Least Squares Superposition Codes of Moderate Dictionary Size, Reliable at Rates up to Capacity

Andrew R Barron, Antony Joseph
Department of Statistics, Yale University
Email: andrew.barron@yale.edu, antony.joseph@yale.edu

Presented at the *IEEE International Symposium on Information Theory*, Austin, TX, June 13-18, 2010.

***Abstract*— Sparse superposition codes are developed for the additive white Gaussian noise channel with average codeword power constraint. Codewords are linear combinations of subsets of vectors, with the possible messages indexed by the choice of subset. Decoding is by least squares, tailored to the assumed form of linear combination. Communication is shown to be reliable with error probability exponentially small for all rates up to the Shannon capacity.**

I. Introduction

We introduce classes of superposition codes for the additive white Gaussian noise channel (AWGN) and analyze their properties. We show superposition codes from polynomial size dictionaries with least squares decoding achieve exponentially small error probability for any communication rate less than the Shannon capacity. A companion paper [1] provides a fast decoding method and its analysis.

The familiar communication problem is as follows. An encoder is required to map input bit strings $u = (u_1, u_2, \dots, u_K)$ of length K into codewords which are length n strings of real numbers $c_1, c_2, \dots, c_n$, with norm expressed via the power $(1/n)\sum_{i=1}^n c_i^2$. We constrain the average of the power across the 2^K codewords to be not more than P. The channel adds independent $N(0, \sigma^2)$ noise to the selected codeword yielding a received length n string Y. A decoder is required to map it into an estimate $\hat{u}$ which we want to be a correct decoding of u. Block error is the event $\hat{u} \neq u$, bit error at position i is the event $\hat{u}_i \neq u_i$, and the bit error rate is $(1/K)\sum_{i=1}^K 1_{\{\hat{u}_i \neq u_i\}}$. The reliability requirement is that, with sufficiently large n, the bit error rate is small with high probability or, more stringently, the block error probability is small, averaged over input strings u as well as the distribution of Y. The communication rate $R = K/n$ is the ratio of the input length to the codelength for communication across the channel.

The supremum of reliable rates of communication is the channel capacity $C = (1/2)\log_2(1{+}P/\sigma^2)$, by traditional information theory [13],[9],[5]. For practical coding the challenge is to achieve rates close to capacity, while guaranteeing reliable decoding in manageable computation time.

Our framework is as follows. Start with a list (or book) $X_1, X_2, \dots, X_N$ of vectors, each with n coordinates, for which the codeword vectors take the form of superpositions $\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_N X_N$. The received vector is in accordance with the statistical linear model $Y = X\beta + \varepsilon$, where X is the matrix whose columns are the vectors $X_1, X_2, \dots, X_N$ and ε is the noise vector distributed Normal$(0, \sigma^2 I)$.

Our coefficient vectors β are arranged to be of a specified form which we call partitioned superposition coding. In this case, the book X is split into L sections of size B, with one term selected from each, yielding L terms in each codeword out of a dictionary of size $N = LB$. Likewise, the coefficient vector β is split into sections, with one coordinate non-zero in each section to indicate the selected term. We draw entries of X independently from a normal distribution with mean zero and a variance P/L and set the non-zero coefficients to have magnitude 1. Note, since a linear combination of two codewords does not correspond to a codeword, these are not linear codes in the traditional algebraic coding sense.

Most convenient is the case that the sizes of these sections are powers of two. Then an input bit string of length $K = L\log_2 B$ splits into L substrings of size $\log_2 B$. The mapping from u to β is then obtained 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. With one term from each section, the number of codewords is B^L.

Optimal decoding for minimal average probability of error consists of finding the codeword $X\beta$ with coefficient vector β of the assumed form that minimizes $\|Y - X\beta\|^2$.

For a subset $S = \{j_1, j_2, ..., j_L\}$ with j_l in section L, the codeword is $X_S = \sum_{l=1}^L X_{j_l}$. The least squares decoder produces estimates $\hat{j}_1, \hat{j}_2, ..., \hat{j}_L$. We seek to control the distribution of the section mistake rate $(1/L)\sum_{l=1}^L 1_{\{\hat{j}_l \neq j_l\}}$.

The heart of the analysis shows that competing codewords that differ in a fraction of at least α_0 terms are exponentially unlikely to have smaller distance from Y than the true codeword, provided that the section size $B = L^a$ is polynomially large in the number of sections L, where a sufficient value of a is determined. There is a positive constant c such that for rates R less than the capacity C with a gap $\Delta = C{-}R$ not too large, for any $\alpha_0 > 0$, the probability that the fraction of mistakes is at least α_0 is not more than $\exp\{-nc\min\{\Delta^2, \alpha_0\}\}$. Thus for prescribed error probability ϵ, if $\alpha_0 \leq (1/nc)\log(1/\epsilon)$ the rate R is of the form $C - \sqrt{1/(nc)}\sqrt{\log(1/\epsilon)}$, in agreement with the best possible order as in [11].

Moreover, an approach is discussed which we corrects the small fraction of remaining mistakes by arranging sufficient distance between the subsets, using composition with an outer Reed-Solomon (RS) code of rate near one.

Regarding relationships with previous work, standard approaches to dealing with AWGN channels involve separate coding and signal constellation shaping [8], whereas we build the shaping directly into the sparse superposition code. Superposition codes for Gaussian channels began with Cover [4] in the context of determination of the rate region for multiple-user channels, whereas here the idea is put to use to simplify the original Shannon single-user problem. The analysis of concatenated codes in Forney [7] is also a forerunner to the development we give here. He identified benefits of an outer Reed-Solomon code paired with an optimal inner code.

Our conclusions also complement recent work on sparse signal recovery [14],[6],[3]. Their work shows that for reliable determination of L terms from noisy measurements, having the number of such measurements n be of order $L \log B$ is sufficient, and is achieved by convex optimization with an ℓ_1 control on the coefficients. This translates into saying that the communication rate with this procedure is positive even though a precise value for its rate is not identified there. Wainwright [15] uses information-theoretic techniques to give converse as well as achievability bounds on sparse signal recovery. Ours differs from this in that while he deals with exact support recovery, our sparse regression setup is used as an inner code to obtain a low section error rate allowing us to achieve rates closer to capacity. Further, partitioning renders more explicit answers with regards to the relationships between the distribution of mistakes, the communication rate, and the blocklength for a given signal to noise ratio.

Our ideas of sparse superposition coding are adapted to Gaussian vector quantization in [10].

II. Preliminaries

For vectors a, b of length n, 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. The logarithms are taken to be base e, unless otherwise specified.

We make repeated use of the following fact- If Z and $\tilde{Z}$ are normal with means equal to 0, variances equal to 1, and correlation coefficient ρ then $\mathbb{E}(e^{(\lambda/2)(Z^2-\tilde{Z}^2)})$ takes the value $1/[1-\lambda^2(1-\rho^2)]^{1/2}$ when $\lambda^2 < 1/(1-\rho^2)$ and infinity otherwise. For positive Δ we define the quantity $D = D(\Delta, 1-\rho^2)$ given by

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

It is not hard to see that D is an increasing function of the ratio $\Delta^2/(1-\rho^2)$. The expression corresponding to D but with the maximum restricted to $0 \leq \lambda \leq 1$ is denoted $D_1 = D_1(\Delta, 1-\rho^2)$. Note that it is obvious from the above that both D and D_1 are non-negative.

III. Performance of Least Squares

For a subset S of size L we measure how different it is from S^*, the subset that was sent. Note that $\ell = card(S - S^*)$ is the number of sections incorrectly decoded. Further, let $\hat{S}$ be the least squares solution, or an approximate least squares solution, achieving $|Y - X_{\hat{S}}|^2 \leq |Y - X_{S^*}|^2 + \delta_0$ with $\delta_0 \geq 0$.

Let $C_\alpha = \frac{1}{2}\log(1+\alpha v)$ 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 amount $\alpha(C-R)$, positive when the rate R is less than the Shannon capacity C.

Our first result on the distribution of the number of mistakes is the following.

Lemma 1: Set $\alpha = \ell/L$ for an $\ell \in \{1, 2, \ldots, L\}$. For approximate least squares with $0 \leq \delta_0 \leq 2\sigma^2(C_\alpha - \alpha R)$, the probability of a fraction $\alpha = \ell/L$ mistakes is upper bounded by

$$\binom{L}{\alpha L} \exp\left\{-nD_1(\Delta_\alpha, \alpha v/(1+\alpha v))\right\},$$

where $\Delta_\alpha = C_\alpha - \alpha R - (\delta_0/2\sigma^2)$ and v is the signal-to-noise ratio.

Proof: To incur ℓ mistakes, there must be an allowed subset S of size L which differs from the subset S^* sent in an amount $card(S^* - S) = \ell$ which undesirably has squared distance $|Y - X_S|^2$ less than or equal to the value $|Y - X_{S^*}|^2 + \delta_0$ achieved by S^*.

The analysis proceeds by considering an arbitrary such S, bounding the probability that $|Y - X_S|^2 \leq |Y - X_{S^*}|^2 + \delta_0$, and then using an appropriately designed union bound to put such probabilities together.

Consider the statistic $T = T(S)$ given by

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

We set a threshold for this statistic equal to $t = \delta_0/(2\sigma^2)$. The event of interest is that $T \leq t$.

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 $\ell = \alpha L$. Given $(X_j : j \in S)$ the actual density of Y is normal with mean $X_{S_1} = \sum_{j \in S_1} X_j$ and variance $(\sigma^2 + \alpha P)I$ and we denote this density $p(Y|X_{S_1})$. In particular, there is conditional independence of Y and X_{S_2} given X_{S_1}.

Consider the alternative hypothesis of a conditional distribution for Y given X_{S_1} and X_{S_2} which is Normal$(X_S, \sigma^2 I)$. It is the distribution which would have governed Y if S were sent. Let $p_h(Y|X_{S_1}, X_{S_2}) = p_h(Y|X_S)$ be the associated conditional density. 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 above test statistic as

$$\frac{1}{2}\left[\frac{|Y - X_{S_1}|^2}{\sigma^2 + \alpha P} - \frac{|Y - X_{S^*}|^2}{\sigma^2}\right] + \frac{1}{2}\left[\frac{|Y - X_S|^2}{\sigma^2} - \frac{|Y - X_{S_1}|^2}{\sigma^2 + \alpha P}\right].$$

Calling the two parts T_1 and T_2, respectively, 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

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

We are examining the event E_ℓ that there is an allowed subset $S = S_1 \cup S_2$ (with $S_1 = S \cap S^*$ of size $L - \ell$ and $S_2 = S - S_1$ of size ℓ) such that that $T(S)$ is less than t. For positive λ the indicator of this event satisfies

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

Here the outer sum is over $S_1 \subset S^*$ of size $L - \ell$. 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 obtain that $\mathbb{P}[E_\ell]$ is not more than

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

A simplification here is that the true density for X_{S_2} is independent of the conditioning variables Y, X_{S_1} and X_{S^*}.

We arrange for λ to be not more than 1. Then by Jensen's inequality, the conditional expectation may be brought inside the λ power and inside the inner sum, yielding

$$\mathbb{P}[E_\ell] \leq \sum_{S_1} \mathbb{E} e^{-n\lambda(T_1(S_1)-t)} \left(\sum_{S_2} \mathbb{E}_{X_{S_2}|Y,X_{S_1}} e^{-nT_2(S)} \right)^{\lambda}.$$

We also note that

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

where the above follows from Bayes rule by noting equality of $\frac{p_h(X_{S_2}|Y,X_{S_1})}{p(X_{S_2}|X_{S_1})}$ and $\frac{p_h(Y|X_{S_1},X_{S_2})}{p(Y|X_{S_1})}$ and that the true density for X_{S_2} is independent of the conditioning variables. So when we take the expectation of this ratio we cancel the denominator leaving the numerator density which integrates to 1. Consequently, the resulting expectation of $e^{-nT_2(S)}$ is not more than e^{-nC_α}. The sum over S_2 entails less than $B^\ell = e^{nR\ell/L}$ choices so the bound is

$$\mathbb{P}[E_\ell] \leq \sum_{S_1} \mathbb{E} e^{-n\lambda T_1(S_1)} e^{-n\lambda[C_\alpha - \alpha R - t]}.$$

Now $nT_1(S_1)$ 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 the expectation $\mathbb{E}e^{-n\lambda T_1(S_1)}$ is found to be equal to $[1/[1 - \lambda^2 \alpha v/(1+\alpha v)]]^{n/2}$. When plugged in above it yields the claimed bound optimized over λ in $[0, 1]$. We recognize that the exponent takes the form $D_1(\Delta, 1-\rho^2)$ with $1-\rho^2 = \alpha v/(1+\alpha v)$ as discussed in the preliminaries. This completes the proof of Lemma 1.

A difficulty with the Lemma 1 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. The following lemma proves to be useful in this regard.

Lemma 2: Let a positive integer $\ell \leq L$ be given and let $\alpha = \ell/L$. Suppose $0 \leq t < C_\alpha - \alpha R$. As above let E_ℓ be the event that there is an allowed L term subset S with $S - S^*$ of size ℓ such that $T(S)$ is less than t. Then $\mathbb{P}[E_\ell]$ is bounded by the minimum for t_α in the interval between t and $C_\alpha - \alpha R$ of the following

$$\binom{L}{L\alpha} \exp\left\{-nD_1(C_\alpha - \alpha R - t_\alpha, 1-\rho_\alpha^2)\right\}$$
$$+ \exp\big\{ -nD(t_\alpha - t, \alpha^2 v/(1+\alpha^2 v)]\big\}.$$

where $1-\rho_\alpha^2 = \alpha(1-\alpha)v/(1+\alpha v)$.

Proof Sketch: 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

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

Likewise we split the threshold $t = \tilde{t} + t^*$ where $t^* = -(t_\alpha - t)$ is negative and $\tilde{t} = t_\alpha$ is positive. The event that $T(S) < t$ is contained in the union of the two events $\tilde{T}(S) < \tilde{t}$ and $T^* < t^*$. Now proceed as in Lemma 1. See [2] for details.

IV. Sufficient Section Size

We come to the matter of sufficient conditions on the section size B for our exponential bounds to swamp the combinatorial coefficient, for partitioned superposition codes.

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. Here the size of a controls the polynomial size of the dictionary $N = L^{a+1}$.

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 let's 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 swamp the combinatorial coefficient.

Define $D_{\alpha,v} = D_1(\triangle_\alpha, 1-\rho_\alpha^2)$ and $\tilde{D}_{\alpha,v} = D_1(\tilde{\triangle}_\alpha, 1-\rho_\alpha^2)$. 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 wash out the affect of the log combinatorial coefficient $\log\binom{L}{L\alpha}$. That is, we want to arrange for the nonnegativity of the difference

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

This difference is small for α near 0 and 1. Its constituent quantities have a shape comparable to multiples of $\alpha(1-\alpha)$.

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

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

where the maximum is for α in $\{1/L, 2/L, \ldots, 1-1/L\}$.

We show that $a_{v,L}$ is fairly insensitive to L, with the value close to a limit a_v characterized by the ratio in the vicinity of $\alpha = 1$.

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

Lemma 3: The section size rate $a_{v,L}$ has a continuous limit $a_v = \lim_{L\to\infty} a_{v,L}$ which is given, for $0 < v < v^*$, by

$$a_v = \frac{R}{[(1+v)\log(1+v) - v]^2/[8v(1+v)]}$$

and for $v \geq v^*$ by

$$a_v = \frac{R}{[(1+v)\log(1+v) - 2v]/[2(1+v)]}$$

where v is the signal-to-noise ratio. With R replaced by $C = (1/2)\log(1+v)$ and using log base e, in the case $0<v<v^*$, it is

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

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

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

which asymptotes to the value 1 for large v.

Proof Sketch: It is not hard to see that for large L the maximum in the expression for $a_{v,L}$ occurs at α near 0 or 1. Taking the limit and using L'Hopital's rule we get

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

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. Simplifying this we get the above expression. See [2] for details.

While a_v is large for small v, it has reasonable values for moderate 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 proscribed high fraction of C, say $R = 0.8C$, for α_0 a proscribed small target fraction of mistakes, say $\alpha_0 = 0.1$, and for ϵ to be a small target probability, so as to obtain $\min\{P[E_\ell], P[\tilde{E}_\ell] + P[E^*_\ell]\} \leq \epsilon$, taking the minimum over allowed values of t_α, for every $\alpha = \ell/L$ at least α_0. This is illustrated in Figure 1 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.

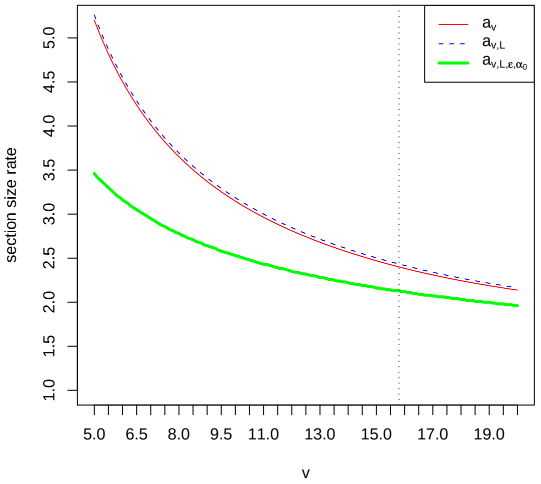

Fig. 1. Sufficient section size rate a as a function of the signal-to-noise ratio v. The dashed curve shows $a_{v,L}$ at $L = 64$. Just below it the thin solid curve is the limit for large L. For section size $B \geq L^a$ the error probabilities are exponentially small for all $R < C$ and any $\alpha_0 > 0$. 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 $L = 64$.

V. Confirming Exponentially Small Probability

Here we put the above conclusions together to demonstrate the reliability of approximate least squares. The probability of the event of more than any small positive fraction of mistakes $\alpha_0 = \ell_0/L$ is shown to be exponentially small.

Let *mistakes* denote the number of mistakes in the approximate least square solution.Suppose the threshold $t = \frac{\delta_0}{2\sigma}$ is not more than $(1/2)\min_{\alpha\geq\alpha_0}\{\alpha(C-R) + (\tilde{\Delta}_\alpha - \Delta^{\min}_\alpha)\}$, where $\Delta^{\min}_\alpha$ is the solution to the equation

$$nD_1(\Delta^{\min}_\alpha, 1-\rho^2_\alpha) = \log\binom{L}{L\alpha}.$$

Some natural choices for the threshold include $t = 0$ and $t = (1/2)\alpha_0(C-R)$. For positive x let $g(x) = \min\{x, x^2\}$.

Theorem 4: Suppose the section size rate a is at least $a_{v,L}$, that the communication rate R is less than the capacity C with codeword length $n = (1/R)aL\log L$, and that we have an approximate least squares estimator. For ℓ_0 between 1 and L, the probability $\mathbb{P}[mistakes \geq \ell_0]$ is bounded by the sum over integers ℓ from ℓ_0 to L of $\mathbb{P}[E_\ell]$ using the minimum of the bounds from Lemmas 1 and 2. It follows that there is a positive constant c, such that for all α_0 between 0 and 1,

$$\mathbb{P}[mistakes \geq \alpha_0 L] \leq 2L\exp\{-nc\min\{\alpha_0, g(C-R)\}\}.$$

Hence, for any fixed α_0, a, and R, not depending on L, satisfying $\alpha_0 > 0$, $a > a_v$ and $R < C$, we conclude that this probability is exponentially small.

Proof Outline: Consider the simple case when t is less than a fixed fraction of $\alpha_0(C-R)$. Let $\Delta_\alpha = \tilde{\Delta}_\alpha + \alpha(C–R) - t_\alpha$, with t_α between t and $\alpha(C-R)$. Consider the exponent $D_{\alpha,v} = D_1(\Delta_\alpha, 1-\rho_\alpha^2)$ as given in the preceding section.

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,

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

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

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

when $\Delta < (1-\rho_\alpha^2)/\rho_\alpha^2$, and this derivative is equal to 1 otherwise.

Now lower bound the components of this tangent line. First lower bound the derivative $D' = D_1'(\Delta)$ evaluated at $\Delta = \tilde{\Delta}_\alpha$. As in our developments in previous sections $\tilde{\Delta}_\alpha^2/(1-\rho_\alpha^2)$ is a bounded function of α. Moreover, $\tilde{\Delta}_\alpha$ and $1-\rho_\alpha^2$ are positive functions of order $\alpha(1-\alpha)$ in the unit interval, with ratio tending to positive values as α tends to 0 and 1, so their ratio is uniformly bounded away from 0. Consequently $w_v = \min_\alpha D_1'(\tilde{\Delta}_\alpha)$ is strictly positive.

Now we are in position to apply the above lemmas. If the section size rate a is at least $a_{v,L}$ we have that $n\tilde{D}_{\alpha,v}$ cancels the combinatorial coefficient and hence the first term in the $\mathbb{P}[E_\ell]$ bound (the part controlling $\mathbb{P}[\tilde{E}_\ell]$) is not more than $\exp\{-n[\Delta_\alpha - \tilde{\Delta}_\alpha]\, D'\}$, where $\alpha = \ell/L$. This yields $\mathbb{P}[E_\ell]$ not more than the sum of

$$\exp\{-n[\alpha(C-R) - t_\alpha]\, D'\}$$

and

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

for any choice of t_α between t and $\alpha(C-R)$. For instance one may choose t_α to be half way between t and $\alpha(C-R)$. Now since t is less than a fixed fraction of $\alpha_0(C-R)$, we have arranged for both $\alpha(C–R) - t_\alpha$ and $t_\alpha - t$ to be of order $\alpha(C-R)$ uniformly for $\alpha \geq \alpha_0$.

Accordingly, the first of the two parts in the bound has exponent exceeding a quantity of order $\alpha_0(C–R)$. The second part has exponent related to a function of the ratio $h = (\alpha(C-R))^2/[\alpha^2 v/(1+\alpha^2 v)]$ which is of order h for small h and order $\sqrt{h}$ for large h. Here h is of order $(C-R)^2$ uniformly in α. It follows that there is a constant c (depending on v) such that

$$\mathbb{P}[E_\ell] \leq 2\exp\{-nc\min\{\alpha_0, g(C-R)\}\}.$$

VI. From Small Fraction of Mistakes to Small Probability of Any Mistake

Theorem 5: To demonstrate a practical partitioned superposition code with small block error probability it is enough to have demonstrated such a code for which the bit error rate is small with high probability.

Proof Sketch: Place an RS code with rate near 1 between the original message stream and the input to the superposition code and likewise an RS decoder after the output of the superposition code. If $R_{outer} = 1 - \delta$ be the rate of the RS code with $0 < \delta < 1$, then section error rates less than α_0 can be corrected provided $2\alpha_0 < \delta$. An efficient choice is to make the target fraction to correct a fixed fraction of the gap $C - R$ between the capacity and the rate of the inner superposition code. Further, if R_{inner} be the rate associated with our inner (superposition) code then the composite rate is given by $R_{inner}R_{outer}$. The end result, using our theory for the distribution of the fraction of mistakes of the superposition code, is that the block error probability is within a log factor of being exponentially small. One may regard the composite code as a superposition code in which the subsets are forced to maintain at least a certain minimal separation, so that decoding to within a certain distance from the true subset implies exact decoding.

References

[1] A.R. Barron, A. Joseph, "Toward Fast Reliable Communication at Practical Rates Near Capacity with Gaussian Noise," *Proc. IEEE Int. Symp. Inform. Theory* Austin, TX, June 13-18, 2010.

[2] A.R. Barron, A. Joseph, "Least Squares Superposition Codes of Moderate Dictionary Size, Reliable at Rates up to Capacity," Submitted to the *IEEE Trans. Inform. Theory* June 4, 2010.

[3] E. Candes and Palm, "Near-ideal model selection by ℓ_1 minimization," *Annals of Statistics*, 2009.

[4] T.M. Cover, "Broadcast channels," *IEEE Trans. Inform. Theory*, vol.18, pp.2-14, 1972.

[5] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, New York, Wiley-Interscience, 2006.

[6] D.L. Donoho, M. Elad, and V.M. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Inform. Theory*, vol.52, no.1, pp.6-18, Jan. 2006.

[7] G.D. Forney, Jr. *Concatenated Codes*, Research Monograph No. 37, Cambridge, Massachusetts, M.I.T. Press. 1966.

[8] G.D. Forney, Jr., G. Ungerboeck, "Modulation and Coding for Linear Gaussian Channels," *IEEE Trans. Inform. Theory*, vol. 44, no.6, pp.2384-2415, October 1998.

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

[10] I. Kontoyiannis, S. Gitzenis, and K.R. Rad, "Superposition codes for Gaussian vector quantization," *2010 IEEE Information Theory Workshop*, Cairo, Egypt, Jan. 2010.

[11] Y. Polyanskiy, H.V. Poor, S. Verdú, "Channel Coding Rate in the Finite Blocklength Regime." *IEEE Trans. Inform. Theory*, 2010.

[12] I.S. Reed and G. Solomon, "Polynomial codes over certain finite fields." *J. SIAM*, vol.8, pp.300-304, June 1960.

[13] C.E. Shannon, "A mathematical theory of communication." *Bell Syst. Tech. J.*, vol.27, pp.379-423 and 623-656.

[14] M.J. Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso)." *IEEE Trans. Inform. Theory*, vol.55, no.5, pp.2183-2202, May 2009.

[15] M. J. Wainwright, "Information-theoretic limitations on sparsity recovery in the high-dimensional and noisy setting." *IEEE Transactions on Information Theory*, vol.55, pp.5728–5741, December 2009