APPROXIMATE ITERATIVE BAYES OPTIMAL ESTIMATES FOR HIGH-RATE SPARSE SUPERPOSITION CODES

Sanghee Cho and Andrew Barron

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

ABSTRACT

This paper is concerned with sparse superposition codes with iterative term selection for additive white Gaussian noise channel with power control. In particular, we consider a soft decision decoder with Bayes optimal estimates at each step, presuming uniform prior on the choice of the terms that are sent. Bayes optimal estimates are formulated and shown to have a Martingale property that provides alternative representations of a posterior probability of error. Since the Bayes optimal estimates are infeasible, an approximation method is suggested. We analyze the performance of the approximation method in comparison with the infeasible estimates.

1. INTRODUCTION

Superposition codes use specific linear combinations of a given set of vectors. With a dictionary X consisting of vectors $X_1, X_2, \ldots, X_N$, each of n coordinates, the codeword vectors are superpositions $\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_N X_N$. The vectors X_j provide components of the codewords with coefficients β_j. By design, each entry of these vectors X_j is independent standard normal. The choice of codeword is conveyed through the coefficients by the choice of L non-zero coefficients of β. For a channel with additive white Gaussian noise(AWGN) with superposition coding, $Y = X\beta + \epsilon$ is received. It is a vector of length n and ϵ is the noise vector which is normal random variable $N(0, \sigma^2 I)$.

By design, we split the coefficient vector into L sections each size of $M = N/L$. Only one coefficient in each section is nonzero so that we have M^L choice of codewords. With M power of two, an input bit string u of length $K = L \log_2 M$ splits into L substrings of size $\log_2 M$ which index the term chosen to be non-zero. Denote the terms sent as $\{j_1, \ldots, j_L\}$ which takes values $\beta_{j_\ell} = \sqrt{P_\ell}$, with $\sum_\ell P_\ell = P$. The total power is controlled by P.

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

These sparse superposition codes with an adaptive successive decoder for the Gaussian noise channel are computationally fast with any fixed rate below capacity with exponentially small error probability. See [1, 2, 3] for this conclusion and relationship to other literature.

The adaptive successive decoder uses iteratively obtained test statistics related to inner products of the X_j with the part of previous fits or the residuals. For each step, the decoder accepts terms for which the statistics above the threshold. The threshold is chosen to be high enough to avoid false alarms.

The previous conference paper [8] is motivated by the same type of the test statistics and improve performance using a soft decision decoder with Bayes optimal statistics. The conditional distribution of the statistics is approximately normal random variable, shifted for the true term. The amount of the shift increases as we successively decode. The soft decision decoder is based on posterior probability of the term j sent with uniform prior on the choice of the term sent in each section. Also, that paper introduces an update function $g(x)$ which gives the expected success rate on a step if the success rate of the previous step is x.

This paper continues to study the soft decision decoder for AGWN. In the following section, we review background motivation and findings from [8], the formulation of the statistics, their distribution, Bayes optimal estimates and numerical simulation of the performance improvement. In the next section, we introduce a way to construct a Bayes optimal estimates with idealized ingredients which is simplified version of our statistics. Finally, we construct the approximate Bayes optimal estimates. One way is to combine our test statistics with the same weights of combination as in the third section. Another way is to combine our statistics to have approximately desired form, shifted standard normal. However, the weights of combinations of such can't be calculated not knowing which terms are sent. So we estimate the weights of combinations and try to find a better error bound than the first way of combination.

If there were no computational restriction, the optimal decoder would be based on the conditional distribution of β given X and Y with mean $\hat{\beta}_{opt} = \mathbb{E}[\beta|X, Y]$, using a uniform prior on β. Practical iterative decoders produce a sequence of statistics, $stat_k$ using inner products of the columns of X with components of the residuals of previous fits. Correspondingly, optimal coefficient estimate of β would be $\hat{\beta}_{k,opt} = \mathbb{E}[\beta|stat_1, \ldots, stat_{k-1}]$. In this paper, armed with iterative distributional properties of ingredients of decoding statistics, we explore the formation

of approximation to such Bayes optimal estimates for decoding.

2. FRAMEWORK FOR ITERATIVE DECODER, BAYES OPTIMAL ESTIMATES

The decoder develops a sequence of estimates $\hat{\beta}_k$ of the true coefficient vector β. In [8], we introduced a framework for a iterative decoder, the distribution of the corresponding statistics. Let $\hat{\beta}_k$ be any sequence of estimates. For the initial step, $G_0 = Y$. For $k \geq 1$, let $F_k = X\hat{\beta}_k$ and let G_k be a part of F_k which is orthogonal to $G_0, G_1, \ldots, G_{k-1}$. We assume new fits are not in a linear span of previous fits so that we have $\|G_k\| > 0$. Let $\mathcal{Z}_{k,j} = X_j^T G_k/\|G_k\|$ for $j = 1, \ldots, N$. These are essential ingredients in the statistics that we form. For analysis purposes, let $\mathcal{Z}_{k,N+1} = (\epsilon/\sigma)^T G_k/\|G_k\|$ and we define $\beta_e, \hat{\beta}_{1,e}, \ldots, \hat{\beta}_{k,e}$ as extended vectors in R^{N+1} appending one more coordinate to the vectors. The σ is appended for β_e and zero for others. Let $b_{0,e}, b_{1,e}, \ldots, b_{k,e}$ be a result of successive Gram-Schmidt orthogonalization of $\beta_e, \hat{\beta}_{1,e}, \ldots, \hat{\beta}_{k,e}$. 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 $\beta_e, \hat{\beta}_{1,e}, \ldots, \hat{\beta}_{k,e}$. Let Σ_k denote the upper left $N \times N$ portion of this matrix. Then [8] states a lemma as following.

Lemma 1. *For $k \geq 0$, the conditional distribution $\mathbb{P}_{\mathcal{Z}_k|\mathcal{F}_{k-1}}$ of $\mathcal{Z}_k$ given $\mathcal{F}_{k-1} = (\mathcal{Z}_0, \|G_0\|, \ldots, \mathcal{Z}_{k-1}, \|G_{k-1}\|)$ is determined by the representation*

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

where $Z_k^{red} = (Z_{k,j}^{red} : j \in J)$ 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, $\mathcal{X}_{n-k}^2 = \|G_k\|^2/\sigma_k^2$ is distributed as a Chi-square$(n-k)$ random variable independent of the Z_k and the past $\mathcal{F}_{k-1}$.

In this representation $\mathcal{Z}_k = \mathcal{X}_{n-k}b_k + Z_k^{red}$, the b_k is the most important part for the properties that we seek, while Z_k^{red} (red stands for reduced) has no component of vector β_e nor in the direction of the fits.

Related to the distribution $\mathbb{P}_{Z_k^{red}|\mathcal{F}_{k-1}}$ is the distribution $Q_{Z_k^{red}|\mathcal{F}_{k-1}}$ which makes the Z_k^{red} be Normal$(0, I - Proj_k)$ where $Proj_k$ is projection matrix onto the linear space orthogonal to the estimates, $\hat{\beta}_1, \ldots, \hat{\beta}_k$. This $Proj_k$ does not provide orthogonality to β. 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 $\sqrt{1+snr}$. We can add $Proj_k \tilde{Z}_k$ where $\tilde{Z}_k$ are auxiliary independent standard normal vectors provided to the sample space for P and Q to $\mathcal{Z}_k$. Then, with respect to Q, given $\mathcal{F}_{k-1}$, the $\mathcal{Z}_k^{clean} = \mathcal{Z}_k + Proj_k \tilde{Z}_k$ have a representation $\mathcal{X}_{n-k}b_k + Z_k$ with the Z_k distributed Normal$(0, I)$. For abbreviation, we will denote $\mathcal{Z}_k^{clean}$ as $\mathcal{Z}_k$ and later when needed denote original $\mathcal{Z}_k$ as $\mathcal{Z}_k^{precleaning}$. It is because we will combine $\mathcal{Z}_k^{clean}$ in the decoder for analysis purposes.

As long as the number of step k is small compare to n, the Chi distribution $\mathcal{X}_{n-k}^2/n$ is close to constant one except in events of exponentially small probability. Thus, our $\mathcal{Z}_k^{clean}$ have a representation approximately $\sqrt{n}b_k + Z_k$. Also, certain combination of these $\mathcal{Z}_k^{clean}$ makes nearly constant shifted normal. The paper [8] provides several motivations of combinations of these components to produce our statistics $stat_k$. The $stat_k$ take the following form, for some choice of vector $\underline{\lambda}_k = (\lambda_{0,k}, \lambda_{1,k}, \ldots, \lambda_{k,k})$ with unit square norm and some c_k typically between σ^2 and $\sigma^2 + P$,

$$stat_k = \mathcal{Z}_k^{comb} + \frac{\sqrt{n}}{\sqrt{c_k}}\hat{\beta}_k \tag{1}$$

where $\mathcal{Z}_k^{comb} = \lambda_{0,k}\mathcal{Z}_0 + \lambda_{1,k}\mathcal{Z}_1 + \ldots + \lambda_{k,k}\mathcal{Z}_k$. The combination should be such that these statistics have the representation

$$Z_k^{comb} + \frac{\sqrt{n}}{\sqrt{c_k}}\beta \tag{2}$$

with the desired shift $\frac{\sqrt{n}}{\sqrt{c_k}}\beta$. The two representations (1) and (2) look similar. The equation (1) provides the definition of $stat_k$ in terms of quantities computed by the decoder $\mathcal{Z}_0, \mathcal{Z}_1, \ldots, \mathcal{Z}_k$ and $\hat{\beta}_k$. The second representation (2) is a distributional characterization in terms of the unknown β which is essential to our analysis. This representation only holds for certain choices of $(\lambda_{0,k}, \lambda_{1,k}, \ldots, \lambda_{k,k})$ and $\hat{\beta}_k$. In some cases, this distributional form only holds approximately.

One of the ideal choices of $\underline{\lambda}_k$ is based on coefficients of orthogonal components of the $\hat{\beta}_k$, with $\underline{\lambda}_k$ proportional to

$$\Big((\sigma_Y - b_0^T\hat{\beta}_k), (-b_1^T\hat{\beta}_k), \ldots, (-b_k^T\hat{\beta}_k)\Big). \tag{3}$$

We denote this $\underline{\lambda}_k$ as $\hat{\underline{\lambda}}_k^*$. Set $\hat{c}_k^* = \sigma^2 + \|\beta - \hat{\beta}_k\|^2$ and

$$\hat{stat}_k^* = \sum_{k'=0}^{k} \hat{\lambda}_{k',k}^* \mathcal{Z}_k + \frac{\sqrt{n}}{\sqrt{\hat{c}_k^*}}\hat{\beta}_k. \tag{4}$$

Here, if we approximate $\mathcal{Z}_{k'} = \mathcal{X}_{n-k'}b_{k'} + Z_{k'}$ by $\sqrt{n}\ b_{k'} + Z_{k'}$, it produces the relationship

$$\hat{stat}_k^* \approx Z_k^{comb} + \frac{\sqrt{n}}{\sqrt{\hat{c}_k^*}}\beta.$$

It has the desired representation, but since we do not know β in advance, we cannot use it for our statistics. Instead, we can replace $b_{k'}^T\hat{\beta}_k$ with its estimates. In this paper, we construct reasonable estimates for the weights of combination $\hat{\underline{\lambda}}_k^*$ and show that the statistics based on those estimates are not far from the approximate form.

The approximating distribution of the $stat_{k,j}$ is independent Normal$\big(\alpha_\ell 1_{\{j=j_\ell\}}, 1\big)$, for j in any section ℓ, where $\alpha_\ell = \alpha_\ell(x_k) = \sqrt{nP_\ell/c_k}$ with $c_k = \sigma^2 + P(1 - x_k)$. Supposing this distribution, the paper [8] arranged the updated coefficient estimates $\hat{\beta}_{k+1}$ to be posterior mean

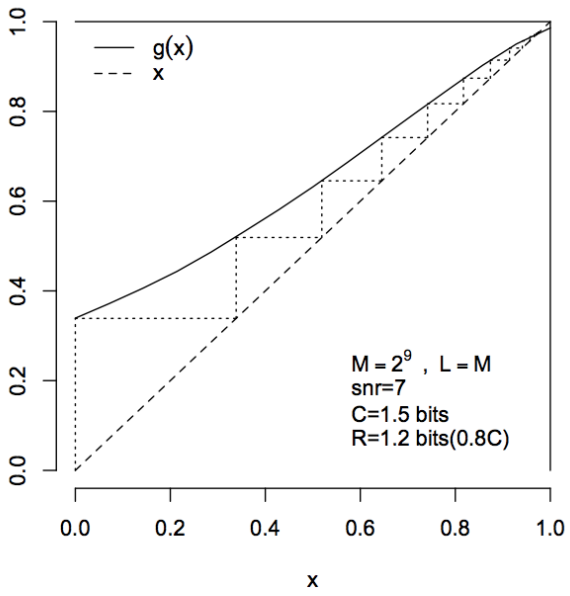

Figure 1. Plot of g(x) and the sequence x_k. It is computed for a grid of fifteen x values by Monte Carlo simulation with replicate size 500.

of β given $stat_k$. With the terms sent are chosen uniformly over M choices in section ℓ, the estimate appropriate to use each step would be, for $j \in sec_\ell$,

$$\hat{\beta}_{k+1,j} = \sqrt{P_\ell}\, w_{k,j} = \sqrt{P_\ell}\, \frac{e^{\alpha_\ell stat_{k,j}}}{\sum_{j' \in sec_\ell} e^{\alpha_\ell stat_{k,j'}}}.$$

Suppose the expected success rate at step k is x_k. Then we define an idealized estimate for β as following. For $j \in sec_\ell$, jth component of the estimate β^*_{k+1} is

$$\beta^*_{k+1,j} = \sqrt{P_\ell}\; w^*_{k,j} = \sqrt{P_\ell} \frac{e^{\alpha_\ell(x_k)(Z_j + \frac{\sqrt{n}}{\sqrt{c_k}}\beta_j)}}{\sum_{j' \in sec_\ell} e^{\alpha_\ell(x_k)(Z_{j'} + \frac{\sqrt{n}}{\sqrt{c_k}}\beta_{j'})}}.$$

Now consider the inner product $\beta^T \beta^*_{k+1}$ for any particular realization $j_1, j_2, \ldots, j_L$. This inner product, takes the form $\sum_{\ell=1}^{L} P_\ell\, w_{k,j_\ell}$. Dividing by P it is the power weighted average of the weights given to the true terms, and it thus interpreted as a success rate. The idealized update function $g^{ideal}(x)$ is defined by

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

where $\alpha_\ell(x) = \sqrt{\frac{nP_\ell}{\sigma^2 + P(1-x)}}$. It is expected success rate given that the previous success rate is x. From this update function, we can study the progression of the idealized estimate.

As in Fig.1 is a realization of the update function for certain parameters indicated in the figure. The dotted steps shows the progression in expectation of the idealized estimates. As long as the update function stays above the 45 degree line, we can update our estimates.

It is believed that these statistics make optimal combination of the ingredients $\mathcal{Z}_k$ that maximizes the shift in each section. A related property was shown in [2] for hard decision estimates. We explore in the next section whether we are making approximately optimal use of the sequence of $stat_1, \ldots, stat_k$.

3. IDEALIZED ESTIMATES

Here, we introduce one way to formulate Bayes optimal estimates. Suppose we have any given sequence of estimates $\hat{\beta}_1, \ldots, \hat{\beta}_k$ and any decreasing sequence $\sigma^2 + P = c_0 > c_1 > \cdots > c_k > \sigma^2$. We construct an idealized sequence of pseudo-statistics $stat^*_k$ with the distribution of $Normal(\sqrt{n/c_k}\beta, I)$. These idealized statistics arise from combining idealized ingredients $\mathcal{Z}^*_{k'}$ for $k' = 0, 1, \ldots, k$. These ingredients are defined by the representation $\mathcal{Z}^*_{k'} = \sqrt{n} b^*_{k'} + Z_{k'}$ with the $Z_{k'}$ independent standard normal vectors. The $b^*_{k'}$ is defined as

$$b^*_{k'} = \frac{\hat{\beta}_{k'} - \hat{\beta}_{k'-1} - \lambda^2_{k',k'}(\beta - \hat{\beta}_{k'-1})}{\lambda_{k',k'}\sqrt{c_{k'}}}$$

with

$$\underline{\lambda}^*_k = \sqrt{c_k}\left(\sqrt{\frac{1}{c_0}}, -\sqrt{\frac{1}{c_1} - \frac{1}{c_0}}, \ldots, -\sqrt{\frac{1}{c_k} - \frac{1}{c_{k-1}}} \right).$$

The $b^*_{k'}$ is intended as a simplification that we shall relate to $b_{k'}$. Recall that b_k is, for $k \geq 1$, an normalized part of the estimate $\hat{\beta}_k$ orthogonal to the previous estimates and to β. Likewise, the numerator of b^*_k is the part of $\hat{\beta}_k$ that remains after subtracting a linear combination of $\hat{\beta}_{k-1}$ and β in the extended coordinates. It arises in the form above because $\hat{\beta}_{k-1}$ can be interpreted as, approximately, a projection of both $\hat{\beta}_k$ and β onto the span of $\hat{\beta}_{k-1}, \ldots, \hat{\beta}_1$.

Because β is unknown, $b^*_{k'}$ like $b_{k'}$ is not known from the received data. Likewise, these ingredients $\mathcal{Z}^*_k$ are not actual statistics from the received data but rather they are approximation to $\mathcal{Z}_{k'}$. From these ingredients, define

$$stat^*_k = \sum_{k'=0}^{k} \lambda^*_{k',k} \mathcal{Z}^*_{k'} + \frac{\sqrt{n}}{\sqrt{c_k}} \hat{\beta}_k.$$

The next lemma shows $stat^*_k$ has the property that $\beta^*_{k+1} = \mathbb{E}[\beta | stat^*_k] = \mathbb{E}[\beta | \mathcal{F}_k]$ where $\mathcal{F}^*_k = (\mathcal{Z}^*_0, \ldots, \mathcal{Z}^*_k)$. Thus, if one had access to the approximate ingredients $\mathcal{Z}^*_0, \ldots, \mathcal{Z}^*_k$, then $stat^*_k$ would be Bayes optimal statistics and β^*_{k+1} would be corresponding Bayes optimal estimates for β given these ingredients.

We use a uniform prior on beta. This prior corresponds to a uniform choice of j_ℓ in each section, independently across sections.

Lemma 2 (Optimal Statistics)**.** *For each step k where $k = 0, 1, \ldots, k^*$, the posterior distribution of β given $\mathcal{F}^*_k$ is independent across the sections with posterior probability that $j_\ell = j$ for $j \in sec_\ell$ equal to $w^*_{k+1,j}$, which is a function only of $(stat^*_{k,j} : j \in sec_\ell)$. The $\beta^*_{k+1} = \mathbb{E}[\beta | stat^*_k] = \mathbb{E}[\beta | \mathcal{F}^*_k]$ is the associated conditional mean of β given $\mathcal{F}^*_k$.*

We can prove the above lemma by examining the joint density $p(\mathcal{Z}_0^*, \mathcal{Z}_1^*, \ldots, \mathcal{Z}_k^* | \beta)$ and identify $stat_k^*$ as a sufficient statistic. In particular, the joint density is proportional to

$$\exp\{\sqrt{n/c_k}\,\beta^T \mathcal{Z}_k^{comb,*} + \frac{n}{c_k}(\beta_{k'}^*)^T\beta\}$$

which is

$$\exp\{\sqrt{n/c_k}\,\beta^T stat_k^*\}$$

representable as a product of factors, one for each section. Recall that β assigns one non-zero term $\beta_j = \sqrt{P_\ell}\ 1_{\{j=j_\ell\}}$ in each section ℓ. Accordingly, with the prior for β providing for independence between the sections, we see that the posterior distribution of β is independent across the sections with $\mathbb{P}[j_\ell = j | \mathcal{F}_k^*]$ reducing to $\mathbb{P}[j_\ell = j | stat_k^*] = w_{k+1,j}^*$ for j in section ℓ. Accordingly, $\mathbb{E}[\beta|\mathcal{F}_k^*]$ is equal to $\mathbb{E}[\beta|stat_k^*]$ which is β_{k+1}^* with coordinates $\beta_{k+1,j}^* = \sqrt{P_\ell}\, w_{k+1,j}^*$ for each j in section ℓ for each ℓ. This completes the proof of Lemma 2.

In the Bayes formulation, in which the expectations are with respect to the joint distribution of β and the statistics, one sees, using iterated expectation, that $\mathbb{E}\left[\beta^T\beta_k^*\right]$ and $\mathbb{E}\left[(\beta_{k+k'}^*)^T\beta_k^*\right]$ with $k' > 1$ are the same as $\mathbb{E}\left[\|\beta_k^*\|^2\right]$. Alternatively, if these expectations are computed conditionally on β, one sees that they are the same for every β. In this way, the Bayes formulation provides alternative representations of x_k if we define $x_k = \mathbb{E}\left[\beta^T\beta_k^*\right]/P$.

4. APPROXIMATE ITERATIVELY OPTIMAL ESTIMATES

4.1. Preliminary

Here, we discuss some properties we need to prove a main theorem. We first state reliability and a concentration property of chi-square random variables. Then, we provide estimates for $b_k^T\hat{\beta}_k$ and evaluate the estimate using Chi-concentration. Finally, we provide upperbound for the distance between two exponential weights using the difference between the exponents.

Lemma 3 (Reliability)**.** *For any β and any $1 \le k < k' \le k^*$, the expectation of $\beta^T\beta_k^*$, $\|\beta_k^*\|^2$ and $\beta_k^{*T}\beta_{k'}^*$ are the same which will be defined by x_kP where P is the power constraint. Also, they are close to their expectation with high probability. If we define the event $A_{\beta,\delta}$ as*

$$A_{\beta,\delta} = \quad \{|\beta^T\beta_k^* - x_kP| < \delta \text{ and } |\|\beta_k^*\|^2 - x_kP| < \delta \text{ and } |\beta_k^{*T}\beta_{k'}^* - x_kP| < \delta\}$$

then for any $\delta > 0$,

$$\mathbb{P}\{A_{\beta,\delta}^c\} \le 6\exp\{-\frac{2}{c^2}L\delta^2\}$$

The three quantities are sum of bounded independent random variables. The sum of squares of the ranges of these random variables is $\sum_{\ell=1}^L P_\ell^2$. Thus by Hoeffding's inequality, the probability that the distance of each quantity and the expectation is greater than δP is not more than

$$2\exp\{-\frac{2}{c^2}L\delta^2\},$$

where $c = L\max(P_\ell/P)$. The union bound would be sum of the tail probability. This completes the proof.

Lemma 4 (Chi-square concentration)**.** *For a Chi-square random variable $\mathcal{X}_d^2$ with $d \ge 1$ degrees of freedom and any $0 < h < 1$, the event $|\mathcal{X}_d^2/d - 1| > h$ has probability bounded by $2e^{-dD_h}$ where $D_h = h - \log(1+h)$ is at least $h^2/4$.*

We can prove the lemma with usual Cramer-Chernoff bound. As a result of the Lemma 4, as log as $nh > k$, it holds that $\mathcal{X}_{n-k'}^2/n$ is h close to 1, for $k' = 0, 1, \ldots, k$ except in an event of probability upper bounded by $2(k+1)e^{-\frac{n-k}{2}D_{h_k}}$, where $h_k = (nh-k)/(n-k)$ matches h to within order k/n. When this concentration holds, note also that $\mathcal{X}_{n-k'}/\sqrt{n}$ is $\tilde{h}$ close to 1 with $\tilde{h} = 1 - \sqrt{1-h}$ less than h. The Chi-square probability bound is exponentially small in n, so we may set a rather small $h = \sqrt{L/n}h^*$. Then the error probability is less than $exp\{-Lh^{*2}/16\}$ as long as $k < nh/\sqrt{2}$.

Corollary 5. *For each step k,the difference between $b_k^T\hat{\beta}_k$ and $(\mathcal{Z}_k^{preclean})^T\hat{\beta}_k$ occurs only from the Chi-square random variable. Here, $\mathcal{Z}_k^{preclean}$ is the original $\mathcal{Z}_k$ in lemma. 1 before the cleaning. Thus, except in an event of probability $k^* exp\{-Lh^2/16\}$,*

$$\left|b_k^T\hat{\beta}_k - (\mathcal{Z}_k^{preclean})^T\hat{\beta}_k/\sqrt{n}\right| \le (b_k^T\hat{\beta}_k)h$$

We use the fact that $(Z_k^{red})^T\hat{\beta}_k$ is zero. Note that Z_k^{res} is distributed Normal with zero mean and covariance $(I - Proj_k)$ so that we can write $(Z_k^{red})^T\hat{\beta}_k = Z(I - Proj_k)\hat{\beta}_k$ for some standard normal distributed Z. Since $(I - Proj_k)$ is orthogonal to $\hat{\beta}_k$, $(Z_k^{red})^T\hat{\beta}_k$ is zero so that the difference between $b_k^T\hat{\beta}_k$ and $(\mathcal{Z}_k^{red})^T\hat{\beta}_k/\sqrt{n}$ occurs only from the Chi-square random variable. Using the union bounds, the probability of the error is not more than $k^* exp\{-Lh^2/16\}$.

Lemma 6. *Suppose we have weights $w_j^* = e^{s_j}/\sum_{j'=1}^M e^{s_{j'}}$, for $j = 1, \ldots, M$. And consider another sets of weights where $w_j = e^{s_j+\epsilon_j}/\sum_{j'=1}^M e^{s_{j'}+\epsilon_{j'}}$, for $j = 1, \ldots, M$. Then,*

$$\left|\sum_{j=1}^M \left(w_j^2 - (w_j^*)^2\right)\right| \le 2\max_{j=1,\ldots,M}|\epsilon_j|$$

and if we pick any $j \in \{1, \ldots, M\}$, denote j_ℓ, then

$$\left|w_{j_\ell} - w_{j_\ell}^*\right| \le 2\max_{j=1,\ldots,M}|\epsilon_j|.$$

Furthermore, suppose there are other sets of weights, say $\{w_{2,j}\}_{j=1}^M$ and $\{w_{2,j}^*\}_{j=1}^M$. We denote the corresponding difference in the exponents $\epsilon_{2,j}$. Then we have,

$$\left|\sum_{j=1}^M \left(w_j w_{2,j} - w_j^* w_{2,j}^*\right)\right| \le 2\max_{j=1,\ldots,M}|\epsilon_j| + 2\max_{j=1,\ldots,M}|\epsilon_{2,j}|.$$

To prove the lemma, we use Taylor expansion with respect to the difference of the exponents. Then, we rearrange the equations and take out the maximum of the difference and bound it with a constant using the fact that the weights are positive and all sum up to one.

4.2. Using the idealized weight of combination

Given all the parameters, we know the sequence of expected success rate of Bayes optimal estimates, $x_1, \ldots, x_k$ from the idealized update function $x_{k+1} = g^{ideal}(x_k)$.

One way to approximate the idealized estimates is to combine $\mathcal{Z}_k$ with the same weights of combinations that we combined $\mathcal{Z}_k^*$ to construct Bayes optimal estimates in the previous section. Define $stat_k$ as

$$\begin{aligned} & \lambda^*_{k,0}\mathcal{Z}_0 + \lambda^*_{k,1}\mathcal{Z}_1 + \ldots + \lambda^*_{k,k}\mathcal{Z}_k + \frac{\sqrt{n}}{\sqrt{c_k}}\hat{\beta}_k \\ = \quad & \frac{\sqrt{c_k}}{\sqrt{c_0}}\mathcal{Z}_0 - \sum_{k'=1}^{k}\sqrt{c_k(\frac{1}{c_{k'}} - \frac{1}{c_{k'-1}})}\,\mathcal{Z}_{k'} + \frac{\sqrt{n}}{\sqrt{c_k}}\hat{\beta}_k \end{aligned}$$

The approximate optimal estimates is defines as, for $j \in sec_\ell$,

$$\hat{\beta}_{k+1,j} = \sqrt{P_\ell}\,\frac{e^{\alpha_{\ell,k} stat_{k,j}}}{\sum_{j' \in sec_\ell} e^{\alpha_{\ell,k} stat_{k,j'}}}$$

where $\alpha_{\ell,k} = \sqrt{nP_\ell/(\sigma^2 + (1-x_k)P)}$. Note that $\mathcal{Z}_k = \mathcal{X}_{n-k}b_k + Z_k$. We will see b_k is close to b_k^* and $\mathcal{X}_{n-k}$ is concentrated its mean $\sqrt{n-k}$ which is also not far from $\sqrt{n}$ if k is small enough so that $\mathcal{Z}_k$ is close to $\mathcal{Z}_k^*$.

Lemma 7. *For $1 \le k < k' \le k^*$, for any $\eta > 0$ and $\delta > 0$, followings hold in the event where the reliability and the Chi-concentration hold with $\delta = h^* = \eta$, for some constant a_k,*

(a) $|\beta^T\hat{\beta}_k - x_kP| \le a_k(n/L)^{k-1/2}\eta$

(b) $\left|\|\hat{\beta}_k\|^2 - x_kP\right| \le a_k(n/L)^{k-1/2}\eta$

(c) $|\hat{\beta}_{k'}^T\hat{\beta}_k - x_kP| \le (a_{k'}(n/L)^{k'-1/2} + a_k(n/L)^{k-1/2})\eta$

Thus, the above hold except an event of probability of

$$k^* exp\{-L\eta^2/16\} + 6k^* exp\{-\frac{2}{c^2}L\eta^2\}$$

For any small $\eta^* > 0$, we have actual success rate $\beta^T\hat{\beta}_k$ to be η^* close to x_kP except an event of probability bounded by $7k^* exp\{-\min(1/16, 2/c^2)L\eta^2\}$ where $\eta = (1/a_k)(n/L)^{-k^*+1/2}\eta^*$. If the number of steps is in constant order, then we can choose L large enough so that the error probability can be exponentially controlled. However, if the number of steps is in increasing order of L or M then it is hard to control the exponentially small error probability. The next method shows a possibility to improve the error bound.

4.3. Cholesky decomposition based estimates

We have seen a motivation from [8] based on coefficients of orthogonal components of the $\hat{\beta}_k$ in equation (4), with

$$\hat{\underline{\lambda}}_k^* = \Big((\sigma_Y - b_0^T\hat{\beta}_k), (-b_1^T\hat{\beta}_k), \ldots, (-b_k^T\hat{\beta}_k)\Big)$$

The Eq.4 has a representation approximately $\hat{stat}_k^* \approx Z_k^{comb} + \frac{\sqrt{n}}{\sqrt{\hat{c}_k^*}}\beta$. However, since we do not know β in advance, we cannot use it for our statistics. Instead, we can replace $b_{k'}^T\hat{\beta}_k$ with its estimates. Here, we study reasonable estimates for the weights of combination $\underline{\lambda}_k$ and see if the statistics based on those estimates are not far from the approximate form. The proof is not completed yet, but we will describe the idea and the strategy to prove.

4.3.1. Estimation of the weights of combination

Suppose we have a sequence of estimates $\hat{\beta}_1, \ldots, \hat{\beta}_k$. Let's consider a matrix $B = [\beta, \hat{\beta}_1, \ldots, \hat{\beta}_k]$ with dimension $(N+1) \times (k+1)$. Since $b_0, b_1, \ldots, b_k$ is Gram-smidts orthogonalization of columns of B with extended version, we can write B as following.

$$\underbrace{\begin{bmatrix} b_0 & b_1 & \cdots & b_k \end{bmatrix}}_{Q} \underbrace{\begin{bmatrix} (b_0^T\beta) & (b_0^T\hat{\beta}_1) & \cdots & (b_0^T\hat{\beta}_k) \\ 0 & (b_1^T\hat{\beta}_1) & \cdots & (b_1^T\hat{\beta}_k) \\ & & \ddots & \vdots \\ 0 & 0 & \cdots & (b_k^T\hat{\beta}_k) \end{bmatrix}}_{R}$$

This is the QR decomposition of the matrix B. Note that elements of the Cholesky factor matrix R are the components of $\hat{\underline{\lambda}}_k^*$. Moreover, if we consider the Cholesky decomposition of B^TB, we can write it as R^TR where

$$B^TB = \begin{bmatrix} \|\beta\|^2 & \beta^T\hat{\beta}_1 & \cdots & \beta^T\hat{\beta}_k \\ \hat{\beta}_1^T\beta & \|\hat{\beta}_1\|^2 & \cdots & \hat{\beta}_1^T\hat{\beta}_k \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\beta}_k^T\beta & \hat{\beta}_k^T\hat{\beta}_1 & \cdots & \|\hat{\beta}_k\|^2 \end{bmatrix}$$

$$= R^T \begin{bmatrix} (b_0^T\beta) & (b_0^T\hat{\beta}_1) & \cdots & (b_0^T\hat{\beta}_k) \\ 0 & (b_1^T\hat{\beta}_1) & \cdots & (b_1^T\hat{\beta}_k) \\ & & \ddots & \vdots \\ 0 & 0 & \cdots & (b_k^T\hat{\beta}_k) \end{bmatrix}$$

On the upper part, we know elements in B^TB with shaded region are from the data. On the lower part, for the diagonals in R(shaded), we have reasonable estimates from Cor. 5. Based on what we know from the data, we can estimate the rest of the components.

For each step k, suppose we know all $b_{k''}^T\hat{\beta}_{k'}$ for $0 \le k'' \le k' < k$ and $(b_k^T\hat{\beta}_k)$ exactly without any error. Then we can recover the rest of the elements by constructing one linear system along with one quadratic equation as following.

$$\begin{bmatrix} (b_1^T\hat{\beta}_1) & 0 & \cdots & 0 \\ (b_1^T\hat{\beta}_2) & (b_2^T\hat{\beta}_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ (b_1^T\hat{\beta}_{k-1}) & (b_2^T\hat{\beta}_{k-1}) & \cdots & (b_{k-1}^T\hat{\beta}_{k-1}) \end{bmatrix} \begin{bmatrix} (b_1^T\hat{\beta}_k) \\ (b_2^T\hat{\beta}_k) \\ \vdots \\ (b_{k-1}^T\hat{\beta}_k) \end{bmatrix}$$

$$= \begin{bmatrix} (\hat{\beta}_1^T\hat{\beta}_k) \\ (\hat{\beta}_2^T\hat{\beta}_k) \\ \vdots \\ (\hat{\beta}_{k-1}^T\hat{\beta}_k) \end{bmatrix} - (b_0^T\hat{\beta}_k) \begin{bmatrix} (b_0^T\hat{\beta}_1) \\ (b_0^T\hat{\beta}_2) \\ \vdots \\ (b_0^T\hat{\beta}_{k-1}) \end{bmatrix} \quad (5)$$

and

$$(b_0^T\hat{\beta}_k)^2+(b_1^T\hat{\beta}_k)^2+\cdots+(b_{k-1}^T\hat{\beta}_k)^2 = \|\hat{\beta}_k\|^2-(b_k^T\hat{\beta}_k)^2. \quad (6)$$

We can write $\left[(b_1^T\hat{\beta}_k),(b_2^T\hat{\beta}_k),\cdots,(b_{k-1}^T\hat{\beta}_k)\right]^T$ as a function of $(b_0^T\hat{\beta}_k)$ from the equation (5). We plug in the function to equation (6) and solve for $(b_0^T\hat{\beta}_k)$. Then, we can solve for the vector $\left[(b_1^T\hat{\beta}_k),(b_2^T\hat{\beta}_k),\cdots,(b_{k-1}^T\hat{\beta}_k)\right]^T$ using the solution in equation (6).

Suppose we have $\sqrt{L/n}\,h$ error for $b_k^T\hat{\beta}_k$. The solution of the algorithm won't be exact anymore, but we can use them as our estimates for the weight of combination $\underline{\lambda}_k^*$ after we normalize them to have a unit square norm and we will denote it $\hat{\lambda}_{k',k}$ for each element and $\hat{\underline{\lambda}}_k$ for the whole vector. The normalizer is $\hat{c}_k = \sum_{k'=1}^k \hat{r}_{k',k}^2 + (\sigma_Y - \hat{r}_{0,k})^2 = \sigma_Y^2 + \|\hat{\beta}_k\|^2 - 2\sigma_Y\hat{r}_{0,k}$ where each $\hat{r}_{k',k}$ is an estimate for $(b_{k'}^T\hat{\beta}_k)$.

Since we have quadratic equation, we need to concern uniqueness and existence of the solution. For exact recovery, we don't have to worry about the existence. For the uniqueness, for each step k, we have a belief that $\hat{\underline{\lambda}}_k^*$ is close to $\underline{\lambda}_k^*$ but not enough to use $\underline{\lambda}_k^*$ as an estimates though. However, it may be enought to give a direction which solution is needed for the analysis. With the errors, there is a possibility that we have no solution for the equation when we have negative sign for some test equation. If we adjust the solution to zero, then we may still have enough closeness to the actual solution.

Lemma 8. *For each $k = 1,\ldots,k^*$, as we discussed in Cor. 5, suppose we use $(\mathcal{Z}_k^{preclean})^T\hat{\beta}_k/\sqrt{n}$ as an estimate for $b_k^T\hat{\beta}_k$ and recover other Cholesky factors in R based on the estimate, then we have*

$$\left|b_{k'}^T\hat{\beta}_k - \hat{r}_{k',k}\right| \le F_{k',k}\sqrt{L/n}\,h$$

except probability, $6k^*\exp(-Lh^2)$. The $F_{R,k',k}$ is a function of elements from the upper $k'\times k$ part of the matrix R.

As we believe that $\hat{\underline{\lambda}}_k^*$ is close to $\underline{\lambda}_k^*$, each component $(b_{k'}^T\hat{\beta}_k)$ is close to some deterministic values $c_k\sqrt{\omega_{k'}}$ with $\omega_{k'} = 1/c_{k'} - 1/c_{k'-1}$. Since $F_{R,k',k}$ is function of elements in R which is close to that of R^* where R^* is replacing $(b_{k'}^T\hat{\beta}_k)$ with $c_k\sqrt{\omega_{k'}}$, $F_{R,k',k}$ is close to $F_{R^*,k',k} = (1-c_k/c_0)/\sqrt{\omega_{k'}} + \sqrt{\omega_{k'}}\,c_k(k-k'+1)$. The proof is not included in this paper, but it works as inductive argument as the matrix B^TB expands each step k.

4.3.2. Strategy for the Analysis

Now, we have an estimate for the weights of combination $\hat{\underline{\lambda}}_k^*$ which is denoted by $\hat{\underline{\lambda}}_k$. With a normalizer $\hat{c}_k$, we define the statistics at step k,

$$\hat{stat}_k = \sum_{k'=0}^{k}\hat{\lambda}_{k',k}\mathcal{Z}_k + \frac{\sqrt{n}}{\sqrt{\hat{c}_k}}\hat{\beta}_k$$

and define the estimate for $\hat{\beta}_{k+1}$ as, for $j\in sec_\ell$,

$$\hat{\beta}_{k+1,j} = \sqrt{P_\ell}\,\frac{e^{\hat{\alpha}_{\ell,k}\hat{stat}_{k,j}}}{\sum_{j'\in sec_\ell}e^{\hat{\alpha}_{\ell,k}\hat{stat}_{k,j'}}}$$

where $\hat{\alpha}_{\ell,k} = \sqrt{nP_\ell/\hat{c}_k}$.

Denote the estimate for β using Eq.4 as $\hat{\beta}_k^*$ although we can't use it as an estimate. This will be used as a bridge between our actual estimate and the expected success rate x_kP.

The main strategy is to show the closeness between $\hat{\beta}_k$ and $\hat{\beta}_k^*$ elementwise. Next, we know that $\hat{\beta}_k^*$ is near their expectation since they are sum of independent sections as we showed reliability earlier. Finally, the expectations are close to that of β_k^*. We try to show that the three quantities $\hat{\beta}_k$, $\|\hat{\beta}_k\|^2$ and $\hat{\beta}_{k'}^T\hat{\beta}_k$ with $k'>k$ are close to x_kP. Inductively, this leads to a confirmation of our belief that $\hat{\underline{\lambda}}_k^*$ is close to the deterministic values $\underline{\lambda}_k^*$. It follows that $\tilde{const}_k$ is bounded by some deterministic constant. Then we can prove that the estimation of the weights of combinations for the next step is close to $\hat{\underline{\lambda}}_{k+1}^*$. The proof works inductively as following.

(a) $\left|\beta^T\hat{\beta}_k - x_kP\right| \le \sqrt{n/L}\,h$

(b) $\left|\|\hat{\beta}_k\|^2 - x_kP\right| \le \sqrt{n/L}\,h$

(c) $\left|\hat{\beta}_{k'}^T\hat{\beta}_k - x_{k'}P\right| \le \sqrt{n/L}\,\eta$ for $k' = 1,\ldots,k-1$.

(d) $\left|b_{k'}^T\hat{\beta}_k - c_k\sqrt{\frac{1}{c_{k'}}-\frac{1}{c_{k'-1}}}\right| \le \sqrt{n/L}\,h$ for $k' = 1,\ldots,k-1$.

(e) $\left|b_{k'}^T\hat{\beta}_k - \hat{r}_{k',k}\right| \le \sqrt{L/n}\,h$ for $k' = 1,\ldots,k-1$.

(f) $\left|\lambda_{k,k'} - \hat{\lambda}_{k,k'}\right| \le \frac{2}{\sqrt{c_0}}\max_{k'}\left|b_{k'}^T\hat{\beta}_k - Est_{k',k}\right|$ $\le \frac{2}{\sqrt{c_0}}\sqrt{L/n}\,h$ for $k' = 1,\ldots,k$.

5. APPENDIX

Proof of Lemma 7. We proof inductively. There are hidden inductive argument as following, in order to have (a)-(c) we need

(A) $\sum_{\ell=1}^{L}\max_{j\in sec_\ell}|b_{k-1,j} - b_{k-1,j}^*|$ $\le d_{k-1}(n/L)^{k-3/2}\sqrt{L}\eta$

(B) $\sum_{\ell=1}^{L} \max_{j\in sec_\ell} |\mathcal{Z}_{k-1,j} - \mathcal{Z}^*_{k-1,j}|$
$\le 2\sqrt{n}Ld_{k-1}(n/L)^{k-3/2}\eta$

(C) $\sum_{\ell=1}^{L} \max_{j\in sec_\ell} |\mathcal{Z}^{comb}_{k-1,j} - \mathcal{Z}^{comb,*}_{k-1,j}|$
$\le A_k\sqrt{n}L(n/L)^{k-1-1/2}\eta$

for some constant d_k and A_k specified in the proof. We start with the difference between $\mathcal{Z}^*_0$ and $\mathcal{Z}_0$. The (A) for k=1 is trivial since $b^*_0 = b_0 = \beta/\sqrt{c_0}$. The only difference of the statistics occurs from the Chi-distribution,

$$|\mathcal{Z}^*_0 - \mathcal{Z}_0| = |\sqrt{n} - \mathcal{X}_n|\ b_0 \le \sqrt{L}h\, b_0.$$

Thus,

$$\sum_{\ell=1}^{L} \max_{j\in sec_\ell} |\mathcal{Z}_{k-1,j} - \mathcal{Z}^*_{k-1,j}| \le \sqrt{L}h \sum_{\ell=1}^{L} \max_{j\in sec_\ell} |b_{0,j}|.$$

This difference is also same as $|stat^*_0 - stat_0|$.

Recall that square norm of b_0 is not more than one since the extended vector $b_{0,e}$ is a unit vector. Thus, the maximum case of $\sum_{\ell=1}^{L} \max_{j\in sec_\ell} |b_{0,j}|$ would occurs when we have one $\sqrt{1/L}$ element in each section and zero elsewhere. Then the sum would be $\sqrt{L}$.

Using Lemma 6, we know that both $\left|\|\hat\beta_1\|^2 - \|\beta^*_1\|^2\right|$ and $\left|\beta^T\hat\beta_1 - \beta^T\beta^*_1\right|$ can be upperbounded by the same amount.

$$\begin{aligned}&\sum_{\ell=1}^{L} P_\ell\alpha_{\ell,0} \max_{j\in sec_\ell} |stat_{0,j} - stat^*_{0,j}|\\ \le\ & \left(\max_\ell P_\ell\alpha_{\ell,0}\right)\sum_{\ell=1}^{L}\max_{j\in sec_\ell}|stat_{0,j} - stat^*_{0,j}|\end{aligned}$$

Recall that $P_\ell\alpha_{\ell,0} = (P_\ell)^{3/2}\sqrt{n}/\sqrt{c_0}$. Thus, maximum of that will be equal to $\left((cP)^{3/2}/\sqrt{c_0}\right)\left(\sqrt{n/L^3}\right)$ where $c = L\,\max(P_\ell/P)$. Thus, we can conclude

$$\left|\|\hat\beta_1\|^2 - \|\beta^*_1\|^2\right| \le \frac{(cP)^{3/2}}{\sqrt{c_0}}\frac{\sqrt{n}}{\sqrt{L}}h.$$

Accordingly,

$$\begin{aligned}\left|\|\hat\beta_1\|^2 - x_1P\right| &\le \left|\|\hat\beta_1\|^2 - \|\beta^*_1\|^2\right| + \left|\|\beta^*_1\|^2 - x_1P\right|\\ &\le \frac{(cP)^{3/2}}{\sqrt{c_0}}\frac{\sqrt{n}}{\sqrt{L}}h + \delta P.\end{aligned}$$

The (a) and (b) both upper bounded by the same amount with $a_k = 2\frac{(cP)^{3/2}}{\sqrt{c_0}}$.

This completes proof for step $k = 1$. Suppose the conclusion is true for up to step k. We will show for step $k+1$ starting with $|b_k - b^*_k|$. we will see the difference of the denominator and numerator separately. We denote

$$b_k = \frac{\hat\beta_k - \sum_{k'=0}^{k-1}(b^T_{k'}\hat\beta_k)b_{k'}}{\sqrt{\|\hat\beta^2_k\|^2 - \sum_{k'=0}^{k-1}(b^T_{k'}\hat\beta_k)^2}} = \frac{num_k}{den_k}$$

and

$$b^*_k = \frac{\hat\beta_k - \hat\beta_{k-1} - \lambda^2_{k,k}(\beta - \hat\beta_{k-1})}{\lambda_{k,k}\sqrt{c_k}} = \frac{num^*_k}{den^*_k}.$$

Now

$$\begin{aligned}|b_k - b^*_k| &= \left|\frac{num_k}{den_k} - \frac{num^*_k}{den^*_k}\right|\\ &\le \frac{|num_k - num^*_k|}{den^*_k} + |b_k|\frac{|den_k - den^*_k|}{den^*_k}.\end{aligned}$$

By rearranging

$$\lambda_{k,0}\, b^*_0 + \lambda_{k,1}\, b^*_1 + \ldots + \lambda_{k,k}\, b^*_k = (\beta - \hat\beta_k)/\sqrt{c_k},$$

we can get

$$num^*_k = \hat\beta_k - (c_0 - c_k)\sqrt{\omega_0}\, b^*_0 - c_k\sum_{k'=1}^{k-1}\sqrt{\omega_{k'}}\, b^*_{k'}.$$

where $\omega_{k'} = 1/c_{k'} - 1/c_{k'-1}$. Thus, $|num_k - num^*_k|$ is equal to

$$\begin{aligned}&|\sum_{k'=0}^{k-1}(b^T_{k'}\hat\beta)b_{k'} - (c_0 - c_k)\sqrt{\omega_0}\, b^*_0 - c_k\sum_{k'=1}^{k-1}\sqrt{\omega_{k'}}\, b^*_{k'}|\\ &\le \left|(b^T_0\hat\beta_k - (c_0 - c_k)\sqrt{\omega_0})\right||b_0| +\\ &\sum_{k'=1}^{k-1}|(b^T_{k'}\hat\beta_k) - c_k\sqrt{\omega_{k'}}||b_{k'}| + \sum_{k'=1}^{k-1}c_k\sqrt{\omega_{k'}}|b_{k'} - b^*_{k'}|\end{aligned}$$

For the first coefficient $|b^T_0\hat\beta_k - (c_0 - c_k)\sqrt{\omega_0}|$, recall that $b_0 = \beta/\sqrt{c_0}$. Then we have

$$\begin{aligned}|b^T_0\hat\beta_k - (c_0 - c_k)\sqrt{\omega_0}| &= |\beta^T\hat\beta_k - x_kP|/\sqrt{c_0}\\ &\le (a_k/\sqrt{c_0})(\log M)^{k-1/2}\eta\end{aligned}$$

For the coeffient for k' where $|(b^T_{k'}\hat\beta_k) - c_k\sqrt{\omega_{k'}}|$ we prove using that $b_{k'}$ is close to $b^*_{k'}$ so that $|(b^T_{k'}\hat\beta_k) - c_k\sqrt{\omega_{k'}}| \le |b^T_{k'}\hat\beta_k - (b^*_{k'})^T\hat\beta_k| + |(b^*_{k'})^T\hat\beta_k - c_k\sqrt{\omega_{k'}}|$. Let's look at the first part on the right side. Note that $|b^T_{k'}\hat\beta_k - (b^*_{k'})^T\hat\beta_k| = \sum_{\ell=1}^{L}\sqrt{P_\ell}\sum_{j\in sec_\ell} w_{k,j}|b_{k',j} - b^*_{k',j}| \le \sum_{\ell=1}^{L}\sqrt{P_\ell}\max_{j\in sec_\ell}|b_{k',j} - b^*_{k',j}|$ by Holder's inequality. This is bounded by $d_{k'}(n/L)^{k'-1/2}\eta$ by the assumption.

Next we show the second part $|(b^*_{k'})^T\hat\beta_k - c_k\sqrt{\omega_{k'}}|$ is small. By simple algebra, we can see that

$$c_k\lambda_{k',k'}\sqrt{\omega_{k'}c_{k'}} = x_{k'}P - \frac{c_{k'}}{c_{k'-1}}x_{k'-1}P - (1 - \frac{c_{k'}}{c_{k'-1}})x_kP$$

Accordingly,

$$\begin{aligned}&|(b^*_{k'})^T\hat\beta_k - c_k\sqrt{\omega_{k'}}|\lambda_{k',k'}\sqrt{c_{k'}}\\ =\ &|\hat\beta^T_{k'}\hat\beta_k - (1 - \lambda^2_{k',k'})\hat\beta^T_{k'-1}\hat\beta_k - \lambda^2_{k',k'}\beta^T\hat\beta_k\\ &-c_k\sqrt{\omega_{k'}\lambda_{k',k'}c_{k'}}|\\ \le\ &|\hat\beta^T_{k'}\hat\beta_k - x_{k'}P| + (1 - \lambda^2_{k',k'})|\hat\beta^T_{k'-1}\hat\beta_k - x_{k'-1}P|\\ &+\lambda^2_{k',k'}|\beta^T\hat\beta_k - x_kP|.\end{aligned}$$

We show above is bounded using the reliability of $(\beta^*_{k'})^T\beta^*_k$. For any k' with $k' < k$,

$$|\hat{\beta}^T_{k'}\hat{\beta}_k - x_{k'}P| \leq |(\beta^*_{k'})^T\beta^*_k - x_{k'}P| + |(\beta^*_{k'})^T\beta^*_k - \hat{\beta}^T_{k'}\hat{\beta}_k|.$$

Recall that the first part on the right side is bounded by δP by Lemma 3. For the second part, using the Lemma. 6, we can bound $|(\beta^*_{k'})^T\beta^*_k - \hat{\beta}^T_{k'}\hat{\beta}_k|$ by

$$\begin{aligned} \leq & \sum_{\ell=1}^{L} P_\ell max_{j\in sec_\ell} |\alpha_{\ell,k'} stat^*_{k',j} - \alpha_{\ell,k'}\hat{s}tat_{k',j}| \\ & + \sum_{\ell=1}^{L} P_\ell max_{j\in sec_\ell} |\alpha_{\ell,k} stat^*_{k,j} - \alpha_{\ell,k}\hat{s}tat_{k,j}| \end{aligned}$$

From (a) and (b), we can conclude

$$|\hat{\beta}^T_{k'}\hat{\beta}_k - x_{k'}P| \leq a_{k'}(\log M)^{k'-1/2}\eta + a_k(\log M)^{k-1/2}\eta.$$

Accordingly, we can bound $\sum_{\ell=1}^{L} \max_{j\in sec_\ell} |num_{k,j} - num^*_{k,j}|$ by $(const_1)(\log M)^{k-1/2}\sqrt{L}\eta$ where the $const_1 = a_k\sqrt{\omega_0} + 2\sum_{k'=1}^{k}\{c_k\sqrt{\omega_{k'}}d_{k'} + 3a_k + (2-\lambda^2_{k',k'}a_{k'})/den^*_{k'}\}$. For denominator,

$$|den^2_k - (den^*_k)^2| \leq |\|\hat{\beta}_k\|^2 - \sum_{k'=0}^{k-1}(b^T_{k'}\hat{\beta}_k)^2 - \lambda_{k,k}c_k|.$$

Note that $\lambda_{k,k}c_k = (c_0 - c_k) - (c_0 - c_k)^2\omega_0 - \sum_{k'=1}^{k-1} c^2_k\omega_{k'}$. Recall that $(b^T_{k'}\hat{\beta}_k)$ was close to $c_k\sqrt{\omega_{k'}}$ for $k' = 1, \ldots, k-1$ and $(b^T_0\hat{\beta}_k)$ is close to $(c_0 - c_k)\sqrt{\omega_0}$. Thus,

$$\begin{aligned} & |den^2_k - (den^*_k)^2| \leq |\|\hat{\beta}_k\|^2 - \sum_{k'=0}^{k-1}(b^T_{k'}\hat{\beta}_k)^2 - \lambda_{k,k}c_k| \\ & \leq |\|\hat{\beta}_k\|^2 - (c_0 - c_k)| + |(b^T_0\hat{\beta}_k)^2 - (c_0 - c_k)^2\omega_0| \\ & + \sum_{k'=1}^{k-1} |(b^T_{k'}\hat{\beta}_k)^2 - c^2_k\omega_{k'}| \\ & \leq (const_2)(\log M)^{k-1/2}\eta \end{aligned}$$

where $(const_2) = (3 - 2c_k/c_0)a_k + 2\sum_{k'=1}^{k} c_k\sqrt{\omega_{k'}}(3a_k + (2-\lambda^2_{k',k'})a_{k'})/den^*_{k'}$ Accordingly, $\sum_{\ell=1}^{L}\max_{j\in sec_\ell}|b_k - b^*_k| \leq d_k(logM)^{k-1/2}\sqrt{L}\eta$ where $d_k = const_1/den^*_k + const_2/(den^*_k)^2$. Next, we evaluate the difference between $\mathcal{Z}_k$ and $\mathcal{Z}^*_k$. We have

$$\begin{aligned} |\mathcal{Z}_k - \mathcal{Z}^*_k| &= |\mathcal{X}_{n-k}b_k - \sqrt{n}b^*_k| \\ &\leq |\mathcal{X}_{n-k} - \sqrt{n}||b_k| + \sqrt{n}|b_k - b^*_k|. \end{aligned}$$

Thus, $\sum_{\ell=1}^{L}\max_{j\in sec_\ell}|\mathcal{Z}_k - \mathcal{Z}^*_k|$ is bounded above by $L\eta + \sqrt{n}Ld_k(\log M)^{k-1/2}\eta$ which is not greater than $2\sqrt{n}Ld_k(\log M)^{k-1/2}\eta$

Accordingly, $\sum_{\ell=1}^{L}\max_{j\in sec_\ell}|\mathcal{Z}^{comb}_{k,j} - \mathcal{Z}^{comb,*}_{k,j}|$ is bounded by

$$\begin{aligned} \leq & \sqrt{1-\lambda^2_{k,k}}\sum_{\ell=1}^{L}\max_{j\in sec_\ell}|\mathcal{Z}^{comb}_{k-1,j} - \mathcal{Z}^{comb,*}_{k-1,j}| \\ & + \lambda_{k,k}\sum_{\ell=1}^{L}\max_{j\in sec_\ell}|\mathcal{Z}_{k,j} - \mathcal{Z}^*_{k,j}| \\ \leq & A_k\sqrt{n}L(\log M)^{k-1/2}\eta \end{aligned}$$

where $A_k = \sqrt{1-\lambda^2_{k,k}}A_{k-1} + 2\lambda_{k,k}d_k$. Finally, we evaluate the difference of square norm and success rate for $\hat{\beta}_{k+1}$. The distance from the square norm and the success rate of $\hat{\beta}_{k+1}$ to $x_{k+1}P$ is not more than

$$(\max_\ell P_\ell\alpha_{\ell,k})\sum_{\ell=1}^{L}\max_{j\in sec_\ell}|\mathcal{Z}^{comb}_{k,j} - \mathcal{Z}^{comb,*}_{k,j}|.$$

Recall that $(\max_\ell P_\ell\alpha_{\ell,k})$ is not more than $\frac{(cP)^{3/2}}{\sqrt{c_k}}(\sqrt{\frac{n}{L^3}})$. Thus, we can conclude the above quantity is not more than $a_{k+1}(\log M)^{k+1/2}\eta$ where $a_{k+1} = A_k\frac{(cP)^{3/2}}{\sqrt{c_k}}$. This completes the proof

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.

[8] A.R. Barron and S. Cho, "High-Rate Sparse Superposition Codes with Iteratively Optimal Estimates." *Proc. IEEE Intern. Symp. Inform. Theory*, Boston, USA, 2012.