High-Rate Sparse Superposition Codes with Iteratively Optimal Estimates

Andrew R. Barron and Sanghee Cho
Department of Statistics, Yale University, New Haven, CT 06520 USA
e-mail: {andrew.barron, sanghee.cho}@yale.edu

Abstract— **Recently sparse superposition codes with iterative term selection have been developed which are mathematically proven to be fast and reliable at any rate below the capacity for the additive white Gaussian noise channel with power control. We improve the performance using a soft decision decoder with Bayes optimal statistics at each iteration, followed by thresholding only at the final step. This presentation includes formulation of the statistics, proof of their distributions, numerical simulations of the performance improvement, and useful identities relating a squared error risk to a posterior probability of error.**

I. Introduction

Sparse superposition codes use a dictionary X consisting of vectors $X_1, X_2, \ldots, X_N$, each of n coordinates. The codeword vectors $X\beta$ are superpositions $\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_N X_N$. The entries of X are independent N$(0,1)$. The codeword is conveyed through the choice of which L of the coefficients are non-zero, where L matches n to within a log factor, yet L is a small fraction of the dictionary size N. For a channel with additive white Gaussian noise (AWGN), with superposition coding, what is received is $Y = X\beta + \varepsilon$, a vector of length n, where ε is the noise vector with distribution N$(0, \sigma^2 I)$.

The coefficient vector is split into L sections each of size $M = N/L$, with one non-zero entry in each section. There are M^L codewords. With M a power of 2, the encoding from an input bit-string $u_1, u_2, \ldots, u_K$, with $K = L \log M$, consists of partitioning the string into L substrings of length $\log M$ which index the terms chosen to be included in the codeword. Denote these indices $\{j_1, j_2, \ldots, j_L\}$. The non-zero coefficient from section ℓ takes value $\beta_{j_\ell} = \sqrt{P_\ell}$, with $\sum_\ell P_\ell = P$ to control the codeword power.

The rate of the code is $R = (L \log M)/n$ and the capacity of the AWGN channel is $\mathcal{C} = (1/2)\log(1+snr)$ where $snr = P/\sigma^2$ is the signal-to-noise ratio.

These sparse superposition codes with an adaptive successive decoder are computationally fast codes for the Gaussian noise channel with any fixed rate below capacity, with error probability proven to be exponentially small. See [1], [2], [3] for this conclusion and the relationship to other literature on compressive sensing, signal recovery, and coding for the Gaussian channel. The adaptive successive decoder uses iteratively obtained test statistics related to inner products of the X_j with residuals of the previous fits. There, for each fit update, the decoder accepts those terms j for which the statistic is above a threshold, chosen high enough to avoid false alarms.

The conditional distribution of such statistics is approximated as a normal random variable shifted, for the true terms compared to the others, by an amount inversely related to the squared distance between the current fit and the true coefficient vector. Accordingly, we seek improved estimates of the coefficients, using the squared error loss, to increase the separation between the distributions of the statistics and thereby improve the reliability of the final decision.

We use the Bayes optimal coefficient estimates based on the distribution of iteratively obtained statistics, with a uniform prior on the choice of the j_ℓ sent from each section. These estimates use computed posterior weights of terms in each section. These weights provide a soft decision decoding, rather than the $\{0,1\}$ valued weights associated with thresholding.

We formulate the statistics, quantify their distribution, and give identities that relate the expected sum of squared distance (Bayes risk) of the estimate to an expected posterior probability of error. Taking 1 minus it quantifies the rate of success.

A function $g(x)$ gives the expected success rate on a step if the previous success rate was x. Numerical evaluations show it is higher than the corresponding function from the threshold-based decisions. Indeed, $g(x)$ stays above x for a longer extent of the interval $[0,1]$ than previously obtained, leading to a smaller fraction of mistakes. As before, this remaining fraction of mistakes are corrected by an optional outer code.

To get rates approaching capacity, a variable power allocation is used, with P_ℓ proportional to $e^{-2\mathcal{C}\,\ell/L}$.

For $R < C$, theoretically optimal codes have exponentially small error probability with exponent of order $(\mathcal{C}-R)^2\, n$. So far, fast sparse superposition codes in [1], [2], [3] achieve comparable exponents of order $(\mathcal{C}–R)^2\, L$, within a logarithmic factor of the optimal form. More specifically, the bounds on the error probability have the exponents $(\mathcal{C}_M - R)^2$ in place of $(\mathcal{C}-R)^2$, where $\mathcal{C}_M$ slowly approaches the capacity $\mathcal{C}$, and the bounds are applicable only for $R < \mathcal{C}_M$. This motivates additional effort, initiated here, to improve the rate and reliable tradeoff of sparse superposition codes by improving the adaptive successive decoder.

II. Results

A. Iterative Decoding Statistics and their Distribution

In the process of estimating which terms were sent, the decoder develops a sequence of estimates $\hat{\beta}_k$ of the true coefficient vector β, and corresponding fits $F_k = X\hat{\beta}_k$ of the codeword $X\beta$. Let $J = \{1,\ldots,N\}$ be the full set of indices.

The initial estimate uses $stat_{0,j} = \mathcal{Z}_{0,j} = X_j^T Y/\|Y\|$ and later estimates use similar inner products with residuals of fits in place of Y. The distribution of $stat_{0,j}$ is approximately a standard normal shifted by $\beta_j \sqrt{n/(\sigma^2 + P)}$. The $\sigma^2 + P$ in the denominator is from the variance of the coordinates

978-1-4673-2579-0/12/$31.00 ©2012 IEEE

of Y. The idea of the steps is to use residuals of successive fits to gradually reduce it to σ^2. This increases the amount by which the distribution is shifted, thereby improving the distinguishability of the true terms from the others.

Define the shift factor $\alpha_\ell = \sqrt{P_\ell\, n/(\sigma^2+D)}$. The shift in section ℓ takes the form $\alpha_\ell\, 1_{\{j=j_\ell\}}$. Initially $D_0 = P$. For subsequent steps, the role of D is played by $\|\hat\beta_k - \beta\|^2$ or its expected value $D_k = \mathbb{E}\|\hat\beta_k - \beta\|^2$, which quantifies for the statistics we develop the level of remaining interference in the residuals due to inaccuracy of $\hat\beta_k$. The associated shift factor is $\alpha_{\ell,k} = \sqrt{P_\ell\, n/(\sigma^2+D_k)}$.

Let $\hat\beta_k$ be any estimate constructed from statistics $stat_{k-1} = (stat_{k-1,j}, j \in J)$ computed on the previous step. For instance $stat_{k-1,j}$ could be the inner products of residuals with the columns of X. Initialize $G_0 = Y$. For $k \geq 1$, let $F_k = X\hat\beta_k$ and let G_k be the part of the F_k orthogonal to $G_0, G_1, \ldots, G_{k-1}$. Assume the current fit $X\hat\beta_k$ is not in the linear span of the previous fits, so $\|G_k\| > 0$. Let $\mathcal{Z}_{k,j} = X_j^T G_k/\|G_k\|$ be the normalized inner product of X_j and G_k.

The $\mathcal{Z}_k = (\mathcal{Z}_{k,j}, j \in J)$ and $\|G_k\|$ are used to update $stat_k$ and then $\hat\beta_{k+1}$. Require that $stat_k$ and $\hat\beta_{k+1}$ be functions of $\mathcal{F}_k = (\mathcal{Z}_0, \|G_0\|, \ldots, \mathcal{Z}_k, \|G_k\|)$. Our first lemma provides the conditional distribution of $\mathcal{Z}_k$ and $\|G_k\|$ given $\mathcal{F}_{k-1}$. For $k = 0$ there is no conditioning $\mathcal{F}_{k-1}$.

For analysis purposes, let $\beta_e, \hat\beta_{1,e}, \ldots, \hat\beta_{k,e}$ be the vectors in R^{N+1} obtained by appending an extra coordinate to the vectors $\beta, \hat\beta_1, \ldots, \hat\beta_k$ in R^N. The value of the extra coordinate for β_e is σ and for the $\hat\beta_{1,e}, \ldots, \hat\beta_{k,e}$ it is 0. The subscript e denotes that the vectors are thus extended.

Likewise X_e denotes an extended dictionary with an additional column ε/σ. Armed with this extension we have opportunity to use a standard linear model trick representing $Y = X_e\beta_e$. Then the $G_0, G_1, \ldots, G_{k-1}$ are the successive orthogonal components of $X_e\beta_e, X_e\hat\beta_{1,e}, \ldots, X_e\hat\beta_{k-1,e}$.

Parallel to the development of these vectors G in R^n, let $b_{0,e}, b_{1,e}, \ldots, b_{k,e}$ be defined as the vectors in R^{N+1} of successive orthonormalization of $\beta_e, \hat\beta_{1,e}, \ldots, \hat\beta_{k,e}$ and let $b_0, b_1, \ldots, b_k$, respectively, be the vectors formed from the corresponding upper N coordinates.

Let $\Sigma_{k,e} = I - (b_{0,e}b_{0,e}^T + b_{1,e}b_{1,e}^T + \ldots + b_{k,e}b_{k,e}^T)$ be the $R^{(N+1)\times(N+1)}$ matrix of projection onto the linear space orthogonal to $\beta_e, \hat\beta_{1,e}, \ldots, \hat\beta_{k,e}$. The upper left $N \times N$ portion of this matrix denoted Σ_k is the conditional covariance matrix below. The suggestion to interpret Σ_k as a portion of a projection matrix was made by our colleague Antony Joseph, who credits [4],[5] for some analogous thinking.

Also let $Proj_k$ be the matrix of projection onto the span of the estimates $\hat\beta_1, \ldots, \hat\beta_k$, and likewise $Proj_{k,e}$ in which 0 is appended to each of these estimates. $\Sigma_{k,e}$ differs from $I - Proj_{k,e}$ by accounting for orthogonality to β_e.

Lemma 1, proved in the appendix, generalizes conclusions from [3], [2] to handle the present generality.

Lemma 1: For $k \geq 0$, the conditional distribution $\mathbb{P}_{\mathcal{Z}_k|\mathcal{F}_{k-1}}$ of $\mathcal{Z}_k$ given $\mathcal{F}_{k-1}$ is determined by the representation

$$\mathcal{Z}_{k,j} = b_{k,j}\frac{\|G_k\|}{\sigma_k} \;+\; Z_{k,j},$$

where $Z_k = (Z_{k,j} : j \in J)$ has conditional distribution Normal$(0, \Sigma_k)$. Here $\sigma_0^2 = \sigma^2 + P$ and for $k \geq 1$ it is $\sigma_k^2 = \hat\beta_k^T \Sigma_{k-1} \hat\beta_k$. Moreover, $\|G_k\|^2/\sigma_k^2$ is distributed as a $\mathcal{X}_{n-k}^2$ random variable independent of the Z_k and the past $\mathcal{F}_{k-1}$.

Related to the distribution $\mathbb{P}_{Z_k|\mathcal{F}_{k-1}}$ is the distribution $Q_{Z_k|\mathcal{F}_{k-1}}$ which makes the Z_k be Normal$(0, I - Proj_k)$. The density ratio between $\mathbb{P}_{Z_k|\mathcal{F}_{k-1}}$ and $Q_{Z_k|\mathcal{F}_{k-1}}$ on R^N is uniformly bounded by the constant $\sqrt{1+snr}$.

The Chi-square random variable divided by n is close to the constant 1, except in events of exponentially small probability, as long as the number of steps k is small compared to n. Thus $\mathcal{Z}_k$ is approximately $\sqrt{n}\, b_k + Z_k$, a normal shifted by $\sqrt{n}\, b_k$. The distribution is further cleaned by addition of an independent normal of covariance $Proj_k$. This makes the cleaned Z_k be distributed N$(0, I)$ with respect to Q. Moreover, as in [2], the boundedness of the density ratio permits replacement of the distribution P with the simplified distribution that arises from Q, when determining events that have exponentially small probability. Henceforth for this summary we presume the cleaned shifted standard normal distribution for the $\mathcal{Z}_k$.

Consider $\mathcal{Z}_k^{comb} = \sqrt{\lambda_{k,0}}\, \mathcal{Z}_0 - \sum_{k'=1}^k \sqrt{\lambda_{k,k'}}\, \mathcal{Z}_{k'}$, where the vector $\underline{\lambda}_k = (\lambda_{k,k'} : 0 \leq k' \leq k)$ satisfies $\sum_{k'=0}^k \lambda_{k,k'} = 1$. These can be interpreted as shifts of the standard normals $Z_k^{comb} = \sqrt{\lambda_{k,0}}\, Z_0 - \sum_{k'=1}^k \sqrt{\lambda_{k,k'}}\, Z_{k'}$, where the shift arises from combinations of the $\sqrt{n} b_{k'}$. The task is to choose the coefficients of combination to produce a $stat_k$ with total shift of the desired form.

Motivation comes from the statistics $(Y - X\hat\beta_{k,-j})^T X_j$, or scalings thereof, where $\hat\beta_{k,-j}$ is the vector $\hat\beta_k$ with the contribution from the current j removed. It takes the form $(Y - X\hat\beta_k)^T X_j + \|X_j\|^2 \hat\beta_{k,j}$. We also find motivation from development of approximate Bayes optimality properties. The $stat_k$ take the following form, for some choice of vector $\underline{\lambda}_k$ and some c_k typically between σ^2 and $\sigma^2 + P$,

$$stat_k = \mathcal{Z}_k^{comb} + \frac{\sqrt{n}}{\sqrt{c_k}}\, \hat\beta_k$$

The combination should be such that these statistics have the representation $Z_k^{comb} + \frac{\sqrt{n}}{\sqrt{c_k}}\, \beta$, with the desired shift $\frac{\sqrt{n}}{\sqrt{c_k}}\, \beta$.

Here are three related examples of such statistics. Idealized case [B] has the desired form and case [C] approximately so. Case [A] is similar, but has additional randomness from weights based on $\mathcal{Z}_{k'}^T \hat\beta_k/\sqrt{n}$ rather than $b_{k'}^T \hat\beta_k$.

[A] **Based on residuals:** Let

$$stat_k = \frac{X^T(Y - X\hat\beta_k)}{\sqrt{n\, c_k}} + \frac{\sqrt{n}}{\sqrt{c_k}}\hat\beta_k,$$

with $nc_k = \|Y - X\hat\beta_k\|^2$, from $\underline{\lambda}_k$ proportional to

$$\Big((\|Y\| - \mathcal{Z}_0^T\hat\beta_k)^2, (\mathcal{Z}_1^T\hat\beta_k)^2, \ldots, (\mathcal{Z}_k^T\hat\beta_k)^2\Big).$$

[B] **Idealized:** Based on coefficients of orthogonal components of the $\hat\beta_k$, with $\underline{\lambda}_k$ proportional to

$$\Big((\sigma_Y - b_0^T\hat\beta_k)^2, (b_1^T\hat\beta_k)^2, \ldots, (b_k^T\hat\beta_k)^2\Big)$$

and $c_k = \sigma^2 + \|\beta - \hat\beta_k\|^2$, producing the relationship

$$stat_k \,=\, Z_k^{comb} \,+\, \frac{\sqrt{n}}{\sqrt{c_k}}\beta$$

for which, in each section ℓ, the shift factor is of the desired form α_ℓ with $D_k = \|\beta - \hat{\beta}_k\|^2$. It suffers from dependence of the weights of combination on the unknown β. The $b_{k'}^T \hat{\beta}_k$ depend on $\beta^T \hat{\beta}_{k'}$, for $k' = 1, 2, \ldots, k$.

[C] **Simplified:** As in [B] but with each occurrence of $\beta^T \hat{\beta}_{k'}$ replaced by its known expected value.

The $\beta^T \hat{\beta}_k$ is close to its expected value, indeed, within any specified small positive η, except in an event of probability exponentially small in $L\eta^2$. This is a consequence of Hoeffding's inequality, interpreting $\beta^T \hat{\beta}_k$ as an average of L bounded independent random variables. Likewise, the $\|\hat{\beta}_k - \beta\|^2$ is close to its expectation, permitting the approximation to its distribution as a shifted normal using $D_k = \mathbb{E}\|\hat{\beta}_k - \beta\|^2$ in defining the shift factor $\alpha_{\ell,k}$ as before.

B. Iteratively Optimal Coefficient Estimates

Consider the choice of the updated coefficient estimates $\hat{\beta}_{k+1}$ as a function of $stat_k$. We arrange these to be the Bayes optimal posterior mean of β given $stat_k$, as derived here. Use the approximating distribution that the $stat_{k,j}$ are independent $\text{Normal}\big(\alpha_\ell 1_{\{j=j_\ell\}}, 1\big)$, for j in any section ℓ, where $\alpha_\ell = \alpha_{\ell,k}$. Let $\phi(s)$ be the standard normal density and note that $\phi(s-\mu)/\phi(s)$ is proportional to $e^{\mu s}$. With the term j_ℓ chosen according to a uniform distribution over the M choices in section ℓ, the posterior distribution of j_ℓ is

$$Prob\{j_\ell = j | stat_k\} = w_{k,j} = e^{\alpha_\ell stat_{k,j}} / \sum_{j \in sec_\ell} e^{\alpha_\ell stat_{k,j}}.$$

Restricted to section ℓ, the $\beta_j = \sqrt{P_\ell} 1_{j_\ell = j}$. Accordingly, the posterior mean of β_j provides the Bayes estimator, $\mathbb{E}[\beta_j | stat_k] = \sum_{j_\ell \in sec_\ell} w_{k,j_\ell} \sqrt{P_\ell} 1_{\{j=j_\ell\}}$ which reduces to

$$\hat{\beta}_{k+1,j} = \sqrt{P_\ell}\, w_{k,j}.$$

This is the estimate appropriate to use each step.

At the final step, in each section, the decoded term $\hat{j}_\ell$ may be taken to be the one the highest posterior weight $w_{k,j}$. The posterior probability of success in a section is the posterior weight of the true term w_{k,j_ℓ}.

C. Relating Squared Error and Expected Posterior Success

The quantity $\hat{\beta}_{k+1}^T \beta / P$ can be interpreted as a posterior success rate $\sum_{\ell=1}^{L} (P_\ell / P)\, w_{k,j_\ell}$, with a power-weighted average across the sections.

***Lemma** 2:* The posterior success rate has the same expectation as the squared norm $\|\hat{\beta}_{k+1}\|^2 / P$. Consequently, the posterior error rate $\sum_{\ell=1}^{L} P_\ell\, (1 - w_{k,j_\ell})$ has the same expectation as the squared distance $\|\hat{\beta}_{k+1} - \beta\|^2$.

The proof of Lemma 2 is in the appendix.

D. Update Function and its Analysis

Analysis of the progression of the adaptive successive decoder is considerably simplified if one finds a recursively updated measure of success. The progress may be tracked using either the expected squared distance $D_k = \mathbb{E}\|\hat{\beta}_k - \beta\|^2$ or the expected posterior success rate which we will denote x_k.

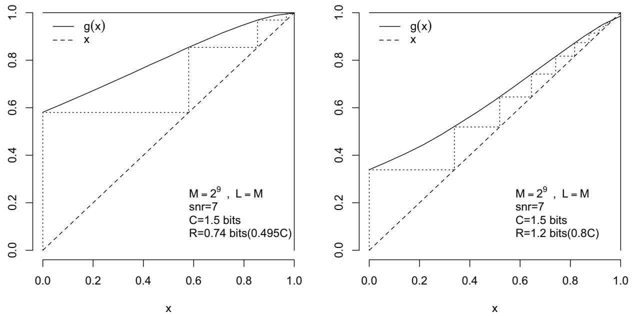

Fig. 1. Plot of g(x) and the sequence x_k. It is computed for a grid of fifteen x values by Monte Carlo simulation with replicate size 500.

The above results show that these two quantities are related by $D_k = (1 - x_k)P$ permitting their recursion as follows.

Consider, for any realization $j_1, j_2, \ldots, j_L$, the next step expected posterior success rate $x_{k+1} = \sum_{\ell=1}^{L} (P_\ell / P)\, \mathbb{E}[w_{k,j_\ell}]$. This expected value is the same no matter which $j_1, j_2, \ldots, j_L$ was chosen, so assume here that the first term in each section was sent. Accordingly, at $\alpha_\ell = \alpha_{\ell,k}$,

$$x_{k+1} = g(x_k) = \sum_{\ell=1}^{L} (P_\ell / P)\, \mathbb{E}\left[\frac{e^{\alpha_\ell^2 + \alpha_\ell Z_1}}{e^{\alpha_\ell^2 + \alpha_\ell Z_1} + \sum_{j=2}^{M} e^{\alpha_\ell Z_j}} \right],$$

where $Z_1, Z_2, \ldots, Z_M$ are independent $\mathrm{N}(0,1)$. What makes this a recursive characterization of progress is that $\alpha_{\ell,k}$ is a function of the preceding x_k via its relationship to the expected squared distance. Indeed, $\alpha_{\ell,k} = \alpha_\ell(x_k)$ where $\alpha_\ell(x)$ is $\sqrt{P_\ell\, n / (\sigma^2 + (1-x)P)}$. Thus our update function is

$$g(x) = \sum_{\ell=1}^{L} (P_\ell / P)\, \mathbb{E}\left[w_1(\alpha_\ell(x)) \right]$$

where $w_1(\alpha) = (e^{\alpha^2 + \alpha Z_1}) / \big[e^{\alpha^2 + \alpha Z_1} + \sum_{j=2}^{M} e^{\alpha Z_j} \big]$. We initiate investigation of this $g(x)$ and compare it to the corresponding update function that arose from the $\{0,1\}$ valued weights of the thresholding method. As in [1], [2], [3], it is given by $g_{\{0,1\}}(x) = \sum_{\ell=1}^{L} (P_\ell / P) \Phi(\alpha_\ell(x) - \tau_a)$, where $\tau_a = \sqrt{2 \log M} + a$ is the threshold. In that scheme $a > 0$ is needed to avoid false alarms.

For the algorithm to update properly, we need $x_{k+1} > x_k$ where $x_{k+1} = g(x_k)$. Thus it is desired to have $g(x)$ stay above x (the 45 degree line). In [1], [2], [3], it is confirmed that, for any fixed rate below the capacity, $g_{\{0,1\}}(x)$ stays above x in an interval $[0, x^*]$, where x^* is near 1, though the gap from 1 was of order $1/\log M$. We evaluate $g(x)$ to study the performance of the soft decision decoder and to compare it with the $\{0,1\}$ valued decoder. Of interest is whether the crossing point x^* is moved substantially closer to 1.

Fig. 1 shows our update function with rates at two different fractions of capacity. Observe that, in both cases, the update function is above x on the most of the interval $[0,1]$. The step function in the gap in Fig. 1 shows progression of the steps. The gap between $g(x)$ and x affects the number of steps to arrive at a success rate near x^*. [Dan Spielman has suggested there is similarity of our use of the function $g(x)$, which is for adaptive successive decoding of sparse superposition codes, with the EXIT charts of [7], used in the study of iterative decoders of turbo codes.]

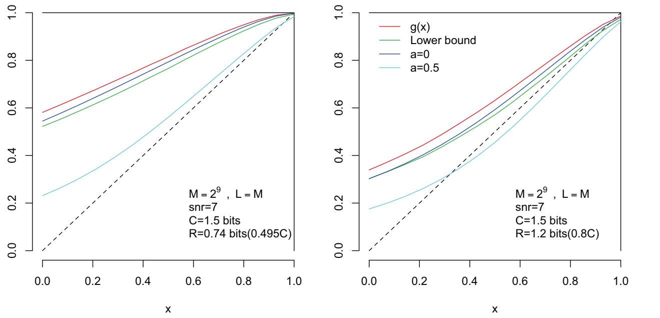

Fig. 2. Comparison of update functions. Red line refers to the $g(x)$ which is calculated by Monte Carlo simulation and the green line refers to a theoretical lower bound of $g(x)$. Blue and light blue lines indicates $\{0,1\}$ decision using the threshold $\sqrt{2\log M}+a$ with respect to the value a as indicated.

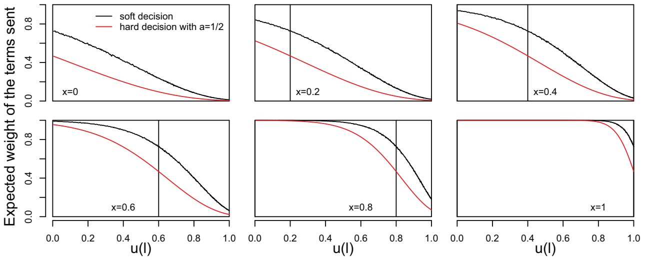

Fig. 3. Transition Plots: $M=2^9$, L=M, C=1.5 bits, R=0.8C and a=0.5. We used Monte Carlo simulation with replicate size 10000. The horizontal axis depicts $u(\ell)=(1-e^{-2C\ell/L})/(1-e^{-2C})$ which is an increasing function of ℓ. The vertical axis gives $g_\ell(x)$. This representation allows the area under the curve to represent $g(x)$. Also, the area of the rectangle to the left of the vertical bar is x. One can see if $g(x)$ is above x by comparing the two areas.

A simplified lower-bound on $g(x)$ is obtained via Jensen's inequality by replacing the $\sum_{j=2}^{M} e^{\alpha Z_j}$ in the denominator above by its expectation $(M-1)e^{\alpha^2/2}$.

Fig. 2 evaluates different update functions. The highest is $g(x)$ of the new procedure. It is much higher than $g_{\{0,1\}}(x)$ at realistic thresholds (e.g. $a=1/2$) and yet still higher than $g_{\{0,1\}}(x)$ with the unrealistic idealized threshold $a=0$. In the realistic case ($a=1/2$) the $g_{\{0,1\}}(x)$ fails to allow rates at 80% of capacity (for $M=2^9$ and $snr=7$) because its curve drops below x at a value far from 1. In [1], [2] good performance at reasonable rates required a much larger section size, such as $M=2^{16}$. In contrast, the new decoder is successful at 80% of capacity with the smaller section size.

The value $\delta=1-x^*$ bounds the likely fraction of mistakes of the final step of the decoder. It controls the closeness to 1 of the rate of an outer Reed-Solomon code that corrects the remain errors (as described in [1], [2]). Our goal in further research is to establish whether the order of the error $1-x^*$ is superior to the order $1/\log M$ established in [2], [3].

Fig. 3 considers the success rate $g_\ell(x)=\mathbb{E}[w_{j_\ell}(\alpha_\ell(x))]$ as a function of the section index ℓ. It shows an increasing wave of closeness to 1 as x increases.

After a suitable number of steps, the decoder will succeed if the weights w_{k,j_ℓ} are large enough. It is recommended on the final step to decode the sections for which the maximal $w_{k,j}$ is at least 1/2. In contrast, when $\max_{j\in sec_\ell} w_{k,j}<1/2$, the posterior probability of error is more likely than not. In that case it is recommended to leave the section undecoded as an erasure to be corrected by the outer R.S code.

Appendix

Proof of Lemma 1: Consider the representation of the collection of vectors X_j, for $1\le j\le N$, augmented by one additional vector $X_{N+1}=\varepsilon/\sigma$. The $\mathcal{Z}_{k',j}=X_j^T G_{k'}/\|G_{k'}\|$ for $k'<k$ are the coefficients of the representation of X_j in the span of the orthonormal $G_0/\|G_0\|,\ldots,G_{k-1}/\|G_{k-1}\|$, with an orthogonal residual vector $V_{k,j}$, for j in $J_e=\{1,\ldots,N,N+1\}$. Collecting these into a matrix decomposition, it takes the form

$$X=\frac{G_0}{\|G_0\|}\mathcal{Z}_0^T+\frac{G_1}{\|G_1\|}\mathcal{Z}_1^T+\ldots+\frac{G_{k-1}}{\|G_{k-1}\|}\mathcal{Z}_{k-1}^T+V_k,$$

where the vectors $\mathcal{Z}_{k'}=(\mathcal{Z}_{k',j}:j\in J)$ extend to $\mathcal{Z}_{k',e}=(\mathcal{Z}_{k',j}:j\in J_e)$ when representing X_e.

Using these $G_0,G_1,\ldots,G_{k-1}$ and the columns of the identity, Gram-Schmidt fills out a basis of R^n with n orthornormal vectors $\xi_{k,0},\xi_{k,1},\ldots,\xi_{k,n-1}$, in which the residuals $V_{k,j}$ have representation $\sum_{i=k}^{n-1}V_{k,j,i}\xi_{k,i}$, using the last $n-k$ of these orthonormal vectors, with $V_{k,j,i}=V_{k,j}^T\xi_{k,i}$.

With the columns of X_e assumed to be independent standard normal vectors, we solve for the evolution of the conditional distributions of the $\mathcal{Z}_{k,e}$ and $\|G_k\|$, using the above representation. The conditional distribution of the $\mathcal{Z}_{k,e}$ and $\|G_k\|$ given $\mathcal{F}_{k-1,e}=(\mathcal{Z}_{0,e},\|G_0\|,\ldots,\mathcal{Z}_{k-1,e},\|G_{k-1}\|)$ has $\mathcal{X}_{n-k}^2=\|G_k\|^2/\sigma_k^2$ distributed chi-square$(n-k)$ and $\mathcal{Z}_{k,e}=b_{k,e}\mathcal{X}_{n-k}+Z_{k,e}$ with $Z_{k,e}$ distributed N$(0,\Sigma_{k,e})$. The conclusion of the lemma then follows from noting for the $\mathcal{Z}_k$ that the conditional distribution given $\mathcal{F}_{k-1,e}$ only depends on $\mathcal{F}_{k-1}$, under the assumption that successively the estimates $\hat{\beta}_k$ are computed only from the information $\mathcal{F}_{k-1}$ available to the decoder (without knowledge of the noise).

Moreover, it is claimed that conditionally given $\mathcal{F}_{k-1,e}$, the coordinates $V_{k,j,i}$ of the vectors $V_{k,j}$ in the basis $\xi_{k,i}$, for $i=k,k+1,\ldots,n-1$, are conditionally mean-zero Normal random variables, independent across i, and jointly across $j\in J_e$, having covariance $\Sigma_{k-1,e}$ [where for $k=0$ the $\Sigma_{k-1,e}$ is replaced by the identity matrix].

The number of columns is arbitrary. Henceforth in the proof there is no need to make a distinction between the cases with and without the extension, so drop the subscript e.

Prove this claim inductively on $k\ge 0$. Initially, $V_{0,j}=X_j$ and the normality of the X_j provides for the validity of the distributional claim for $V_{k,j}$ for $k=0$. For the induction, assume the claim to be true at step k and derive from it that it is true at the next step $k+1$. Along the way, the conditional distribution properties of the $\|G_k\|$ and $\mathcal{Z}_k$ in the lemma are established as consequences.

Concerning G_k, note $\|G_0\|^2/\sigma_0^2$ is $\mathcal{X}_n^2$ distributed. For $k\ge 1$, the G_k as the part of $X\hat{\beta}_k$ orthogonal to the previous parts $G_0,\ldots,G_{k-1}$ is equal to $G_k=V_k\hat{\beta}_k=\sum_j\hat{\beta}_{k,j}V_{k,j}$ since V_k is the part of X with columns orthogonal to the previous parts. Representing G_k in the basis $\xi_{k,0},\ldots,\xi_{k,n-1}$ it has coordinates $G_{k,i}$ equal to 0 for $0\le i\le k-1$ and equal to $\sum_j V_{k,j,i}\hat{\beta}_{k,j}$ for $k\le i\le n-1$. From the induction hypothesis, these $(V_{k,j,i}:j\in J)$ have conditional distribution Normal$(0,\Sigma_{k-1})$. Accordingly, these $G_{k,i}$ are independent

Normal$(0,\sigma_k^2)$ where $\sigma_k^2 = \hat{\beta}_k^T \Sigma_{k-1} \hat{\beta}_k$, from which it follows that $\|G_k\|^2/\sigma_k^2$ is $\mathcal{X}_{n-k}^2$ distributed, independent of $\mathcal{F}_{k-1}$.

Next, for each j, seek $b_{k,j}$ as a regression coefficient based on the joint distribution of the $V_{k,j}$ and G_k (given $\mathcal{F}_{k-1}$) to obtain the representation of the vectors

$$V_{k,j} = b_{k,j}\frac{G_k}{\sigma_k} + U_{k,j}.$$

This is done in the basis $\xi_{k,k}, \ldots, \xi_{k,n-1}$ where the coordinates $V_{k,j,i}$ and $G_{k,i}$ are jointly normal (where across $i = k, \ldots, n\text{–}1$ they are independent and identically distributed, conditionally given $\mathcal{F}_{k-1}$, so they share the same regression coefficient $b_{k,j}$). The coordinates of $U_{k,j,i}$ are conditionally normal random variables, independent of the $G_{k,i}$, and independent for $k \le i \le n-1$. For $k = 0$ the coefficient $b_{k,j} = E[V_{k,j,i}G_{k,i}/\sigma_k]$ simplifies to $E[X_{j,i}Y_i/\sigma_Y] = \beta_j/\sigma_Y$.

For $k \ge 1$ the $b_{k,j} = E[V_{k,j,i}G_{k,i}/\sigma_k]$ may be expressed as $E[V_{k,j,i}\sum_{j'} V_{k,j',i}\hat{\beta}_{j'}]$ where the expectation is with respect to the Normal$(0,\Sigma_{k-1})$ distribution for the $(V_{k,j,i} : j \in J)$. Accordingly, summarize the solution for these coefficients as the vector $b_k = \Sigma_{k-1}\hat{\beta}_k/\sigma_k$.

As for the parameters of the distribution of the $(U_{k,j,i} : j \in J)$, use the identity $U_{k,j,i} = V_{k,j,i} - b_{k,j}G_{k,i}/\sigma_k$ and the conditional distribution of the V and G coordinates to conclude that it has mean 0 and conditional variance $\Sigma_{k-1} - b_k b_k^T$, in agreement with Σ_k.

Note that $\mathcal{Z}_{k,j} = X_j^T G_k/\|G_k\|$ reduces to $V_{k,j}^T G_k/\|G_k\|$, which by the above representation of $V_{k,j}$ takes the form

$$\mathcal{Z}_{k,j} = b_{k,j}\frac{\|G_k\|}{\sigma_k} + \frac{U_{k,j}^T G_k}{\|G_k\|}.$$

The latter term is what we call $Z_{k,j}$. The inner product is preserved by switching to the basis $\xi_{k,0}, \ldots, \xi_{k,n-1}$. Thus $Z_{k,j} = \sum_{i=0}^{n-1} \alpha_i U_{k,j,i}$, with $\alpha_i = G_{k,i}/\|G_k\|$, which is 0 for $0 \le i \le k-1$. The sum of squares of the α_i is equal to 1. Proceed conditionally on $\mathcal{F}_{k-1}$. For any fixed α with sum of squares equal to 1, the $\sum_{i=k}^{n-1} \alpha_i U_{k,j,i}$ shares the $\mathrm{N}(0,\Sigma_k)$ distribution, as a result of the independence across i. Accordingly, with $\alpha_i = G_{k,i}/\|G_k\|$, the conditional distribution of Z_k given G_k is as indicated, and it does not depend on G_k, so the Z_k and G_k are independent given $\mathcal{F}_{k-1}$.

Use G_k to update the orthonormal basis of $\mathbb{R}^n$ by Gram-Schmidt, replacing $\xi_{k,k}, \xi_{k,k+1}, \ldots, \xi_{k,n-1}$ with $G_k/\|G_k\|, \xi_{k+1,k+1}, \ldots, \xi_{k+1,n-1}$.

The coefficients of $U_{k,j}$ in this updated basis are $U_{k,j}^T G_k/\|G_k\|$, $U_{k,j}^T \xi_{k+1,k+1}, \ldots, U_{k,j}^T \xi_{k+1,n-1}$, which are denoted $U_{k+1,j,k} = Z_{k,j}$ and $U_{k+1,j,k+1}, \ldots, U_{k+1,j,n-1}$, respectively. Recalling the conditional distribution of the $U_{k,j}$, these coefficients $(U_{k+1,j,i} : k \le i \le n-1, j \in J)$ are also normally distributed, conditional on $\mathcal{F}_{k-1}$ and G_k, independent across i from k to $n-1$; moreover, for each i from k to $n-1$, the $(U_{k+1,j,i} : j \in J)$ inherit a joint $N(0,\Sigma_k)$ conditional distribution from the conditional distribution that the $(U_{k,j,i} : j \in J)$ have.

Specializing the conclusion, separating off the $i = k$ case where the $U_{k+1,j,i}$ is $Z_{k,j}$, the remaining $(U_{k+1,j,i} : k+1 \le i \le n,\ j \in J)$ have the specified conditional distribution and are conditionally independent of G_k and Z_k given $\mathcal{F}_{k-1}$. It follows that the conditional distribution of $(U_{k+1,j,i} : k+1 \le i \le n-1,\ j \in J)$ given $\mathcal{F}_k = (\mathcal{F}_{k-1}, \|G_k\|, Z_k)$ is identified.

Likewise, the vector $V_{k,j} = b_{k,j}\, G_k/\sigma_k + U_{k,j}$ has representation in this updated basis with coefficient $\mathcal{Z}_{k,j}$ in place of $Z_{k,j}$ and with $V_{k+1,j,i} = U_{k+1,j,i}$ for i from $k{+}1$ to $n-1$. So these coefficients $(V_{k+1,j,i} : j \in J)$ have the normal $N(0,\Sigma_k)$ distribution for each i, independently across i from $k+1$ to n, conditionally given $\mathcal{F}_k$. Thus the induction is established, which completes the proof of Lemma 1.

Proof of Lemma 2: The random variables in question are sums across the sections. We show equality of the expectations in each section. Fix a section ℓ and a step k and let $stat = (stat_{k,j} : j \in sec_\ell)$ be the relevant part of the statistics, with index set sec_ℓ regarded as $\{1, 2, \ldots, M\}$.

The random variables in question have the same expected value no matter which terms j_ℓ was sent. Accordingly, the expectation taken conditionally on any particular realization j_ℓ match what is obtained if alternatively one averages with respect to the uniform prior on j_ℓ. Let $P_j = P_{stat|j_\ell = j}$ be the conditional distributions and $P = (1/M)\sum_{j=1}^M P_{stat|j_\ell = j}$ be the marginal distribution of $stat$ in section ℓ, and likewise let $\mathbb{E}_j$ and $\mathbb{E}$, respectively, denote corresponding expectations of functions of $stat$. Now $w_j = w_{k,j}$ is the posterior probability $P[j_\ell = j|stat]$. Show that w_{j_ℓ} and $\|w\|^2 = \sum_{j=1}^M w_j^2$ have the same expectation.

The P_j and P have likelihood ratio Mw_j. Set $j = 1$. Calculate the expectation $\mathbb{E}_1[w_1]$ using the measure P rather than P_1 by incorporating the factor Mw_1. Thus $\mathbb{E}_1[w_1] = M\mathbb{E}[w_1^2]$. By symmetry, $\mathbb{E}[w_j^2]$ is same across all j and so $M\mathbb{E}[w_1^2]$ equals $\mathbb{E}[\sum_{j=1}^M w_j^2] = \mathbb{E}[\|w\|^2]$ which is $(1/M)\sum_{j=1}^M \mathbb{E}_j[\|w\|^2]$. Each term in this sum is the same so it is $\mathbb{E}_1[\|w\|^2]$ as claimed. This completes the proof of Lemma 2.

Space does not permit full listing and discussion of the relationship of sparse superposition codes to past work in information theory, compressive sensing, and coding. For such the reader is invited to see the discussion and reference lists in [1], [2], [3], [6].

References

[1] A.R. Barron and A. Joseph, "Toward fast reliable communication at rates near capacity with Gaussian noise, *Proc. IEEE Intern. Symp. Inform. Theory*, Austin, TX, June 13-18, 2010.

[2] A.R. Barron and A. Joseph, "Sparse superposition codes: Fast and reliable at rates approaching capacity with Gaussian noise," March 2011. Available from www.stat.yale.edu/∼arb4/publications.html

[3] A.R. Barron and A. Joseph "Analysis of fast sparse superposition codes," *Proc. IEEE Intern. Symp. Inform. Theory*, St. Petersburg, Russia, 2011.

[4] M. Bayati and A. Montanari, "The dynamics of message passing on dense graphs, with applications to compressed sensing," *IEEE Trans. Inform. Theory*, vol.57, no.2, pp.764–785, Feb. 2011.

[5] M. Bayati and A. Montanari, "The LASSO risk for gaussian matrices," *IEEE Trans. Inform. Theory*, vol.58, no.4, pp.1997-2017, April 2012.

[6] A. Joseph and A.R. Barron, "Least squares superposition codes of moderate dictionary size are reliable at rates up to capacity," *IEEE Trans. Inform. Theory*, vol.58, no.5, pp.2541-2557, May 2012.

[7] S. ten Brink, "Convergence behavior of iteratively decoded parallel concatenated codes," *IEEE Trans. Commun.*, vol.49, no.10, pp.1727-1737, Oct 2001.