is be enough to show that $g(x) - x$ is at least $2\mathcal{C}/(L\nu)$.

Furthermore, in view of the relationship between $g(x)$ and $g_{num}(x)$, work with $g(x) - x$ in the form

$$g(x) - x = \frac{h(x)}{1+D(\delta_c)/snr},$$

where

$$h(x) = g_{num}(x) - x - xD(\delta_c)/snr$$

and establish the required positivity of $g(x) - x$ via the positivity of $h(x)$ or its lower bound $h_{low}(x) = g_{low}(x) - x - xD(\delta_c)/snr$. Indeed, if either of these is greater than or equal to a positive value gap_{num} on an interval $[0, x^*]$, then also on such interval,

$$g(x) - x \geq gap = \frac{gap_{num}}{1+D(\delta_c)/snr}.$$

As above we use a rate of the form

$$R = \frac{\mathcal{C}'}{(1+\delta_a)^2(1+r/\log B)} = \frac{\mathcal{C}'}{(1+\delta_b)^2}.$$

Recall that $z_x = z_x^{low}$ is a strictly increasing function of x, with values in the interval $I_0 = [z_0, z_1]$ for $0 \leq x \leq 1$. For values z in I_0, let $x = x(z)$ be the choice for which $z_x = z$.

Also the distance of $x = x(z)$ from 1 can be expressed as a function of z as

$$1 - x = \frac{1}{snr} \left[\frac{(1+\delta_c)(1+\delta_b)^2 - (1+\delta_a+\delta_z)^2}{(1+\delta_a+\delta_z)^2} \right].$$

Then if z_{up} be the value of z at which $(1+\delta_a+\delta_z)^2 = (1+\delta_c)(1+\delta_a)^2$ and $x_{up} = x(z_{up})$ be the corresponding value of x. Then from the above we get that

$$1 - x_{up} \;=\; \frac{1}{snr} \left[\frac{r}{\log B} \right].$$

We remark that the value of x_{up} is the same for all values of $c \geq 0$. The value of x_{up} is used in giving an upper end point of a range of sufficient positivity of $g(x) - x$. As it is desirable that this upper endpoint be not far from 1, one may restrict attention to cases with $snr \geq 1/\sqrt{2 \log B}$, say, so that $1 - x_{up}$ remains not more than $2r/\sqrt{2 \log B}$.

Let $(1+\delta_{b^*})^2 = (1+\delta_a)^2 + 1/(2 \log B)$ and let r_0 be such that

$$(1+\delta_{b^*})^2 = (1 + \delta_a)^2 (1 + r_0/\log B).$$

Further, let $d = (r - r_0)(1+\delta_a)$.

At $x = x(z)$ it is fairly straightforward to see that one can express $g_{low}(x)$ in terms of z as

$$g_{low}(x(z)) = x(z) \;+\; \frac{1-\nu}{\nu} G(z)$$

where

$$\begin{aligned} G(z) &= \frac{(1+\delta_c)(1+\delta_a)}{(1+\delta_a+\delta_z)^2} \frac{d}{\log B} \\ &+ \left[\frac{(1+\delta_c)(1+\delta_{b^*})^2}{(1+\delta_a+\delta_z)^2} - 1 + D(\delta_c) \right] \Phi(z) \\ &- \frac{2(1+\delta_a)(1+\delta_c)}{(1+\delta_a+\delta_z)^2} \frac{\phi(z)}{\sqrt{2 \log B}} \;-\; \frac{(1+\delta_c)}{(1+\delta_a+\delta_z)^2} \frac{z\phi(z)}{2 \log B}. \end{aligned} \tag{4.37}$$

Likewise, $g_{low}(x) - x$ simplifies to $G(z)/snr$.

For this next lemma, we take the power allocation $P_{(\ell)}$ to be proportional to $u_\ell = \exp\{-2\mathcal{C}\frac{\ell-1}{L}\}$ and consider the case of no leveling, that is, $c = 0$.

Lemma 21. Positivity of $g_{low}(x) - x$ with no leveling. *Let rate R be of the form* (4.36)*, with*

$r \geq r_0$. *Let*

$$d_1 = \sqrt{2 \log B}/\sqrt{2\pi}.$$

Then, for $0 \leq x \leq x_{up}$ *the difference* $g_{low}(x) - x$ *is greater than or equal to*

$$\frac{1}{snr} \left[\frac{(d - d_1)(1+\delta_a) + 1/4}{(1+\delta_a)^2 \log B} \right].$$

In particular, expressing the rate as $R = \mathcal{C}'/(1+\delta_b)^2$, *suppose*

$$b = a + \frac{1}{\sqrt{2\pi}} + \frac{extra}{2\sqrt{2 \log B}} \tag{4.38}$$

with $extra > (1/2)(1 - 1/\pi)$. *Then, for* $0 \leq x \leq x_{up}$ *the difference* $g_{low}(x) - x$ *is greater than or equal to*

$$\frac{1}{snr} \left[\frac{extra - (1/2)(1 - 1/\pi)}{(1+\delta_a)^2 (2 \log B)} \right]. \tag{4.39}$$

Proof of Lemma 21. With $c = 0$ the functions $g_{num}(x)$ and $g(x)$ are the same and have lower bound $g_{low}(x)$. By corollary 20, $g_{low}(x)$ has derivative bounded by

$$\int_{-\infty}^{\infty} \frac{\left(1+\delta_a+\delta_z\right)^2}{(1+\delta_b)^2} \phi(z) dz = \frac{(1+\delta_{b^*})^2}{(1+\delta_b)^2}$$

where $(1+\delta_{b^*})^2 = (1+\delta_a)^2 + 1/(2 \log B)$. This bound is less than 1 for $b \geq b^*$, that is for $r \geq 1/[2(1+\delta_a)]$. Thus $g_{low}(x) - x$, like $g(x) - x$, is decreasing as it has a negative derivative. Likewise by the correspondence between x and z, it follows that $G(z)$ is decreasing in z, so $G(z) \geq G(0)$ for $z \leq 0$.

To complete the proof, evaluate $g_{low}(x) - x$ at the point $x = x_{up}$. With $c = 0$ the point x_{up} is the choice where $z_x = 0$. Accordingly, by (4.37), the value of $g_{low}(x) - x$ there is $[(1-\nu)/\nu]G(0)$, where the value of $G(0)$ is given by the following

$$\frac{d\,(1+\delta_a)}{(1+\delta_a)^2 \log B} + \frac{\Phi(0)}{(1+\delta_a)^2 (2 \log B)} - \frac{2\,\phi(0)}{(1+\delta_a)\sqrt{2 \log B}},$$

which is

$$\frac{(d - d_1)(1+\delta_a) + 1/4}{(1+\delta_a)^2 \log B}.$$

The first term of the above arises from the representation of $(1+\delta_b)^2 - (1+\delta_{b^*})^2$ divided by $(1+\delta_a)^2$

as $d/\log B$. Alternatively, this difference $(1+\delta_b)^2 - (1+\delta_a)^2 - 1/(2\log B)$ may be expanded as $(\delta_b - \delta_a)(2 + \delta_b + \delta_a) - 1/(2\log B)$ so that the $G(0)(1+\delta_a)^2$ may be written

$$\left[\frac{2(b-a-1/\sqrt{2\pi}\,)}{\sqrt{2\log B}} + \frac{b^2 - a^2 - 1/2 - 2a/\sqrt{2\pi}}{2\log B}\right].$$

It is positive by choosing b slightly larger than $a+1/\sqrt{2\pi}$. Indeed, setting b as in (4.38), we have at $x = x_{up}$ that $g_{low}(x)-x$ is at least the quantity in (4.39), which is positive for $extra > (1/2)(1-1/\pi)$.

Consequently, in view of the monotonicity, it follows for $0 \le x \le x_{null}$, that $g_{low}(x) - x$ is at least that same value. This completes the proof. □

In the above form of b, the terms are small enough that it provides a practical rate $\mathcal{C}'/\big(1 + b/\sqrt{2\log B}\,\big)^2$ reasonably close to capacity with moderate B. Nevertheless, it would be nice to remove the $1/\sqrt{2\pi}$ part so that with suitable a a rate closer to capacity is achieved for large B. Another way to say it is that we would like to arrange for r to be of smaller order. For that reason we next take advantage of the modification to the power allocation in which it is slightly leveled using a positive c.

Monotonicity of $g(x) - x$ and of $g_{low}(x) - x$ is demonstrated in the above proof for the $c = 0$ case. It what follows we take advantage of more detailed shape properties that include the case of positive c.

The function of interest is $g_{num}(x) - x - xD(\delta_c)/snr$. Work with the lower bound $g_{low}(x) - x - D(\delta_c)/snr$ which evaluates to

$$\frac{1}{snr}\big[G(z) - D(\delta_c)\big].$$

Now write

$$G(z) = \frac{(1+\delta_c)}{(1+\delta_a+\delta_z)^2}\ \frac{A(z)}{2\log B}$$

where

$$A(z) = 2d\,(1+\delta_a)\ -\ 2(1+\delta_a)\phi(z)\sqrt{2\,\log B}\ -\ z\,\phi(z)\ +$$

$$\Big[(1+\delta_{b^*})^2 - (1-\Delta_c)(1+\delta_a+\delta_z)^2\Big]\Phi(z)\,2\log B,$$

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

The following lemma characterizes the shape of $A(z)$.

Lemma 22. Shape properties. *Define* $\Delta_c = \log(1+\delta_c)$. *Consider three cases. When* $\Delta_c \le$

$2/\big(\tau^2/4+3\big)$, *the function* $A(z)$ *is decreasing and concave for* $z \geq -\tau$ *and, moreover,* $G(z)$ *and* $G(z)-x(z)D(\delta_c)$ *are decreasing there. When* $\Delta_c \geq 2/\big(\tau^2/4+2\big)$, *the function* $A(z)$ *is unimodal for* $z \geq -\tau/2$. *When* Δ_c *is between* $2/\big(\tau^2/4+3\big)$ *and* $2/\big(\tau^2/4+2\big)$, *the function* $A(z)$ *is unimodal for* $z \geq -\tau/2+1$.

Proof of Lemma 22. The expression $[\delta_c + D(\delta_c)]/(1+\delta_c)$ simplifies to $\Delta_c = \log(1+\delta_c)$ using the definition of $D(\delta_c)$. Differentiating and collecting terms obtain that $a(z) = A'(z)$ is

$$-2(1-\Delta_c)\,(1+\delta_a+\delta_z)\,\Phi(z)\sqrt{2\log B}$$

$$+\ \Delta_c\,(1+\delta_a+\delta_z)^2\,\phi(z)\,2\log B\,.$$

Consider values of z in $I_\tau = (-\tau, \infty)$ to the right of $-\tau$, which is where the factor $(1+\delta_a+\delta_z)$ is strictly positive. This includes $[z_0, z_1]$. Factoring out $(1+\delta_a+\delta_z)\sqrt{2\log B}$, the sign behavior of $a(z)$ is determined by

$$M(z) \;=\; -2(1-\Delta_c)\,\Phi(z) \;+\; \Delta_c\,(1+\delta_a+\delta_z)\,\phi(z)\,\sqrt{2\log B}.$$

Consider c with $\Delta_c < 1$. Note that this function $M(z)$ is continuous, starts out negative at $z = -\tau$, and is also negative for large z as it converges to $-2(1-\Delta_c)$. Thus $A(z)$ is initially decreasing in I_τ and also decreasing for large z. If c is 0 the function $M(z)$ reduces to $-2\Phi(z)$ which is negative and whence $A(z)$ is decreasing in all of I_τ. So consider positive c with $\Delta_c < 1$. Consider the derivative of $M(z)$ given by

$$M'(z) = -\big[2-3\Delta_c \;+\; \Delta_c\,(z/\delta)(1+\delta_a+\delta_z)\big]\,\phi(z),$$

where $\delta = 1/\sqrt{2\log B}$. The expression in brackets is a quadratic function of z centered and extremal at $z_{cent} = -\tau/2$. This quadratic attains the value 0 only if c is large enough that $\Delta_c\tau^2/4$ is at least $2-3\Delta_c$, that is, for Δ_c at least $2/(\tau^2/4+3)$. Let c^* be the value where these two quantities match.

For $c < c^*$, the $M'(z)$ stays negative and consequently $M(z)$ is decreasing, so $M(z)$ and $a(z)$ remains negative for $z > -\tau$, so $A(z)$ is decreasing in I_τ. Consequently $G(z)$ and $G(z) - x(z)D(\delta_c)$ are decreasing in I_τ and hence in the domain of interest $[z_0, z_1]$ where $x(z)$ is between 0 and 1. Moreover, $a(z) = (1+\delta_a+\delta_z)M(z)\sqrt{2\log B}$ has

$$a'(z) = M(z) + (1+\delta_a+\delta_z)M'(z)\sqrt{2\log B},$$

also negative in I_τ, so $A(z)$ is concave there.

For $c \geq c^*$, for which the function $M'(z)$ does cross 0, this $M'(z)$ is positive in an interval of values of z centered at $z_{cent} = -\tau/2$ and heading up to a point z^* in I where $z(1+\delta_a+\delta_z) = -(2-3\Delta_c)\delta/\Delta_c$. This point z^* is found to be

$$\frac{-\tau + \sqrt{\tau^2 + 4(3 - 2/\Delta_c)}}{2}.$$

In this interval the function $M(z)$ is increasing.

Let's see whether $M(z)$ is positive, at least at z_{cent}. Use the inequality $\Phi(z) \leq \phi(z)/(-z)$ for $z < 0$. This lower bound is sufficient to demonstrate positivity in the interval of values of z centered at the same point, provided $\Delta_c\tau^2/4$ is at least $2(1-\Delta_c)$, that is, $\Delta_c \geq 2/(\tau^2/4+2)$. Let's call c^{**} the point where these match. For $c \geq c^{**}$, this interval is where the same quadratic $z(1+\delta_a+\delta_z)$ is less than $-2(1-\Delta_c)\delta/\Delta_c$. For such c, the $M(z)$ is positive at $-\tau/2$ and furthermore increasing from there up to z^*, while, to the right of z^*, 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 $-\tau/2$.

Whereas if c is between c^* and 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^*, that $M(z)$ is decreasing there and hence it and $a(z)$ has at most one root to the right of z^*, so $A(z)$ is unimodal there. In the notion of unimodality (a single peak in an interval) we are allowing for the possibility of monotonicity, that is, the peak may be at the end point z^*. Being less than c^{**}, the value of c is small enough that $2/\Delta_c > \tau^2/4 + 2$, and hence z^* is not more than $[-\tau + \sqrt{4}]/2$ which is $-\tau/2 + 1$. This completes the proof of lemma 22. □

The following lemma characterizes the minimum value of $A(z)$ and gives lower bound for $h_{low}(x)$ for a suitable interval.

Lemma 23. The minimum value of the gap. *Set*

$$z_{up} = \zeta = \sqrt{2\log_+\Big(\frac{\sqrt{2\log B}}{d_0\sqrt{2\pi}}\Big)}.$$

for some positive $d_0 \leq \sqrt{2\log(B)}/\sqrt{2\pi}$.

(a) If

$$\Delta_c \geq 2/(\tau^2/4+2)$$

then the minimum value of $A(z)$ *on* $[-\tau, z_{up}]$ *is equal to that over* $[-\tau,-\tau/2]\cup z_{up}$.

(b) Furthermore define

$$\mathit{diff} = \frac{2(1+\delta_a)d_0 - 1/2 - \tau^2 D(\delta_c)\Phi(z_{up})}{2(1+\delta_a)}.$$

Assume $d > d_1$, *where*

$$d_1 = \tau^2 D(\delta_c)/(2(1+\delta_a)) + (\mathit{diff})_+ + \epsilon_B.$$

Here $\epsilon_B = 2\tau\phi(-\tau/2)/(2(1+\delta_a))$ *is a small term that is polynomially in* $1/B$. *For such choices of* d, $h_{low}(x)$ *is positive for* $-\tau \leq z \leq z_{up}$ *and is at least*

$$gap_{num} = \frac{2(1+\delta_a)(d-d_1)}{snr\,\tau^2}.$$

Proof. Recall that from Lemma 22, for $\Delta_c \geq 2/(\tau^2/4+2)$, we have that $A(z)$ is unimodal for $z \geq \tau/2$. Consequently, the minimum of $A(z)$ over $[-\tau, z_{up}]$ is equal to that of over $[-\tau,-\tau/2]\cup z_{up}$. This proves part (a).

Let's examine $A(z)$ for $-\tau \leq z \leq -\tau/2$. For z in this range $A(z)$ is at least $2d(1+\delta_a) - 2\tau\phi(z)$. This is seen by observing that in the expression for $A(z)$ given in the beginning of the proof of Lemma 22, the third and fourth terms are positive for $z \leq 0$.

Noting that $2\tau\phi(z) \leq 2\tau\phi(-\tau/2)$ for z in this interval, we have that the minimum of $A(z)$ on $[-\tau,-\tau/2]$ is at least $2(d-\epsilon_B)(1+\delta_a)$, where ϵ_B is a term that is polynomially small in $1/B$.

Next, let's evaluate the value of $A(z)$ at z_{up}. Write the function $A(z)$ as

$$A_0(z) + \Big[(1+\delta_a)^2 - (1+\delta_a+\delta_z)^2/(1+\delta_c)\Big]\Phi(z)(2\log B)$$

$$+ D(\delta_c)(1+\delta_a+\delta_z)^2\,\Phi(z)\,(2\log B)/(1+\delta_c)$$

where

$$A_0(z) \;=\; 2(1+\delta_a)[d - \phi(z)\sqrt{2\log B}] \;+\; \Phi(z) - z\,\phi(z).$$

Recalling that $(1+\delta_a+\delta_{z_{up}})^2 = (1+\delta_c)(1+\delta_a)^2$, we have that at $z = z_{up}$, $A(z)$ reduces to

$$A_0(z_{up}) + \tau^2 D(\delta_c)\Phi(z_{up}).$$

The relationship between δ_c and z_{up} can also be expressed as

$$1+\delta_c \;=\; (1+z_{up}/\tau)^2.$$

Concerning $A_0(z)$ observe that for z non-negative, $\Phi(z)$ is at least $\Phi(0)+z\phi(z)$, because the normal probability of the interval from 0 to z is at least the width of the interval times the density minimum. Also $\Phi(0) = 1/2$. Thus the last two terms of $A_0(z)$ which consists of $\Phi(z) - z\,\phi(z)$ is always at least $1/2$, and tends to 1 for large z.

Consider $\zeta = \zeta(\phi) = \sqrt{2\log(1/(2\pi\phi))}$, the positive inverse of the standard normal density function. Note that since $d_0 \le \sqrt{2\log B}/\sqrt{2\pi}$, we have that $d_0/\sqrt{2\log(B)}$ is in the range of ϕ. Defining z_{up} to be ζ evaluated at $\phi = d_0/\sqrt{2\log B}$, we get that

$$z_{up} \;=\; \zeta \;=\; \sqrt{2\log\left(\sqrt{2\log B}/(\sqrt{2\pi}\,d_0)\right)}.$$

Also, since $\phi(z_{up}) = \phi(\zeta) = d_0/\sqrt{2\log B}$, it follows from the above expression for $A_0(z_{up})$ that

$$A_0(z_{up}) \ge 2(1+\delta_a)(d-d_0) \;+\; 1/2$$

and $A(z)$ is at least,

$$2(1+\delta_a)[d-d_0] \;+ 1/2 + \tau^2 D(\delta_c)\Phi(z_{up}).$$

which is

$$2(1+\delta_a)(d - \mathit{diff})$$

Thus combining this with part (a) and the result on the minimum from $[-\tau, -\tau/2]$, we get that the minimum of $A(z)$ on $[-\tau, z_{up}]$ is at least,

$$2(1+\delta_a)(d - (\mathit{diff})_+ - \epsilon_B).$$

We now use this to lower bound h_{low} on $[-\tau, z_{up}]$. Recall that,

$$h_{low}(x) = \frac{1-\nu}{\nu}\left(\frac{(1+\delta_c)}{(\tau+z)^2}A(z) - x_{up}D(\delta_c)\right)$$

for $-\tau \leq z \leq z_{up}$. The above can be written as

$$\frac{1-\nu}{\nu}\frac{(1+\delta_c)}{(\tau+z)^2}\left(A(z) - \frac{(\tau+z)^2}{(1+\delta_c)}x_{up}D(\delta_c)\right)$$

which is at least

$$\frac{1-\nu}{\nu}\frac{A^*(z)}{\tau^2},$$

where $A^*(z) = A(z) - \tau^2 x_{up}D(\delta_c)$. The above follows since for $-\tau \leq z \leq z_{up}$ we have that $(\tau+z)^2 \leq \tau^2(1+\delta_c)$.

The result on gap_{num} immediately follows from using the lower bound on $A(z)$ developed above and noting that $(1-\nu)/\nu$ is $1/snr$. □

4.18.4 Choices of a, r, c that optimize the overall rate drop

Here we focus on evaluation of a, r, and c that optimize our summary expressions for the rate drop, based on the lower bounds on $g_L(x) - x$. For the time being let's assume that for a particular d_0, yet to be specified and the conditions in part (a) of the above lemma are satisfied.

Recall that the rate of our inner code is

$$R = \mathcal{C}\frac{1-h}{(1+\delta_{sum}^2)(1+\delta_a)^2(1+r/\log B)}.$$

With $0 < \eta < gap_L$ and $f > f^*$ satisfying $4f \leq (gap_L - \eta)^2$, per our theory, we reliably have that the weighted fraction of correct terms decoded rises to $g(x_{up}) - \eta$, which is at least $x_{up} + gap - \eta$. Further, the weighted fraction of false alarms is not more than mf. This is accomplished in not more than $m = 1/inc$ steps, where if we arrange $f = (gap_L - \eta)^2/4$ then the increment each step is at least $inc = (gap_L - \eta)/2$. By the above development the weighted failed detection rate is $(1 - x_{up}) - (gap - \eta)$ and the weighted false alarm rate is

$$mf = (gap_L - \eta)/2.$$

With sufficient size L, it is possible to arrange smaller η and f closer to f^* while retaining reliability. To arrange the indicated value of $f = (gap_L - \eta)^2/4$ and hence of f^*, one solves for a using the inverse of the normal distribution to produce $f^* = (B-1)\bar{\Phi}(\sqrt{2\log B} + a)$. Or, for simplicity, arrange instead to use the f^* upper bound

$$\frac{1}{\sqrt{2\pi}\sqrt{2\log B}} e^{-a\sqrt{2\log B}},$$

by setting

$$a = \frac{\log\big(1/(f^*\sqrt{2\pi}\sqrt{2\log B})\big)}{\sqrt{2\log B}},$$

which, when evaluated at the targeted f^*, e.g. $f/2$, produces a value of δ_a of order $(\log\log B)/(\log B)$.

Bounds on un-weighted fractions of failed detections and false alarms are obtained by multiplying the weighted fractions by the factor

$$\text{fac} = \frac{snr\,(1+\delta_{sum}^2)}{2\mathcal{C}\,(1+\delta_c)}.$$

To see this notice that for a given weighted fraction, the maximum possible un-weighted fraction would be if we assume that all the failed detection or false alarms came from the section with the smallest weight. This would correspond to the section with weight $\pi_{(L)}$, where we recall that $\pi_{(L)} = 2\mathcal{C}\,(1+\delta_c)/(L\,snr\,(1+\delta_{sum}^2))$.

For large L, δ_{sum}^2 would be like $\delta_c^2/2$ so we can assume that fac is bounded by $snr/(2\mathcal{C})$. Notice that $snr/(2\mathcal{C})$ is greater that 1 and is near 1 for small snr. For large L, δ_{sum}^2 would be like $\delta_c^2/2$ so we can assume that fac is bounded by $snr/(2\mathcal{C})$. Notice that $snr/(2\mathcal{C})$ is greater that 1 and is near 1 for small snr.

So with high reliability, the total fraction of mistakes is bounded by

$$\delta_{mis} = \frac{snr}{2\mathcal{C}}\left[(1-x_{up}) \;-\; (gap_L - \eta)/2\right].$$

If the outer Reed-Solomon code has distance $d_{RS}/L = \delta$ designed to be at least $2\delta_{mis}$ then any occurrences of a fraction of mistakes less than δ_{mis} are corrected. The overall rate of the code is $R_{total} = (1-\delta)R$.

Sensible values of the parameters a, r, and c can be obtained by optimizing the overall rate under a presumption of small error probability, using simplifying approximations of our expressions.

Reference values (corresponding to large L) are obtained by considering what the parameters become with $\eta = 0$ and $f = f^*$. We also take gap_L replaced by gap, h by 0 and δ_{sum}^2 replaced by $D(\delta_c)/snr$ in the expression for the rate of the inner code.

One may optimize this expression to obtain reference values of the parameters. Set a such that

$$a\sqrt{2\log B} = \log\left[1/\left(f^*\sqrt{2\pi}\sqrt{2\log B}\right)\right].$$

Now, the minimum gap for $0 \leq x \leq x_{up}$ is given by $gap = gap_{num}/(1 + D(\delta_c)/snr)$, which using the expression from Lemma 23 is given by

$$gap = \frac{2(1+\delta_a)(d-d_1)}{snr\,\tau^2(1+D(\delta_c)/snr)}.$$

which can also be written as

$$gap = \frac{(r-r_1)}{snr\,\log(B)(1+D(\delta_c)/snr)}.$$

where $r_1 = r_0 + d_1/(1+\delta_a)$

Then using $1 - x^* = 1 - x_{up}$ and the expression given above for the gap, the bound on the fraction of un-weighted failed detections would be

$$\delta_u = \frac{1}{2\mathcal{C}}\frac{r}{\log B}$$

and the bound on the total fraction of false alarms would be

$$\delta_e = \frac{1}{2\mathcal{C}}\left[\frac{r-r_1}{(2\log B)(1+D(\delta_c)/snr)}\right].$$

The role of $D(\delta_c)$ is captured primarily by its appearance in the inner code rate, whereas here in the mistake rate its effect is of smaller order than the other terms in the numerator and in the denominator, so let's drop it from δ_e. This yields a simplified approximate expression for the mistake rate

$$\delta_{mis} = \frac{1}{4\mathcal{C}\,\log B}\left[r + r_1\right].$$

Express the overall communication rate

$$R_{total} = (1-2\delta_{mis})\frac{\mathcal{C}}{(1+D(\delta_c)/snr)(1+\delta_a)^2(1+r/\log B)}$$

as

$$R_{total} = \frac{\mathcal{C}}{1+drop}.$$

Ignoring negligible terms, the drop from capacity is

$$\big[(1/snr)D(\delta_c) + 2\delta_a + r/\log B\big] + 2\delta_{mis}.$$

Thus we find that

$$drop = \frac{drop_{num}}{2\log B},$$

where $drop_{num}$ is given by

$$\frac{c^2}{2\,snr} + 2a\sqrt{2\log B} + 2r.$$

Recall that f^* chosen to match $gap^2/4$ which is nearly

$$\frac{(r-r_1)^2}{snr^2(2\log(B))^2},$$

ignoring the effect of the $1 + D(\delta_c)/snr$ term. Thus the $2a\sqrt{2\log B}$ term becomes

$$4\log\left(\frac{snr(2\log(B))}{r-r_1}\right) - \log B - \log\big(4\pi\big).$$

Next let's determine the best choice of r. The part of the rate drop that depends on r is proportional to

$$4\log 1/[r-r_1] + 2\,r.$$

Taking the derivative with respect to r and setting it to zero reveals a best value of r of

$$r^* = r_1 + 2$$

at which the optimum choice of a becomes

$$a = (3/2)\log(\log(B))/\sqrt{2\log(B)} + a_{ext},$$

where,

$$a_{ext} = \frac{2\log\left[snr/((4\pi)^{.25})\right]}{\sqrt{2\log(B)}}.$$

Also with the above choice of r we get that $drop_{num}$ is $3\log\log B$ plus the following

$$c^2/(2\,snr) \;+\; 4\log\big(snr\big) - \log(4\pi) + 4 + 2r_1.$$

It is instructive to approximate this drop in the when snr large compared to $\log\log B$. Using the approximation that $(1/2)\log(snr)$ nearly equal to $(1/2)\log(1+snr)$ and that for such moderately large snr cases, we can ignore terms of order $(1/snr)$. The approximate rate drop then is,

$$\frac{3\log\log B + 2\,(1+1/\mathcal{C})\,r_1 + 8\mathcal{C} + 1.46}{2\log B},$$

From the above, if one wants the rate close to the capacity, it requires $\log B$ large compared to $4\mathcal{C}$. This entails having B large compared to $(1+snr)^2$.

Further for snr values near 1, the approximate rate drop becomes,

$$\frac{3\log\log B + c^2/2 + 2r_1 + 1.46}{2\log B}.$$

We see that for such snr terms although there is no $8\mathcal{C}$ term, the term $c^2/2$ comes into play. This term is of order $\log\log B$ as we will see below.

We now specify the value of d_0 which goes into determining the value of z_{up} and hence c. We also need to ensure that for this specified value the requirement in part (a) of Lemma 23 is satisfied.

Before doing this, recall that $D(\delta_c)$ is near $\delta_c^2/2$. From $1+\delta_c = (1+\zeta/\tau)^2$, we have that δ_c is near $2\zeta/\tau$, making $D(\delta_c)$ near $2\zeta^2/\tau^2$. Thus we get that *diff* is near

$$\frac{2(1+\delta_a)d_0 - 1/2 - 2\zeta^2\Phi(z_{up})}{2(1+\delta_a)}.$$

Now using $\Phi(z_{up}) \geq 1/2$ we get that the above approximation for *diff* can be bounded by

$$\frac{2(1+\delta_a)d_0 - 1/2 - \zeta^2}{2(1+\delta_a)}.$$

We think a good value of d_0 is

$$d_0 = .5\log(\log(B))/(1+\delta_a),$$

as this choice would make *diff* near zero. To see this, substituting this value of d_0 in the approximation for *diff* given above we get that it is equal to

$$\frac{-1/2 - 2\log\left((2\sqrt{2}/\sqrt{2\pi})(1+\delta_a)\right) + 2\log\log\log(B)}{2(1+\delta_a)},$$

which ignoring the effects of the $1+\delta_a$ term is

$$-.37 + \log\log\log(B).$$

For our purposes we can bound $\log\log\log(B)$ by 1. Indeed, for B as high as 10^7, $\log\log\log(B)$ is 1.02. This means that *diff* can be bounded .63, small for our purposes. Further, the above choice of d_0 makes Δ_c near $\sqrt{\log\log B}/\sqrt{\log B}$. This is easily greater than $2/(\tau^2/4+3)$, which is of the order of $1/\log B$.

We now evaluate the above rate drop for a toy example with $B = e^{12.5}$ and *snr* fixed at say 20. For d_0 value as specified above and for these B and *snr* values, we verified that the condition in part (a) is indeed satisfied. The *drop* is evaluated to be 1.13 which provides a rate $.47\,\mathcal{C}$ or 47% of capacity.

Now consider the case for *snr* values near 1. The evaluation for large L, with the same B as before give a drop of of 1.09 which corresponds to a rate of again around 48% of $\mathcal{C}$.

We now give a more explicit expression for the rate drop by approximating the d_1 term appearing in the expression. We see that from the above approximation for $D(\delta_c)$ and *diff*, ignoring the $1+\delta_a$ and ϵ_B terms, d_1 is near $\zeta^2 + (\mathit{diff})_+$. This can be bounded by $\zeta^2 + .63$ using the bound for *diff* given above. Also notice that ζ^2 can be written as

$$\log(\log(B)) + 2\log\left((2\sqrt{2}/\sqrt{2\pi}\right) - \log\log\log(B).$$

Assume that $\log\log\log(B) \geq 0$ which is true for $B \geq 20$, we get that ζ^2 is at most $\log(\log(B)) + .25$. This gives that d_1 can be bounded by $\log(\log(B)) + .88$, making r_1 bounded by $\log(\log(B)) + 1.38$.

Thus the rate drop for the large signal to noise case can be approximated by,

$$\frac{5\log\log(B) + 8\mathcal{C} + 4.12}{2\log B}.$$

For *snr* near 1, using arguments similar to that above we see that $c^2/2$ is near $2\zeta^2$, which can be bounded by $2\log(\log(B)) + 4\log\left((2\sqrt{2}/\sqrt{2\pi}\right)$. This is $2\log(\log(B)) + .48$. Thus the rate drop can be approximated by

$$\frac{7\log\log(B) + 4.12}{2\log B}.$$

Remark: The above explicit expressions are given to highlight the nature of dependence of the rate drop on *snr* and B. These are quite conservative. For more accurate evaluation of the see subsection 4.17 on computational illustrations.

4.18.5 Definition of $\mathcal{C}^*$ and Proof of Proposition 12

In the previous subsection we determined the optimum r, denoted by r^* that maximized, in an approximate sense, the outer code rate for given *snr* and B values and for large L. This led to explicit expressions for the maximal achievable outer code rate as a function of *snr* and B. We define $\mathcal{C}^*$ to be the inner code rate corresponding to this maximum achievable outer code rate. Thus,

$$\mathcal{C}^* = \mathcal{C}\frac{1-h}{(1+\delta_{sum}^2)(1+\delta_a)^2\left[1 + r^*/\log B\right]}.$$

Similar to above, $\mathcal{C}^*$ can be written a $\mathcal{C}/(1 + drop^*)$ where $drop^*$ can be approximated by

$$\frac{3\log\log B + 2r_1 + 8\mathcal{C} + 5.47}{2\log B}$$

for *snr* large compared to $\log\log B$ and is

$$\frac{3\log\log B + c^2/2 + 2r_1 + 5.47}{2\log B}$$

for *snr* near 1.

Like before, the above drop can be approximated by $(5\log\log B + 8\mathcal{C} + 8.23)/\log B$ for large

snr and by $(7 \log\log B + 8.71)/\log B$ for *snr* near 1.

We now give a proof of our main result.

Proof of Proposition 12: Take $r = r^* + \kappa$. Using

$$(1 + \kappa/\log B)(1 + r^*/\log B) \geq (1 + r/\log B),$$

we find that for the rate R as in Proposition 12, *gap* is at least $(r - r_1)/(snr \log B)$ for $x \leq x_{up}$ with $x_{up} = r/(snr \log B)$.

Take $f^* = (r^* - r_1)^2/(2snr \log B)^2$, so that a is the same as given in the previous subsection. Now, we need to select $c > 1$ and $\eta > 0$ so that

$$cf^* \leq (gap - \eta)^2/4.$$

One sees that we can satisfy the above requirement by taking η as $(1/2)\kappa/(snr \log B)$ and $c = (1 + \kappa\omega/2)^2$.

This choice makes $f_{1,m}$ at most $(gap - \eta)/2$ which is $\sqrt{c}/(2\omega snr \log B)$. Also we select

$$h = \frac{\kappa}{(2 \log B)^{3/2}},$$

which is indeed at most $(2 f_{1,m}\, snr)/\sqrt{2 \log B}$ as required.

The fraction of mistakes,

$$\delta_{mis} = \frac{snr}{2\mathcal{C}} \left[\frac{r}{snr \log B} - (gap - \eta)/2 \right]$$

is calculated as in the previous subsection, except here we have to account for the positive η. Substituting the expression for *gap* and η gives the expression for δ_{mis} as in the proposition.

Now let's look at the error probability. The error probability is given by

$$m e^{-2L_\pi \eta^2 + m c_0} + m e^{-L_\pi f^*_{1,m} c D(c)} + m e^{mh^2/2} e^{-nh^2/2}.$$

Notice that $nh^2/2$ is at least $(L_\pi \log B)h^2/(2\mathcal{C}^*)$, where we use that $L \geq L_\pi$ and $R \leq \mathcal{C}^*$.Thus the

above probability is less than

$$\kappa_1 \exp\{-L_\pi \min\{2\eta^2,\, f^*_{1,m} cD(c),\, h^2 \log B/(2\mathcal{C}^*)\}\}$$

with

$$\kappa_1 = 3m\, e^{m \max\{c_0, 1/2\}},$$

where for the above we use $h < 1$.

Substituting, we see that $2\eta^2$ is $(1/2)\kappa^2/(snr \log B)^2$ and $h^2 \log B/(2\mathcal{C}^*)$ is

$$\frac{1}{16\mathcal{C}^*} \frac{\kappa^2}{(\log B)^2}.$$

Also, one sees that $cD(c)$ is at least $2(\sqrt{c}-1)^2$. Thus the term $f^*_{1,m} cD(c)$ is at least

$$\frac{\kappa^2 \omega}{4(1+\kappa\omega/2) snr \log B}.$$

We bound from below the above quantity by considering two cases viz. $\kappa \leq 2/\omega$ and $\kappa > 2/\omega$. For the first case we have $1+\kappa\omega/2 \leq 2$, so this quantity is bounded from below by $\kappa^2\omega/(8snr \log B)$.For the second case use $\kappa/(1+\kappa\omega/2)$ is bounded from below by $1/\omega$, to get that this term is at least $\kappa/(4snr \log B)$.

Now we bound from below the quantity $\min\{2\eta^2, f^*_{1,m} cD(c),\, h^2 \log B/(2\mathcal{C}^*)\}$ appearing in the exponent. For $\kappa \leq 2/\omega$ this quantity is bounded from below by

$$\kappa_3 \frac{\kappa^2}{(\log B)^2},$$

with κ_3 as in the proposition. For $\kappa > 2/\omega$ this is quantity is at least

$$\min\left\{\kappa_3 \frac{\kappa^2}{(\log B)^2}, \kappa_4 \frac{\kappa}{\log B}\right\}.$$

Also notice that $\mathcal{C}^* - R$ is at most $\mathcal{C}^*\kappa/\log B$. Thus we have that

$$\min\{2\eta^2, f^*_{1,m} cD(c),\, h^2 \log B/(2\mathcal{C}^*)\}$$

is at least

$$\min\left\{\kappa_3\frac{(\mathcal{C}^*-R)^2}{(\mathcal{C}^*)^2},\kappa_4\frac{\mathcal{C}^*-R}{\mathcal{C}^*}\right\}.$$

Further, recalling that $L_\pi = L\nu(1+D(\delta_c)/snr)/(2\mathcal{C})$, we get that $\kappa_2 = \nu(1+D(\delta_c)/snr)/(2\mathcal{C})$, which is near $\nu/(2\mathcal{C})$.

Also regarding the value of m, recall that m is at most $2/(gap-\eta)$. Using the above we get that m is at most $(2\omega snr)\log B$. Thus ignoring the $3m$ term κ_1 is polynomial in B with power $2\omega snr\max\{c_0,1/2\}$.

Part II follows from the properties of Reed Solomon code as given in Chapter 2.

4.19 Proof of lemma 14

For each $k\geq 2$, express X as,

$$X = \frac{G_1}{\|G_1\|}\mathcal{Z}_1^{\mathrm{T}} + \ldots + \frac{G_{k-1}}{\|G_{k-1}\|}\mathcal{Z}_{k-1}^{\mathrm{T}} + \xi_k V_k,$$

where $\xi_k = [\xi_{k,k} : \ldots : \xi_{k,n}]$ is an $n\times(n-k+1)$ orthonormal matrix, with the vectors $\xi_{k,i}$, for $i=k,\ldots,n$, being orthogonal to $G_1,\ldots,G_{k-1}$. There is flexibility in the choice of the $\xi_{k,i}$'s – the only requirement being that they depend on only $G_1,\ldots,G_{k-1}$ and no other random quantities. For convenience, we take these $\xi_{k,i}$'s to come from the Grahm-Schmidt orthogonalization of $G_1,\ldots,G_{k-1}$ and the columns of the identity matrix.

The matrix V_k, which is $(n-k+1)\times N$ dimensional, is also denoted as,

$$V_k = [V_{k,1} : V_{k,2} : \ldots : V_{k,N}].$$

The columns $V_{k,j}$, where $j=1,\ldots,N+1$ gives the coefficients of the expansion of the column X_j in the basis $\xi_{k,k},\xi_{k,k+1},\ldots,\xi_{k,n}$. We also denote the entries of V_k as $V_{k,i,j}$, where $i=k,\ldots,n$ and $j=1,\ldots,N$.

Let's prove that conditional on $\mathcal{F}_{k-1}$, the distribution of the $V_{k,i,j}$ is independent across i from k to n, and for each such i the joint distribution of $(V_{k,i,j} : j\in J_{k-1})$ is Normal $N(0,\Sigma_{k-1})$. The proof is by induction. The stated property is true initially, at $k=2$, from lemma 13. Recall that the rows of the matrix U in lemma 13 are i.i.d. $N(0,\Sigma_1)$. Correspondingly, since for each $i=2,\ldots,n$, the element $V_{2,i,j}$ is simply the projection of the j th column of U in the direction $\xi_{2,i}$, where $j\in J$,

it follows that the orthonormality of $\xi_{2,i}$'s that the stated property holds for $k = 2$.

Presuming the stated conditional distribution property to be true at k, we conduct analysis, from which its validity will be demonstrated at $k+1$. 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 we first obtain them by explicit recursions and then verify the stated form.

Denote as

$$G_{k,i}^{coef} = -\sum_{j \in dec_{k-1}} \sqrt{P_j}\, V_{k,i,j} \quad \text{for } i = k, \ldots, n\ .$$

Also denote as,

$$G_k^{coef} = (G_{k,k}^{coef},\, G_{k,k+1}^{coef}, \ldots, G_{k,n}^{coef})^{\mathrm{T}}.$$

The vector G_k^{coef} gives the representation of G_k in the basis consisting of columns vectors of ξ_k. In other words, $G_k = \xi_k G_k^{coef}$.

Notice that,

$$\mathcal{Z}_{k,j} = V_{k,j}^{\mathrm{T}} G_k^{coef} / \|G_k^{coef}\|.$$

The representation

$$V_{k,j} = b_{k-1,j}\, G_k^{coef}/\sigma_k + U_{k,j}$$

is used with values of $b_{k-1,j}$ following an update rule that will be specified (depending on $\mathcal{F}_{k-1}$).Denote as $U_k = [U_{k,1} : U_{k,2} : \ldots : U_{k,N}]$, which is an $(n-k+1) \times N$ dimensional matrix like V_k. The entries of U_k are denoted as $U_{k,i,j}$, where $i = k, \ldots, n$ and $j = 1, \ldots, N$.

For the conditional distribution of $G_{k,i}^{coef}$ given $\mathcal{F}_{k-1}$, independence across i, conditional normality and conditional mean 0 are properties inherited from the corresponding properties of the $V_{k,i,j}$. To obtain the conditional variance of $G_{k,i}^{coef} = -\sum_{j \in dec_{k-1}} \sqrt{P_j}\ V_{k,i,j}$, use the conditional covariance $\Sigma_{k-1} = I - \delta_{k-1}\delta_{k-1}^T$ of $V_{k,i,j}$ 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 each $G_{k,i}^{coef}$, for $i = k, \ldots, n$ is

$$\sigma_k^2 \;=\; \left[\, \hat{q}_{k-1} + \hat{f}_{k-1} - \hat{q}_{k-1}^2\, \nu_{k-1} \,\right] P.$$

Conditional on $\mathcal{F}_{k-1}$, the distribution of $\|G_k^{coef}\|^2 = \sum_{i=k}^n (G_{k,i}^{coef})^2$ 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 we compute $b_{k-1,j}$, which is the value of

$$\mathrm{E}[V_{k,i,j} G_{k,i}^{coef} | \mathcal{F}_{k-1}]/\sigma_k$$

for any of the coordinates $i = k, \ldots, n$. Consider the product $V_{k,i,j}\, G_{k,i}^{coef}$ in the numerator. Use the representation of $G_{k,i}^{coef}$ as a sum of the $-\sqrt{P_{j'}}\, V_{k,i,j'}$ for $j' \in dec_{k-1}$. Accordingly, the numerator is $-\sum_{j' \in dec_{k-1}} \sqrt{P_{j'}} \left[1_{j'=j} - \delta_{k-1,j}\delta_{k-1,j'}\right]$, which simplifies to $-\sqrt{P_j} \left[1_{j \in dec_{k-1}} - \nu_{k-1}\hat{q}_{k-1} 1_{j\ sent}\right]$. So for j in $J_k = J_{k-1} - dec_{k-1}$, we have the simplification

$$b_{k-1,j} = \frac{\hat{q}_{k-1}\, \nu_{k-1} \beta_j}{\sigma_k},$$

for which the product for j, j' in J_k takes the form

$$b_{k-1,j} b_{k-1,j'} = \delta_{k-1,j}\delta_{k-1,j'} \frac{\hat{q}_{k-1}\nu_{k-1}}{1 + \hat{f}_{k-1}/\hat{q}_{k-1} - \hat{q}_{k-1}\nu_{k-1}}.$$

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 $U_{k,i,j} = V_{k,i,j} - b_{k-1,j}\, G_{k,i}^{coef}/\sigma_k$ 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 $\mathrm{E}[U_{k,i,j}U_{k,i,j'}|\mathcal{F}_{k-1}]$ by the choice of $b_{k-1,j}$ reduces to $\mathrm{E}[V_{k,i,j}V_{k,i,j'}|\mathcal{F}_{k-1}] - b_{k-1,j}b_{k-1,j'}$ which, for $j \in J_k$, is

$$1_{j=j'} - \delta_{k-1,j}\delta_{k-1,j'} - b_{k-1,j}b_{k-1,j'}.$$

That is, for each i, the $(U_{k,i,j} : j \in J_k)$ 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

$$\nu_k \;=\; \frac{\nu_{k-1}}{1 \;-\; \hat{q}^{adj}_{k-1}\,\nu_{k-1}}.$$

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^{coef}_k, given $\mathcal{F}_{k-1}$. After all, the

$$Z_{k,j} = U^{\mathrm{T}}_{k,j} G^{coef}_k \,/\|G^{coef}_k\|$$

have this $N_{J_k}(0, \Sigma_k)$ distribution, conditional on G^{coef}_k and $\mathcal{F}_{k-1}$, but since this distribution does not depend on G^{coef}_k we have the stated conditional independence.

Now $\mathcal{Z}_{k,j} = X_j^T G_k \,/\|G_k\|$ reduces to $V^T_{k,j} G^{coef}_k \,/\|G^{coef}_k\|$ by the orthogonality of the $G_1, \ldots, G_{k-1}$ and ξ_k. So using the representation $V_{k,j} = b_{k-1,j}\, G^{coef}_k/\sigma_k + U_{k,j}$ one obtains

$$\mathcal{Z}_{k,j} \;=\; b_{k-1,j}\,\|G^{coef}_k\|/\sigma_k \;+\; Z_{k,j}.$$

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 $\mathcal{X}^2_{n-k+1} = \|G^{coef}_k\|^2/\sigma^2_k$. Using the form of $b_{k-1,j}$, for j in J_k, the location shift $b_{k-1,j}\,\mathcal{X}_{n-k+1}$ 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 $n\, b^2_{k,j}/C_{j,R,B}$. 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 $b^2_{k-1,j} = P_j \hat{q}^{adj}_{k-1}\,\nu^2_{k-1}/[1 - \hat{q}^{adj}_{k-1}\nu_{k-1}]$. So this $\hat{w}_k$ may be expressed as

$$\hat{w}_k \;=\; \frac{\nu_{k-1}}{\nu}\;\frac{\hat{q}^{adj}_{k-1}\,\nu_{k-1}}{1 - \hat{q}^{adj}_{k-1}\nu_{k-1}},$$

which, using the update rule for ν_{k-1}, is seen to equal

$$\hat{w}_k = \frac{\nu_{k-1} - \nu_k}{\nu}.$$

We need to prove that conditional on $\mathcal{F}_k$ that the rows of V_{k+1}, for $j \in J_k$, are i.i.d. $N_{J_k}(0, \Sigma_k)$. Recall that $V_{k+1} = \xi_{k+1}^{\mathrm{T}} X$. Since the column span of ξ_{k+1} is contained in that of ξ_k, one may also write V_{k+1} as $\xi_{k+1}^{\mathrm{T}} \xi_k V_k$. Similar to the representation $G_k = \xi_k G_k^{coef}$, express the columns of ξ_{k+1} in terms of the columns of ξ_k as $\xi_{k+1} = \xi_k \xi_k^{coef}$. where ξ_k^{coeff} is an $(n-k+1) \times (n-k)$ dimensional matrix. Using this representation one gets that $V_{k+1} = (\xi_k^{coef})^{\mathrm{T}} V_k$.

Notice that ξ_k is $\mathcal{F}_{k-1}$ measurable and that ξ_{k+1} is $\sigma\{\mathcal{F}_{k-1}, G_k\}$ measurable. Correspondingly, ξ_k^{coef} is also $\sigma\{\mathcal{F}_{k-1}, G_k\}$ measurable. Further, because of the orthonormality of ξ_k and ξ_{k+1}, one gets that ξ_k^{coef} is also orthonormal. Further, as G_k is orthonormal to ξ_{k+1}, one has the G_k^{coef} is orthogonal to the columns of ξ_k^{coef} as well.

Accordingly, one has that $V_{k+1} = (\xi_k^{coef})^{\mathrm{T}} U_k$. Consequently, using the independence of U_k and G_k^{coef}, and the above, one gets that conditional on $\sigma\{\mathcal{F}_{k-1}, G_k\}$, for $j \in J_k$, the rows of V_{k+1} are i.i.d. $N_{J_k}(0, \Sigma_k)$.

We need to prove that conditional on $\mathcal{F}_k$, the distribution of V_{k+1} is as above, where recall that $\mathcal{F}_k = \sigma\{\mathcal{F}_{k-1}, G_k, \mathcal{Z}_k\}$ or equivalently, $\sigma\{\mathcal{F}_{k-1}, G_k, Z_k\}$. This claim follows from the conclusion of the previous paragraph by noting that V_{k+1} is independent of $Z_k = (G_k^{coef})^{\mathrm{T}} U_k$, conditional on $\sigma\{\mathcal{F}_{k-1}, G_k\}$ as G_k^{coef} is orthogonal to ξ_k^{coeff}.

Finally, repeatedly apply $\nu_{k'}/\nu_{k'-1} = 1/(1 - \hat{q}_{k'-1}^{adj} \nu_{k'-1})$, for k' from k to 2, each time substituting the required expression on the right and simplifying to obtain

$$\frac{\nu_k}{\nu_{k-1}} \;=\; \frac{1 \;-\; (\hat{q}_1^{adj} + \ldots + \hat{q}_{k-2}^{adj})\; \nu}{1 - (\hat{q}_1^{adj} + \ldots + \hat{q}_{k-2}^{adj} + \hat{q}_{k-1}^{adj})\, \nu}.$$

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 14 is complete.

4.A 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

$$D_\alpha(\mathbb{P} \| \mathbb{Q}) = \frac{1}{\alpha - 1} \log \mathrm{E}_{\mathbb{Q}}[(p(Z)/q(Z))^\alpha].$$

Its limit for large α is $D_\infty(\mathbb{P} \| \mathbb{Q}) = \log \|p/q\|_\infty$.

Lemma 24. *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. For convex f, as in Csiszar's f-divergence inequality, from Jensen's inequality applied to the decomposition of $\mathrm{E}_{\mathbb{Q}}[f(p(Z)/q(Z))]$ using the distributions conditional on A and its complement,

$$\mathbb{Q}A\, f(\mathbb{P}A/\mathbb{Q}A) \;+\; \mathbb{Q}A^c\, f(\mathbb{P}A^c/\mathbb{Q}A^c) \;\le\; \mathrm{E}_{\mathbb{Q}} f(p(Z)/q(Z)).$$

Using in particular $f(r) = r^\alpha$ and throwing out the non-negative A^c part, yields

$$(\mathbb{P}A)^\alpha \;\le\; (\mathbb{Q}A)^{\alpha-1} \mathrm{E}_{\mathbb{Q}}[(p(Z)/q(Z))^\alpha].$$

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 25. *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. 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 25. □

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 15. 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, $\mathbb{P}[A] = \mathrm{E}\,[\mathbb{P}[A|\mathcal{F}_{k-1}]]$. 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

$$\mathbb{P}[A] = \mathrm{E}_{\mathbb{P}}\left[\mathbb{P}_{\mathcal{X}^2_{n-k+1}, Z_{k,J_k}|\mathcal{F}_{k-1}}[A]\right],$$

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 14 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 25, 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 15. □

4.B Proof of Lemma 16

The claim regarding the increase can be checked by induction as follows. For $k=1$, the $q_{1,1} = g(0)-\eta$ is at least $gap-\eta$ above $q_{1,k-1}$ taken to be 0, initially. Subsequently, suppose we are at step $k>1$ with $q^{adj}_{1,k-1} \leq x_r$ and suppose the claim holds for $k-1$. We have $f_{1,k-1} \leq (k-1)f$ and $q_{1,k-1} \geq (k-1)\Lambda$, so the ratio satisfies $f_{1,k-1}/q_{1,k-1} \leq f/\Lambda$. Consequently, $q^{adj}_{1,k-1}$ is at least $q_{1,k-1}/(1+f/\Lambda)$, which may be expressed as $q_{1,k-1} - (f/\Lambda)q_{1,k-1}/(1+f/\Lambda)$. Then consider $q_{1,k} = g(q^{adj}_{1,k-1}) - \eta$. Since the argument of g is less than x_r, it follows that this is at least $q^{adj}_{1,k-1} + gap - \eta$, which in turn is at least $q_{1,k-1} - (f/\Lambda)x_r/(1+f/\Lambda) + gap - \eta$, which is at least $q_{1,k-1} + \Lambda$.

The increase is indeed at least Λ each step, and the number of steps required for $q^{adj}_{1,m-1}$ to first exceed x_r is not more than $m-1 = x_r/\Lambda \leq (1-\Lambda)/\Lambda = 1/\Lambda - 1$. At that point $q_{1,m} = g(q^{adj}_{1,m-1}) - \eta$ exceeds $g(x_r) - \eta$. This completes the proof of Lemma 16.

4.C Distribution of $\mathcal{Z}_k$

Here we characterize the distribution of $\mathcal{Z}_{k,j}$, for all $j \in J$, and in the process provide an alternative proof lemma 14. The proof of lemma 27, which has parallels with recent work by Bayati and

Montanari [8], has been given in Barron and Cho [3]. For completeness, we provide here a slight variation of the proof. Lemma 28 provides an alternative proof of lemma 14 using the claims of lemma 27.

Since it is just a matter of a change of scale, in this section we assume for convenience that the noise variance $\sigma^2 = 1$. Let $\tilde{\beta}_0 = (\beta^{\mathrm{T}} : 1)^{\mathrm{T}}$ and for $k \geq 1$, define, $\tilde{\beta}_k = (\tilde{\beta}_{k,1}, \ldots, \tilde{\beta}_{k,N+1})$, as

$$\tilde{\beta}_{k,j} = \begin{cases} -\sqrt{P_j} & \text{if } dec_k \cap J \\ 0 & \text{if } j = N+1 \end{cases}$$

Notice that $G_1 = Y$ is equal to $\tilde{X}\tilde{\beta}_0$, where $\tilde{X} = [X : \epsilon]$ is the matrix formed by adding the column comprising the noise vector ϵ to the X matrix. For $k \geq 1$, the vector $\tilde{\beta}_k$ is non-zero at positions corresponding terms decoded in step k. The following simple lemma will come in handy.

Lemma 26. *For each $k \geq 1$, the following holds.*

(i) $\tilde{\beta}_0^{\mathrm{T}} \tilde{\beta}_k = -P\hat{q}_k$

(ii) $\|\tilde{\beta}_k\|^2 = P(\hat{q}_k + \hat{f}_k)$

Proof. We first prove part (i). From the definition of $\tilde{\beta}_k$, one has $\tilde{\beta}_{k,N+1}$ is 0. Correspondingly, one has,

$$\tilde{\beta}_0^{\mathrm{T}} \tilde{\beta}_k = -\sum_{j \in J \cap dec_k} \beta_j \sqrt{P_j}.$$

Now, $\beta_j \sqrt{P_j}$ is P_j for $j \in sent \cap dec_k$, and is equal to 0 otherwise. Correspondingly, the above sum is equal to $-\sum_{j \in sent \cap dec_k} P_j$, which is $-P\hat{q}_k$.

For part (ii), notice that,

$$\|\tilde{\beta}_k\|^2 = \sum_{j \in others \cap dec_k} P_j + \sum_{j \in sent \cap dec_k} P_j,$$

which is equal $P(\hat{q}_k + \hat{f}_k)$. This completes the proof. □

Further, let b_0, b_1, …, b_k be orthonormal vectors obtained by successive Grahm-Schmidt orthonormalization of the vectors $\tilde{\beta}_0$, $\tilde{\beta}_1$, …, $\tilde{\beta}_k$. Notice that $\tilde{\beta}_1$, …, $\tilde{\beta}_{k-1}$ are $\mathcal{F}_{k-1}$ measurable as they are completely specified from knowledge of the decoded sets $dec_1, \ldots, dec_{k-1}$, which are $\mathcal{F}_{k-1}$ measurable. Consequently, the same is true for the vectors b_0, b_1, …, b_{k-1}.

Let $\tilde{W}_k = I - b_0 b_0^{\mathrm{T}} - \ldots - b_{k-1} b_{k-1}^{\mathrm{T}}$ be the $\mathbb{R}^{(N+1)\times(N+1)}$ projection matrix for the space orthogonal to that spanned by $\tilde{\beta}_0, \ldots, \tilde{\beta}_{k-1}$. The sub-matrix $W_k = (\tilde{W}_k)_J$, formed by taking the first N rows and columns of $\tilde{W}_k$ gives the covariance matrix of the noise term in the conditional distribution of $\mathcal{Z}_k$ given $\mathcal{F}_{k-1}$. We formally state this in the following lemma. For convenience take $\mathcal{F}_0$ to be the empty σ-field and W_0 to be the $N \times N$ identity matrix.

Lemma 27. *For $k \geq 1$, the conditional distribution $\mathbb{P}_{\mathcal{Z}_k|\mathcal{F}_{k-1}}$ of $\mathcal{Z}_k$ given $\mathcal{F}_{k-1}$ can be representation as,*

$$\mathcal{Z}_k = b_{k-1} \frac{\|G_k\|}{\sigma_k} + Z_k,$$

where Z_k is distributed as $N(0, W_k)$ conditional on $\mathcal{F}_{k-1}$, where W_k is as given above.

Further, $\sigma_k^2 = \tilde{\beta}_{k-1} \tilde{W}_{k-1} \tilde{\beta}_{k-1}$, for $k \geq 1$. Moreover, conditioned on $\mathcal{F}_{k-1}$, the quantity $\|G_k\|/\sigma_k$ is independent of Z_k and has a Chi distribution with $n - k + 1$ degrees of freedom.

Proof. The notation for some of the quantities used here will be different from that used in the proof of lemma 14 given in section 4.19. Let

$$\tilde{X} = [X : \epsilon]$$

be the $n \times (N+1)$ matrix formed by adding the column comprising of the noise vector ϵ to the dictionary matrix X. Further, for each $k \geq 1$, let

$$\tilde{\mathcal{Z}}_k = \tilde{X}^{\mathrm{T}} \frac{G_k}{\|G_k\|}. \tag{4.40}$$

Notice that $\mathcal{Z}_k = [\tilde{\mathcal{Z}}_k]_J$. In other words, $\mathcal{Z}_k$ corresponds to the first N entries of the the length $N+1$ vector $\tilde{\mathcal{Z}}_k$.

Also, denote $\tilde{\mathcal{F}}_{k-1} = \sigma\{G_1, \ldots, G_{k-1}, \tilde{\mathcal{Z}}_1, \ldots, \mathcal{Z}_{k-1}\}$. Notice that $\tilde{\mathcal{F}}_{k-1}$ is a larger σ-field that $\mathcal{F}_{k-1}$ since it comprises of $\tilde{\mathcal{Z}}_{k'}$, for $k' = 1, \ldots, k-1$, instead of $\mathcal{Z}_{k'}$'s in $\mathcal{F}_{k-1}$.

We prove prove lemma 27 by showing that the conditional distribution of $\tilde{\mathcal{Z}}_k$ given $\tilde{\mathcal{F}}_{k-1}$ may be represented as,

$$\tilde{\mathcal{Z}}_k = b_{k-1} \frac{\|G_k\|}{\sigma_k} + \tilde{Z}_k, \tag{4.41}$$

where $\tilde{Z}_k$ is distribution as $N(0, \tilde{W}_k)$, conditional on $\tilde{\mathcal{F}}_{k-1}$. Also, conditional on $\tilde{\mathcal{F}}_{k-1}$, the quantity $\|G_k\|/\sigma_k$ follows $\mathcal{X}_{n-k+1}$ and is independent of Z_k.

Lemma 27 immediately follows from the above as the quantities b_{k-1}, σ_k, as well as the covariance matrix $\tilde{W}_k$ as all $\mathcal{F}_{k-1}$ measurable. This follows since these quantities are functions of $\tilde{\beta}_0, \ldots, \tilde{\beta}_{k-1}$, which are $\mathcal{F}_{k-1}$ measurable.

For each $k \geq 1$, express $\tilde{X}$ as

$$\tilde{X} = \frac{G_1}{\|G_1\|}\tilde{\mathcal{Z}}_1^{\mathrm{T}} + \ldots + \frac{G_{k-1}}{\|G_{k-1}\|}\tilde{\mathcal{Z}}_{k-1}^{\mathrm{T}} + \xi_k V_k,$$

where $\xi_k = [\xi_{k,k} : \ldots : \xi_{k,n}]$ is an $n \times (n-k+1)$ orthonormal matrix, with the vectors $\xi_{k,i}$, for $i = k, \ldots, n$, being orthogonal to $G_1, \ldots, G_{k-1}$. This ξ_k is chosen exactly similarly as in the proof of lemma 14 in section 4.19.

The j th column, where $j = 1, \ldots, N+1$, of the $(n-k+1) \times (N+1)$ dimensional matrix V_k gives coefficient of the expansion of the column $\tilde{X}_j$ in the basis $\xi_{k,k}, \xi_{k,k+1}, \ldots, \xi_{k,n}$. We first prove inductively that for $k' \geq 1$, conditional on $\mathcal{F}_{k'-1}$, the rows of $V_{k'}$ are i.i.d. $N(0, \tilde{W}_{k'-1})$.

Clearly the hypothesis is true for $k' = 1$, as $V_1 = \tilde{X}$. Let's assume that hypothesis holds for a $k' = k$, for some $k \geq 1$. We need to show that it holds for $k' = k+1$ as well.

Notice that $G_k = \xi_k V_k \tilde{\beta}_{k-1}$. Denote as,

$$G_k^{coef} = (G_{k,k}^{coef}, \ldots, G_{k,n}^{coef})^{\mathrm{T}} = V_k \tilde{\beta}_{k-1}.$$

In others words, the vector G_k^{coef} gives the coefficients in the expansion of G_k using the columns of ξ_k. As a consequence of the induction hypothesis, conditional on $\tilde{\mathcal{F}}_{k-1}$, the $G_{k,i}^{coef}$, for $i = k, \ldots, n$, are i.i.d. $N(0, \sigma_k^2)$, where $\sigma_k^2 = \tilde{\beta}_{k-1}^{\mathrm{T}} \tilde{W}_{k-1} \tilde{\beta}_{k-1}$. Here we make use of the fact the $\tilde{\beta}_{k-1}$ is $\tilde{\mathcal{F}}_{k-1}$ measurable.

Notice that V_k and G^{coef} are jointly normal conditional on $\tilde{\mathcal{F}}_{k-1}$. Denote the entries of V_k as $V_{k,i,j}$, where $i = k, \ldots, n$ and $j = 1, \ldots, N+1$. Correspondingly, similar to lemma 13, through conditioning on G_k^{coef} (and $\tilde{\mathcal{F}}_{k-1}$), may be expressed as,

$$V_{k,i,j} = b_{k-1,j} \frac{G_{k,i}^{coef}}{\sigma_k} + U_{k,i,j}, \tag{4.42}$$

where $b_{k-1,\,j} = \mathrm{E}(V_{k,i,j}\, G_{k,i}^{coef}/\sigma_k)$, where the expectation is assumed to be conditional on $\tilde{\mathcal{F}}_{k-1}$. Denoting $b_{k-1} = (b_{k-1,j} : j = 1, \ldots, N+1)$, one sees that $b_k = \tilde{W}_{k-1}\tilde{\beta}_{k-1}/\sigma_k$, which has norm 1, since $\sigma_k^2 = \|\tilde{W}_{k-1}\tilde{\beta}_{k-1}\|^2$, as $\tilde{W}_{k-1}$ is idempotent being a projection matrix. Correspondingly, this

definition of b_{k-1} is consistent with that given previously.

Further, denoting U_k as the $(n-k+1)$ dimensional matrix with entries $U_{k,i,j}$, one has that U_k is independent of G_k^{coef}, conditional on $\tilde{\mathcal{F}}_{k-1}$, and further, the rows of U_k are i.i.d. $N(0, \tilde{W}_k)$, where

$$\tilde{W}_k = \tilde{W}_{k-1} - b_{k-1} b_{k-1}^{\mathrm{T}}.$$

This definition of $\tilde{W}_k$ is also consistent with that given previously.

Assuming that the induction hypothesis is true, (4.42) allows us to prove our claim regarding the condition distribution of $\tilde{\mathcal{Z}}_k$ given $\tilde{\mathcal{F}}_{k-1}$ given in (4.41). To see this, recall that $\tilde{\mathcal{Z}}_k = \tilde{X}^{\mathrm{T}} G_k / \|G_k\|$, which, using the expansion in the basis represented by columns of ξ_k, is also equal to $V_k^{\mathrm{T}} G^{coef} / \|G^{coef}\|$. Accordingly, using (4.42) one gets that

$$\tilde{\mathcal{Z}}_k = b_{k-1} \frac{\|G^{coef}\|}{\sigma_k} + \tilde{Z}_k$$

where $\tilde{Z}_k = U_k^{\mathrm{T}} G^{coef} / \|G^{coef}\|$. Since, conditional on $\tilde{\mathcal{F}}_{k-1}$, the entries of G_k^{coef} are i.i.d. $N(0, \sigma_k^2)$, one has that $\|G^{coef}\|^2 / \sigma_k^2$ follows $\mathcal{X}^2_{n-k+1}$. Further, as U_k is independent of G_k^{coef}, using the same reasoning as in the proof of lemma 13, one has that $\tilde{Z}_k$ follows $N(0, \tilde{W}_k)$ and is independent of G_k^{coef} (conditional on $\tilde{\mathcal{F}}_{k-1}$).

What remains to be proven is the conditional distribution of V_{k+1} given $\tilde{\mathcal{F}}_k$. The proof is similar to that given in section 4.19. We write it verbatim, only changing notations whenever necessary.

Recall that $V_{k+1} = \xi_{k+1}^{\mathrm{T}} \tilde{X}$. Since the column span of ξ_{k+1} is contained in that of ξ_k, one may also write V_{k+1} as $\xi_{k+1}^{\mathrm{T}} \xi_k V_k$. Similar to the representation $G_k = \xi_k G_k^{coef}$, express the columns of ξ_{k+1} in terms of the columns of ξ_k as $\xi_{k+1} = \xi_k \xi_k^{coef}$. where ξ_k^{coeff} is an $(n-k+1) \times (n-k)$ dimensional matrix. Using this representation one gets that $V_{k+1} = (\xi_k^{coef})^{\mathrm{T}} V_k$.

Notice that ξ_k is $\tilde{\mathcal{F}}_{k-1}$ measurable and that ξ_{k+1} is $\sigma\{\tilde{\mathcal{F}}_{k-1}, G_k\}$ measurable. Correspondingly, ξ_k^{coef} is also $\sigma\{\tilde{\mathcal{F}}_{k-1}, G_k\}$ measurable. Further, because of the orthonormality of ξ_k and ξ_{k+1}, one gets that ξ_k^{coef} is also orthonormal. Further, as G_k is orthonormal to ξ_{k+1}, one has the G_k^{coef} is orthogonal to the columns of ξ_k^{coef} as well.

Accordingly, using (4.42) and the fact that $(\xi_k^{coef})^{\mathrm{T}} G_k^{coef} = 0$, one has that $V_{k+1} = (\xi_k^{coef})^{\mathrm{T}} U_k$. Consequently, using the independence of U_k and G_k^{coef}, and the above, one gets that conditional on $\sigma\{\tilde{\mathcal{F}}_{k-1}, G_k\}$, the rows of V_{k+1} are i.i.d. $N(0, \tilde{W}_k)$.

We need to prove that conditional on $\tilde{\mathcal{F}}_k$, the distribution of V_{k+1} is as above, where recall that

$\tilde{\mathcal{F}}_k = \sigma\{\tilde{\mathcal{F}}_{k-1}, G_k, \tilde{\mathcal{Z}}_k\}$ or equivalently, $\sigma\{\tilde{\mathcal{F}}_{k-1}, G_k, \tilde{Z}_k\}$. This claim follows from the conclusion of the previous paragraph by noting that V_{k+1} is independent of $\tilde{Z}_k = (G_k^{coef})^{\mathrm{T}} U_k$, conditional on $\sigma\{\tilde{\mathcal{F}}_{k-1}, G_k\}$ as G_k^{coef} is orthogonal to ξ_k^{coeff}. This completes the proof of the lemma. □

Lemma 14 is an immediate consequence of the above lemma. Notice that lemma 14 gives explicit expressions for the conditional distribution of $\mathcal{Z}_{k,J_k} = (\mathcal{Z}_{k,j} : j \in J_k)$, given $\mathcal{F}_{k-1}$. From lemma 27 one gets that conditional on $\mathcal{F}_{k-1}$, one has

$$\mathcal{Z}_{k,j} = \sqrt{n}\, b_{k-1,j}(\mathcal{X}_{d_k}/\sqrt{n}) + Z_{k,j} \qquad \text{for } j \in J_k.$$

Here $\mathbb{P}_{Z_{k,\,J_k}|\mathcal{F}_{k-1}}$ is $N(0, \Sigma_k)$, where $\Sigma_k = (\tilde{W}_k)_{J_k}$. Accordingly, lemma 14 is proven one we evaluate the quantities $b_{k-1,j}$, for $j \in J_k$ and $(\tilde{W}_k)_{J_k}$. This is done in the next lemma.

Lemma 28. *The covariance $\Sigma_k = (\tilde{W}_k)_{J_k}$ has the simplified representation as given in lemma 14. Further,*

$$b_{k-1,\,j} = \frac{\sqrt{\hat{w}_k C_{j,R,B}}}{\sqrt{n}} 1_{\{j \in sent\}} \qquad \textit{for } j \in J_k.$$

Proof. We first prove the expression for $\Sigma_k = \left(\tilde{W}_k\right)_{J_k}$. Recall that $\tilde{W}_k = I - P_{B_k}$, where P_{B_k} is the projection matrix in the space spanned by $B_k = [\tilde{\beta}_0 : \tilde{\beta}_1 : \ldots : \tilde{\beta}_{k-1}]$. We denote B_k as B whenever there is no ambiguity. We compute P_B by exhibiting an orthonormal basis V for B. Consequently, $\tilde{W}_k$ becomes $I - VV^{\mathrm{T}}$.

Notice that in $B = [\tilde{\beta}_0 : \tilde{\beta}_1 : \ldots : \tilde{\beta}_{k-1}]$, the vectors $\tilde{\beta}_1, \tilde{\beta}_2, \ldots, \tilde{\beta}_{k-1}$ are orthonormal, since the location of non-zeroes in these vectors occur at disjoint locations. Correspondingly, we are only left with the task making the vector $\tilde{\beta}_0 = (\beta^{\mathrm{T}}, 1)^{\mathrm{T}}$ orthogonal to $\tilde{\beta}_1, \tilde{\beta}_2, \ldots, \tilde{\beta}_{k-1}$.

We see that $b_{k-1}^{\|}$, the projection of $\tilde{\beta}_0$ onto space spanned by $\tilde{\beta}_1, \tilde{\beta}_2, \ldots, \tilde{\beta}_{k-1}$ is given by

$$b_{k-1}^{\|} = -\sum_{k'=1}^{k-1} \frac{\hat{q}_{k'}}{\hat{q}_{k'} + \hat{f}_{k'}} \tilde{\beta}_{k'}.$$

The above follows from using that $t b_0^{\mathrm{T}} \tilde{\beta}_{k'} = -\hat{q}_{k'}$ and $\|\tilde{\beta}_{k'}\|^2 = \hat{q}_{k'} + \hat{f}_{k'}$, as proved in lemma 26. The above may also be expressed as,

$$b_{k-1,j}^{\|} = \begin{cases} \frac{\hat{q}_{k'}}{\hat{q}_{k'} + \hat{f}_{k'}} \sqrt{P_j} & \text{if } j \in dec_{k'} \text{ and } 1 \le k' \le k-1 \\ 0 & \text{otherwise} \end{cases}$$

From this we get that $b^{\perp}_{k-1} = (b^{\perp}_{k-1,1}, \ldots, b^{\perp}_{k-1,N+1})^{\mathrm{T}}$, the projection of $\tilde{\beta}_0$ orthogonal to $\tilde{\beta}_1, \tilde{\beta}_2, \ldots, \tilde{\beta}_{k-1}$, which is also equal to $\tilde{\beta}_0 - b^{\|}_{k-1}$, is

$$b^{\perp}_{k-1,j} = \begin{cases} \beta_j & \text{if } j \in J_k \\ \frac{\hat{f}_{k'}}{\hat{q}_{k'}+\hat{f}_{k'}}\sqrt{P_j} & \text{if } j \in dec_{k'} \cap sent \text{ and } 1 \le k' \le k-1 \\ -\frac{\hat{q}_{k'}}{\hat{q}_{k'}+\hat{f}_{k'}}\sqrt{P_j} & \text{if } j \in dec_{k'} \cap others \text{ and } 1 \le k' \le k-1 \\ 1 & \text{if } j = N+1 \end{cases} \tag{4.43}$$

We now find the norm of $b^{\perp}_{k-1,j}$. Notice that,

$$\begin{aligned} \sum_{j \in J_k} \beta_j^2 &= P - \sum_{j \in sent \cap dec_{1,k-1}} \beta_j^2 \\ &= P(1 - \sum_{k'=1}^{k-1} \hat{q}_{k'}). \end{aligned}$$

The first expression follows from using that $J_k = J - dec_{1,k-1}$ and the fact that β_j is non-zero only if $j \in sent$. From this and using (4.43), it is seen that

$$\begin{aligned} \|b^{\perp}_{k-1}\|^2 &= 1 + P(1 - \sum_{k'=1}^{k-1} \hat{q}_{k'}) + \sum_{k'=1}^{k-1} \frac{\hat{f}^2_{k'}}{(\hat{q}_{k'}+\hat{f}_{k'})^2} \sum_{j \in dec_{k'} \cap sent} P_j + \sum_{k'=1}^{k-1} \frac{\hat{q}^2_{k'}}{(\hat{q}_{k'}+\hat{f}_{k'})^2} \sum_{j \in dec_{k'} \cap others} P_j \\ &= 1 + P(1 - \sum_{k'=1}^{k-1} \hat{q}_{k'}) + P \sum_{k'=1}^{k-1} \frac{\hat{f}^2_{k'}}{(\hat{q}_{k'}+\hat{f}_{k'})^2} \hat{q}_{k'} + P \sum_{k'=1}^{k-1} \frac{\hat{q}^2_{k'}}{(\hat{q}_{k'}+\hat{f}_{k'})^2} \hat{f}_{k'} \\ &= 1 + P(1 - \sum_{k'=1}^{k-1} \hat{q}_{k'}) + P \sum_{k'=1}^{k-1} \frac{\hat{f}_{k'}\hat{q}_{k'}}{\hat{q}_{k'}+\hat{f}_{k'}} \\ &= 1 + P(1 - \hat{q}^{adj,tot}_{k-1}), \end{aligned}$$

where $\hat{q}^{adj,tot}_k$ is as in (4.20). The last expression may also be written as, $\|b^{\perp}_{k-1}\|^2 = (1+P)\left(1 - \nu\hat{q}^{adj,tot}_{k-1}\right)$. Thus,

$$\|b^{\perp}_{k-1}\|^2 = \frac{1}{(1-\nu)\hat{s}_k},$$

where $\hat{s}_k$ is as in lemma 14.

The above calculations lead us to the orthonormal matrix $V = [b^{\perp}_{k-1}/\|b^{\perp}_{k-1}\| : \tilde{\beta}_1/\|\tilde{\beta}_1\| : \ldots : \tilde{\beta}_{k-1}/\|\tilde{\beta}_{k-1}\|]$, with the same column space as B. Recall that $\|\tilde{\beta}_k\|^2 = P(\hat{q}_k + \hat{f}_k)$ for $k \ge 1$, from

lemma 26. Consequently, the projection matrix P_B can be expressed as

$$P_B = \frac{b_{k-1}^{\perp}}{\|b_{k-1}^{\perp}\|}\frac{(b_{k-1}^{\perp})^{\mathrm{T}}}{\|b_{k-1}^{\perp}\|} + \sum_{k'=1}^{k-1}\frac{\tilde{\beta}_{k'}\tilde{\beta}_{k'}^{\mathrm{T}}}{P(\hat{q}_{k'}+\hat{f}_{k'})}.$$

From the above expressions for the vectors in V, one sees that $j \in others \cap J_k$ and $j' \in J$, $(P_B)_{j,j'} = (P_B)_{j',j} = 0$. This follows since for $j \in others$ we have $b_{k-1,j}^{\perp} = 0$ and since the j th row of $[\tilde{\beta}_1 : \ldots : \tilde{\beta}_{k-1}]$ is all 0 for $j \in J_k$ as the index j has not been decoded in previous steps.

Further, for j, $j' \in sent \cap J_k$ we have,

$$(P_B)_{j,j'} = (1-\nu)\hat{s}_k(\beta_j\beta_{j'}).$$

This is seen by noting that $b_{k-1,j}^{\perp} = \beta_j$ for $j \in sent \cap J_k$ and as before the j rows of $\tilde{\beta}_{k'}$, $1 \le k' \le k-1$ are all zero for $j \in J_k$.

Thus Z_{k,J_k} has covariance matrix $I - \Sigma_k$, where

$$\Sigma_k = I - \nu\hat{s}_k(\beta\beta^{\mathrm{T}})_{J_k}/P,$$

using $(1-\nu)P = \nu$. Also, as before, $(\beta\beta^{\mathrm{T}})_{J_k}$ refers to the sub-matrix of $\beta\beta^{\mathrm{T}}$ comprising of rows and column indices from J_k.

Next, we now prove the claim regarding the expression for $b_{k-1,j}$. Notice that,

$$b_{k-1} = \frac{(I - P_{B_{-1}})\tilde{\beta}_{k-1}}{\|(I - P_{B_{-1}})\tilde{\beta}_{k-1}\|},$$

where for convenience we write B_{k-1} as B_{-1}. Notice that the denominator of the above expression is simply σ_k.

Let's evaluate $(I - P_{B_{-1}})\tilde{\beta}_{k-1}$. Using the same reasoning as before, one has that,

$$P_{B_{-1}} = \frac{b_{k-2}^{\perp}}{\|b_{k-2}^{\perp}\|}\frac{(b_{k-2}^{\perp})^{\mathrm{T}}}{\|b_{k-2}^{\perp}\|} + \sum_{k'=1}^{k-2}\frac{\tilde{\beta}_{k'}\tilde{\beta}_{k'}^{\mathrm{T}}}{P(\hat{q}_{k'}+\hat{f}_{k'})}.$$

Consequently,

$$P_{B_{-1}}\tilde{\beta}_{k-1} = \frac{\tilde{\beta}_{k-1}^{\mathrm{T}}b_{k-2}^{\perp}}{\|b_{k-2}^{\perp}\|^2}b_{k-2}^{\perp},$$

The above follows as $\tilde{\beta}_{k'}^{\mathrm{T}}\tilde{\beta}_{k-1} = 0$ for $1 \le k' \le k-2$. Now notice from the form of $b_{k-1}^{\perp}$ given in

(4.43), that $\tilde{\beta}_{k-1}^{\mathrm{T}} b_{k-2}^{\perp} = -P\hat{q}_{k-1}$. The above evaluates to

$$P_{B_{-1}}\tilde{\beta}_{k-1} = -(1-\nu)\hat{s}_{k-1}P\hat{q}_{k-1}b_{k-2}^{\perp}. \tag{4.44}$$

Accordingly, $\sigma_k^2 = \|(I - P_{B_{-1}})\tilde{\beta}_{k-1}\|^2$ is equal to $\|\tilde{\beta}_{k-1}\|^2 - \tilde{\beta}_{k-1}^{\mathrm{T}} P_{B_{-1}}\tilde{\beta}_{k-1}$, which using (4.44) evaluates to

$$P(\hat{q}_{k-1} + \hat{f}_{k-1}) - (1-\nu)\hat{s}_{k-1}(P\hat{q}_{k-1})^2,$$

which simplifies to,

$$P(\hat{q}_{k-1} + \hat{f}_{k-1})\frac{\hat{s}_{k-1}}{\hat{s}_k}$$

For $j \in J_k$, we have $\tilde{\beta}_{k-1,j} = 0$ and $b_{k-2,j}^{\perp} = \beta_j$. Consequently,

$$\begin{aligned} b_{k-1,j} &= (1-\nu)\hat{s}_{k-1}\frac{P\hat{q}_{k-1}}{\sigma_k}\beta_j \\ &= \sqrt{P}(1-\nu)\sqrt{\hat{s}_{k-1}\hat{s}_k}\sqrt{\hat{q}_{k-1}^{adj}}\,\beta_j \\ &= \frac{\sqrt{C_{j,R,B}\,\hat{w}_k}}{\sqrt{n}}\,1_{\{j\in sent\}}, \end{aligned}$$

where we use that $\hat{s}_{k-1}\hat{s}_k(\hat{q}_{k-1}^{adj}\nu) = \hat{w}_k$. This completes the proof of lemma 28. □

4.D Tails for weighted Bernoulli sums

Lemma 29. *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 29. 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 [**?**] we have that

$$P(\mathcal{A}) \;=\; \exp\{-D\big(P_{\underline{W}|\mathcal{A}}\|P_{\underline{W}}\big)\}$$

$$\le\; \exp\big\{-\sum_j D\big(P_{W_j|\mathcal{A}}||P_{W_j}\big)\big\}$$

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

$$\sum_j D\big(P_{W_j|\mathcal{A}}\|P_{W_j}\big) \ge N_\alpha \sum_j \alpha_j D\big(P_{W_j|\mathcal{A}}\|P_{W_j}\big).$$

Furthermore, the convexity of the relative entropy implies that

$$\sum_j \alpha_j D(P_{W_j|\mathcal{A}} \parallel P_{W_j}) \ge D\left(\sum_j \alpha_j P_{W_j|\mathcal{A}} \parallel \sum_j \alpha_j P_{W_j}\right).$$

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

$$r_e = \mathrm{E}\big[\sum_j \alpha_j W_j \,\big|\, \mathcal{A}\big] = \mathrm{E}\big[\hat{r}\,\big|\,\mathcal{A}\big].$$

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 29. □

4.E Lower Bounds on D

Lemma 30. *For $p \geq p^*$, the relative entropy between Bernoulli(p) and Bernoulli(p^*) distributions has the succession of lower bounds*

$$D_{Ber}(p\|p^*) \geq D_{Poi}(p\|p^*) \geq 2\big(\sqrt{p} - \sqrt{p^*}\big)^2 \geq \frac{(p-p^*)^2}{2p}$$

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.

Proof of Lemma 30. 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 $(p-p^*)^2/(2p)$. □

Bibliography

[1] M. Akçakaya and V. Tarokh. Shannon-theoretic limits on noisy compressive sampling. *IEEE Trans. Inform. Theory*, 56(1):492–504, 2010.

[2] A.R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inform. Theory*, 39(3):930–945, 1993.

[3] A.R. Barron and S. Cho. High-rate sparse superposition codes with iteratively optimal estimates. In *Proc. Int. Sypm. Inform. Theory, to appear*, 2012.

[4] A.R. Barron and A. Joseph. Least squares superposition coding of moderate dictionary size, reliable at rates up to channel capacity. *http://arxiv.org/abs/1006.3780*, 2010.

[5] A.R. Barron and A. Joseph. Sparse superposition codes: Fast and reliable at rates approaching capacity with gaussian noise. Technical report, Yale University, 2010.

[6] A.R. Barron and A. Joseph. Least squares superposition coding of moderate dictionary size, reliable at rates up to channel capacity. *IEEE Trans. Inform. Theory*, 58:2541 – 2557, 2012.

[7] A.R. Barron, A. Cohen, W. Dahmen, and R.A. DeVore. Approximation and learning by greedy algorithms. *Ann. Statist.*, 36(1):64–94, 2008.

[8] M. Bayati and A. Montanari. The lasso risk for gaussian matrices. *IEEE Trans. Inform. Theory*, 58(4):1997, 2012.

[9] C. Berrou, A. Glavieux, and P. Thitimajshima. Near shannon limit error-correcting coding and decoding: Turbo-codes. 1. In *IEEE Int. Conf. Commun.*, volume 2, pages 1064–1070. IEEE, 1993.

[10] E.J. Candès and Y. Plan. Near-ideal model selection by l1 minimization. *Ann. Statist.*, 37 (5A):2145–2177, 2009.

[11] E.J. Candes and T. Tao. Decoding by linear programming. *IEEE Trans. Inform. Theory*, 51 (12):4203–4215, 2005.

[12] E.J. Candes and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Trans. Inform. Theory*, 52(12):5406–5425, 2006.

[13] S.S. Chen, D.L. Donoho, and M.A. Saunders. Atomic decomposition by basis pursuit. *SIAM review*, pages 129–159, 2001.

[14] T. Cover. Broadcast channels. *IEEE Trans. Inform. Theory*, 18(1):2–14, 1972.

[15] T.M. Cover, J.A. Thomas, J. Wiley, et al. *Elements of information theory*, volume 6. Wiley Online Library, 1991.

[16] D.L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52(4):1289–1306, 2006.

[17] D.L. Donoho. For most large underdetermined systems of equations, the minimal l1-norm near-solution approximates the sparsest near-solution. *Communications on pure and applied mathematics*, 59(7):907–934, 2006.

[18] R. Gallager. Low-density parity-check codes. *IRE Trans. Inform. Theory*, 8(1):21–28, 1962.

[19] R.G. Gallager. *Information theory and reliable communication*, volume 15. Wiley, 1968.

[20] AC Gilbert and JA Tropp. Applications of sparse approximation in communications. In *Proc. Int. Symp. Inform. Theory*, pages 1000–1004. IEEE, 2005.

[21] C. Huang, G.H.L. Cheang, and A.R. Barron. Risk of penalized least squares, greedy selection and l1 penalization for flexible function libraries. *Submitted to Ann. Statist.*, 2008.

[22] L. Jones. A simple lemma for optimization in a hilbert space, with application to projection pursuit and neural net training. *Ann. Statist.*, 20:608–613, 1992.

[23] A. Joseph. Variable selection in high dimensions with random designs and orthogonal matching pursuit. *Available at arXiv:1109.0730*, 2011.

[24] W.S. Lee, P.L. Bartlett, and R.C. Williamson. Efficient agnostic learning of neural networks with bounded fan-in. *IEEE Trans. Inform. Theory*, 42(6):2118–2132, 1996.

[25] S. Lin and D.J. Costello. *Error control coding: fundamentals and applications*. Pearson Education India, 2004.

[26] S.G. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

[27] N. Meinshausen and P. Buhlmann. High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462, 2006.

[28] Y.C. Pati, R. Rezaiifar, and PS Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Conf. Rec. 27th Asilomar Conf. Sig., Sys. and Comput.*, pages 40–44. IEEE, 1993.

[29] Y. Polyanskiy, H.V. Poor, and S. Verdú. Channel coding rate in the finite blocklength regime. *IEEE Trans. Inform. Theory*, 56(5):2307–2359, 2010.

[30] I.S. Reed and G. Solomon. Polynomial codes over certain finite fields. *Journal of the Society for Industrial and Applied Mathematics*, 8(2):300–304, 1960.

[31] G. Reeves and M. Gastpar. Sampling bounds for sparse support recovery in the presence of noise. In *Proc. Int. Symp. Inform. Theory*, pages 2187–2191. IEEE, 2008.

[32] C.E. Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

[33] S.J. Szarek. Condition numbers of random matrices. *Journal of Complexity*, 7(2):131–149, 1991.

[34] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, pages 267–288, 1996.

[35] J.A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Trans. Inform. Theory*, 50(10):2231–2242, 2004.

[36] J.A. Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Trans. Inform. Theory*, 52(3):1030–1051, 2006.

[37] J.A. Tropp and A.C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Trans. Inform. Theory*, 53(12):4655–4666, 2007.

[38] M.J. Wainwright. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Trans. Inform. Theory*, 55(12):5728–5741, 2009.

[39] M.J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE Trans. Inform. Theory*, 55(5):2183–2202, 2009.

[40] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T.S. Huang, and S. Yan. Sparse representation for computer vision and pattern recognition. *Proceedings of the IEEE*, 98(6):1031–1044, 2010.

[41] C.H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, 38(2):894–942, 2010.

[42] T. Zhang. Adaptive forward-backward greedy algorithm for sparse learning with linear models. NIPS, 2008.

[43] P. Zhao and B. Yu. On model selection consistency of lasso. *J. Mach. Learn. Res.*, 7:2541–2563, 2006.