High-dimensional regression with random design, including sparse superposition codes

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

by
Sanghee Cho

Dissertation Director: Andrew R. Barron

Mar 14, 2014

Abstract

High-dimensional regression with random design, including sparse superposition codes

Sanghee Cho

2014

The dissertation studies variable selection for a linear regression model. We focus on high-dimensional setting where the number of explanatory variables is much greater than the number of observations. For the independent variables, we assume a random design.

The first chapter of the dissertation evaluates the performance of an out-of-sample prediction estimator when the complexity of candidate models is controlled relative to the sample size. We refine an upperbound to a tail probability for the distance between the out-of-sample prediction and its estimate in Leeb (2008) in order to deal with models of varying complexity relative to the sample size. We construct a modified model selection criterion that will allow us to guarantee the performance over a large class of candidate models.

The second chapter of the dissertation is an application of variable selection to the mathematical theory of communication. We focus on a problem of high rate sparse superposition codes (sparse regression codes) for the additive white Gaussian noise channel with a power control. This problem can be interpreted as a variable selection with a special structure of the sparse coefficient vector β. We propose a fast and reliable algorithm of iterative updates for estimating the coefficient vector motivated by Bayes optimal estimates.

Contents

Acknowledgements

I would like to express the deepest appreciation to my advisor Andrew R. Barron. Without his endless support, encouragement and guidance, this dissertation would not have been possible. I wish to thank Harrison Zhou and Sahand Negahban for insightful comments and suggestions. Also I am very grateful to all the professors and staffs in Statistics department for their support. Special thanks go to Hyunjoong Kim who initiated my interest in Statistics.

I would like to thank my dear friends Heejin Kim, Hyunkyu Kim, Kyounghee Kim, Meekyoung Kim, Sunkyoung Kim, Jinhee Kwon, Jiyoung Ko, Eunhee Lee and Jihyung Lee for their priceless support.

I am extremely grateful and indebted to my parents who have always believed in me and encouraged me. Lastly, I would like to give a deepest gratitude to my husband, Byoung Gun Park who has always there for me as my love and my best friend even when I had a tough time. His encouragement and support was invaluable.

For simulations, it was supported in part by the facilities and staff of the Yale University Faculty of Arts and Sciences High Performance Computing Center, and by the National Science Foundation under grant # CNS 08-21132, National Institute of Health under grants RR19895 and RR029676-01 that partially funded acquisition of the facilities.

Introduction

The dissertation consists of two parts. The first part *Loss bounds for out-of-sample prediction* studies loss bounds for out-of-sample prediction and proposes a model selection criterion considering model complexity. The second part *Sparse superposition codes for the Gaussian channel with approximate iterative optimal estimates* presents an application of variable selection to information theory.

We look at variable selection under the general linear model setting as following

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

Here, ϵ is a Gaussian noise with mean zero and variance σ^2. The sample size will be denoted by n. The total number of explanatory variables that can be related to the response can be finite or infinite. The goal is to find a good subset of these variables that can be related to Y. In the first chapter, we look for a subset of explanatory variables that minimizes out-of-sample mean square error prediction. The second chapter is an application to a decoder for sparse superposition codes with Gaussian Channel. The goal is to find a subset of the elements of β with non-zero values in the above model.

The first chapter studies a problem of selecting a model out of a class of candidate models. When there are more variables than the sample size, the classical approach is to set a class of candidate models where each candidate model consists of a subset of variables where the cardinality of the subset is much smaller than the sample size. The idea is to look for a subset that has the minimum mean square error prediction. However, we can't calculate the minimum mean square error prediction not knowing the true parameter β. Instead, we use its estimate as a model selection criterion. We wish to study properties of this estimate, in order to be confident about our model selection criterion.

Based on a random design setting, Leeb (2008) showed that the true out-of-sample predictive performance is well approximated by the generalized cross validation (GCV) or S_p criteria with high probability, uniformly over the whole candidate models, especially if the logarithm of number of candidate models is of smaller order than the sample size.

In this chapter, an improved upperbound to the tail probability is established for the distance between the out of sample prediction and its estimate relevant to model selection. Some examples of refined GCV as a model selection criterion is provided taking into account the model complexity. In this way, I propose a modified model selection criterion, which allows us to guarantee the performance over a large class of candidate models.

The second chapter studies a problem of variable selection applied to mathematical theory of communication. Here, we focus on developing fast and reliable decoder for high rate sparse superposition codes with the additive white Gaussian noise channel with a power control. This coding scheme for AWGN channel is first developed by Joseph and Barron (2012) and it coincides with the problem of variable selection for a linear model $Y = X\beta + \epsilon$ with some special structure of coefficient vector.

The message is conveyed through the sparse coefficient vector β. It is designed to be partitioned into L sections with only one non-zero term in each section. We

assign the non-zero values in each section to be exponentially decaying as the section index increase. The assigned non-zero value for each section is shared by sender and the receiver so that the goal of the decoder is to estimate the indices of the non-zero terms.

We propose an adaptive successive decoder using soft decisions motivated by Bayes optimal estimates at each step with a uniform prior on the terms sent. We examine the performance of the decoder based on the optimal estimates. However, since the exact Bayes optimal estimates are infeasible, we propose two methods to approximate the Bayes optimal estimates and show their reliability and that the performance of the estimate is not far from the optimal ones. We use the method of the nearby measure for the distributional analysis. It allows us to use a convenient distribution that is not far from the true distribution when the Renyi relative entropy between the two distributions is bounded.

Our theory shows that for any fixed rate below the capacity C there is a fast and reliable communication with exponentially small error probability.

Numerical simulation shows that the decoder follows the update rule that we have examined with the optimal estimates and it is improved from the one with the thresholding decoder.

Chapter 1

Loss bounds for out-of-sample prediction

1.1 Introduction

In the procedure of constructing a model, selecting a model out of a class of candidate models is a challenging problem in statistics. This includes choosing a subset of explanatory variables which can be related to our response Y. Not knowing the true model, we need to make a decision based on the data set we have. Numerous researchers have been working on constructing and analyzing criteria, which give a way to select a model. Often these criteria are based on a Mean Square Prediction Error (MSPE), which can be estimated by a function of the Residual Sum of Squares (RSS).

Various methods of including certain penalty terms were suggested such as Mallow's C_p. They are functions of RSS to which a penalty term is added. The penalty term is a function of the number of parameters and the sample size. Also, there are criteria such as Akaike's Final Prediction Error (FPE), Craven and Wahba's Generalized Cross Validation (GCV) and Tukey's S_p criterion that are multiplications of RSS and a penalty term.

Several of these criteria were originally explored as unbiased estimators for the mean square prediction error. The FPE and C_p are the ones for fixed design and S_p is the one for random design. These can be used as a method of comparing the performances of the models so that we can select one of the models for interpretation or prediction.

However, unbiasedness does not guarantee that a criterion works well as a method of model selection. That is, unbiasedness won't guarantee that the MSPE of the minimizer of the criterion would be the minimum MSPE over all candidate models. Several researchers have worked on how the criteria perform over a class of candidate models. Some classical references are Shibata (1981), Breiman and Freedman (1983) and Li (1987). Leeb (2008) pointed out that the classical performance analysis of the criteria does not give a clear picture as to what method is preferable. Based on a random design setting, Leeb (2008) showed that the true out-of-sample predictive performance is well approximated by the GCV or S_p criteria with high probability, uniformly over the whole candidate models, if the logarithm of number of candidate models is of smaller order than the sample size.

However, Leeb (2008)'s theorem requires the logarithm of number of candidate models to be of smaller order than the sample size. This strong condition allows the use of a union bound on a tail probability, to establish uniformity over all candidate models. In this chapter, we will consider a more flexible tail bound on each candidate model which is an improvement on the bound in Leeb (2008). It allows us to guarantee GCV's performance over a somewhat larger class of candidate models. We describe the model selection setting in Section 1.2 and review classical approaches in Section 1.3. In section 1.4, we summarize Leeb (2008)'s approach and develop our improved upperbound to the tail probability for the distance between the out-of-sample prediction and its estimate. In Section 1.5, how the penalty should depend on whether leading term or arbitrary subset models are being considered. In Section 1.6, we modify the GCV using a technique of Barron (1991) and Barron et al. (1999) using model complexity. Also, we show that a loss of the minimizer of the criteria is bounded by a minimum loss over all candidate models except in an event of a small probability. Some possible further research is discussed and we conclude in Section 1.7.

1.2 Setting

There are some additional notations for this chapter. The true model would be, for one observation,

$$y = \sum_{j=1}^{K} x_j \beta_j + \epsilon. \tag{1.1}$$

Here, ϵ is a random noise with mean zero and variance σ^2 independent from any explanatory variable. The total number of explanatory variables that can be related to the response, which is denoted by K, can be finite or infinite and each x_j for $j = 1, \dots, K$ can be fixed or random. If they are random, the sequence of explanatory variables $x = (x_j)_{j=1}^{K}$ are normally distributed with mean zero and variance/covariance matrix $\Sigma = [\mathbb{E}(x_i x_j)]_{i,j \geq 1}$. We define $\mathcal{M}_n$ as a class of candidate models. For each finite subset m of the variables, let $\hat{\beta}(m)$ be the least squares estimate based on a sample (X, Y) of size n only using the finite subset of explanatory variables m. The size of a specified subset m is denoted as $p = |m|$.

1.3 Classical Approach

The basic idea of a model selection is this: based on the information we have, select a model out of the candidate models that best approximates the true function. For measuring how well a candidate model approximates the true function, ideally we can use the mean square prediction error (MSPE),

$$E(y^f - \hat{y}^f)^2,$$

where y^f is a future response and $\hat{y}^f$ is a fitted value based on the training data set. Since MSPE is an unknown population value, we estimate it based on training data and select a model that minimizes the estimate. We expect that the model we select will approximate the true function as well as the minimizer of the mean square prediction error. Lots of model selection criteria, such as FPE, S_p, Mallow's C_p, AIC and cross validation are constructed in this way.

Several model selection criteria look similar in terms of that they are all functions of Residual sum of squares (RSS) and of the number of parameters. One issue that makes a difference is that we can assume that the explanatory variables are fixed or random. We also need to think about whether the future realization of the explanatory variables is the same as the training data set or not. This affects what is an estimator of the mean square prediction error.

For fixed design, suppose we are considering n future responses with the same explanatory variables as training data set, where Y^f would be a $(n \times 1)$ vector, X_m would be a $(n \times |m|)$ matrix which is a collection of the column vectors of the considered subset, X_q is the rest of the columns and β_q is the vector of the coefficients corresponding to X_q. Then the mean square error prediction is,

$$\frac{1}{n}E\|Y^f - \hat{Y}^f_m\|^2 = \frac{1}{n}\beta_q X_q(I - P_0)X_q\beta_q + \sigma^2 + \frac{m}{n}\sigma^2, \tag{1.2}$$

where P_0 is projection matrix of X_m and $\hat{Y}^f_m$ is a least square fit from the training data set (X, Y), but only using the subset m of explanatory variables.

In contrast, for random design, if we consider only one future response,

$$\begin{aligned} E(y^f - \hat{y}^f_m)^2 &= (\sigma^2_m + \sigma^2)(1 + \frac{|m|}{n - 1 - |m|}) && (1.3)\\ &= \sigma^2_m + \sigma^2 + (\sigma^2_m + \sigma^2)\frac{|m|}{n - 1 - |m|}, && (1.4) \end{aligned}$$

where $\sigma^2_m = var(\sum_{j \notin I_m} x_j\beta^*_j | x_i, i \in I_m)$ and I_m is an index set of a model m with β^*_j as a j-th element of $\arg\min_{\beta_1,\ldots,\beta_N} E(y - \sum_{j=1}^N x_j\beta_j)^2$. Note that σ^2_m is not random, since x_j's are jointly Gaussian. For (1.4), Breiman and Freedman (1983) pointed out that the first term indicates error from omitted variables, which will get small as $|m|$ increases. The second term indicates the future error and the third term measure the effect of the estimation. So the third term would gets large as the model get large. Thus, we can see that there is a trade off between accuracy and the number of parameters.

We can see that their interpretation also works for (1.2). Note that, for fixed design, we can see that the first and the third term is due to the model we are considering, but the second term is from the future error which is not related to the model selection.

1.3.1 Fixed Design

Akaike (1969, 1970) proposed a practical procedure of predictor identification, called

Final Prediction Error (FPE),

$$FPE(m) = \left(1 + \frac{|m|}{n}\right) \frac{RSS_m}{n - |m|}.$$

If our true model is on a space of a subset of X_m, then FPE will be an unbiased estimator for MSPE, since the first term in (1.2) would be 0 and $RSS_m/(n - |m|)$ is an unbiased estimator for σ^2.

Daniel and Wood (1971) recommended C_p given by Mallows, Mallows (1973) as a measure of 'total squared error'. It measures a sum of the squared biases plus the squared random errors in Y at all data points. The C_p statistic is defined as,

$$C_p(m) = \frac{RSS_m}{\hat{\sigma}^2} - n + 2|m|.$$

This is an unbiased estimate of

$$\frac{1}{n}\mathbb{E}\|Y^f - \hat{Y}^f\|^2 - \sigma^2. \tag{1.5}$$

For estimate of σ^2, we can use a residual sum of squares from the full model, which contains all the candidate explanatory variables, which would be $RSS(N)/(n - N)$. Here, we would need an assumption that $N < n$.

By disregarding all the terms that don't change over all candidate models, we can think FPE and C_p as

$$\begin{aligned} FPE^*(m) &= \left(1 + \frac{2|m|}{n - |m|}\right) RSS_m, \\ C_p^*(m) &= RSS_m + 2|m|\hat{\sigma}^2. \end{aligned}$$

If $\hat{\sigma}^2$ is close to $RSS/(n - |m|)$ which happens when model m approximates the true model well, then C_p will be close to FPE. Thus, if the true model is a subset of m, then both FPE^* and C_p^* would be an unbiased estimate for (1.2). But if not, FPE^* would have some bias term whereas C_p^* is still an unbiased estimator.

Shibata (1981) analyzed the asymptotic optimality of a criterion

$$\frac{(n+2|m|)}{n} RSS_m.$$

The motivation was based on the loss $\|X\beta - X\hat{\beta}\|^2$ among least squares estimate with subset of explanatory variables, m. The goal is to find a model which minimize the expected loss or to find a model which has an expected loss close to the minimum. The expected loss here is equal to (1.5) which was a motivation for C_p.

The key assumption in Shibata (1981), which was something different from the previous criteria, is that the number of explanatory variables is infinite or increases with the sample size. And the criterion is not an exact estimate of the risk. Shibata (1981) also proved that it has asymptotic optimality. If we denote a model $\hat{m}$ which minimize the above criterion, the criteria has a property of

$$\lim_{n\to\infty} \frac{E\|X\beta - X_{\hat{m}}\hat{\beta}(\hat{m})\|^2}{min_m E\|X\beta - X_m\hat{\beta}(m)\|^2} = 1. \tag{1.6}$$

And the paper mentioned that C_p, FPE and AIC have the same asymptotic optimality.

1.3.2 Generalized Cross Validation

Craven and Wahba (1978) suggested a criterion called generalized cross-validation (GCV) based on spline smoothing. The main purpose of the paper was suggesting an effective method for estimating the optimum amount of smoothing from the data without a knowledge of a variance of error. GCV is designed for regression estimates that are linear in the vector of observed response values, $Y = (y_1, y_2, \ldots, y_n)'$. If $A(\lambda)$ is a matrix for a certain model with smoothing factor λ, which satisfies $\hat{Y} = A(\lambda)Y$, then GCV is defined as,

$$V(\lambda) = n\|(I - A(\lambda))y\|^2 / tr(I - A(\lambda))^2.$$

If we set $\hat{Y} = A(\lambda)Y$ as a projection of Y onto a linear space of X_m, above criterion would be

$$GCV_m = \frac{RSS_m}{n - |m|} \frac{n}{n - |m|}.$$

Craven and Wahba proved that if we use a smoothing factor λ which minimizes $V(\lambda)$, the risk of the certain model will be asymptotically same as the minimum risk. GCV estimator has some common points with other criteria that we discussed above. When p/n is small, then GCV is close to FPE. Li (1987) explored the asymptotic behaviors of model selection procedures of GCV and C_p. Golub et al. (1979) pointed out that GCV has an advantage that we don't have to estimate σ^2. Thus it also can be used when the number of degrees of freedom for estimating σ^2 is small.

Even though, it is motivated under a fixed design, the name GCV is from the fact that origin of $V(\lambda)$ is from cross validation. Later, Leeb (2008) pointed out that it can be a good estimate for the conditional mean square prediction error in random design setting, which will be discussed in Section 1.4.

1.3.3 Random Design

Breiman and Freedman (1983) assumed that the error ϵ and the explanatory variables are jointly Gaussian and ϵ is independent of all the explanatory variables. As the number of explanatory variables gets large, the minimizer of S_p criterion would have minimum risk asymptotically for special case of nested models. The S_p is defined by,

$$S_p(m) = \left(1 + \frac{|m|}{n-1-|m|}\right) \frac{RSS_m}{n-|m|},$$

which is an unbiased estimator of (1.4). The S_p criterion is first given explicitly by Hocking (1976) and further explored by Thompson (1978).

The motivation for S_p in the paper is based on a loss which would be viewed as a l^2-distance between a future response and its prediction, with respect to the conditional probability of future values conditioning on the training data (X, Y). The loss for specific model m would be the distance between the future response and the prediction based on the linear projection of Y onto a space of X_m. It is actually equivalent to the conditional MSPE,

$$\rho^2(m) = \mathbb{E}[(y^{(f)} - \hat{y}_m^{(f)})^2 | Y, X], \tag{1.7}$$

where $\hat{y}_m^{(f)}$ is prediction of future y based on a model m. By unbiased property, we can say that

$$E[S_p(m)] = E[\rho^2(m)].$$

Let $\hat{m}$ is a minimizer of $S_p(m)$ over all candidate models in $\mathcal{M}_n$. In nested models case, under an assumption that $\sigma_m^2 > 0$ for all m, Breiman and Freedman (1983) showed that

$$\frac{\rho^2(\hat{m}) - \sigma^2}{\min_m [E(y^f - \hat{y}_m^f)^2 - \sigma^2]} \to 1 \text{ in prob}$$

and

$$\frac{\min_m [\rho^2(m) - \sigma^2]}{\min_m [E(y^f - \hat{y}_m^f)^2 - \sigma^2]} \to 1 \text{ in prob.}$$

That is, a loss of a model $\hat{m}$ will be close to a minimum MSPE and also to a minimum $\rho^2(m)$, eventually. These are analogues for random design of proportion of asymptotic optimality in fixed design studied by Shibata (1981) and Li (1987). Later, Leeb (2008) also evaluated some aspects of models' performances by the conditional MSPE over a larger class of models, not necessarily nested models. He stated that S_p can be a good estimate of conditional MSPE.

1.3.4 From Risk Estimation to Model Selection

First, researchers focused on estimating the fit of a single model. One of the main goals of the criteria was to estimate the risk and adjust some bias. However, Shibata (1981) and Breiman and Freedman (1983) were interested in the criteria as a method for comparing the models among all candidate models. That is, if we select a model which minimizes the criteria, then the risk of the model we select would be close to the minimum risk.

When it comes to a performance analysis on model selection procedure, one of the common ways is finding a good upperbound to the tail probability for the distance between the risk and its estimate. If we can show that the tail probability gets small as n grows over all candidate models, we can discuss consistency.

Leeb (2008) studied GCV and S_p to prove consistency in this way. However, it is only guaranteed when the logarithm of number of candidate models is of smaller order than the sample size. It is due to an upperbound and uniform tail probability that he was considering. In the following sections, we will explore an improved upperbound to a tail probability and flexible error on each candidate models and show the consistency over a larger class of candidate models.

1.4 Improved Upperbound to a Tail Probability

Leeb (2008) considered a problem where the number of candidate models is relatively large and the random design as in Section 1.2. Our goal is to find a model with 'good' out-of-sample predictive performance.

In Leeb (2008), he considered out-of-sample prediction with loss given by the conditional mean squared error of the corresponding predictor, where the conditioning is on the training sample,

$$\rho^2(m) = \mathbb{E}[(y^{(f)} - \hat{y}_m^{(f)})^2 | Y, X] \tag{1.8}$$

and a natural estimator of it,

$$\hat{\rho}^2(m) = \frac{RSS_m}{n - |m|} \frac{n+1}{n - |m|} \tag{1.9}$$

given that $n - 1 > |m|$. This is equivalent criteria to GCV since it is monotone function of GCV. As discussed in Leeb (2008), GCV, S_p and $\hat{\rho}^2(m)$ work well in selecting a model, even if the candidate models are complex when compared to sample size and also if the number of candidate models is much larger than sample size.

One can select a model by selecting a model which minimizes (1.9). To obtain desirable properties of this model selection procedure, we need to establish that $\rho^2(m)$ is close to $\hat{\rho}^2(m)$ with high probability, not only for a fixed model m, but for an entire collection of candidate models. This allows us to say that $\rho^2(m)$ of selected model is close to the minimum $\rho^2(m)$ over all candidate models. To evaluate the performance of the estimator, Leeb first found an upperbound of the tail probability of the distance between $\rho^2(m)$ and $\hat{\rho}^2(m)$,

$$P_{n,\beta,\sigma,\Sigma}\big(sup_{m\in\mathcal{M}_n}|\hat{\rho}^2(m)-\rho^2(m)|>\epsilon\big) \tag{1.10}$$

$$\leq 4\exp\left[-n(1-\frac{|m|}{n})\Psi(\frac{\epsilon}{2\sigma^2(m)}(1-\frac{|m|}{n}))\right] \tag{1.11}$$

where $\sigma^2(m)=\sigma_m^2+\sigma^2$ and $\Psi(\cdot)$ is defined by $\Psi(x)=(x/(x+1))^2/8$ for $x\geq 0$.

Extracting the essence of his theorem, using union bounds, one has

$$P_{n,\beta,\sigma,\Sigma}\big(\exists m\in\mathcal{M}_n \text{ s.t } |\hat{\rho}^2(m)-\rho^2(m)|>\epsilon\big) \tag{1.12}$$

$$\leq 4\#\mathcal{M}_n\exp\left[-n(1-\gamma_n)\Psi((\epsilon/(2c))(1-\gamma_n))\right] \tag{1.13}$$

with $\gamma_n = sup_{m\in\mathcal{M}_n}\frac{|m|}{n}$ and under the assumption $Var_{\beta,\sigma,\Sigma}[y]\leq c$. Leeb(2008) established that if the order of logarithm of number of candidate models is smaller than the sample size, then the upperbound of the probability goes to zero.

However, a function $\Psi(x)$ is bounded above by a constant 1/8. We would want the exponent,

$$\log(\#\mathcal{M}_n)-n(1-\gamma_n)\Psi((\epsilon/(2c))(1-\gamma_n)),$$

to be infinity as n grows. But $\Psi(x)$ can prevent the exponent going to infinity when $\#\mathcal{M}_n$ is large. If we can find an increasing function which can replace $\Psi(x)$ as in the following theorem, we can improve the tail probability.

Theorem 1. *Consider a candidate model m. For each $\epsilon_m>0$,*

$$P(|\rho^2(m)-\hat{\rho}^2(m)|>\epsilon_m)\leq 5\exp\left[-\frac{n-|m|}{2}\mathcal{L}(\frac{\epsilon_m(n-|m|+1)}{2\sigma^2(m)(n+1)})\right].$$

The function $\mathcal{L}(\cdot)$ is given by

$$\mathcal{L}(c) = c - \log(1 + c).$$

Here, $\mathcal{L}(c)$ is an increasing function. It behaves like a quadratic function when c is small and an unbounded linear function when c gets large. The proof is in Appendix A.2

1.5 Model Selection using Arbitrary ϵ_m

In Leeb(2008), in order to show the overall performance, he used a constant ϵ to bound the tail probability. He appealed to such a strong condition to guarantee the performance when the logarithm of the number of candidate models is of smaller order than the sample size. We can also consider different errors for different models. That is, we can have ϵ as a function of m. For example, if the data reveal that a simple model works better than a complex model, then we would like the criterion to reveal that behavior. So, we allow a small error for a small model and relatively high error on a complex model. In this section, we will explore that even though the logarithm of the number of candidate models is larger or has the same order as the sample size, we can bound the probability with some population quantity that depends on the sample size in a way that permits us to control the size of the bound. We will use an upperbound in Leeb (2008) and provide some examples. We presume that the models of interest use at most the first K parameters, $\beta_1, \ldots, \beta_K$, subset models are specified by sequence in $\{0, 1\}^K$ that specify which terms may be non-zero.

1.5.1 Examples

Leading term submodel

In this theory, we make the set of models $\mathcal{M}_n$ such that each $|m| < n-1$ for all $m \in \mathcal{M}_n$. For example, we might have $K = \frac{n}{2}$ and consider all models of the form

$$M_n = \{(0,...,0),(1,0,..,0),...,(1,1,...,1)\} \tag{1.14}$$

with the number of ones not more than $K = \frac{n}{2}$. This gives $\frac{n}{2}$ candidate models.

Lemma 1. *If we set* $\varepsilon_m = \sqrt{\frac{\delta}{(1-\frac{|m|}{n})^3}}$ *which increases as* $|m|$ *increases,*

$$\begin{aligned}(1.13) \quad &\leq \sum_{|m|=0}^{n/2} 4\exp(-\frac{n}{16\sigma^4}\delta) \\ &= \exp(-\frac{n}{16\sigma^4}\delta + \log\frac{n}{2}) \to 0 \quad \textit{as } n \to \infty. \qquad (1.15)\end{aligned}$$

Submodels with r parameters (out of n)

Let's consider submodels with r parameters. That is, we are considering the r explanatory variables out of K predictors. Then a number of candidate model will be $\binom{K}{r}$. Let's say a number of potential explanatory variables are same as the number of sample, n. Let's consider when r increases as n increases, $r = \log n$.

Lemma 2. *Consider* $r = \log n$. *Using Sterling's formula,* $n! \approx \sqrt{2\pi}n^{n+\frac{1}{2}}e^{-n}(1+o(\frac{1}{n}))$*,*

$$\begin{aligned}\binom{n}{r} &\approx \frac{1}{\sqrt{2\pi}}\exp((\log n)^2(1+o(1))) \\ &\leq C\exp((\log n)^2) \qquad (1.16)\end{aligned}$$

Lemma 3. *If we set* $\epsilon_m = \sqrt{\frac{\delta}{(1-\frac{|m|}{n})^3}}$ *which increases as* m *increases.*

$$
\begin{aligned}
(1.13) \quad \leq \quad & \sum_{m \in \mathcal{M}_n} 4 \exp(-\frac{n}{16\sigma^4}\delta) \\
\leq \quad & 4c \exp\left(-n(\frac{1}{16\sigma^4}\delta - \frac{(\log n)^2}{n})\right). \qquad (1.17)
\end{aligned}
$$

Thus, if $\delta \gg \frac{(\log n)^2}{n}$, *then* (1.12) *goes to zero as* n *gets large.*

1.5.2 Estimating Good Out-of-sample Prediction

By allowing a flexible error for each model, the difference $\rho^2(m)$ and $\hat{\rho}^2(m)$ can be small enough except in an event of a small probability, even though a number of candidate models is relatively high with respect to a sample size. Given the training data (X, Y), an ideal model m^* would be a model that minimizes $\rho^2(m)$. In its place, one may use a model selection $\tilde{m}$ that minimizes $\hat{\rho}^2(m)$. If $\rho^2(\tilde{m})$ is very close to $\rho^2(m^*)$, then we are choosing a model that has similar predictive performance. Let's denote $\check{m} = \arg\min_{m \in \mathcal{M}_n}[\hat{\rho}^2(m) - \epsilon_m]$ where the performance of $\check{m}$ is not far from $\tilde{m}$ or m^*.

Theorem 2. *If* $|\rho^2(m) - \hat{\rho}^2(m)| < \epsilon_m$ *for each* $m \in \mathcal{M}_n$, *then*

$$
|\rho^2(m^*) - \rho^2(\tilde{m})| < \max(2\epsilon_{\tilde{m}}, \epsilon_{\tilde{m}} + \epsilon_{\check{m}}) \qquad (1.18)
$$

If there is a case that we can control ϵ_m small enough just for the models with small $\rho^2(m)$, then we can say that $\hat{\rho}^2$ works well as a model selection criterion. That is, it is enough if we can control ϵ_m small enough for the models with small $\rho^2(m)$ or $\hat{\rho}^2(m)$ instead of controlling ϵ over the whole candidate models.

1.6 Model Selection using Complexity

In Barron (1991), he defined general complexity regularization criteria and establish bounds on the statistical risk of the estimated functions. Also, these bounds establish consistency, yield rates of convergence and demonstrate the near asymptotic optimality of the model selection criterion. Here, we consider ϵ_m as a function of a complexity and the sample size, similar to Section 1.5. Also, I combine the bounding techniques of Leeb (2008) with Barron (1991)'s. We look at the performance of $\hat{\rho}^2(m)$ controlling the complexity and get an upperbound on $\rho^2(\hat{m})$ in terms of $\min_{m\in\mathcal{M}_n}\{\rho^2(m)(1+\delta_m)^2\}$ where $\hat{m} = \arg\min_{m\in\mathcal{M}_n}\{\hat{\rho}^2(m)(1+\delta_m)\}$. Based on the upperbound we found above in Section 1.4, we can finally build a model selection using complexity.

Theorem 3. *For each $m \in \mathcal{M}_n$, let's consider a model selection criteria*

$$\hat{m} = \arg\min_{m\in\mathcal{M}_n}[\hat{\rho}^2(m)(1+\delta_m)]$$

for

$$\delta_m = \frac{4f_m}{1-2f_m} \qquad \textit{with} \qquad f_m = f\left(\frac{2(C_n(m)+\log 1/\delta)}{n-|m|}\right)$$

where $f(x) = \log\{e^x + \sqrt{e^x+1}\sqrt{e^x-1}\}$. Here, we select complexity term $C_n(m)$ which satisfies $\sum_{m\in\mathcal{M}_n}\exp(-C_n(m)) < 1$ and $2(C_n(m)+\log 1/\delta) < 9(n-|m|)$. Then, except an event of a probability 5δ, we have

$$\rho^2(\hat{m}) \leq \min_{m\in\mathcal{M}_n}\{\rho^2(m)(1+\delta_m)^2\}$$

The detailed proof is in Appendix A.3. The main tool of the proof is the same as the one we used for Section 1.4. We want the criteria to be more accurate for the models when out-of-sample prediction is small and model complexity is controlled. From the above theorem, we can say that the loss of the minimizer of $\hat{\rho}^2(m)$ with complexity term is upperbounded by $\min_{m \in \mathcal{M}_n}\{\rho^2(m)(1+\delta_m)^2\}$.

1.7 Conclusion

We consider a model selection using an estimator $\hat{\rho}^2$ and see their performance in overall sense when we allow a different error for different candidate models. In section 1.5, we manually determined ϵ. Motivated from that, we allow a different error as a function of complexity. Using an upperbound to a tail probability for the distance between out-of-sample prediction and its estimate, refined model selection is constructed. The loss of the minimizer of the criteria were upperbounded by the minimum loss plus small error term, which can be controlled by some fixed parameter values, complexity and the sample size. We can further explore the criterion by looking for simpler form of error term that can quantify the risk bound.

Chapter 2

Sparse superposition codes for the Gaussian channel with approximate iterative optimal estimates

2.1 Introduction

Sparse superposition codes for additive white Gaussian noise (AWGN) channel were developed in Joseph and Barron (2012). For a sparse superposition code the message is carried by choosing L non-zero terms out of N choices where L/N would be a small fraction of the dictionary size. Particularly, we use partitioned codes, where we split the dictionary into L sections with a section size $M = N/L$ with one term from each section chosen to be non-zero. There are M^L choices of codewords.

With a power of 2, the input bits would be $u_1, u_2, \ldots, u_k$ with $K = L \log M$. The dictionary consisted of vectors $X_1, X_2, \ldots, X_N$ each contains n coordinates of independent standard normal random variables. The codeword is a linear combination of those vectors chosen by the input bit strings. Thus the codeword takes the form of $X\beta$ where X would be a $n \times N$ matrix filled with standard normal random variables and β is a length N vector partitioned into L sections and only one term is chosen to be non-zero in each section. The power allocation in section ℓ is denoted by P_ℓ with $\sum_\ell P_\ell = P$ where P is a power constraint. Thus, we have the coefficient value of β for the term sent in section ℓ as $\sqrt{P_\ell}$ and zero for others so that we have the power constraint $\|\beta\|^2 = P$. From the channel with additive white Gaussian noise, what we receive is $Y = X\beta + \epsilon$ with n coordinates and ϵ is a noise vector where each element is an independent Gaussian random variable with mean zero and variance σ^2.

The bits per transmission of the code is the rate $R = K/n$ and the supremum of all achievable rates is the capacity which for the AWGN channel is $\mathcal{C} = (1/2)\log(1+snr)$ where $snr = P/\sigma^2$ is the signal-to-noise ratio. Any rate less than the capacity there are codes with arbitrary small error probability for sufficiently large blocklength n (See e.g. Cover and Thomas (2012)). Various researchers put an effort on developing practically efficient codes that approach the Shannon capacity. Polar codes in Arikan (2009) and Arikan and Telatar (2009) were the first feasible code that achieved the capacity using binary input channel with low encoding and decoding complexities. The error bound that they had is in an order of square root of the blocklength n. The polar codes schemes have been adapted to Gaussian channel in Abbe and Barron (2011).

The sparse superposition codes with partitioning was developed in Joseph and Barron (2012), showing that with maximum likelihood decoder we achieve exponentially small error probability for any rate less than capacity. However, as such decoder is computationally infeasible, Joseph and Barron (2014) developed an adaptive successive decoder which is fast and reliable for the Gaussian noise channel with arbitrary fixed rate below capacity and error probability proven to be exponentially small. See Barron and Joseph (2010) for more conclusions and discussions.

The setting that we are considering can be also seen as a high dimensional linear model with a sparse signal. The difference would be that the coefficient vector β is partitioned and we know the exact non-zero values. This is not usually assumed in problems in searching for a sparse solution for high dimensional regression models. However, some works for signal recovery problem involves the inner products of each X_j with the residuals at each step. For instance, Maleki and Donoho (2010) and Donoho et al. (2009) uses this type of quantity for updating the algorithm using soft and hard thresholding. This quantity is also related to the statistics that the adaptive successive decoder uses in Joseph and Barron (2014) and our work.

The conditional distribution of such statistics is approximately normal distributed with a shift for the terms that are sent. The shift is related to the amount we successfully decoded at the current step. It can be also interpreted as a signal to noise plus interference ratio as we view the remaining amount to decode as interference. The decoder assigns the non-zero coefficient, which is square root of the power allocation for the section when the test statistics for the term is above a threshold. The threshold is chosen to be high enough to avoid false alarms at each step.

In this paper, we are motivated by the same type of the statistics but we consider the Bayes optimal coefficient estimates based on the distribution of desired form of iteratively obtained statistics. These estimates are motivated by computing posterior probability of the term j is sent with a uniform prior on the choice of the terms that are sent. It provides the soft decision decoder with weights for each section rather than the $\{0, 1\}$ valued weights associated with the thresholding in the previously studied decoder (Joseph and Barron, 2014).

In the first following section, we explain how the statistics are motivated and formulated. We show the analysis of their distribution and the desired form of the statistics. The method of nearby measure enables us to work with a convenient distribution that is not far from the true distribution. In Section 3, we introduce Bayes optimal estimates with the given statistics. We show an identity relating the expected posterior error probability with the expected square distance of the estimate from the truth. The next section examines the progression of the decoder based on the Bayes optimal estimates introduced in Section 3. We provide an update function $g_L(x)$ which evaluates the expected fraction of success rate on a step when the previous one was x. In Section 5, we show two ways to construct an estimate that is approximately Bayes optimal so that the performance of the constructed estimate is not far from the theoretical performances we have seen from Section 4. In Section 6, we evaluate the final performance of the decoder. Numerical simulations reveal that the performance of the soft decoder is higher than that of the threshold based method. We conclude with a discussion in the last section.

2.2 Framework for the Decoder and Its Analysis

Define a set of indices of the terms from the dictionary chosen for the codeword sent across the channel as $\{j_1, j_2, \ldots, j_L\}$ and suppose that the decoder develops a sequence of estimates $\hat{\beta}_k$ of the true coefficient vector β.

The initial estimate for the first step is based on $stat_{0,j} = \mathcal{Z}_{0,j} = X_j^T Y / \|Y\|$. Its distribution is found to be approximately that of a standard normal shifted by $\beta_j \sqrt{n/(\sigma^2 + P)}$. The P in the denominator is from the presence of the terms in the codeword and the fact that nothing has been decoded yet. Those not decoded act like noise to this initial statistic so that this shift can be interpreted as a signal to noise plus interference ratio. The iteratively updated statistics we form successively reduce the interference so that the amount of shift for the true terms increases compared to the others.

We construct an estimate $\hat{\beta}_k$ as a function of statistics $stat_{k-1} = (stat_{k-1,j}, j \in J)$ which is computed by the information of the previous step. For example, $stat_{k-1}$ could be a function of $X^T(Y - X\hat{\beta}_{k-1}) + n\hat{\beta}_{k-1}$ or closely related $stat_{k-1,j}$ could be $X_j^T(Y - X\hat{\beta}_{k-1,-j})$ where the $-j$ refers to the fit with the jth term removed, so that $X\hat{\beta}_{k-1,-j}$ provides the removal of the interference of the current fit. For notation, we start with $G_0 = Y$. 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 that the current fit $X\hat{\beta}_k$ is not in the linear span of the previous such fits, so that $\|G_k\| > 0$. Let $\mathcal{Z}_{k,j} = X_j^T G_k / \|G_k\|$ be the normalized inner product of X_j and G_k. With the vector $stat_k$ a function of $\mathcal{F}_k = (\mathcal{Z}_0, \|G_0\|, \ldots, \mathcal{Z}_k, \|G_k\|)$, our first lemma analyzes the conditional distribution of $\mathcal{Z}_k$ and $\|G_k\|$ given $\mathcal{F}_{k-1}$. For $k = 0$, it is an unconditional distribution of $\mathcal{Z}_0$ and $\|G_0\|$.

For analysis purposes, we consider an extended version of the true coefficient vector as $\beta_e = (\beta, \sigma)$. This comes from the representation of $Y = X\beta + \epsilon$ as $Y = [X : \epsilon/\sigma]^T \beta_e$. For the estimates, we append an extra coordinate of value 0 and denote $\hat{\beta}_{k,e}$. The subscript e denotes that the vectors are extended.

Parallel to the sequence of G_k as orthogonal components of the fits $X\hat{\beta}_k$, we have $b_{0,e}, b_{1,e}, \ldots, b_{k,e}$ as a sequence of vectors in R^{N+1} that is obtained by successive Gram-Schmidt orthonormalization of the vectors $\beta_e, \hat{\beta}_{1,e}, \ldots, \hat{\beta}_{k,e}$. As we did for $\hat{\beta}_k$, let $b_0, \ldots, b_k$ be the vectors in R^N obtained by dropping the last coordinate from $b_{0,e}, b_{1,e}, \ldots, b_{k,e}$.

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 $\beta_e, \hat{\beta}_{1,e}, \ldots, \hat{\beta}_{k,e}$. The upper left $N \times N$

portion of this matrix denoted Σ_k plays the role of a conditional covariance matrix below. We work with the extension because of the usefulness of its projection interpretation. This is suggested by our colleague Antony Joseph who credits Bayati and Montanari (2011) and Bayati and Montanari (2012) for some analogous thinking.

Lemma 4 generalizes the conclusion from the corresponding Lemma in Barron and Joseph (2010); Joseph and Barron (2014) to handle the present generality.

Lemma 4. *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*

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

where $Z_k^{red} = (Z_{k,j}^{red} : j \in J)$ has conditional distribution Normal$(0, \Sigma_k)$.Here $\sigma_0^2 = \sigma_Y^2 = \sigma^2 + P$ 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 Chi-square$(n-k)$ random variable independent of the Z_k^{red} and the past $\mathcal{F}_{k-1}$.

The detailed proof is in Appendix B.1. The superscript *red*, an abbreviation of reduced, refers to the fact that Σ_k is of rank $N-k$ rather than full rank N. The shift b_k plays the key role as we combine the components $\mathcal{Z}_k$ while Z_k^{red} has no component of the true coefficient β_e nor the estimates.

We use the method of nearby measure here to use an approximating distribution rather than the true distribution. If the approximating distribution is not far from the true distribution in some sense, an event exponentially unlikely in the approximating distribution is also exponentially unlikely in the true distribution. We approximate the distribution of $\mathcal{Z}_k$ to a simpler distribution to analyze in two ways.

First, we will relate the Normal$(0, \Sigma_k)$ distribution $\mathbb{P}_{Z_k^{red}|\mathcal{F}_{k-1}}$ to $\mathbb{Q}_{Z_k^{red}|\mathcal{F}_{k-1}}$ which makes the Z_k^{red} have the Normal$(0, I - Proj_k)$ distribution. The $Proj_k$ is the matrix of projection onto the linear span of the estimates $\hat{\beta}_1, \ldots, \hat{\beta}_k$, as well as $Proj_{k,e}$

appending 0 to each of these estimates. $\Sigma_{k,e}$ differs from $I - Proj_{k,e}$ because of the orthogonality to β_e. We consider a random vector

$$\mathcal{Z}_k^{clean} = \mathcal{Z}_k \,+\, Proj_k\,\tilde{Z}_k$$

obtained by adding $Proj_k\,\tilde{Z}_k$, where the $\tilde{Z}_k$ are auxiliary independent standard normal vectors provided to the sample space for $\mathbb{P}$ and $\mathbb{Q}$. Then with respect to $\mathbb{Q}$, given $\mathcal{F}_{k-1}$, the $\mathcal{Z}_k^{clean}$ have the representation

$$b_k \mathcal{X}_{n-k} + Z_k,$$

with the Z_k distributed Normal$(0, I)$. One may think it is unfortunate to add independent normals in forming the $\mathcal{Z}_k^{clean}$, but by this representation it will considerably simplify the analysis.

Another idea is that the Chi random variable χ_{n-k} divided by n is concentrated around the constant 1. The expected square of $(\mathcal{X}_{n-k} - \sqrt{n})$ is bounded by a constant as long as the number of steps k is small compared to n. We approximate the distribution of $\mathcal{Z}_k$ given $\mathcal{F}_{k-1}$ further where the shift is $\sqrt{n}\,b_k$ rather than $\mathcal{X}_{n-k}b_k$.

Thus $\mathcal{Z}_k^{clean}$ is approximately $\sqrt{n}\,b_k + Z_k$, a normal shifted by $\sqrt{n}\,b_k$. Equip $\mathbb{Q}$, like $\mathbb{P}$, with the independent chi-square distribution for the $\mathcal{X}_{n-k}^2 = \|G_k\|^2/\sigma_k^2$. This approximation permits the replacement of the distribution with one that provides for independence, when determining events that have exponentially small probability. Also, certain combinations of these $\mathcal{Z}_k^{clean}$ are found to have nearly constant shifts so that the unconditional distribution of the $\mathcal{Z}_k^{comb}$ is approximated by that of a shifted normal.

The following lemma reveals how much penalty we need to pay for using the approximating distribution.

Lemma 5. *For any event A that is determined by the random variables,*

$$\|G_{k'}\| \text{ and } \mathcal{Z}_{k'} \text{ for } k' = 0, \ldots, k$$

we have

$$\mathbb{P}A \leq (\mathbb{Q}Ae^{k(2+k^2/n+C)})^{1/2}$$

The discussion on the method of nearby measure technique in the Appendix B.2 and the detailed proof is on the Appendix B.3. We use the Renyi relative entropy from the true distribution to its nearby distribution to relate the probability of an event as such. If an event A is exponentially unlikely under the nearby measure and the Renyi relative entropy is bounded by a constant or an amount of a smaller order than the exponent of the tail probability under the $\mathbb{Q}$, then we can say that it is also exponentially unlikely under the true distribution. Now, we can construct a decoder and analyze under the approximating distribution represented by $\mathcal{Z}_k^{clean} = \sqrt{n}b_k + Z_k$ where Z_k is a independent standard normal random variable.

We motivate particular forms of combinations of these components to produce our statistics $stat_k$. Initial motivation comes from the statistics $(Y - X\hat{\beta}_k)^T X_j + \|X_j\|^2 \hat{\beta}_{k,j}$ which is equal to $(Y - X\hat{\beta}_{k,-j})^T X_j$. We also find a motivation by combining the $\mathcal{Z}_k$ in a way to maximize the shift for the true term compared to others. The $stat_k$ take the following form, for some choice of vector $\underline{\lambda}_k = (\lambda_{k,0}, \lambda_{k,1}, \ldots, \lambda_{k,k})$ with unit square norm and some c_k typically between σ^2 and $\sigma^2 + P$,

$$stat_k = \mathcal{Z}_k^{comb} + \frac{\sqrt{n}}{\sqrt{c_k}} \hat{\beta}_k \tag{2.1}$$

where

$$\mathcal{Z}_k^{comb} = (\lambda_{k,0}\mathcal{Z}_0 + \lambda_{k,1}\mathcal{Z}_1 + \ldots + \lambda_{k,k}\mathcal{Z}_k).$$

This will be the definition of $stat_k$ as a function of the quantities computed by the decoder. Each $\mathcal{Z}_{k'}$ for $k' = 0, \ldots, k$ can be replaced with $\mathcal{Z}_{k'}^{clean}$. The desired representation of the statistics is as following

$$Z_k^{comb} + \frac{\sqrt{n}}{\sqrt{c_k}}\beta \tag{2.2}$$

with the desired shift $\frac{\sqrt{n}}{\sqrt{c_k}}\beta$. The two representations (2.1) and (2.2) look similar. The Eq. (2.1) provides the definition of $stat_k$ while the second representation (2.2) is a desired distributional characterization. This representation only holds for certain choices of $(\lambda_{k,0}, \lambda_{k,1}, \ldots, \lambda_{k,k})$ and $\hat{\beta}_k$. In some cases, this distributional form only holds approximately.

Notice that these statistics have non-zero shift only for the terms that are sent and the amount of shift represents the signal to noise plus interference ratio. Define the shift factor $\alpha_{\ell,k} = \sqrt{P_\ell\, n/c_k}$ where c_k quantifies the remaining noise plus interference. Then the shift of the desired representation in (2.2) takes the form $\alpha_{\ell,k}\, 1_{\{j=j_\ell\}}$. The c_k can take various forms for example $c_k = \sigma^2 + (1 - x_k)P$ where x_k measures the fraction of success at step k. The $(1 - x_k)P$ quantifies the remaining interference due to the inaccuracy of $\hat{\beta}_k$.

Here are related examples of such statistics. The first example has the form of the first motivation that we discussed. If we combine $\mathcal{Z}_{k'}$ with $\underline{\lambda}_k$ proportional to

$$\left(\|Y\| - \mathcal{Z}_0^T\hat{\beta}_k, -\mathcal{Z}_1^T\hat{\beta}_k, \ldots, -\mathcal{Z}_k^T\hat{\beta}_k\right),$$

then we have

$$stat_k = \frac{X^T(Y - X\hat{\beta}_k)}{\sqrt{\|Y - X\hat{\beta}_k\|^2}} + \frac{\sqrt{n}}{\sqrt{\|Y - X\hat{\beta}_k\|^2/n}}\hat{\beta}_k.$$

This first example shows a form of motivation, but it is hard to analyze the distribution of such statistics.

Another example uses a similar weights of combination where we take the inner product of $\hat{\beta}_k$ only with the shift part of $\mathcal{Z}_k$. In this example we take the weights of combination $\underline{\lambda}_k$ proportional to

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

If we combine $\mathcal{Z}_k^{clean}$ with these weights, then under the appropriate approximating distribution, we can produce the desired distributional relationship

$$stat_k = Z_k^{comb} + \frac{\sqrt{n}}{\sqrt{\hat{c}_k}}\beta.$$

as we desired with $\hat{c}_k = \sigma^2 + \|\beta - \hat{\beta}_k\|^2$. We call these weights of combination oracle weights.

It has the desired representation in its distribution, but since we do not know β in advance, we cannot calculate such weights of combination. Although we cannot directly calculate such weights of combination, under the appropriate approximating distribution and the successive decoding steps, we figure out how to find substitutes for those weights of combination.

2.3 Iteratively Optimal Statistics

Concerning the choice of the updated coefficient estimates $\hat{\beta}_{k+1}$, fundamental to our reasoning is the use of the approximating distribution that the $stat_{k,j}$ be independent Normal$\big(\alpha_\ell 1_{\{j=j_\ell\}}, 1\big)$, for j in any section ℓ, where $\alpha_\ell = \alpha_\ell(x_k)$. Here, our choice of c_k would be $\sigma^2 + (1 - x_k)P$ where x_k is the expected fraction of success which will be described more later in this section. Denote $\phi(s)$ as the standard normal density.

The density ratio $\phi(s-\mu)/\phi(s)$ is proportional to $e^{\mu s}$. With the term j_ℓ chosen according to a uniform prior over the M choices in each section ℓ, the posterior distribution of j_ℓ is

$$Prob\{j_\ell = j | stat_k\} = w_{k+1,j} = \frac{e^{\alpha_\ell stat_{k,j}}}{\sum_{j' \in sec_\ell} e^{\alpha_\ell stat_{k,j'}}}.$$

Furthermore, each element of the coefficient vector β is $\sqrt{P_\ell}\, 1_{\{j=j_\ell\}}$ for j in sec_ℓ. Accordingly, the posterior mean of β_j provides the Bayes estimator

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

This is the form of the estimate that we will use as a adaptive successive decoder with a soft decision. At the final step, for each section, we decode the term with the highest weight as the term sent.

When our $stat_k$ is exactly distributed $Normal(\sqrt{n/c_k}\,\beta, I)$, it can be interpreted as Bayes optimal estimates. The inner product $\beta^T \hat{\beta}_k / P$ can be interpreted as a posterior success rate since it takes the form $\sum_{\ell=1}^{L} (P_\ell / P) w_{k,j_\ell}$, with a power-weighted average across the sections.

Lemma 6. *The posterior success rate $\beta^T \hat{\beta}_k / P$ has the same expectation as the squared norm $\|\hat{\beta}_k\|^2 / P$. Consequently, the posterior error rate given by $\sum_{\ell=1}^{L} P_\ell (1 - w_{k,j_\ell})$ has the same expectation as the squared distance $\|\hat{\beta}_k - \beta\|^2$.*

Proof of Lemma 6: The random variables we are dealing with here are sums across the sections. So, we show that the quantities share their expectation for a fixed section ℓ and for a fixed step k. Also, the expected value will not change no matter which terms j_ℓ was sent. Thus the expectation taken conditionally on any realization would be the same if we take average with respect to the uniform prior on j_ℓ.

Let $P_j = P_{stat|j_\ell=j}$ be the conditional distributions of *stat* in section ℓ and $P = (1/M)\sum_{j\in sec_\ell} P_{stat|j_\ell=j}$ be the marginal distribution. We also denote the expectation $\mathbb{E}_j$ and $\mathbb{E}$ correspondingly. We use the fact that the likelihood ration of P_j and P is Mw_j. Set $j=1$. If we calculate the expectation $\mathbb{E}_1[w_{k,1}]$ using the measure P rather than P_1, then we get $M\mathbb{E}[w_{k,1}^2]$. By symmetry, $\mathbb{E}[w_{k,j}^2]$ is same across all j so that we have $M\mathbb{E}[w_{k,1}^2] = \mathbb{E}[\sum_{j\in sec_\ell} w_{k,j}^2]$ which is an average over particular realization $(1/M)\sum_{j\in sec_\ell}\mathbb{E}_j[\|w\|^2]$. Each term in the summation is the same so it is $\mathbb{E}_1[\|w\|^2]$. Thus, we can conclude that the weights for the term sent and the square norm of the weights in a fixed section shares their expectation. This completes the proof.

From the above identity, the square norm of the estimates in each step can be used as an estimate for the posterior success rate.

2.4 Update Function and Its Analysis

We evaluate the progression of the decoder with a recursive update rule. If the current fraction of success is x_k then we want to measure the expected fraction of success x_{k+1} as a function of x_k.

In previous section, we introduced Bayes optimal estimates with $stat_{k,j}$ being independent Normal$\big(\alpha_\ell 1_{\{j=j_\ell\}}, 1\big)$, for j in any section ℓ, where $\alpha_\ell = \alpha_\ell(x_k)$. We can see that the expected progression of the estimates based on the $stat_k$ depends on α_ℓ. Given power allocation, the progression only depends on the current success rate x_k. Thus we can write the expected success rate for the next step

$$x_{k+1} = \mathbb{E}\sum_{\ell=1}^{L}(P_\ell/P)w_{k+1,j_\ell}$$

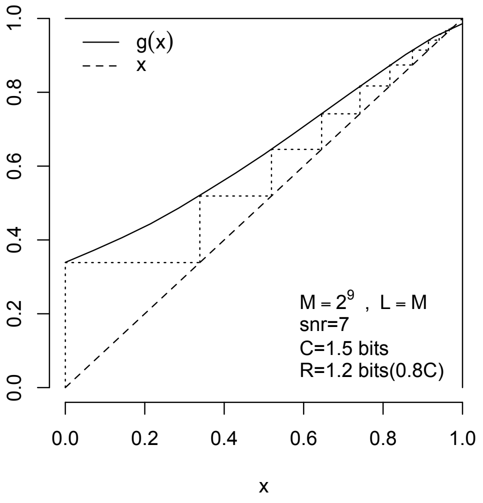

Figure 2.1: Plot of $g_L(x)$ and the sequence x_k.

as a function of x_k recursively.

Accordingly, we define the idealized update function as an expected success rate given that the previous success rate is x,

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

where $\alpha_\ell(x) = \sqrt{\frac{nP_\ell}{\sigma^2+P(1-x)}}$. Using this representation, we can write $x_{k+1} = g(x_k)$.

Fig 2.4 shows one example of an update function for given parameters. In the figure, the dotted line indicates the recursive update rule for the decoder where

$$x_{k+1} = g(x_k)$$

so that we can get a sequence of success rate $x_0, \ldots, x_k$. we will call this sequence as theoretical success rate whereas we define an empirical success rate as

$$\hat{x}_k = \sum_{\ell}^{L} (P_\ell/P) \hat{w}_{k,j_\ell}$$

for any estimate for β.

As long as $g_L(x)$ stays above the x, that is $g_L(x)$ stays above the 45 degree, there is a chance of a new update. If we can believe that the empirical success rate for our estimate is not far from the theoretical ones, examining the first crossing point of the function $g_L(x)$ with the 45 degree line can be a measure of how successful our decoder is. In this section, we will examine the update function to confirm that $g_L(x)$ stays above the 45 degree line in an interval $[0, x^*]$ and how close x^* to one.

2.4.1 Alternative Representation for Update Function

Define $g(u(\ell), x)$ as an expected weight for the true term in section ℓ

$$g(u(\ell), x) = g(\alpha = \alpha(u(\ell), x)) = \mathbb{E}[\frac{e^{\alpha^2+\alpha Z_1}}{e^{\alpha^2+\alpha Z_1} + \sum_{j=2}^{M} e^{\alpha Z_j}}]$$

where

$$\alpha = \alpha(u(\ell), x) = \tau\sqrt{\frac{(1 + 1/snr - u(\ell))\tilde{C}/R}{1 + 1/snr - x}}$$

with $u(\ell) = \frac{1-e^{-2C(\ell-1)/L}}{1-e^{-2C}}$ and $\tilde{C} = L(1 - e^{-2C/L})/2$. The $\tilde{C}$ comes from the approximation of $2\tilde{C}/L = (1 - e^{-2C/L})$. The $u(\ell)$ is an increasing function of ℓ. Notice that $\alpha = \alpha(u(\ell), x)$ matches $\alpha_\ell = \sqrt{\frac{nP_\ell}{\sigma^2+P(1-x)}}$ for each section ℓ. We can write the update function $g_L(x)$ as weighted average of $g(u(\ell), x)$ which is an expected success rate for each section.

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

Using the reparameterizing u_ℓ and using the fact that the number of section L is large, we can write the update function as an expectation with respect to a uniform random variable U as following.

Lemma 7. *Suppose the power allocation is* $P_\ell \propto e^{-2C\ell/L}$ *and* $L >> 2C$*. Define* $g(x)$ *as an expectation of* $g(U, x)$ *where a uniform random variable* U*. Then the update function* $g_L(x)$ *can be approximated by* $g(x)$ *within an order of* $(1/L)$*. Furthermore, it satisfies the lowerbound*

$$g_L(x) \geq \frac{\tilde{C}}{C}\, g(x).$$

with $(1 - C/L) \leq \tilde{C}/C \leq 1$

The detailed proof is in Appendix B.4.1. We use the Riemann sums for the approximation and we change the variable to $u = u(t) = (1 - e^{-2Ct})/(1 - e^{-2C})$.

From the reparameterization to u_ℓ, we can interpret the progression plot in terms of the update function. Fig 2.2 is one example of the progression plot with the same parameters as in Fig 2.4. The progression plot represents the expected weight for the true term for each section for a given success rate x. So as x grows, we can see how the expected success rate for each section progresses. If we rescale the horizontal axis in $u(\ell)$, the area under the curve is approximately $g(x)$. For a fixed x, suppose we take a vertical line where $u = x$ in a progression plot for a given x. Since the rectangle area left to the line is x, we can compare the area of the rectangle and the area under the curve to see if $g(x)$ is greater than x meaning that there is a chance of a new update. As we can see in the figures, for a small x we have chance to have some help from the area under the curve where $u > x$ so that we have more chance to have $g(x)$ greater than x. However, for x near 1, it is harder to gain some area from the right side of the line since the plot is cut off at one.

There is another representation of the update function using logit choice probability.

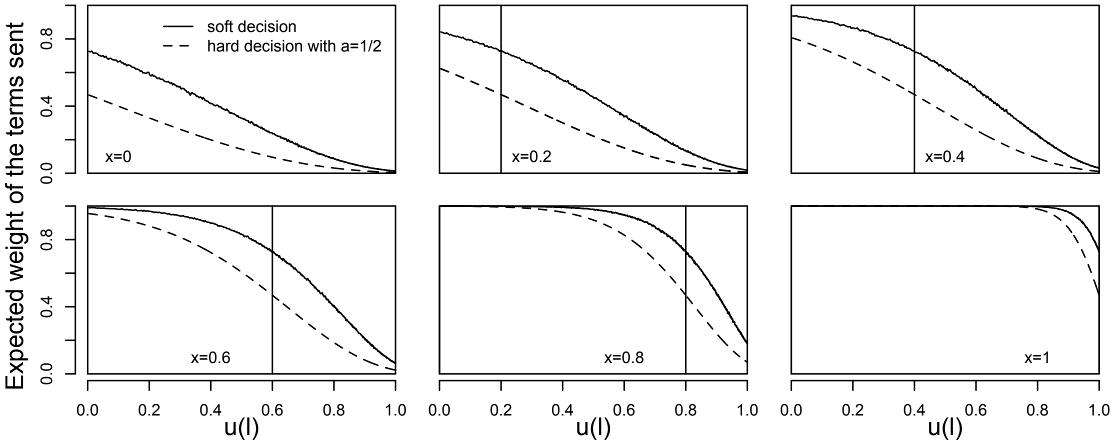

Figure 2.2: Progression Plots. $M = 2^9$, L=M, C=1.5 bits, R=0.8C and a=0.5. We used Monte Carlo simulation with replicate size 10000. The horizontal axis depicts $u(\ell)$. The vertical axis gives $g(u(\ell), x)$, the expected weight of the term sent for each section. This representation allows the area under the curve to be approximately $g(x)$. Also, the area of the rectangle to the left of the vertical bar is x. One can see if $g_L(x) \approx g(x)$ is above x by comparing the two areas.

Lemma 8 (Representation using the logit choice probability)**.** *Suppose*

$$\alpha = \alpha(U) = \tau\sqrt{\frac{(1 + 1/snr - U)\tilde{C}/R}{1 + 1/snr - x}}.$$

Suppose that Z_j for $j = 1, \ldots, m$ are independent standard normal random variables and υ_j for $j = 1, \ldots, m$ are independent Gumbel distributed random variables. We can express the update function as,

$$g(x) = \mathbb{P}_{Z_1,\ldots,Z_m,\upsilon_1,\ldots,\upsilon_m,U}\{\alpha^2 + \alpha Z_1 + \upsilon_1 \geq max_{2\leq j\leq m}(\alpha Z_j + \upsilon_j)\}.$$

It satisfies the lower bound

$$g(x) \geq \mathbb{P}\{\alpha \geq V_a\},$$

where

$$V_a = max_{2\leq j\leq m}\{-\frac{Z_1 - Z_j}{2} + \sqrt{[\upsilon_j - \upsilon_1 + \frac{(Z_1 - Z_j)^2}{4}]_+}\}.$$

This representation is shown in Appendix B.4.1 using McFadden and Zarembka (1974). Furthermore, we provide a simulation result and intuition why this representation can be an approximation of $g(x)$. We provide a similar type of representation which lowerbounds the update function and simpler to analyze.

2.4.2 Lowerbound for the Update Function

Using the Jensen's inequality, we provide a lower bound of $g(u(\ell), x)$ as well as $g(x)$.

Lemma 9. *Suppose $\alpha = \alpha(U) = \tau\sqrt{\frac{(1+1/snr-U)C/R}{1+1/snr-x}}$. Also, suppose that Z_1 is standard Normal random variable and ξ is logistic distributed independently. From the convexity of the function $1/(1+X)$, we have a lower bound for $g(u(\ell), x)$*

$$g(u(\ell), x) \geq \mathbb{E}_{Z_1}\left\{\frac{1}{1+e^{-\alpha^2/2+\tau^2/2-\alpha Z_1}}\right\}.$$

For the update function we have a similar form from Lemma 8,

$$g_L(x) \geq \frac{\tilde{C}}{C}\, g(x) \geq \frac{\tilde{C}}{C}\,\mathbb{P}\{\alpha \geq V_{low}\},$$

where

$$V_{low} = -Z_1 + \sqrt{(\tau^2 + 2\xi + Z_1^2)_+}.$$

Proof of Lemma 9. From the convexity of $1/(1+X)$, we can take expectation on $\sum_{j=2}^{M} e^{\alpha Z_j}$ to get a lower bound as following,

$$g(u(\ell), x) \geq \mathbb{E}_{Z_1}\{\frac{e^{\alpha^2+\alpha Z_1}}{e^{\alpha^2+\alpha Z_1}+(m-1)e^{\alpha^2/2}}\} \geq \mathbb{E}_{Z_1}\{\frac{1}{1+e^{-\alpha^2/2+\tau^2/2-\alpha Z_1}}\}.$$

The above representation is also interpreted as $\mathbb{E}_{Z_1,\xi}\{\xi \leq \alpha^2/2 - \tau^2/2 + \alpha Z_1\}$, where ξ is logistic distributed random variable of which the distribution function is

$1/(1+e^{-\xi})$. Also, we have a corresponding lowerbound for $g(x)$ using lemma 7 as $\mathbb{E}_{Z_1,\xi,U}\{\xi \le \alpha^2/2 - \tau^2/2 + \alpha Z_1\}$ for $\alpha = \alpha(U)$.

Similar to Lemma 8, we can rearrange the inequality in terms of α. Note that

$$\begin{aligned}
& \{\xi \le \alpha^2/2 - \tau^2/2 + \alpha Z_1\} \\
= \ & \{(\alpha + Z_1)^2 \ge (\tau^2 + 2\xi + Z_1^2)_+\} \\
= \ & \left\{\alpha \ge -Z_1 + \sqrt{(\tau^2 + 2\xi + Z_1^2)_+}\right\} + \left\{\alpha \le -Z_1 - \sqrt{(\tau^2 + 2\xi + Z_1^2)_+}\right\} \\
\ge \ & \left\{\alpha \ge -Z_1 + \sqrt{(\tau^2 + 2\xi + Z_1^2)_+}\right\}
\end{aligned}$$

so that

$$g(x) \ge \mathbb{P}\{\alpha \ge V_{low}\}.$$

This completes the proof of Lemma 9.

Suppose the actual decoder that we develop here progress close to the theoretical update rule. Then, by evaluating an interval where $g_L(x) > x$, we can simply measure the performance of the decoder. More analysis will be studied later in the paper relating the reliability of the decoder.

2.5 Approximate Optimal Statistics

We have seen how the progression we would expect from the update function and the performance that we can expect if our actual empirical success rate follows the update rule. Also, we have seen the distributional analysis on the orthogonal components of the given estimates and the motivation and goal of our statistics.

Next, we construct statistics by combining the orthogonal components to be close to the desired form $\sqrt{n/c_k}\,\beta + Z$. We introduce two methods to construct weights of combination to estimate the β. One is to use a deterministic weights of combination that is inversely related to the theoretical success rate. The other method is to recover oracle weights of combination using the advantage of the nearby measure.

We first state some preliminary lemmas that we use as tools for the proofs. Then we provide an alternative interpretation for the oracle weights. This alternative interpretation can be one of the motivations for the two methods of constructing weights of combination. Then we introduce two ways to combine the orthogonal components and we evaluate the reliability of the estimates by comparing the progression of the estimates to the theoretical update rule.

2.5.1 Preliminary

Here, we discuss some properties we need to prove the main results. We first state reliability of estimates when $stat_k$ is distributed $N(\alpha_{\ell,k}\, 1_{\{j=j_\ell\}}, 1)$. Also we study the tail probability of sum over the maximum of $Z_{k,j}$ in the section for each k, $\sum_\ell \max_{j \in sec_\ell} Z_{k,j}$. Finally, we provide an upperbound for the distance between two exponential weights using the difference between the exponents. The probability measure that we consider here is the approximating distribution $\mathbb{Q}$ rather than the true distribution

Lemma 10. *For any β, suppose we have deterministic x and deterministic $\underline{\lambda}$ with unit square norm and length $k \in \mathbb{N}$. We define $\alpha_\ell(x) = \sqrt{n\, P_\ell/(\sigma^2 + (1-x)P)}$ and $Z_j^{comb} = \sum_{k'=0}^{k} \lambda_{k'} Z_{k',j}$ which will be independent standard Normal distributed. We define, for $j \in sec_\ell$, jth element of $\beta^* = \beta^*(x, \underline{\lambda})$ as*

$$\sqrt{P_\ell}\, w_j^* = \sqrt{P_\ell}\, \frac{e^{\alpha_\ell(x)(\alpha_\ell(x) 1_{\{j=j_\ell\}} + Z_j^{comb})}}{\sum_{j' \in sec_\ell} e^{\alpha_\ell(x)(\alpha_\ell(x) 1_{\{j'=j_\ell\}} + Z_{j'}^{comb})}}.$$

Then the expectation of $\beta^T \beta^*$, $\|\beta^*\|^2$ are the same which is $g(x)P$ where P is the *power constraint. Also, they are close to their expectation with high probability. Indeed, if we define the event $A_{\beta,\delta}$ as*

$$A_{\beta,\delta} = \left\{ \left| \frac{\beta^T \beta^*}{P} - g(x) \right| > \delta \right\} \cup \left\{ \left| \frac{\|\beta - \beta^*\|^2}{P} - (1 - g(x)) \right| > \delta \right\}$$

then for any $\delta > 0$,

$$\mathbb{P}\{A_{\beta,\delta}\} \leq 4 \exp\left\{ -\frac{L}{2c^2} \delta^2 \right\}$$

where $c^2 = L \max(P_\ell / P)$ with value near $\frac{2C}{1-e^{-2C}}$ if we use the variable power allocation.

Proof for Lemma 10: We have already revealed in Lemma 6 that the success rate $\beta^T \beta^*$ and the square norm $\|\beta^*\|^2$ share their expectation. The independence across sections allow us to say that each quantity is close to its expectation $g(x)P$. The

$$\frac{\beta^T \beta^*}{P} = \sum_{\ell=1}^{L} \frac{P_\ell}{P} w^*_{j_\ell}$$

is a sum of bounded independent random variables. The sum of squares of the ranges of these random variables is $\sum_{\ell=1}^{L} \left(\frac{P_\ell}{P}\right)^2$. Likewise $\frac{\|\beta - \beta^*\|^2}{P}$ is also sum of bounded independent random variables where

$$\frac{\|\beta - \beta^*\|^2}{P} = \sum_{\ell=1}^{L} \frac{P_\ell}{P} \|e_{j_\ell} - w^*\|^2,$$

with e_{j_ℓ} is the vector of length M with 1 in position j_ℓ and 0 in the other entries. The sum of squares of this random variable is bounded by $4 \sum_{\ell=1}^{L} \left(\frac{P_\ell}{P}\right)^2$. Thus by Hoeffding's inequality, the probability that the distance of each quantity and the expectation is greater than δ is not more than

$$2\exp\left\{-\frac{2\delta^2}{4\sum_{\ell=1}^{L}(P_\ell/P)^2}\right\} \leq 2\exp\left\{-\frac{L\delta^2}{2c^2}\right\},$$

The union bound would be sum of the tail probability. This completes the proof.

Note that if we are using constant power allocation, c^2 will be 1 and if we are using variable power allocation with $P_\ell \propto e^{-2C\ell/L}$ then c^2 would be $\frac{L(1-e^{-2C/L})}{1-e^{-2C}}$ which is approximately $\frac{2C}{1-e^{-2C}}$.

Next, we evaluate the tail probability of the sum of maximum of independent normal random variables.

Lemma 11. *Let $Z_1, Z_2, \ldots, Z_N$ iid Normal random variables. The sum of the maximum of Z_js in each section $\ell = 1, 2, \ldots, L$ is less than $2L\sqrt{\log M}$ with high probability*

$$\mathbb{P}\left\{\sum_{\ell=1}^{L} \max_{j \in sec_\ell} Z_j > 2L\sqrt{\log M}\right\} \leq \exp\{-L\log M\}.$$

Proof of Lemma 11. For simplicity of the notation, we define $\delta = 2L\sqrt{\log M}$ and $D_\ell = max_{j\in sec_\ell} Z_{k,j}$. By the Cramer-Chernoff technique,

$$\mathbb{P}\{\sum_{\ell=1}^{L} \max Z_j > \delta\} \leq \inf_{t>0} \exp\{-\delta t + \log \mathbb{E}e^{t\sum D_\ell}\}$$

Using the independence of the normal random variables and the fact that the maximum is less than the sum,

$$\begin{aligned}
\mathbb{E}e^{t\sum D_\ell} &= \prod_{\ell=1}^{L} \mathbb{E}e^{t\max Z_{\ell,j}} = \prod_{\ell=1}^{L} \mathbb{E}\max e^{tZ_{\ell,j}} \\
&\leq \prod_{\ell=1}^{L}\sum_{j=1}^{M} \mathbb{E}e^{tZ_{\ell,j}} \\
&\leq (exp\{t^2/2 + \log M\})^L.
\end{aligned}$$

Then, the exponent of the probability bound becomes

$$-\delta t + L(t^2/2 + \log M)$$

and the infimum occurs when $t = \delta/L$. so that

$$\begin{aligned} \mathbb{P}\{\sum_{\ell=1}^{L} \max Z_j > \delta\} &\leq \inf_{t>0} \exp\{-\delta t + \log \mathbb{E} e^{t\sum D_\ell}\} \\ &= \exp\{-L(\delta^2/(2L^2) - \log M)\} = \exp\{-L \log M\} \end{aligned}$$

This completes the proof.

Corollary 4. *For each step $k = 0, \ldots, k^*$, we can upperbound $\sum_{\ell=1}^{L} max_{j \in sec_\ell} |Z_{k,j}|$ by $4L\sqrt{\log M}$ except an event of the probability not more than $2k^* \exp\{-L \log M\}$.*

This can be proved by the fact that $\max |Z_{k,j}| \leq \max Z_{k,j} + \max(-Z_{k,j})$ and the symmetry of the normal distribution. We use Lemma 11 and the error probability can be obtained by union bounds.

Next, we evaluate the distance between the two exponential weights using the difference of the exponents. This is one of the key tools for the proof. We evaluate the difference between the two estimates by comparing the statistics that we used for the estimates because the estimates are in forms of exponential weights.

Lemma 12. *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,*

$$\left| \sum_{j=1}^{M} \left(w_j^2 - (w_j^*)^2 \right) \right| \leq 4 \max_{j=1,\ldots,M} |\epsilon_j|$$

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

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

Alternatively, we have

$$|\sum_{j=1}^{M}(w_j^2 - 2w_{j_\ell} - (w_j^*)^2 + 2w^*_{j_\ell})| \leq 4 \max_{j=1,\ldots,M} |\epsilon_j|$$

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,

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

The detailed proof in the Appendix B.5. The key idea of the proof is to use first order Taylor expansion and use the fact that the sum of the weights is one.

2.5.2 Oracle Weights

Recall from that the oracle weight which is proportional to

$$\left((\sigma_Y - b_0^T\hat{\beta}_k), (-b_1^T\hat{\beta}_k), \ldots, (-b_k^T\hat{\beta}_k)\right)$$

approximately forms an idealized statistics where

$$stat_k = \frac{\sqrt{n}}{\sqrt{\hat{c}_k}}\beta + Z_k^{comb}$$

where

$$
\begin{aligned}
\hat{c}_k &= (\sigma_Y - b_0^T \hat{\beta}_k)^2 + \sum_{k'=1}^{k} (b_{k'}^T \hat{\beta}_k)^2 \\
&= \sigma_Y^2 - 2 b_0^T \hat{\beta}_k + \hat{\beta}_k^T \left(b_0 b_0^T + \cdots + b_k b_k^T \right) \hat{\beta}_k \\
&= \sigma^2 + \|\beta - \hat{\beta}_k\|^2 .
\end{aligned}
$$

It is because the shift part yields

$$
\begin{aligned}
& \frac{\sqrt{n}}{\sqrt{\hat{c}_k}} \left((\sigma_Y - b_0^T \hat{\beta}_k) b_0 - (b_1^T \hat{\beta}_k) b_1 - \cdots - (b_k^T \hat{\beta}_k) b_k + \hat{\beta}_k \right) \\
= & \frac{\sqrt{n}}{\sqrt{\hat{c}_k}} \left(\beta + \hat{\beta}_k - (b_0 b_0^T + \cdots + b_k b_k^T) \hat{\beta}_k \right) = \frac{\sqrt{n}}{\sqrt{\hat{c}_k}} \beta
\end{aligned}
$$

Since we cannot calculate these quantities, we want to estimate or replace with some values that we can actually compute.

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 $(N+1) \times (k+1)$. The oracle weight also arises from the QR decomposition of the matrix B.

Since $b_0, b_1, \ldots, b_k$ is Gram-Schmidt orthogonalization of columns of B, we can write B as

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

This is the QR decomposition of the matrix B. Note that the components of the oracles weights are actually the elements in the Cholesky factor matrix R. Moreover, if we consider the Cholesky decomposition of $B^T B$, we can write it as $R^T R$ as

$$\begin{bmatrix} \|\beta\|^2 & \beta^T\hat{\beta}_1 & \cdots & \beta^T\hat{\beta}_k \\ \hat{\beta}_1^T\beta & \|\hat{\beta}_1\|^2 & \cdots & \hat{\beta}_1^T\hat{\beta}_k \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\beta}_k^T\beta & \hat{\beta}_k^T\hat{\beta}_1 & \cdots & \|\hat{\beta}_k\|^2 \end{bmatrix} = R^T \begin{bmatrix} (b_0^T\beta) & (b_0^T\hat{\beta}_1) & \cdots & (b_0^T\hat{\beta}_k) \\ 0 & (b_1^T\hat{\beta}_1) & \cdots & (b_1^T\hat{\beta}_k) \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & (b_k^T\hat{\beta}_k) \end{bmatrix}$$

This representation of the oracle weights has a key role in the analysis of constructing the weights of combination. The deterministic method comes from the fact that the quantities on the right side are close to some deterministic values with high probability. The Cholesky decomposition based method comes from figuring out what elements we know from the data so that we can recover the Cholesky factor matrix R.

2.5.3 Deterministic Weights 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 update function $x_{k+1} = g_L(x_k)$.

The deterministic weights of combination is defined as

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

The approximate optimal estimates are defined as, for j in sec_ℓ,

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

where $\alpha_{\ell,k} = \sqrt{nP_\ell/(\sigma^2 + (1 - x_k)P)}$. This estimate has an advantage of a simple computation although we have been able to show it is reliable only when the number of steps is a constant not depending on the section size.

In order to show the reliability of the estimates using the deterministic weights of combination, we first introduce pseudo-statistics $stat_k^*$. These pseudo-statistics are distributed $Normal(\sqrt{n/c_k}\,\beta, I)$ and formulated by combining $\mathcal{Z}_{k'}^*$ which is defined by $\sqrt{n}\,b_{k'}^* + Z_{k'}$ with $Z_{k'}$ is the ingredient that we had from $\mathcal{Z}_k^{clean}$ and the vector $b_{k'}^*$ is defined as

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

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 of $b_{k'}$. Recall that b_k is, for $k \geq 1$, a part of the estimate $\hat{\beta}_k$ orthogonal to the previous estimates and to the 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. We relate these two components b_k and b_k^* 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, $b_{k'}^*$ as well as $\mathcal{Z}_k^*$ is not known from the received data. These are not actual statistics from the received data but rather they are approximations to $\mathcal{Z}_{k'}$. From these ingredients, define

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

We can see that $stat_k^* = \sqrt{n/c_k}\,\beta + Z_k^{comb}$ because, for the shift part,

$$\begin{aligned} &\sum_{k'=0}^{k} \lambda_{k',k}^* \sqrt{n} b_{k'} + \frac{\sqrt{n}}{\sqrt{c_k}}\hat{\beta}_k \\ = \;& \frac{\sqrt{n}}{\sqrt{c_k}}\left(\frac{c_k}{c_0}\beta - \sum_{k'=1}^{k} \frac{c_k}{\sqrt{c_{k'}}}\lambda_{k',k'}^* b_{k'}^* + \hat{\beta}_k\right) = \frac{\sqrt{n}}{\sqrt{c_k}}\,\beta. \end{aligned}$$

We use the fact that $\lambda^*_{k',k} = \sqrt{c_k/c_{k'}}\,\lambda^*_{k',k'}$. Define β^*_{k+1} as an estimate using $stat^*_k$. 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.

Lemma 13 (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 $\beta^*_{k+1} = \mathbb{E}\left[\beta|stat^*_k\right] = \mathbb{E}\left[\beta|\mathcal{F}^*_k\right]$ is the associated conditional mean of β given $\mathcal{F}^*_k$.*

We prove the 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

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

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, due to the independence between the sections, we see that the posterior distribution of β is independent across the sections with $\mathbb{Q}\left[j_\ell = j|\mathcal{F}^*_k\right]$ reducing to $\mathbb{Q}\left[j_\ell = j|stat^*_k\right] = w^*_{k+1,j}$ for j in section ℓ. Accordingly, $\mathbb{E}\left[\beta|\mathcal{F}^*_k\right]$ is equal to $\mathbb{E}\left[\beta|stat^*_k\right]$ which is β^*_{k+1} with coordinates $\beta^*_{k+1,j} = \sqrt{P_\ell}\,w^*_{k+1,j}$ for each j in section ℓ. This completes the proof of Lemma 13.

In the Bayes formulation, when we consider the expectations are with respect to the joint distribution of β and the statistics, using iterated expectation, we have, for

$k' > 1$,

$$\mathbb{E}\left[(\beta^*_{k+k'})^T \beta^*_k\right] = \mathbb{E}\left[(\beta^*_k)^T \mathbb{E}[\beta | \mathcal{F}^*_{k+k'-1}]\right] = \mathbb{E}\left[\beta^T \beta^*_k\right]$$

which is also same as $\mathbb{E}\left[\|\beta^*_k\|^2\right]$ and $x_k P$. Alternatively, if these expectations are computed conditionally on β then they are the same for every β.

The deterministic weights of combination we use also arise from the Cholesky decomposition that we discussed in the last section. If we replace the inner products among $(\beta, \hat{\beta}_1, \ldots, \hat{\beta}_k)$ with the deterministic values to which we believe that they are close. The values are their expectations when the $\hat{\beta}_k$ is the Bayes optimal estimates. Then the components of the Cholesky factor matrix of the replaced matrix will be filled with the elements of the deterministic weights of combination.

Motivated by the pseudo-statistics, we estimate β by constructing the statistics by combining the orthogonal components $\mathcal{Z}^{clean}_k$ using the deterministic weights of combination. In order to evaluate the performance of the estimates as in the update rule, we need to examine if $\mathcal{Z}^{clean}_k$ is close to $\mathcal{Z}^*_k$. Note that $\mathcal{Z}^{clean}_k = \sqrt{n}\, b_k + Z_k$ under the $\mathbb{Q}$ measure. We will see if b_k is close to b^*_k so that $\mathcal{Z}^{clean}_k$ is close to $\mathcal{Z}^*_k$.

Lemma 14. *For $k = 1, \ldots, k^*$ and for any $\eta > 0$, we define an event A_k*

$$A_k = \{|\beta^T \hat{\beta}_k / P - x_k| > a_k \eta\} \cup \{|\|\hat{\beta}_k\|^2 / P - x_k| > a_k \eta\}$$

and denote $A^k_1 = \cup^k_{k'=1} A_k$. Then we have

$$\mathbb{Q}\{A^k_1\} \leq \sum_{k'=1}^{k} 6(k'+1) \exp\{-\frac{2}{c^2} L \delta^2_{k'}\}$$

where $\delta_{k'} = (a_k/2)(n/L)^{k-1}\eta$ and $c^2 = L \max(P_\ell / P)$.

The detailed proof is in Appendix B.6. By the method of nearby measure, we already approximate the $\mathcal{Z}_k$ with a shift of $\sqrt{n}\, b_k$ with simple independent normal distribution. The key idea of the proof is to show b_k and b_k^* are close to each other as b_k^* is a simplified form of b_k. We use the idea that the inner product between any $\hat{\beta}_k$ with $\hat{\beta}_{k+k'}$ is close to the deterministic value $x_k P$ with high probability.

For any small $\eta^* > 0$, we have actual success rate $\beta^T \hat{\beta}_k$ to be η^* close to $x_k P$ except an event of probability bounded by $7k^* exp\{-\min(1/16, 2/c^2) L \eta^2\}$ where $\eta \sim (1/a_k)(\log M)^{-k^*+1/2}\eta^*$. If the number of steps is a constant that is not depending on the section size M, then we can choose L large enough than $(\log M)^{k^*}$ 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.

One of the advantages of the deterministic weights of combination is the simple computation. The weights of combination has a recursive relationship

$$\underline{\lambda}_k = \left((1 - \lambda_{k,k})\, \underline{\lambda}_{k-1}, \lambda_{k,k} \right).$$

with $\lambda_{k,k} = -\sqrt{1 - (c_k / c_{k-1})}$ so that we can compute the combination $\mathcal{Z}_k^{comb}$ using $\mathcal{Z}_{k-1}^{comb}$ and $\mathcal{Z}_k^{clean}$ where

$$\mathcal{Z}_k^{comb} = \sqrt{1 - \lambda_{k,k}^2}\, \mathcal{Z}_{k-1}^{comb} + \lambda_{k,k} \mathcal{Z}_k^{clean}.$$

The next method we provide somewhat more complicated calculation of weights of combination but with better reliability.

2.5.4 Cholesky Decomposition Based Method

Recall the Cholesky decomposition of the matrix B^TB.

$$\begin{bmatrix} \|\beta\|^2 & \beta^T\hat{\beta}_1 & \cdots & \beta^T\hat{\beta}_k \\ \hat{\beta}_1^T\beta & \|\hat{\beta}_1\|^2 & \cdots & \hat{\beta}_1^T\hat{\beta}_k \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\beta}_k^T\beta & \hat{\beta}_k^T\hat{\beta}_1 & \cdots & \|\hat{\beta}_k\|^2 \end{bmatrix} = R^T \begin{bmatrix} (b_0^T\beta) & (b_0^T\hat{\beta}_1) & \cdots & (b_0^T\hat{\beta}_k) \\ 0 & (b_1^T\hat{\beta}_1) & \cdots & (b_1^T\hat{\beta}_k) \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & (b_k^T\hat{\beta}_k) \end{bmatrix}$$

On the left side, we know the values of elements in B^TB with shaded region. If we know the diagonals in R which is depicted by a shaded region on the right side, we can recover the rest of the elements in the matrix R.

For each step k, suppose we know all $b_{k''}^T\hat{\beta}_{k'}$ for $0 \leq k'' \leq k' < k$ and $(b_k^T\hat{\beta}_k)$ exactly without any error. We want to recover $b_{k'}^T\hat{\beta}_k$ for $k' \leq k$. We can construct one linear system along with one quadratic equation from the Cholesky decomposition as following.

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

$$= \begin{bmatrix} (\hat{\beta}_1^T\hat{\beta}_k) \\ (\hat{\beta}_2^T\hat{\beta}_k) \\ \vdots \\ (\hat{\beta}_{k-1}^T\hat{\beta}_k) \end{bmatrix} - (b_0^T\hat{\beta}_k) \begin{bmatrix} (b_0^T\hat{\beta}_1) \\ (b_0^T\hat{\beta}_2) \\ \vdots \\ (b_0^T\hat{\beta}_{k-1}) \end{bmatrix} \tag{2.3}$$

and

$$(b_0^T\hat{\beta}_k)^2 + (b_1^T\hat{\beta}_k)^2 + \cdots + (b_{k-1}^T\hat{\beta}_k)^2 = \|\hat{\beta}_k\|^2 - (b_k^T\hat{\beta}_k)^2. \tag{2.4}$$

From the Eq.2.3, 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 $(b_0^T\hat{\beta}_k)$. We plug in the function to Eq.2.4 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 Eq. 2.4.

Under the $\mathbb{Q}$ measure, $\mathcal{Z}_k^T\hat{\beta}_k/\sqrt{n} = b_k^T\hat{\beta}_k$ so that we have a diagonal elements of the Cholesky factor matrix R. It is because we have a representation of $Z_k^{red} = (I - Proj_k)\tilde{Z}_k^{red}$ where $\tilde{Z}_k^{red}$ is some independent standard normal random variables. Since $\hat{\beta}_k$ is orthogonal to $Proj_k$, we have $b_k^T Z_k^{red}$ equal to zero

$$\mathcal{Z}_k^T\hat{\beta}_k/\sqrt{n} = \hat{\beta}_k^T(b_k + Z_k^{red}/\sqrt{n}) = b_k^T\hat{\beta}_k.$$

Using this information, we can recover the rest of the Cholesky factor matrix R, which would be same in distribution to the oracle weights of combination which we will denote $\underline{\hat{\lambda}}_k$

Next, we define our estimate $\hat{\beta}_k$. Using the weights we recovered, we combine $\mathcal{Z}_k^{clean}$ to construct the statistics

$$\begin{aligned} stat_k &= \sum_{k'=0}^{k}\hat{\lambda}_{k',k}\mathcal{Z}_{k'}^{clean} + \sqrt{\frac{n}{\hat{c}_k}}\hat{\beta}_k \\ &= \sqrt{\frac{n}{\hat{c}_k}}\beta + Z_k^{comb} \end{aligned}$$

where $\hat{c}_k = \sigma^2 + \|\beta - \hat{\beta}_k\|^2 = \sigma^2 + (1-\hat{x}_k)P$. Notice that $\hat{c}_k$ can be calculated from the Cholesky factor matrix since $\hat{c}_k = (\sigma_Y - b_0^T\hat{\beta}_k)^2 + \sum_{k'=1}^{k}(b_{k'}^T\hat{\beta}_k)^2$. This is the desired form except that we have a random value $\hat{c}_k$ and we combine the standard normals Z_k with a random weights. The jth component in sec_ℓ of $\hat{\beta}_{k+1}$ is defined by

$$\sqrt{P_\ell}\,\frac{e^{\hat{\alpha}_{\ell,k}stat_{k,j}}}{\sum_{j'\in sec_\ell} e^{\hat{\alpha}_{\ell,k}stat_{k,j'}}}$$

Now, we evaluate the reliability of the estimate $\hat{\beta}_k$. Define an event A_k

$$A_k = \{|\beta^T \hat{\beta}_k/P - x_k| > a_k\eta\} \cup \{|\|\beta - \hat{\beta}_k\|^2/P - (1 - x_k)| > a_k\eta\}.$$

We evaluate the reliability by looking at the probability of an event A_k under the $\mathbb{Q}$ measure.

Lemma 15. *Suppose we have a Lipschitz condition on the update function so that*

$$|g_L(x_1) - g_L(x_2)| \leq c_{Lip}|x_1 - x_2|.$$

For $k = 1, \ldots, k^*$, for some $a_k = 1 + c_{Lip}a_{k-1}$ and $a_1 = 1/2$, we define an event A_k

$$A_k = \{|\beta^T \hat{\beta}_k/P - x_k| > a_k\eta\} \cup \{|\|\beta - \hat{\beta}_k\|^2/P - (1 - x_k)| > a_k\eta\}$$

and denote $A_1^k = \cup_{k'=1}^k A_{k'}$. Then, we have

$$\mathbb{Q}\{A_1^k\} \leq \exp(-\frac{L}{8c^2}\eta^2 + \log(4\sum_{k'} Gr_{k'})) + 2k\exp(-L\log M)$$

where $c^2 = L\max_\ell(P_\ell/P)$ and $Gr_k = Const_k \left(\frac{n}{L\eta}\right)^{k+1}$. Notice that if $c_{Lip} \leq 1$, then $a_k \leq (k+1)$.

2.6 Performance of the Decoder

We have studied that the actual decoder reliably follows the theoretical update rule. Next, we examine the final performance of the decoder. We consider the Cholesky decomposition based method. Numerical simulation for the Cholesky decomposition based method shows that the progression follows the update rule of the theoretical

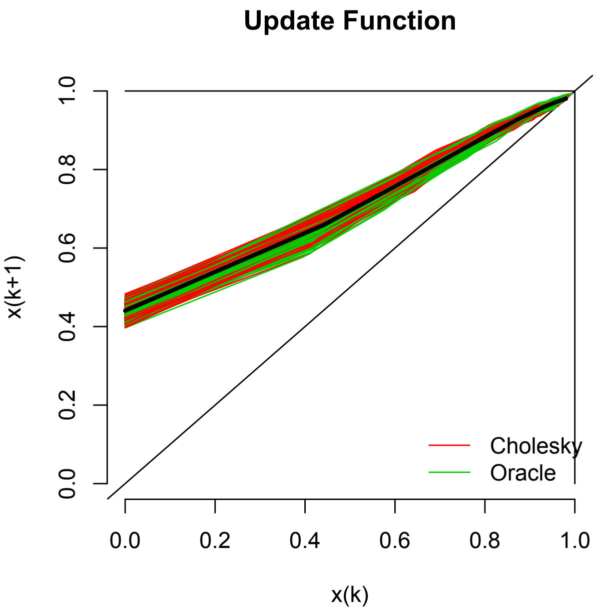

Figure 2.3: L = 512, M = 64, snr = 7, C = 1.5 bits, R = 0.7C, blocklength = 2926. Dark colored (Cholesky decomposition based weights); light colored (oracle weights of combination). A thick black line indicates the theoretical update function. Ran 1000 experiment for each. It is hard to distinguish the two method since the lines overlap each other.

update function $g_L(x)$ as in Fig. 2.3.

We examine the update function with the lowerbound using Jensen's inequality. The most of the argument works the same for the one that we discussed in Lemma 8. We first show that $g_{low}(x) - x$ is monotone non-increasing function. This will be a useful property since we can say that the number of crossing point between $g_{low}(x)$ and x is at most one. Once the monotonicity is established, showing that $g_{low}(x)$ is above x on an interval $[0, x^*]$ will be enough to check at the right end by proving $g_{low}(x^*) > x^*$.

Theorem 5. *Denote the Jensen's lower bound for $g(x)$ as $g_{low}(x)$. With $R \leq \tilde{C}/(1+2/\tau^2)$, the function $g_{low}(x) - x$ is a monotone decreasing function of x. Furthermore,* *if* $R \leq \tilde{C}/(1+drop^*)$ *where*

$$drop^* = \frac{\mathbb{E}((V_+)^2 - \tau^2)1_{B^*}}{\tau^2}$$

with $B^* = \{\tau\sqrt{\tilde{C}/R} \leq V_+ \leq \tau\sqrt{1+snr}\}$ *and* $\tau = \sqrt{2\log M}$*, then for*

$$1 - x^* = \frac{R}{\tilde{C}}\frac{drop^*}{snr}(1 - \frac{R}{\tilde{C}}drop^*)^{-1},$$

we have $g_{low}(x^*) > x^* - e$*. Here,* e *is a value polynomially small in the section size* M *and the* $drop^*$ *is an order of* $1/\tau$*.*

The detailed proof is in Appendix B.8. From above Theorem, we can learn that crossing point is more close to one when the ratio of the rate and the Capacity is small, signal-to-noise ratio is high and the section size M increases. The key idea of the proof is to represent the update function as

$$\begin{aligned}\mathbb{P}_{V,U}\{\alpha(U) \geq V\} &= \mathbb{P}\{\alpha^2 \geq (V_+)^2\} \\ &= \mathbb{P}_{V,U}\{U \leq 1 + \frac{1}{snr} - (1 + \frac{1}{snr} - x)\frac{(V_+)^2}{\tau^2}\frac{R}{\tilde{C}}\} \\ &= \mathbb{E}_V \max(1, (-\frac{(V_+)^2}{\tau^2}\frac{R}{\tilde{C}}(1 + \frac{1}{snr} - x) + 1 + \frac{1}{snr})_+).\end{aligned}$$

The crossing point x^* yields the target mistake rate $(1 - x^*)$. If we assign a non-zero coefficient to the term whose weight is the highest among the others in the section, we call it a mistake when the weight of the true term is not chosen as the highest. Notice that $1_{\{\hat{w}_{j_\ell} < \frac{1}{2}\}} \leq 2(1 - \hat{w}_{j_\ell})$. Thus, we have an empirical mistake rate

$\hat{e}_{mis}$

$$\begin{aligned}\hat{e}_{mis} &= \frac{1}{L}\sum_{\ell=1}^{L} 1_{\{\hat{w}_{j_\ell} \text{ is not the highest}\}} \leq \frac{1}{L}\sum_{\ell=1}^{L} 1_{\{\hat{w}_{j_\ell}<1/2\}} \\ &\leq \frac{1}{Lmin_\ell(P_\ell/P)}\sum_{\ell=1}^{L}\frac{P_\ell}{P}2(1-\hat{w}_{j_\ell}) \leq \frac{2}{Lmin_\ell(P_\ell/P)}(1-\hat{x}_{k^*})\end{aligned}$$

Notice that $\frac{2}{Lmin_\ell(P_\ell/P)}$ is approximately snr/C. Thus,

$$\hat{e}_{mis} \lesssim \frac{snr}{C}(1-\hat{x}^*)$$

If we set $\eta = (1-x^*)$ then we have

$$\hat{e}_{mis} \lesssim \frac{2snr}{C}(1-x^*)$$

except an event of probability that is an order of

$$\big(Const(\eta)\exp(-L\eta^2) + 2k^*\exp(-L\log M)\big)\exp(-k^*(k^* + (k^*)^2/n + C))$$

where $Const(\eta)$ $(1/\eta)^{k^*+1}$ and k^* indicates the total number of steps. This is exponentially small in $L(C-R)^2$. Suppose we have the rate approaching the capacity in order of $(1/\sqrt{\log M})$ where

$$R = C/(1+r/\tau)$$

where $r/\tau > drop^*$ as we specified in the previous lemma. Then we have shown that crossing point from one is approximately $\frac{1}{snr}\frac{r}{\tau}$.

Fig. 2.4 shows the update functions for the given parameters. The highest is the update function of the soft decision decoder. We can see from the plot the

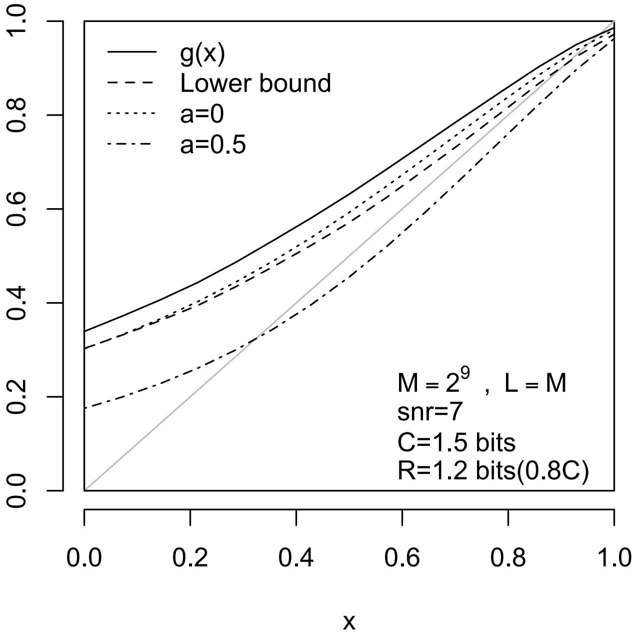

Figure 2.4: Comparison of update functions. The lines with a specified value of a indicates $\{0, 1\}$ decision using the threshold $\tau = \sqrt{2 \log M} + a$

lowerbound using Jensen's inequality is not far from the actual update function. It is much higher than the update function of the hard decision decoder when we set a realistic threshold with $a = 1/2$. For good enough performance, hard decision decoder requires much larger section size while soft decision decoder is successful at 80 percent of capacity with the smaller section size.

The next figures show the bit error rate and the block error rate for the decoder using the oracle weights with comparison to the hard decision decoder in Joseph and Barron (2014). For hard decision decoder, we use the inner product between the columns of X and the residuals for statistics and we do not declare error when there are more than one terms above the threshold within the section. Instead, we decode the term with the maximum statistics within the section to be the one that is sent. We have seen from the simulation this decoder reliably follows the update function

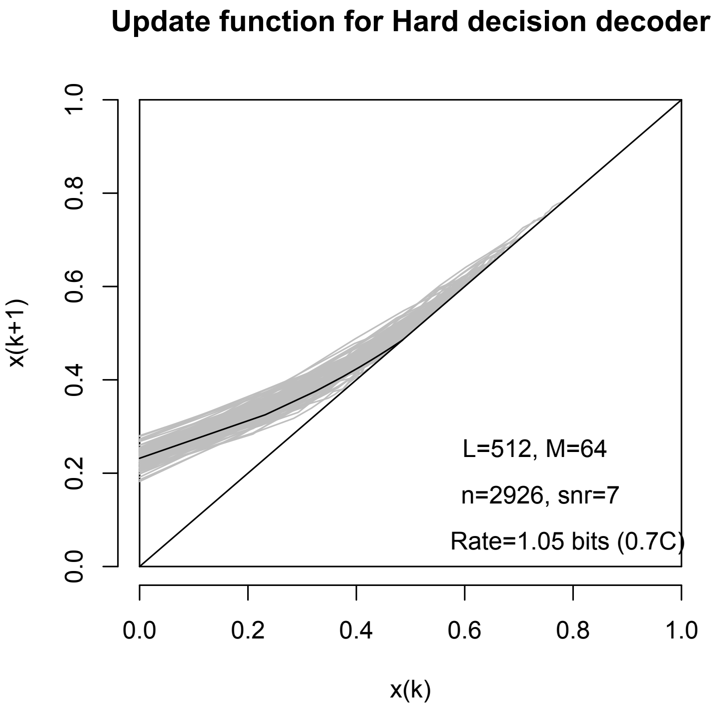

Figure 2.5: We can see that the decoder we used for the comparison follows the update rule well. Each gray line indicates one trial of the decoder and we ran 100 times.

as in Fig. 2.5.

We fix the rate and simulated the bit error rate and the block error rate as the signal to noise ratio increases. To fix the power allocation, we used variable power allocation proportional to $e^{-2C\ell/L}$ with $C = 1/2\log(1+7)$. As in Fig. 2.6, the bit error rate counts the number of sections that is not correctly decoded out of L sections. The log of the error probability drops almost linearly as the capacity increases. Also, we can see that the bit error rate improved a lot compared to the previously studied decoder in Joseph and Barron (2014).

The block error for the decoder occurs when $\hat{u}$ does not match with the input bit string u. The outer Reed-Solomon codes helps us to have a smaller block error probability. With $0 < \delta < 1$, a small enough mistake rate $2\hat{e}_{mis} < \delta$ can be corrected

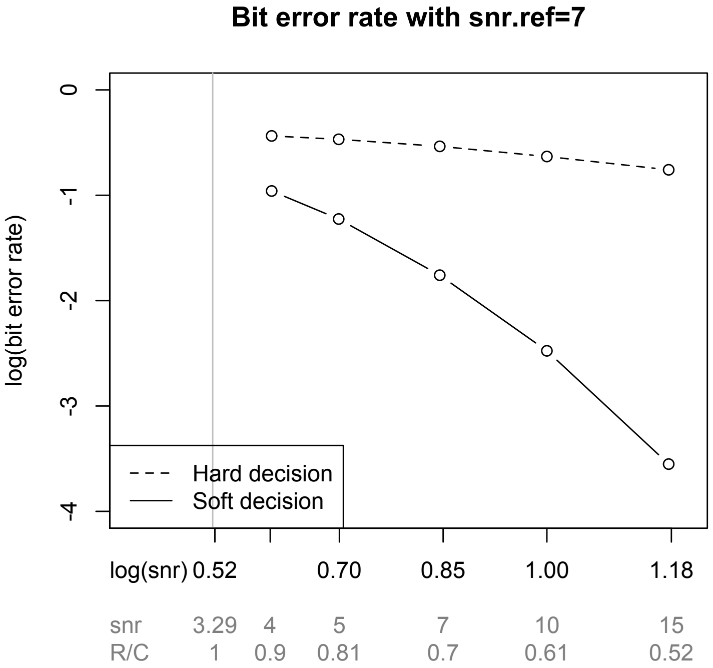

Figure 2.6: L = 512, M = 64, R = 1.05 bits, blocklength n = 2926, snr .ref = 7, snr = (4, 5, 7, 10, 15). Ran 10,000 trials. Average of error count out of 512 sections.

by outer Reed Solomon code. The Fig. 2.7 shows the block error rate for given δ specified in the figure. The total rate would be corrected by $R_{tot} = (1 - \delta)R$. Using the specified δ in the figure, the block error rate for hard decision decoder was one for all signal-to-noise ratio.

2.7 Conclusion

We developed the adaptive successive decoder with soft decision for additive white Gaussian noise channel. The soft decision decoder is motivated by the Bayes optimal estimates for a given statistics. The update function is provided for evaluating the iterative progression of the decoder as well as the final performance of the decoder in expectation.

We develop the approximate optimal estimate where the orthogonal components

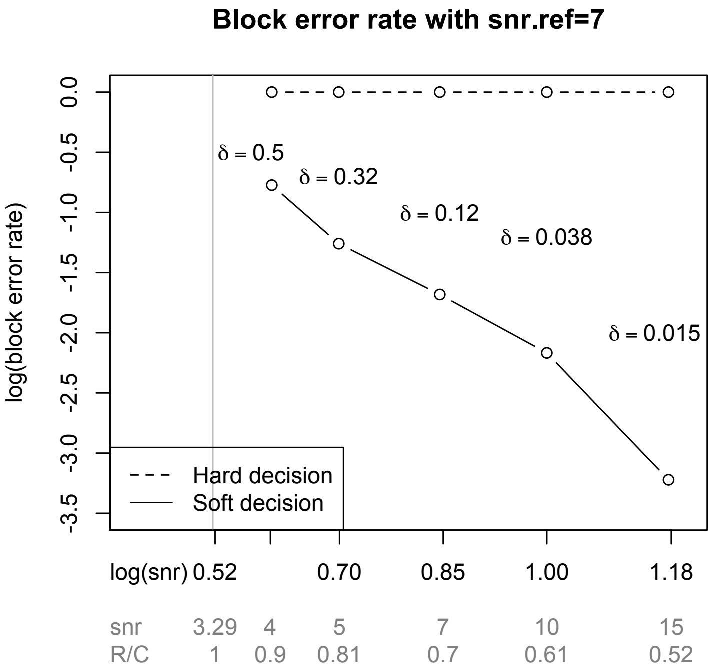

Figure 2.7: L = 512, M = 64, R = 1.05 bits, blocklength n = 2926, snr .ref = 7, snr = (4, 5, 7, 10, 15). Ran 10,000 trials.

are combined to form a statistics approximately to the desired form that is a standard independent normal with a shift only for the true terms. We show that the estimates reliably follow the update rule that we have examined from the update function $g_L(x)$ with the numerical simulations.

The decoder allows us to communicate with any fixed rate below the capacity with error probability that is exponentially small in $L(C-R)^2$. If we consider a communication with rate approaching the capacity, we can achieve any rate at least an order of $1/\sqrt{\log M}$ drop from the capacity.

The simulation shows that there is an improvement of the performance where we have reliable decoder with smaller section size M with a soft decision instead of a hard decision. Also, the performance of the decoder can be improved by future research alternating the power allocation or consider another corrections of the decoder at the final step to push the rate approaches faster to the capacity with exponentially small error probability.

Appendix A

Appendix for Chapter 1

A.1 Preliminary for the proof of Theorem 1

I'm going to state a lemma which Leeb(2008) used in his proof of Thm 1 and some new lemmas which will be used later in the proof of Theorem 1.

Lemma 16. *(Leeb, 2008, lemma A.2) Let B be distributed as χ^2_b, with $b \in \mathbb{N}$. For each $\epsilon > 0$, we then have*

$$P(\frac{B}{b} - 1 > \epsilon) \leq e^{-(b/2)\mathcal{L}(\epsilon)}$$

and

$$P(\frac{B}{b} - 1 < -\epsilon) \leq \begin{cases} e^{-(b/2)\mathcal{L}(\epsilon)}, & \textit{if } \epsilon < 1, \\ 0, & \textit{otherwise.} \end{cases}$$

The function $\mathcal{L}(\cdot)$ is given by

$$\mathcal{L}(c) = c - \log(1 + c)$$

for $c > -1$.

Lemma 17. *Let A and B be independent random variable distributed as χ_a^2 and χ_b^2, respectively, with $a, b \in \mathbb{N}$. For each $\epsilon > 0$ and $\tilde{\epsilon} > 0$, we have*

$$P(|\frac{A}{B} - \frac{a}{b}| > \epsilon) \leq P(|\frac{A}{a} - \frac{B}{b}| > \epsilon\frac{b}{a}(1+\tilde{\epsilon})) + P(B > b(1+\tilde{\epsilon}))$$

Proof

$$\begin{aligned} & P(|\frac{A}{B} - \frac{a}{b}| > \epsilon) \\ = \; & P(\{|\frac{A}{B} - \frac{a}{b}| > \epsilon\} \cap \{B < b(1+\tilde{\epsilon})\}) + P(\{|\frac{A}{B} - \frac{a}{b}| > \epsilon\} \cap \{B > b(1+\tilde{\epsilon})\}) \end{aligned}$$

Note that

$$\begin{aligned} & P(\{|\frac{A}{B} - \frac{a}{b} > \epsilon\} \cap \{B < b(1+\tilde{\epsilon})\}) \\ \leq \; & P(|\frac{A}{B} - \frac{a}{b} > \epsilon | B < b(1+\tilde{\epsilon})) \\ \leq \; & P(\frac{A}{a} - \frac{B}{b} > \epsilon\frac{b}{a}(1+\tilde{\epsilon})) + P(\frac{A}{a} - \frac{B}{b} < -\epsilon\frac{b}{a}(1+\tilde{\epsilon})) \end{aligned}$$

and

$$P(\{|\frac{A}{B} - \frac{a}{b}| > \epsilon\} \cap \{B > b(1+\tilde{\epsilon})\}) \leq P(B > b(1+\tilde{\epsilon}))$$

Thus, we can conclude that

$$\begin{aligned} & P(|\frac{A}{B} - \frac{a}{b}| > \epsilon) \\ \leq \; & P(\frac{A}{a} - \frac{B}{b} > \epsilon\frac{b}{a}(1+\tilde{\epsilon})) + P(\frac{A}{a} - \frac{B}{b} < -\epsilon\frac{b}{a}(1+\tilde{\epsilon})) + P(B > b(1+\tilde{\epsilon})) \end{aligned}$$

This completes the proof of Lemma 17.

Lemma 18. *Let A and B be independent random variable distributed as χ^2_a and χ^2_b, respectively, with $a, b \in \mathbb{N}$. For each $K > 0$, we then have*

$$P(\frac{A}{a} - \frac{B}{b} > K) \leq \exp\left(-t_1 K - \frac{a}{2}\log(1 - \frac{2t_1}{a}) - \frac{b}{2}\log(1 + \frac{2t_1}{b})\right),$$

where

$$t_1 = \frac{1}{4}\Big((a-b) - \frac{a+b}{K} + \sqrt{(b - a + \frac{a+b}{K})^2 + 4ab}\Big)$$

and

$$P(\frac{A}{a} - \frac{B}{b} < -K) \leq \exp\left(t_2 K - \frac{a}{2}\log(1 - \frac{2t_2}{a}) - \frac{b}{2}\log(1 + \frac{2t_2}{b})\right),$$

where

$$t_2 = \frac{1}{4}\Big((a-b) + \frac{a+b}{K} - \sqrt{(b - a - \frac{a+b}{K})^2 + 4ab}\Big)$$

Proof.

$$\begin{aligned}
& P\{\frac{A}{a} - \frac{B}{b} > K\} \\
= \;& P\{\exp(t(\frac{A}{a} - \frac{B}{b})) > \exp(tK)\} \text{ ,where } t \text{ is positive} \\
\leq \;& e^{-tK} E\{\exp(t(\frac{A}{a} - \frac{B}{b}))\} \text{ ,by Markov inequality} \\
= \;& e^{-tK} E\{\exp(\frac{tA}{a})\} E\{\exp(\frac{-tB}{b})\} \text{ ,by independence} \\
= \;& \exp\{-tK - \frac{a}{2}\log(1 - \frac{2t}{a}) - \frac{b}{2}\log(1 + \frac{2t}{b})\}
\end{aligned}$$

Let's look at exponent part and take derivative w.r.t t to get a minimum point of the upperbound.

$$\frac{d}{dt} = 0$$

is equivalent to

$$\frac{2t(a+b)}{(a-2t)(b+2t)} = K$$

The positive part of the solution is

$$4t^* = (a-b) - \frac{a+b}{K} + \sqrt{((b-a) + \frac{a+b}{K})^2 + 4ab}$$

Similarly, we can get the second inequality. This completes the proof of Lemma 18.

Lemma 19. *Let t_1, t_2 and K be*

$$t_1 = \frac{1}{4}\big(- b + a - \frac{a+b}{K} + \sqrt{(b-a+\frac{a+b}{K})^2 + 4ab}\big),$$

$$t_2 = \frac{1}{4}\big(- b + a + \frac{a+b}{K} - \sqrt{(b-a-\frac{a+b}{K})^2 + 4ab}\big)$$

and

$$K = e\frac{b}{a}\big(1 + \frac{be}{a+b}\big),$$

respectively, with $e > 0$ and $a, b \in \mathbb{N}$. Note that $t_1 > 0$ and $t_2 < 0$.

Then we have inequalities as following;

$$-t_1K - \frac{a}{2}\log(1-\frac{2t_1}{a}) - \frac{b}{2}\log(1+\frac{2t_1}{b}) \le -\frac{b}{2}\big(\frac{be}{a+b} - \log(1+\frac{be}{a+b})\big),$$

$$t_2K - \frac{a}{2}\log(1-\frac{2t_2}{a}) - \frac{b}{2}\log(1+\frac{2t_2}{b}) \le -\frac{b}{2}\big(\frac{be}{a+b} - \log(1+\frac{be}{a+b})\big),$$

Proof. For the first inequality, we want to show that,

$$t_1K + \frac{a}{2}\log(1-\frac{2t_1}{a}) + \frac{b}{2}\log(1+\frac{2t_1}{b}) - \frac{b}{2}\big(\frac{be}{a+b} - \log(1+\frac{be}{a+b})\big) \ge 0.$$

First, we are going to show that

$$\frac{a}{2}\log(1-\frac{2t_1}{a})+\frac{b}{2}\log(1+\frac{2t_1}{b})\geq 0.$$

If we multiply by $2/(a+b)$, then we have

$$log\big[(1-\frac{2t_1}{a})^{\frac{a}{a+b}}(1+\frac{2t_1}{b})^{\frac{b}{a+b}}\big].$$

It is enough to show that

$$(1-\frac{2t_1}{a})^{\frac{a}{a+b}}(1+\frac{2t_1}{b})^{\frac{b}{a+b}}\geq 1.$$

Using Jensen's inequality,

$$\begin{aligned} & (1-\frac{2t_1}{a})^{\frac{a}{a+b}}(1+\frac{2t_1}{b})^{\frac{b}{a+b}} \\ \geq\ & \frac{a}{a+b}(1-\frac{2t_1}{a})+\frac{b}{a+b}(1+\frac{2t_1}{b})=1. \end{aligned}$$

Now, we need to show that

$$t_1K-\frac{b}{2}\big(\frac{be}{a+b}-\log(1+\frac{be}{a+b})\big)\geq 0.$$

Since, $log(1+x)\leq x-\frac{x^2}{2}$ when x is non-negative,

$$\frac{b}{2}\log(1+\frac{be}{a+b})\leq\frac{b}{2}[\frac{be}{a+b}-\frac{1}{2}(\frac{be}{a+b})^2]=\frac{b^2e}{2(a+b)}-\frac{b^3e^2}{4(a+b)^2}.$$

We can deduce that

$$\begin{aligned}
& t_1K - \frac{b}{2}\big(\frac{be}{a+b} - \log(1+\frac{be}{a+b})\big) \\
\geq\ & \frac{1}{4}\big(4t_1K - \frac{b^3e^2}{(a+b)^2}\big) \\
=\ & \frac{1}{4}\big((a-b)K - (a+b) + \sqrt{((b-a)K+(a+b))^2+4abK^2} - \frac{b^3e^2}{(a+b)^2}\big)
\end{aligned}$$

which is equivalent to proving

$$\sqrt{((b-a)K+(a+b))^2+4abK^2} \geq \frac{b^3e^2}{(a+b)^2} + (b-a)K + (a+b)$$

For e where the right-hand side is negative, above inequality is always true. And for e where the right-hand side is positive, above inequality is equivalent if we square both terms;

$$\begin{aligned}
& ((b-a)K+(a+b))^2+4abK^2 \geq \big(\frac{b^3e^2}{(a+b)^2} + (b-a)K + (a+b)\big)^2 \\
\Leftrightarrow\ & 4abK^2 - \frac{b^6e^4}{(a+b)^4} - \frac{2b^3e^2}{(a+b)^2}((b-a)K+(a+b)) \geq 0
\end{aligned}$$

Substituting K itself,

$$\begin{aligned}
& 4abK^2 - \frac{b^6e^4}{(a+b)^4} - \frac{2b^3e^2}{(a+b)^2}((b-a)K+(a+b)) \\
=\ & \frac{e^4\,b^5}{a(a+b)^4}\{6a^2+2b^2+7ab\} + \frac{e^3\,b^4}{a(a+b)^2}\{10a+6b\} + \frac{2\,e^2\,b^3}{a(a+b)}\{a+2b\}
\end{aligned}$$

Since e and all the coefficients are non-negative, we can conclude that the above equation is non-negative.

For the second inequality, we want to show that

$$-t_2K + \frac{a}{2}\log(1-\frac{2t_2}{a}) + \frac{b}{2}\log(1+\frac{2t_2}{b}) - \frac{b}{2}\big(\frac{be}{a+b} - \log(1+\frac{be}{a+b})\big) \geq 0$$

By the first argument, we know that

$$\frac{a}{2}\log(1-\frac{2t_2}{a}) + \frac{b}{2}\log(1+\frac{2t_2}{b}) \geq 0$$

We can deduce the problem as

$$-4t_2K - \frac{b^3e^2}{(a+b)^2} \geq 0.$$

By similar argument, the problem is deduced as,

$$4abK^2 - \frac{b^6e^4}{(a+b)^4} + \frac{2b^3e^2}{(a+b)^2}((b-a)K-(a+b)) \geq 0$$

Substituting K itself,

$$\begin{aligned} & 4abK^2 - \frac{b^6e^4}{(a+b)^4} + \frac{2b^3e^2}{(a+b)^2}((b-a)K-(a+b)) \\ = \quad & \frac{e^4\,b^5}{a(a+b)^4}\{2a^2+6b^2+7ab\} + \frac{e^3\,b^4}{a(a+b)^2}\{6a+10b\} + \frac{2\,e^2\,b^3}{a(a+b)}\{a+2b\} \end{aligned}$$

Since e is positive and all the coefficients are positive, the above equation is positive. This completes the proof of Lemma 19.

A.2 Proof of Theorem 1

We are going to establish an upperbound to a probability for the distance between $\rho^2(m)$ and $\hat{\rho}^2(m)$ by using the fact that $\rho^2(m) \sim \sigma^2(m)(1 + \frac{\chi^2_{|m|}}{\chi^2_{n-|m|+1}})$ and RSS(m) $\sim \sigma^2(m)\chi^2_{n-|m|}$. We can decompose the probability as,

$$\begin{aligned} & P(|\rho^2(m) - \hat{\rho}^2(m)| > \epsilon_m) \\ \leq\ & P(|\rho^2(m) - \frac{\sigma^2(m)(n+1)}{n-|m|+1}| > \epsilon_m/2) + P(|\frac{\sigma^2(m)(n+1)}{n-|m|+1} - \hat{\rho}^2(m)| > \epsilon_m/2). \end{aligned}$$

Let's look at the first term.

$$\begin{aligned} & P(|\rho^2(m) - \sigma^2(m)\frac{n+1}{n-|m|+1}| > \epsilon_m/2) \\ =\ & P(|\frac{\rho^2(m)}{\sigma^2(m)} - \frac{n+1}{n-|m|+1}| > \frac{\epsilon_m}{2\sigma^2(m)}) \\ =\ & P(|\frac{\rho^2(m)}{\sigma^2(m)} - 1 - \frac{|m|}{n-|m|+1}| > \frac{\epsilon_m}{2\sigma^2(m)}). \end{aligned}$$

Let's define $A := \chi^2_a$ and $B := \chi^2_b$, where $a := |m|$ and $b := n - |m| + 1$. Since $\rho^2(m) \sim \sigma^2(m)(1 + \frac{\chi^2_{|m|}}{\chi^2_{n-|m|+1}})$, we can rewrite the above equation as,

$$P(|\frac{A}{B} - \frac{a}{b}| > \frac{\epsilon_m}{2\sigma^2(m)}).$$

Using Lemma 17, it is bounded by

$$P(\frac{A}{a} - \frac{B}{b} > \frac{\epsilon_m b}{2a\sigma^2(m)}(1+\tilde{\epsilon}_m)) + P(\frac{A}{a} - \frac{B}{b} < -\frac{\epsilon_m b}{2a\sigma^2(m)}(1+\tilde{\epsilon}_m)) + P(B > b(1+\tilde{\epsilon}_m))$$

Since we can set $\tilde{\epsilon}_m > 0$ arbitrary, let's define $\tilde{\epsilon}_m$ as,

$$\tilde{\epsilon}_m := \frac{\epsilon_m}{2\sigma^2(m)} \frac{b}{a+b}$$

By Lemma 18 and Lemma 19, we can show that the first two terms are bounded by

$$\exp(-\frac{b}{2}\mathcal{L}(\frac{\epsilon_m}{2\sigma^2(m)} \frac{b}{a+b}))$$

Also, by lemma 16, we know that following is true.

$$P(B > b(1+\tilde{\epsilon}_m)) \leq \exp(-\frac{b}{2}\mathcal{L}(\frac{\epsilon_m}{2\sigma^2(m)} \frac{b}{a+b}))$$

Thus, we can deduce that the first term has following upperbound,

$$P(|\rho^2(m) - \sigma^2(m)\frac{n+1}{n-|m|+1}| > \epsilon_m/2) \leq 3\exp(-\frac{b}{2}\mathcal{L}(\frac{\epsilon_m}{2\sigma^2(m)} \frac{b}{a+b})).$$

For the second term, we can find an upperbound by using Lemma 16,

$$\begin{aligned} & P(|\sigma^2(m)\frac{n+1}{n-|m|+1} - \hat{\rho}^2(m)| > \epsilon_m/2) \\ = \; & P(|\frac{RSS(m)}{\sigma^2(m)(n-|m|)} - 1| > \frac{\epsilon_m(n-|m|+1)}{2\sigma^2(m)(n+1)}) \\ \leq \; & 2\exp[-\frac{n-|m|}{2}\mathcal{L}(\frac{\epsilon_m(n-|m|+1)}{2\sigma^2(m)(n+1)})]. \end{aligned}$$

Therefore, we can conclude that

$$\begin{aligned} & P(|\rho^2(m) - \sigma^2(m)\frac{n+1}{n-|m|+1}| > \epsilon_m/2) \\ \leq \; & 5\exp(-\frac{n-|m|}{2}\mathcal{L}(\frac{\epsilon_m}{2\sigma^2(m)} \frac{b}{a+b})) \end{aligned}$$

This completes the proof of Theorem 1.

A.3 Proof of Lemma 3

Let's say we select a model which achieves a minimum of $\hat{\rho}^2(m)(1+\epsilon_m)$. Let's define ϵ_m as,

$$\delta_m = \frac{4f_m}{1-2f_m} \quad \text{with} \quad f_m = f\left(\frac{2(C_n(m)+\log 1/\delta)}{n-|m|}\right)$$

where $f(x) = \log\{e^x + \sqrt{e^x+1}\sqrt{e^x-1}\}$.

We use the same tools as Theorem 1. For each model $m \in \mathcal{M}$, consider a tail probability as following. Using the fact that $\rho^2(m) \sim \sigma^2(m)\left(1+\chi^2_m/\chi^2_{n-m+1}\right)$ and $\hat{\rho}^2(m) \sim \left(\chi^2_{n-|m|}/(n-|m|)\right)((n+1)/(n-|m|))$, we have

$$\begin{aligned}
& P\left\{\rho^2(m) > \hat{\rho}^2(m)(1+\delta_m)\right\} \\
= \ & P\left\{\left(1+\frac{\chi^2_m}{\chi^2_{n-m+1}}\right)\frac{n-|m|}{n+1} - \frac{\chi^2_{n-|m|}}{n-|m|}(1+\delta_m) > 0\right\} \\
\leq \ & P\left\{\frac{\chi^2_m}{\chi^2_{n-m+1}} - \frac{m}{n-|m|+1} > \frac{n+1}{n-|m|}(A-1)\right\} + P\left\{\frac{\chi^2_{n-|m|}}{n-|m|} - 1 < \frac{A}{1+\delta_m} - 1\right\}
\end{aligned}$$

with $A = 1 + \frac{\delta_m}{2+\delta_m}$. This yields,

$$\begin{aligned}
P\left\{\rho^2(m) > \hat{\rho}^2(m)(1+\delta_m)\right\} &\leq 3\exp\left(-\frac{n-|m|}{2}\mathcal{L}(\frac{\delta_m}{2+\delta_m})\right) \\
&\leq 3\exp\left(-\frac{n-|m|}{2}\mathcal{L}(2f_m)\right).
\end{aligned}$$

Note that,

$$\mathcal{L}^{-1}(x) \leq 2f(x).$$

Using similar arguments, we can get

$$\begin{aligned}
& P\left\{\hat{\rho}^2(m) > \rho^2(m)(1+\delta_m)\right\} \cup \left\{\rho^2(m) > \hat{\rho}^2(m)(1+\delta_m)\right\} \\
\leq\ & P\left\{\left|\frac{\chi^2_m}{\chi^2_{n-m+1}} - \frac{m}{n-|m|+1}\right| > \frac{n+1}{n-|m|}(A-1)\right\} \\
& \quad +P\left\{\left|\frac{\chi^2_{n-|m|}}{n-|m|} - 1\right| > 1 - \frac{A}{1+\delta_m}\right\} \\
\leq\ & 5\exp\left(-\frac{n-|m|}{2}\mathcal{L}(2f_m)\right).
\end{aligned}$$

with $A = 1 + \frac{\delta_m}{2+\delta_m}$. Now, using the union bound,

$$\begin{aligned}
& P_{n,\beta,\sigma,\Sigma}\left\{\exists m \in \mathcal{M}_n \text{ s.t } \rho^2(m) > \hat{\rho}^2(m)(1+\delta_m) \text{ or } \hat{\rho}^2(m) > \rho^2(m)(1+\delta_m)\right\} \\
\leq\ & \sum_{m\in\mathcal{M}_n} 5\exp[-\frac{n-|m|}{2}\mathcal{L}(2f_m)] \\
\leq\ & \sum_{m\in\mathcal{M}_n} 5\exp\left[-C_n(m) - \log 1/\delta\right] \\
\leq\ & 5\delta
\end{aligned}$$

We can deduce that, except for a probability of 5δ,

$$\begin{aligned}
\rho^2(\hat{m}) &\leq \hat{\rho}^2(\hat{m})(1+\delta_{\hat{m}}) \\
&\leq \hat{\rho}^2(m)(1+\delta_m) \qquad \text{for any } m \in \mathcal{M}_n \\
&\leq \rho^2(m)(1+\delta_m)^2 \\
&\leq \min_{m\in\mathbb{M}_n}\left\{\rho^2(m)(1+\delta_m)^2\right\}
\end{aligned}$$

Appendix B

Appendix for Chapter 2

B.1 Proof of Lemma 4

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 orthonormal 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 $\mathcal{X}^2_{n-k} = \|G_k\|^2/\sigma_k^2$ 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 $\mathrm{N}(0, \Sigma_{k,e})$. 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 $k \geq 0$. 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 $\mathcal{X}^2_n$ distributed. For $k \geq 1$, 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 $0 \leq i \leq k-1$ and equal to $\sum_j V_{k,j,i}\hat{\beta}_{k,j}$ for $k \leq i \leq n-1$. From the induction hypothesis, these $(V_{k,j,i} : j \in J)$ have conditional distribution $\mathrm{Normal}(0, \Sigma_{k-1})$. 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 $i = k, \ldots, n-1$ 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, keeping $G_0/\|G_0\|, \ldots, G_{k-1}/\|G_{k-1}\|$, but 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 4.

B.2 The method of nearby measure

The method of nearby measure using Renyi Relative entropy is stated by Barron and Joseph (2010) and it is currently explored further by e.g. Rush and Barron (2013). The Renyi relative entropy of order $\alpha > 1$ of two probability measures $\mathbb{P}$ and $\mathbb{Q}$ with density functions $p(Z)$ and $q(Z)$ for a random vector Z is defined by

$$D_\alpha(\mathbb{P}\|\mathbb{Q}) = \frac{1}{\alpha-1} \log \mathbb{E}_\mathbb{Q}[(p(Z)/q(Z))^\alpha].$$

Lemma 20 (Lemma 44. in Barron and Joseph (2010))**.** *Let $\mathbb{P}$ and $\mathbb{Q}$ be a pair of probability measures with finite $D_\alpha(\mathbb{P}\|\mathbb{Q})$. For any event A, and $\alpha > 1$,*

$$\mathbb{P}[A] \leq \left[\mathbb{Q}[A] e^{D_\alpha(\mathbb{P}\|\mathbb{Q})}\right]^{(\alpha-1)/\alpha}.$$

If $D_\alpha(\mathbb{P}\|\mathbb{Q}) \leq c_0$ for all α, then the following bound holds, taking the limit of large α,

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

In this case the density ratio $p(Z)/q(Z)$ is uniformly bounded by e^{c_0}.

Proof. For convex f, as in Csiszar's f-divergence inequality, from Jensen's inequality applied to the decomposition of $\mathbb{E}[f(p(Z)/q(Z))]$ using the distributions conditional on A and its complement,

$$\mathbb{Q}Af(\mathbb{P}A/\mathbb{Q}A) + \mathbb{Q}A^c f(\mathbb{P}A^c/\mathbb{Q}A^c) \leq \mathbb{E}_{\mathbb{Q}} f(p(Z)/q(Z))$$

Using in particular $f(r) = r^\alpha$ and throwing out the non-negative A^c part, yields

$$(\mathbb{P}[A])^\alpha \leq (\mathbb{Q}[A])^{\alpha-1}\mathbb{E}_{\mathbb{Q}}[(p(Z)/q(Z))^\alpha].$$

It is also seen as Holder's inequality applied to $\int q(p/q)1_A$. Taking the α root produces the stated inequality. The proof for Lemma 20 is completed.

When the true distribution is complicated to analyze, but when we know a convenient distribution which not far from the true one, this method can make the analysis much simpler. If an event A is exponentially unlikely under the approximating distribution and the Renyi relative entropy is bounded by a constant or an amount of a smaller order than the exponent of the tail probability under the approximating distribution, then we can say that it is also exponentially unlikely under the true distribution.

B.3 Proof of Lemma 5.

The true distribution of $\mathcal{Z}_k$ given $\mathcal{F}_{k-1}$ is proven to be

$$\mathcal{X}_{n-k} b_k + Z_k$$

where $Z_k \sim N(0, \Sigma_k)$. We will approximate the distribution to

$$\sqrt{n} b_k + Z_k$$

where $Z_k \sim N(0, \tilde{\Sigma}_k)$ with $\tilde{\Sigma}_k = I - Proj_k$.

Using the Lemma. 3 in Barron and Joseph (2010), if the Renyi relative entropy for $\alpha = 2$ between $\mathbb{P}_{\mathcal{Z}_k|\mathcal{F}_{k-1}}$ and the corresponding $\mathbb{Q}$ is bounded by $(C + 2 + k^2/n)$ for each k, then the event A which is determined by $\mathcal{F}_k$ is bounded by

$$(\mathbb{Q}Ae^{k(2+k^2/n+C)})^{1/2}$$

Thus it is enough to show the Renyi relative entropy for a fixed step k.

Lemma 21. *The Renyi relative entropy for $\alpha = 2$ between the two distribution $\mathbb{P}_{\mathcal{Z}_k|\mathcal{F}_{k-1}}$ and $\mathbb{Q}_{\mathcal{Z}_k|\mathcal{F}_{k-1}}$ is bounded by $2 + k^2/n + C$*

Proof: The true distribution of $\mathcal{Z}_k$ given $\mathcal{F}_{k-1}$ is normal distribution with mean $(\|G_k\|/\sigma_k)b_k$ where $\|G_k\|/\sigma_k$ is $\mathcal{X}$ distributed with degree of freedom $n - k$ and the covariance matrix is upper $N \times N$ part of

$$\Sigma_{k,e} = I - b_{0,e}b_{0,e}^T - \cdots - b_{k-1,e}b_{k-1,e}^T$$

which will be denoted by Σ_k.

The approximating distribution is also a normal distribution with mean $\sqrt{n}\, b_k$ and the same covariance matrix. Thus,

$$\begin{aligned} D(\mathbb{P}\|\mathbb{Q}) &= D\left(\mathbb{E}\phi(. - (\|G_k\|/\sigma_k)b_k)\|\phi(. - \sqrt{n}b_k)\right) \\ &\leq \mathbb{E}D\left(\phi(. - (\|G_k\|/\sigma_k)b_k)\|\phi(. - \sqrt{n}b_k)\right) \end{aligned}$$

In our case, the covariance matrix is not a full rank matrix. When the covariance matrix of a multivariate normal distribution is of low rank, there is no closed form of the density under the Lebesgue measure. Suppose we can decompose the covariance matrix with rank r to QDQ^T where Q is $N \times r$ orthonormal matrix and D is a diagonal matrix. Then, we can transform the distribution to the space of $\mathbb{R}^r$.

Define $\tilde{Q}$ as a $N \times (N - r)$ matrix the orthonormal column vectors which is orthogonal to Q so that $U^T = [Q : \tilde{Q}]$ is a orthogonal vectors for $\mathbb{R}^N$. For any vector $x \in \mathbb{R}^N$, Ux would be a rotation. Thus if x is a normal random variable with some covariance matrix, Ux preserves its distribution.

Suppose the covariance matrix can be decomposed by $Q\tilde{D}Q^T$. Then, instead of comparing $Z_1 \sim N(B, \Sigma)$ for $\mathbb{P}$ and $Z_2 \sim N(b, \tilde{\Sigma})$ for $\mathbb{Q}$, we compare first r elements of $\tilde{Z}_2 \sim N(UB - u, U\tilde{\Sigma}U^T)$ from $\tilde{Z}_1 \sim N(Ub - u, U\Sigma U^T)$. By this transformation, they preserve their distribution. The u is the shift of the mean to make the last few components of mean to be zero so that the last $(r + 1)$ to N elements of $\tilde{Z}_1$ and $\tilde{Z}_2$ will be zero and the the covariance would be D for the first r elements and zero elsewhere. Thus, the distribution lies in $\mathbb{R}^r$ space rather than a some subspace of $\mathbb{R}^N$.

Thus the Renyi relative entropy would be

$$D_\alpha(Z_1 \| Z_2) = D_\alpha(\tilde{Z}_1 \| \tilde{Z}_2)$$

Here, $\tilde{Z}_1$ and $\tilde{Z}_2$ is the first r elements.

For the true distribution, the covariance matrix is an upper $N \times N$ part of projection matrix onto space orthogonal to $(\beta_e, \hat{\beta}_{1,e}, \ldots, \hat{\beta}_{k,e})$. This can be represented as

$$\Sigma_k = I - Proj_k - c_0 q_0 q_0^T$$

where $c_0 = \|\tilde{\Sigma}_k \beta\|^2 / (\|\tilde{\Sigma}_k \beta\|^2 + \sigma^2)$ and $q_0 \propto \tilde{\Sigma}_k \beta$ with appropriately normalized.

Notice that the covariance matrix for the approximating distribution is

$$\tilde{\Sigma}_k = I - Proj_k$$

If we write $Proj_k = q_1 q_1^T + \cdots + q_k q_k^T$, by successive orthogonal Gram-Schmidt procedure, we can write the identity matrix as

$$q_1 q_1^T + \cdots + q_k q_k^T + q_0 q_0^T + q_{k+1} q_{k+1}^T + \cdots + q_{N-1} q_{N-1}^T.$$

Then we can write the two covariance matrix Σ_k and $\tilde{\Sigma}_k$ in terms of those orthogonal vectors as following

$$\Sigma_k = (1 - c_0) q_0 q_0^T + q_{k+1} q_{k+1}^T + \cdots + q_{N-1} q_{N-1}^T = Q D_k Q^T$$

and

$$\tilde{\Sigma}_k = q_0 q_0^T + q_{k+1} q_{k+1}^T + \cdots + q_{N-1} q_{N-1}^T = Q \tilde{D}_k Q^T$$

with $D_k = I - c_0 e_1 e_1^T$ and $\tilde{D}_k = I$. From the above representation, we can see that rank of both matrices is $(N - k)$. As a result, we get the Renyi relative entropy $D_\alpha(\mathbb{P}\|\mathbb{Q})$ is as following. The derivation is in the following subsection.

$$\frac{1}{2} \log \frac{1}{(1-c_0)} - \frac{1}{2(\alpha-1)} \log(1 + c_0(\alpha - 1)) + \frac{\alpha}{2} (\tilde{B} - \tilde{b})^T (I - \frac{c_0(\alpha - 1)}{1-c_0 + \alpha c_0} e_1 e_1^T)(\tilde{B} - \tilde{b})$$

where $\tilde{B}$ is the first $N - k$ element of $\mathcal{X}_{n-k} U b_k$ and $\tilde{b}$ is also the first $N - k$ element of $\sqrt{n}\, U b_k$.

For the first part, if we plug in the value c_0, it is upperbounded by the Capacity. We can drop the second term since it is negative. For the third term, Since it is a quadratic terms, we can bound the above by $\frac{\alpha}{2} \|\tilde{B} - \tilde{b}\|^2$ by ignoring the negative part corresponds to $\frac{c_0(\alpha-1)}{1-c_0+\alpha c_0} e_1 e_1^T$. Thus, the Renyi relative $D_\alpha(\mathbb{P}\|\mathbb{Q})$ is not more than $\frac{\alpha}{2}(\chi_{n-k}-\sqrt{n})^2$.

Next, we prove $\mathbb{E}(\mathcal{X}_{n-k}-\sqrt{n})^2 \leq 2+k^2/n$. We use the fact that

$$|A-a| = \frac{|A^2-a^2|}{|A+a|} \leq \frac{|A^2-a^2|}{a^2}.$$

Then,

$$\begin{aligned}
\mathbb{E}(\mathcal{X}_{n-k}-\sqrt{n})^2 &\leq \frac{\mathbb{E}(\mathcal{X}_{n-k}^2-n)^2}{n} \\
&\leq \frac{\mathbb{E}(\mathcal{X}_{n-k}^2-(n-k))^2+(n-k-n)^2}{n} \\
&\leq \frac{2(n-k)+k^2}{n} \leq 2+\frac{k^2}{n}
\end{aligned}$$

This completes the proof.

B.3.1 The derivation of the Renyi relative entropy

Since the two covariance matrices are diagonal, the normals are independent. So we can calculate the Renyi relative entropy using the marginals. $D_\alpha(\mathbb{P}\|\mathbb{Q})$ is defined by

$$\frac{1}{\alpha-1}\log\int(2\pi)^{-r/2}\frac{|\tilde{D}_k|^{(\alpha-1)/2}}{|D_k|^{\alpha/2}}e^{-\frac{1}{2}\left(\alpha(x-\tilde{B})^T D_k^{-1}(x-\tilde{B})-(\alpha-1)(x-\tilde{b})^T\tilde{D}_k^{-1}(x-\tilde{b})\right)}$$

Here, since $D_k = I-c_0e_1e_1^T$, the determinant is $(1-c_0)$ and the inverse is $I+\frac{c_0}{1-c_0}e_1e_1^T$. For $\tilde{D}_k$, since it is an identity matrix in $\mathbb{R}^{N-k}$, the determinant is one and the inverse is also an identity matrix. So now we have

$$\frac{1}{\alpha-1}\log\int(2\pi)^{-r/2}\frac{1}{(1-c_0)^{\alpha/2}}e^{-\frac{1}{2}\left(\alpha(x-\tilde{B})^T(I+\frac{c_0}{1-c_0}e_1e_1^T)(x-\tilde{B})-(\alpha-1)(x-\tilde{b})^T(x-\tilde{b})\right)}.$$

Now we look at the -2 of exponent to ignore the $-1/2$ factor for a moment. We can write it as

$$\alpha(1+\frac{c_0}{1-c_0})(x_1-\tilde{B}_1)^2-(\alpha-1)(x_1-\tilde{b}_1)^2+\sum_{i=2}^{N-k}\alpha(x_i-\tilde{B}_i)^2-(\alpha-1)(x_i-\tilde{b}_i)^2.$$

For $i=2,\ldots,N-k$, we can rearrange the equation, for some $b_{new,i}$ as a function of $\tilde{B}_i$, $\tilde{b}_i$ and α, as following.

$$(x_i-b_{new,i})^2-\alpha(\alpha-1)(\tilde{B}_i-\tilde{b}_i)^2.$$

For $i=1$, from $c_0e_1e_1^T$ part, we get

$$\frac{1-c_0+\alpha c_0}{1-c_0}(x_1-b_{new,1})^2-\frac{\alpha(\alpha-1)}{1-c_0+\alpha c_0}(\tilde{B}_1-\tilde{b}_1)^2.$$

After fitting into the integral and integrate out all the x_i, then we have

$$\frac{1}{\alpha-1}\log\left(\frac{(1-c_0)^{-(\alpha-1)/2}}{(1-c_0+\alpha c_0)^{-1/2}}\exp\{\frac{\alpha(\alpha-1)}{2}(\tilde{B}-\tilde{b})^T(I-\frac{c_0(\alpha-1)}{1-c_0+\alpha c_0}e_1e_1^T)(\tilde{B}-\tilde{b})\}\right)$$

This is

$$-\frac{1}{2}\log(1-c_0)-\frac{1}{2(\alpha-1)}\log(1-c_0+\alpha c_0)+\frac{\alpha}{2}(\tilde{B}-\tilde{b})^T(I-\frac{c_0(\alpha-1)}{1-c_0+\alpha c_0}e_1e_1^T)(\tilde{B}-\tilde{b})$$

B.4 Proof of Lemma for update functions

B.4.1 Proof of Lemma 7

We use the Riemann sums for the approximation and we change the measure $u=u(t)=(1-e^{-2Ct})/(1-e^{-2C})$. For a monotone non-increasing function $G(t)$ for

$0 < t < 1$, the Riemann sum as following is lowerbounded by Riemann integral

$$\sum_{\ell=1}^{L} \frac{1}{L} G(\frac{\ell-1}{L}) \geq \int_0^1 G(t)dt$$

The gap between the inequality is bounded by $G(1)/L$ as following

$$\begin{aligned}\sum_{\ell=1}^{L} \frac{1}{L} G(\frac{\ell-1}{L}) - \int_0^1 G(t)dt &\leq \sum_{\ell=1}^{L} \frac{1}{L} \left(G(\frac{\ell}{L}) - G(\frac{\ell-1}{L}) \right) \\ &= \frac{1}{L} \left(G(1) - G(0) \right) \leq \frac{G(1)}{L}.\end{aligned}$$

Next, we examine the update function. Notice that

$$\frac{P_\ell}{P} = \frac{e^{-2C(\ell-1)/L}}{1-e^{-2C}} \frac{2\tilde{C}}{L} = \frac{2\tilde{C}}{L} \left(1 + \frac{1}{snr} - u(\ell) \right).$$

Thus, we have

$$\begin{aligned} g_L(x) &= \sum_{\ell=1}^{L} \frac{P_\ell}{P} g(u(\ell), x) \\ &= \sum_{\ell=1}^{L} \frac{2\tilde{C}}{L} \left(1 + \frac{1}{snr} - u(\ell) \right) g(u(\ell), x) \\ &\leq \int_0^1 G(u(t), x) dt \end{aligned}$$

where

$$G(u(t), x) = 2\tilde{C} \left(1 + \frac{1}{snr} - u(t) \right) g(u(t), x)$$

with $u(t) = (1 - e^{-2Ct})/(1 - e^{-2C})$. It is non-increasing function in t and $u(t)$ is increasing function in t. Next, we change the measure w.r.t. u so that

$$g_L(x) \geq \frac{\tilde{C}}{C} g(x).$$

Furthermore, we approximate $g_L(x)$ with $g(x)$ within an order of $1/L$.

$$\begin{aligned}|g_L(x)-g(x)| &\leq \left|g_L(x)-\frac{\tilde{C}}{C}g(x)\right|+\left|\frac{\tilde{C}}{C}g(x)-g(x)\right|\\ &\leq \frac{G(1,x)}{L}+g(x)(1-\frac{\tilde{C}}{C})\leq\frac{P_L}{P}+\frac{C}{L}\end{aligned}$$

which is an order of $1/L$ when $L >> 2C$.

Proof of Lemma 8

We first prove the following by McFadden and Zarembka (1974).

$$\mathbb{P}_{\upsilon_1,\ldots,\upsilon_M|Y_1,\ldots,Y_M}\{Y_1+\upsilon_1\geq max_{2\leq j\leq m}(Y_j+\upsilon_j)\}=\frac{e^{Y_1}}{\sum_{j=1}^M e^{Y_j}}$$

so that $\mathbb{P}\{Y_1+\upsilon_1 \geq max_{2\leq j\leq m}(Y_j+\upsilon_j)\} = \mathbb{E}_{Y_1,\ldots,Y_M}\frac{e^{Y_1}}{\sum_{j=1}^M e^{Y_j}}$ is true. Suppose $Y_1,\ldots,Y_M$ is given as a constant.

$$\begin{aligned}&\mathbb{P}_{\upsilon_1,\ldots,\upsilon_M}\{Y_1+\epsilon_1\geq max_{2\leq j\leq m}(Y_j+\upsilon_j)\}\\ =\;&\mathbb{P}_{\upsilon_1}\{\prod_{j=2}^M\mathbb{P}_{\upsilon_2}\{Y_1+\upsilon_1\geq Y_2+\upsilon_2|\upsilon_1\}\}\\ =\;&\mathbb{P}_{\upsilon_1}exp(-\sum_{j=2}^M e^{-(\upsilon_1+Y_1-Y_j)})\\ =\;&\int_{-\infty}^{\infty}exp(-\sum_{j=2}^M e^{-(y+Y_1-Y_j)}-e^{-y})exp(-y)dy\\ =\;&\int_{-\infty}^{\infty}exp(-t(\sum_{j=2}^M e^{-(Y_1-Y_j)}+1))dt\\ =\;&\frac{1}{\sum_{j=2}^M e^{-(Y_1-Y_j)}+1}=\frac{e^{Y_1}}{\sum_{j=1}^M e^{Y_j}}.\end{aligned}$$

The fourth equality follows by changing the variable $e^{-y} = t$. Set the independent random sequence Y_j as $Y_1 = \alpha^2 + \alpha Z_1$ and $Y_j = \alpha Z_j$ for $j = 2, \ldots, M$. Then, using lemma 7,

$$\begin{aligned} g(x) &= \mathbb{E}_U g(U,x) = \mathbb{E}_U \mathbb{E}[\frac{e^{\alpha^2+\alpha Z_1}}{e^{\alpha^2+\alpha Z_1} + \sum_{j=2}^{M} e^{\alpha Z_j}}] \\ &= \mathbb{P}_{Z_1,\ldots,Z_m,\upsilon_1,\ldots,\upsilon_m,U}\{\alpha^2 + \alpha Z_1 + \upsilon_1 \geq max_{2\leq j\leq m}(\alpha Z_j + \upsilon_j)\} \end{aligned}$$

This proves the first expression in the lemma.

If we consider an event $\{\alpha^2 + \alpha Z_1 + \upsilon_1 \geq max_{2\leq j\leq m}(\alpha Z_j + \upsilon_j)\}$, we can rearrange the inequality and get a lower bound with α on one side and random variables only depends on m on the other side.

$$\begin{aligned} & \{\alpha^2 + \alpha Z_1 + \epsilon_1 \geq max_{2\leq j\leq m}(\alpha Z_j + \upsilon_j)\} \\ = & \prod_{2\leq j\leq m} \{\alpha^2 + \alpha Z_1 + \upsilon_1 \geq \alpha Z_j + \upsilon_j\} \\ = & \prod_{2\leq j\leq m} \{\alpha^2 + \alpha(Z_1 - Z_j) \geq \upsilon_j - \upsilon_1\} \\ = & \prod_{2\leq j\leq m} [\{\alpha \geq -\frac{Z_1 - Z_j}{2} + \sqrt{[\upsilon_j - \upsilon_1 + \frac{(Z_1 - Z_j)^2}{4}]_+}\} \\ & + \{\alpha \leq -\frac{Z_1 - Z_j}{2} - \sqrt{[\upsilon_j - \upsilon_1 + \frac{(Z_1 - Z_j)^2}{4}]_+}\}] \\ \geq & \{\alpha \geq max_{2\leq j\leq m}[-\frac{Z_1 - Z_j}{2} + \sqrt{[\upsilon_j - \upsilon_1 + \frac{(Z_1 - Z_j)^2}{4}]_+}]\} \end{aligned}$$

The fourth inequality follows by throwing away the second event for each j.

If we rearrange the set $\{Y_1 \geq max_{2\leq j\leq m} Y_j\}$ in terms of α, then it is an intersection of intervals, $(-\infty, V_a(j, left)] \cup [V_a(j, right), \infty) = (V_a(j, left), V_a(j, right))^c$, where $V_a(j, left)$ and $V_a(j, right)$ is corresponding end points for each j as in the

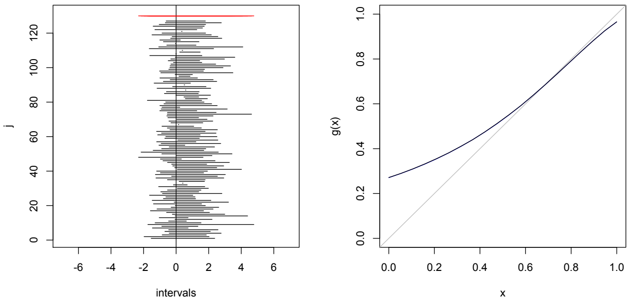

Figure B.1: The first figure is a realization of (m-1) intervals for $m = 2^7$. The second plot is $g(x)$ with the first lowerbound. It is hard to see the gap between the two curves. We used MC simulation with replicate size 10000.

third expression in the proof. Note that an intersection of compliment sets is a compliment of a union of them. Figure B.1 shows a realization of the intervals $(V_a(j, left), V_a(j, right))^c$ for $m = 2^7$. The line at the top indicates the union of those intervals. If there are overlaps among the intervals, the union set is more likely to be one continuous interval which would be $(\min V_a(j, left), \max V_a(j, right))$, as in the figure. Thus, as m increases, the red line is more likely to be one continuous interval. Also, the minimum decreases and is likely to be non-positive. This makes the gap between the lowerbound and the actual update function small. This completes the proof of lemma 8.

B.5 Proof of Lemma 12

The key tools for the proof would be the first order Taylor expansion with the property of the weights that the each element is positive and between 0 and 1 and also the weights are sum to one. First, for the square norm difference, we have

$$\begin{aligned}
\left|\sum_{j=1}^{M}\left(w_j^2-(w_j^*)^2\right)\right| &= 2\left|\sum_{j=1}^{M}\tilde{w}_j^2\left(\epsilon_j-\sum_{j'=1}^{M}\tilde{w}_j\epsilon_j\right)\right| \\
&= 2\left|\sum_{j=1}^{M}\tilde{w}_j\epsilon_j\left(\tilde{w}_j-\sum_{j'=1}^{M}\tilde{w}_{j'}^2\right)\right| \\
&\leq 2\max_j|\epsilon_j|\sum_{j=1}^{M}|\tilde{w}_j\left(\tilde{w}_j-\|\underline{\tilde{w}}\|^2\right)| \\
&\leq 2\max_j|\epsilon_j|
\end{aligned}$$

where $\tilde{w}_j = e^{s_j+\tilde{\epsilon}_j}/\sum_{j'=1}^{M} e^{s_{j'}+\tilde{\epsilon}_{j'}}$ where $\tilde{\epsilon}_j$ is between ϵ_j and zero. The first equality follows from the Taylor expansion with respect to ϵ_j for $j=1,\ldots,M$. We rearrange the summand and take out the maximum of ϵ_j. Finally the last inequality holds since the $\underline{\tilde{w}}$ sums to one and all the elements are positive.

Similarly, we evaluate the difference of one specific weight. As we shall see, first we use the Taylor expansion with respect to the difference of the exponents. Then we get an upperbound using the Holder's inequality and the fact that $\tilde{w}_j$ sums to one and is positive.

$$\begin{aligned}
\left|w_{j_\ell}-w_{j_\ell}^*\right| &= \left|\tilde{w}_{j_\ell}\left(\epsilon_{j_\ell}-\sum_{j'=1}^{M}\tilde{w}_{j'}\epsilon_{j'}\right)\right| \\
&\leq \left|\epsilon_{j_\ell}-\sum_{j'=1}^{M}\tilde{w}_{j'}\epsilon_{j'}\right| \\
&\leq |\epsilon_{j_\ell}|+\left|\sum_{j'=1}^{M}\tilde{w}_{j'}\epsilon_{j'}\right| \\
&\leq 2\max_j|\epsilon_j|
\end{aligned}$$

Indeed, we prove the following similarly. 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.

$$\begin{aligned}
& \left|\sum_{j=1}^{M}\left(w_j w_{2,j} - w_j^* w_{2,j}^*\right)\right| \\
= & \left|\sum_{j=1}^{M} w_j(w_{2,j} - w_{2,j}^*) + w_{2,j}^*(w_j - w_j^*)\right| \\
\leq & \left|\sum_{j=1}^{M} w_j \tilde{w}_{2,j}(\epsilon_{2,j} - \sum_{j'=1}^{M}\tilde{w}_{2,j'}\epsilon_{2,j'})\right| + \left|\sum_{j=1}^{M} w_{2,j}^* \tilde{w}_j(\epsilon_j - \sum_{j'=1}^{M}\tilde{w}_{j'}\epsilon_{j'})\right| \\
\leq & \ 2\left(\max_j |\epsilon_{2,j}| + \max_j |\epsilon_j|\right)
\end{aligned}$$

Similarly, we evaluate for $\|w\|^2 - 2w_{j_\ell}$ using above statements. Again, we use the first order Taylor expansion.

$$\begin{aligned}
& |\sum_{j=1}^{M}(w_j^2 - 2w_{j_\ell} - (w_j^*)^2 + 2w_{j_\ell}^*)| \\
= & \left|2\sum_{j=1}^{M}\tilde{w}_j^2\left(\epsilon_j - \sum_{j'=1}^{M}\tilde{w}_j\epsilon_j\right) - 2\tilde{w}_{j_\ell}\left(\epsilon_{j_\ell} - \sum_{j'=1}^{M}\tilde{w}_{j'}\epsilon_{j'}\right)\right| \\
= & \ 2\left|\sum_{j=1}^{M}\epsilon_j\tilde{w}_j\{\underbrace{2\tilde{w}_j - \sum_{j'=1}^{M}\tilde{w}_j^2 + w_{j_\ell} - 1_{\{j=j_\ell\}}}_{(*)}\}\right|
\end{aligned}$$

The $(*)$ part is not more than 2 in absolute value. Thus, we have

$$|\sum_{j=1}^{M}(w_j^2 - 2w_{j_\ell} - (w_j^*)^2 + 2w_{j_\ell}^*)| \leq 4\max_j |\epsilon_j|$$

B.6 Proof of Lemma 14

For the proof, we have a corollary from the Lemma 10 using the Bayes optimal estimates.

Corollary 6 (Reliability)**.** *For any β and any $1 \leq k < k' \leq k^*$, the expectation of $\beta^T \beta_k^*$, $\|\beta_k^*\|^2$ and $\beta_k^{*T} \beta_{k'}^*$ are the same which will be defined by $x_k P$ 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*

$$A_{\beta,\delta} = \{|\beta^T \beta_k^* - x_k P| < \delta_k \textit{ and } |\|\beta_k^*\|^2 - x_k P| < \delta_k \textit{ and } |\beta_k^{*T} \beta_{k'}^* - x_k P| < \delta_k\}$$

then for any $\delta > 0$,

$$\mathbb{P}\{A_{\beta,\delta}^c\} \leq 6(k+1) \exp\{-\frac{2}{c^2} L \delta_k^2\},$$

where $c^2 = L \max(P_\ell)$ with value near $P \frac{2C}{1-e^{-2C}}$ if we use the variable power allocation.

Proof for Cor 6: We already reveal that the success rate $\beta^T \beta_k^*$ and the square norm $\|\hat{\beta}_k^*\|^2$ and $\beta_k^{*T} \beta_{k'}^*$ share their expectation. As we have seen in Lemma 10, those quantities have the independence across sections and they are also sum of bounded random variables by $\sum_{\ell=1}^{L} P_\ell^2$. The union bound would be sum of the three tail probability which makes $6 \exp\{-\frac{2}{c^2 P} L \delta^2\}$. This completes the proof.

Next, if we consider the difference between the estimates and the theoretical success rate, we can use the Bayes optimal estimates as a bridge. By triangular inequality, we have

$$\left| \frac{\beta^T \hat{\beta}_k}{P} - x_k \right| \leq \left| \frac{\beta^T \hat{\beta}_k}{P} - \frac{\beta^T \beta_k^*}{P} \right| + \left| \frac{\beta^T \beta_k^*}{P} - x_k \right|$$

where β^*_{k+1} is a function of $stat^*_k = \sum_{k'=0}^k \lambda^*_{k,k'} \mathcal{Z}^*_k + \sqrt{n/c_k}\,\hat{\beta}_k$. The similar argument works for the square norm. The second part on the right side will be controlled by reliability we discussed in Cor 6. The first part is difference between the two weighted averages of the exponential weights. We can bound it by the difference among the exponents by Lemma 12.

$$
\begin{aligned}
&\max\left(\left|\frac{\beta^T\hat{\beta}_{k+1}}{P} - \frac{\beta^T\beta^*_{k+1}}{P}\right|, \left|\frac{\|\hat{\beta}_{k+1}\|^2}{P} - \frac{\|\beta^*_{k+1}\|^2}{P}\right|\right) \\
&\leq 4\sum_\ell \frac{P_\ell}{P} \max_{j\in sec_\ell} |\alpha_{\ell,k}\hat{stat}_{k,j} - \alpha_{\ell,k} stat^*_{k,j}| \\
&\leq 4\sum_\ell \frac{P_\ell}{P} \alpha_{\ell,k} \max_{j\in sec_\ell} |\mathcal{Z}^{comb}_{k,j} - \mathcal{Z}^{comb,*}_{k,j}| \\
&\leq 4\sum_\ell \frac{P_\ell}{P} \alpha_{\ell,k} \max_{j\in sec_\ell} |\mathcal{Z}^{comb}_{k-1,j} - \mathcal{Z}^{comb,*}_{k-1,j}| + \lambda_{k,k}|\mathcal{Z}_{k,j} - \mathcal{Z}^*_{k,j}| \\
&\leq 4\sum_\ell \frac{P_\ell}{P} \alpha_{\ell,k} \max_{j\in sec_\ell} |\mathcal{Z}^{comb}_{k-1,j} - \mathcal{Z}^{comb,*}_{k-1,j}| + 4\sum_\ell \frac{P_\ell}{P} \alpha_{\ell,k}\lambda_{k,k} \max_{j\in sec_\ell} |\mathcal{Z}_{k,j} - \mathcal{Z}^*_{k,j}|
\end{aligned}
$$

We prove inductively and there are hidden inductive arguments as following, in order to have (a)-(c) we need

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

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

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

(D) $\sum_{\ell=1^L} \frac{P_\ell}{P}\alpha_{\ell,k-1} \max_{j\in sec_\ell} |\mathcal{Z}^{comb}_{k-1,j} - \mathcal{Z}^{comb,*}_{k-1,j}| \leq \frac{\sqrt{P}}{\sqrt{c_{k-1}}} c^3 A_{k-1}(n/L)^{k-1}\eta$

for some constant $A_k = A_{k-1}/(n/L) + \lambda_{k,k} d_k$ and d_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 $b^*_0 = b_0 = \beta/\sqrt{c_0}$. Thus, we have $\mathcal{Z}_0 = \mathcal{Z}^*_0$ and so does $|stat^*_0 - stat_0|$.

Thus,

$$\begin{aligned}\left|\frac{\|\hat{\beta}_1\|^2}{P} - x_1\right| &= \left|\frac{\|\beta_1^*\|^2}{P} - x_1\right| \\ &\leq \delta = a_1\eta.\end{aligned}$$

The (a) and (b) both upper bounded by the same amount with $a_k = 1$.

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

$$b_k = \frac{\hat{\beta}_k - \sum_{k'=0}^{k-1}(b_{k'}^T\hat{\beta}_k)b_{k'}}{\sqrt{\|\hat{\beta}_k^2\|^2 - \sum_{k'=0}^{k-1}(b_{k'}^T\hat{\beta}_k)^2}} = \frac{num_k}{den_k}$$

and

$$b_k^* = \frac{\hat{\beta}_k - \hat{\beta}_{k-1} - \lambda_{k,k}^2(\beta - \hat{\beta}_{k-1})}{\lambda_{k,k}\sqrt{c_k}} = \frac{num_k^*}{den_k^*}.$$

Now

$$\begin{aligned}|b_k - b_k^*| &= \left|\frac{num_k}{den_k} - \frac{num_k^*}{den_k^*}\right| \\ &\leq \frac{|num_k - num_k^*|}{den_k^*} + |b_k|\frac{|den_k - den_k^*|}{den_k^*}.\end{aligned}$$

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

$$\begin{aligned}
&|\sum_{k'=0}^{k-1}(b_{k'}^T\hat{\beta})b_{k'} - (c_0 - c_k)\sqrt{\omega_0}\,b_0^* - c_k\sum_{k'=1}^{k-1}\sqrt{\omega_{k'}}\,b_{k'}^*| \\
&\leq \left|(b_0^T\hat{\beta}_k - (c_0 - c_k)\sqrt{\omega_0})\right||b_0| + \\
&\sum_{k'=1}^{k-1}|(b_{k'}^T\hat{\beta}_k) - c_k\sqrt{\omega_{k'}}||b_{k'}| + \sum_{k'=1}^{k-1}c_k\sqrt{\omega_{k'}}|b_{k'} - b_{k'}^*|
\end{aligned}$$

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

$$\begin{aligned}
|b_0^T\hat{\beta}_k - (c_0 - c_k)\sqrt{\omega_0}| &= |\beta^T\hat{\beta}_k - x_k P|/\sqrt{c_0} \\
&\leq (P/\sqrt{c_0})a_k(n/L)^{k-1}\eta
\end{aligned}$$

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

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

Accordingly,

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

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 6. For the second part, using the Lemma. 12, we can bound $|(\beta^*_{k'})^T\beta^*_k - \hat{\beta}^T_{k'}\hat{\beta}_k|$ by

$$\begin{aligned}
\leq \ & 2\sum_{\ell=1}^{L} P_\ell max_{j\in sec_\ell}|\alpha_{\ell,k'}stat^*_{k',j} - \alpha_{\ell,k'}\hat{s}tat_{k',j}| \\
& +2\sum_{\ell=1}^{L} P_\ell max_{j\in sec_\ell}|\alpha_{\ell,k}stat^*_{k,j} - \alpha_{\ell,k}\hat{s}tat_{k,j}|
\end{aligned}$$

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

$$|\hat{\beta}^T_{k'}\hat{\beta}_k - x_{k'}P| \leq Pa_{k'}(n/L)^{k'-1}\eta + Pa_k(n/L)^{k-1}\eta.$$

Accordingly, we can upperbound

$$\sum_{\ell=1}^{L}\max_{j\in sec_\ell}|num_{k,j} - num^*_{k,j}| \leq P(const_1)(\log M)^{k-1/2}\sqrt{L}\eta$$

where the $const_1$ is equal to $a_k\sqrt{\omega_0}+\sum_{k'=1}^{k-1}\{(c_k\sqrt{\omega_{k'}}+\frac{c}{\sqrt{P}})d_{k'}(\frac{n}{L})^{k'-k}+(a_k+a_{k'}(\frac{n}{L})^{k'-k}+a_{k'-1}(\frac{n}{L})^{k'-k-1})/den^*_{k'}\}$. For denominator,

$$|den_k^2-(den_k^*)^2|\leq |\|\hat{\beta}_k\|^2-\sum_{k'=0}^{k-1}(b_{k'}^T\hat{\beta}_k)^2-\lambda_{k,k}c_k|.$$

Note that $\lambda_{k,k}c_k=(c_0-c_k)-(c_0-c_k)^2\omega_0-\sum_{k'=1}^{k-1}c_k^2\omega_{k'}$. Recall that $(b_{k'}^T\hat{\beta}_k)$ was close to $c_k\sqrt{\omega_{k'}}$ for $k'=1,\ldots,k-1$ and $(b_0^T\hat{\beta}_k)$ is close to $(c_0-c_k)\sqrt{\omega_0}$. Thus,

$$\begin{aligned}&|den_k^2-(den_k^*)^2|\leq |\|\hat{\beta}_k\|^2-\sum_{k'=0}^{k-1}(b_{k'}^T\hat{\beta}_k)^2-\lambda_{k,k}c_k|\\&\leq |\|\hat{\beta}_k\|^2-(c_0-c_k)|+|(b_0^T\hat{\beta}_k)^2-(c_0-c_k)^2\omega_0|\\&+\sum_{k'=1}^{k-1}|(b_{k'}^T\hat{\beta}_k)^2-c_k^2\omega_{k'}|\\&\leq (const_2)(\log M)^{k-1/2}\eta\end{aligned}$$

where $(const_2)=(P+2P^2/c_0)a_k+\sum_{k'=1}^{k-1}\{(c\sqrt{P}+c_k\sqrt{\omega_{k'}})\{c/\sqrt{P}d_{k'}(n/L)^{k'-k}+(a_k+a_{k'}(n/L)^{k'-k}+a_{k'-1}(n/L)^{k'-k-1})/den^*_{k'}\}\}$.

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

Accordingly, $\sum_{\ell=1}^{L}\max_{j\in sec_\ell}|b_k-b_k^*|\leq d_k(n/L)^{k-1}\sqrt{L}\eta$ 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

$$|\mathcal{Z}_k-\mathcal{Z}_k^*| \;=\; \sqrt{n}|b_k-b_k^*|.$$

Thus, $\sum_{\ell=1}^{L}\max_{j\in sec_\ell}|\mathcal{Z}_k-\mathcal{Z}_k^*|$ is bounded above by $\sqrt{nL}d_k(N/L)^{k-1}\eta$

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

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

where $A_k = \sqrt{1-\lambda_{k,k}^2} A_{k-1}/(n/L) + \lambda_{k,k} d_k$. Finally, we evaluate the difference of square norm and success rate for $\hat{\beta}_{k+1}$. We have

$$\begin{aligned}
&\max\left(\left| \frac{\beta^T \hat{\beta}_{k+1}}{P} - \frac{\beta^T \beta_{k+1}^*}{P} \right|, \left| \frac{\|\hat{\beta}_{k+1}\|^2}{P} - \frac{\|\beta_{k+1}^*\|^2}{P} \right| \right) \\
\le \; & 4c^3 \sqrt{P}\, A_k (n/L)^k \eta \le (a_{k+1}/2)(n/L)^k \eta
\end{aligned}$$

where $a_{k+1} = 4c^3 \sqrt{P}\, A_k$. If we set δ_{k+1} in Cor 6 as $a_{k+1}/2$, we have

$$\max\left(\left| \frac{\beta^T \hat{\beta}_{k+1}}{P} - x_{k+1} \right|, \left| \frac{\|\hat{\beta}_{k+1}\|^2}{P} - x_{k+1} \right| \right) \le a_{k+1} (n/L)^k \eta$$

except an event of probability not greater than

$$\sum_{k'=1}^{k+1} 6(k'+1) \exp\{-\frac{2}{c^2} L \delta_{k'}^2\}$$

as desired. This completes the proof.

B.7 Proof of Lemma 15.

The initial step is a special case since we know the initial c_0 and we are not combining any other orthogonal components here. We have an idealized form for the initial statistics,

$$stat_0 = \mathcal{Z}_0 = \sqrt{n}b_0 + Z_0 = \sqrt{\frac{n}{c_0}}\beta + Z_0$$

Using the Lemma 10, we have

$$\mathbb{Q}\{A_1\} \leq 4\exp(-\frac{L}{8c^2}\eta^2)$$

Next, we prove inductively for $k = 2, \ldots, k^*$. As we have seen above $stat_k$ has a form of

$$stat_k = \sqrt{\frac{n}{\hat{c}_k}}\beta + Z_k^{comb}$$

This is an approximate form of the idealized statistics. We define $\hat{\alpha}_{k,\ell} = \sqrt{nP_\ell/\hat{c}_k}$ and the weight for the term j in sec_ℓ will be

$$w_{k+1,j} = \frac{e^{\hat{\alpha}_{k,\ell} stat_{k,j}}}{\sum_{j' \in sec_\ell} e^{\hat{\alpha}_{k,\ell} stat_{k,j'}}}$$

We want to consider the above estimate as $\beta^*(x, \underline{\lambda})$ in Lemma 10 with x as $\hat{x}_k$ instead of x_k since we know that $\hat{x}_k$ is not far from x_k with high probability from the previous steps. However, $\hat{x}_k$ is random and $\{\hat{\lambda}^*_{k,k'}\}_{k'=0}^{k}$ is also random.

We prove the assertion by the following steps. First, Consider a fine grid on $[x_k - a_k\eta, x_k + a_k\eta]$ with width $d_k\eta$ for d_k as a function of k and M which will be specified later in the proof. Also, consider another fine grid on a surface of a unit ball with $const_k\eta$ close to $\underline{\lambda}_k^*$ in ℓ_∞-norm for some $const_k$ specified later in the proof. For given $\hat{x}_k$ and $\underline{\lambda}_k$, we restrict the quantities to the grid points by rounding up. This permits union bounds to show that, for determination of whether events are exponentially unlikely, it suffices to treat it as deterministic value for each grid point. The estimate which is restricted to the grid points, will be denoted $\hat{\beta}_k^r$ and the each estimates with replacement of each grid point instead of $\tilde{x}_k$ and $\hat{\underline{\lambda}}_k$ will be denoted

$\hat{\beta}_k(grid_i)$ for each grid point $grid_i$. Here, we will show that error from the rounding is small by controlling the size of grid. Next, we claim that for each grid point, the quantity regarding $\hat{\beta}_k(grid_i)$ is not far from their expectation using Lemma 10. Finally, we show that each expectation of the grid point is not far from $x_{k+1}P$ using the Lipschitz condition as we assumed.

Define

$$A_{k,1} = \{\left|\frac{\beta^T\hat{\beta}_{k+1}}{P} - \frac{\beta^T\hat{\beta}^r_{k+1}}{P}\right| > \frac{1}{2}\eta\} \cup \{\left|\frac{\|\beta-\hat{\beta}_{k+1}\|^2}{P} - \frac{\|\beta-\hat{\beta}^r_{k+1}\|^2}{P}\right| > \frac{1}{2}\eta\}$$

and

$$\begin{aligned} A_{k,2} &= \{\left|\frac{\beta^T\hat{\beta}^r_{k+1}}{P} - x_{k+1}\right| > (\frac{1}{2} + c_{Lip}a_k)\eta\} \\ &\qquad \cup\{\left|\frac{\|\beta-\hat{\beta}^r_{k+1}\|^2}{P} - (1-x_{k+1})\right| > (\frac{1}{2} + c_{Lip}a_k)\eta\}. \end{aligned}$$

so that we can write $A_k = A_{k,1} \cup A_{k,2}$. We define $S_k = \{\sum_{\ell=1}^{L} max_{j\in sec_\ell}|Z_{k,j}| > 4L\log M\}$ and $S_0^k = \cup_{k'=1}^{k} S_{k'}$. We show inductively that

$$A_1^k \subseteq A_1 \cup \left(\cup_{k'=2}^{k} A_{k',2}\right) \cup S_0^{k-1}.$$

We know that from Cor 4 $\mathbb{Q}\{S_0^{k-1}\} \leq 2k\exp(-L\log M)$ and we next show for each k',

$$\mathbb{Q}\{A_{k',2}\} \leq 4Gr_{k'}\exp(-\frac{L}{8c^2}\eta^2)$$

so that we can conclude as desired.

First, we show that for any $k = 2, \ldots, k^*$ we have

$$A_{k,1} \subseteq A_{k-1} \cup N_0^{k-1}$$

Notice that from Lemma 12,

$$\begin{aligned}
& \max\left(\left|\beta^T\hat{\beta}_{k+1}/P - \beta^T\hat{\beta}^r_{k+1}/P\right|, \left|\|\beta - \hat{\beta}_{k+1}\|^2/P - \|\beta - \hat{\beta}^r_{k+1}\|^2/P\right|\right) \\
\le\ & 4\sum_{\ell=1}^{L}\frac{P_\ell}{P}\max_{j\in sec_\ell}\left|\hat{\alpha}_\ell \hat{stat}_{k,j} - \hat{\alpha}^r_\ell \hat{stat}^r_{k,j}\right| \\
\le\ & 4\sum_{\ell=1}^{L}\frac{P_\ell}{P}\max_{j\in sec_\ell}\left|\hat{\alpha}^*_\ell(\hat{\alpha}^*_\ell 1_{\{j=j_\ell\}} + Z^{comb}_{k,j}) - \hat{\alpha}^r_\ell(\hat{\alpha}^r_\ell 1_{\{j=j_\ell\}} + Z^{comb,r}_{k,j})\right| \\
\le\ & \underbrace{4\sum_{\ell=1}^{L}\frac{P_\ell}{P}|(\hat{\alpha}^*_\ell)^2 - (\hat{\alpha}^r_\ell)^2|}_{(D)} + \underbrace{4\sum_{\ell=1}^{L}\frac{P_\ell}{P}|\hat{\alpha}^*_\ell - \hat{\alpha}^r_\ell|\max_{j\in sec_\ell}|Z^{comb,r}_{k,j}|}_{(E)} \\
& \qquad + \underbrace{4\sum_{\ell=1}^{L}\frac{P_\ell}{P}\hat{\alpha}^*_\ell\max_{j\in sec_\ell}\left|\sum_{k'=0}^{k}|\hat{\lambda}^*_{k,k'} - \lambda^r_{k,k'}|Z_{k',j}\right|}_{(F)}
\end{aligned}$$

Suppose we have

(a) $|\hat{x}_k - \hat{x}^r_k| \le d_k\eta$ where $d_k = \frac{L}{n}\frac{1}{24c^2 snr}\min(\frac{1}{snr}, \frac{1}{3(k+1)c\sqrt{R\,snr}})$,

(b) $\max_{k'=0,\ldots,k}|\hat{\lambda}^*_{k,k'} - \lambda^r_{k,k'}| \le const_k\eta$ where $const_k = \frac{n}{24L}\frac{1}{4(k+1)c^3\sqrt{R\,snr}}$

(c) $\sum_{\ell=1}^{L} max_{j\in sec_\ell}|Z_{k',j}| \le 4L\log M$ for $k' = 0,\ldots,k$.

We can check the above three conditions for every $A_{k+1,1}$ which evaluates whether $\hat{\beta}_{k+1}$ is close enough to the $\hat{\beta}^r_{k+1}$. The condition (a) is satisfied when $\hat{x}_k$ lies inside of the interval $[x_k - a_k\eta, x_k + a_k\eta]$ which is equivalent to $|\hat{x}_k - x_k| \le a_k\eta$. The condition (b) is satisfied for all $\underline{\hat{\lambda}}_k$ and (c) is satisfied on the event of $(N^k_0)^c$. Using the three conditions, we can upperbound $4((D)+(E)+(F))$ by $\frac{1}{2}\eta$ as following. Also, we can say that $A_{k,1} \subseteq A_{k-1} \cup S^{k-1}_0$ For (D),

$$
\begin{aligned}
\sum_{\ell=1}^{L} P_\ell |(\hat{\alpha}_\ell)^2 - (\hat{\alpha}_\ell^r)^2| &\leq \frac{n(\max_\ell P_\ell)}{(\sigma^2)^2} |\hat{c}_k - \hat{c}_k^r| \sum_{\ell=1}^{L} P_\ell \leq \frac{nc^2\, snr^2}{L} |\hat{x}_k - \hat{x}_k^r| \\
&\leq \frac{nc^2\, snr^2}{L} d_k \eta \leq \frac{1}{24}\eta.
\end{aligned}
$$

For (E),

$$
\begin{aligned}
\sum_{\ell=1}^{L} P_\ell |\hat{\alpha}_\ell - \hat{\alpha}_\ell^r| \max_{j \in sec_\ell} |Z_{k,j}^{comb,r}| &\leq (\max_\ell P_\ell |\hat{\alpha}_\ell - \hat{\alpha}_\ell^r|) \sum_{k'=0}^{k} \hat{\lambda}_{k,k'} \sum_{\ell=1}^{L} \max_{j \in sec_\ell} |Z_{k,j}| \\
&\leq \frac{4(k+1)c^3 P\sqrt{P}}{\sigma^2\sqrt{\sigma^2}} \sqrt{n/L}\sqrt{\log M} |\hat{x}_k^* - \hat{x}_k^r| \\
&\leq 4(k+1)c^3\, snr^{3/2} d_k \sqrt{n/L}\sqrt{\log M}\, \eta \leq \frac{1}{24}\eta.
\end{aligned}
$$

For (F),

$$
\begin{aligned}
\sum_{\ell=1}^{L} P_\ell \hat{\alpha}_\ell^* \max_{j \in sec_\ell} \left| \sum_{k'=0}^{k} |\hat{\lambda}_{k,k'}^* - \lambda_{k,k'}^r| Z_{k',j} \right| &\leq const_k \eta (\max_\ell P_\ell \hat{\alpha}_\ell^*) \sum_{k'=0}^{k} \sum_{\ell=1}^{L} \max_{j \in sec_\ell} |Z_{k',j}| \\
&\leq const_k 4(k+1)c^3 \sqrt{snr}\sqrt{n/L}\sqrt{\log M}\, \eta \\
&\leq \frac{1}{24}\eta.
\end{aligned}
$$

Thus, we can conclude,

$$
\begin{aligned}
&\max\left(\left| \beta^T \hat{\beta}_{k+1} - \beta^T \hat{\beta}_{k+1}^r \right|, \left| \|\beta - \hat{\beta}_{k+1}\|^2/P - \|\beta - \hat{\beta}_{k+1}^r\|^2/P \right| \right) \\
\leq\ & 4\left((D) + (E) + (F)\right) \\
\leq\ & \frac{1}{2}\eta
\end{aligned}
$$

For $k = 2$, since we have $A_{2,1} \subseteq A_1 \cup S_0^1$,

$$\begin{aligned} A_1^2 &\subseteq A_1 \cup A_{2,1} \cup A_{2,2} \\ &\subseteq A_1 \cup S_0^1 \cup A_{2,2} \end{aligned}$$

This completes the proof for $k = 2$. Next, we suppose that this is true for the first k step and prove for the $(k+1)$ step. We know that

$$\begin{aligned} A_1^{k+1} &\subseteq A_1^k \cup A_{k+1,1} \cup A_{k+1,2} \\ &\subseteq A_1^k \cup S_0^k \cup A_{k+1,2} \end{aligned}$$

Since, we assumed that $A_1^k \subseteq A_1 \cup \left(\cup_{k'=2}^{k} A_{k',2}\right) \cup S_0^{k-1}$, we have

$$A_1^{k+1} \subseteq A_1^k \cup S_0^k \cup A_{k+1,2} \subseteq A_1 \cup \left(\cup_{k'=2}^{k+1} A_{k',2}\right) \cup S_0^k$$

as we desired.

Next, we show that for each k,

$$\mathbb{Q}\{A_{k,2}\} \leq 4Gr_k \exp(-\frac{L}{8c^2}\eta^2)$$

We use union bound over the grids as following,

$$\begin{aligned}
&\mathbb{Q}\{\Big|\|\beta-\hat{\beta}^r_{k+1}\|^2/P-(1-x_{k+1})\Big|>(1/2+c_{Lip}a_k)\eta\} \\
\leq\ & \sum_{grid_i}\mathbb{Q}\{\Big|\|\beta-\hat{\beta}_{k+1}(grid_i)\|^2/P-(1-x_{k+1})\Big|>(1/2+c_{Lip}a_k)\eta\} \\
\leq\ & \sum_{grid_i}\Big[\mathbb{Q}\{\Big|\|\beta-\hat{\beta}_{k+1}(grid_i)\|^2/P-\mathbb{E}\|\beta-\hat{\beta}_{k+1}(grid_i)\|^2/P\Big|>\frac{1}{2}\eta\} \\
&\qquad\qquad +\mathbb{Q}\{\Big|\mathbb{E}\|\hat{\beta}_{k+1}(grid_i)\|^2/P-(1-x_{k+1})\Big|>c_{Lip}a_k\eta\}\Big]
\end{aligned}$$

The first sum will be bounded by the error probability in Cor.6 multiplied by the number of grid point. For the second term of above equation, we know that for each grid point, x and λ is deterministic. It allows us to say that the combined normal random variable is distributed standard normal. Thus, we can write $\mathbb{E}\|\hat{\beta}_{k+1}(grid_i)\|^2$ as $g(x_k(grid_i))$. By the Lipschitz condition as we assumed, $|g(x_k(grid_i))-g(x_k)|\leq c_{Lip}|x_k(grid_i)-x_k|\leq c_{Lip}a_k\eta$ so that the probability of the event is zero. For simplicity, we showed the assertion only with $\|\beta-\hat{\beta}^r_k\|^2$. However, the same assertion works with the union set regarding $\beta^T\hat{\beta}^r_k$.

Thus,

$$\begin{aligned}
\mathbb{Q}\{A^k_1\} &\leq \mathbb{Q}\{A_1\cup\big(\cup^k_{k'=2}A_{k',2}\big)\cup S^{k-1}_0\} \\
&\leq \mathbb{Q}\{A_1\}+\sum_{k'=2}^{k}\mathbb{Q}\{A_{k',2}\}+\mathbb{Q}\{S^{k-1}_0\} \\
&\leq (4\sum_{k'=1}^{k}Gr_{k'})\exp(-\frac{L}{8c^2}\eta^2)+2k\exp(-L\log M)
\end{aligned}$$

Here $Gr_1 = 1$. For $k = 2, \ldots, k^*$, we have $Gr_k =$ (#of grid) and

$$(\#\text{of grid for } x_k) \times (\#\text{of grid for } \underline{\lambda}) \leq \frac{2c_{Lip} a_k \eta}{d_k \eta} \times \left(\frac{1}{const_k \eta} \right)^k = Const_k \left(\frac{n}{L\eta} \right)^{k+1}$$

B.8 Proof of Lemma 5

First, we show the function $g_{low}(x) - x$ is monotone non-decreasing which is equivalent to $g'_{low} \leq 1$. We can represent the update function $g(x)$ as

$$\begin{aligned} \mathbb{P}_{V,U}\{\alpha(U) \geq V\} &= \mathbb{P}\{\alpha^2 \geq (V_+)^2\} \\ &= \mathbb{P}_{V,U}\{U \leq 1 + \frac{1}{snr} - (1 + \frac{1}{snr} - x)\frac{(V_+)^2}{\tau^2}\frac{R}{\tilde{C}}\} \\ &= \mathbb{E}_V \max(1, (-\frac{(V_+)^2}{\tau^2}\frac{R}{\tilde{C}}(1 + \frac{1}{snr} - x) + 1 + \frac{1}{snr})_+). \end{aligned}$$

If we take derivative with respect to x, we have

$$\mathbb{E}_V \frac{(V_+)^2}{\tau^2}\frac{R}{\tilde{C}} 1_{\tilde{B}},$$

where

$$\tilde{B} = \{-\frac{(V_+)^2}{\tau^2}\frac{R}{\tilde{C}}(1 + \frac{1}{snr} - x) + 1 + \frac{1}{snr} \in [0,1]\} = \{\alpha_1(x) \leq V \leq \alpha_0(x)\}$$

where $\alpha_0(x) = \alpha(u = 0, x)$ and $\alpha_1(x) = \alpha(u = 1, x)$. We can check the monotonicity by evaluating the above equation. Since we are taking expectation with some positive random variable, the expectation with the indicator function is not more than the one without it. Also, we use the fact that $(V+)^2 \leq V^2$. Thus, we have

$$g'_{low}(x) = \mathbb{E}_{V_{low}} \frac{((V_{low})_+)^2}{\tau^2}\frac{R}{\tilde{C}} 1_{\tilde{B}} \leq \frac{R}{\tilde{C}\tau^2}\mathbb{E}V_{low}^2 1_{\tilde{B}} \leq \frac{R}{\tilde{C}\tau^2}\mathbb{E}V_{low}^2$$

Now, we evaluate $\mathbb{E}V_{low}^2$. Note that the expectation of odd power of standard normal distribution is zero and so does the logistic distribution. Also, Z and ξ are independent.

$$\begin{aligned}
\mathbb{E}V_{low}^2 &= \mathbb{P}\{2Z^2 + 2\xi + \tau^2 - 2Z\sqrt{(\tau^2 + 2\xi + Z^2)_+}\} \\
&= 2 + \tau^2
\end{aligned}$$

This follows from the fact that $2Z\sqrt{(\tau^2 + 2\xi + Z^2)_+}$ is an odd function in Z around zero. Thus, we have

$$g'_{low}(x) \leq \frac{R}{\tilde{C}}(1 + \frac{2}{\tau^2})$$

The above quantity is not greater than 1 when $R \leq \tilde{C}/(1 + \frac{2}{\tau^2})$.

Next, we evaluate the crossing point. We denote V_a as V for simplicity. We can alternatively express $g_{low}(x)$ as

$$\begin{aligned}
g_{low}(x) &= \mathbb{E}_V \max(1, (-\frac{(V_+)^2}{\tau^2}\frac{R}{\tilde{C}}(1 + \frac{1}{snr} - x) + 1 + \frac{1}{snr})_+) \\
&= \mathbb{E}(-\frac{(V_+)^2}{\tau^2}\frac{R}{\tilde{C}}(1 + \frac{1}{snr} - x) + 1 + \frac{1}{snr})1_{\tilde{B}} \\
&\qquad\qquad + \mathbb{P}\{-\frac{(V_+)^2}{\tau^2}\frac{R}{\tilde{C}}(1 + \frac{1}{snr} - x) + 1 + \frac{1}{s\tilde{n}r} > 1\} \\
&= (1 + \frac{1}{snr})\mathbb{P}\tilde{B} - (1 + \frac{1}{snr} - x)\mathbb{E}(\frac{(V_+)^2}{\tau^2}\frac{R}{\tilde{C}})1_{\tilde{B}} + \mathbb{P}\{\alpha_1(x) \geq V_+\} \\
&= g(0,x) + \frac{1}{snr}\mathbb{P}\tilde{B} - (1 + \frac{1}{snr} - x)\frac{R}{\tilde{C}}\mathbb{E}(\frac{(V_+)^2}{\tau^2})1_{\tilde{B}} \\
&= g(0,x) - \frac{R}{\tilde{C}}\left(1 + \frac{1}{snr} - x\right)\mathbb{E}(\frac{(V_+)^2 - a(x)\tau^2}{\tau^2})1_{\tilde{B}}
\end{aligned}$$

with $a(x) = \frac{\tilde{C}}{R}\frac{1/snr}{1+1/snr-x}$. Let's first consider $1 - y^* = (\frac{\tilde{C}}{R} - 1)/snr$ which is equivalent to

$$\left(1 + \frac{1}{snr} - y^*\right)\frac{R}{\tilde{C}} = \frac{1}{snr}$$

which makes $B^* \subseteq \tilde{B}(x = y^*)$ and $a(y^*) = 1$. Thus, we have

$$\begin{aligned} 1 - g_{low}(y^*) &= \{1 - g_{low}(0, y^*)\} + \frac{1}{snr} \frac{\mathbb{E}((V_+)^2 - \tau^2) 1_{\tilde{B}(x=y^*)}}{\tau^2} \\ &\leq \{1 - g_{low}(0, y^*)\} + \frac{1}{snr} \frac{\mathbb{E}((V_+)^2 - \tau^2) 1_{B^*}}{\tau^2} \end{aligned}$$

The inequality comes from the fact that $B^* \subseteq \tilde{B}(x = y^*)$ and $(V_+)^2 > \tau^2$ when the indicator function is one. By definition, the second term is $drop^*/snr$ which is not greater than $1 - y^*$. For y^* near 1, the first term is exponentially small.

Next, let's consider term $y^* < x^* < 1$ to find more accurate crossing point where

$$1 - x^* = \min\left(\frac{R}{C}\frac{drop^*}{snr}(1 - \frac{R}{C}drop^*)^{-1}, 1 - y^*\right)$$

which gives us $1 \leq a(x^*) \leq \tilde{C}/R$ and $B^* \subseteq \tilde{B}(x = x^*)$. Similar to the argument for $1 - y^*$,

$$\begin{aligned} 1 - g_{low}(x^*) &= \{1 - g_{low}(0, x^*)\} + (1 - x^*) \\ &\leq \{1 - g_{low}(0, y^*)\} + (1 - x^*) \end{aligned}$$

Thus, we can say that $g(x^*) \geq x^* - e$ with e is an order of $(1/M)$ which will be shown at the end of the proof. To evaluate the first term $\{1 - g_{low}(0, y^*)\}$, we have a lemma to evaluate the gap between $g(u, x)$ and one when $u < x$.

Lemma 22. *Suppose our current success rate is s^*. Then, for $u(\ell) = s^* - \delta$ with $\delta > 0$, the gap between one and $g(u(\ell), x)$ is polynomially small in M.*

Proof. For $u(\ell) = s^* - \delta$, we have α larger than τ. We can also write α as $\alpha = \tau\sqrt{\frac{1+1/snr-u(\ell)}{1+1/snr-s^*}} = \tau\sqrt{1 + \frac{\delta}{1+1/snr-s^*}}$. If we write $\alpha = \tau\sqrt{1 + \tilde{\delta}}$ with $\tilde{\delta} = \delta/(1 + 1/snr - s^*)$. Note that $F(\xi)$ is less than e^{ξ} for $\xi < 0$ and $1 - \frac{1}{2}e^{-\xi}$ for $\xi > 0$. Also, we will use the fact that the tail probability of normal distribution $\bar{\Phi}(z)$ is bounded above with $\frac{1}{2}e^{-z^2/2}$ for $z > 0$.

$$\begin{aligned}
& 1 - g_{low}(u(\ell), x) \\
= \; & \mathbb{P}_{\xi,Z}\{\xi \geq \alpha^2/2 - \tau^2/2 + \alpha Z\} \\
= \; & \mathbb{P}_Z F(-\alpha^2/2 + \tau^2/2 - \alpha Z) \\
= \; & \int_{Z<-\frac{y}{\alpha}} F(-\frac{\alpha^2}{2} + \frac{\tau^2}{2} - \alpha Z)\phi(z) + \int_{Z>-\frac{y}{\alpha}} F(-\frac{\alpha^2}{2} + \frac{\tau^2}{2} - \alpha Z)\phi(z) \\
\leq \; & \int_{Z<-\frac{y}{\alpha}} (1 - \frac{1}{2}e^{\alpha^2/2-\tau^2/2+\alpha Z})\phi(z) + \int_{Z>-\frac{y}{\alpha}} e^{-\alpha^2/2+\tau^2/2-\alpha Z}\phi(z) \\
\leq \; & \exp\{-\frac{\tau^2}{8}(\frac{d^2-1}{2d})^2\} \leq \exp\{-\frac{\tau^2}{8}\tilde{\delta}\}
\end{aligned}$$

where $d^2 = \alpha^2/\tau^2$.

Here, we have $u = 0$ and $x = x^*$ so that $\tilde{\delta} > snr$. Thus, we can say that it is exponentially small in τ. This completes the proof of Lemma 22.

Thus, to show that the crossing point is on the right side of x^*, we need to show that the gap $1 - g(x^*)$ less than $1 - x^*$. The above argument is true for V_a as well as V_{low}. To evaluate the order of $drop^*$, we consider when $V = V_{low}$. Then, we have

$$drop^* = \frac{\mathbb{E}((V_+)^2 - \tau^2)1_{B^*}}{\tau^2} = \frac{2\mathbb{E}(Z^2 + \xi - Z\sqrt{(\tau^2 + 2\xi + Z^2)_+})B^*}{\tau^2},$$

Using repetitive Cauchy-Swartz inequality, we can get the above equation is less than $(\tau + 9)/\tau^2$. Thus, we can see that it would be order of $1/\tau$.

Bibliography

Abbe, E. and A. Barron (2011). Polar coding schemes for the AWGN channel. In *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on*, pp. 194–198.

Akaike, H. (1969). Statistical predictor indentification. *Ann. Inst. Statist. Math 21*, 243–247.

Akaike, H. (1970). Fitting autoregressive models for prediction. *Ann. Inst. Statist. Math 22*, 203–217.

Arikan, E. (2009). Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels. *Information Theory, IEEE Transactions on 55*(7), 3051–3073.

Arikan, E. and E. Telatar (2009). On the rate of channel polarization. In *Information Theory, 2009. ISIT 2009. IEEE International Symposium on*, pp. 1493–1495.

Barron, A., L. Birg, and P. Massart (1999). Risk bounds for model selection via penalization. *Probability theory and related fields 113*(3), 301413.

Barron, A. R. (1991). *Complexity regularization with application to artificial neural networks.* Kluwer Academic Publisher.

Barron, A. R. and A. Joseph (2010). Sparse superposition codes: Fast and reliable at rates approaching capacity with gaussian noise. *Manuscript. Available at http://www. stat. yale. edu/ arb4*.

Bayati, M. and A. Montanari (2011). The dynamics of message passing on dense graphs, with applications to compressed sensing. *Information Theory, IEEE Transactions on 57*(2), 764–785.

Bayati, M. and A. Montanari (2012). The LASSO risk for gaussian matrices. *Information Theory, IEEE Transactions on 58*(4), 1997–2017.

Breiman, L. and D. Freedman (1983). How many variables should be entered in a regression equation? *Journal of the American Statistical Association 78*(381), 131–136.

Cover, T. M. and J. A. Thomas (2012). *Elements of information theory.* John Wiley & Sons.

Craven, P. and G. Wahba (1978). Smoothing noisy data with spline functions. estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math 31*, 377–403.

Daniel, C. and F. Wood (1971). *Fitting equations to data: computer analysis of multifactor data for scientists and engineers.* Wiley series in probability and mathematical statistics. Applied probability and statistics. Wiley-Interscience.

Donoho, D. L., A. Maleki, and A. Montanari (2009). Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences 106*(45), 18914–18919.

Golub, G. H., M. Heath, and G. Wahba (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics 21*(2), 215–223.

Hocking, R. R. (1976). A biometrics invited paper. the analysis and selection of variables in linear regression. *Biometrics 32*(1), 1–49.

Joseph, A. and A. Barron (2012). Least squares superposition codes of moderate dictionary size are reliable at rates up to capacity. *IEEE Transactions on Information Theory 58*(5), 2541–2557.

Joseph, A. and A. Barron (2014). Fast sparse superposition codes have near exponential error probability for $R < C$. *Information Theory, IEEE Transactions on 60*(2), 919–942.

Leeb, H. (2008). Evaluation and selection of models for out-of-sample prediction when the sample size is small relative to the complexity of the data-generating process. *Bernoulli 14*(3), 661–690.

Li, K.-C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set. *The Annals of Statistics 15*(3), 958–975.

Maleki, A. and D. L. Donoho (2010). Optimally tuned iterative reconstruction algorithms for compressed sensing. *Selected Topics in Signal Processing, IEEE Journal of 4*(2), 330–341.

Mallows, C. L. (1973). Some comments on C_P. *Technometrics 15*(4), 661–675.

McFadden, D. and P. Zarembka (1974). Conditional logit analysis of qualitative choice behavior. In *Frontiers in econometrics.*

Rush, C. and A. R. Barron (2013). Using the method of nearby measures in superposition coding with a bernoulli dictionary. *Proc. Workshop on Information Theoretic Methods in Science and Engineering*.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika 68*(1), 45–54.

Thompson, M. L. (1978). Selection of variables in multiple regression: Part i. a review and evaluation. *International Statistical Review / Revue Internationale de Statistique 46*(1), 1–19.