Abstract

Iterative Algorithms for Inference and Optimization, with Applications in Communications and Compressed Sensing

Cynthia Rush

2016

This work studies the high-dimensional statistical linear regression model,

$$Y = X\beta + \epsilon, \tag{1}$$

for output $Y \in \mathbb{R}^n$, design matrix $X \in \mathbb{R}^{n \times N}$, noise $\epsilon \in \mathbb{R}^n$, and unknown message $\beta \in \mathbb{R}^N$ when N larger than the sample size n. The aim is to recover the message with knowledge of the output Y, the design X, and the distribution of the noise ϵ. In the high-dimensional setting, it is necessary that β have an underlying structure for successful recovery to be possible. We study this problem under two different assumptions on the distributional properties of the unknown message β motivated by practical applications.

The first application studied is communication over a noisy channel. We propose Approximate Message Passing, or AMP, as a fast decoding strategy for sparse regression codes, introduced by Barron and Joseph [1, 2]. We prove that this scheme is asymptotically capacity-achieving with error probabilities approaching zero in the large system limit and good empirical performance at practical block lengths.

In many applications, one wishes to study the model given in (1.1), when the only assumption made on the message β is that its entries are i.i.d. according to some prior distribution. In this case Approximate Message Passing, or AMP, has been proposed [4–8] as a fast, iterative algorithm to recover β. In [6] it is shown that the performance of AMP can be characterized in the large system limit, meaning as $n, N \to \infty$ simultaneously, via a simple scalar iteration called *state evolution*. This dissertation analyzes the finite-sample performance of AMP, demonstrating that state evolution still accurately characterizes the algorithm's performance for practically-sized n.

Iterative Algorithms for Inference and Optimization, with Applications in Communications and Compressed Sensing

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

by
Cynthia Rush

Dissertation Director: Andrew Barron

May, 2016

Contents

List of Figures

Acknowledgements

A lot of people are awesome.

Chapter 1

Introduction

This dissertation studies the high-dimensional statistical linear regression model,

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

for output $Y \in \mathbb{R}^n$, design matrix $X \in \mathbb{R}^{n \times N}$, measurement noise $\epsilon \in \mathbb{R}^n$, and unknown message vector $\beta \in \mathbb{R}^N$. This model is high-dimensional in that the dimension N is possibly larger than the sample size n.

Our aim is to recover the unknown message vector with knowledge of the output Y, the design X, and the distribution of the noise ϵ, such that the estimate we produce of the unknown message, labeled $\hat{\beta}$, is close to the true message in some sense, for example, in ℓ_2 distance. In the high-dimensional setting, it is necessary that β have some underlying structure, such as sparsity, for successful recovery to be possible. In what follows we study this problem under two different assumptions on the distributional properties of the unknown message β motivated by practical applications.

1.1 Communications

In the communications problem, we wish to create practical encoding and decoding schemes to reliably communicate information over a noisy channel. Using sparse regression codes, introduced by Barron and Joseph [1, 2], it is possible analyze the channel coding problem

using the statistical framework of high-dimensional regression modeled in (1.1), with β assumed to be $L-$sparse, meaning β has some number of non-zero values L that is small compared to its length N.

In this framework, theoretical bounds on the rate at which information can be transmitted across a channel correspond to lower bounds on the sample size n necessary for successful support recovery. The goal then of the communications is to recover β with exponentially small probability of error in the sample size n, for any n greater than this theoretical minimum value, with the additional knowledge that β belongs to some known, finite set. Practically such recovery must be computationally efficient.

In [1], Barron and Joseph demonstrate that the maximum likelihood decoder, corresponding to the least squares decoder, is theoretically optimal but impractical. Barron and Joseph [2] and Barron and Cho [3] additionally proposed efficient, asymptotically capacity-achieving iterative decoding schemes with exponentially small error probabilities. Despite the strong theoretical guarantees, the rates that are achievable at practical block lengths with these decoders are much less than capacity.

In this dissertation we propose Approximate Message Passing, or AMP, as a fast decoding strategy that is provably asymptotically capacity-achieving with error probabilities approaching zero in the large system limit and good empirical performance at practical block lengths.

1.2 Compressed Sensing and Other Applications

In many applications, one wishes to study the high-dimensional regression model given in (1.1), when the only assumption made on the message β is that its entries are i.i.d. according to some prior distribution. When this is the case Approximate Message Passing, or AMP, has been proposed [4–8] as a fast, iterative algorithm to recover β. AMP is derived as an approximation to loopy belief propagation algorithms, like min-sum or sum-product, but meant for problems with dense factor graph representation corresponding to (1.1).

When the design X is Gaussian, the performance of AMP in the large system limit, meaning as $n, N \to \infty$ simultaneously (with n/N constant) has been analyzed in [6]. In their

analysis, Bayati and Montenari [6] show that the performance of AMP can be characterized in the large system limit via a simple scalar iteration called *state evolution*. In particular, if $\beta^1, \beta^2, \ldots$ are the estimates produced by AMP, their result implies that performance measures such as the ℓ_2-error $\frac{1}{N}\|\beta^t - \beta\|^2$ and the ℓ_1-error $\frac{1}{N}\|\beta^t - \beta\|_1$ converge almost surely to constants that can be easily computed via the distribution of β.

This dissertation analyzes the finite-sample performance of AMP in this setting, when the design matrix is Gaussian and the under sampling ratio, n/N is constant. We derive a concentration result that implies that probability of deviation the between $\frac{1}{N}\|\beta^t - \beta\|^2$ and its limiting constant value falls exponentially in n.

1.3 Disseration Structure

In Chapter 2 we introduce the communications problem in more detail, and we rigorously analyze the performance of Approximate Message Passing, a computationally-efficient iterative algorithm for recovering β in the communications setting. This work was first presented in [9]. In Chapter 3 we present the rigorous finite-sample analysis of AMP and give examples of its applications. Finally in Chapter 4, we present work aiming to provide an understanding of the performance of iterative decoding schemes for channel communication when the design matrix is equiprobable Bernoulli as opposed to the traditionally-studied Gaussian. This work was originally presented in [10].

Notation: The ℓ_2-norm of the vector x is denoted $\|x\|$. The notation A^* indicates the transpose of matrix A. For a positive integer t, $[t]$ denotes the set $\{1, \ldots, t\}$. Logarithms are denoted as log and ln for base 2 and base e, respectively. The following notation is used for limiting statements: $f(x) = o(g(x))$ means $\lim_{x\to\infty} f(x)/g(x) = 0$.

Chapter 2

Channel Communication with Approximate Message Passing Decoding

Wired and wireless communication using cell phones or smart devices is ubiquitous, creating a pressing need for low-complexity, high data-rate communication schemes. The additive white Gaussian noise (AWGN) channel is a practical model of this sort of communication. The AWGN channel is introduced in Section 2.1. Sparse regression codes, or SPARCs, were introduced as an encoding scheme over the AWGN channel by Barron and Joseph [1, 2], allowing for analysis of the channel coding problem using the framework of high-dimension statistical regression. Section 2.2 introduces SPARCs codes and Section 2.3 introduces decoders that, along with SPARCs, are provably capacity-achieving with small error probabilities in the case of Gaussian dictionaries, or Gaussian design matrices.

This chapter rigorously analyzes the performance of approximate message passing (AMP) as a decoding scheme for the additive white Gaussian noise channel along with sparse regression codes. AMP decoding is proposed as a computationally efficient alternative to the decoders of Section 2.3. In Section 2.6 we prove that the probability of decoding error for AMP goes to zero with growing block length for all fixed rates $R < \mathcal{C}$ and we provide simulation results which demonstrate the strong performance of the decoder at finite block

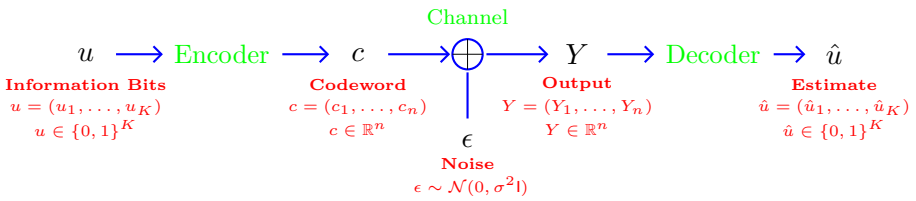

Figure 2.1: Additive White Gaussian Noise Channel

lengths. The approximate message passing algorithm is introduced in Section 2.4 and the form of AMP adapted to SPARCs encoding over the AWGN channel is presented in Section 2.5. Section 2.5 additionally provides some intuition as to how the decoder functions. Finally Section 2.7 provides the proof of the main performance guarantees of the algorithm which are given in Section 2.6. This work was first presented in [9].

2.1 The Additive White Gaussian Noise Channel

The Additive White Gaussian Noise (AWGN) channel is frequently studied as a model of many everyday communications channels including wired and wireless television, satellite, and telephone and is the most basic model of communication of this sort. Noise in such channels may be due to a variety of causes, and by the central limit theorem, the cumulative effect of a large number of small random effects will be approximately normal, so the additive, Gaussian assumption is valid in a large number of situations.

The AWGN channel is shown in Figure 2.1 and basic communication over this channel proceeds as follows. An encoder maps bit strings $u = (u_1, u_2, \ldots, u_K) \in \{0,1\}^K$ of length K (representing the information to be communicated over the channel) into real-valued codewords $c = (c_1, c_2, \ldots, c_n) \in \mathbb{R}^n$ of length n, called the block length. The set of all possible codewords considered by the encoder is called the codebook. The codeword is sent through the channel requiring n transmissions of the discrete-time channel. Generally, the energy, or the power, of the codeword is constrained in some way, and here we consider an average power constraint. The power of the codeword is its ℓ_2-norm and the average power constraint takes the following form $\frac{1}{n}\|c\|^2 = \frac{1}{n}\sum_{i=1}^n c_i^2 < P$ where P is the power constraint. The rate of communication under this scheme is the ratio of the amount of information communicated per channel use, or $R = K/n$, where R stands for the 'rate'.

The decoder receives output $Y = (Y_1, Y_2, \ldots, Y_n) \in \mathbb{R}^n$ which is the sum of the codeword and independent, Gaussian noise:

$$Y_i = c_i + \epsilon_i \text{ for } i = 1, 2, \ldots, n \tag{2.1}$$

with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d. noise and $c = (c_1, c_2, \ldots, c_n)$ the transmitted codeword. With knowledge of the output and the encoding scheme, meaning knowledge of the codebook, the decoder would like to map the output Y into an accurate estimate of the input bit string; this estimate is denoted $\hat{u} = (\hat{u}_1, \ldots, \hat{u}_k)$. A block error is made if the decoding is unsuccessful, meaning if $\hat{u} \neq u$, and a communication rate is considered reliable if, for large n, the probability of a block error is small when averaged over all possible input strings u and the distribution of the output string Y.

The fundamental limit on the rate at which information can be passed over a channel is called the capacity of the channel, which is the supremum over all possible reliable rates of communication over that channel. In the case of the additive Gaussian white noise channel, the capacity is equal to

$$\mathcal{C} = \frac{1}{2} \log_2 (1 + \mathsf{snr}), \tag{2.2}$$

where $\mathsf{snr} = P/\sigma^2$ is the signal-to-noise ratio [11, 12].

The aim of research in the area of channel coding is to produce encoding and decoding schemes with reliable communication rates close to the fundamental limit, the capacity. Moreover, it is required that these schemes be practical to implement computationally, meaning that encoding and decoding computations should be able to proceed rapidly. The work that follows studies the sparse regression coding scheme as a low-complexity, capacity-achieving encoder for communication over the additive white Gaussian noise channel.

2.2 Sparse Regression Codes

Sparse Regression Codes (SPARCs), also called sparse superposition codes, were introduced by Barron and Joseph [1, 2] for communication over the AWGN channel in (2.1). SPARCs are defined in terms of a dictionary or design matrix X of dimension $n \times ML$, with entries

which are i.i.d. $\mathcal{N}(0, \frac{1}{n})$. Here n is the block length, and M, L are positive integers with values specified below in terms of n and the rate R. As demonstrated in Figure 2.2, the dictionary X can be thought of as divided into L sections with M columns each section. SPARCs codewords are constructed as linear combinations of L columns of the dictionary, with one column from each section.

Chapter 4 studies the performance of SPARCs when instead the dictionary X has independent Bernoulli $\pm\frac{1}{\sqrt{n}}$ random variables as entries, meaning they take values in $\left\{+\frac{1}{\sqrt{n}}, -\frac{1}{\sqrt{n}}\right\}$ with equal probability. This change increases the computational efficiently of the scheme. In this chapter, though, we consider the i.i.d. Gaussian dictionary and the results which have been shown in this case.

Formally, a SPARCs codeword is expressed as $X\beta$, where $\beta = (\beta_1, \ldots, \beta_{ML})$ is a vector of length ML with the following property: there is exactly one non-zero β_j for $1 \leq j \leq M$, one non-zero β_j for $M + 1 \leq j \leq 2M$, and so forth. In other words, if we consider β as divided into L sections with M elements in each section, like the dictionary, there is one non-zero value in each section. The non-zero value of β in section $\ell \in \{1, 2, \ldots, L\}$ is set to $\sqrt{nP_\ell}$ called the power allocation, where the positive constants P_ℓ satisfy $\sum_{\ell=1}^{L} P_\ell = P$. Denote the set of all β's that satisfy this property by $\mathcal{B}_{M,L}(P_1, \ldots, P_L)$, which we denote as $\mathcal{B}_{M,L}$ for short when the power allocation is understood.

For simplicity in encoding, assume that M is a power of 2 and that the length of the input $K = L \log_2 M$. The encoder splits its stream of input bits into L sections of $\log_2 M$ bits in each and the decimal equivalent of section ℓ of the input determines the location of the single non-zero value in section ℓ of the vector β. Therefore each possible input string corresponds to a unique subset of columns of the dictionary X used in the linear combination $X\beta$, with one column from each of the L sections of X.

Thus the encoder is a map from the set of all input bit strings, $u \in \{0, 1\}^K$, to the set $\mathcal{B}_{M,L}(P_1, \ldots, P_L)$ and the codewords then take the form

$$X_1\beta_1 + X_2\beta_2 + ... + X_N\beta_N, \tag{2.3}$$

with exactly one column in each of the L sections of the dictionary contributing to the sum.

The received output then follows the familiar statistical linear regression model

$$Y = X\beta + \epsilon \tag{2.4}$$

where $\epsilon \sim N(0, \sigma^2 \mathsf{I})$ independent of the codeword.

Each of the L sections of X contains M columns, so the size of the codebook is M^L. To obtain a communication rate R, we need

$$M^L = 2^{nR} \quad \text{or} \quad L \log M = nR. \tag{2.5}$$

There are several choices for the pair (M, L) which satisfy (2.5). For example, $L = 1$ and $M = 2^{nR}$ recovers the Shannon random codebook for X with 2^{nR} columns. In our construction, we choose $M = L^b$, for some constant $b > 0$. In this case, (2.5) becomes

$$bL \log L = nR. \tag{2.6}$$

This means $L = \Theta(\frac{n}{\log n})$, and the size of the design matrix X (given by $n \times ML = n \times L^{b+1}$) grows polynomially in n.

The power allocation $\{P_\ell\}_{\ell=1}^L$ has been shown to play an important role in determining the performance of various decoders when used with SPARCs encoding. We will consider two different power allocations, or values of $P_{(l)}$ for $l \in L$. In the next section we discuss how constant power allocation, meaning $P_{(l)} = \frac{P}{L}$, has been used to achieve reliable rates up to capacity when least squares decoding is used, and how variable power allocation, meaning $P_{(l)} \propto e^{-\kappa \ell / L}$ for parameter $\kappa > 0$, has been needed to show that all rates up to capacity are reliable when using adaptive successive decoding. We will also show in Section 2.6 that a 'modified' power allocation, which is a combination of the two, gives the best performance when decoding at finite block lengths via Approximate Message Passing. For both power allocations, $P_\ell = \Theta(\frac{1}{L})$ and $\frac{1}{n}||\beta||^2 = \frac{1}{n}\sum_{i=1}^N \beta_i^2 = P$ holds. Therefore, for each $\beta \in \mathcal{B}_{M,L}$ the expected codeword power $\frac{1}{n}\mathbb{E}||X\beta||^2 = P$ (this is true for both the Bernoulli $\pm\frac{1}{\sqrt{n}}$ or the Gaussian dictionary). Moreover, the expected codeword power averaged over

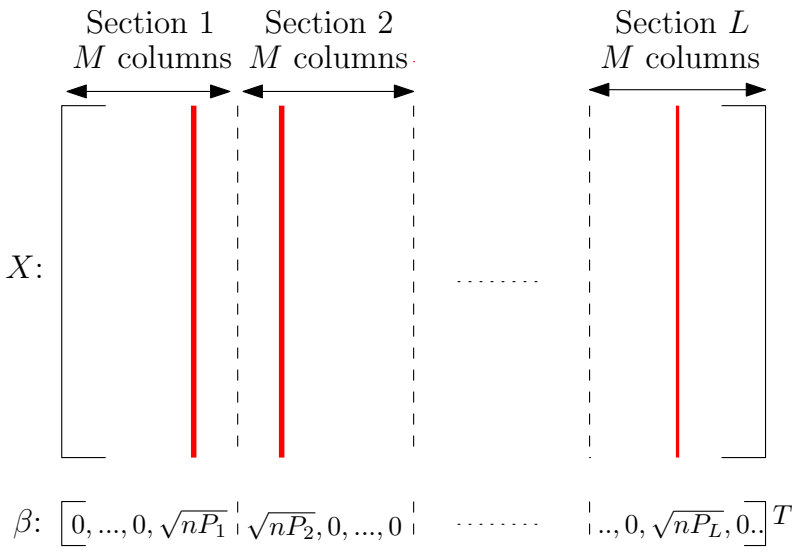

Figure 2.2: X is an $n \times ML$ matrix and β is a $ML \times 1$ vector. The red columns of the dictionary X correspond to the positions of the non-zeros in β. These columns are summed to form the codeword $X\beta$.

all possible codewords,

$$\frac{1}{2^K} \sum_{\beta \in \mathcal{B}_{M,L}} \frac{1}{n} \mathbb{E}||X\beta||^2 = P,$$

and so with high probability the codeword power averaged over all possible codewords is close to P. Both the design matrix X and the power allocation $\{P_\ell\}_{\ell=1}^L$ are known to the encoder and the decoder before communication begins.

2.3 Decoders

In this section we introduce a few of the decoders used with SPARCs encoding over the AWGN channel.

Least Squares Decoding Using a Gaussian design and constant power allocation, Barron and Joseph [13] prove reliable communication at rates approaching capacity using a maximum likelihood decoder, or least squares decoder, with SPARCs encoding. The decoder produces estimates as any $\hat{\beta}$ in the solution set of the following:

$$\hat{\beta} = \arg \min_{\beta \in \mathcal{B}_{M,L}} ||Y - X\beta||^2, \tag{2.7}$$

Joseph and Barron rigorously analyze the performance of this decoder, showing that for

any fixed rate $R < \mathcal{C}$, the probability of decoding error decays to zero exponentially in n, the block length of the code. While theoretically optimal, this scheme is computationally inefficient. A computational improvement is given by Takeishi, Kawakita, and Takeuchi [14]: using a Bernoulli $\pm\frac{1}{\sqrt{n}}$ design, they again prove that least squares decoding is reliable, as conjectured by Joseph and Barron, with exponentially small probability of error.

Adaptive Successive and Soft-Decision Iterative Decoding Barron and Joseph [2] additionally propose an efficient decoding algorithm called 'adaptive successive decoding', for which they show that for any fixed rate $R < \mathcal{C}$, the probability of decoding error decays to zero exponentially in $\frac{n}{\log n}$, where n is the block length of the code. Despite the strong theoretical performance guarantees, the rates achieved by this decoder at practical block lengths are significantly less than the capacity. Subsequently, a soft-decision iterative decoder was proposed by Cho and Barron [3, 15], with theoretical guarantees similar to the adaptive successive decoder but improved empirical performance for finite block lengths.

In Chapter 4 we provide work towards analyzing the performance of adaptive successive decoding using a Bernoulli $\pm\frac{1}{\sqrt{n}}$ dictionary. A shift to the Bernoulli dictionary, not only increases computational efficiency but also decreases memory requirement for the storage of the dictionary.

Approximate Message Passing Decoder In the rest of this chapter, we propose an approximate message passing (AMP) decoder for SPARCs. We rigorously analyze its asymptotic performance and prove that for all fixed rates $R < \mathcal{C}$ the probability of decoding error goes to zero as the block length increases. The AMP decoder performs better at practical block lengths than either the adaptive successive of soft-decision iterative decoders. An AMP decoder for SPARCs was proposed by Barbier and Krzakala in [16] with different update rules from the decoder proposed in the following Chapter. Their performance analysis of the decoder suggested it was unable to achieve rates beyond some threshold rate strictly smaller than $\mathcal{C}$. Barbier et al [17] reported empirical results which show that the performance of the decoder in [16] can be improved by using spatially coupled Hadamard design matrices. The computational aspects of these two methods are compared in [18].

Both the adaptive successive [2] and iterative soft-threshold decoder [3,15] have probability of error decreasing like $\frac{n}{\log n}$ for any fixed rate $R < \mathcal{C}$, however the iterative soft-threshold

decoder has better empirical performance. In analyzing the performance of the AMP decoder we prove that the probability of error goes to zero for all $R < \mathcal{C}$ but we don't provide the rate at which this happens, meaning that while we can make qualitative comparisons of the performance of the two decoders, we are unable to make theoretical comparisons.

In both AMP and iterative soft-threshold decoding, successive estimates of the message are based on the values of 'test statistics' at each time $t = 0, 1, \ldots$ and the main difference between the two is in how the test statistics are generated. At step t, the iterative soft-thresholding decoder generates a test statistic based on an orthonormalization of the observed vector Y and previous 'fits' $X\beta^1, \ldots, X\beta^t$. In contrast, a modified version of the residual $(Y - X\beta^t)$ generates the test statistic for the AMP decoder. Despite these differences, test statistics for both decoders have a similar distributional structure: they are asymptotically equivalent to an observation of β corrupted by additive Gaussian noise with variance decreasing in t. AMP test statistics, however, are computed more quickly at each step making it feasible to implement the decoder for larger block lengths, which in turn results in lower (empirical) probability of decoding error.

2.4 Approximate Message Passing Introduction

In this section we introduce, generally, the approximate message passing algorithm and in Section 2.5 we specialize the algorithm for SPARCs decoding.

Consider the statistical high-dimensional regression problem, where the goal is to estimate a vector $\beta_0 \in \mathbb{R}^N$ from a noisy measurement $Y \in \mathbb{R}^n$ given by

$$Y = X\beta_0 + \epsilon. \tag{2.8}$$

Here X is a known $n \times N$ measurement matrix where it is possible that $n < N$, and $\epsilon \in \mathbb{R}^n$ is the measurement noise. The ratio $\frac{n}{N} \in (0, \infty)$ is denoted by δ.

Approximate message passing (AMP) [4–8] is a widely-studied class of low-complexity, scalable algorithms to solve (2.8), under suitable assumptions on β_0. Because the factor graph representing (2.8) is dense, the use of traditional message passing algorithms is in-

feasible since these methods use as messages complicated, real-valued functions. AMP, on the other hand, passes on scalar parameters summarizing the more complicated functions, thereby side-stepping this problem. For example, if the original functions are posterior distributions, the scalars might be the mean and variance. Using such approximations to the more complicated functions, the message passing updates become a set of simple rules for computing successive estimates of β_0.

AMP Update Rules Given the observed vector $Y = X\beta_0 + \epsilon$, the AMP decoder generates successive estimates of the unknown vector β_0, denoted by $\{\beta^t\}$, where $\beta^t \in \mathbb{R}^N$ for $t = 1, 2, \ldots$. Set the initial estimate $\beta^0 = 0$, the all-zeros vector. For $t = 0, 1, \ldots$, compute

$$z^t = Y - X\beta^t + \frac{z^{t-1}}{n} \sum_{i=1}^{N} \eta'_{t-1}([X^* z^{t-1}]_i + \beta_i^{t-1}), \tag{2.9}$$

$$\beta^{t+1} = \eta_t(\beta^t + X^* z^t), \tag{2.10}$$

using an appropriately-chosen sequence of functions $\{\eta_t\}_{t \geq 0} : \mathbb{R} \to \mathbb{R}$. In (2.9) and (2.10), X^* denotes the transpose of X, η_t acts component-wise when applied to a vector, and η'_t denotes its (weak) derivative. Quantities with a negative index are set to zero. The derivation of AMP updates (2.9) and (2.10) from a traditional message passing algorithm is demonstrated in full in [6] and [8, 19], among others, provide comprehensive lists of work related to AMP.

AMP Performance Guarantees For a Gaussian measurement matrix X with entries that are i.i.d. $\sim \mathcal{N}(0, 1/n)$, a constant undersampling ratio $\frac{n}{N}$, and message β_0 assumed to be i.i.d. according to some known prior, it was rigorously proven [6, 20] that the performance of AMP can be characterized in the large system limit via a simple scalar iteration called *state evolution*. In particular, the result implies that the ℓ_2-error $\frac{1}{N}\|\beta_0 - \beta^t\|^2$ and the ℓ_1-error $\frac{1}{N}\|\beta_0 - \beta^t\|_1$ converge almost surely to constants that can be computed using the prior distribution of β_0. (The large system limit is defined as $n, N \to \infty$ such that $\frac{n}{N}$ is constant.)

Dissertation Outline for AMP Results In the following chapter we give a finite-

sample version of the above result. We derive a concentration result that implies that the probability of Δ-deviation between $\frac{1}{N}\|\beta_0 - \beta^t\|^2$ and its limiting constant value falls exponentially in n. Empirical findings have previously shown accuracy of the state evolution equations for practical n, for example of the order of several hundreds [4], and the work presented in the next chapter provides theoretical support of such findings.

In the rest of the current chapter, we propose an AMP decoder for sparse regression codes, which is derived as an approximation of a min-sum-like message passing algorithm. The full details of the approximation can be found in [9]. In the following Section 2.5 we demonstrate how to adapt the AMP updates of (2.9) and (2.10) to the channel coding problem and then the main performance results are provided in Section 2.6. Theorem 1 shows that the probability of decoding error goes to zero as the block length tends to infinity, for all rates $R < \mathcal{C}$, and simulation results demonstrate good performance at finite block lengths. We also show that smart choices for the power allocation can significantly improve the empirical performance the decoder at rates not close to $\mathcal{C}$ and that Hadamard design matrices greatly reduce decoding complexity without impeding performance; again the full details are in [9].

2.5 Approximate Message Passing for SPARCs

Recall from Section 2.2 equation (2.4) that the received codeword is given as $Y = X\beta_0 + \epsilon$, where $\beta_0 \in \mathcal{B}_{M,L}(P_1, \ldots, P_L)$, the set of vectors of length ML having a single non-zero value equal to $\sqrt{nP_\ell}$ in each section $\ell \in [L]$. Here we refer to the true message vector as β_0 which should be understood as a *realization* of the random vector β, which is uniformly distributed over $\mathcal{B}_{M,L}(P_1, \ldots, P_L)$.

While this model is similar to the one which traditional analysis of AMP considers, given in (2.8), there are two main differences in the SPARCs model. The first is that the under sampling ratio $n/N \to 0$ as the block size increases while in the original analysis of AMP n/N is constant. Secondly, in the original analysis of AMP, the prior on β is i.i.d. across the elements, while in the SPARCs model, β is assumed to have a prior which is uniform over all $\beta \in \mathcal{B}_{M,L}$. So in this case, β is section-wise i.i.d. with dependence within

each section. For these reasons, the analysis of the AMP decoder does not follow directly from the results in [6, 20].

Notation: Indices i, j will denote specific entries of β, while the index ℓ will denote the entire section ℓ of β. Thus β_i, β_j are scalars, while β_ℓ is a length M vector. Set $N = ML$. Performance guarantees for the SPARC decoder are given in the large system limit as the dictionary size goes to ∞. We write $\lim x$ to denote the limit of the quantity x as SPARC parameters $n, L, M \to \infty$ simultaneously, according to the relationship established in (2.6): $M = L^b$ and $bL \log L = nR$.

The AMP Decoder The AMP decoder generates successive estimates of the message, denoted $\{\beta^t\}$, where $\beta^t \in \mathbb{R}^N$ for $t = 1, 2, \ldots$. Set $\beta^0 = 0$, the all-zeros vector, and for $t = 0, 1, \ldots$, compute

$$z^t = Y - X\beta^t + \frac{z^{t-1}}{\tau^2_{t-1}} \left(P - \frac{\|\beta^t\|^2}{n} \right), \tag{2.11}$$

$$\beta^{t+1}_i = \eta^t_i(\beta^t + X^* z^t), \quad \text{for } i = 1, \ldots, N = ML, \tag{2.12}$$

where quantities with negative indices are set equal to zero. The constants $\{\tau_t\}$, and the estimation functions $\eta^t_i(\cdot)$ are defined as follows for $t = 0, 1, \ldots$.

- Define

$$\tau^2_0 = \sigma^2 + P, \qquad \tau^2_{t+1} = \sigma^2 + P(1 - x_{t+1}), \quad t \geq 0, \tag{2.13}$$

  where

$$x_{t+1} = \sum_{\ell=1}^{L} \frac{P_\ell}{P} \mathbb{E} \left[\frac{\exp\left(\frac{\sqrt{nP_\ell}}{\tau_t} (U^\ell_1 + \frac{\sqrt{nP_\ell}}{\tau_t}) \right)}{\exp\left(\frac{\sqrt{nP_\ell}}{\tau_t} (U^\ell_1 + \frac{\sqrt{nP_\ell}}{\tau_t}) \right) + \sum_{j=2}^{M} \exp\left(\frac{\sqrt{nP_\ell}}{\tau_t} U^\ell_j \right)} \right]. \tag{2.14}$$

  In (2.14), $\{U^\ell_j\}$ are i.i.d. $\mathcal{N}(0, 1)$ random variables for $j \in [M],\ \ell \in [L]$.

- For $i \in [N]$, define

$$\eta^t_i(s) = \sqrt{nP_\ell} \frac{\exp\left(\frac{s_i \sqrt{nP_\ell}}{\tau^2_t} \right)}{\sum_{j \in \text{sec}_\ell} \exp\left(\frac{s_j \sqrt{nP_\ell}}{\tau^2_t} \right)}, \quad \text{if } i \in \text{sec}_\ell,\ 1 \leq \ell \leq L. \tag{2.15}$$

The notation $j \in \text{sec}_\ell$ is used as shorthand for "index j in section ℓ", i.e., $j \in \{(\ell-1)M + 1, \ldots, \ell M\}$. Notice that $\eta_i^t(s)$ depends on all the components of s in the *section* containing i. For brevity, the argument of η_i^t in (2.12) is written as $X^* z^t + \beta^t$, with the understanding that only the components in the section containing i play a role in computing η_i^t. The AMP decoder proposed above is derived via a first-order approximation of a min-sum-like message passing algorithm, the details of which can be found in [9].

State Evolution In agreement with terminology from the original AMP analysis in [4, 6], the recursive relationship (2.13), describing how τ_{t+1} is obtained from τ_t, is called *state evolution*. The state evolution constants can be iteratively computed using (2.13) and (2.14) offline, before decoding begins, via Monte Carlo simulation to calculate expectations in (2.14) for given values of M, L, n.

Closed form expressions for x^{t+1} and τ_t^2 as $n \to \infty$ are shown in Section 2.6, but for now, it suffices to note that for any fixed $R < \mathcal{C}$, terms τ_t strictly decreas with t for a finite number of steps which we call T_n, at which point we have $\tau_{T_{n+1}} \geq \tau_{T_n}$. Having computed $\tau_0, \tau_1, \ldots, \tau_{T_n}$ before decoding begins, the decoder iteratively computes estimates $\beta^1, \ldots, \beta^{T_n}$ using (2.11) and (2.12) and terminating at time T_n. For the final estimate β^{T_n}, in each section $\ell \in [L]$, set the maximum value to $\sqrt{nP_\ell}$ and other entries to 0 to obtain the decoded message $\hat{\beta}$.

Test Statistics For an intuitive understanding of the AMP update rules ((2.11) and (2.12)), first consider (2.12), which generates an updated estimate β^{t+1} based on the value of the test statistic:

$$s^t := \beta^t + A^* z^t.$$

The form of this update step is motivated by the following key property of the test statistic, which is ultimately the reason why AMP 'works': *s^t is asymptotically (as $n \to \infty$) distributed as $\beta + \bar{\tau}_t Z$, where $\bar{\tau}_t$ is the limit of τ_t, and Z is an i.i.d. $\mathcal{N}(0,1)$ random vector independent of the message vector β.* This property of the test statistic, which we prove rigorously in Section 2.6 is due to the presence of the "Onsager" correction term in residual update step (2.11):

$$\frac{z^{t-1}}{\tau_{t-1}^2}\left(P - \frac{\|\beta^t\|^2}{n}\right).$$

Intuition about role of the Onsager correction term in the standard AMP algorithm is provided in [6, Section I-C].

In light of the above property, we generate β^{t+1} from $s^t = s$ as Bayes optimal estimate of β conditional on the value of the test statistic:

$$\beta^{t+1}(s) = \mathbb{E}[\beta \,|\, \beta + \tau_t Z = s], \tag{2.16}$$

For $i \in \text{sec}_\ell$, $\ell \in [L]$, we have

$$\begin{aligned}
\beta_i^{t+1}(s) &= \mathbb{E}[\beta_i \mid \beta + \tau_t Z = s] = \mathbb{E}[\beta_i \mid \{\beta_j + \tau_t Z_j = s_j\}_{j \in \text{sec}_\ell}] \\
&= \sqrt{nP_\ell}\; P(\beta_i = \sqrt{nP_\ell} \mid \{\beta_j + \tau_t Z_j = s_j\}_{j \in \text{sec}_\ell}) \\
&= \sqrt{nP_\ell}\; \frac{f(\{\beta_j + \tau_t Z_j = s_j\}_{j \in \text{sec}_\ell} \mid \beta_i = \sqrt{nP_\ell})\, P(\beta_i = \sqrt{nP_\ell})}{\sum_{k \in \text{sec}_\ell} f(\{\beta_j + \tau_t Z_j = s_j\}_{j \in \text{sec}_\ell} \mid \beta_k = \sqrt{nP_\ell})\, P(\beta_k = \sqrt{nP_\ell})}
\end{aligned} \tag{2.17}$$

where we have used Bayes Theorem with f denoting the joint density function of $\{\beta_j + \tau_t Z_j\}_{j \in \text{sec}_\ell}$. Since β and Z are independent, with Z having i.i.d. $\mathcal{N}(0,1)$ entries, for each $k \in \text{sec}_\ell$ we have

$$\begin{aligned}
f(\{\beta_j + \tau_t Z_j = s_j\}_{j \in \text{sec}_\ell} \mid \beta_k = \sqrt{nP_\ell}) &\propto e^{-(s_k - \sqrt{nP_\ell})^2/2\tau_t^2} \prod_{j \in \text{sec}_\ell, j \neq k} e^{-s_j^2/2\tau_t^2} \\
&= e^{s_k\sqrt{nP_\ell}/\tau_t^2}\, e^{-nP_\ell/2\tau_t^2} \prod_{j \in \text{sec}_\ell} e^{-s_j^2/2\tau_t^2}.
\end{aligned} \tag{2.18}$$

Using (2.18) in (2.17), together with the fact that $P(\beta_k = \sqrt{nP_\ell}) = \frac{1}{M}$ for each $k \in \text{sec}_\ell$, we obtain

$$\beta_i^{t+1}(s) = \mathbb{E}[\beta_i \mid \beta + \tau_t Z = s] = \sqrt{nP_\ell}\, \frac{\exp\left(\frac{s_i\sqrt{nP_\ell}}{\tau_t^2}\right)}{\sum_{j \in \text{sec}_\ell} \exp\left(\frac{s_j\sqrt{nP_\ell}}{\tau_t^2}\right)}, \tag{2.19}$$

which is the expression in (2.15).

Thus, under the distributional assumption that s^t equals the true message plus independent Gaussian noise with variance determined by state evolution, β^{t+1} is the minimum expected squared error estimate of the message vector β (based on s^t). Also, for $i \in \text{sec}_\ell$, $\beta_i^{t+1}/\sqrt{nP_\ell}$ is the posterior probability of β_i being the non-zero entry in section ℓ, conditioned on the observation $s^t = \beta + \tau_t Z$.

2.5.1 Consequences of State Evolution

Many of the state evolution parameters of (2.13) and (2.14) have nice interpretations that aid understanding of the algorithm and we discuss these situations in what follows. These parameters are also key in determining when the algorithm should be terminated. We first discuss the role of the quantity x_{t+1} in equations (2.13) and (2.14).

Proposition 2.5.1. *[9, Proposition 3.1] Under the assumption that $s^t = \beta + \tau_t Z$, where Z is i.i.d. $\sim \mathcal{N}(0,1)$ and independent of β, the quantity x^{t+1} defined in (2.14) satisfies*

$$x_{t+1} = \frac{1}{nP}\mathbb{E}[\beta^* \beta^{t+1}], \quad 1 - x_{t+1} = \frac{1}{nP}\mathbb{E}[\|\beta - \beta^{t+1}\|^2], \tag{2.20}$$

and consequently, $\tau_{t+1}^2 = \sigma^2 + \frac{\mathbb{E}[\|\beta - \beta^{t+1}\|^2]}{n}$.

Proof. Proof in Appendix A.1. □

Proposition 2.5.1 tells us that x_{t+1} can be interpreted as the expectation of the (power-weighted) fraction of correctly decoded sections in step $t+1$, however this interpretation is accurate only in the limit when s^t is exactly distributed as $\beta + \bar{\tau}_t Z$, with $\bar{\tau}_t := \lim \tau_t$. In what follows we specify the limiting values of the state evolution parameters (2.13) and (2.14) under exponentially decaying power allocation and show how these values guide in when to terminate the algorithm. The performance of the AMP decoder, and the result of the following Lemma, will be analyzed with the following exponentially decaying power allocation:

$$P_\ell = P \cdot \frac{2^{2\mathcal{C}/L} - 1}{1 - 2^{-2\mathcal{C}}} \cdot 2^{-2\mathcal{C}\ell/L}, \quad \ell \in [L]. \tag{2.21}$$

Lemma 1. *[9, Lemma 2] For the power allocation $\{P_\ell\}$ given in (2.21), we have for $t = 0, 1, \ldots$:*

$$\bar{x}_t := \lim x_t = \frac{(1 + \mathit{snr}) - (1 + \mathit{snr})^{1 - \xi_{t-1}}}{\mathit{snr}} \tag{2.22}$$

$$\bar{\tau}_t^2 := \lim \tau_t^2 = \sigma^2 + P(1 - \bar{x}_t) = \sigma^2 \, (1 + \mathit{snr})^{1 - \xi_{t-1}} \tag{2.23}$$

where $\xi_{-1} = 0$*, and for* $t \geq 0$*,*

$$\xi_t = \min\left\{\left(\frac{1}{2\mathcal{C}}\log\left(\frac{\mathcal{C}}{R}\right) + \xi_{t-1}\right),\ 1\right\}. \tag{2.24}$$

Proof. Proof in Appendix A.2. □

Considering results (2.22) and (2.24) it is clear that $\bar{x}_t$ strictly increases with t until it reaches one, which occurs in a finite number of steps we label T^* where $T^* = \left\lceil \frac{2\mathcal{C}}{\log(\mathcal{C}/R)} \right\rceil$ and $\bar{x}_{T^*} = 1$. Similarly, $\bar{\tau}_t^2$, the variance of the "noise" in the large system distribution of the AMP test statistic, decreases monotonically from $\bar{\tau}_0^2 = \sigma^2 + P$ down to $\bar{\tau}_{T^*}^2 = \sigma^2$. In other words, the initial observation $Y = X\beta + \epsilon$ is effectively transformed by the AMP decoder into a 'denoised' statistic $s^{T^*} = \beta + \epsilon'$, where ϵ' is Gaussian with the *same* variance as the measurement noise ϵ. AMP has effectively converted the Gaussian design X into the identity matrix.

Moreover, in the limit, the constants $\{\xi_t\}_{t\geq 0}$ can be interpreted as follows: at the end of step $t+1$, the first ξ_t fraction of sections in β^{t+1} will be correctly decodable with high probability, i.e. the correct location of the non-zero entry in these sections will have almost all the posterior probability mass. The other $(1-\xi_t)$ fraction of sections will not be correctly decodable from β^{t+1} as the power allocated to these sections is not large enough. In each step until T^*, an additional $\frac{1}{2\mathcal{C}}\log\left(\frac{\mathcal{C}}{R}\right)$ fraction of sections become correctly decodable, and at step T^* all the sections are correctly decodable with high probability.

As noted earlier, the termination step T_n is the smallest t for which $\tau_t^2 \leq \tau_{t+1}^2$. Now Lemma 1 shows that in the large system limit, the number of steps until the AMP decoder terminates is $\lim T_n = T^*$. Since T^n and T^* are both integers, $\lim T^n = T^*$ implies that for sufficiently large n we will have $T^n = T^*$, and so we allow T^* to determine the termination point of the algorithm. Recalling $T^* = \left\lceil \frac{2\mathcal{C}}{\log(\mathcal{C}/R)} \right\rceil$, we see that as the rate approaches capacity, the algorithm requires more steps to terminate.

In summary, from Lemma 1, we see that the algorithm terminates in a finite number of steps, namely T^*. Then using Proposition 2.5.1, at termination step T^*, the large system

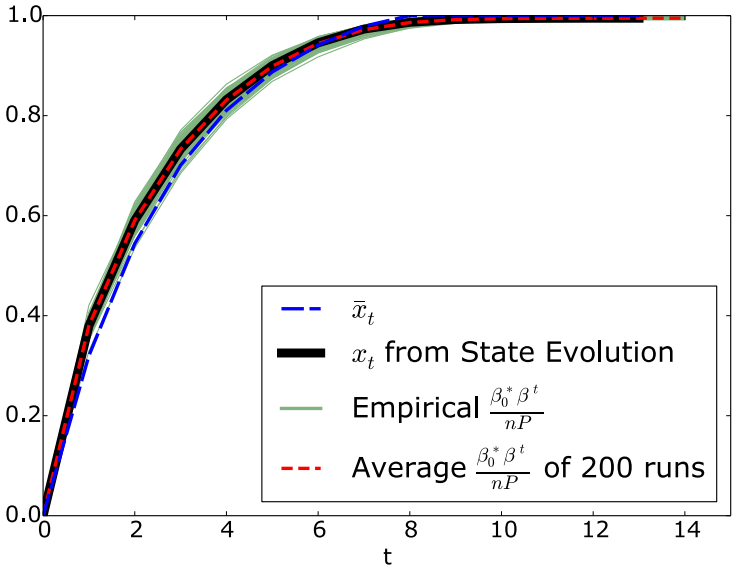

Figure 2.3: Comparison of state evolution and AMP. The SPARC parameters are $M = 512, L = 1024, \mathsf{snr} = 15, R = 0.7\mathcal{C}$, $P_\ell \propto 2^{-2\mathcal{C}\ell/L}$. The average of the 200 trials (green curves) is the dashed red curve, which is almost indistinguishable from the state evolution prediction (black curve).

limit $\lim \frac{1}{n}\mathbb{E}\|\beta - \beta^{T^*}\|^2$ equals zero.

Unfortunately, though, for finite-sized dictionaries, the test statistic s^t is not exactly distributed as $\beta + \tau_t Z$, and so the interpretations of the state evolution parameters given above will not hold exactly. Nevertheless, computing x_{t+1} numerically via the state evolution equations (2.13) and (2.14) yields an estimate for the expected weighted fraction of correctly decoded sections after each step, and simulations in Section 2.6 indicate that the behavior of the AMP is close to that predicted by state evolution for moderately large values of n, M, L. For example, Figure 2.3[1] shows the trajectory of x_t vs t for a SPARC with the parameters specified in the figure. The empirical average of $(\beta_0^* \beta^t)/nP$ matches almost exactly with x_t, as does the theoretical limit $\bar{x}_t$ given in (2.22).

Statistical Behavior of AMP The distributional behavior of the AMP decoder can be summarized as follows. The test statistic $s^t = \beta^t + A^* z^t$ that is used for the β-update in (2.10) is asymptotically distributed as $\beta + \bar{\tau}_t Z$, where Z has i.i.d. standard Gaussian entries and is independent of the message vector β. For any $R < \mathcal{C}$, the variance of the "noise" in the test statistic, $\bar{\tau}_t^2$, decreases monotonically from $\sigma^2 + P$ to σ^2 in a finite number of steps

1. Many thanks to Adam Grieg for this figure and for empirical study of the performance of the AMP decoder at finite block lengths.

we label T^*. In other words, the initial observation $Y = X\beta + \epsilon$ is effectively transformed by the AMP decoder into a cleaner statistic $s^{T^*} = \beta + \epsilon'$, where ϵ' is Gaussian with the same variance as the measurement noise ϵ.

2.6 Performance of AMP

To analyze the performance of AMP as a decoder for SPARCs, we use the framework of Bayati and Montanari [6], who in turn built on techniques introduced by Bolthausen [21]. However, the analysis of the proposed algorithm does not follow directly from the results in [6, 22]. The main reason is that the under sampling ratio $n/N \to 0$ as the block size increases while in the original analysis of AMP n/N is constant. Secondly, in the original analysis of AMP, the prior on β is i.i.d. across the elements, while in the SPARCs model, β is assumed to have a prior which is uniform over all $\beta \in \mathcal{B}_{M,L}$. So in this case, β is section-wise i.i.d. with dependence within each section. For these reasons, the analysis of the AMP decoder does not follow directly from the results in [6, 20].

Our main result is proved for the following slightly modified AMP decoder, which runs for exactly T^* steps. Set $\beta^0 = 0$ and compute

$$z^t = y - A\beta^t + \frac{z^{t-1}}{\bar{\tau}^2_{t-1}} \left(P - \frac{\|\beta^t\|^2}{n} \right), \tag{2.25}$$

$$\beta^{t+1}_i = \eta^t_i(\beta^t + A^* z^t), \quad \text{for } i \in [N] \tag{2.26}$$

where for $i \in \text{sec}_\ell,\ \ell \in [L]$,

$$\eta^t_i(s) = \sqrt{nP_\ell}\, \frac{\exp\left(s_i \sqrt{nP_\ell}/\bar{\tau}^2_t\right)}{\sum_{j \in \text{sec}_\ell} \exp\left(s_j \sqrt{nP_\ell}/\bar{\tau}^2_t\right)}. \tag{2.27}$$

The only difference from the earlier decoder described in (2.11)–(2.15) is that we replace τ^2_t with its limiting value $\bar{\tau}^2_t$ defined in Lemma 1.

The algorithm terminates after generating β^{T^*} and the decoded codeword $\hat{\beta} \in \mathcal{B}_{M,L}(P_1, \ldots, P_L)$ is obtained by setting the maximum of β^{T^*} in each section $\ell \in [L]$ to $\sqrt{nP_\ell}$ and the remaining entries to 0.

The *section error rate* of a decoder for a SPARC $\mathcal{S}$ is defined as

$$\mathcal{E}_{sec}(\mathcal{S}) := \frac{1}{L}\sum_{\ell=1}^{L} \mathbf{1}\{\hat{\beta}_\ell \neq \beta_{0_\ell}\}. \tag{2.28}$$

Theorem 1. *[9, Theorem 1] Fix any rate $R < \mathcal{C}$, and $b > 0$. Consider a sequence of rate R SPARCs $\{\mathcal{S}_n\}$ indexed by block length n, with design matrix parameters L and $M = L^b$ determined according to* (2.6)*, and an exponentially decaying power allocation given by* (2.21)*. Then the section error rate of the AMP decoder (described in* (2.25)*–*(2.27)*, and run for T^* steps) converges to zero almost surely, i.e., for any $\epsilon > 0$,*

$$\lim_{n_0\to\infty} P\left(\mathcal{E}_{sec}(\mathcal{S}_n) < \epsilon,\ \forall n \geq n_0\right) = 1. \tag{2.29}$$

Proof. The proof of Theorem 1 is given in Section 2.7. □

Remarks:

1. The probability measure in (2.29) is over the Gaussian design matrix X, the Gaussian channel noise ϵ, and the message β distributed uniformly in $\mathcal{B}_{M,L}(P_1, \ldots, P_L)$.

2. As in [2], we can construct a concatenated code with an inner SPARC of rate R and an outer Reed-Solomon (RS) code of rate $(1 - 2\epsilon)$. If M is a prime power, a RS code defined over a finite field of order M defines a one-to-one mapping between a symbol of the RS codeword and a section of the SPARC. The concatenated code has rate $R(1 - 2\epsilon)$, and decoding complexity that is polynomial in n. The decoded message $\hat{\beta}$ equals β whenever the section error rate of the SPARC is less than ϵ. Thus for any $\epsilon > 0$, the theorem guarantees that the probability of *message* decoding error for a sequence of rate $R(1 - 2\epsilon)$ SPARC-RS concatenated codes will tend to zero, i.e.,

$$\lim P(\hat{\beta} \neq \beta) = 0.$$

2.6.1 Empirical Performance at Finite Blocklengths

We suggest two modifications to the algorithm used in Theorem 1 that have been demonstrated empirically to increase computational efficiency. First, a 'modified' power allocation yields several orders of magnitude improvement in section error rate for rates R that are not very close to the capacity $\mathcal{C}$. Second, a Hadamard design matrix (instead of Gaussian), facilitates a decoder with $O(N \log N)$ running time and a memory requirement of $O(N)$. In comparison, a Gaussian design matrix has $O(nN)$ running time and memory of the AMP decoder. Other work [17] independently considered an AMP decoder with a spatially coupled Hadamard-based design matrix.

The power allocation in (2.21) is effective at rates just less than $\mathcal{C}$ but can be improved for lower rates, where it otherwise over-allocates power to initial sections such that not enough power is left for decoding at the end. When considering the power allocation, there are two conflicting objectives. One needs enough power in the beginning sections making it such that these sections are more likely to decode correctly, which in turn decreases the effective noise variance $\bar{\tau}_t^2$ in subsequent AMP iterations. On the other hand, we must ensure that the final sections have enough power to be decoded correctly. We suggest using a modified power allocation, which uses a steeper exponential decay in the beginning but flattening at the end: a combination of both flat and exponentially decaying power allocations. For a more in depth discussion of the modified power allocation, we refer the reader to [9].

The computational complexity of the decoder in (2.25)–(2.27) is determined by the matrix-vector multiplications $X\beta^t$ and $X^* z^t$, whose running time is $O(nN)$ if performed in the straightforward way. The remaining operations are $O(N)$. As the number of iterations is finite, the decoding complexity scales linearly with the size of the design matrix. With a Gaussian design matrix, the memory requirement is also proportional to nN as the entire matrix has to be stored. This is the major bottleneck in scaling the AMP decoder to work with large design matrices.

To reduce the decoding complexity and the required memory, we generate X from a Hadamard matrix, by randomly selecting n rows of an $N \times N$ Hadamard matrix. More details are given in [9]. For X generated in this manner, the matrix-vector multiplications

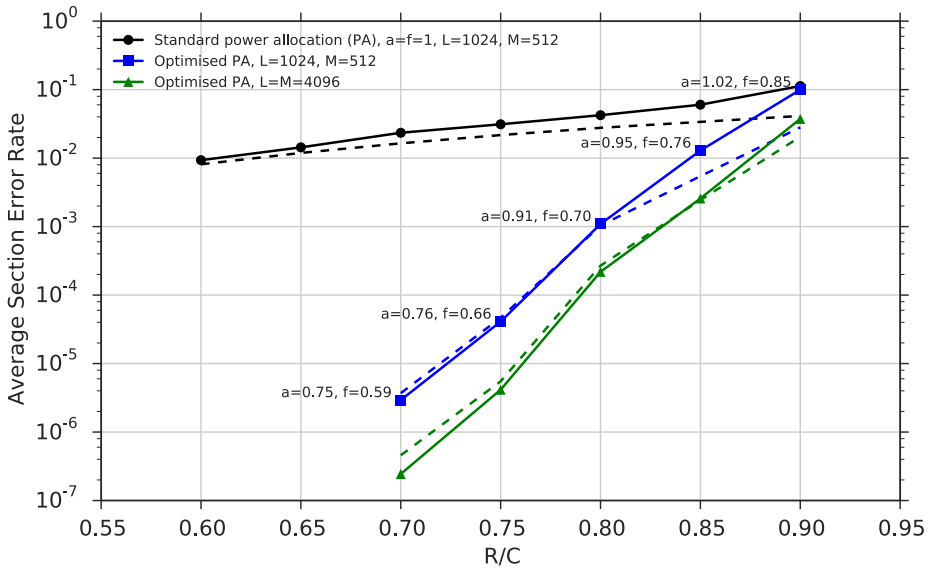

Figure 2.4: Section error rate vs $R/\mathcal{C}$ at $\mathsf{snr} = 15, \mathcal{C} = 2$ bits. The top solid black curve shows the average section error rate of the AMP over 1000 trials with exponentially decaying power allocation. The solid blue curve in the middle shows the section error rate using a modified power allocation. The SPARC parameters for both these curves are $M = 512, L = 1024$. The bottom solid green curve shows the section error rate with a modified power allocation, but $L = M = 4096$. In all cases, the dashed lines show the section error rate predicted by state evolution. Missing points at $R = 0.6\mathcal{C}$ and $0.65\mathcal{C}$ indicate no errors observed over 1000 trials.

$X\beta^t$ and $X^* z^t$ can be performed efficiently using the *fast Walsh-Hadamard Transform* (WHT) [23], which has $O(N \log N)$ running time. Further, we do not need to store X; only the vectors β^t and z^t need to be kept in memory. Hence the running time and memory requirement of the decoder are now $O(N \log N)$ and $O(N)$, respectively. These substantial improvements allow the use of much larger dictionaries (e.g., $M = L = 4096$) for which AMP decoding with Gaussian matrices is infeasible with standard computing resources. For given values of n, M, L and power allocation $\{P_\ell\}$, we found the empirical performance with a Hadamard dictionary to be very similar to the Gaussian case.

Experimental Results: Figure 2.4[2] shows the performance of the AMP at different rates nearing the capacity. Given the values of M, L, the block length n is determined by the rate R according to (2.5). For example, with $M = 512, L = 1024$, we have $n = 7680$ for $R = 0.6\mathcal{C}$, and $n = 5120$ for $R = 0.9\mathcal{C}$.

2. Many thanks to Adam Grieg for this figure and for empirical study of the performance of the AMP decoder at finite block lengths.

The solid black curve at the top of the shows the average section error rate of the AMP (over 1000 runs) with an exponentially decaying power allocation where $P_\ell \propto 2^{-2\mathcal{C}\ell/L}$. However, the solid blue curve in the middle shows the average section error rate when the modified power allocation discussed above is employed. Clearly a smart choice of power allocation can greatly improve empirical performance at rates far from capacity. The green solid curve at the bottom shows the average section error rate when using a larger dictionary with $L = M = 4096$, and the modified power allocation.

In all cases, the decoder described in (2.25)–(2.27) was used. The constants $\{\bar{\tau}_t^2\}$ required by the decoder are specified by Lemma 1 for the exponential allocation, and their corresponding versions given explicitly in [9] for the modified allocation. The simulations for Fig. 2.4 were run using Hadamard design matrices.

Across trials, we observed good concentration around the average section error rates. For example, with $M = 512, L = 1024$ and $R = 0.75\mathcal{C}$, 958 of the 1000 trials had zero errors, and the remaining 42 had only one section in error, for an average section error rate of 4.10×10^{-5}. Further, all the section errors were in the flat part of the power allocation, as expected. Increasing L tends to improve this concentration, while increasing M reduces the average section error rate. This improvement in the section error rate is illustrated by the bottom curve in Fig. 2.4. The dashed curves in Fig. 2.4 show the section error rate predictions for the two power allocations obtained from state evolution.

2.7 Technical Lemma

The proof of Theorem 1 relies on the following technical lemma. Presented below, Lemma 2 shows that the state evolution equations (2.22) and (2.23) accurately predict the performance of the AMP decoder, at least in the large system limit. One consequence of Lemma 2 is that the ℓ_2-error $\frac{1}{n}\|\beta^t - \beta\|^2$ converges almost surely to $P(1 - \bar{x}_t)$, for $0 \leq t \leq T^*$.

For consistency and ease of comparison, we use notation similar to [6]. Define the

following column vectors recursively for $t \geq 0$, starting with $\beta^0 = 0$ and $z^0 = y$.

$$\begin{aligned} h^{t+1} &= \beta_0 - (X^* z^t + \beta^t), \qquad q^t = \beta^t - \beta_0, \\ b^t &= \epsilon - z^t, \qquad m^t = -z^t. \end{aligned} \tag{2.30}$$

Recall that β_0 is the true message vector. Due to the symmetry of the code construction, we can assume that the non-zeros of β_0 are in the first entry of each section. The vector h^{t+1} is the noise in the test statistic $X^* z^t + \beta^t$ and q^t is the error in the current estimate.

Define $\mathscr{S}_{t_1,t_2}$ to be the sigma-algebra generated by

$$b^0, ..., b^{t_1-1}, m^0, ..., m^{t_1-1}, h^1, ..., h^{t_2}, q^0, ..., q^{t_2}, \text{ and } \beta_0, w.$$

Lemma 2 recursively computes the conditional distributions $b^t|_{\mathscr{S}_{t,t}}$ and $h^{t+1}|_{\mathscr{S}_{t+1,t}}$, as well as the limiting values of various inner products involving h^{t+1}, q^t, b^t, and m^t. A key ingredient in proving the lemma is the conditional distribution of the design matrix X given $\mathscr{S}_{t_1,t_2}$. For $t \geq 1$, let

$$\lambda_t = \frac{-1}{\tilde{\tau}^2_{t-1}} \left(P - \frac{\|\beta^t\|^2}{n} \right). \tag{2.31}$$

Define matrices

$$M_t = [m^0 \mid \ldots \mid m^{t-1}], \qquad Q_t = [q^0 \mid \ldots \mid q^{t-1}]. \tag{2.32}$$

The notation $[c_1 \mid c_2 \mid \ldots \mid c_k]$ is used to denote a matrix with columns $c_1, \ldots, c_k$. Note that M_0 and Q_0 are the all-zero vector. We use the notation $m^t_\parallel$ and $q^t_\parallel$ to denote the projection of m^t and q^t onto the column space of M_t and Q_t, respectively. Let $\vec{\alpha}_t = (\alpha_0, \ldots, \alpha_{t-1})$ and $\vec{\gamma}_t = (\gamma_0, \ldots, \gamma_{t-1})$ be the coefficient vectors of these projections, i.e.,

$$m^t_\parallel = \sum_{i=0}^{t-1} \alpha_i m^i, \quad q^t_\parallel = \sum_{i=0}^{t-1} \gamma_i q^i. \tag{2.33}$$

The projections of m^t and q^t onto the orthogonal complements of M^t and Q^t, respectively, are denoted by

$$m^t_\perp = m^t - m^t_\parallel, \quad q^t_\perp = q^t - q^t_\parallel \tag{2.34}$$

Given two random vectors A, B and a sigma-algebra $\mathscr{S}$, $A|_{\mathscr{S}} \stackrel{d}{=} B$ implies that the conditional distribution of A given $\mathscr{S}$ equals the distribution of B. For random variables A, B, the notation $A \stackrel{a.s.}{=} B$ means that A and B are equal almost surely. We use the notation $\vec{o}_t(n^{-\delta})$ to denote a vector in $\mathbb{R}^t$ such that each of its coordinates is $o(n^{-\delta})$ (here t is fixed). The $t \times t$ identity matrix is denoted by $\mathsf{I}_{t\times t}$, and the $t \times s$ all-zero matrix is denoted by $\mathsf{0}_{t\times s}$.

The notation 'lim' is used to denote the large system limit as $n, M, L \to \infty$; recall that the three quantities are related as $L \log M = nR$, with $M = L^b$. We keep in mind that (given R and b) the block length n uniquely determines the dimensions of all the quantities in the system including $X, \beta_0, \epsilon, h^{t+1}, q^t, b^t, m^t$. Thus we have a sequence indexed by n of each of these random quantities, associated with the sequence of SPARCs $\{\mathcal{S}_n\}$.

Finally, we recall the definition of *pseudo-Lipschitz* functions from [6].

Definition 2.7.1. *A function $\phi : \mathbb{R}^m \to \mathbb{R}$ is pseudo-Lipschitz of order k (denoted by $\phi \in PL(k)$) if there exists a constant $C > 0$ such that for all $x, y \in \mathbb{R}^m$,*

$$|\phi(x) - \phi(y)| \leq C(1 + \|x\|^{k-1} + \|y\|^{k-1})\|x - y\|. \tag{2.35}$$

We will use the fact that when $\phi \in PL(k)$, there is a constant C' such that $\forall x \in \mathbb{R}^m$,

$$|\phi(x)| \leq C'(1 + \|x\|^k). \tag{2.36}$$

2.7.1 Asymptotics Lemma

In the lemma below, $\delta \in (0, \frac{1}{2})$ is a generic positive number whose exact value is not required. The value of δ in each statement of the lemma may be different. We will say that a sequence x_n converges to a constant c at rate $n^{-\delta}$ if $\lim_{n\to\infty} n^{\delta}(x_n - c) = 0$.

Lemma 2. *The following statements hold for $0 \leq t \leq T^*$, where $T^* = \left\lceil \frac{2\mathcal{C}}{\log(\mathcal{C}/R)} \right\rceil$.*

(a)

$$h^{t+1}|_{\mathscr{S}_{t+1,t}} \stackrel{d}{=} \sum_{i=0}^{t-1} \alpha_i h^{i+1} + \tilde{X}^* m_\perp^t + \tilde{Q}_{t+1}\vec{o}_{t+1}(n^{-\delta}), \tag{2.37}$$

$$b^t|_{\mathscr{S}_{t,t}} \stackrel{d}{=} \sum_{i=0}^{t-1} \gamma_i b^i + \tilde{X} q_\perp^t + \tilde{M}_t\vec{o}_t(n^{-\delta}) \tag{2.38}$$

where $\tilde{X}$ is an independent copy of X and the columns of the matrices $\tilde{Q}_t$ and $\tilde{M}_t$ form an orthogonal basis for the column space of Q_t and M_t, respectively, such that

$$\tilde{Q}_t^*\tilde{Q}_t = \tilde{M}_t^*\tilde{M}_t = n\mathrm{I}_{t\times t}. \tag{2.39}$$

(b) i) Consider the following functions ϕ_h defined on $\mathbb{R}^M \times \mathbb{R}^M \times \mathbb{R}^M \to \mathbb{R}$:

$$\phi_h(h_\ell, \tilde{h}_\ell, \beta_\ell) = \begin{cases} h_\ell^*\tilde{h}_\ell/M, & \\ \|\eta^r(\beta_\ell - h_\ell)\|^2/\log M, & 0 \le r \le t, \\ [\eta^r(\beta_\ell - h_\ell) - \beta_\ell]^*[\eta^s(\beta_\ell - \tilde{h}_\ell) - \beta_\ell]/\log M, & 0 \le r \le s \le t, \\ h_\ell^*[\,\eta^r(\beta_\ell - h_\ell) - \beta_\ell]/\log M, & 0 \le r \le t, \end{cases} \tag{2.40}$$

For each function in (2.40) and arbitrary constants $(a_0, \ldots, a_t, b_0, \ldots, b_t)$, we have:

$$\lim n^\delta \left[\frac{1}{L}\sum_{\ell=1}^{L} \phi_h\left(\sum_{r=0}^{t} a_r h_\ell^{r+1}, \sum_{s=0}^{t} b_s h_\ell^{s+1}, \beta_{0_\ell}\right) - \lim \frac{1}{L}\sum_{\ell=1}^{L} \mathbb{E}\left\{\phi_h\left(\sum_{r=0}^{t} a_r \bar{\tau}_r Z_{r_\ell}, \sum_{s=0}^{t} b_s \bar{\tau}_s Z_{s_\ell}, \beta_\ell\right)\right\}\right] \stackrel{a.s.}{=} 0, \tag{2.41}$$

where $\bar{\tau}_r$ is defined in Lemma 1 and $Z_0, ..., Z_t$ are length-N Gaussian random vectors independent of β, with Z_{r_ℓ} denoting the ℓth section of Z_r. For $0 \le s \le t$, $\{Z_{s,j}\}_{j\in[N]}$ are i.i.d. $\sim \mathcal{N}(0,1)$, and for each $i \in [N]$, $(Z_{0,i}, \ldots, Z_{t,i})$ are jointly Gaussian. The inner limit in (2.41) exists and is finite for each ϕ_h in (2.40).

ii) For all pseudo-Lipschitz functions $\phi_b : \mathbb{R}^{t+2} \to \mathbb{R}$ of order two, we have

$$\lim n^\delta \left[\frac{1}{n}\sum_{i=1}^{n} \phi_b(b_i^0, ..., b_i^t, \epsilon_i) - \mathbb{E}\{\phi_b(\bar{\sigma_0}\hat{Z}_0, ..., \bar{\sigma}_t\hat{Z}_t, \sigma Z_\epsilon)\}\right] \stackrel{a.s}{=} 0. \tag{2.42}$$

where for $s \geq 0$,

$$\bar{\sigma}_s^2 := \bar{\tau}_s^2 - \sigma^2 = P(1 - \bar{x}_s), \tag{2.43}$$

with $\bar{x}_s$ *defined in Lemma 1. The random variables* $(\hat{Z}_0, ..., \hat{Z}_t)$ *are jointly Gaussian with* $\hat{Z}_s \sim \mathcal{N}(0,1)$ *for* $0 \leq s \leq t$. *Further,* $(\hat{Z}_0, ..., \hat{Z}_t)$ *are independent of* $Z_\epsilon \sim \mathcal{N}(0,1)$.

(c) For all $0 \leq r \leq s \leq t$,

$$\lim \frac{(h^{r+1})^* h^{s+1}}{N} \overset{a.s}{=} \lim \frac{(m^r)^* m^s}{n} \overset{a.s.}{=} \mathbb{E}[(\bar{\sigma}_r \hat{Z}_r - \sigma Z_\epsilon)(\bar{\sigma}_s \hat{Z}_s - \sigma Z_\epsilon)], \tag{2.44}$$

$$\lim \frac{(b^r)^* b^s}{n} \overset{a.s}{=} \lim \frac{(q^r)^* q^s}{n} \overset{a.s.}{=} \bar{\sigma}_s^2, \tag{2.45}$$

where the random variables $\hat{Z}_r, \hat{Z}_s, Z_\epsilon$ *are those in* (B.12), *and* $\bar{\sigma}_s$ *is defined in* (2.43). *The convergence rate in both* (B.13) *and* (B.14) *is* $n^{-\delta}$.

(d) For all $0 \leq r \leq s \leq t$,

$$\lim \frac{(h^{r+1})^* q^{s+1}}{n} \overset{a.s}{=} \lim \lambda_{s+1} \lim \frac{(m^r)^* m^s}{n} \overset{a.s.}{=} \frac{-\bar{\sigma}_{s+1}^2}{\sigma^2 + \bar{\sigma}_s^2} \mathbb{E}[(\bar{\sigma}_r \hat{Z}_r - \sigma Z_w)(\bar{\sigma}_s \hat{Z}_s - \sigma Z_w)], \tag{2.46}$$

$$\lim \frac{(b^r)^* m^s}{n} \overset{a.s}{=} \lim \frac{(b^r)^* b^s}{n} \overset{a.s.}{=} \bar{\sigma}_s^2. \tag{2.47}$$

The convergence rate in both (2.46) *and* (2.47) *is* $n^{-\delta}$.

(e)

$$\lim \frac{(h^{t+1})^* q^0}{n} \overset{a.s.}{=} 0. \tag{2.48}$$

(f) The following hold almost surely.

$$\lim \frac{\|q_\perp^0\|^2}{n} = \bar{\sigma}_0^2 = P, \quad \lim \frac{\|q_\perp^r\|^2}{n} = \bar{\sigma}_r^2 \left(1 - \frac{\bar{\sigma}_r^2}{\bar{\sigma}_{r-1}^2}\right) \quad \textit{for} \;\; 1 \leq r \leq t, \tag{2.49}$$

$$\lim \frac{\|m_\perp^0\|^2}{n} = \bar{\tau}_0^2 = \sigma^2 + P, \quad \lim \frac{\|m_\perp^s\|^2}{n} = \bar{\tau}_s^2 - u^* C^{-1} u, \quad \textit{for} \;\; 1 \leq s \leq t-1, \tag{2.50}$$

where for $1 \leq i, j \leq s$,

$$u_i = \mathbb{E}\left[(\bar{\sigma}_s \hat{Z}_s - \sigma Z_\epsilon)(\bar{\sigma}_{i-1}\hat{Z}_{i-1} - \sigma Z_\epsilon)\right],\ C_{ij} = \mathbb{E}\left[(\bar{\sigma}_{i-1}\hat{Z}_{i-1} - \sigma Z_\epsilon)(\bar{\sigma}_{j-1}\hat{Z}_{j-1} - \sigma Z_\epsilon)\right].$$

The limits in (B.21) *and* (B.22) *are strictly positive for* $r, s < T^*$.

The full proof of the above lemma can be found in [9] and is included here in Appendix A.3. The main difference between Lemma 2 and [6, Lemma 1] is part (b).i, which is a key ingredient in proving Theorem 1. The functions involving η we study in (2.40) all act *section-wise* when applied to vectors in $\mathbb{R}^N$, in contrast to the *component-wise* functions considered in [6] (and in part (b).ii above). This is due to the fact that the prior on β we consider is section-wise i.i.d. instead of entry-wise i.i.d.Ṫo prove (2.41) for the section-wise functions as the section size $M \to \infty$, we need that the limits in the other parts of the lemma (particularly in (B.9) and (B.10)) have convergence rates of $n^{-\delta}$ for some $\delta > 0$. Minimum rates of convergence were not needed for [6, Lemma 1].

2.7.2 Proof of Theorem 1

From the definition in (2.28), the event that the section error rate is larger than Δ can be written as

$$\{\mathcal{E}_{sec}(\mathcal{S}_n) > \Delta\} = \left\{\sum_{\ell=1}^{L} \mathbf{1}\{\hat{\beta}_\ell \neq \beta_{0_\ell}\} > L\Delta\right\}. \tag{2.51}$$

When a section ℓ is decoded in error, the correct non-zero entry has no more than half the total mass of section ℓ at the termination step T^*. That is,

$$\beta^{T^*}_{\mathsf{sent}(\ell)} \leq \frac{1}{2}\sqrt{nP_\ell} \tag{2.52}$$

where $\mathsf{sent}(\ell)$ is the index of the non-zero entry in section ℓ of the true message β_0. Since $\beta_{0_{\mathsf{sent}(\ell)}} = \sqrt{nP_\ell}$, we have

$$\mathbf{1}\{\hat{\beta}_\ell \neq \beta_{0_\ell}\} \quad \Rightarrow \quad \|\beta^{T^*}_\ell - \beta_{0_\ell}\|^2 \geq \frac{nP_\ell}{4}, \quad \ell \in [L]. \tag{2.53}$$

Hence when (2.51) holds, we have

$$\|\beta^{T^*} - \beta_0\|^2 = \sum_{\ell=1}^{L} \|\beta_\ell^{T^*} - \beta_{0_\ell}\|^2 \overset{(a)}{\geq} \sum_{\ell=1}^{L} \mathbf{1}\{\hat{\beta}_\ell \neq \beta_{0_\ell}\} \frac{nP_\ell}{4} \overset{(b)}{\geq} L\Delta \frac{nP_L}{4} \overset{(c)}{\geq} \frac{n\,\Delta\,\sigma^2 \ln(1+\mathsf{snr})}{4}, \tag{2.54}$$

where (a) follows from (2.53); (b) is obtained using (2.51), and the fact that $P_\ell > P_L$ for $\ell \in [L-1]$ for the exponentially decaying power allocation in (2.21); (c) is obtained using the first-order Taylor series lower bound $LP_L \geq \sigma^2 \ln(1 + \frac{P}{\sigma^2})$. We therefore conclude that

$$\{\mathcal{E}_{sec}(\mathcal{S}_n) > \Delta\} \;\Rightarrow\; \left\{ \frac{\|\beta^{T^*} - \beta_0\|^2}{n} \geq \frac{\Delta\,\sigma^2 \ln(1+\mathsf{snr})}{4} \right\}. \tag{2.55}$$

Now, from (B.14) of Lemma 2(c), we know that

$$\lim \frac{\|\beta^{T^*} - \beta_0\|^2}{n} = \lim \frac{\|q^{T^*}\|^2}{n} \overset{a.s.}{=} P(1 - \bar{x}_{T^*}) \overset{(a)}{=} 0, \tag{2.56}$$

where (a) follows from Lemma 1, which implies that $\xi_{T^*-1} = 1$ for $T^* = \left\lceil \frac{2\mathcal{C}}{\log(\mathcal{C}/R)} \right\rceil$, and hence $\bar{x}_{T^*} = 1$. Thus we have shown in (2.56) that $\frac{\|\beta^{T^*} - \beta_0\|^2}{n}$ converges almost surely to zero, i.e.,

$$\lim_{n_0 \to \infty} P\left(\frac{\|\beta^{T^*} - \beta_0\|^2}{n} < \Delta,\ \forall n \geq n_0 \right) = 1 \tag{2.57}$$

for any $e > 0$. From (2.55), this implies that for $\Delta' = \frac{4\Delta}{\sigma^2 \ln(1+\mathsf{snr})}$,

$$\lim_{n_0 \to \infty} P\left(\mathcal{E}_{sec}(\mathcal{S}_n) \leq \Delta',\ \forall n \geq n_0 \right) = 1. \tag{2.58}$$

Chapter 3

Finite-sample Analysis of Approximate Message Passing

Approximate Message Passing was introduced in Section 2.4 and the rest of Chapter 2 rigorously analyzes its performance as a decoder for sparse regression codes over the additive white Gaussian noise channel. In this Chapter we analyze finite-sample performance of the AMP algorithm, showing that for n of practical sizes, the simple scalar iteration called state evolution still accurately predicts the performance of the algorithm. Specifically we show that probability of deviation between the actual performance and the state evolution prediction falls exponentially in n, the sample size of the problem. In Section 3.1 we remind the reader of the framework for the AMP decoder, which was previously described in Section 2.4. Note that in this chapter we general formulation of AMP, not the specific usage of AMP as a decoder for SPARCs. In Section 3.2 we provide our main result, Theorem 2, analyzing the performance of the algorithm. Finally in Section 3.3 we prove Theorem 2 using a technical lemma which tracks the step-by-step distributional properties of the algorithm.

3.1 Approximate Message Passing Background

Recall model (2.8) from Section 2.4 that is considered by AMP: the goal is to estimate a vector $\beta_0 \in \mathbb{R}^N$ from noisy measurements $Y \in \mathbb{R}^n$ given by

$$Y = X\beta_0 + \epsilon. \tag{3.1}$$

Here X is a known $n \times N$ measurement matrix, and $\epsilon \in \mathbb{R}^n$ is the measurement noise. The ratio $\frac{n}{N} \in (0, \infty)$ is denoted by δ and is constant.

Given the observed vector $Y = X\beta_0 + \epsilon$, the AMP decoder generates successive estimates of the unknown vector β_0, with the estimates denoted by $\{\beta^t\}$, where $\beta^t \in \mathbb{R}^N$ for $t = 1, 2, \ldots$. Set the initial estimate $\beta^0 = 0$, the all-zeros vector. For $t = 0, 1, \ldots$, compute

$$z^t = Y - X\beta^t + \frac{z^{t-1}}{n} \sum_{i=1}^{N} \eta'_{t-1}([X^* z^{t-1}]_i + \beta_i^{t-1}), \tag{3.2}$$

$$\beta^{t+1} = \eta_t(\beta^t + X^* z^t), \tag{3.3}$$

using an appropriately-chosen sequence of functions $\{\eta_t\}_{t \geq 0} : \mathbb{R} \to \mathbb{R}$. In (3.2) and (3.3), η_t acts component-wise when applied to a vector, η'_t denotes its (weak) derivative, and quantities with a negative index are set to zero.

For a Gaussian measurement matrix X with entries that are i.i.d. $\sim \mathcal{N}(0, 1/n)$, it was rigorously proven [6, 20] that the performance of AMP can be characterized in the large system limit via a simple scalar iteration called *state evolution*. In the work that follows, we give a finite-sample version of this result. We derive a concentration result (Theorem 2) that implies that the probability of Δ-deviation between $\frac{1}{N}\|\beta_0 - \beta^t\|^2$ and its limiting constant value falls exponentially in n. Empirical findings have previously shown accuracy of the state evolution equations for practically-sized n, for example of the order of several hundreds [4], and the work presented in the next chapter provides theoretical support of such findings.

3.2 AMP Performance

3.2.1 Assumptions

Throughout the chapter we will make the following assumptions.

- **Signal:** The entries of the signal β_0 are i.i.d. according to a sub-Gaussian[1] distribution referred to as p_β.
- **Measurement Matrix:** The entries of measurement matrix $X \in \mathbb{R}^{n \times N}$ are i.i.d. $\sim \mathcal{N}(0, 1/n)$.
- **Measurement Noise:** Assume that the measurement noise ϵ has entries distributed i.i.d. according to p_ϵ with mean 0 and $\mathbb{E}[\epsilon_i^2] = \sigma^2 < \infty$ for $i \in [n]$. Moreover we assume for $\Delta \in (0, 1)$ and positive constant κ,

$$\mathbf{Pr}\left(\left|\frac{\|\epsilon\|^2}{n} - \sigma^2\right| \geq \Delta\right) \leq e^{-\kappa n \Delta^2}. \tag{3.4}$$

  This is true when the entries of ϵ are i.i.d. sub-Gaussian, though (3.4) holds more generally.

- **The Functions η_t:** The de-noising functions, $\eta_t : \mathbb{R} \to \mathbb{R}$, used in (3.3) are Lipschitz continuous for each $t \geq 0$ and, therefore, are also weakly differentiable with weak derivative denoted η_t'. Further, η_t' is assumed to be differentiable, except possibly at a finite number of points, with bounded derivative everywhere it exists.

In what follows, $\kappa > 0$ is an arbitrary constant and $\Delta > 0$ an arbitrarily small value that does not depend on n.

3.2.2 State Evolution

We next show that knowledge of the signal distribution p_β and the noise distribution p_ϵ can help choose good denoting functions $\{\eta_t\}$, however, the performance results hold for

1. A random variable X is sub-Gaussian if there exist positive constants c, κ such that $P(|X| > t) \leq ce^{-\kappa t^2}$, $\forall t > 0$. Examples of sub-Gaussian random variables include zero-mean Gaussian and bounded random variables [24].

any choice of functions $\{\eta_t\}$. Additionally we introduce a simple scalar iteration called state evolution, which predicts the performance of AMP in the large system limit. Scalar iteration was previously discussed in Section 2.5 for the specific case of the AMP decoder. Given p_β, let $\beta \in \mathbb{R} \sim p_\beta$. Let $\sigma_0^2 = \mathbb{E}\left\{\beta^2\right\}/\delta > 0$, and define quantities $\{\tau_t^2\}_{t\geq 0}$ and $\{\sigma_t^2\}_{t\geq 0}$ as follows.

$$\sigma_t^2 = \frac{1}{\delta}\mathbb{E}\left\{(\eta_{t-1}(\beta + \tau_{t-1}Z) - \beta)^2\right\}, \tag{3.5}$$

$$\tau_t^2 = \sigma^2 + \sigma_t^2, \tag{3.6}$$

where $\beta \sim p_\beta$ and $Z \sim \mathcal{N}(0,1)$ are independent random variables.

Similarly to the case of the AMP decoder in Section 2.5, the AMP update for the estimate (3.3) is underpinned by the following key property of the vector $X^*z^t + \beta^t$, which as before is called the 'test statistic': *for large n, the test statistic $X^*z^t + \beta^t$ is approximately distributed as $\beta_0 + \tau_t Z$, where Z is an i.i.d. $\mathcal{N}(0,1)$ random vector independent of β_0 and τ_t is given in* (3.6). In light of this property, a natural way to generate β^{t+1} from the "effective observation" $X^*z^t + \beta^t = s$ is via the conditional expectation:

$$\beta^{t+1}(s) = \mathbb{E}[\beta \mid \beta + \tau_t Z = s], \tag{3.7}$$

i.e., β^{t+1} is the minimum mean square error estimate of β_0 given the noisy observation $\beta_0 + \tau_t Z$. Thus if p_β is known, the Bayes-optimal choice for $\eta_t(s)$ is the conditional expectation in (3.7).

In the definition of the "modified residual" z^t given in (3.2), the third term, often call the 'Onsager' correction term, is crucial to ensure that the effective observation $X^*z^t + \beta^t$ has the above distributional property. For intuition about the role of this 'Onsager' term, the reader is referred to [6, Section I-C].

We now review two examples to illustrate how full or partial knowledge of p_β can guide the choice of the denoising function η_t. Note that the work in Section 2.5 defines denoising functions $\{\eta_t\}_{t\geq 0}$ in the case of the AMP decoder using property (3.7). The assumptions made for the AMP decoder, however, are slightly different than those we make in this

chapter.

In the first example, suppose we know that each element of β_0 is chosen uniformly at random from the set $\{+1, -1\}$. Computing the conditional expectation in (3.7) with this p_β, we obtain $\eta_t(s) = \tanh(s/\tau_t^2)$ [6]. The constants τ_t^2 are determined iteratively from the state evolution equations (3.5)-(3.6).

As a second example, consider the compressed sensing problem, where $\delta < 1$, and p_β is such that the probability that any entry of β_0 equals 0 is $1 - \xi$. The parameter $\xi \in (0, 1)$ determines the sparsity of β_0, with β_0 expected to have $N\xi$ non-zeros. For this problem, the authors in [4, 5] suggested the choice $\eta_t(s) = \eta(s; \theta_t)$, where the soft-thresholding function η is defined as

$$\eta(s, \theta) = \begin{cases} (s - \theta), & \text{if } s > \theta, \\ 0 & \text{if } -\theta \leq s \leq \theta, \\ (s - \theta), & \text{if } s < -\theta. \end{cases}$$

The threshold θ_t at step t is set to $\theta_t = \alpha\tau_t$, where α is a tunable constant and τ_t is determined by (3.6). However, computing τ_t using (3.6) requires full knowledge of p_β. In the absence of such knowledge, we can estimate τ_t^2 by $\frac{\|z^t\|^2}{n}$: our concentration result (Lemma 5(f)) shows that this approximation is very good for large n. To fix α, one could run the AMP with several different values of α, and choose the one that gives the smallest value of $\frac{\|z^t\|^2}{n}$ for large t.

We note that in each of the two above examples η_t is Lipschitz, and its derivative satisfies the assumption stated above.

3.2.3 AMP Performance Guarantees

Recall the definition of *pseudo-Lipschitz* functions from [6].

Definition 3.2.1. *A function $\phi : \mathbb{R}^m \to \mathbb{R}$ is pseudo-Lipschitz (of order 2) if there exists a constant $L > 0$ such that for all $x, y \in \mathbb{R}^m$,*

$$|\phi(x) - \phi(y)| \leq L(1 + \|x\| + \|y\|)\|x - y\|, \tag{3.8}$$

where $\|\cdot\|$ denotes the Euclidean norm.

Our result, Theorem 2, is a concentration inequality for pseudo-Lipschitz loss functions.

Theorem 2. *With the assumptions stated in Subsection 3.2.1, the following holds for any pseudo-Lipschitz function $\phi : \mathbb{R}^2 \to \mathbb{R}$, $\Delta < \Delta_0$, and $t \geq 0$:*

$$P\left(\left|\frac{1}{N}\sum_{i=1}^{N}\phi(\beta_i^{t+1}, \beta_{0_i}) - \mathbb{E}\left[\phi\left(\eta_t\left(\beta + \tau_t Z\right), \beta\right)\right]\right| \geq \Delta\right) \leq Ke^{-\kappa_t n\Delta^2}. \tag{3.9}$$

The expectation in (3.9) *is computed with independent random variables $\beta \sim p_\beta$ and $Z \sim \mathcal{N}(0,1)$, and τ_t is given by* (3.5)-(3.6)*.*

The positive constants $\Delta_0 < 1$ and K, κ_t do not depend on n, but their values are not exactly specified.

The probability in (3.9) is with respect to the product measure on the space of the measurement matrix A, signal β_0, and the noise ϵ.

Remarks:

1. By considering the pseudo-Lipschitz function $\phi(a,b) = (a-b)^2$, Theorem 2 proves that state evolution tracks the mean square error of the AMP estimates with exponentially small probability of error in the sample size n. Indeed, for all $t \geq 0$ and $\Delta < \Delta_0$,

$$P\left(\left|\frac{\|\beta^{t+1} - \beta_0\|^2}{N} - \delta\sigma_{t+1}^2\right| \geq \Delta\right) \leq Ke^{-\kappa_t n\Delta^2}, \tag{3.10}$$

   where σ_t^2 is given by (3.5).

   Similarly, taking $\phi(a,b) = |a-b|$, the theorem implies that the normalized ℓ_1-error $\frac{1}{N}\|\beta^{t+1} - \beta_0\|_1$ is concentrated around $\mathbb{E}|\eta_t\left(\beta + \tau_t Z\right) - \beta|$.

2. Asymptotic convergence results of the kind given in [6, 20] are implied by Theorem 2. Indeed, from Theorem 2 we have

$$\sum_{N=1}^{\infty} P\Big(\Big|\frac{1}{N}\sum_{i=1}^{N}\phi(\beta_i^{t+1}, \beta_{0_i}) - \mathbb{E}\left[\phi(\eta_t\left(\beta + \tau_t Z\right), \beta)\right]\Big| \geq \Delta\Big) < \infty.$$

Therefore the Borel-Cantelli lemma implies that

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \phi(\beta_i^{t+1}, \beta_{0_i}) \overset{a.s.}{=} \mathbb{E}\left[\phi\left(\eta_t\left(\beta + \tau_t Z\right), \beta\right)\right].$$

Though the concentration result, Theorem 2, is proved for the high-dimensional regression model (3.1), we expect that it can be extended to other settings where it has been rigorously proven that state evolution accurately characterizes the AMP performance in the asymptotic limit, e.g. the LASSO normalized risk [20], robust high-dimensional M-estimation [26], AMP with spatially coupled matrices [19], and Generalized Approximate Message Passsing [22, 27]. These extensions will be discussed in a future paper.

3.3 Technical Lemma

The main ingredients in the proof of Theorem 2 are two technical lemmas (Lemmas 4 and 5). In what follows we introduce notation that will be used in the lemmas, state the two lemmas, and finally use them to prove Theorem 2. The proofs of the lemmas are included in the appendix, and we provide here some intuition and comments about the statements. Detailed proofs can also be found in [25, Sections 3, 5].

For consistency and ease of comparison, we use notation similar to [6], and consequently similar to that used in providing performance guarantees for the AMP decoder in Section 2.7. Define the following column vectors recursively for $t \geq 0$, starting with $\beta^0 = 0$ and $z^0 = y$.

$$\begin{aligned} h^{t+1} &= \beta_0 - (X^* z^t + \beta^t), \qquad & q^t &= \beta^t - \beta_0, \\ b^t &= \epsilon - z^t, & m^t &= -z^t. \end{aligned} \tag{3.11}$$

Recall that $\beta_0 \in \mathbb{R}^N$ is the vector we would like to recover and $\epsilon \in \mathbb{R}^n$ is the measurement noise. The vector h^{t+1} is the noise in the effective observation $X^* z^t + \beta^t$, while q^t is the error in the estimate β^t. Lemma 5 will show that h^t and m^t are approximately i.i.d. $\mathcal{N}(0, \tau_t^2)$, while q^t and b^t are approximately i.i.d. $\mathcal{N}(0, \sigma_t^2)$.

Let

$$\lambda_t := -\frac{1}{n}\sum_{i=1}^{N} \eta'_{t-1}(\beta_{0_i} - h_i^t), \tag{3.12}$$

and for $t > 0$, define the matrices

$$\begin{aligned} M_t &:= [m^0 \mid \ldots \mid m^{t-1}], \quad Q_t := [q^0 \mid \ldots \mid q^{t-1}], \\ B_t &:= [b^0 \mid \ldots \mid b^{t-1}], \quad H_t := [h^1 \mid \ldots \mid h^t]. \end{aligned} \tag{3.13}$$

The notation $[c_1 \mid \ldots \mid c_k]$ is used to denote a matrix with columns $c_1, \ldots, c_k$. Note that M_0, B_0, H_0, and Q_0 are the all-zero vector.

We use the notation $m_\parallel^t$ and $q_\parallel^t$ to denote the projection of m^t and q^t onto the column space of M_t and Q_t, respectively. Let

$$\alpha^t := (\alpha_0^t, \ldots, \alpha_{t-1}^t)^*, \quad \gamma^t := (\gamma_0^t, \ldots, \gamma_{t-1}^t)^* \tag{3.14}$$

be the coefficient vectors of these projections, i.e.,

$$m_\parallel^t := \sum_{r=0}^{t-1} \alpha_r^t m^r, \qquad q_\parallel^t := \sum_{r=0}^{t-1} \gamma_r^t q^r. \tag{3.15}$$

The projections of m^t and q^t onto the orthogonal complements of M^t and Q^t, respectively, are denoted by

$$m_\perp^t = m^t - m_\parallel^t, \quad q_\perp^t = q^t - q_\parallel^t \tag{3.16}$$

Lemma 5 shows that for large n, the entries of α^t and γ^t concentrate around constants. We now specify these constants. Let $\{\tilde{Z}_t\}$, $t \geq 0$ be a sequence of zero-mean jointly Gaussian random variables such that for $r, t \geq 0$ the covariance

$$\mathbb{E}[\tilde{Z}_r \tilde{Z}_t] = \frac{(\sigma^2 + E_{r,t})}{\tau_r \tau_t}, \tag{3.17}$$

where

$$E_{r,t} := \frac{\mathbb{E}[(\eta_{r-1}(\beta + \tau_{r-1}\tilde{Z}_{r-1}) - \beta)(\eta_{t-1}(\beta + \tau_{t-1}\tilde{Z}_{t-1}) - \beta)]}{\delta} \tag{3.18}$$

with $\eta_{-1}(\cdot) = 0$. From the definitions of τ_t, σ_t in (3.5) and (3.6), note that $E_{t,t} = \sigma_t^2$ and thus $\mathbb{E}[\tilde{Z}_t^2] = 1$ for $t \geq 0$. Define matric $C^t \in \mathbb{R}^{t \times t}$ such that

$$C^t_{i+1,j+1} = E_{i,j}, \quad 0 \leq i, j \leq t-1. \tag{3.19}$$

With this definitions, the concentrating values for γ^t and α^t are

$$\hat{\gamma}^t := (C^t)^{-1} E_t, \quad \text{and} \quad \hat{\alpha}^t := (\sigma^2 + C^t)^{-1}(\sigma^2 + E_t), \tag{3.20}$$

where

$$E_t := (E_{0,t} \ldots, E_{t-1,t})^*. \tag{3.21}$$

Finally, let $(\sigma_0^\perp)^2 := \sigma_0^2$ and $(\tau_0^\perp)^2 := \tau_0^2$, and for $t > 0$ define

$$\begin{aligned}
(\sigma_t^\perp)^2 &:= \sigma_t^2 - (\hat{\gamma}^t)^* E_t = \sigma_t^2 - E_t^* (C^t)^{-1} E_t, \\
(\tau_t^\perp)^2 &:= \tau_t^2 - (\hat{\alpha}^t)^*(\sigma^2 + E_t) = \tau_t^2 - (\sigma^2 + E_t)^*(\sigma^2 + C^t)^{-1}(\sigma^2 + E_t).
\end{aligned} \tag{3.22}$$

Lemma 3. *For $t > 0$, matrices C^t and $\sigma^2 + C^t$ are invertible where C^t is defined in* (3.19)*. For $t > 0$, $(\sigma_t^\perp)^2 > 0$ and $(\tau_t^\perp)^2 > 0$ using the definitions in* (3.22)*.*

Proof. The proof can be found in Appendix B.3 □

The proof of Theorem 2 consists of two main lemmas. Lemma 4 specifies the conditional distribution of the vectors h^{t+1} and b^t given the matrices in (3.13) as well as β_0, ϵ. This conditional distribution shows that h^{t+1} and b^t can each be expressed as the sum of an i.i.d. Gaussian random vector and a deviation term. Lemma 5 provides concentration results showing that the deviation terms in Lemma 4 are small with high probability, as well as concentration inequalities for various inner products and functions involving $\{h^{t+1}, q^t, b^t, m^t\}$ all defined in (3.11).

3.3.1 Conditional Distribution Lemma

Define $\mathscr{S}_{t_1,t_2}$ to be the sigma-algebra generated by

$$b^0, ..., b^{t_1-1}, m^0, ..., m^{t_1-1}, h^1, ..., h^{t_2}, q^0, ..., q^{t_2}, \text{ and } \beta_0, \epsilon.$$

The following lemma specifies the conditional distributions of b^t and h^{t+1} given $\mathscr{S}_{t,t}$ and $\mathscr{S}_{t+1,t}$, respectively.

Lemma 4. *For the vectors h^{t+1} and b^t defined in* (3.11)*, the following hold for $t \geq 1$:*

$$b^0|_{\mathscr{S}_{0,0}} \overset{d}{=} \sigma_0^{\perp} Z_0', \quad h^1|_{\mathscr{S}_{1,0}} \overset{d}{=} \tau_0^{\perp} Z_0 + \Delta_{1,0},$$
$$b^t|_{\mathscr{S}_{t,t}} \overset{d}{=} \sum_{r=0}^{t-1} \hat{\gamma}_r^t b^r + \sigma_t^{\perp} Z_t' + \Delta_{t,t}, \quad h^{t+1}|_{\mathscr{S}_{t+1,t}} \overset{d}{=} \sum_{r=0}^{t-1} \hat{\alpha}_r^t h^{r+1} + \tau_t^{\perp} Z_t + \Delta_{t+1,t}.$$

where $Z_0, Z_t \in \mathbb{R}^N$ and $Z_0', Z_t' \in \mathbb{R}^n$ are i.i.d. $\mathcal{N}(0,1)$ random vectors that are independent of the corresponding conditioning sigma algebras. The deviation terms are

$$\Delta_{0,0} := Z_0' \left(\frac{\|q^0\|}{\sqrt{n}} - \sigma_0^{\perp} \right), \tag{3.23}$$
$$\Delta_{1,0} := Z_0 \left(\frac{\|m^0\|}{\sqrt{n}} - \tau_0 \right) - \frac{\|m^0\|}{\sqrt{n}} \frac{q^0}{\|q^0\|} \bar{Z}_0 + q^0 \left(\frac{\|q^0\|^2}{n} \right)^{-1} \left(\frac{(b^0)^* m_0}{n} - \frac{\|q^0\|^2}{n} \right), \tag{3.24}$$

where $\bar{Z}_0$ is a standard Gaussian random variable. For $t > 0$,

$$\begin{aligned} \Delta_{t,t} := & \sum_{r=0}^{t-1} (\gamma_r^t - \hat{\gamma}_r^t) b^r + Z_t' \left(\frac{\|q_\perp^t\|}{\sqrt{n}} - \sigma_t^{\perp} \right) - \frac{\|q_\perp^t\| \tilde{M}_t \bar{Z}_t'}{n} \\ & + M_t \left(\frac{M_t^* M_t}{n} \right)^{-1} \left(\frac{H_t^* q_\perp^t}{n} - \frac{M_t^*}{n} \left[\lambda_t m^{t-1} - \sum_{r=1}^{t-1} \lambda_r \gamma_r^t m^{r-1} \right] \right), \end{aligned} \tag{3.25}$$
$$\begin{aligned} \Delta_{t+1,t} := & \sum_{r=0}^{t-1} (\alpha_r^t - \hat{\alpha}_r^t) h^{r+1} + Z_t \left(\frac{\|m_\perp^t\|}{\sqrt{n}} - \tau_t^{\perp} \right) - \frac{\|m_\perp^t\| \tilde{Q}_{t+1} \bar{Z}_t}{n} \\ & + Q_{t+1} \left(\frac{Q_{t+1}^* Q_{t+1}}{n} \right)^{-1} \left(\frac{B_{t+1}^* m_t^{\perp}}{n} - \frac{Q_{t+1}^* q_\perp^t}{n} \right). \end{aligned} \tag{3.26}$$

In the above, $\bar{Z}_t \in \mathbb{R}^{t+1}$ and $\bar{Z}_t' \in \mathbb{R}^t$ are random vectors with i.i.d. $\mathcal{N}(0,1)$ entries. ($\bar{Z}_t$ is defined via a projection of Z_t, and $\bar{Z}_t'$ via a projection of Z_t'.) The terms $\hat{\gamma}_i^t$ and $\hat{\alpha}_i^t$

for $0 \le i \le t-1$ *are defined in* (3.20)*, and* $(\tau_t^\perp)^2$ *and* $(\sigma_t^\perp)^2$ *in* (3.22)*. Matrices* $\tilde{Q}_t$ *and* $\tilde{M}_t$ *form an orthogonal basis for the column space of* Q_t *and* M_t*, respectively, such that* $\tilde{Q}_t^* \tilde{Q}_t = \tilde{M}_t^* \tilde{M}_t = n\mathsf{I}_{t\times t}$*.*

Proof. The proof, which can be found in Appendix B.4, is based on the conditional distribution of X given $\mathscr{S}_{t,t}$ or $\mathscr{S}_{t+1,t}$, which was derived in [6, Lemmas 10, 12] and presented here in Appendix B.4. □

The conditional distribution representation in Lemma 4 implies that for each $t \ge 0$, h^{t+1} is the sum of an i.i.d. $\mathcal{N}(0, \tau_t^2)$ random vector plus a deviation term. This is straightforward to verify for the case where denoising function $\eta(\cdot)$ is chosen as the conditional expectation of β given the noisy observation $\beta + \tau_t Z$, as in (3.7). In this case, it can be shown that $E_{r,t}$ in (3.18) equals σ_t^2 for $0 \le r \le t$. This is shown by applying the orthogonality principle to the definition, after verifying that the following Markov property holds for the jointly Gaussian $\tilde{Z}_r, \tilde{Z}_t$ with covariance given by (3.17):

$$\mathbb{E}[\beta \mid \beta + \tau_t \tilde{Z}_t,\ \beta + \tau_r \tilde{Z}_r] = \mathbb{E}[\beta \mid \beta + \tau_t \tilde{Z}_t], \quad 0 \le r \le t.$$

With $E_{r,t} = \sigma_t^2$ for $r \le t$, the quantities in (3.20)–(3.22) simplify to the following for $t > 0$:

$$\begin{aligned} &\hat{\gamma}^t = [0, \dots, 0, \sigma_t^2/\sigma_{t-1}^2], \quad \hat{\alpha}^t = [0, \dots, 0, \tau_t^2/\tau_{t-1}^2], \\ &(\sigma_t^\perp)^2 := \sigma_t^2 \left(1 - \frac{\sigma_t^2}{\sigma_{t-1}^2}\right), \quad (\tau_t^\perp)^2 := \tau_t^2 \left(1 - \frac{\tau_t^2}{\tau_{t-1}^2}\right). \end{aligned} \tag{3.27}$$

Using (3.27) in Lemma 4, we get

$$h^{t+1}|_{\mathscr{S}_{t+1,t}} \overset{d}{=} \frac{\tau_t^2}{\tau_{t-1}^2} h^t + \tau_t^\perp Z_t + \Delta_{t+1,t} \tag{3.28}$$

Assuming $h^t \overset{d}{=} \tau_{t-1}\tilde{Z}_{t-1} + \Delta_t$, then substituting in (3.28) gives

$$h^{t+1} \overset{d}{=} \frac{\tau_t^2}{\tau_{t-1}} \tilde{Z}_{t-1} + \tau_t^\perp Z_t' + \Delta_t + \Delta_{t+1,t} \overset{d}{=} \tau_t \tilde{Z}_t + \Delta_t + \Delta_{t+1,t}.$$

To obtain the last equality above, we combine the independent Gaussians $\tilde{Z}_{t-1}$, Z'_t using the expression for $\tau_t^\perp$ in (3.27). It can be similarly seen that b^t is the sum of an i.i.d. $\mathcal{N}(0, \sigma_t^2)$ random vector plus a deviation term. The next lemma shows that these deviation terms are small with high probability.

3.3.2 Concentration Lemma

We use the shorthand $X_n \doteq c$ to denote the concentration inequality $P(|X_n - c| \geq \Delta) \leq Ke^{-\kappa_t n \Delta^2}$.

Lemma 5. *With the $\doteq$ notation defined above, the following statements hold for $t \geq 0$.*

(a)

$$\frac{\|\Delta_{t+1,t}\|^2}{N} \doteq 0, \quad \frac{\|\Delta_{t,t}\|^2}{n} \doteq 0.$$

(b)

$$\frac{(h^{t+1})^* q^0}{n} \doteq 0, \quad \frac{(b^t)^* \epsilon}{n} \doteq 0, \quad \frac{(m^t)^* \epsilon}{n} \doteq -\sigma^2.$$

(c) i) *For pseudo-Lipschitz functions $\phi_h : \mathbb{R}^{t+2} \to \mathbb{R}$*

$$\frac{1}{N} \sum_{i=1}^{N} \phi_h \left(h_i^1, \ldots, h_i^{t+1}, \beta_{0_i} \right) \doteq \mathbb{E}\left[\phi_h \left(\tau_0 \tilde{Z}_0, \ldots, \tau_t \tilde{Z}_t, \beta \right) \right].$$

The random variables $\tilde{Z}_0, \ldots, \tilde{Z}_t$ are jointly Gaussian with zero mean and covariance given by (3.17).

ii) *Let $\psi_h : \mathbb{R} \to \mathbb{R}$ be a bounded function that is almost everywhere differentiable, with bounded derivative where it exists. Then for finite constants $(a_0, \ldots, a_t)$,*

$$\frac{1}{N} \sum_{i=1}^{N} \psi_h(\beta_{0_i} - \sum_{r=0}^{t} a_r h_i^{r+1}) \doteq \mathbb{E}\left[\psi_h(\beta - \sum_{r=0}^{t} a_r \tau_r \tilde{Z}_r) \right].$$

(d) For all $0 \leq r \leq t$,

$$\frac{(q^0)^* q^{t+1}}{n} \doteq \sigma_{t+1}^2, \ \frac{(q^{r+1})^* q^{t+1}}{n} \doteq \sigma_{t+1}^2, \ \frac{(b^r)^* b^t}{n} \doteq \sigma_t^2.$$

(e) Define $\hat{\lambda}_{t+1} = -\frac{1}{\delta}\mathbb{E}[\eta_t'(\beta - \tau_t \tilde{Z}_t)]$. *For all* $0 \le r \le t$,

$$\frac{(h^{t+1})^* q^{r+1}}{n} \doteq \hat{\lambda}_{r+1}\tau_t^2, \quad \frac{(h^{r+1})^* q^{t+1}}{n} \doteq \hat{\lambda}_{t+1}\tau_t^2,$$
$$\frac{(b^r)^* m^t}{n} \doteq \sigma_t^2, \quad \frac{(b^t)^* m^r}{n} \doteq \sigma_t^2.$$

(f) For all $0 \le r \le t$,

$$\frac{(h^{r+1})^* h^{t+1}}{N} \doteq \tau_t^2, \quad \frac{(m^r)^* m^t}{n} \doteq \tau_t^2$$

(g) For $0 \le k \le t$ *and* $0 \le k' \le t-1$,

$$\gamma_k^{t+1} \doteq \hat{\gamma}_k^{t+1}, \quad \alpha_{k'}^t \doteq \hat{\alpha}_{k'}^t,$$

where $\hat{\gamma}_k^{t+1}, \hat{\alpha}_{k'}^t$ *are defined in* (3.20).

(h)

$$\frac{\|q_\perp^{t+1}\|^2}{n} \doteq (\sigma_{t+1}^\perp)^2, \quad \frac{\|m_\perp^t\|^2}{n} \doteq (\tau_t^\perp)^2,$$

where $\sigma_{t+1}^\perp, \tau_t^\perp$ *are defined in* (3.22).

Many of the statements in Lemma 5 are similar to those in [6, Lemma 1], but we provide concentration inequalities rather than asymptotic convergence statements. The proof of the lemma is given in [25] and in Appendix B.5 of this document. It is based on induction starting at time $t = 0$, sequentially proving the statements (a)–(h). Though the proof of Theorem 2 below requires only the concentration result (c) above, the remaining concentration inequalities are required for the inductive proof.

We hope in the future to obtain explicit bounds for the constants in K, κ_t, and Δ_0 in Theorem 2. Such bounds would make the non-asymptotic result more powerful. The main difficulty here is tracking the constants throughout the induction step in Lemma 5 : the concentration inequalities we derive for each time step t depend on those proved for the previous step.

Recalling that h^t is the noise in the effective observation $X^* z^t + \beta^t$, and q^t is the estimation error $\beta^t - \beta_0$, the lemma specifies the correlation between these vectors in different steps of the AMP algorithm.

3.3.3 Proof of Theorem 2

Applying Part (c)(i) of Lemma 5 to a pseudo-Lipschitz (PL) function of the form $\phi_h(h^{t+1}, \beta_0)$, we get

$$P\left(\left|\frac{1}{N}\sum_{i=1}^{N}\phi_h(h_i^{t+1}, \beta_{0_i}) - \mathbb{E}\left[\phi(\tau_t Z, \beta)\right]\right| \geq \Delta\right) \leq Ke^{-\kappa_t n\Delta^2} \tag{3.29}$$

where the random variables $Z \sim N(0,1)$ and $\beta \sim p_\beta$ are independent. Now let

$$\phi_h(h_i^{t+1}, \beta_{0_i}) := \phi(\eta_t(\beta_{0_i} - h_i^{t+1}), \beta_{0_i}), \tag{3.30}$$

where ϕ is the PL function in the statement of the Theorem. The function $\phi_h(h_i^{t+1}, \beta_{0_i})$ in (3.30) is PL since ϕ is PL and η_t is Lipschitz. We therefore obtain

$$P\left(\left|\frac{1}{N}\sum_{i=1}^{N}\phi(\eta_t(\beta_{0_i} - h_i^{t+1}), \beta_{0_i}) - \mathbb{E}\left[\phi(\eta_t(\beta - \tau_t Z), \beta)\right]\right| \geq \Delta\right) \leq Ke^{-\kappa_t n\Delta^2}. \tag{3.31}$$

The proof is completed by noting from (3.3) and (3.11) that

$$\beta^{t+1} = \eta_t(X^* z^t + \beta^t) = \eta_t(\beta_0 - h^{t+1}).$$

Chapter 4

Channel Communication with a Bernoulli Dictionary

In this chapter we present work analyzing the performance of the adaptive successive decoder, introduced in Section 2.3, when the design matrix is equiprobable Bernoulli instead of the traditionally-studied i.i.d. Gaussian. A Bernoulli dictionary reduces the computational complexity and memory requirements of the coding scheme providing better performance. In Section 4.1 we discuss the challenges associated with coding with the Bernoulli dictionary and we introduce the Method of Nearby Measures, a powerful tool that greatly simplifies our analysis. In Section 4.2 we discuss work towards analyzing the performance of the adaptive successive decoder in the Bernoulli dictionary case and specifically we give a distributional analysis of the first step of the algorithm.

4.1 The Bernoulli Dictionary Case

In this chapter we consider the case of a Bernoulli $\left\{-\frac{1}{\sqrt{n}}, +\frac{1}{\sqrt{n}}\right\}$ dictionary, instead of the Gaussian dictionary, and present a distributional analysis of the key components of the decoding task when this design matrix is used. A statistical understanding of the decoder is necessary if one wants to provide performance guarantees for the adaptive successive and soft-decision iterative decoders in this setting. The Bernoulli dictionary is of interest, because its use would reduce memory storage requirements of the coding scheme and increase

computational efficiency since multiplication and division operations become addition and subtraction. With a Bernoulli dictionary, the output is modeled as follows:

$$Y = X\beta + \epsilon \tag{4.1}$$

where $\epsilon \sim N(0, \sigma^2 \mathsf{I})$ and the entries of X are i.i.d. equiprobable $\left\{-\frac{1}{\sqrt{n}}, +\frac{1}{\sqrt{n}}\right\}$, making the output, Y, a linear combination of independent Bernoulli random variables and noise. Recall from Section 2.2 that $\beta \in \mathcal{B}_{M,L}(P_1, \ldots, P_L)$, and so β is a vector of zeros with a single non-zero value in each section, with the non-zero value equal to $\sqrt{nP_\ell}$. Then considering the model (4.1), the output $Y \in \mathbb{R}^n$ has the following representation:

$$Y = \sum_{\ell=1}^{L} \sqrt{nP_\ell}\, X_\ell + \epsilon \tag{4.2}$$

where X_ℓ for $\ell \in [L]$ is the column of X in section ℓ 'sent' in the codeword.

Statistical decoding requires the study of the conditional distribution of the columns of the dictionary, meaning the Bernoulli random variables, given the output, Y. Considering (4.2), this is the conditional distribution of summands given the sum of independent random variables. This sort of distributional analysis often arises in science and engineering applications and this conditional distribution has been studied extensively in both statistical mechanics and mathematical statistics, by Cover and Campenhout [29] and Csiszár [30]. It has been shown that given the sum, the summands are approximately independent with exponentially tilted distributions.

In what follows we present the method of nearby measures as way to use this information. By bounding the Rényi relative entropy between the true distribution and the independent, exponentially tilted distribution it can be shown that events which are rare under the approximate distribution are also rare under the true distribution. This allows for calculations to be computed using the approximate, usually much simpler, distribution. The following Subsection 4.1.1 introduces the method of nearby measures and the Rényi relative entropy, and Section 4.1.2 establishes upper bounds on the Shannon relative entropy and the Rényi relative entropy of order α between the true distribution and the approximate

distribution.

Using the method of nearby measures, we will show that when analyzing the statistical properties of key components of the decoding algorithm, we can act as if, conditional on the output, the elements of the design matrix are i.i.d. according to an exponentially tilted distribution. This is an approximate distribution which is much easier to study than the true distribution of the elements which would have a complicated dependence structure.

4.1.1 The Method of Nearby Measures

Decoding using the adaptive successive decoder, developed by Joseph and Barron [2], requires the study of the conditional distribution of the columns $(X_1, ..., X_L)$ given the output Y. Notice that each entry in Y is independent of the others, so we focus on a single row of the output given in (4.2): for $i \in [n]$,

$$Y_i = \sum_{\ell=1}^{L} \sqrt{nP_\ell}\ X_{i,\ell} + \epsilon_i. \tag{4.3}$$

We will study the conditional distribution of $(X_{i,1}, ..., X_{i,L})$ given the output Y_i, but in what follows we drop the subscript i when discussing the one-dimensional random variables $X_1, ..., X_L, Y$.

The distribution of summands given the sum of independent random variables has been studied extensively in statistical mechanics motivated by the original work of Boltzmann (see, for example, Lanford [31]) and others in statistics and information theory, for example Cover and van Campenhout [29] and Csiszár [30]. This work states that conditionally given the sum, the summands are distributed approximately independently according to the maximum entropy distribution subject to the mean constraint, which takes the form of exponentially tilted distributions. The statistical decoding problem involving the conditional distribution of $(X_1, ..., X_L)$ given the output Y, is analogous to this situation. Motivated by this work, we hope to be able to approximate the true conditional distribution of $(X_1, ..., X_L)$ given Y by the product of independent, exponentially tilted Bernoulli $\pm\frac{1}{\sqrt{n}}$ distributions, meaning that if an event is rare under the approximate distribution then it remains rare under the true distribution.

Consider the true distribution of $X_1, ..., X_L$ which we define by independent q such that

$$q(x_\ell) = \begin{cases} \frac{1}{2}, & \text{if } x_\ell = \frac{1}{\sqrt{n}} \\ \frac{1}{2}, & \text{if } x_\ell = -\frac{1}{\sqrt{n}} \end{cases} \tag{4.4}$$

for each $\ell \in [L]$. Observation of Y gives rise to $q_{X|Y}$ which we approximate as the product of independent, exponentially tilted distributions. We let $q^a_{X|Y}$ be the tilted distributions given Y. For each $\ell \in [L]$,

$$q^a_{X_\ell|Y}(x_\ell) = \begin{cases} \frac{\exp\{aY\sqrt{P_\ell}\}}{\exp\{aY\sqrt{P_\ell}\}+\exp\{-aY\sqrt{P_\ell}\}}, & \text{if } x_\ell = \frac{1}{\sqrt{n}} \\ \frac{\exp\{-aY\sqrt{P_\ell}\}}{\exp\{aY\sqrt{P_\ell}\}+\exp\{-aY\sqrt{P_\ell}\}}, & \text{if } x_\ell = -\frac{1}{\sqrt{n}} \end{cases} \tag{4.5}$$

where a is an appropriate constant. Let us define $\mathbb{Q}_L$ as the measure associated with the true joint distribution (joint probability mass function) of $(X_1, ..., X_L, Y)$. Similarly, let $\mathbb{Q}^a_L$, a for approximate, be the measure associated with the joint probability distribution of $(X_1, ..., X_L, Y)$ when the conditional distribution of $(X_1, ..., X_L)$ given Y is the product of exponentially tilted distributions. Finally, let q_L and q^a_L be the probability distributions associated with each measure. Specifically,

$$q^a_L(x_1, \ldots, x_L, y) = p_Y(y) \prod_{\ell=1}^{L} q^a_{X_\ell|Y}(x_\ell), \tag{4.6}$$

where $p_Y(y)$ is the probability mass function of Y. Let us define the Rényi relative entropy of order $\alpha > 1$ between these measures, denoted $D_\alpha(\mathbb{Q}_L||\mathbb{Q}^a_L)$, as

$$D_\alpha(\mathbb{Q}_L||\mathbb{Q}^a_L) = \frac{1}{\alpha - 1} \log \mathbb{E}_{\mathbb{Q}_L} \left[\left(\frac{q_L(X_1, ..., X_L, Y)}{q^a_L(X_1, ..., X_L, Y)} \right)^{\alpha-1} \right]. \tag{4.7}$$

If $D_\alpha(\mathbb{Q}_L||\mathbb{Q}^a_L)$ is finite for some $\alpha > 1$, then we can relate probabilities under the true measure $\mathbb{Q}_L$ to probabilities under the approximate measure $\mathbb{Q}^a_L$. This relationship is summarized in the following Lemma.

Lemma 6. *Consider an event A. If the Rényi relative entropy between the two measures is finite for some order $\alpha > 1$, meaning $D_\alpha(\mathbb{Q}_L||\mathbb{Q}^a_L) \leq c_0$ for some constant c_0, then the*

probability of the event under the true measure is upper bounded using the probability under the approximate distribution with the following inequality.

$$\mathbb{Q}_L(A) \leq (e^{c_0}\mathbb{Q}_L^a(A))^{\frac{\alpha-1}{\alpha}} . \tag{4.8}$$

Proof.

$$\begin{aligned}\mathbb{Q}_L(A) = \int q_L(\underline{x})\mathbb{1}_{\{\underline{x}\in A\}}d\underline{x} &= \int \frac{q_L(\underline{x})}{q_L^a(\underline{x})} \cdot q_L^a(\underline{x})\mathbb{1}_{\{\underline{x}\in A\}}d\underline{x} \\ &\overset{(a)}{\leq} \left(\int q_L(\underline{x})\left[\frac{q_L(\underline{x})}{q_L^a(\underline{x})}\right]^{\alpha-1} d\underline{x}\right)^{\frac{1}{\alpha}} (\mathbb{Q}_L^a(A))^{\frac{\alpha-1}{\alpha}} \\ &= \left(e^{D_\alpha(\mathbb{Q}_L||\mathbb{Q}_L^a)}\mathbb{Q}_L^a(A)\right)^{\frac{\alpha-1}{\alpha}} .\end{aligned}$$

upper bound (a) follows from Holder's inequality. □

In the following section we demonstrate how to obtain bounds for both the Shannon relative entropy and the Rényi relative entropy between the two measures for all signal-to-noise ratios, in order that we are able to make use of Lemma 6 when considering statistical decoding.

4.1.2 Bounding Relative Entropy

The Shannon relative entropy between the true distribution $\mathbb{Q}_L$ and the approximate distribution $\mathbb{Q}_L^a$ is defined to be

$$D(\mathbb{Q}_L||\mathbb{Q}_L^a) = \mathbb{E}_{\mathbb{Q}_L}\left[\log \frac{q_L(X_1, ..., X_L, Y)}{q_L^a(X_1, ..., X_L, Y)}\right]. \tag{4.9}$$

This is also the Rényi relative entropy of order $\alpha = 1$. Because the Rényi relative entropy is continuous in α, the upper bound for α just above 1 should be close to the Shannon entropy between the two measures. Before we demonstrate a bound for the Rényi relative entropy, we show that the Shannon relative entropy upper bound is finite for all snr.

Consider the true joint distribution,

$$q_L(x_1, ..., x_L, y) = \phi_\epsilon \left(y - \sum_{\ell=1}^{L} \sqrt{nP_\ell}\ x_\ell \right) \prod_{\ell=1}^{L} q(x_\ell), \tag{4.10}$$

where ϕ_ϵ is the probability mass function associated with $\epsilon \sim N(0, \sigma^2)$, and the approximate joint distribution from (4.6),

$$q_L^a(x_1, ..., x_L, y) = p_Y(y) \prod_{\ell=1}^{L} q^a_{X_\ell|Y}(x_\ell) = p_Y(y) \prod_{\ell=1}^{L} \left(\frac{q(x_\ell) \exp\{ayx_\ell\sqrt{nP_\ell}\}}{\frac{1}{2}\exp\{ay\sqrt{P_\ell}\} + \frac{1}{2}\exp\{-ay\sqrt{P_\ell}\}} \right). \tag{4.11}$$

The following theorem provides an upper bound for the Shannon relative entropy between these two distributions.

Lemma 7. *For any constant a,*

$$D(\mathbb{Q}_L||\mathbb{Q}_L^a) \leq \frac{1}{2}\log(1 + \mathit{snr}) + \frac{1}{2}a^2P(\sigma^2 + P) - aP, \tag{4.12}$$

which is minimized by choosing $a = \frac{1}{\sigma^2+P}$ making the upper bound

$$D(\mathbb{Q}_L||\mathbb{Q}_L^a) \leq \frac{1}{2}\log(1 + \mathit{snr}) - \frac{\mathit{snr}}{2(1 + \mathit{snr})}. \tag{4.13}$$

Proof. The proof can be found in Appendix C.1. □

Notice that the upper bound stated in Theorem 7 is positive for all values of snr, as we would expect of the Shannon relative entropy. This is can be seen by remembering that $\log(1 + x) \leq x$ for all $x > -1$ and so

$$\log(1 + \mathsf{snr}) = -\log\left(1 - \frac{\mathsf{snr}}{1 + \mathsf{snr}}\right) \geq \frac{\mathsf{snr}}{1 + \mathsf{snr}}. \tag{4.14}$$

We next demonstrate bounds for the Rényi relative entropy. We first choose work with the Rényi relative entropy of order $\alpha = 2$ for simplicity. Recall the definition of relative entropy

in (4.7), from which it follows that for $\alpha = 2$,

$$D_2(\mathbb{Q}_L||\mathbb{Q}_L^a) = \log \mathbb{E}_{\mathbb{Q}_L}\left[\left(\frac{q_L(X_1,...,X_L,Y)}{q_L^a(X_1,...,X_L,Y)}\right)\right]. \tag{4.15}$$

The following Theorem upper bounds this relative entropy.

Lemma 8. *For any* $\mathsf{snr} \leq .58$*, there exists a range of* γ *values in the interval* $0 < \gamma < 1 - \frac{\mathsf{snr}}{(1+\mathsf{snr})^2}$ *such that*

$$D_2(\mathbb{Q}_L||\mathbb{Q}_L^a) \leq \log\frac{20}{3} + \left(1+\frac{1}{\gamma}\right)2\mathsf{snr} - \frac{1}{2}\log\left(1-\gamma-\frac{\mathsf{snr}}{(1+\mathsf{snr})^2}\right). \tag{4.16}$$

Proof. The proof can be found in Appendix C.2. □

While the Shannon relative entropy upper bound held for all snr, the Rényi relative entropy upper bound at order $\alpha = 2$ is limited to only small snr. In allowing α to approach 1, the Rényi relative entropy approaches the Shannon relative entropy, and bounds are obtained for all values of snr. The following Theorem bounds the Rényi relative entropy for all values of the snr by allowing α to be arbitrarily small.

Lemma 9. *For any* snr *and any* γ *in the range* $0 < \gamma < \frac{1}{2}$*, there exists a* $\delta = \alpha - 1 > 0$ *such that*

$$D_\alpha(\mathbb{Q}_L||\mathbb{Q}_L^a) \leq \log\frac{4(5)^{1/\delta}}{3} + \left(1+\frac{1}{\gamma}\right)2\mathsf{snr} - \frac{1}{2\delta}\log(\delta(1-\gamma-a^2\sigma^2P)). \tag{4.17}$$

Proof. The proof can be found in Appendix C.3. □

Using this bound and Lemma 6, we are able to demonstrate an upper bound on the error accrued when approximating the true distribution with the tilted distribution. Using knowledge of the distributional behavior of the summands given the sum of independent random variables, and the closeness of measures established by finite Rényi relative entropy, we are able to approximate a distribution which is statistically difficult to analyze with a much simpler distribution with a constant error rate, thus simplifying statistical decoding of superposition coding over the Gaussian white noise channel.

4.2 Decoding with the Bernoulli Dictionary

In this section we discuss how weighted sums of Bernoulli random variables are sub-Gaussian, meaning we show that rare events under the Gaussian measure still have small probabilities of occurrence under the Bernoulli measure. Then we use this information and the Method of Nearby Measures from Section 4.1 to provide distributional analysis of the first step of the adaptive successive decoding algorithm.

Let μ_n be the Bernoulli ± 1 measure which assigns mass $\left(\frac{1}{2}\right)^n$ to each point on the unit cube $\{-1,+1\}^n$ and $\mathbb{P}$ be the standard normal measure on $\mathbb{R}$ having density

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

with respect to the Lebesgue measure. Similarly, the cumulative distribution function of $\mathbb{P}$ is $\Phi(x) = \mathbb{P}((-\infty, x])$. Further define the probability measure $\mathbb{P}_{\mu,\sigma^2}$, density ϕ_{μ,σ^2}, and cumulative distribution Φ_{μ,σ^2} of the normal random variable with mean μ and variance σ^2.

Define $S_{n,a}(x) = \sum_{i=1}^n a_i x_i$ to be a function taking values on $\{-1,+1\}^n$ with respect to μ_n with constants such that $\sum_{i=1}^n a_i^2 = 1$. For τ taking some value larger than the expectation of $S_{n,a}$, which equals 0, we wish to bound Bernoulli tail measure $\mu_n\{S_{n,a} > \tau\}$ by the probability that a standard normal takes a value larger than τ.

Pinelis [32] was the first to give a proof of a stronger form of Eaton's conjecture: for all $\tau > 0$

$$\mu_n\{S_{n,a} \geq \tau\} \leq \kappa(1 - \Phi(\tau)) \leq \frac{\kappa}{\tau}\phi(\tau), \tag{4.18}$$

when $\kappa = \frac{2}{9}e^3 \approx 4.46$ (Eaton's constant) using a proof method which compared moments of the distribution of $S_{n,a}$ with moments of the standard normal. Some years later using a simple induction proof, Bobkov, Götze, and Houdré [33] established inequality (4.18) for $\kappa = [2\,(1 - \Phi(\sqrt{3}))]^{-1} \approx 12$. The most recent development is that of Pinelis [34], who demonstrates that the best possible constant κ for (4.18) falls in the range $\kappa \in [3.18, 3.22]$.

In Lemma 1, we further demonstrate that the distribution of the sum of Bernoulli ± 1 random variables remains sub-gaussian, meaning we can find an upper bound like (4.18), when an independent normal random variable is added to it. This will be needed in what

follows to establish the distributional properties of the test statistics in the first step of the algorithm.

Lemma 10. *Consider the sum of independent Bernoulli random variables,* $\sum_{i=1}^{n} a_i X_i$ *where* $X = (X_1, ..., X_n)$ *has distribution induced by Bernoulli measure* μ_n*. Define* $S = \sum_{i=1}^{n} a_i X_i +$ Z *where* Z *is a mean-zero, normal random variable independent of* X*. Let* S *be distributed according to the probability measure which is the convolution* $(\mu_n \star \mathbb{P}_{0,\sigma_Z^2})$*. Then for* $\tau \geq 0$*,*

$$\boldsymbol{Pr}(S \geq \tau) \leq \kappa(1 - \Phi_{0,\sigma_S^2}(\tau)),$$

where $\sigma_S^2 = \sum_{i=1}^{n} a_i^2 + \sigma_Z^2$ *is the variance of the sum* S*. The symmetric result is also true.*

Proof. See Appendix C.4 for the proof. □

4.2.1 Distributional Analysis of the First Step

The basic decoding problem is this: how does one determine the sent codeword with only knowledge of the received string Y and the dictionary X? We When using sparse superposition coding this task corresponds to determining which columns j of the dictionary X are those belonging to the set $sent = \{j : \beta_j \neq 0\}$. We define the other indices as $other = \{j : \beta_j = 0\}$.

In the first step of adaptive successive decoding, test statistics $\mathcal{Z}_{1,j}$ are computed as the normalized inner product of the j^{th} column of the dictionary with the received vector Y for each index $j \in J = \{1, 2, ..., N\}$. In previous work [2], normalization with $||Y||$ was used because it allowed the $\mathcal{Z}_{1,j}$ test statistics to be normally distributed. Since we no longer have normality when working with the Bernoulli dictionary, we instead normalize using $\sigma_Y \sqrt{n}$ for analytic simplicity. So for each $j = 1, 2, ..., N$,

$$\mathcal{Z}_{1,j} = \frac{X_j^T Y}{\sigma_Y}.$$

These $\mathcal{Z}_{1,j}$ test statistics are then compared to some threshold τ and those indices for which $\mathcal{Z}_{1,j}$ is above the threshold are collected in the decoded set dec_1, that is $dec_1 = \{j \in J : \mathcal{Z}_{1,j} > \tau\}$. Details about subsequent steps of the decoding algorithm can be found in [2],

but we don't consider them here. Instead we complete a full distributional analysis of the firs step.

Define sets

$$\hat{q}_1 = \sum_{j \in sent \cap dec_1} \frac{P_j}{P} \quad \text{and} \quad \hat{f}_1 = \sum_{j \in other \cap dec_1} \frac{P_j}{P}$$

as, respectively, a weighted measure of correct detections and the section average count of false alarms which occur in step (1) of the algorithm. Note that in the above $P_j = P_\ell$ for $j \in \text{sec}(\ell)$. To establish reliability of our decoder, we wish to upper bound the probabilities of the following exception events

$$A_1 = \{\hat{q}_1 < q_1\} \text{ and } B_1 = \{\hat{f}_1 > f_1\}. \tag{4.19}$$

When the probabilities of these events are small, we are ensured at least q_1 correct detections and at most f_1 false alarms occur in the first step with high probability. These bounds would then allow us to establish an upper bound for the fraction of section mistakes. In the definition of the exception events (4.19), the deterministic values q_1 and f_1 are chosen such that $q_1 < \mathbb{E}\hat{q}_1$ and $f_1 > \mathbb{E}\hat{f}_1$ so it is unlikely that these events will occur. To bound the exception events, we take advantage of the properties of the marginal distributions of the test statistics $\mathcal{Z}_{1,j}$. We establish these distribution properties in the following lemma.

Lemma 11. *The test statistic $\mathcal{Z}_{1,j}$ can be represented as*

$$\frac{\beta_j}{\sigma_Y} + \sum_{\substack{j' \in sent \\ j' \neq j}} \frac{\beta_{j'}}{n\sigma_Y} \sum_{i=1}^{n} B_{i,j'} + \frac{\sigma}{\sigma_Y} Z, \tag{4.20}$$

where $B_{i,j'}$ are i.i.d. equiprobable $\{+1, -1\}$ for $i \in [n]$ and $j' \in sent$ with $j' \neq j$ and Z is independent standard normal.

Therefore, for $j \in other$,

$$\boldsymbol{Pr}(\mathcal{Z}_{1,j} \geq \tau) \leq \kappa(1 - \Phi(\tau))$$

where κ is some constant. Let $\sigma^2_{sent} = 1 - \frac{\beta_j^2}{nP}\left(\frac{1}{1+\mathsf{snr}^{-1}}\right)$. Then for $j \in sent$,

$$\boldsymbol{Pr}(\mathcal{Z}_{1,j} \leq \tau) \leq \kappa\, \Phi_{0,\sigma^2_{sent}}(\tau - shift_1)$$

for the same constant κ and a positive shift

$$shift_1 = \frac{\beta_j}{\sqrt{\sigma^2 + P}} = \sqrt{\frac{n(P_j/P)}{1 + \mathsf{snr}^{-1}}}. \tag{4.21}$$

Proof. The proof of Lemma 11 can be found in C.5. □

To find bounds for the probabilities exception events A_1 and B_1 defined in (4.19), we study two joint distributions, that of the random variables $\mathcal{Z}_{1,j}$ for $j \in sent$ and that of the random variables $\mathcal{Z}_{1,j}$ for $j \in other$. Define $\bar{\mathcal{Z}}_{1,sent} \in \mathbb{R}^L$ and $\bar{\mathcal{Z}}_{1,other} \in \mathbb{R}^{L(M-1)}$, as the random vectors holding the two collections of test statistics. In what follows we analyze the distributions of these vectors and find bounds on the probability of exception events (4.19).

Exception Event B_1. Recall from (4.19) that $B_1 = \{\hat{f}_1 > f_1\}$ where

$$\hat{f}_1 = \sum_{j \in\, other \cap dec_1} \frac{P_j}{P} = \sum_{j \in other} \frac{P_j}{P}\, \mathsf{I}\{\mathcal{Z}_{1,j} \geq \tau\}. \tag{4.22}$$

Conditional on the output Y, the elements of the vector $\bar{\mathcal{Z}}_{1,other}$ are independent since the columns X_j for $j \in other$ are independent. Therefore (4.22) is just the weighted sum of independent Bernoulli trials with success probability which can be upper bounded as in by (4.18). Therefore we will find here results similar to those obtained by Barron and Joseph.

Exception Event A_1. Recall from (4.19) that $A_1 = \{\hat{q}_1 < q_1\}$ where

$$\hat{q}_1 = \sum_{j \in\, sent \cap dec_1} \frac{P_j}{P} = \sum_{j \in sent} \frac{P_j}{P}\, \mathsf{I}\{\mathcal{Z}_{1,j} \geq \tau\} = 1 - \sum_{j \in sent} \frac{P_j}{P}\, \mathsf{I}\{\mathcal{Z}_{1,j} \leq \tau\}. \tag{4.23}$$

Again notice that if the elements of the vector $\bar{\mathcal{Z}}_{1,sent}$ were independent when conditioned on Y then we could handle the probability of the event A_1 as in Joseph and Barron since the sum on the right side (4.23) is a sum of independent Bernoulli trials with success probably

which can be upper bounded as in Lemma 11.

Unfortunately the terms of the vector $\bar{\mathcal{Z}}_{1,sent}$ are dependent even when conditioned on Y. In order to deal with the dependence, we will use the method of nearby measures which tells us that if we can bound the Rényi relative entropy between the true joint distribution and an approximate distribution then events that are rare under the approximate distribution are also rare under the true distribution. This is discussed in Section 4.1. As an approximate conditional distribution we choose the distribution such that for j, j' both in *sent*, that $\mathcal{Z}_{1,j}$ and $\mathcal{Z}_{1,j'}$ are independent with their respective marginal distributions so that we have the scenario described in the previous paragraph. The true distribution is of course the true joint distribution of $\bar{\mathcal{Z}}_{1,sent}$ when conditioned on the output.

We will show that the Rényi relative entropy is bounded between the two distributions using the following lemma.

Lemma 12 relates the Rényi relative entropy to the Kullback-Liebler divergence.

Lemma 12. *Consider independent random vectors $U_1, \ldots, U_n \in \mathbb{R}^L$ which are element-wise dependent down $j \in [L]$. We give an upper bound for the Rényi relative entropy between the distribution of the weighted sum $S^n = \sum_{i=1}^n U_i a_i$ for constants $a_1, \ldots, a_n$ and an approximate distribution which assumes element-wise independence of the vectors. Let $\mathbb{P}_{S^n}$ be the measure associated with the true joint probability mass function of the sum over n and let $\mathbb{Q}_{S^n}$ be the measure associated with the approximate distribution. Then,*

$$D_\alpha\left(\mathbb{P}_{S^n}||\mathbb{Q}_{S^n}\right) \leq D_\alpha\left(\mathbb{P}_{S^{n-1}}||\mathbb{Q}_{S^{n-1}}\right) + D_\alpha\left(\mathbb{P}_{S^1}||\mathbb{Q}_{S^1}\right). \tag{4.24}$$

Using result (4.24) *repeatedly we find the following.*

$$D_\alpha\left(\mathbb{P}_{S^n}||\mathbb{Q}_{S^n}\right) \leq \sum_{i=1}^{n} D_\alpha\left(\mathbb{P}_{a_i U^i}||\mathbb{Q}_{a_i U^i}\right) + D_\alpha\left(\mathbb{P}_{S^1}||\mathbb{Q}_{S^1}\right), \tag{4.25}$$

where the Rényi relative entropy on the right side of the above is between the true distribution of the vector $a_i U^i$ and the approximate measure for which the elements are independently distributed according to their marginals.

Proof. The proof of Lemma 12 can be found in Appendix C.6. □

We describe here how Lemma 12 relates to our problem. Since we are only working with indices $j \in sent$,we will assume that for $\ell \in [L]$, the index ℓ refers to the sent column in section ℓ, First notice:

$$\bar{\mathcal{Z}}^n_{1,sent} = \begin{bmatrix} \mathcal{Z}_{1,1} \\ \vdots \\ \mathcal{Z}_{1,L} \end{bmatrix} = \begin{bmatrix} \frac{X_1^* Y}{\sigma_Y} \\ \vdots \\ \frac{X_L^* Y}{\sigma_Y} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n X_{i,1} \left(\frac{Y_i}{\sigma_Y}\right) \\ \vdots \\ \sum_{i=1}^n X_{i,L} \left(\frac{Y_i}{\sigma_Y}\right) \end{bmatrix} = \frac{1}{\sigma_Y} \sum_{i=1}^n \tilde{X}_i Y_i, \tag{4.26}$$

where $\tilde{X}_i$ is the i^{th} row of the dictionary X for only the sent columns. Write $S^n := \frac{1}{\sigma_Y} \sum_{i=1}^n \tilde{X}_i Y_i$ and similarly $S^n_\ell := \frac{1}{\sigma_Y} \sum_{i=1}^n \tilde{X}_{i,\ell} Y_i$ for $\ell \in [L]$.

Then denote the true joint density function of the vector $\bar{\mathcal{Z}}^n_{1,sent}$ when conditioned on Y as $p_{S^n|Y}$ making explicit the dependence on n and the approximate as $\prod_{\ell=1}^L p_{S^n_\ell|Y}$ where $p_{S^n_\ell|Y}$ is the density function for the sum S^n_ℓ when conditioned on Y. Let $\mathbb{P}_{S^n}$ be the measure associated with the true joint density function of $\bar{\mathcal{Z}}^n_{1,sent}$ when conditioned on Y and $\mathbb{Q}_{S^n}$ be the approximate measure such that when j, j' both in $sent$, $\mathcal{Z}_{1,j}$ and $\mathcal{Z}_{1,j'}$ are independent when conditioned on Y with their respective marginal distributions.

This is equivalent to the scenario in Lemma 12 since under true measure, rows $\tilde{X}_1, \ldots, \tilde{X}_L$ are independent but the elements within each row are not independent (when conditioned on Y). Under the approximate measure we assume that the elements within each row are independent according to their marginals when conditioned on Y, making the elements of $\bar{\mathcal{Z}}^n_{1,sent}$ independent as well. Then by Lemma 12,

$$D_\alpha \left(\mathbb{P}_{S^n} || \mathbb{Q}_{S^n}\right) \leq \sum_{i=1}^n D_\alpha \left(\mathbb{P}_{\frac{\tilde{X}_i Y_i}{\sigma_Y}|Y} || \mathbb{Q}_{\frac{\tilde{X}_i Y_i}{\sigma_Y}|Y_i}\right). \tag{4.27}$$

Note that the Rényi relative entropies on the right side of (4.27) are the divergence between the true joint distribution of the summands conditional on the sum and the approximate distribution where each summand is independently distributed according to its marginal distribution. Such relative entropies were studied in Section 4.1.

Appendix A

Chapter 2 Appendix

A.1 Proof of Proposition 2.5.1

For convenience of notation, relabel the N i.i.d. random variables $\{Z_k\}_{k\in[N]}$ as $\{U_j^\ell\}_{j\in[M],\ell\in[L]}$. For any ℓ, U^ℓ denotes the length M vector $\{U_j^\ell\}_{j\in[M]}$, and U is the length N vector $\{U^\ell\}_{\ell\in[L]}$. We have

$$\begin{aligned}
\frac{1}{nP}\mathbb{E}[\beta^*\beta^{t+1}] = \frac{1}{nP}\mathbb{E}[\beta^*\,\eta^t(\beta+\tau_t U)] &\overset{(a)}{=} \frac{1}{nP}\sum_{\ell=1}^{L}\mathbb{E}[\sqrt{nP_\ell}\;\eta^t_{\mathsf{sent}(\ell)}(\beta_\ell+\tau_t U^\ell)] \\
&\overset{(b)}{=} \frac{1}{nP}\sum_{\ell=1}^{L}\mathbb{E}\left[\sqrt{nP_\ell}\cdot\sqrt{nP_\ell}\,\frac{\exp\left(\frac{(\sqrt{nP_\ell}+\tau_t U_1^\ell)\sqrt{nP_\ell}}{\tau_t^2}\right)}{\exp\left(\frac{(\sqrt{nP_\ell}+\tau_t U_1^\ell)\sqrt{nP_\ell}}{\tau_t^2}\right)+\sum_{j=2}^{M}\exp\left(\frac{\tau_t U_j^\ell\sqrt{nP_\ell}}{\tau_t^2}\right)}\right] \\
&= \sum_{\ell=1}^{L}\frac{P_\ell}{P}\,\mathbb{E}\left[\frac{\exp\left((\frac{\sqrt{nP_\ell}}{\tau_t}+U_1^\ell)\frac{\sqrt{nP_\ell}}{\tau_t}\right)}{\exp\left((\frac{\sqrt{nP_\ell}}{\tau_t}+U_1^\ell)\frac{\sqrt{nP_\ell}}{\tau_t}\right)+\sum_{j=2}^{M}\exp\left(U_j^\ell\frac{\sqrt{nP_\ell}}{\tau_t}\right)}\right] = x_{t+1}.
\end{aligned} \tag{A.1}$$

In (a) above, the index of the non-zero term in section ℓ is denoted by $\mathsf{sent}(\ell)$. (b) is obtained by assuming that $\mathsf{sent}(\ell)$ is the first entry in section ℓ — this assumption is valid because the prior on β is uniform over $\mathcal{B}_{M,L}(P_1,\ldots,P_L)$.

Next, consider

$$\frac{1}{nP}\mathbb{E}[\|\beta-\beta^{t+1}\|^2] = 1+\frac{\mathbb{E}[\|\beta^{t+1}\|^2]-2\mathbb{E}[\beta^*\beta^{t+1}]}{nP}. \tag{A.2}$$

Under the assumption that $s^t = \beta + \tau_t Z$, recall from Section **??** that β^{t+1} can be expressed as $\beta^{t+1} = \mathbb{E}[\beta \mid s^t]$. We therefore have

$$\mathbb{E}[\|\beta^{t+1}\|^2] = \mathbb{E}[\,\|\mathbb{E}[\beta|s^t]\|^2\,] = \mathbb{E}[\,(\mathbb{E}[\beta|s^t] - \beta + \beta)^*\mathbb{E}[\beta|s^t]\,] \overset{(a)}{=} \mathbb{E}[\,\beta^*\mathbb{E}[\beta|s^t]\,] = \mathbb{E}[\,\beta^*\beta^{t+1}\,], \tag{A.3}$$

where step (a) follows because $\mathbb{E}[\,(\mathbb{E}[\beta|s^t] - \beta)^*\mathbb{E}[\beta|s^t]\,] = 0$ due to the orthogonality principle. Substituting (A.3) in (A.2) and using (A.1) yields

$$\frac{1}{nP}\mathbb{E}[\|\beta - \beta^{t+1}\|^2] = 1 - \frac{\mathbb{E}[\,\beta^*\beta^{t+1}\,]}{nP} = 1 - x_{t+1}.$$

A.2 Proof of Lemma 1

Treating x_{t+1} in (2.14) as a function of τ, we can define

$$x(\tau) := \sum_{\ell=1}^{L} \frac{P_\ell}{P}\, \mathbb{E}\left[\frac{\exp\left(\frac{\sqrt{nP_\ell}}{\tau}\,(U_1^\ell + \frac{\sqrt{nP_\ell}}{\tau})\right)}{\exp\left(\frac{\sqrt{nP_\ell}}{\tau}\,(U_1^\ell + \frac{\sqrt{nP_\ell}}{\tau})\right) + \sum_{j=2}^{M}\exp\left(\frac{\sqrt{nP_\ell}}{\tau}U_j^\ell\right)}\right], \tag{A.4}$$

where $\{U_j^\ell\}$ are i.i.d. $\sim \mathcal{N}(0,1)$ for $j \in [M],\ \ell \in [L]$. We use the following Lemma, which is proved below, to complete the proof of Lemma 1.

Lemma 13. *For $t = 0, 1, \ldots$, we have*

$$\bar{x}(\tau) := \lim x(\tau) = \lim_{L\to\infty} \sum_{\ell=1}^{L} \frac{P_\ell}{P}\, \mathbf{1}\{c_\ell > 2(\ln 2)R\,\tau^2\} \tag{A.5}$$

where $c_\ell := \lim_{L\to\infty} LP_\ell$.

Proof. From (A.4), $x(\tau)$ can be written as

$$x(\tau) := \sum_{\ell=1}^{L} \frac{P_\ell}{P}\, \mathcal{E}_\ell \tag{A.6}$$

where

$$\mathcal{E}_\ell = \mathbb{E}\left[\frac{\exp\left(\frac{\sqrt{nP_\ell}}{\tau}\,U_1^\ell\right)}{\exp\left(\frac{\sqrt{nP_\ell}}{\tau}\,U_1^\ell\right) + \exp\left(-\frac{nP_\ell}{\tau^2}\right)\sum_{j=2}^{M}\exp\left(\frac{\sqrt{nP_\ell}}{\tau}U_j^\ell\right)}\right]. \tag{A.7}$$

We will prove the lemma by showing that for $\ell = 1, \ldots, L$:

$$\lim \mathcal{E}_\ell = \begin{cases} 1, & \text{if } c_\ell > 2(\ln 2) R\tau^2, \\ 0, & \text{if } c_\ell < 2(\ln 2) R\tau^2, \end{cases} \tag{A.8}$$

where for the power allocation in (2.21),

$$c_\ell = \lim_{L\to\infty} LP_\ell = 2(\ln 2)\mathcal{C}(P+\sigma^2) \lim_{L\to\infty} \left(\frac{\sigma^2}{\sigma^2+P}\right)^{\ell/L}. \tag{A.9}$$

[1] Using the relation $nR = \frac{L \ln M}{\ln 2}$, we can write

$$\frac{nP_\ell}{\tau^2} = \nu_\ell \ln M, \tag{A.10}$$

where $\nu_\ell = \frac{LP_\ell}{R\tau^2 \ln 2}$. Hence $\mathcal{E}_\ell$ in (A.7) can be written as

$$\mathcal{E}_\ell = \mathbb{E}\left[\frac{\exp\left(\sqrt{\ln M}\sqrt{\nu_\ell}\, U_1^\ell\right)}{\exp\left(\sqrt{\ln M}\sqrt{\nu_\ell}\, U_1^\ell\right) + M^{-\nu_\ell}\sum_{j=2}^M \exp\left(\sqrt{\ln M}\sqrt{\nu_\ell}\, U_j^\ell\right)}\right] \tag{A.11}$$

$$= \mathbb{E}\left[\mathbb{E}\left[\frac{\exp\left(\sqrt{\ln M}\sqrt{\nu_\ell}\, U_1^\ell\right)}{\exp\left(\sqrt{\ln M}\sqrt{\nu_\ell}\, U_1^\ell\right) + M^{-\nu_\ell}\sum_{j=2}^M \exp\left(\sqrt{\ln M}\sqrt{\nu_\ell}\, U_j^\ell\right)}\Big| U_1^\ell\right]\right]. \tag{A.12}$$

The inner expectation in (A.12) is of the form

$$\mathbb{E}\left[\frac{\exp\left(\sqrt{\ln M}\sqrt{\nu_\ell}\, U_1^\ell\right)}{\exp\left(\sqrt{\ln M}\sqrt{\nu_\ell}\, U_1^\ell\right) + M^{-\nu_\ell}\sum_{j=2}^M \exp\left(\sqrt{\ln M}\sqrt{\nu_\ell}\, U_j^\ell\right)}\Big| U_1^\ell\right] = \mathbb{E}_X\left[\frac{c}{c+X}\right], \tag{A.13}$$

where $c = \exp\left(\sqrt{\ln M}\sqrt{\nu_\ell}\, U_1^\ell\right)$ is treated as a positive constant, and the expectation is with respect to the random variable

$$X := M^{-\nu_\ell}\sum_{j=2}^M \exp\left(\sqrt{\ln M}\sqrt{\nu_\ell}\, U_j^\ell\right). \tag{A.14}$$

1. We can also prove that $\lim \mathcal{E}_\ell = \frac{1}{2}$ if $c_\ell = 2(\ln 2) R\tau^2$, but we do not need this for the exponentially decaying power allocation since c_ℓ exactly equals $2(\ln 2) R\tau^2$ for only a vanishing fraction of sections. Since $\mathcal{E}_\ell \in [0,1]$, these sections do not affect the value of $\lim x(\tau)$ in (A.6).

Case 1: $\lim \nu_l > 2$. Since $\frac{c}{c+X}$ is a convex function of X, applying Jensen's inequality we get

$$\mathbb{E}_X\left[\frac{c}{c+X}\right] \geq \frac{c}{c+\mathbb{E}X}. \tag{A.15}$$

The expectation of X is

$$\mathbb{E}X = M^{-\nu_\ell}\sum_{j=2}^{M}\mathbb{E}\left[\exp\left(\sqrt{\ln M}\sqrt{\nu_\ell}\,U_j^\ell\right)\right] \overset{(a)}{=} M^{-\nu_\ell}(M-1)M^{\nu_\ell/2} \leq M^{1-\nu_\ell/2} \tag{A.16}$$

where (a) is obtained using the moment generating function of a Gaussian random variable. We therefore have

$$1 \geq \mathbb{E}_X\left[\frac{c}{c+X}\right] \geq \frac{c}{c+\mathbb{E}X} \geq \frac{c}{c+M^{1-\nu_\ell/2}}. \tag{A.17}$$

Recalling that $c = \exp\left(\sqrt{\ln M}\sqrt{\nu_\ell}\,U_1^\ell\right)$, (A.17) implies that

$$\mathbb{E}_X\left[\frac{\exp\left(\sqrt{\ln M}\sqrt{\nu_\ell}\,U_1^\ell\right)}{\exp\left(\sqrt{\ln M}\sqrt{\nu_\ell}\,U_1^\ell\right)+X}\;\middle|\; U_1^\ell\right] \geq \frac{1}{1+M^{1-\nu_\ell/2}\exp\left(-\sqrt{\ln M}\sqrt{\nu_\ell}\,U_1^\ell\right)}. \tag{A.18}$$

When $\{U_1^\ell > -(\ln M)^{1/4}\}$, the RHS of (A.18) is at least $[1+M^{1-\nu_\ell/2}\exp\left((\ln M)^{3/4}\sqrt{\nu_\ell}\right)]^{-1}$. Using this in (A.12), we obtain that

$$1 \geq \mathcal{E}_\ell \geq P(U_1^\ell > -(\ln M)^{1/4})\cdot\frac{1}{1+M^{1-\nu_\ell/2}\exp\left((\ln M)^{3/4}\sqrt{\nu_\ell}\right)} \overset{M\to\infty}{\longrightarrow} 1 \;\text{ since }\; \nu_\ell > 2.$$

Hence $\mathcal{E}_\ell \to 1$ when $\lim \nu_\ell > 2$.

Case 2: $\lim \nu_l < 2$**.** The random variable X in (A.14) can be bounded from below as follows.

$$X \geq M^{-\nu_\ell}\max_{j\in\{2,\ldots,M\}}\exp\left(\sqrt{\ln M}\sqrt{\nu_\ell}\,U_j^\ell\right) = M^{-\nu_\ell}\exp\left(\left[\max_{j\in\{2,\ldots,M\}}U_j^\ell\right]\sqrt{\ln M}\sqrt{\nu_\ell}\right). \tag{A.19}$$

Using standard bounds for the standard normal distribution, it can be shown that

$$P\left(\max_{j\in\{2,\ldots,M\}}U_j^\ell\; < \sqrt{2\ln M}(1-\epsilon)\right) \leq \exp(-M^{\epsilon(1-\epsilon)}), \tag{A.20}$$

for $\epsilon = \omega\left(\frac{\ln\ln M}{\ln M}\right)$. [2] Combining (A.20) and (A.19), we obtain that

$$\begin{aligned}\exp(-M^{\epsilon(1-\epsilon)}) &\geq P\left(\max_{j\in\{2,\ldots,M\}} U_j^\ell \ < \sqrt{2\ln M}(1-\epsilon)\right) \\ &\geq P\left(X < M^{-\nu_\ell}\exp\left(\sqrt{2\ln M}(1-\epsilon)\sqrt{\ln M}\sqrt{\nu_\ell}\right)\right) = P\left(X < M^{\sqrt{2\nu_\ell}(1-\epsilon)-\nu_\ell}\right).\end{aligned} \tag{A.21}$$

Since $\lim \nu_\ell < 2$ and $\epsilon > 0$ can be arbitrary small, there exists a strictly positive constant δ such that $\delta < \sqrt{2\nu_\ell}(1-\epsilon) - \nu_\ell$ for all sufficiently large L. Therefore, for sufficiently large M, the expectation in (A.13) can be bounded as

$$\begin{aligned}\mathbb{E}_X\left[\frac{c}{c+X}\right] &\leq P(X < M^\delta)\cdot 1 + P(X \geq M^\delta)\cdot\frac{c}{c+M^\delta} \\ &\leq \exp(-M^{\epsilon(1-\epsilon)}) + 1\cdot\frac{c}{c+M^\delta} \leq \frac{2}{1+c^{-1}M^\delta}.\end{aligned} \tag{A.22}$$

Recalling that $c = \exp\left(\sqrt{\ln M}\sqrt{\nu_\ell}\,U_1^\ell\right)$, and using the bound of (A.22) in (A.12), we obtain

$$\begin{aligned}\mathcal{E}_\ell &\leq \mathbb{E}\left[\frac{1}{1+M^\delta\exp\left(-\sqrt{\ln M}\sqrt{\nu_\ell}\,U_1^\ell\right)}\right] \\ &\leq P(U_1^\ell > (\ln M)^{1/4})\cdot 1 + P(U_1^\ell \leq (\ln M)^{1/4})\cdot\frac{1}{1+M^\delta\exp(-\sqrt{\nu_\ell}(\ln M)^{3/4})} \\ &\overset{(a)}{\leq} \exp(-\tfrac{1}{2}(\ln M)^{1/2}) + 1\cdot\frac{1}{1+\exp\left(\delta\ln M - \sqrt{\nu_\ell}(\ln M)^{3/4}\right)} \overset{(b)}{\longrightarrow} 0 \text{ as } M\to\infty.\end{aligned} \tag{A.23}$$

In (A.23), (a) is obtained using the bound $\Phi(x) < \exp(-x^2/2)$ for $x \geq 0$, where $\Phi(\cdot)$ is the Gaussian cdf; (b) holds since δ and $\lim\nu_\ell$ are both positive constants.

This proves that $\mathcal{E}_\ell \to 0$ when $\lim\nu_\ell < 2$. The proof of Lemma 13 is complete since we have proved both statements in (A.8). □

We remark that the $\ln 2$ term appears in Lemma 13 because R and $\mathcal{C}$ are measured in bits. For $t = 0$, $\tau_0^2 = \sigma^2 + P$. From Lemma 13, we have

$$\bar{x}_1 = \lim\sum_{\ell=1}^{L}\frac{P_\ell}{P}\,\mathbf{1}\{c_\ell > 2(\ln 2)R\,(\sigma^2+P)\} = \lim_{L\to\infty}\sum_{\ell=1}^{L}\frac{P_\ell}{P}\,\mathbf{1}\left\{\frac{\ell}{L} < \frac{\log(\mathcal{C}/R)}{2\mathcal{C}}\right\}, \tag{A.24}$$

2. Recall that $f(n) = \omega(g(n))$ if for each $k > 0$, $|f(n)|/|g(n)| \geq k$ for sufficiently large n,

where the second equality is obtained using the expression for c_ℓ in (A.9) and simplifying. Substituting $\frac{\log(\mathcal{C}/R)}{2\mathcal{C}} = \xi_0$, and using the geometric series formula

$$\sum_{\ell=1}^{k} P_\ell = (P+\sigma^2)(1-2^{-2\mathcal{C}k/L}), \tag{A.25}$$

(A.24) becomes

$$\bar{x}_1 = \lim_{L\to\infty} \sum_{\ell=1}^{\lfloor \xi_0 L \rfloor} \frac{P_\ell}{P} = \frac{P+\sigma^2}{P}(1-2^{-2\mathcal{C}\xi_0}) = \frac{(1+\mathsf{snr})-(1+\mathsf{snr})^{1-\xi_0}}{\mathsf{snr}}. \tag{A.26}$$

The expression for $\bar{\tau}_1^2$ is a straightforward simplification of $\sigma^2 + P(1-\bar{x}_1)$.

Assume towards induction that (2.22) and (2.23) hold for $\bar{x}_t, \bar{\tau}_t^2$. For step $(t+1)$, from Lemma 13 we have

$$\bar{x}_{t+1} = \lim \sum_{\ell=1}^{L} \frac{P_\ell}{P} \mathbf{1}\{c_\ell > 2(\ln 2) R\, \bar{\tau}_t^2\} = \lim_{L\to\infty} \sum_{\ell=1}^{L} \frac{P_\ell}{P} \mathbf{1}\left\{ \frac{\ell}{L} < \frac{1}{2\mathcal{C}} \log \frac{\mathcal{C}(P+\sigma^2)}{R\,\bar{\tau}_t^2} \right\}, \tag{A.27}$$

where the second equality is obtained using the expression for c_ℓ in (A.9) and simplifying. Using the induction hypothesis for $\bar{\tau}_t^2$, we get

$$\frac{(P+\sigma^2)}{\bar{\tau}_t^2} = \frac{(P+\sigma^2)}{\sigma^2\,(1+\mathsf{snr})^{1-\xi_{t-1}}} = (1+\mathsf{snr})^{\xi_{t-1}} = 2^{2\mathcal{C}\xi_{t-1}}. \tag{A.28}$$

Hence

$$\frac{1}{2\mathcal{C}} \log \frac{\mathcal{C}(P+\sigma^2)}{R\bar{\tau}_t^2} = \underbrace{\frac{1}{2\mathcal{C}} \log\left(\frac{\mathcal{C}}{R}\right) + \xi_{t-1}}_{\xi_t}. \tag{A.29}$$

Using (A.29) in (A.27), we obtain

$$\bar{x}_{t+1} = \lim_{L\to\infty} \sum_{\ell=1}^{\lfloor \xi_t L \rfloor} \frac{P_\ell}{P} = \frac{P+\sigma^2}{P}(1-2^{-2\mathcal{C}\xi_t}) = \frac{(1+\mathsf{snr})-(1+\mathsf{snr})^{1-\xi_t}}{\mathsf{snr}}. \tag{A.30}$$

The proof is concluded by using (A.30) to compute $\bar{\tau}_{t+1}^2 = P + \sigma^2(1-\bar{x}_{t+1})$.

A.3 Proof of Lemma 2

A.3.1 Useful Probability and Linear Algebra Results

We now list some results that will be used in the proof of Lemma 2. Most of these can be found in [6, Section III.G], but we summarize them here for completeness.

Fact 1. *Let $u \in \mathbb{R}^N$ and $v \in \mathbb{R}^n$ be deterministic vectors such that $\lim_{n\to\infty}\|u\|^2/n$ and $\lim_{n\to\infty}\|v\|^2/n$ both exist and are finite. Let $\tilde{X} \in \mathbb{R}^{n\times N}$ be a matrix with independent $\mathcal{N}(0,1/n)$ entries. Then:*

(a)

$$\tilde{X}u \stackrel{d}{=} \frac{\|u\|}{\sqrt{n}} Z_u \quad \text{and} \quad \tilde{X}^* v \stackrel{d}{=} \frac{\|v\|}{\sqrt{n}} Z_v, \tag{A.31}$$

where $Z_u \in \mathbb{R}^n$ and $Z_v \in \mathbb{R}^N$ are Gaussian random vectors distributed as $\mathcal{N}(0, \mathsf{I}_{n\times n})$ and $\mathcal{N}(0, \mathsf{I}_{N\times N})$, respectively. Consequently,

$$\lim_{n\to\infty} \frac{\|\tilde{X}u\|^2}{n} \stackrel{a.s.}{=} \lim_{n\to\infty} \frac{\|u\|^2}{n} \sum_{i=1}^{n} \frac{Z_{u,i}^2}{n} \stackrel{a.s.}{=} \lim_{n\to\infty} \frac{\|u\|^2}{n} \tag{A.32}$$

$$\lim_{n\to\infty} \frac{\|\tilde{X}^*v\|^2}{N} \stackrel{a.s.}{=} \lim_{n\to\infty} \frac{\|v\|^2}{n} \sum_{j=1}^{N} \frac{Z_{v,j}^2}{N} \stackrel{a.s.}{=} \lim_{n\to\infty} \frac{\|v\|^2}{n} \tag{A.33}$$

(b) Let $\mathcal{W}$ be a d-dimensional subspace of $\mathbb{R}^n$ for $d \le n$. Let $(w_1, ..., w_d)$ be an orthogonal basis of $\mathcal{W}$ with $\|w_i\|^2 = n$ for $i \in [d]$, and let $\mathsf{P}_{\mathcal{W}}$ denote the orthogonal projection operator onto $\mathcal{W}$. Then for $D = [w_1 \mid \ldots \mid w_d]$, we have $\mathsf{P}_{\mathcal{W}}\tilde{X}u \stackrel{d}{=} \frac{\|u\|}{\sqrt{n}} Dx$ where $x \in \mathbb{R}^d$ is a random vector with i.i.d.

$\mathcal{N}(0,1/n)$ entries. Therefore $\lim_{n\to\infty} n^\delta \|x\| \stackrel{a.s.}{=} 0$ for any constant $\delta \in [0, 0.5)$. (The limit is taken with d fixed.)

Fact 2 (Strong Law for Triangular Arrays)**.** *Let $\{X_{n,i} : i \in [n], n \ge 1\}$ be a triangular array of random variables such that for each n $(X_{n,1}, \ldots, X_{n,n})$ are mutually independent, have zero mean, and satisfy*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}|X_{n,i}|^{2+\kappa} \le c n^{\kappa/2} \quad \text{for some } \kappa \in (0,1) \text{ and } c < \infty. \tag{A.34}$$

Then $\frac{1}{n}\sum_{i=1}^{n} X_{n,i} \to 0$ *almost surely as* $n \to \infty$.

Fact 3. *Let* $v \in \mathbb{R}^n$ *be a random vector with i.i.d.*

entries $\sim p_V$ *where the measure* p_V *has bounded second moment. Then for any function* ψ *that is pseudo-Lipschitz of order two:*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \psi(v_i) \overset{a.s.}{=} \mathbb{E}_{p_V}[\psi(V)] \tag{A.35}$$

with convergence rate $n^{-\delta}$*, for some* $\delta \in (0, 1/4)$.

Fact 4. *Let* $Z_1, \ldots, Z_t$ *be jointly Gaussian random variables with zero mean and an invertible covariance matrix* C*. Then*

$$Var(Z_t \mid Z_1, \ldots, Z_{t-1}) = \mathbb{E}[Z_t^2] - u^* C^{-1} u,$$

where for $i \in [t-1]$, $u_i = \mathbb{E}[Z_t Z_i]$.

Fact 5. *Let* $Z_1, \ldots, Z_t$ *be jointly Gaussian random variables such that for all* $i \in [t]$,

$$\mathbb{E}[Z_i^2] \le K \quad \text{and} \quad Var(Z_i \mid Z_1, \ldots, Z_{i-1}) \ge c_i,$$

for some strictly positive constants $K, c_1, \ldots, c_t$*. Let* Y *be a random variable defined on the same probability space, and let* $g : \mathbb{R}^2 \to \mathbb{R}$ *be a Lipschitz function with* $z \to g(z, Y)$ *non-constant with positive probability. Then there exists a positive constant* c_t' *such that*

$$\mathbb{E}[(g(Z_t, Y))^2] - u^* C^{-1} u > c_t',$$

where $u \in \mathbb{R}^{t-1}$ *and* $C \in \mathbb{R}^{(t-1)\times(t-1)}$ *are given by*

$$u_i = \mathbb{E}[g(Z_t, Y) g(Z_i, Y)],\ C_{ij} = \mathbb{E}[g(Z_i, Y) g(Z_j, Y)],\ i, j \in [t-1].$$

(The constant c_t' *depends only on the* K*, the random variable* Y *and the function* g*.)*

Fact 6 (Stein's lemma)**.** *For zero-mean jointly Gaussian random variables* Z_1, Z_2*, and any*

function $f : \mathbb{R} \to \mathbb{R}$ *for which* $\mathbb{E}[Z_1 f(Z_2)]$ *and* $\mathbb{E}[f'(Z_2)]$ *both exist, we have* $\mathbb{E}[Z_1 f(Z_2)] = \mathbb{E}[Z_1 Z_2]\mathbb{E}[f'(Z_2)]$.

Fact 7. *Let* $v_1, \ldots, v_t$ *be a sequence of vectors in* $\mathbb{R}^n$ *such that for* $i \in [t]$

$$\frac{1}{n}\|v_i - \mathsf{P}_{i-1}(v_i)\|^2 \geq c,$$

where c *is a positive constant and* P_{i-1} *is the orthogonal projection onto the span of* $v_1, \ldots, v_{i-1}$. *Then the matrix* $C \in \mathbb{R}^{t \times t}$ *with* $C_{ij} = v_i^* v_j / n$ *has minimum eigenvalue* $\lambda_{\min} \geq c'$, *where* c' *is a strictly positive constant (depending only on* c *and* t*).*

Fact 8. *Let* $\{S_n\}_{n \geq 1}$ *be a sequence of* $t \times t$ *matrices such that* $\lim_{n\to\infty} S_n = S_\infty$ *where the limit is element-wise. Then if* $\liminf_{n\to\infty} \lambda_{\min}(S_n) \geq c$ *for a positive constant* c, *then* $\lambda_{\min}(S_\infty) \geq c$.

A.3.2 Inductive Proof

A key ingredient in the proof is the distribution of X conditioned on the sigma algebra $\mathscr{S}_{t_1,t}$ where t_1 is either $t+1$ or t. We then have

$$b^t + \lambda_t m^{t-1} = X q^t, \tag{A.36}$$

which follows from (2.9) and (2.30). We also have

$$h^{t+1} + q^t = X^* m^t. \tag{A.37}$$

From (A.36) and (A.37), we have the matrix equations

$$A_t = X^* M_t, \quad Y_t = X Q_t, \tag{A.38}$$

where M_t and Q_t are defined in (2.32) and

$$A_t = [h^1 + q^0 \mid h^2 + q^1 \mid \ldots \mid h^t + q^{t-1}], \qquad Y_t = [b^0 \mid b^1 + \lambda_1 m^0 \mid \ldots \mid b^{t-1} + \lambda_{t-1} m^{t-2}], \tag{A.39}$$

The notation $[c_1 \mid c_2 \mid \ldots \mid c_k]$ is used to denote a matrix with columns $c_1, \ldots, c_k$.

Observing that conditioning on $\mathscr{S}_{t_1,t}$ is equivalent to conditioning on the linear constraints

$$XQ_{t_1} = Y_{t_1},\ X^*M_t = A_t,$$

the following lemma from [6] specifies the conditional distribution $X|_{\mathscr{S}_{t_1,t}}$. [3]

Lemma 14. *[6, Lemma 10] For $t_1 = t+1$ or t, the conditional distribution of the random matrix X given $\mathscr{S}_{t_1,t}$ satisfies*

$$X|_{\mathscr{S}_{t_1,t}} \stackrel{d}{=} \mathbb{E}_{t_1,t} + \mathsf{P}^{\perp}_{M_t} \tilde{X} \mathsf{P}^{\perp}_{Q_{t_1}}.$$

Here $\tilde{X} \stackrel{d}{=} X$ is random matrix independent of $\mathscr{S}_{t_1,t}$, and $\mathsf{P}^{\perp}_{M_t} = \mathsf{I} - \mathsf{P}_{M_t}$ where $\mathsf{P}_{M_t} = M_t(M_t^*M_t)^{-1}M_t^*$ is the orthogonal projection matrix onto the column space of M_t; similarly, $\mathsf{P}^{\perp}_{Q_{t_1}} = \mathsf{I} - \mathsf{P}_{Q_{t_1}}$, where $\mathsf{P}_{Q_{t_1}} = Q_{t_1}(Q_{t_1}^*Q_{t_1})^{-1}Q_{t_1}^*$. The matrix $\mathbb{E}_{t_1,t} = \mathbb{E}[X|\mathscr{S}_{t_1,t}]$ is given by

$$\begin{aligned} \mathbb{E}_{t_1,t} &= \mathbb{E}[X\mathsf{P}_{Q_{t_1}} + \mathsf{P}_{M_t} X \mathsf{P}^{\perp}_{Q_{t_1}} \mid XQ_{t_1} = Y_{t_1},\ X^*M_t = A_t] \\ &= Y_{t_1}(Q_{t_1}^*Q_{t_1})^{-1}Q_{t_1}^* + M_t(M_t^*M_t)^{-1}A_t^* - M_t(M_t^*M_t)^{-1}M_t^*Y_{t_1}(Q_{t_1}^*Q_{t_1})^{-1}Q_{t_1}^*. \end{aligned} \tag{A.40}$$

3. While conditioning on the linear constraints, we emphasize that only X is treated as random.

Lemma 15. *[6, Lemma 12] For the matrix $\mathbb{E}_{t_1,t}$ defined in Lemma 14, the following hold:*

$$\mathbb{E}^*_{t+1,t} m^t = A_t (M_t^* M_t)^{-1} M_t^* m_\parallel^t \; + \; Q_{t+1} (Q_{t+1}^* Q_{t+1})^{-1} Y_{t+1}^* m_\perp^t, \tag{A.41}$$

$$\mathbb{E}_{t,t} q^t = Y_t (Q_t^* Q_t)^{-1} Q_t^* q_\parallel^t \; + \; M_t (M_t^* M_t)^{-1} A_t^* q_\perp^t, \tag{A.42}$$

where $m_\parallel^t, m_\perp^t, q_\parallel^t, q_\perp^t$ are defined in (3.15) *and* (3.16).

We mention that Lemmas 14 and 15 can be applied only when $M_t^* M_t$ and $Q_{t_1}^* Q_{t_1}$ are invertible.

We are now ready to prove Lemma 2. The proof proceeds by induction on t. We label as $\mathcal{H}^{t+1}$ the results (B.9), (2.41), (B.13), (2.46), (B.19), (B.21) and similarly as $\mathcal{B}^t$ the results (B.10), (B.12), (B.14), (2.47), (B.22). The proof consists of four steps:

1. $\mathcal{B}_0$ holds.
2. $\mathcal{H}_1$ holds.
3. If $\mathcal{B}_r, \mathcal{H}_s$ holds for all $r < t$ and $s \leq t$, then $\mathcal{B}_t$ holds.
4. if $\mathcal{B}_r, \mathcal{H}_s$ holds for all $r \leq t$ and $s \leq t$, then $\mathcal{H}_{t+1}$ holds.

Step 1: Showing $\mathcal{B}_0$ holds

We wish to show that (B.10), (B.12), (B.14), (2.47), and (B.22) hold when $t = 0$.

(a) The sigma-algebra $\mathscr{S}_{0,0}$ is generated by $q^0 = -\beta_0$ and w. Both M_0 and Q_0 are empty matrices, and therefore $\tilde{M}_0$ is an empty matrix and $q_\perp^0 = q^0$. The result follows by noting that $b^0 = -X\beta_0 = Xq_0$, from the definitions in (2.30).

(b) We will first use Fact 2 to show that

$$\lim n^\delta \left[\frac{1}{n} \sum_{i=1}^n \phi_b(b_i^0, \epsilon_i) - \frac{1}{n} \sum_{i=1}^n \mathbb{E}_X \left\{ \phi_b(b_i^0, \epsilon_i) \right\} \right] \overset{a.s.}{=} 0. \tag{A.43}$$

To apply Fact 2, we need to verify that

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E} | n^\delta \phi_b(b_i^0, \epsilon_i) - n^\delta \mathbb{E}_X \left\{ \phi_b(b_i^0, \epsilon_i) \right\} |^{2+\kappa} \leq \; c n^{\kappa/2}. \tag{A.44}$$

for some $0 \leq \kappa \leq 1$ and c some constant. Using $b^0 = Xq^0$,

$$\begin{aligned}
&\mathbb{E}|\phi_b(b_i^0, \epsilon_i) - \mathbb{E}_X\left\{\phi_b(b_i^0, \epsilon_i)\right\}|^{2+\kappa} = \mathbb{E}_{\tilde{X}}|\phi_b([\tilde{X}q^0]_i, \epsilon_i) - \mathbb{E}_X\left\{\phi_b([Xq^0]_i, \epsilon_i)\right\}|^{2+\kappa} \\
&\overset{(a)}{\leq} \mathbb{E}_{\tilde{X},X}\left|\phi_b([\tilde{X}q^0]_i, \epsilon_i) - \phi_b([Xq^0]_i, \epsilon_i)\right|^{2+\kappa} \\
&\overset{(b)}{\leq} c'\,\mathbb{E}_{\tilde{X},X}\left\{|[\tilde{X}q^0]_i - [Xq^0]_i|^{2+\kappa}\left(1 + |[\tilde{X}q^0]_i| + |\epsilon_i| + |[Xq^0]_i|\right)^{2+\kappa}\right\} \\
&\leq c_0\left[\mathbb{E}_{\tilde{X},X}\left\{|[\tilde{X}q^0]_i - [Xq^0]_i|^{2+\kappa}\left(1 + |[\tilde{X}q^0]_i|^{2+\kappa} + |[Xq^0]_i|^{2+\kappa}\right)\right\} + |\epsilon_i|^{2+\kappa}\mathbb{E}_{\tilde{X},X}\left\{|[\tilde{X}q^0]_i - [Xq^0]_i|^{2+\kappa}\right\}\right] \\
&\overset{(c)}{\leq} c_1 + c_2|\epsilon_i|^{2+\kappa},
\end{aligned} \tag{A.45}$$

where c', c_0, c_1, c_2 are positive constants. In the chain above, (a) uses Jensen's inequality, (b) holds because $\phi_b \in PL(2)$, and (c) is obtained using the fact that $[Xq_0]_i = -[X\beta_0]_i \overset{d}{=} \sqrt{P}Z$, and $[\tilde{X}q_0]_i \overset{d}{=} \sqrt{P}\tilde{Z}$, where $Z, \tilde{Z}$ are i.i.d.

$\mathcal{N}(0,1)$. Using (A.45) in (A.44), we obtain

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}|n^{\delta}\phi_b(b_i^0, \epsilon_i) - n^{\delta}\mathbb{E}_X\left\{\phi_b(b_i^0, \epsilon_i)\right\}|^{2+\kappa} \leq \frac{n^{\delta(\kappa+2)}}{n}\sum_{i=1}^{n}(c_1 + c_2|\epsilon_i|^{2+\kappa}) \leq cn^{\kappa/2}, \tag{A.46}$$

for $\delta < \frac{\kappa/2}{\kappa+2}$ since the ϵ_i's are i.i.d.

$\mathcal{N}(0, \sigma^2)$. Thus (A.43) holds.

Since $b^0 = Xq^0 \overset{d}{=} \sqrt{P}Z$, where $Z \in \mathbb{R}^n$ is i.i.d.

$\sim \mathcal{N}(0,1)$, we have

$$\mathbb{E}_X\left\{\phi_b(b_i^0, \epsilon_i)\right\} = \mathbb{E}_X\left\{\phi_b([Xq^0]_i, \epsilon_i)\right\} = \mathbb{E}_{Z_0}\left\{\phi_b(\bar{\sigma}_0 Z_0, \epsilon_i)\right\}, \tag{A.47}$$

where $\bar{\sigma}_0^2 = P$ and $Z_0 \sim \mathcal{N}(0,1)$. Thus

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_X[\phi_b(b_i^0, \epsilon_i)] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{Z_0}[\phi_b(\bar{\sigma}_0 Z_0, \epsilon_i)] \overset{n\to\infty}{\longrightarrow} \mathbb{E}[\phi_b(\bar{\sigma}_0 Z_0, \sigma Z_\epsilon)]\ a.s., \tag{A.48}$$

due to Fact 3, which also guarantees that the convergence rate in (A.48) is $o(n^{-\delta})$. Combining (A.43) and (A.48) yields the result.

(c) Using the definition $b^0 = Xq^0$ and conditioning on $q^0 = -\beta_0$, we have using Fact

9(a):

$$\frac{\|b^0\|^2}{n} = \frac{\|Xq^0\|^2}{n} \stackrel{d}{=} \frac{\|q^0\|^2}{n} \sum_{i=1}^{n} \frac{Z_i^2}{n}, \tag{A.49}$$

where $Z_1, \ldots, Z_n$ are i.i.d.

$\mathcal{N}(0,1)$. Taking the limit of (A.49) gives the desired result since $\|q^0\|^2/n = P$ and by the central limit theorem, $\frac{1}{n}\sum_{i=1}^{n} Z_i^2 - 1$ is $o(n^{-\delta})$ almost surely for any $\delta \in (0, 1/2)$.

(d) Since $m^0 = b^0 - w$, $(b^0)^* m^0 = \|b^0\|^2 - (b^0)^* w$. By Step 1(c) above, $\frac{\|b^0\|^2}{n} \to P$ almost surely at rate $n^{-\delta}$. Using using Fact 9(a), we have

$$(b^0)^*\epsilon = (Xq^0)^*\epsilon = \frac{(q^0)^* X^* \epsilon}{n} \stackrel{d}{=} \frac{\|q^0\|}{\sqrt{n}} \frac{\|\epsilon\|}{\sqrt{n}} \frac{Z}{\sqrt{n}} = \sqrt{P} \frac{\|\epsilon\|}{\sqrt{n}} \frac{Z}{\sqrt{n}} \tag{A.50}$$

where the random variable $Z \sim \mathcal{N}(0,1)$ is independent of ϵ. The result follows by noting that $(\frac{\|\epsilon\|}{\sqrt{n}} - \sigma)$ is $o(n^{-\delta})$ almost surely.

(f) Since M_0 is the empty matrix, $m_\perp^0 = m_0 = (b^0 - \epsilon)$. Applying $\mathcal{B}_0$(b) to the function $\phi_b(b_i^0, \epsilon_i) = (b_i^0 - \epsilon_i)^2$, we obtain

$$\lim \frac{\|m^0\|^2}{n} = \lim \frac{1}{n} \sum_{i=1}^{n} (b_i^0 - \epsilon_i)^2 = \lim \frac{1}{n} \sum_{i=1}^{n} \phi_b(b_i^0, \epsilon_i) \stackrel{a.s.}{=} \mathbb{E}\left\{(\bar{\sigma}_0 Z_0 - \sigma Z_\epsilon)^2\right\} = \sigma^2 + \bar{\sigma}_0^2. \tag{A.51}$$

Step 2: Showing $\mathcal{H}_1$ holds

(a) The conditioning sigma-algebra $\mathscr{S}_{1,0}$ is generated by $b^0, m^0, q^0 = -\beta_0$ and ϵ. From Lemmas 14 and 15, we have

$$X|_{\mathscr{S}_{1,0}} \stackrel{d}{=} Y_1 (Q_1^* Q_1)^{-1} Q_1^* + \tilde{X} P_{Q_1}^{\perp} = \frac{b^0 q^{0^*}}{\|q^0\|^2} + \tilde{X} P_{q^0}^{\perp} \tag{A.52}$$

as M_0 and Q_0 are empty matrices, and $Q_1 = q^0$. Since $h^1 = X^* m^0 - q^0$, (A.52) implies

$$h^1|_{\mathscr{S}_{1,0}} \stackrel{d}{=} \frac{q^0 b^{0^*} m^0}{\|q^0\|^2} + P_{q^0}^{\perp} \tilde{X}^* m^0 - q^0. \tag{A.53}$$

First note that

$$\frac{q^0 b^{0^*} m^0}{\|q^0\|^2} - q^0 = \frac{q^0}{P}\left(\frac{b^{0^*} m^0}{n} - P\right) \stackrel{a.s.}{=} \frac{q^0}{P} o(n^{-\delta}), \tag{A.54}$$

where the last equality follows from $\mathcal{B}_0$(d). Substituting (A.54) in (A.53), we see that the result follows if we prove that

$$P_{q^0}^{\perp}\tilde{X}^* m^0 \stackrel{d}{=} \tilde{X}^* m_{\perp}^0 + \frac{q^0}{\sqrt{P}} o(n^{-\delta}). \tag{A.55}$$

To show (A.55), we observe that $P_{q^0}^{\perp}\tilde{X}^* m^0 = \tilde{X}^* m^0 - P_{q^0}^{\parallel}\tilde{X}^* m^0$. Further, since M^0 is an empty matrix $\tilde{X}^* m^0 = \tilde{X}^* m_{\perp}^0$. Thus, all that is left to show is that $P_{q^0}^{\parallel}\tilde{X}^* m^0 = \frac{q^0}{\sqrt{P}} o(n^{-\delta})$ almost surely. Since q^0, m^0 are in the conditioning sigma-algebra and are independent of $\tilde{X}$, we obtain using Fact 9(a),

$$P_{q^0}^{\parallel}\tilde{X}^* m^0 = \frac{q^0 q^{0^*}}{\|q^0\|^2}\tilde{X}^* m^0 = \frac{q^0\|m^0\|}{\|q^0\|}\left(\frac{q^{0^*}}{\|q^0\|}\tilde{X}^*\frac{m^0}{\|m^0\|}\right) \stackrel{d}{=} \frac{q^0}{\sqrt{P}}\left(\frac{\|m^0\|}{\sqrt{n}}\frac{Z}{\sqrt{n}}\right), \tag{A.56}$$

where Z is a standard normal random variable. It was shown in (A.51) that $\frac{\|m^0\|^2}{n} \stackrel{a.s.}{\to} \sigma^2 + \bar{\sigma}_0^2 = \bar{\tau}_0^2$, which implies that that $\frac{Z}{\sqrt{n}}\frac{\|m^0\|}{\sqrt{n}} = o(n^{-\delta})$ almost surely.

(c) From $\mathcal{H}_1$(a) shown above, $h^1\big|_{\mathscr{S}_{1,0}} \stackrel{d}{=} \tilde{X}^* m^0 + \frac{q^0}{\sqrt{P}} o(n^{-\delta})$, and so

$$\frac{\|h^1\|^2}{N}\Big|_{\mathscr{S}_{1,0}} \stackrel{d}{=} \frac{\|\tilde{X}^* m^0\|^2}{N} + \frac{\|q^0\|^2}{NP}\vec{o}_1(n^{-2\delta}) - 2\frac{(q^0)^*\tilde{X}^* m^0}{N\sqrt{P}} o(n^{-\delta}). \tag{A.57}$$

The last two terms in (A.57) are $o(n^{-\delta})$. Indeed, $\|q^0\|^2 = nP$, $\frac{n}{N} = \Theta(\frac{\log M}{M})$, and by Fact 9(a),

$$\frac{(q^0)^*\tilde{X}^* m^0}{N\sqrt{P}} \stackrel{d}{=} \frac{\|m^0\|}{\sqrt{n}}\frac{\|q^0\|}{\sqrt{NP}}\frac{Z}{\sqrt{N}} \quad \text{where } Z \sim \mathcal{N}(0,1). \tag{A.58}$$

It was shown in (A.51) that $\frac{\|m^0\|}{\sqrt{n}} \stackrel{a.s.}{\to} \bar{\tau}_0$, hence the term in (A.58) is $o(n^{-\delta})$.

Applying Fact 9(a) to the first term in (A.57), we obtain

$$\lim \frac{\|\tilde{X}^* m^0\|^2}{N} \stackrel{a.s.}{=} \lim \frac{\|m_0\|^2}{n}\frac{\|Z\|^2}{N} \stackrel{a.s.}{=} \bar{\tau}_0^2 \cdot 1 \tag{A.59}$$

where $Z \in \mathbb{R}^N$ is i.i.d. $\mathcal{N}(0,1)$. By $\mathcal{B}_0$(b) and the central limit theorem, the convergence rate in (A.59) is $n^{-\delta}$.

(b) The proof of this part involves several claims which are fairly straightforward but

tedious to verify, so we only give the main steps, referring the reader to [35] for the details. From $\mathcal{H}_1$(a),

$$h^1|_{\mathscr{S}_{1,0}} \stackrel{d}{=} \tilde{X}^* m^0 + \tilde{Q}_1 \vec{o}_1(n^{-\delta}), \tag{A.60}$$

where $\tilde{X}$ is an independent copy of X and $\tilde{Q}_1 = \frac{q^0}{\sqrt{P}}$. Then

$$\phi_h(a_0 h^1_\ell,\, b_0 h^1_\ell,\, \beta_{0_\ell})|_{\mathscr{S}_{1,0}} \stackrel{d}{=} \phi_h\left(a_0[\tilde{X}^* m^0]_\ell + a_0[\tilde{Q}_1\vec{o}_1(n^{-\delta})]_\ell,\, b_0[\tilde{X}^* m^0]_\ell + b_0[\tilde{Q}_1\vec{o}_1(n^{-\delta})]_\ell,\, \beta_{0_\ell}\right).$$

First, we show that the error term $\tilde{Q}_1\vec{o}_1(n^{-\delta'})$ can be dropped. For each section $\ell \in [L]$, let $h_\ell = a_0[\tilde{X}^* m^0]_\ell$ and $\Delta_\ell = a_0[\tilde{Q}_1\vec{o}_1(n^{-\delta'})]_\ell$. Similarly define $\tilde{h}_\ell$ and $\tilde{\Delta}_\ell$, with a_0 replaced by b_0. Then it is shown in [35] that for each of the functions in (2.40), we almost surely have

$$\frac{1}{L}\sum_{\ell=1}^{L}\left|\phi_h(h_\ell + \Delta_\ell, \tilde{h}_\ell + \tilde{\Delta}_\ell, \beta_{0_\ell}) - \phi_h(h_\ell, \tilde{h}_\ell, \beta_{0_\ell})\right| = o(n^{-\delta'}\log M). \tag{A.61}$$

for some $\delta' > 0$. Choosing $\delta \in (0, \delta')$ ensures that we can drop the $\tilde{Q}_{t+1}\vec{o}_{t+1}(n^{-\delta})$ terms.

In what follows, we use the notation $h_\ell[\tilde{X}] = a_0[\tilde{X}^* m^0]_\ell$ and $\tilde{h}_\ell[\tilde{X}] = b_0[\tilde{X}^* m^0]_\ell$, making explicit the dependence on $\tilde{X}$. We will appeal to Fact 2 to show that

$$\lim n^\delta\left[\frac{1}{L}\sum_{\ell=1}^{L}\phi_h\left(h_\ell[\tilde{X}], \tilde{h}_\ell[\tilde{X}], \beta_{0_\ell}\right) - \frac{1}{L}\sum_{\ell=1}^{L}\mathbb{E}_{\tilde{X}}\left\{\phi_h\left(h_\ell[\tilde{X}], \tilde{h}_\ell[\tilde{X}], \beta_{0_\ell}\right)\right\}\right] \stackrel{a.s.}{=} 0 \tag{A.62}$$

To invoke Fact 2 (conditionally on $\mathscr{S}_{1,0}$), we need to verify that

$$\frac{1}{L}\sum_{\ell=1}^{L}\mathbb{E}_{\hat{X}}\left|n^\delta\phi_h\left(h_\ell[\hat{X}], \tilde{h}_\ell[\hat{X}], \beta_{0_\ell}\right) - n^\delta\mathbb{E}_{\tilde{X}}\left\{\phi_h\left(h_\ell[\tilde{X}], \tilde{h}_\ell[\tilde{X}], \beta_{0_\ell}\right)\right\}\right|^{2+\kappa} \le cL^{\kappa/2} \tag{A.63}$$

for some $0 \le \kappa \le 1$ and some constant c. In (A.63), $\hat{X}, \tilde{X}$ are i.i.d. copies of X. From Jensen's inequality, we have

$$\begin{aligned}&\mathbb{E}_{\hat{X}}\left|\phi_h\left(h_\ell[\hat{X}], \tilde{h}_\ell[\hat{X}], \beta_{0_\ell}\right) - \mathbb{E}_{\tilde{X}}\left\{\phi_h\left(h_\ell[\tilde{X}], \tilde{h}_\ell[\tilde{X}], \beta_{0_\ell}\right)\right\}\right|^{2+\kappa}\\ &\le \mathbb{E}_{\hat{X},\tilde{X}}\left|\phi_h\left(h_\ell[\hat{X}], \tilde{h}_\ell[\hat{X}], \beta_{0_\ell}\right) - \phi_h\left(h_\ell[\tilde{X}], \tilde{h}_\ell[\tilde{X}], \beta_{0_\ell}\right)\right|^{2+\kappa},\end{aligned}$$

and in [35], it is shown that for each function in (2.40),

$$\mathbb{E}_{\hat{X},\tilde{X}}\left|\phi_h\left(h_\ell[\hat{X}],\tilde{h}_\ell[\hat{X}],\beta_{0_\ell}\right)-\phi_h\left(h_\ell[\tilde{X}],\tilde{h}_\ell[\tilde{X}],\beta_{0_\ell}\right)\right|^{2+\kappa}\overset{a.s.}{=}O((\log M)^{2+\kappa}),\quad \ell\in[L]. \tag{A.64}$$

The bound in (A.64) implies (A.63) holds if $\delta(2+\kappa)$ is chosen to be smaller than $\kappa/2$. (Recall that $L=\Theta(n/\log n)$). We have thus shown (A.62).

Recall that for each $\ell\in[L]$, we have $[\tilde{X}^*m^0]_\ell\overset{d}{=}(\|m^0\|/\sqrt{n})Z_{0_\ell}$ where $Z_{0_\ell}\sim\mathcal{N}(0,\mathsf{I}_{M\times M})$. Therefore, in (A.62), $h_\ell[\tilde{X}]\overset{d}{=}a_0\frac{\|m^0\|}{\sqrt{n}}Z_{0_\ell}$, and $\tilde{h}_\ell[\tilde{X}]\overset{d}{=}b_0\frac{\|m^0\|}{\sqrt{n}}Z_{0_\ell}$. We will next show that

$$\lim n^\delta\left[\frac{1}{L}\sum_{\ell=1}^{L}\mathbb{E}_{Z_0}\left|\ \phi_h\left(a_0\frac{\|m^0\|}{\sqrt{n}}Z_{0_\ell},b_0\frac{\|m^0\|}{\sqrt{n}}Z_{0_\ell},\beta_{0_\ell}\right)-\phi_h\left(a_0\bar{\tau}_0Z_{0_\ell},b_0\bar{\tau}_0Z_{0_\ell},\beta_{0_\ell}\right)\right|\right]\overset{a.s.}{=}0. \tag{A.65}$$

Let us redefine $h_\ell=a_0\frac{\|m^0\|}{\sqrt{n}}Z_{0_\ell}$ and $\Delta_\ell=a_0\left(\bar{\tau}_0-\frac{\|m^0\|}{\sqrt{n}}\right)Z_{0_\ell}$. Define $\tilde{h}_\ell$ and $\tilde{\Delta}_\ell$ similarly with b_0 replacing a_0. Then (A.65) can be written as

$$\lim n^\delta\left[\frac{1}{L}\sum_{\ell=1}^{L}\mathbb{E}_{Z_0}\left|\ \phi_h\left(h_\ell,\tilde{h}_\ell,\beta_{0_\ell}\right)-\phi_h\left(h_\ell+\Delta_\ell,\tilde{h}_\ell+\tilde{\Delta}_\ell,\beta_{0_\ell}\right)\right|\right]\overset{a.s.}{=}0. \tag{A.66}$$

Note that from $\mathcal{H}_1$(c) and the fact that $Z_{0_\ell}\sim\mathcal{N}(0,\mathsf{I}_{M\times M})$,

$$\begin{aligned}\max_{j\in sec(\ell)}|h_{\ell_j}|&=|a_0|\frac{\|m^0\|}{\sqrt{n}}\max_{j\in sec(\ell)}|Z_{0_{\ell_j}}|\overset{a.s.}{=}\Theta(\sqrt{\log M}),\\ \max_{j\in sec(\ell)}|\Delta_{\ell_j}|&=|a_0|\left|\bar{\tau}_0-\frac{\|m^0\|}{\sqrt{n}}\right|\max_{j\in sec(\ell)}|Z_{0_{\ell_j}}|\overset{a.s.}{=}\Theta(n^{-\delta'}\sqrt{\log M})\end{aligned} \tag{A.67}$$

for some $\delta'>0$. The almost-sure equality in each line of (A.67) holds for sufficiently large M. (This can be shown using the standard normal distribution of Z_0 and the Borel-Cantelli lemma). Similarly $\max_{j\in sec(\ell)}|\tilde{h}_{\ell_j}|=\Theta(\sqrt{\log M})$ and $\max_{j\in sec(\ell)}|\tilde{\Delta}_{\ell_j}|=\Theta(n^{-\delta'}\sqrt{\log M})$. Using (A.67), it is shown in [35] that

$$\left|\ \phi_h\left(h_\ell,\tilde{h}_\ell,\beta_{0_\ell}\right)-\phi_h\left(h_\ell+\Delta_\ell,\tilde{h}_\ell+\tilde{\Delta}_\ell,\beta_{0_\ell}\right)\right|\overset{a.s.}{=}o(n^{-\delta'}\log M) \tag{A.68}$$

for some $\delta' > 0$. Thus (A.65) holds for $\delta < \delta'$. To complete the proof, we need to show that

$$\lim n^{\delta} \left[\frac{1}{L} \sum_{\ell=1}^{L} \mathbb{E}_{Z_0} \left[\phi_h \left(a_0 \bar{\tau}_0 Z_{0_\ell}, b_0 \bar{\tau}_0 Z_{0_\ell}, \beta_{0_\ell} \right) \right] - \frac{1}{L} \sum_{\ell=1}^{L} \mathbb{E}_{(Z_0,\beta)} \left[\phi_h \left(a_0 \bar{\tau}_0 Z_{0_\ell}, b_0 \bar{\tau}_0 Z_{0_\ell}, \beta_{\ell} \right) \right] \right] \overset{a.s.}{=} 0 \tag{A.69}$$

But (A.69) holds because the uniform distribution of the non-zero entry in β_ℓ over the M possible locations and the i.i.d. distribution of Z_0 together ensure that for all $\beta_0 \in \mathcal{B}_{M,L}$, we have

$$\mathbb{E}_{Z_0} \left[\phi_h \left(a_0 \bar{\tau}_0 Z_{0_\ell}, b_0 \bar{\tau}_0 Z_{0_\ell}, \beta_{0_\ell} \right) \right] = \mathbb{E}_{(Z_0,\beta)} \left[\phi_h \left(a_0 \bar{\tau}_0 Z_{0_\ell}, b_0 \bar{\tau}_0 Z_{0_\ell}, \beta_{\ell} \right) \right], \quad \forall \ell \in [L].$$

(d) By definition $q^1 = \eta^0(\beta_0 - h^1) - \beta_0$, and hence $\frac{(h^1)^* q^1}{n} = \frac{1}{n} \sum_{\ell=1}^{L} \phi_h(h_\ell^1, \beta_{0_\ell})$, where the function $\phi_h : \mathbb{R}^M \times \mathbb{R}^M \to \mathbb{R}$ is $\phi_h(h_\ell^1, \beta_{0_\ell}) := (h_\ell^1)^* [\eta_\ell^0(\beta_0 - h^1) - \beta_{0_\ell}]$. Applying $\mathcal{H}_1$(b) to ϕ_h yields

$$\lim n^{\delta} \left[\frac{1}{n} \sum_{\ell=1}^{L} \phi_h(h_\ell^1, \beta_{0_\ell}) - \lim \frac{1}{n} \sum_{\ell=1}^{L} \mathbb{E}\{ \bar{\tau}_0 Z_{0_\ell}^* [\eta_\ell^0(\beta_0 - \bar{\tau}_0 Z_0) - \beta_{0_\ell}] \} \right] \overset{a.s.}{=} 0. \tag{A.70}$$

Consider a single term in the expectation in (A.70), say $\ell = 1$. We have

$$\mathbb{E}\{ \bar{\tau}_0 Z_{0_{(1)}}^* [\eta_{(1)}^0(\beta_0 - \bar{\tau}_0 Z_0) - \beta_{0_{(1)}}] \} = \bar{\tau}_0 \sum_{i=1}^{M} \mathbb{E}\{ Z_{0_i} [\eta_i^0(\beta_0 - \bar{\tau}_0 Z_0) - \beta_{0_i}] \} \tag{A.71}$$

where $\beta_{0_{(1)}} = (\beta_{0_1}, \beta_{0_2}, \ldots, \beta_{0_M})$ and $Z_{0_{(1)}} = (Z_{0_1}, Z_{0_2}, .., Z_{0_M})$. Note that for each i, the function $\eta_i^0(\cdot)$ depends on all the M indices in the section containing i. For each $i \in [M]$, we evaluate the expectation on the RHS of (A.71) using the law of iterated expectations:

$$\mathbb{E}\{ Z_{0_i} [\eta_i^0(\beta_0 - \bar{\tau}_0 Z_0) - \beta_{0_i}] \} = \mathbb{E}\left[\mathbb{E}\left\{ Z_{0_i} [\eta_{0_i}(\beta_0 - \bar{\tau}_0 Z_0) - \beta_{0_i}] | \beta_{0_{(1)}}, Z_{0_{(1)\setminus i}} \right\} \right] \tag{A.72}$$

where the inner expectation is over Z_{0_i} conditioned on $\{\beta_{0_{(1)}}, Z_{0_{(1)\setminus i}}\}$. Since Z_{0_i} is independent of $\{\beta_{0_{(1)}}, Z_{0_{(1)\setminus i}}\}$, the latter just act as constants in the inner expectation, which is

over $Z_{0_i} \sim \mathcal{N}(0,1)$. Applying Stein's lemma (Fact 10) to the inner expectation, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{E}\left\{Z_{0_i}[\eta_i^0(\beta_0 - \bar{\tau}_0 Z_0) - \beta_{0_i}] \mid \beta_{0_{(1)}}, Z_{0_{(1)}\setminus i}\right\}\right] = \mathbb{E}\left[\mathbb{E}\left\{\frac{\partial}{\partial Z_{0_i}}[\eta_i^0(\beta_0 - \bar{\tau}_0 Z_0) - \beta_{0_i}] \mid \beta_{0_{(1)}}, Z_{0_{(1)}\setminus i}\right\}\right] \\
&\overset{(a)}{=} -\frac{\bar{\tau}_0}{\bar{\tau}_0^2}\mathbb{E}\left[\mathbb{E}\left\{\eta_i^0(\beta_0 - \bar{\tau}_0 Z_0)\left(\sqrt{nP_1} - \eta_i^0(\beta_0 - \bar{\tau}_0 Z_0)\right)|\beta_{0_{(1)}}, Z_{0_{(1)}\setminus i}\right\}\right] \\
&\overset{(b)}{=} -\frac{1}{\bar{\tau}_0}\mathbb{E}\left[\eta_i^0(\beta_0 - \bar{\tau}_0 Z_0)\left(\sqrt{nP_1} - \eta_i^0(\beta_0 - \bar{\tau}_0 Z_0)\right)\right]
\end{aligned} \tag{A.73}$$

where (a) holds because the definition of η_i^t in (2.15) implies that

$$\frac{\partial \eta_i^t(s)}{\delta s_i} = \frac{\eta_i^t(s)}{\bar{\tau}_t^2}\left(\sqrt{nP_\ell} - \eta_i^t(s)\right) \text{ for } i \in \text{section } \ell,$$

and (b) follows from the law of iterated expectation. Using (A.73) in (A.72) and (A.71), we have

$$\mathbb{E}\left\{\bar{\tau}_0 Z^*_{0_{(1)}}[\eta^0_{(1)}(\beta_0 - \bar{\tau}_0 Z_0) - \beta_{0_{(1)}}]\right\} = \sum_{i=1}^{M}\mathbb{E}\left[\eta_i^0(\beta_0 - \bar{\tau}_0 Z_0)\left(\eta_i^0(\beta_0 - \bar{\tau}_0 Z_0) - \sqrt{nP_1}\right)\right]. \tag{A.74}$$

The argument above can be repeated for each section $\ell \in [L]$ to obtain a relation analogous to (A.74). Using this for the expectation in (A.70), we obtain

$$\lim n^\delta \left[\frac{1}{n}\sum_{\ell=1}^{L}\phi_h(h_\ell^1, \beta_{0_\ell}) - \lim\left(\frac{\mathbb{E}\left\{\|\eta^0(\beta_0 - \bar{\tau}_0 Z_0)\|^2\right\}}{n} - P\right)\right] \overset{a.s.}{=} 0. \tag{A.75}$$

It is shown in Appendix A.3.3 that $\lim\left(P - \frac{\mathbb{E}\left\{\|\eta^0(\beta_0 - \bar{\tau}_0 Z_0)\|^2\right\}}{n}\right) = \bar{\sigma}_1^2$. Therefore (A.75) becomes

$$\lim n^\delta \left[\frac{1}{n}\sum_{\ell=1}^{L}(h^1)^*[\eta^0(\beta_0 - h^1) - \beta_0] + \bar{\sigma}_1^2\right] \overset{a.s.}{=} 0, \tag{A.76}$$

where we have used $\phi_h(h_\ell^1, \beta_{0_\ell}) = (h_\ell^1)^*[\eta_\ell^0(\beta_0 - h^1) - \beta_{0_\ell}]$.

To complete the proof, recall from $\mathcal{H}_1$(c) that $\frac{\|m^0\|^2}{n} \overset{a.s.}{\to} \sigma^2 + \bar{\sigma}_0^2$ at rate $n^{-\delta}$. Further,

from (2.31), we observe that

$$\lambda_1 = \frac{1}{\bar{\tau}_0^2}\left(\frac{\|\beta^1\|^2}{n} - P\right) \overset{a.s.}{\to} \lim \frac{1}{\bar{\tau}_0^2}\left(\frac{\mathbb{E}\left\{\|\eta^0(\beta_0 - \bar{\tau}_0 Z_0)\|^2\right\}}{n} - P\right) = \frac{-\bar{\sigma}_1^2}{\bar{\tau}_0^2} = \frac{-\bar{\sigma}_1^2}{\sigma^2 + \bar{\sigma}_0^2}, \tag{A.77}$$

where the convergence at rate $n^{-\delta}$ follows from $\mathcal{H}_1$(b) applied to the function $\frac{\|\eta^0(\beta - h^1)\|^2}{n} = \frac{\|\beta^1\|^2}{n}$.

(**e**) We use $\mathcal{H}_1$(a) to represent

$$h^1\big|_{\mathscr{S}_{1,0}} \overset{d}{=} \tilde{X}^* m_\perp^0 + \tilde{Q}_1\, o(n^{-\delta}) = \tilde{X}^* m^0 + \frac{q^0}{\sqrt{P}} o(n^{-\delta}). \tag{A.78}$$

Therefore

$$\frac{(q^0)^* h^1}{n}\Big|_{\mathscr{S}_{1,0}} \overset{d}{=} \frac{(q^0)^* \tilde{X}^* m^0}{n} + \frac{\|q^0\|^2}{n\sqrt{P}} o(n^{-\delta}) \overset{d}{=} \sqrt{P}\frac{\|m^0\|}{\sqrt{n}}\frac{Z}{\sqrt{n}} + \sqrt{P} o(n^{-\delta}), \tag{A.79}$$

where we have used Fact 9(a) as q^0, m^0 are in the sigma-field and independent of $\tilde{X}$. By $\mathcal{H}_1$(c), $\lim \frac{\|m^0\|^2}{n} \overset{a.s.}{=} \bar{\tau}_0^2$ and therefore (A.79) goes to zero almost surely in the limit at rate $n^{-\delta}$.

(**f**) Since Q^0 is the empty matrix, $q_\perp^0 = q^0$ and so $\lim \frac{\|q_\perp^0\|^2}{n} = \lim \frac{\|q^0\|^2}{n} = P$.

Step 3: Showing $\mathcal{B}_t$ holds

(f) By the induction hypothesis $\mathcal{B}_{t-1}$, (B.22) is true for $0 \le s \le t-2$, so we prove the $s = t-1$ case. Let $\mathsf{P}_{M_{t-1}} = M_{t-1}(M_{t-1}^* M_{t-1})^{-1} M_{t-1}^*$ be the projection matrix onto the column space of M_{t-1}. Then,

$$\frac{\|m_\perp^{t-1}\|^2}{n} = \|(\mathsf{I} - \mathsf{P}_{M_{t-1}})m^{t-1}\|^2 = \frac{\|m^{t-1}\|^2}{n} - \frac{(m^{t-1})^* M_{t-1}}{n}\left(\frac{M_{t-1}^* M_{t-1}}{n}\right)^{-1}\frac{M_{t-1}^* m^{t-1}}{n}. \tag{A.80}$$

Consider the matrix inverse in (A.80). By the induction hypothesis $\mathcal{B}_{t-1}$(f),

$$\lim \frac{\|m_\perp^r\|^2}{n} = \lim \frac{\|m^r - \mathsf{P}_{M_{r-1}} m^r\|^2}{n} > \varsigma_r \quad \text{for } 0 \le r \le t-2, \tag{A.81}$$

for positive constants ς_r. Using (A.81), Facts 11 and 12 imply that the smallest eigenvalue of $\lim \frac{M_{t-1}^* M_{t-1}}{n}$ is greater than some positive constant; hence its inverse exists.

Using the induction hypothesis $\mathcal{H}_t(\mathrm{c})$, we have for $0 \leq r, s \leq t-1$:

$$\lim \frac{(m^r)^* m^s}{n} \stackrel{a.s.}{=} \mathbb{E}\left[(\bar{\sigma}_r \hat{Z}_r - \sigma Z_\epsilon)(\bar{\sigma}_s \hat{Z}_s - \sigma Z_\epsilon)\right] \tag{A.82}$$

where $(\hat{Z}_r, \hat{Z}_s)$ are jointly Gaussian with $\mathcal{N}(0,1)$ marginals, and independent of Z_ϵ. Using (A.82) in (A.80), we obtain

$$\lim \frac{\|m_\perp^{t-1}\|^2}{n} = \mathbb{E}\left[(\bar{\sigma}_{t-1} \hat{Z}_{t-1} - \sigma Z_\epsilon)^2\right] - u^* C^{-1} u \tag{A.83}$$

where for $1 \leq i, j \leq (t-1)$,

$$u_i = \mathbb{E}\left[(\bar{\sigma}_{t-1} \hat{Z}_{t-1} - \sigma Z_\epsilon)(\bar{\sigma}_{i-1} \hat{Z}_{i-1} - \sigma Z_\epsilon)\right], \quad C_{ij} = \mathbb{E}\left[(\bar{\sigma}_{i-1} \hat{Z}_{i-1} - \sigma Z_\epsilon)(\bar{\sigma}_{j-1} \hat{Z}_{j-1} - \sigma Z_\epsilon)\right]. \tag{A.84}$$

Now the result follows from Fact 14 if we can show that there exists strictly positive constants $c_1, \ldots, c_{t-1}$ such that $\mathrm{Var}(\bar{\sigma}_r Z_r | \bar{\sigma}_0 Z_0, \ldots, \bar{\sigma}_{r-1} Z_{r-1}) \geq c_r$, for $1 \leq r \leq (t-1)$. Indeed, we will now prove that

$$\mathrm{Var}(\bar{\sigma}_r Z_r | \bar{\sigma}_0 Z_0, \ldots, \bar{\sigma}_{r-1} Z_{r-1}) = \bar{\sigma}_r^2 \left(1 - \frac{\bar{\sigma}_r^2}{\bar{\sigma}_{r-1}^2}\right). \tag{A.85}$$

Since $\bar{\sigma}_r^2 = \sigma^2 \left((1+\mathsf{snr})^{1-\xi_{r-1}} - 1\right)$, the definition of ξ_{r-1} in (2.24) implies that the RHS of (A.85) is strictly positive for $r \leq T^* - 1$, where $T^* = \left\lceil \frac{2\mathcal{C}}{\log(\mathcal{C}/R)} \right\rceil$.

For $r \in [t-1]$, we have

$$\lim \frac{\|b_\perp^r\|^2}{n} = \lim \frac{\|b^r\|^2}{n} - \frac{(b^r)^* B_r}{n} \left(\frac{B_r^* B_r}{n}\right)^{-1} \frac{B_r^* b^r}{n} = \lim \frac{\|q^r\|^2}{n} - \frac{(q^r)^* Q_r}{n} \left(\frac{Q_r^* Q_r}{n}\right)^{-1} \frac{Q_r^* q^r}{n} \tag{A.86}$$

where the second equality follows from the induction hypothesis $\mathcal{B}_{t-1}(c)$ which says that

$$\lim \frac{(b^{r'})^* b^r}{n} = \lim \frac{(q^{r'})^* q^r}{n} = \bar{\sigma}_r^2 \ \text{ for } 0 \leq r' \leq r \leq (t-1). \tag{A.87}$$

Denoting $\lim \frac{B_r^* B_r}{n} = \lim \frac{Q_r^* Q_r}{n}$ by $\tilde{C}$, we have $\tilde{C}_{ij} = \tilde{C}_{ji} = \lim \frac{(q^i)^* q^j}{n} = \bar{\sigma}_j^2$, for $0 \leq i \leq j \leq (r-1)$. The induction hypothesis $\mathcal{H}_t(\mathrm{f})$ guarantees that $\frac{\|q_\perp^r\|^2}{n}$ is strictly positive for $0 \leq r \leq t-1$. Consequently, Facts 11 and 12 imply that $\tilde{C}$ is invertible. Hence

$$\lim \frac{\|b_\perp^r\|^2}{n} = \lim \frac{\|q^r\|^2}{n} - \frac{(q^r)^* Q_r}{n} \left(\frac{Q_r^* Q_r}{n} \right)^{-1} \frac{Q_r^* q^r}{n} \overset{(a)}{=} \bar{\sigma}_r^2 - \bar{\sigma}_r^2 (\mathsf{e}_r^* \tilde{C}^{-1} \mathsf{e}_r) \bar{\sigma}_r^2 \overset{(b)}{=} \bar{\sigma}_r^2 \left(1 - \frac{\bar{\sigma}_r^2}{\bar{\sigma}_{r-1}^2} \right). \tag{A.88}$$

In (A.88), (a) is obtained using (A.87) with $\mathsf{e}_r \in \mathbb{R}^r$ denoting the all-ones column vector. The equality (b) is obtained using the fact that $\tilde{C}^{-1}\mathsf{e}_r$ is the solution to $\tilde{C}x = \mathsf{e}_r$: since all the entries in the last column of $\tilde{C}$ are equal to $\bar{\sigma}_{r-1}^2$, by inspection the solution to $\tilde{C}x = \mathsf{e}_r$ is $x = [0, \ldots, 0, (\bar{\sigma}_{r-1}^2)^{-1}]^*$, which yields equality (b) in (A.88).

Using the induction hypothesis $\mathcal{B}_{t-1}(b)$ for the $PL(2)$ function $\phi_b(x, y) = xy$, we have

$$\lim \frac{1}{n} (b^r)^* b^s = \lim \sum_{i=1}^{n} \frac{1}{n} b_i^r b_i^s = \mathbb{E}[\bar{\sigma}_r \hat{Z}_r \bar{\sigma}_s \hat{Z}_s], \ 0 \leq r, s \leq (t-1). \tag{A.89}$$

Using this, we obtain

$$\lim \frac{\|b_\perp^r\|^2}{n} = \lim \frac{\|b^r\|^2}{n} - \frac{(b^r)^* B_r}{n} \left(\frac{B_r^* B_r}{n} \right)^{-1} \frac{B_r^* b^r}{n} = \bar{\sigma}_r^2 - v^* D^{-1} v \overset{(a)}{=} \mathrm{Var}(\bar{\sigma}_r \hat{Z}_r | \bar{\sigma}_0 \hat{Z}_0, \ldots, \bar{\sigma}_{r-1} \hat{Z}_{r-1}) \tag{A.90}$$

where for $0 \leq i, j \leq (r-1)$, $v_i = \mathbb{E}\left[\bar{\sigma}_r \bar{\sigma}_i \hat{Z}_r \hat{Z}_i\right]$, and $D_{ij} = \mathbb{E}\left[\bar{\sigma}_i \bar{\sigma}_j \hat{Z}_i \hat{Z}_j\right]$. Equality (a) in (A.90) follows from Fact 13. We have proved (A.85) via (A.88) and (A.90), which completes the proof of $\mathcal{B}_t(\mathrm{f})$.

We now state a couple of lemmas that will be useful for proving the remainder of $\mathcal{B}_t$ and $\mathcal{H}_{t+1}$.

Lemma 16. *For $t \leq T^*$, the vectors of coefficients in* (3.15)*, given by*

$$\vec{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_{t-1}) = \left(\frac{M_t^* M_t}{n} \right)^{-1} \frac{M_t^* m^t}{n}, \qquad \vec{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_{t-1}) = \left(\frac{Q_t^* Q_t}{n} \right)^{-1} \frac{Q_t^* q^t}{n}$$

converge to finite limits at rate $n^{-\delta}$ as $n \to \infty$.

Proof. From the induction hypothesis $\mathcal{H}_t(\mathrm{c})$, $\frac{(m^r)^* m^s}{n}$ converges almost surely to a constant

at rate $n^{-\delta}$ for $r, s \leq (t-1)$. Further, $\mathcal{B}_t(\mathrm{f})$ proved above and Fact 11 together imply that the smallest eigenvalue of the matrix $\frac{M_t^* M_t}{n}$ is bounded from below by a positive constant for all n; then Fact 12 implies that its inverse has a finite limit. Further, the inverse converges to its limit at rate $n^{-\delta}$ as each entry in $\frac{M_t^* M_t}{n}$ converges at this rate. The statement for $\vec{\gamma}$ is proved in an analogous manner using the induction hypotheses $\mathcal{B}_{t-1}(\mathrm{c})$ and $\mathcal{H}_t(\mathrm{f})$, together with Facts 11 and 12. □

Lemma 17. *The following statements hold for $t \leq T^*$:*

$$h^{t+1}|_{\mathscr{S}_{t+1,t}} \stackrel{d}{=} H_t (M_t^* M_t)^{-1} M_t^* m_{\|}^t + P_{Q_{t+1}}^{\perp} \tilde{X}^* m_{\perp}^t + Q_{t+1} \vec{o}_{t+1}(n^{-\delta}), \tag{A.91}$$

$$b^t|_{\mathscr{S}_{t,t}} \stackrel{d}{=} B_t (Q_t^* Q_t)^{-1} Q_t^* q_{\|}^t + P_{M_t}^{\perp} \tilde{X} q_{\perp}^t + M_t \vec{o}_t(n^{-\delta}), \tag{A.92}$$

where $B_t = [b^0 \mid \ldots \mid b^{t-1}]$ and $H_t = [h^1 \mid \ldots \mid h^t]$.

Proof. The proof is very similar to that of [6, Lemma 13]. We use Lemmas 14 and 15 to write

$$\begin{aligned} b^t|_{\mathscr{S}_{t,t}} &= (Xq^t - \lambda_t m^{t-1})|_{\mathscr{S}_{t,t}} \stackrel{\mathrm{d}}{=} Y_t (Q_t^* Q_t)^{-1} Q_t^* q_{\|}^t + M_t (M_t^* M_t)^{-1} X_t^* q_{\perp}^t + P_{M_t}^{\perp} \tilde{X} q_{\perp}^t - \lambda_t m^{t-1} \\ &= B_t (Q_t^* Q_t)^{-1} Q_t^* q_{\|}^t + [0|M_{t-1}] \Lambda_t (Q_t^* Q_t)^{-1} Q_t^* q_{\|}^t + M_t (M_t^* M_t)^{-1} H_t^* q_{\perp}^t + P_{M_t}^{\perp} \tilde{X} q_{\perp}^t - \lambda_t m^{t-1}, \end{aligned} \tag{A.93}$$

where $\Lambda_t = \mathrm{diag}(\lambda_0, \ldots, \lambda_{t-1})$. The last equality above is obtained using $Y_t = B_t + [0|M_{t-1}]\Lambda_t$, and $X_t = H_t + Q_t$. Thus, to show (A.92), we need to prove that

$$[0|M_{t-1}] \Lambda_t \vec{\gamma} + M_t (M_t^* M_t)^{-1} H_t^* q_{\perp}^t - \lambda_t m^{t-1} = M_t \, \vec{o}_t(n^{-\delta}). \tag{A.94}$$

Observe that each side of (A.94) is a linear combination of $\{m^k\}, 0 \leq k \leq (t-1)$. The coefficient of m^k on the LHS equals

$$\begin{aligned} &\lambda_{k+1}\gamma_{k+1} + \left[\left(\tfrac{M_t^* M_t}{n} \right)^{-1} \tfrac{H_t^* q_{\perp}^t}{n} \right]_{k+1} && \text{for } 0 \leq k \leq t-2, \\ &-\lambda_t + \left[\left(\tfrac{M_t^* M_t}{n} \right)^{-1} \tfrac{H_t^* q_{\perp}^t}{n} \right]_t, && \text{for } k = t-1. \end{aligned} \tag{A.95}$$

We prove (A.94) by showing that each of the coefficients above is $o(n^{-\delta})$. Indeed, for

$1 \le i \le t$,

$$\begin{aligned}\left[\frac{H_t^* q_\perp^t}{n}\right]_i = \frac{(h^i)^* q_\perp^t}{n} = \frac{(h^i)^*(q^t - q_\parallel^t)}{n} = \frac{(h^i)^* q^t}{n} - \sum_{r=0}^{t-1} \gamma_r \frac{(h^i)^* q^r}{n} \\ \overset{a.s.}{\to} \lim \left[\lambda_t \frac{(m^{i-1})^* m^{t-1}}{n} - \sum_{r=1}^{t-1} \gamma_r \lambda_r \frac{(m^{i-1})^* m^{r-1}}{n}\right]\end{aligned} \tag{A.96}$$

where the convergence (at rate $n^{-\delta}$) follows from $\mathcal{H}_t$(d); Lemma 16 guarantees the convergence of the γ_r coefficients. Therefore

$$\left[\frac{H_t^* q_\perp^t}{n}\right] \overset{a.s.}{\to} \lim \left[\lambda_t \frac{(M_t)^* m^{t-1}}{n} - \sum_{r=0}^{t-2} \gamma_{r+1} \lambda_{r+1} \frac{(M_t)^* m^r}{n}\right] \quad \text{at rate } n^{-\delta}. \tag{A.97}$$

Substituting (A.97) in (A.95) yields (A.94), and completes the proof of (A.92). The other part of the lemma, (A.91), is proved in a similar manner. □

(a) From Lemma 17, we have

$$b^t|_{\mathscr{S}_{t,t}} \overset{\mathrm{d}}{=} B_t (Q_t^* Q_t)^{-1} Q_t^* q_\parallel^t + P_{M_t}^\perp \tilde{X} q_\perp^t + M_t \vec{o}_t(n^{-\delta}). \tag{A.98}$$

First notice that

$$B_t (Q_t^* Q_t)^{-1} Q_t^* q_\parallel^t = B_t \left(\frac{Q_t^* Q_t}{n}\right)^{-1} \frac{Q_t^* q^t}{n} = B_t \vec{\gamma} = \sum_{i=0}^{t-1} \gamma_i b^i, \tag{A.99}$$

where $\vec{\gamma}$ is defined in Lemma 16. Next, observe that $P_{M_t}^\perp \tilde{X} q_\perp^t = \tilde{X} q^t - P_{M_t}^\parallel \tilde{X} q^t$. Hence the result follows if we can show that $P_{M_t}^\parallel \tilde{X} q^t = \tilde{M}_t \vec{o}_t(n^{-\delta})$. Indeed, using Fact 9(b), we see that

$$P_{M_t}^\parallel \tilde{X} q_\perp^t \overset{d}{=} \frac{\|q_\perp^t\|}{\sqrt{n}} \tilde{M}_t \vec{o}_t(n^{-\delta}) \overset{d}{=} \tilde{M}_t \vec{o}_t(n^{-\delta})$$

where the last equality follows since $\frac{\|q_\perp^t\|^2}{n} \le \frac{\|q^t\|^2}{n} \le 2P$.

(c) By the induction hypothesis, the result holds for all $r, s < t$, so we only consider the $r < t, s = t$ and $r = s = t$ cases. From $\mathcal{B}_t$(a) above, we have

$$b^t|_{\mathscr{S}_{t,t}} \overset{d}{=} \sum_{i=0}^{t-1} \gamma_i b^i + \tilde{X} q_\perp^t + M_t \vec{o}_t(n^{-\delta}) \tag{A.100}$$

For $r < t, s = t$, we have from (A.100):

$$\frac{(b^r)^* b^t}{n}\Big|_{\mathscr{S}_{t,t}} \stackrel{d}{=} \sum_{i=1}^{t-1} \gamma_i \frac{(b^r)^* b^i}{n} + \frac{(b^r)^* \tilde{X} q_\perp^t}{n} + \sum_{i=0}^{t-1} o(n^{-\delta}) \frac{(b^r)^* m^i}{n}. \tag{A.101}$$

Applying Fact 9(a), the second term in (A.101) is $\frac{(b^r)^* \tilde{X} q_\perp^t}{n} \stackrel{d}{=} \frac{\|b^r\| \|q_\perp^t\|}{n} \frac{Z}{\sqrt{n}}$, where $Z \sim \mathcal{N}(0,1)$. Therefore the last two terms in (A.101) are $o(n^{-\delta})$ since $\frac{\|q_\perp^t\|}{n} \leq 2P$ and $\mathcal{B}_{t-1}$(c), (d) imply that $\frac{\|b^r\|}{n}$ and $\frac{(b^r)^* m^i}{n}$ converge to finite limits. Using $\mathcal{B}_{t-1}$(c) again, the limit of first term in (A.101) can be written as

$$\lim \sum_{i=0}^{t-1} \gamma_i \frac{(b^r)^* b^i}{n} \stackrel{a.s.}{=} \lim \sum_{i=1}^{t-1} \gamma_i \frac{(q^r)^* q^i}{n} = \lim \frac{(q^r)^* q_\parallel^t}{n} \stackrel{(a)}{=} \lim \frac{(q^r)^* q^t}{n} \stackrel{(b)}{=} \bar{\sigma}_t^2 \ a.s. \tag{A.102}$$

where the γ_i's have finite limits due to Lemma 16. Equality (a) in (A.102) holds because $q_\perp^t \perp q^r$, while (b) is obtained by applying $\mathcal{H}_t$(b) to the function

$$\phi_h(h_\ell^r, h_\ell^s, \beta_{0_\ell}) := [\eta_\ell^{r-1}(\beta_0 - h^r) - \beta_{0_\ell}]^* [\eta_\ell^{s-1}(\beta_0 - h^s) - \beta_{0_\ell}] = (q_\ell^r)^* q_\ell^s,$$

which yields

$$\lim \frac{(q^r)^* q^t}{n} \stackrel{a.s.}{=} \lim \frac{1}{n} \mathbb{E}\{[\eta^{r-1}(\beta - \tau_{r-1} Z_{r-1}) - \beta]^* [\eta^{t-1}(\beta - \tau_{t-1} Z_{t-1}) - \beta]\} = \bar{\sigma}_t^2,$$

where the second equality above is proved in Appendix A.3.3. From $\mathcal{B}_{t-1}$(c) and $\mathcal{H}_{t-1}$(b), it follows that the rate of convergence in (A.102) is $n^{-\delta}$.

For $r = s = t$, using (A.100), we have

$$\begin{aligned}\frac{\|b^t\|^2}{n}|_{\mathscr{S}_{t,t}} \stackrel{d}{=} & \sum_{i=0}^{t-1} \sum_{i'=0}^{t-1} \gamma_i \gamma_{i'} \frac{(b^i)^* b^{i'}}{n} + \frac{\|\tilde{X} q_\perp^t\|^2}{n} \\ & + 2 \sum_{i=0}^{t-1} \gamma_i \frac{(b^i)^* \tilde{X} q_\perp^t}{n} + 2 \sum_{i=0}^{t-1} \gamma_i \frac{(b^i)^* M_t \vec{o}_t(n^{-\delta})}{n} + 2 \frac{(\tilde{X} q_\perp^t)^* M_t \vec{o}_t(n^{-\delta})}{n} + \frac{\|M_t \vec{o}_t(n^{-\delta})\|^2}{n}.\end{aligned} \tag{A.103}$$

Using arguments similar to those for the $r < t$ case, the last four terms in (A.103) can be shown to be $o(n^{-\delta})$, and by Fact 9 $\frac{\|\tilde{X} q_\perp^t\|^2}{n} = \frac{\|q_\perp^t\|^2}{n} \frac{\|Z\|^2}{n}$ where $Z \in \mathbb{R}^n$ is i.i.d. standard

normal. Therefore,

$$\frac{\|b^t\|^2}{n}|_{\mathscr{S}_{t,t}} \stackrel{d}{=} \sum_{i=0}^{t-1}\sum_{i'=0}^{t-1}\gamma_i\gamma_{i'}\frac{(b^i)^* b^{i'}}{n} + \frac{\|q_\perp^t\|^2}{n} + o(n^{-\delta})$$
$$\stackrel{a.s.}{\to} \lim \sum_{i=0}^{t-1}\sum_{i'=0}^{t-1}\gamma_i\gamma_{i'}\frac{(q^i)^* q^{i'}}{n} + \frac{\|q_\perp^t\|^2}{n} = \lim \frac{\|q_\parallel^t\|^2}{n} + \frac{\|q_\perp^t\|^2}{n} = \lim \frac{\|q^t\|^2}{n},$$

where the convergence at rate $n^{-\delta}$ follows from $\mathcal{B}_{t-1}$(c).

(b) Using the characterization for b^t obtained in $\mathcal{B}_t$(a) above, we have

$$\phi_b(b_i^0,\ldots,b_i^t,\epsilon_i)\Big|_{\mathscr{S}_{t,t}} \stackrel{d}{=} \phi_b\left(b_i^0,\ldots,b_i^{t-1},\left[\sum_{r=0}^{t-1}\gamma_r b^r + \tilde{X}q_\perp^t + \tilde{M}_t\vec{o}_t(n^{-\delta'})\right]_i,\epsilon_i\right). \quad \text{(A.104)}$$

for some $\delta' > 0$. The term $\tilde{M}_t\vec{o}_t(n^{-\delta'})$ in the RHS can be dropped. Indeed, defining

$$a_i = \left(b_i^0,\ldots,b_i^{t-1},\left[\sum_{r=0}^{t-1}\gamma_r b^r + \tilde{X}q_\perp^t + \tilde{M}_t\vec{o}_t(n^{-\delta'})\right]_i,\epsilon_i\right), \quad c_i = \left(b_i^0,\ldots,b_i^{t-1},\left[\sum_{r=0}^{t-1}\gamma_r b^r + \tilde{X}q_\perp^t\right]_i,\epsilon_i\right),$$

we can show that

$$\frac{1}{n}\left|\sum_{i=1}^n \phi_b(a_i) - \sum_{i=1}^n \phi_b(c_i)\right| \le \frac{1}{n}\sum_{i=1}^n |\phi_b(a_i) - \phi_b(c_i)| \stackrel{(a)}{\le} \frac{C}{n}\sum_{i=1}^n (1+\|a_i\|+\|c_i\|)\left|\left(\tilde{M}_t\vec{o}_t(n^{-\delta'})\right)_i\right|$$
$$\stackrel{(b)}{\le} C\sqrt{\sum_{i=1}^n \frac{(1+\|a_i\|+\|c_i\|)^2}{n}}\sqrt{\sum_{r=0}^{t-1}\frac{\|\tilde{m}^r\|^2}{n}2^t o(n^{-\delta'})} \stackrel{(c)}{=} o(n^{-\delta'}). \quad \text{(A.105)}$$

In (A.105), (a) holds because $\phi_b \in PL(2)$. (b) is obtained using Hölder's inequality and the fact that $\sum_{i=1}^n \left[\tilde{M}_t\vec{o}_t(n^{-\delta'})\right]_i^2 \le 2^t\vec{o}_1(n^{-\delta'})\sum_{r=0}^{t-1}\|\tilde{m}^r\|^2$. Equality (c) can be shown by verifying that $\sum_{i=1}^n \frac{\|a_i\|^2}{n}$ and $\sum_{i=1}^n \frac{\|c_i\|^2}{n}$are bounded and finite. The details are similar to [6, $\mathcal{B}_t$(b)] and are omitted. Thus by choosing $\delta < \delta'$, we can work with c_i instead of a_i.

Next, we use Fact 2 to show that

$$\lim n^\delta\left[\frac{1}{n}\sum_{i=1}^n \phi_b(c_i) - \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\tilde{X}}\{\phi_b(c_i)\}\right] \stackrel{a.s.}{=} 0, \quad \text{(A.106)}$$

To appeal to Fact 2, we need to verify that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left|n^{\delta}\phi_b\left(c_i\right)-\mathbb{E}_{\tilde{X}}\left\{n^{\delta}\phi_b\left(c_i\right)\right\}\right|^{2+\kappa}\leq cn^{\kappa/2}. \tag{A.107}$$

Using steps similar to (A.45), we can show that

$$\begin{aligned}\mathbb{E}\left|\phi_b\left(c_i\right)-\mathbb{E}_{\tilde{X}}\left\{\phi_b\left(c_i\right)\right\}\right|^{2+\kappa} &\leq \kappa'\mathbb{E}_{\tilde{X}',\tilde{X}}\left\{|[\tilde{X}'q_{\perp}^{t}]_i-[\tilde{X}q_{\perp}^{t}]_i|^{2+\kappa}\left(1+|[\tilde{X}'q_{\perp}^{t}]_i|^{2+\kappa}+|[\tilde{X}q_{\perp}^{t}]_i|^{2+\kappa}\right)\right\}\\ &\quad+\kappa'\left(\sum_{r=0}^{t-1}(|1+\gamma_r||b_i^r|)^{2+\kappa}+|\epsilon_i|^{2+\kappa}\right)\mathbb{E}_{\tilde{X}',\tilde{X}}\left\{|[\tilde{X}'q_{\perp}^{t}]_i-[\tilde{X}q_{\perp}^{t}]_i|^{2+\kappa}\right\}\\ &\overset{(a)}{\leq}\kappa_1+\kappa_2\left(|\epsilon_i|^{2+\kappa}+\sum_{r=0}^{t-1}(1+\gamma_r)^{2+\kappa}|b_i^r|^{2+\kappa}\right)\end{aligned} \tag{A.108}$$

for some constants $\kappa',\kappa_1,\kappa_2>0$, where $\tilde{X},\tilde{X}'$ are independent copies of X. In (A.108), (a) holds because $\tilde{X}q_{\perp}^{t}\overset{d}{=}\frac{\|q_{\perp}^{t}\|}{n}\tilde{Z}$ and $\frac{\|q_{\perp}^{t}\|}{n}\leq\frac{\|q^{t}\|}{n}\leq P$; similarly, $\tilde{X}'q_{\perp}^{t}\overset{d}{=}\frac{\|q_{\perp}^{t}\|}{n}\tilde{Z}'$, where $\tilde{Z},\tilde{Z}'$ are $\mathcal{N}(0,1)$. Substituting (A.108) in the LHS of (A.107), and applying induction hypothesis $\mathcal{B}_t(b)$ shows that the condition (A.107) is satisfied if $\delta<\frac{\kappa/2}{\kappa+2}$.

Thus we now need to show that

$$\lim\frac{n^{\delta}}{n}\sum_{i=1}^{n}\left[\mathbb{E}_{\tilde{X}}\left\{\phi_b(b_i^0,\ldots,b_i^{t-1},\sum_{r=0}^{t-1}\gamma_r b_i^r+[\tilde{X}q_{\perp}^{t}]_i,\epsilon_i)\right\}-\mathbb{E}\{\phi_b(\bar{\sigma}_0Z_0,\ldots,\bar{\sigma}_tZ_t,\sigma Z_\epsilon)\}\right]\overset{a.s.}{=}0. \tag{A.109}$$

Recalling that $[\tilde{X}q_{\perp}^{t}]_i\overset{d}{=}\frac{\|q_{\perp}^{t}\|}{\sqrt{n}}\tilde{Z}$ where $\tilde{Z}\sim\mathcal{N}(0,1)$, we have

$$\mathbb{E}_{\tilde{X}}\left\{\phi_b(b_i^0,\ldots,b_i^{t-1},\sum_{r=0}^{t-1}\gamma_r b_i^r+[\tilde{X}q_{\perp}^{t}]_i,\epsilon_i)\right\}=\mathbb{E}_{\tilde{Z}}\left\{\phi_b(b_i^0,\ldots,b_i^{t-1},\sum_{r=0}^{t-1}\gamma_r b_i^r+\frac{\|q_{\perp}^{t}\|}{\sqrt{n}}\tilde{Z},\epsilon_i)\right\}. \tag{A.110}$$

Define the function

$$\phi_b^{NEW}(b_i^0,\ldots,b_i^{t-1},\epsilon_i):=\mathbb{E}_{\tilde{Z}}\left\{\phi_b(b_i^0,\ldots,b_i^{t-1},\sum_{r=0}^{t-1}\gamma_r b_i^r+\frac{\|q_{\perp}^{t}\|}{\sqrt{n}}\tilde{Z},\epsilon_i)\right\}. \tag{A.111}$$

It can be verified that $\phi_b^{NEW}\in PL(2)$, and hence the induction hypothesis $\mathcal{B}_{t-1}$(b) implies

that

$$\lim n^{\delta}\left[\frac{1}{n}\sum_{i=1}^{n}\phi_b^{NEW}(b_i^0,\ldots,b_i^{t-1},\epsilon_i)-\mathbb{E}\left\{\phi_b^{NEW}(\bar{\sigma}_0\hat{Z}_0,\ldots,\bar{\sigma}_{t-1}\hat{Z}_{t-1},\sigma Z_\epsilon)\right\}\right]\overset{a.s.}{=}0. \tag{A.112}$$

Thus from (A.110) – (A.112), we see that

$$\begin{aligned}\lim\frac{n^{\delta}}{n}\Bigg[&\sum_{i=1}^{n}\mathbb{E}_{\tilde{X}}\left\{\phi_b(b_i^0,\ldots,b_i^{t-1},\sum_{r=0}^{t-1}\gamma_r b_i^r+[\tilde{X}q_\perp^t]_i,\epsilon_i)\right\}\\&-\mathbb{E}\mathbb{E}_{\tilde{Z}}\left\{\phi_b(\bar{\sigma}_0\hat{Z}_0,\ldots,\bar{\sigma}_{t-1}\hat{Z}_{t-1},\sum_{r=0}^{t-1}\gamma_r\bar{\sigma}_r\hat{Z}_r+\frac{\|q_\perp^t\|}{\sqrt{n}}\tilde{Z},\sigma Z_\epsilon)\right\}\Bigg]\overset{a.s.}{=}0.\end{aligned} \tag{A.113}$$

In (A.113), Lemma 16 implies that the γ_r's converge to a finite limit as $n\to\infty$. Further,

$$\frac{\|q_\perp^t\|^2}{n}=\frac{\|q^t\|^2}{n}-\frac{\|q_\parallel^t\|^2}{n}=\frac{\|q^t\|^2}{n}-\frac{\|\sum_{r=1}^{t-1}\gamma_r q^r\|^2}{n}=\frac{\|q^t\|^2}{n}-\frac{\sum_{r,s=1}^{t-1}\gamma_r\gamma_s(q^r)^*q^s}{n}.$$

Hence $\frac{\|q_\perp^t\|}{\sqrt{n}}$ also converges to a finite limit due to $\mathcal{B}_t(\mathrm{c})$, proved above. The final step is to show that the variance of the Gaussian random variable $\left(\sum_{r=0}^{t-1}\gamma_r\bar{\sigma}_r\hat{Z}_r+\frac{\|q_\perp^t\|}{\sqrt{n}}\tilde{Z}\right)$ converges to $\bar{\sigma}_t^2$ at rate $n^{-\delta'}$ for some $\delta'>0$. Applying (A.113) to the $PL(2)$ function $\phi_b(b_i^0,\ldots,b_i^t,\epsilon_i):=(b_i^t)^2$, we obtain

$$\lim n^{\delta}\left[\frac{\|b^t\|^2}{n}-\mathbb{E}\left\{\left(\sum_{r=0}^{t-1}\gamma_r\bar{\sigma}_r\hat{Z}_r+\frac{\|q_\perp^t\|}{\sqrt{n}}\tilde{Z}\right)^2\right\}\right]\overset{a.s.}{=}0. \tag{A.114}$$

Using the induction hypothesis $\mathcal{H}_t(\mathrm{b})$ for the function $\phi_\ell(h_\ell,\beta_\ell)=\|\eta_\ell^{t-1}(\beta-h^t)-\beta_\ell\|^2=\|q_\ell^t\|^2$, we have

$$\lim n^{\delta}\left[\frac{\|q^t\|^2}{n}-\mathbb{E}\left\{\frac{\|\eta^{t-1}(\beta-\bar{\tau}_{t-1}Z_{t-1})-\beta\|^2}{n}\right\}\right]\overset{a.s.}{=}\lim n^{\delta}\left[\frac{\|q^t\|^2}{n}-\bar{\sigma}_t^2\right]\overset{a.s.}{=}0 \tag{A.115}$$

since Appendix A.3.3 shows that $\lim\mathbb{E}\{\|\eta^{t-1}(\beta-\bar{\tau}_{t-1}Z_{t-1})-\beta\|^2/n\}=\bar{\sigma}_t^2$. Further, induction hypothesis $\mathcal{B}_t(\mathrm{c})$ implies that $\lim n^{\delta}\left[\frac{\|b^t\|^2}{n}-\frac{\|q^t\|^2}{n}\right]\overset{a.s.}{=}0$. Combining this with (A.114) and (A.115) completes the proof.

(d) By definition $m^s=b^s-w$ and so

$$\lim \frac{(b^r)^* m^s}{n} = \lim \frac{(b^r)^* b^s}{n} - \lim \frac{(b^r)^* w}{n}.$$

By $\mathcal{B}_t$(c), $\frac{(b^r)^* b^s}{n}$ converges almost surely to $\bar{\sigma}_s^2$ at rate $n^{-\delta}$. Hence the result follows if it can be shown that $\frac{(b^r)^* w}{n}$ approaches 0 almost surely at rate $n^{-\delta}$. Applying $\mathcal{B}_t$(b) to the function $\phi_b(b_i^r, \epsilon_i) = b_i^r \epsilon_i$, we obtain

$$\lim n^{\delta} \left[\frac{1}{n} \sum_{i=1}^{n} \phi_b(b_i^r, \epsilon_i) - \mathbb{E}\left\{ \phi_b(\bar{\sigma}_r \hat{Z}_r, \sigma Z_\epsilon) \right\} \right] \stackrel{a.s}{=} 0. \tag{A.116}$$

The result holds since $\mathbb{E}\left\{ \phi_b(\bar{\sigma}_r \hat{Z}_r, \sigma Z_\epsilon) \right\} = \mathbb{E}\left\{ \bar{\sigma}_r \sigma \hat{Z}_r Z_\epsilon \right\} = 0$ as $\hat{Z}_r$ is independent of Z_ϵ.

Step 4: Showing $\mathcal{H}_{t+1}$ holds

(f) By the induction hypothesis, $\mathcal{H}_t$(f) is true for $0 \leq r \leq (t-1)$. For $r = t$, we have

$$\lim \frac{\|q_\perp^t\|^2}{n} = \lim \frac{\|q^t\|^2}{n} - \frac{(q^t)^* Q_t}{n} \left(\frac{Q_t^* Q_t}{n} \right)^{-1} \frac{(Q_t)^* q^t}{n} \tag{A.117}$$

We note the matrix inverse in (A.117) exists almost surely. Indeed, from the induction hypothesis $\mathcal{H}_t$(f) we have

$$\lim \frac{\|q_\perp^r\|^2}{n} = \bar{\sigma}_r^2 \left(1 - \frac{\bar{\sigma}_r^2}{\bar{\sigma}_{r-1}^2} \right) > 0 \quad \text{for } 0 \leq r \leq (t-1).$$

Then Facts 11 and 12 imply that the matrix $\lim \frac{Q_t^* Q_t}{n}$ is invertible.

From $\mathcal{B}_t$(c), we know that $\frac{(q^r)^* q^s}{n} \stackrel{a.s.}{\to} \bar{\sigma}_s^2$ for $0 \leq r \leq s \leq t$. Using this in (A.117), and via arguments identical to those used to prove (A.88) in $\mathcal{B}_t(f)$, we obtain

$$\lim \frac{\|q^t\|^2}{n} - \frac{(q^t)^* Q_t}{n} \left(\frac{Q_t^* Q_t}{n} \right)^{-1} \frac{Q_t^* q^t}{n} = \bar{\sigma}_t^2 \left(1 - \frac{\bar{\sigma}_t^2}{\bar{\sigma}_{t-1}^2} \right). \tag{A.118}$$

Since $\bar{\sigma}_t^2 = \sigma^2 \left((1 + \mathsf{snr})^{1 - \xi_{t-1}} - 1 \right)$, the definition of ξ_{t-1} in (2.24) implies that the RHS of (A.118) is strictly positive for $t \leq T^* - 1$.

(a) We start with the characterization for h^{t+1} in (A.91) of Lemma 17. The proof from there on is along the same lines as $\mathcal{B}_t$(a), with (H_t, M_t, m^t, Q_{t+1}) replacing (B_t, Q_t, q^t, M_t),

respectively.

(c) From $\mathcal{H}_{t+1}$(a), we have

$$h^{t+1}|_{\mathscr{S}_{t+1,t}} \stackrel{d}{=} \sum_{i=0}^{t-1} \alpha_i h^{i+1} + \tilde{X}^* m_\perp^t + \tilde{Q}_{t+1}\vec{o}_{t+1}(n^{-\delta}) \tag{A.119}$$

where we have used $Q_{t+1}\vec{o}_{t+1}(n^{-\delta}) = \tilde{Q}_{t+1}\vec{o}_{t+1}(n^{-\delta})$. For $r < t, s = t$, we have

$$\frac{(h^{r+1})^* h^{t+1}}{N}\Big|_{\mathscr{S}_{t+1,t}} \stackrel{d}{=} \sum_{i=1}^{t-1} \alpha_i \frac{(h^{r+1})^* h^{i+1}}{N} + \frac{(h^{r+1})^* \tilde{X}^* m_\perp^t}{N} + \sum_{i=0}^{t} o(n^{-\delta}) \frac{(h^{r+1})^* q^i}{N}. \tag{A.120}$$

Applying Fact 9(a), the second term in (A.120) is $\frac{(h^{r+1})^* \tilde{X} q_\perp^t}{N} \stackrel{d}{=} \frac{\|h^{r+1}\| \|m_\perp^t\|}{N} \frac{Z}{\sqrt{n}}$, where $Z \sim \mathcal{N}(0,1)$. Therefore, $\mathcal{H}_t(c)$ and $\mathcal{B}_t$(f) imply that the second term is $o(n^{-\delta})$. The third term is also $o(n^{-\delta})$ since $\mathcal{H}_t$(e) implies that the inner products $\frac{(h^{r+1})^* q^i}{N}$ go to zero. Using $\mathcal{H}_t$(c) and Lemma 16, the first term in (A.120) converges at rate $n^{-\delta}$ to

$$\begin{aligned} \lim \sum_{i=0}^{t-1} \alpha_i \frac{(h^{r+1})^* h^{i+1}}{N} \stackrel{a.s.}{=} \lim \sum_{i=0}^{t-1} \alpha_i \frac{(m^r)^* m^i}{n} = \lim \frac{(m^r)^* m_\parallel^t}{n} = \lim \frac{(m^r)^* m^t}{n} \\ \stackrel{a.s.}{=} \mathbb{E}[(\bar{\sigma}_r \hat{Z}_r - \sigma Z_\epsilon)(\bar{\sigma}_t \hat{Z}_t - \sigma Z_\epsilon)], \end{aligned} \tag{A.121}$$

where the last equality is obtained by applying $\mathcal{B}_t$(b) to $\phi_b(b_i^r, b_i^t, \epsilon_i) = (b_i^r - \epsilon_i)(b_i^t - \epsilon_i) = m_i^r m_i^t$.

For $r = s = t$, using (A.119) we have

$$\begin{aligned} \frac{\|h^{t+1}\|^2}{N}|_{\mathscr{S}_{t+1,t}} \stackrel{d}{=} & \sum_{i=0}^{t-1}\sum_{j=0}^{t-1} \alpha_i \alpha_j \frac{(h^{j+1})^* h^{i+1}}{N} + \frac{\|\tilde{X}^* m_\perp^t\|^2}{N} \\ & + 2\sum_{i=0}^{t-1} \alpha_i \frac{(h^{i+1})^* \tilde{X}^* m_\perp^t}{N} + 2\sum_{i=0}^{t-1} \alpha_i \frac{(h^{i+1})^* Q_{t+1}\vec{o}_{t+1}(n^{-\delta})}{N} \\ & + 2\frac{[Q_{t+1}\vec{o}_{t+1}(n^{-\delta})]^* \tilde{X}^* m_\perp^t}{N} + \frac{\|Q_{t+1}\vec{o}_{t+1}(n^{-\delta})\|^2}{N}. \end{aligned} \tag{A.122}$$

Using arguments similar to those for the $r < t$ case, the last four terms in (A.122) can be shown to be $o(n^{-\delta})$, and by Fact 9(a) $\frac{\tilde{X}^* m_\perp^t}{N} \stackrel{d}{=} \frac{\|m_\perp^t\|}{\sqrt{n}} \frac{Z}{N}$ where $Z \in \mathbb{R}^N$ is i.i.d. standard

normal. Therefore,

$$\lim \frac{\|h^{t+1}\|^2}{N}|_{\mathscr{S}_{t,t}} \overset{a.s.}{=} \lim \sum_{i=0}^{t-1}\sum_{j=0}^{t-1}\alpha_i\alpha_j\frac{(h^{j+1})^*h^{i+1}}{N} + \frac{\|m_\perp^t\|^2}{n}$$
$$\overset{a.s.}{=} \lim \sum_{i=0}^{t-1}\sum_{j=0}^{t-1}\alpha_i\alpha_j\frac{(m^j)^*m^i}{n} + \frac{\|m_\perp^t\|^2}{n} = \lim \frac{\|m_\parallel^t\|^2}{n} + \frac{\|m_\perp^t\|^2}{n} = \lim \frac{\|m^t\|^2}{n},$$

where the second equality is obtained using $\mathcal{H}_t$(c), which together with the central limit theorem also gives the $n^{-\delta}$ rate of convergence.

(b) From $\mathcal{H}_{t+1}$(a), we have

$$h^{t+1}|_{\mathscr{S}_{t+1,t}} \overset{d}{=} \sum_{u=0}^{t-1}\alpha_u h^{u+1} + \tilde{X}^* m_\perp^t + \tilde{Q}_{t+1}\vec{o}_{t+1}(n^{-\delta'}), \tag{A.123}$$

where $\tilde{X}$ is an independent copy of X and the columns of the matrix $\tilde{Q}_{t+1}$ form an orthogonal basis for the columns of Q_{t+1} with $\tilde{Q}_{t+1}^*\tilde{Q}_{t+1} = n\mathsf{I}_{t\times t}$. We therefore have

$$\phi_h\left(\sum_{u=0}^{t} a_u h_\ell^{u+1}, \sum_{v=0}^{t} b_v h_\ell^{v+1}, \beta_{0_\ell}\right)\Bigg|_{\mathscr{S}_{t+1,t}} \overset{d}{=} \phi_h(h_\ell + \Delta_\ell, \tilde{h}_\ell + \tilde{\Delta}_\ell, \beta_{0_\ell}), \tag{A.124}$$

where $h_\ell = \sum_{u=0}^{t-1}(a_u + a_t\alpha_u)h_\ell^{u+1} + a_t[\tilde{X}^* m_\perp^t]_\ell$ and $\Delta_\ell = a_t[\tilde{Q}_{t+1}\vec{o}_{t+1}(n^{-\delta'})]_\ell$. Similarly define $\tilde{h}_\ell$ and $\tilde{\Delta}_\ell$, with the b_v's replacing the a_u's. Note that for each $r \geq 0$, we have $\|q_\ell^r\| \leq c\sqrt{nP_\ell} = \Theta(\sqrt{\log M})$. Therefore, $\max_{j\in[M]}|\Delta_{\ell_j}| = \Theta(n^{-\delta'}\sqrt{\log M})$ for $\ell \in [L]$. Using this, it is shown in [35] that for each of the functions in (2.40), we have

$$\frac{1}{L}\sum_{\ell=1}^{L}\left|\phi_h(h_\ell + \Delta_\ell, \tilde{h}_\ell + \tilde{\Delta}_\ell, \beta_{0_\ell}) - \phi_h(h_\ell, \tilde{h}_\ell, \beta_{0_\ell})\right| \overset{(a)}{=} o(n^{-\delta'}\log M). \tag{A.125}$$

for some $\delta' > 0$. Consequently, by choosing $\delta \in (0, \delta')$ we can drop the $[\tilde{Q}_{t+1}\vec{o}_{t+1}(n^{-\delta'})]_\ell$ terms. In what follows, we use the notation $h_\ell[\tilde{X}] = \sum_{u=0}^{t-1}(a_u + a_t\alpha_u)h_\ell^{u+1} + a_t[\tilde{X}^* m_\perp^t]_\ell$ and $\tilde{h}_\ell[\tilde{X}] = \sum_{v=0}^{t-1}(b_v + b_t\alpha_v)h_\ell^{v+1} + b_t[\tilde{X}^* m_\perp^t]_\ell$, making explicit the dependence on $\tilde{X}$. We now appeal to Fact 2 to show that

$$\lim n^\delta\left[\frac{1}{L}\sum_{\ell=1}^{L}\phi_h\left(h_\ell[\tilde{X}], \tilde{h}_\ell[\tilde{X}], \beta_{0_\ell}\right) - \frac{1}{L}\sum_{\ell=1}^{L}\mathbb{E}_{\tilde{X}}\left\{\phi_h\left(h_\ell[\tilde{X}], \tilde{h}_\ell[\tilde{X}], \beta_{0_\ell}\right)\right\}\right] \overset{a.s.}{=} 0 \tag{A.126}$$

To invoke Fact 2 (conditionally on $\mathscr{S}_{t+1,t}$), we need to verify that

$$\frac{1}{L}\sum_{\ell=1}^{L}\mathbb{E}_{\hat{X}}\left|n^{\delta}\phi_h\left(h_\ell[\hat{X}],\tilde{h}_\ell[\hat{X}],\beta_{0_\ell}\right)-\mathbb{E}_{\tilde{X}}\left\{n^{\delta}\phi_h\left(h_\ell[\tilde{X}],\tilde{h}_\ell[\tilde{X}],\beta_{0_\ell}\right)\right\}\right|^{2+\kappa}\leq cL^{\kappa/2} \quad \text{(A.127)}$$

for constants $\kappa\in(0,1)$. In (A.127), $\hat{X},\tilde{X}$ are i.i.d. copies of X. In [35], it is shown that for each function in (2.40),

$$\mathbb{E}_{\hat{X},\tilde{X}}\left|\phi_h\left(h_\ell[\hat{X}],\tilde{h}_\ell[\hat{X}],\beta_{0_\ell}\right)-\phi_h\left(h_\ell[\tilde{X}],\tilde{h}_\ell[\tilde{X}],\beta_{0_\ell}\right)\right|^{2+\kappa}\overset{a.s.}{=}O((\log M)^{2+\kappa}),\quad \ell\in[L]. \tag{A.128}$$

Due to Jensen's inequality, the bound in (A.128) implies that (A.127) holds if δ is chosen such that $\delta(2+\kappa)<\kappa/2$. Hence (A.126) holds.

Recalling that $[\tilde{X}^* m_\perp^t]\overset{d}{=}(\|m_\perp^t\|/\sqrt{n})Z$ where $Z\sim\mathcal{N}(0,\mathsf{I}_{N\times N})$, we have

$$\mathbb{E}_{\tilde{X}}\left\{\phi_h\left(h_\ell[\tilde{X}],\tilde{h}_\ell[\tilde{X}],\beta_{0_\ell}\right)\right\}=\underbrace{\mathbb{E}_Z\left[\phi_h\left(\sum_{u=0}^{t-1}a'_u h_\ell^{u+1}+a_t\frac{\|m_\perp^t\|}{\sqrt{n}}Z_\ell,\ \sum_{v=0}^{t-1}b'_v h_\ell^{v+1}+b_t\frac{\|m_\perp^t\|}{\sqrt{n}}Z_\ell,\ \beta_{0_\ell}\right)\right]}_{\phi_h^{new}\left(\sum_{u=0}^{t-1}a'_u h_\ell^{u+1},\sum_{v=0}^{t-1}b'_v h_\ell^{v+1},\beta_{0_\ell}\right)} \tag{A.129}$$

where we have defined $a'_u=(a_u+a_t\alpha_u)$ and $b'_v=(b_v+b_t\alpha_v)$. Using Jensen's inequality, it can be shown that the induction hypothesis $\mathcal{H}_t(b)$ holds for the function ϕ_h^{new} whenever $\mathcal{H}_t(b)$ holds for the function ϕ_h inside the expectation defining ϕ_h^{new} in (A.129). We therefore have

$$\lim n^{\delta}\left[\frac{1}{L}\sum_{\ell=1}^{L}\phi_h^{new}\left(\sum_{u=0}^{t-1}a'_u h_\ell^{u+1},\sum_{v=0}^{t-1}b'_v h_\ell^{v+1},\beta_{0_\ell}\right)-\frac{1}{L}\sum_{\ell=1}^{L}\mathbb{E}\left[\phi_h^{new}\left(\sum_{u=0}^{t-1}a'_u\bar{\tau}_u Z_{u_\ell},\sum_{v=0}^{t-1}b'_v\bar{\tau}_v Z_{v_\ell},\beta_\ell\right)\right]\right]$$
$$\overset{a.s.}{=}0. \tag{A.130}$$

It is then shown in [35] that

$$\lim \frac{n^{\delta}}{L} \sum_{\ell=1}^{L} \Bigg| \overbrace{\mathbb{E}\, \mathbb{E}_Z \left\{ \phi_h \left(\sum_{u=0}^{t-1} a'_u \bar{\tau}_u Z_{u_\ell} + a_t \frac{\|m_\perp^t\|}{\sqrt{n}} Z_\ell, \sum_{v=0}^{t-1} b'_v \bar{\tau}_v Z_{v_\ell} + b_t \frac{\|m_\perp^t\|}{\sqrt{n}} Z_\ell, \beta_\ell \right) \right\}}^{\phi_h^{new}(\sum_{u=0}^{t-1} a'_u \bar{\tau}_u Z_{u_\ell}, \sum_{v=0}^{t-1} b'_v \bar{\tau}_v Z_{v_\ell}, \beta_\ell)} \\ - \mathbb{E}\mathbb{E}_Z \left\{ \phi_h \left(\sum_{u=0}^{t-1} a'_u \bar{\tau}_u Z_{u_\ell} + a_t \zeta_t Z_\ell, \sum_{v=0}^{t-1} b'_v \bar{\tau}_v Z_{v_\ell} + b_t \zeta_t Z_\ell, \beta_\ell \right) \right\} \Bigg| \overset{a.s.}{=} 0 \tag{A.131}$$

where ζ_t is the limit of $\frac{\|m_\perp^t\|}{\sqrt{n}}$. That ζ_t is well-defined and finite can be seen as follows.

$$\frac{\|m_\perp^t\|^2}{n} = \frac{\|m^t\|^2}{n} - \frac{\|m_\parallel^t\|^2}{n} = \frac{\|m^t\|^2}{n} - \sum_{i=1}^{t-1} \sum_{i'=1}^{t-1} \alpha_u \alpha_{i'} \frac{(m^i)^* m^{i'}}{n}. \tag{A.132}$$

Each of the terms in (A.132) converges to a finite limit at rate $n^{-\delta}$ by $\mathcal{H}_{t+1}(c)$ and Lemma 16. Using the definitions $a'_u = (a_u + a_t \alpha_u)$ and $b'_v = (b_v + b_t \alpha_v)$, we have for $\ell \in [L]$

$$\begin{aligned} &\phi_h \left(\sum_{u=0}^{t-1} a'_u \bar{\tau}_u Z_{u_\ell} + a_t \zeta_t Z_\ell, \sum_{v=0}^{t-1} b'_v \bar{\tau}_v Z_{v_\ell} + b_t \zeta_t Z_\ell, \beta_\ell \right) \\ &= \phi_h \left(\sum_{u=0}^{t-1} a_u \bar{\tau}_u Z_{u_\ell} + a_t (\sum_{u=0}^{t-1} \alpha_u Z_{u_\ell} + \zeta_t Z_\ell), \sum_{v=0}^{t-1} b_v \bar{\tau}_v Z_{v_\ell} + b_t (\sum_{v=0}^{t-1} \alpha_v Z_{v_\ell} + \zeta_t Z_\ell), \beta_\ell \right). \end{aligned} \tag{A.133}$$

Thus the proof is complete if we show that the i.i.d. entries of the Gaussian random vector $\sum_{u=0}^{t-1} \alpha_u Z_u + \zeta_t Z$ have variance $\bar{\tau}_t^2$. To see this, apply the proof thus far (from (A.124) – (A.133)) to the function $\phi_h(h_\ell, \tilde{h}_\ell, \beta_\ell) = \frac{(h_\ell)^* \tilde{h}_\ell}{M}$ with $a_t = b_t = 1$ and $a_u = b_u = 0$ for $0 \le u \le (t-1)$. We thus obtain

$$\lim n^{\delta} \left[\frac{\|h^{t+1}\|^2}{N} - \frac{1}{L} \sum_{\ell=1}^{L} \frac{\mathbb{E} \|\sum_{u=0}^{t-1} \alpha_u \bar{\tau}_u Z_{u_\ell} + \gamma_t Z_\ell\|^2}{M} \right] \overset{a.s}{=} 0. \tag{A.134}$$

Further, since $\sum_{u=0}^{t-1} \alpha_u Z_u + \zeta_t Z$ has i.i.d. entries, $\frac{1}{ML} \sum_{\ell=1}^{L} \mathbb{E} \|\sum_{u=0}^{t-1} \alpha_u \bar{\tau}_u Z_{u_\ell} + \gamma_t Z_\ell\|^2$ equals $\mathbb{E} \left(\sum_{u=0}^{t-1} \alpha_u \bar{\tau}_u Z_{u_i} + \gamma_t Z_i \right)^2$ for any $i \in [N]$. On the other hand, from $\mathcal{H}_{t+1}$(c) we know that $\lim \frac{\|h^{t+1}\|^2}{N} \overset{a.s.}{=} \lim \frac{\|m^t\|^2}{n} \overset{a.s.}{=} \mathbb{E} \left(\bar{\sigma}_t \hat{Z}_t - \sigma Z_\epsilon \right)^2$, all at rate $o(n^{-\delta})$. The result follows since $\mathbb{E} \left(\bar{\sigma}_t \hat{Z}_t - \sigma Z_\epsilon \right)^2 = \bar{\sigma}_t^2 + \sigma^2 = \bar{\tau}_t^2$.

(d) By definition $q^{s+1} = \eta^s(\beta_0 - h^{s+1}) - \beta_0$, and hence

$$\frac{(h^{r+1})^* q^{s+1}}{n} = \frac{1}{n}\sum_{\ell=1}^{L} \phi_h(h_\ell^{r+1}, h_\ell^{s+1}, \beta_{0_\ell})$$

for $\phi_h : \mathbb{R}^M \times \mathbb{R}^M \times \mathbb{R}^M \to \mathbb{R}$ defined as $\phi_h(h_\ell^{r+1}, h_\ell^{s+1}, \beta_{0_\ell}) = (h_\ell^{r+1})^*[\eta_\ell^s(\beta_0 - h^{s+1}) - \beta_{0_\ell}]$. Applying $\mathcal{H}_{t+1}$(b) to ϕ_h yields

$$\lim n^\delta \left[\frac{1}{n}\sum_{\ell=1}^{L} \phi_h(h_\ell^{r+1}, h_\ell^{s+1}, \beta_{0_\ell}) - \lim \frac{1}{n}\sum_{\ell=1}^{L} \mathbb{E}\{\bar{\tau}_r Z_{r_\ell}^*[\eta_\ell^s(\beta_0 - \bar{\tau}_s Z_s) - \beta_{0_\ell}]\}\right] \overset{a.s.}{=} 0. \tag{A.135}$$

Using arguments very similar to those in $\mathcal{H}_1$(d) (iterated expectations and Stein's lemma), we obtain that

$$\mathbb{E}\{\bar{\tau}_r Z_{r_\ell}^*[\eta_\ell^s(\beta_0 - \bar{\tau}_s Z_s) - \beta_{0_\ell}]\} = \frac{\bar{\tau}_r}{\bar{\tau}_s}\mathbb{E}[Z_{r_1} Z_{s_1}]\left(\mathbb{E}\|\eta_\ell^s(\beta - \bar{\tau}_s Z_s)\|^2 - nP_\ell\right), \quad \ell \in [L]. \tag{A.136}$$

Here Z_{r_1}, Z_{s_1} refer to the first entries of the vectors Z_r, Z_s, respectively. Thus (A.135) becomes

$$\lim n^\delta \left[\frac{1}{n}\sum_{\ell=1}^{L} \phi_h(h_\ell^{r+1}, h_\ell^{s+1}, \beta_{0_\ell}) - \lim \frac{\bar{\tau}_r}{\bar{\tau}_s}\mathbb{E}[Z_{r_1} Z_{s_1}]\left(\frac{\mathbb{E}\|\eta^s(\beta - \bar{\tau}_s Z_s)\|^2}{n} - P\right)\right] \overset{a.s.}{=} 0. \tag{A.137}$$

From (2.31), we observe that

$$\lambda_{s+1} = \frac{1}{\bar{\tau}_s^2}\left(\frac{\|\beta^{s+1}\|^2}{n} - P\right) \overset{a.s.}{\to} \lim \frac{1}{\bar{\tau}_s^2}\left(\frac{\mathbb{E}\left\{\|\eta^s(\beta - \bar{\tau}_s Z_s)\|^2\right\}}{n} - P\right) = \frac{-\bar{\sigma}_{s+1}^2}{\bar{\sigma}_s^2 + \sigma^2}, \tag{A.138}$$

where the convergence at rate $n^{-\delta}$ follows from $\mathcal{H}_{t+1}$(b) applied to the function $\frac{\|\eta^s(\beta - h^{s+1})\|^2}{n} = \frac{\|\beta^{s+1}\|^2}{n}$. The last equality in (A.138) holds because $\left(P - \frac{\mathbb{E}\left\{\|\eta^s(\beta_0 - \bar{\tau}_s Z_s)\|^2\right\}}{n}\right) \to \bar{\sigma}_{s+1}^2$ (cf. Appendix A.3.3). Considering (A.138) and (A.137), what remains to be shown is

$$\bar{\tau}_r \bar{\tau}_s \mathbb{E}[Z_{r_1} Z_{s_1}] \overset{a.s.}{=} \lim \frac{(m^r)^* m^s}{n} \overset{a.s.}{=} \mathbb{E}[(\bar{\sigma}_r \hat{Z}_r - \sigma Z_\epsilon)(\bar{\sigma}_s \hat{Z}_s - \sigma Z_\epsilon)]. \tag{A.139}$$

The second equality above is due to $\mathcal{H}_{t+1}(c)$, which also says that $\lim \frac{(m^r)^* m^s}{n} \overset{a.s.}{=} \lim \frac{(h^{r+1})^*(h^{s+1})}{N}$. Then the first equality in (A.139) is obtained by applying $\mathcal{H}_{t+1}$(b) to the function $(h_\ell^{r+1})^* h_\ell^{s+1}$

to see that

$$\lim \frac{(h^{r+1})^*(h^{s+1})}{N} \overset{a.s.}{=} \bar{\tau}_r \bar{\tau}_s \mathbb{E}[Z_{r_1} Z_{s_1}].$$

(e) By $\mathcal{H}_{t+1}$ part (a),

$$\frac{(q^0)^* h^{t+1}}{n}|_{\mathscr{S}_{t+1,t}} \overset{d}{=} \sum_{i=0}^{t-1} \alpha_i \frac{(q^0)^* h^{i+1}}{n} + \frac{(q^0)^* \tilde{X}^* m_\perp^t}{n} + \frac{(q^0)^* \tilde{Q}_{t+1} \vec{o}_{t+1}(n^{-\delta})}{n}. \tag{A.140}$$

We argue that each term on the RHS approaches 0 almost surely with rate $n^{-\delta}$. This is true for the first term by the induction hypothesis $\mathcal{H}_t$(e) and Lemma 16. Next, Fact 9(a) implies that $\frac{(q^0)^* \tilde{X}^* m_\perp^t}{n} \overset{d}{=} \frac{\|q^0\|}{\sqrt{n}} \frac{\|m_\perp^t\|}{\sqrt{n}} \frac{Z}{\sqrt{n}}$ where $Z \sim \mathcal{N}(0,1)$. Thus the second term in (A.140) approaches 0 almost surely with rate $n^{-\delta}$ since $\|q^0\|/\sqrt{n} = \sqrt{P}$ and $\lim \|m_\perp^t\|/\sqrt{n}$ is a constant by $\mathcal{H}_t$(f). For the third term, the result holds because $\frac{(q^0)^* q^r}{n}$ converges to a constant for $r = 0, \ldots, t$, due to $\mathcal{B}_t$(c).

A.3.3 Limit of $\frac{1}{n}\mathbb{E}\{[\eta^r(\beta - \bar{\tau}_r Z_r) - \beta]^*[\eta^s(\beta - \bar{\tau}_s Z_s) - \beta]\}$

Since $\|\beta\|^2 = nP$, the required limit is

$$\lim \frac{1}{n}\mathbb{E}\{[\eta^r(\beta - \bar{\tau}_r Z_r)]^*[\eta^s(\beta - \bar{\tau}_s Z_s)]\} - \frac{1}{n}\mathbb{E}\{\beta^* \eta^r(\beta - \bar{\tau}_r Z_r)\} - \frac{1}{n}\mathbb{E}\{\beta^* \eta^s(\beta - \bar{\tau}_s Z_s)\} + P. \tag{A.141}$$

For $r \leq s$, we prove that the limit in (A.141) equals $\bar{\sigma}_{s+1}^2 = \sigma^2\left((1+\mathsf{snr})^{1-\xi_s} - 1\right)$ by showing the following:

$$\lim \frac{1}{n}\mathbb{E}\{\beta^* \eta^r(\beta - \bar{\tau}_r Z_r)\} = \sigma^2\left((1+\mathsf{snr}) - (1+\mathsf{snr})^{1-\xi_r}\right), \tag{A.142}$$

$$\frac{1}{n}\mathbb{E}\{\|\eta^r(\beta - \bar{\tau}_r Z_r)]\|^2\} = \frac{1}{n}\mathbb{E}\{\beta^* \eta^r(\beta - \bar{\tau}_r Z_r)\}, \tag{A.143}$$

$$\lim \frac{1}{n}\mathbb{E}\{[\eta^r(\beta - \bar{\tau}_r Z_r)]^*[\eta^s(\beta - \bar{\tau}_s Z_s)]\} = \lim \frac{1}{n}\mathbb{E}\{\beta^* \eta^r(\beta - \bar{\tau}_r Z_r)\}, \quad \text{for } r < s. \tag{A.144}$$

Since β is distributed uniformly over the set $\mathcal{B}_{M,L}$, the expectation in (A.142) can be computed by assuming that β has a non-zero in the first entry of each section. Thus

$$\lim \frac{1}{n}\mathbb{E}\{\beta^* \eta^r(\beta - \bar{\tau}_r Z_r)\} = \lim \sum_{l=1}^{L} P_\ell \, \mathbb{E}\left[\frac{\exp\left(\frac{nP_\ell}{\bar{\tau}_r^2}\right)\exp\left(\frac{\sqrt{nP_\ell}}{\bar{\tau}_r}\, U_1^\ell\right)}{\exp\left(\frac{nP_\ell}{\bar{\tau}_r^2}\right)\exp\left(\frac{\sqrt{nP_\ell}}{\bar{\tau}_r}\, U_1^\ell\right) + \sum_{j=2}^{M}\exp\left(\frac{\sqrt{nP_\ell}}{\bar{\tau}_r} U_j^\ell\right)}\right]$$
$$\stackrel{(a)}{=} \sum_{l=1}^{L} P_\ell \mathbf{1}\{c_\ell > 2(\ln 2) R\bar{\tau}_r^2\} \stackrel{(b)}{=} \sigma^2\left((1+\mathsf{snr}) - (1+\mathsf{snr})^{1-\xi_r}\right). \tag{A.145}$$

In (A.145), $\{U_j^\ell\}$ with $\ell \in [L], j \in [M]$ is just a relabeled version of $-Z_r$, and is thus i.i.d. $\mathcal{N}(0,1)$. The equality (a) is obtained from (A.7) and (A.8) in Appendix A.2, noting that $c_\ell = \lim LP_\ell$ while (b) follows from Lemmas 13 and 1 (cf. (A.5) and (2.22)).

Since $\beta^{r+1}(s) = \eta^r(s)$, (A.143) was proved in Proposition 2.5.1 (cf. (A.2) and (A.3)). Next, from the Cauchy-Schwarz inequality, we have

$$\frac{1}{n}\mathbb{E}\{(\eta^r(\beta - \bar{\tau}_r Z_r))^* \eta^s(\beta - \bar{\tau}_s Z_s)\} \le \frac{1}{n}\sum_{l=1}^{L}\left(\mathbb{E}\{\|\eta_\ell^r(\beta_\ell - \bar{\tau}_r Z_{r_\ell})\|^2\}\,\mathbb{E}\{\|\eta_\ell^s(\beta_\ell - \bar{\tau}_s Z_{s_\ell})\|^2\}\right)^{1/2}$$
$$\stackrel{(a)}{=} \sum_{\ell} P_\ell \, \mathbf{1}\{c_\ell > 2(\ln 2) R\bar{\tau}_r^2\}\, \mathbf{1}\{c_\ell > 2(\ln 2) R\bar{\tau}_s^2\} \stackrel{(b)}{=} \sum_{\ell} P_\ell \, \mathbf{1}\{c_\ell > 2(\ln 2) R\bar{\tau}_r^2\}, \tag{A.146}$$

where (a) follows from (A.143) and (A.145), and (b) holds because $\bar{\tau}_r^2 > \bar{\tau}_s^2$ since $r < s$.

Since β is distributed uniformly over the set $\mathcal{B}_{M,L}$, the expectation $\mathbb{E}\{[\eta_\ell^r(\beta_\ell - \bar{\tau}_r Z_{r_\ell})]^*[\eta_\ell^s(\beta_\ell - \bar{\tau}_s Z_{s_\ell})]\}$ can be computed by assuming that β has a non-zero in the first entry of each section:

$$\frac{1}{n}\mathbb{E}\{(\eta^r(\beta - \bar{\tau}_r Z_r))^* \eta^s(\beta - \bar{\tau}_s Z_s)\} = \frac{1}{n}\sum_{\ell}\mathbb{E}\{[\eta_\ell^r(\beta_\ell - \bar{\tau}_r Z_{r_\ell})]^*[\eta_\ell^s(\beta_\ell - \bar{\tau}_s Z_{s_\ell})]\} = \sum_{\ell} P_\ell \, \mathcal{E}_{rs,\ell} \tag{A.147}$$

where

$$\mathcal{E}_{rs,\ell} = \mathbb{E}\Bigg[\frac{\exp\left(c_{r,\ell}^2\right)\exp\left(c_{r,\ell}\,U_{r1}^{\ell}\right)}{\exp\left(c_{r,\ell}^2\right)\exp\left(c_{r,\ell}\,U_{r1}^{\ell}\right)+\sum_{j=2}^{M}\exp\left(c_{r,\ell}U_{rj}^{\ell}\right)}\cdot\frac{\exp\left(c_{s,\ell}^2\right)\exp\left(c_{s,\ell}\,U_{s1}^{\ell}\right)}{\exp\left(c_{s,\ell}^2\right)\exp\left(c_{s,\ell}\,U_{s1}^{\ell}\right)+\sum_{j=2}^{M}\exp\left(c_{s,\ell}U_{sj}^{\ell}\right)}$$
$$+\sum_{i=2}^{M}\frac{\exp\left(c_{r,\ell}^2\,U_{ri}^{\ell}\right)}{\exp\left(c_{r,\ell}^2\right)\exp\left(c_{r,\ell}\,U_{r1}^{\ell}\right)+\sum_{j=2}^{M}\exp\left(c_{r,\ell}U_{rj}^{\ell}\right)}\cdot\frac{\exp\left(c_{s,\ell}\,U_{si}^{\ell}\right)}{\exp\left(\frac{nP_\ell}{\bar{\tau}_s^2}\right)\exp\left(c_{s,\ell}\,U_{s1}^{\ell}\right)+\sum_{j=2}^{M}\exp\left(c_{s,\ell}U_{sj}^{\ell}\right)}\Bigg], \tag{A.148}$$

with $c_{r,\ell}=\frac{\sqrt{nP_\ell}}{\bar{\tau}_r}$ and $c_{s,\ell}=\frac{\sqrt{nP_\ell}}{\bar{\tau}_s}$. In (A.148), the pairs of random variables $\{(U_{rj}^{\ell},U_{sj}^{\ell})\}$, $j\in[M]$ are i.i.d. across index j, and for each j, U_{rj}^{ℓ} and U_{sj}^{ℓ} are jointly Gaussian with $\mathcal{N}(0,1)$ marginals.

The expectation of the first term on the right-hand side of (A.148) can be written as

$$\mathbb{E}\Bigg[\mathbb{E}\Bigg[\left(\frac{\exp(c_{r,\ell}U_{r1}^{\ell})}{\exp(c_{r,\ell}U_{r1}^{\ell})+\exp(-c_{r,\ell}^2)\sum_{j=2}^{M}\exp(c_{r,\ell}U_{rj}^{\ell})}\right)\left(\frac{\exp(c_{s,\ell}U_{s1}^{\ell})}{\exp(c_{s,\ell}U_{s1}^{\ell})+\exp(-c_{s,\ell}^2)\sum_{j=2}^{M}\exp(c_{s,\ell}U_{sj}^{\ell})}\right)\Big|\;U_{r1}^{\ell},U_{s1}^{\ell}\Bigg]\Bigg]$$
$$\overset{(a)}{\geq}\mathbb{E}\left[\left(\frac{\exp(c_{r,\ell}U_{r1}^{\ell})}{\exp(c_{r,\ell}U_{r1}^{\ell})+M\exp(-\frac{c_{r,\ell}^2}{2})}\right)\left(\frac{\exp(c_{s,\ell}U_{s1}^{\ell})}{\exp(c_{s,\ell}U_{s1}^{\ell})+M\exp(-\frac{c_{s,\ell}^2}{2})}\right)\right]$$
$$=\mathbb{E}\left[\left(1+M\exp(-\frac{c_{r,\ell}^2}{2})\exp(-c_{r,\ell}U_{r1}^{\ell})\right)^{-1}\left(1+M\exp(-\frac{c_{s,\ell}^2}{2})\exp(-c_{s,\ell}U_{s1}^{\ell})\right)^{-1}\right]$$
$$\geq P\left(U_{r1}^{\ell}>-\sqrt{c_{r,\ell}}\right)P\left(U_{s1}^{\ell}>-\sqrt{c_{s,\ell}}\right)\cdot\left(1+M\exp\left(-\frac{c_{r,\ell}^2}{2}-c_{r,\ell}^{\frac{3}{2}}\right)\right)^{-1}\left(1+M\exp\left(-\frac{c_{s,\ell}^2}{2}-c_{s,\ell}^{\frac{3}{2}}\right)\right)^{-1}$$
$$\overset{(b)}{\longrightarrow}1\;\text{ as }\;M\to\infty\;\text{ if }\;\lim_{M\to\infty}\frac{c_{r,\ell}^2}{2}\cdot\frac{1}{\ln M}>1. \tag{A.149}$$

In (A.149), (a) is obtained as follows. The inner expectation on the first line of the form $\mathbb{E}_{X,Y}[f(X,Y)]$ with $f(X,Y)=\frac{\kappa_1}{\kappa_1+X}\cdot\frac{\kappa_2}{\kappa_2+Y}$, where κ_1,κ_2 are positive constants. Since f is a convex function of (X,Y), Jensen's inequality implies $\mathbb{E}[f(X,Y)]\geq f(\mathbb{E}X,\mathbb{E}Y)$, with $\mathbb{E}[\exp(c_{r,\ell}U_{r1}^{\ell})]=\exp(\frac{c_{r,\ell}^2}{2})$.

Since $\mathcal{E}_{rs,\ell}$ in (A.148) lies in $[0,1]$, (A.149) implies that

$$\lim\mathcal{E}_{rs,\ell}=1\;\text{ if }\;\lim_{M\to\infty}\frac{c_{r,\ell}^2}{2}\cdot\frac{1}{\ln M}=\frac{c_\ell}{2R\bar{\tau}_r^2\ln 2}>1. \tag{A.150}$$

where we have used $nR=L\log M$ and $c_{r,\ell}=\frac{\sqrt{nP_\ell}}{\bar{\tau}_r}$ noting that $c_\ell:=\lim LP_\ell$. Using this

in (A.147), we conclude that $\frac{1}{n}\mathbb{E}\{(\eta^r(\beta - \bar{\tau}_r Z_r))^* \eta^s(\beta - \bar{\tau}_s Z_s)\} \geq \sum_\ell P_\ell \mathbf{1}\{c_\ell > 2(\ln 2) R\bar{\tau}_r^2\}$. Together with the upper bound in (A.146), this proves (A.144), and hence completes the proof.

Appendix B

Chapter 3 Appendix

B.1 Mathematical Preliminaries

We first list some results that will be used in the proof of Lemmas 4 and 5 which are given below. Some of these can be found in [6, Section III.G], but we summarize them here for completeness.

Fact 9. *Let $u \in \mathbb{R}^N$ be a deterministic vector such that $\|u\|^2/n$ is finite. Let $\tilde{X} \in \mathbb{R}^{n\times N}$ be a matrix with independent $\mathcal{N}(0, 1/n)$ entries. Let $\mathcal{W}$ be a d-dimensional subspace of $\mathbb{R}^n$ for $d \leq n$. Let $(w_1, ..., w_d)$ be an orthogonal basis of $\mathcal{W}$ with $\|w_i\|^2 = n$ for $i \in [d]$, and let $\mathsf{P}_{\mathcal{W}}$ denote the orthogonal projection operator onto $\mathcal{W}$. Then for $D = [w_1 \mid \ldots \mid w_d]$, we have $\mathsf{P}_{\mathcal{W}}\tilde{X}u \stackrel{d}{=} \frac{\|u\|}{\sqrt{n}}Dx$ where $x \in \mathbb{R}^d$ is a random vector with i.i.d. $\mathcal{N}(0, 1/n)$ entries.*

Fact 10 (Stein's lemma)**.** *For zero-mean jointly Gaussian random variables Z_1, Z_2, and any function $f : \mathbb{R} \to \mathbb{R}$ for which $\mathbb{E}[Z_1 f(Z_2)]$ and $\mathbb{E}[f'(Z_2)]$ both exist, we have $\mathbb{E}[Z_1 f(Z_2)] = \mathbb{E}[Z_1 Z_2]\mathbb{E}[f'(Z_2)]$.*

Fact 11. *Let $v_1, \ldots, v_t$ be a sequence of vectors in $\mathbb{R}^n$ such that for $i \in [t]$*

$$\frac{1}{n}\|v_i - \mathsf{P}_{i-1}(v_i)\|^2 \geq c,$$

where c is a positive constant and P_{i-1} is the orthogonal projection onto the span of $v_1, \ldots, v_{i-1}$. Then the matrix $C \in \mathbb{R}^{t\times t}$ with $C_{ij} = v_i^* v_j/n$ has minimum eigenvalue $\lambda_{\min} \geq c'$, where c' is a strictly positive constant (depending only on c and t).

Fact 12. *Let $\{S_n\}_{n\geq 1}$ be a sequence of $t \times t$ matrices such that $\lim_{n\to\infty} S_n = S_\infty$ where the limit is element-wise. Then if $\liminf_{n\to\infty} \lambda_{\min}(S_n) \geq c$ for a positive constant c, then $\lambda_{\min}(S_\infty) \geq c$.*

Fact 13. *Let $Z_1, \ldots, Z_t$ be jointly Gaussian random variables with zero mean and an invertible covariance matrix C. Then*

$$Var(Z_t \mid Z_1, \ldots, Z_{t-1}) = \mathbb{E}[Z_t^2] - u^* C^{-1} u,$$

where for $i \in [t-1]$, $u_i = \mathbb{E}[Z_t Z_i]$.

Fact 14. *Let $Z_1, \ldots, Z_t$ be jointly Gaussian random variables such that for all $i \in [t]$,*

$$\mathbb{E}[Z_i^2] \leq K \quad \text{and} \quad Var(Z_i \mid Z_1, \ldots, Z_{i-1}) \geq c_i,$$

for some strictly positive constants $K, c_1, \ldots, c_t$. Let Y be a random variable defined on the same probability space, and let $g : \mathbb{R}^2 \to \mathbb{R}$ be a Lipschitz function with $z \to g(z, Y)$ non-constant with positive probability. Then there exists a positive constant c'_t such that

$$\mathbb{E}[(g(Z_t, Y))^2] - u^* C^{-1} u > c'_t,$$

where $u \in \mathbb{R}^{t-1}$ and $C \in \mathbb{R}^{(t-1)\times(t-1)}$ are given by

$$u_i = \mathbb{E}[g(Z_t, Y) g(Z_i, Y)],\ C_{ij} = \mathbb{E}[g(Z_i, Y) g(Z_j, Y)],\ i, j \in [t-1].$$

(The constant c'_t depends only on the K, the random variable Y and the function g.)

B.2 Distributional Properties of Key Ingredients

Given two random vectors X, Y and a sigma-algebra $\mathscr{S}$, $X|_{\mathscr{S}} \stackrel{d}{=} Y$ implies that the conditional distribution of X given $\mathscr{S}$ equals the distribution of Y. The $t \times t$ identity matrix is denoted by I_t, and the $t \times s$ all-zero matrix is denoted by $\mathsf{0}_{t\times s}$.

Define $\mathscr{S}_{t_1,t_2}$ to be the sigma-algebra generated by

$$b^0, ..., b^{t_1-1}, m^0, ..., m^{t_1-1}, h^1, ..., h^{t_2}, q^0, ..., q^{t_2}, \text{ and } \beta_0, \epsilon.$$

A key ingredient in the proof is the distribution of X conditioned on the sigma algebra $\mathscr{S}_{t_1,t}$ where t_1 is either $t+1$ or t.

Recall matrices M_t, B_t, Q_t and H_t defined in (3.13). Additionally define matrices $\Lambda_t := \text{diag}(\lambda_0, \ldots, \lambda_{t-1})$,

$$A_t := [h^1 + q^0 \mid h^2 + q^1 \mid \ldots \mid h^t + q^{t-1}], \quad Y_t := [b^0 \mid b^1 + \lambda_1 m^0 \mid \ldots \mid b^{t-1} + \lambda_{t-1} m^{t-2}]. \tag{B.1}$$

where A_0, Y_0 and λ_0 is the all-zero vector. From the definitions (3.13), (B.1), and (3.11), it follows

$$A_t = X^* M_t, \quad Y_t = X Q_t. \tag{B.2}$$

Observing that conditioning on $\mathscr{S}_{t_1,t}$ is equivalent to conditioning on the linear constraints $A_t = X^* M_t$ and $Y_{t_1} = X Q_{t_1}$, the following lemma from [6] specifies the conditional distribution $X|_{\mathscr{S}_{t_1,t}}$. [1]

Lemma 18. *[6, Lemma 10] For $t_1 = t+1$ or t, the conditional distribution of the random matrix A given $\mathscr{S}_{t_1,t}$ satisfies*

$$X|_{\mathscr{S}_{t_1,t}} \stackrel{d}{=} \mathbb{E}_{t_1,t} + \mathsf{P}^{\perp}_{M_t} \tilde{X} \mathsf{P}^{\perp}_{Q_{t_1}}.$$

Here $\tilde{X} \stackrel{d}{=} X$ is random matrix independent of $\mathscr{S}_{t_1,t}$, and $\mathsf{P}^{\perp}_{M_t} = \mathsf{I} - \mathsf{P}_{M_t}$ where $\mathsf{P}_{M_t} = M_t(M_t^* M_t)^{-1} M_t^*$ is the orthogonal projection matrix onto the column space of M_t; similarly,

1. While conditioning on the linear constraints, we emphasize that only X is treated as random.

$\mathsf{P}^{\perp}_{Q_{t_1}} = \mathsf{I} - \mathsf{P}_{Q_{t_1}}$, *where* $\mathsf{P}_{Q_{t_1}} = Q_{t_1}(Q^*_{t_1}Q_{t_1})^{-1}Q^*_{t_1}$. *The matrix* $\mathbb{E}_{t_1,t} = \mathbb{E}[X|\mathscr{S}_{t_1,t}]$ *is given by*

$$\begin{aligned}\mathbb{E}_{t_1,t} &= \mathbb{E}[X\mathsf{P}_{Q_{t_1}} + \mathsf{P}_{M_t}X\mathsf{P}^{\perp}_{Q_{t_1}} \mid XQ_{t_1} = Y_{t_1},\ X^*M_t = A_t] \\ &= Y_{t_1}(Q^*_{t_1}Q_{t_1})^{-1}Q^*_{t_1} + M_t(M^*_tM_t)^{-1}A^*_t - M_t(M^*_tM_t)^{-1}M^*_tY_{t_1}(Q^*_{t_1}Q_{t_1})^{-1}Q^*_{t_1}.\end{aligned} \tag{B.3}$$

Lemma 19. *[6, Lemma 12] For the matrix* $\mathbb{E}_{t_1,t}$ *defined in Lemma 18, the following hold:*

$$\mathbb{E}^*_{t+1,t}m^t = A_t(M^*_tM_t)^{-1}M^*_tm^t_{\|} + Q_{t+1}(Q^*_{t+1}Q_{t+1})^{-1}Y^*_{t+1}m^t_{\perp}, \tag{B.4}$$

$$\mathbb{E}_{t,t}q^t = Y_t(Q^*_tQ_t)^{-1}Q^*_tq^t_{\|} + M_t(M^*_tM_t)^{-1}A^*_tq^t_{\perp}, \tag{B.5}$$

where $m^t_{\|}, m^t_{\perp}, q^t_{\|}, q^t_{\perp}$ *are defined in* (3.15) *and* (3.16).

We mention that Lemmas 4, 18, and 19 can be applied only when $M^*_tM_t$ and $Q^*_{t_1}Q_{t_1}$ are invertible.

B.3 Proof of Lemma 3

We prove this by induction. First we show the result for $t = 0$. By assumption $\sigma_0^2 > 0$ and therefore $\tau_0^2 = \sigma^2 + \sigma_0^2 > 0$. Moreover, $C^0 = \sigma_0^2 > 0$ and $\sigma^2 + C^0 = \tau_0^2 > 0$ and so both are invertible.

Recall from (3.22), $(\sigma_t^{\perp})^2 = E_{t,t} - E_t^*(C^t)^{-1}E_t$ and $(\tau_t^{\perp})^2 = (\sigma^2 + E_{t,t}) - (\sigma^2 + E_t)^*(\sigma^2 + C^t)^{-1}(\sigma^2 + E_t)$. By Fact 14, both are greater than some positive constant if

$$Var(\tau_i\tilde{Z}_i|\tau_1\tilde{Z}_1, \ldots, \tau_{i-1}\tilde{Z}_{i-1}) > c_i$$

for some strictly positive constant c_i for all $i \in [t-1]$. By Fact 13,

$$Var(\tau_i\tilde{Z}_i|\tau_1\tilde{Z}_1, \ldots, \tau_{i-1}\tilde{Z}_{i-1}) = \tau_i^2 - (\hat{\alpha}^i)^*(\sigma^2 + E_i) = (\tau_i^{\perp})^2 > c_i.$$

The inequality in the above follows by inductive hypotheses $(\tau_i^{\perp})^2 > 0$ for $i \in [t-1]$.

Now we show C^t is invertible. First note that we can view the matrix $C^t \in \mathbb{R}^{t\times t}$ as follows

$$C^t = \begin{bmatrix} \mathbf{M}_1 & \mathbf{M}_2 \\ \mathbf{M}_3 & \mathbf{M}_4 \end{bmatrix}$$

where $\mathbf{M}_1 = C^{t-1} \in \mathbb{R}^{(t-1)\times(t-1)}$, $\mathbf{M}_4 = \sigma_{t-1}^2$, and $\mathbf{M}_2^* = \mathbf{M}_3 = E_{t-1}^* \in \mathbb{R}^{1\times(t-1)}$ defined in (3.21). Then using blockwise inversion, C^t is invertible if C^{t-1} is invertible, which is true by the indicative hypothesis, and if

$$\sigma_{t-1}^2 - E_{t-1}^*(C^{t-1})^{-1}E_{t-1} = (\sigma_{t-1}^{\perp})^2 > 0,$$

which is also true by the inductive hypothesis. Showing that $\sigma^2 + C^t$ is invertible is very similar.

B.4 Proof of Lemma 4

We demonstratie result (B.10) and result (B.9) can be shown similarly. By (A.36) it follows

$$b^0|_{\mathscr{S}_{0,0}} \overset{d}{=} Xq_{\perp}^0 \overset{d}{=} \frac{\|q^0\|}{\sqrt{n}} Z_0',$$

where $Z_0' \in \mathbb{R}^n$ is an i.i.d. standard Gaussian random vector, independent of $\mathscr{S}_{0,0}$.

For the case $t \geq 1$, we use Lemmas 18 and 19 to write

$$\begin{aligned} b^t|_{\mathscr{S}_{t,t}} &= (Xq^t - \lambda_t m^{t-1})|_{\mathscr{S}_{t,t}} \overset{\mathrm{d}}{=} Y_t(Q_t^*Q_t)^{-1}Q_t^*q_{\|}^t + M_t(M_t^*M_t)^{-1}A_t^*q_{\perp}^t + \mathsf{P}_{M_t}^{\perp}\tilde{X}q_{\perp}^t - \lambda_t m^{t-1} \\ &= B_t(Q_t^*Q_t)^{-1}Q_t^*q_{\|}^t + [0|M_{t-1}]\Lambda_t(Q_t^*Q_t)^{-1}Q_t^*q_{\|}^t + M_t(M_t^*M_t)^{-1}H_t^*q_{\perp}^t + \mathsf{P}_{M_t}^{\perp}\tilde{X}q_{\perp}^t - \lambda_t m^{t-1}. \end{aligned}$$

The last equality above is obtained using $Y_t = B_t + [0|M_{t-1}]\Lambda_t$, and $A_t = H_t + Q_t$. It follows,

$$b^t|_{\mathscr{S}_{t,t}} \overset{d}{=} \sum_{i=0}^{t-1} \gamma_i^t b^i + (\mathsf{I} - \mathsf{P}_{M_t})\tilde{X}q_{\perp}^t + [0|M_{t-1}]\Lambda_t(Q_t^*Q_t)^{-1}Q_t^*q_{\|}^t + M_t(M_t^*M_t)^{-1}H_t^*q_{\perp}^t - \lambda_t m^{t-1}, \tag{B.6}$$

by noticing that $\mathsf{P}_{M_t}^{\perp}\tilde{X}q_{\perp}^t = (\mathsf{I}-\mathsf{P}_{M_t})\tilde{X}q_{\perp}^t$ and $B_t(Q_t^*Q_t)^{-1}Q_t^*q_{\|}^t = \sum_{i=0}^{t-1}\gamma_i^t b^i$. Using Fact 9,

$$(\mathsf{I}-\mathsf{P}_{M_t}^{\|})\tilde{X}q_{\perp}^t \stackrel{d}{=} \frac{\|q_{\perp}^t\|}{\sqrt{n}}(\mathsf{I}-\mathsf{P}_{M_t}^{\|})Z_t' \stackrel{d}{=} \frac{\|q_{\perp}^t\|}{\sqrt{n}}Z_t' - \frac{\|q_{\perp}^t\|\tilde{M}_t\bar{Z}_t'}{n}, \tag{B.7}$$

where $Z_t' \in \mathbb{R}^n$ and $\bar{Z}_t' \in \mathbb{R}^t$ are random vectors with i.i.d. $\mathcal{N}(0,1)$ entries, and $\tilde{M}_t$ forms an orthogonal basis for the column space of M_t such that $\tilde{M}_t^*\tilde{M}_t = n\mathsf{I}_t$. Notice that Z_t' and $\bar{Z}_t'$ are independent of $\epsilon, (Z_0', Z_1', \ldots, Z_{t-1}')$ and $(Z_0, Z_1, \ldots, Z_{t-1})$ since $\tilde{X}$ is independent of X and ϵ. Now using (B.6) and (B.7),

$$\begin{aligned} &b^t|_{\mathscr{S}_{t,t}} \\ &\stackrel{d}{=} \frac{\|q_{\perp}^t\|}{\sqrt{n}}Z_t' - \frac{\|q_{\perp}^t\|\tilde{M}_t\bar{Z}_t'}{n} + \sum_{i=0}^{t-1}\gamma_i^t b^i + [0|M_{t-1}]\Lambda_t(Q_t^*Q_t)^{-1}Q_t^*q_{\|}^t + M_t(M_t^*M_t)^{-1}H_t^*q_{\perp}^t - \lambda_t m^{t-1}, \end{aligned}$$

Note that all values in the above except for the random parts, Z_t' and $\bar{Z}_t'$, are in the conditioning sigma-field. Equivalently to the above we write

$$b^t|_{\mathscr{S}_{t,t}} \stackrel{d}{=} \sum_{r=0}^{t-1}\hat{\gamma}_r^t b^r + \sigma_t^{\perp}Z_t' + \Delta_{t,t},$$

where

$$\begin{aligned} \Delta_{t,t} = &\sum_{r=0}^{t-1}(\gamma_r^t - \hat{\gamma}_r^t)b^r \;+\; Z_t'\left(\frac{\|q_{\perp}^t\|}{\sqrt{n}} \;-\; \sigma_t^{\perp}\right) \;-\; \frac{\|q_{\perp}^t\|\tilde{M}_t\bar{Z}_t'}{n} \\ &+\; [0|M_{t-1}]\Lambda_t(Q_t^*Q_t)^{-1}Q_t^*q_{\|}^t \;+\; M_t(M_t^*M_t)^{-1}H_t^*q_{\perp}^t \;-\; \lambda_t m^{t-1}. \end{aligned} \tag{B.8}$$

Showing that (B.8) equals (3.25) requires demonstrating

$$[0|M_{t-1}]\Lambda_t(Q_t^*Q_t)^{-1}Q_t^*q_{\|}^t + M_t\left(\frac{M_t^*M_t}{n}\right)^{-1}\frac{M_t^*}{n}\left(\lambda_t m^{t-1} - \sum_{i=0}^{t-2}\lambda_{i+1}\gamma_{i+1}^t m^i\right) - \lambda_t m^{t-1} = 0.$$

To see that this is true, notice that the above is a linear combination of the vectors $(m^0, \ldots, m^t)$. Consider the three terms of the above separately, which we label $T_a - T_c$. Now

$$T_a = \sum_{k=0}^{t-2}\lambda_{k+1}\gamma_{k+1}^t m^k.$$

Similarly, using the fact that for $0 \leq i \leq t-1$ the $t-$length vector $\left(M_t^* M_t\right)^{-1} M_t^* m^i$ equals $\underline{e}_{i+1}$, the vector of 0's with a single 1 in the $(i+1)$ position, it follows:

$$T_b = \lambda_t m^{t-1} - \sum_{i=0}^{t-2} \lambda_{i+1} \gamma_{i+1}^t m^i.$$

From the above, it is clear that $T_a + T_b + T_c = 0$. This completes the proof of (B.10).

B.5 Proof of Lemma 5

Below we label the results of Lemma 2, all of which we will prove.

(a)

$$\mathbf{Pr}\left(\frac{\|\Delta_{t+1,t}\|^2}{N} \geq \Delta\right) \leq K e^{-\kappa_t n \Delta}, \tag{B.9}$$

$$\mathbf{Pr}\left(\frac{\|\Delta_{t,t}\|^2}{n} \geq \Delta\right) \leq K e^{-\kappa_t n \Delta}. \tag{B.10}$$

(b)

$$\frac{(h^{t+1})^* q^0}{n} \doteq 0, \tag{B.11}$$

$$\frac{(b^t)^* \epsilon}{n} \doteq 0, \quad \frac{(m^t)^* \epsilon}{n} \doteq -\sigma^2. \tag{B.12}$$

(c) For all $0 \leq r \leq t$,

$$\frac{(h^{r+1})^* h^{t+1}}{N} \doteq \left(\sigma^2 + E_{r,t}\right), \tag{B.13}$$

$$\frac{(b^r)^* b^t}{n} \doteq E_{r,t}. \tag{B.14}$$

(d) *i)* For pseudo-Lipschitz functions $\phi_h : \mathbb{R}^{t+2} \to \mathbb{R}$

$$\frac{1}{N} \sum_{i=1}^{N} \phi_h\left(h_i^1, \ldots, h_i^{t+1}, \beta_{0_i}\right) \doteq \mathbb{E}\left[\phi_h\left(\tau_0 \tilde{Z}_0, \ldots, \tau_t \tilde{Z}_t, \beta\right)\right]. \tag{B.15}$$

The random variables $\tilde{Z}_0, \ldots, \tilde{Z}_t$ are jointly Gaussian with zero mean and covariance

given by (3.17).

ii) Let $\psi_h : \mathbb{R} \to \mathbb{R}$ be a bounded function that is almost everywhere differentiable, with bounded derivative where it exists. Then for finite constants $(a_0, \ldots, a_t)$,

$$\frac{1}{N}\sum_{i=1}^{N} \psi_h(\beta_{0_i} - \sum_{r=0}^{t} a_r h_i^{r+1}) \doteq \mathbb{E}\left[\psi_h(\beta - \sum_{r=0}^{t} a_r \tau_r \tilde{Z}_r)\right]. \tag{B.16}$$

(e) For all $0 \leq r \leq t$,

$$\frac{(q^0)^* q^{t+1}}{n} \doteq E_{0,t+1}, \quad \frac{(q^{r+1})^* q^{t+1}}{n} \doteq E_{r+1,t+1}, \tag{B.17}$$

$$\frac{(b^r)^* m^t}{n} \doteq E_{r,t}, \quad \frac{(b^t)^* m^r}{n} \doteq E_{r,t}. \tag{B.18}$$

(f) Define $\hat{\lambda}_{t+1} = -\frac{1}{\delta}\mathbb{E}[\eta_t'(\beta - \tau_t \tilde{Z}_t)]$. For all $0 \leq r \leq t$,

$$\frac{(h^{t+1})^* q^{r+1}}{n} \doteq \hat{\lambda}_{r+1}\left(\sigma^2 + E_{r,t}\right), \quad \frac{(h^{r+1})^* q^{t+1}}{n} \doteq \hat{\lambda}_{t+1}\left(\sigma^2 + E_{r,t}\right), \tag{B.19}$$

$$\frac{(m^r)^* m^t}{n} \doteq \left(\sigma^2 + E_{r,t}\right). \tag{B.20}$$

(g) For $0 \leq k \leq t$ and $0 \leq k' \leq t-1$ (when $t \geq 1$),

$$\gamma_k^{t+1} \doteq \hat{\gamma}_k^{t+1}, \tag{B.21}$$

$$\alpha_{k'}^t \doteq \hat{\alpha}_{k'}^t, \tag{B.22}$$

where $\hat{\gamma}_k^{t+1}$ and $\hat{\alpha}_{k'}^t$ are defined in (3.20).

(h)

$$\frac{\|q_\perp^{t+1}\|^2}{n} \doteq (\sigma_{t+1}^\perp)^2, \tag{B.23}$$

$$\frac{\|m_\perp^t\|^2}{n} \doteq (\tau_t^\perp)^2, \tag{B.24}$$

where $\sigma_{t+1}^\perp, \tau_t^\perp$ are defined in (3.22).

We now prove Lemma 2. The proof proceeds by induction on t. We label as $\mathcal{H}^{t+1}$ the results (B.9), (B.11), (B.13), (B.15), (B.16), (B.17), (B.19), (B.21), (B.23) and similarly as $\mathcal{B}^t$ the results (B.10), (B.12), (B.14), (B.18), (B.20), (B.22), (B.24). The proof consists of four steps:

1. $\mathcal{B}_0$ holds.
2. $\mathcal{H}_1$ holds.
3. If $\mathcal{B}_r, \mathcal{H}_s$ holds for all $r < t$ and $s \leq t$, then $\mathcal{B}_t$ holds.
4. if $\mathcal{B}_r, \mathcal{H}_s$ holds for all $r \leq t$ and $s \leq t$, then $\mathcal{H}_{t+1}$ holds.

Step 1: Showing $\mathcal{B}_0$ holds

We wish to show results (a) - (h) in (B.10), (B.12), (B.14), (B.18), (B.20), (B.22), (B.24).

(a)

$$\begin{aligned}\mathbf{Pr}\left(\frac{\|\Delta_{0,0}\|^2}{n} \geq \Delta\right) &\overset{(a)}{\leq} \mathbf{Pr}\left(\left|\frac{\|q^0\|}{\sqrt{n}} - \sigma_0^{\perp}\right| \geq \sqrt{\frac{\Delta}{2}}\right) + \mathbf{Pr}\left(\left|\frac{\|Z_0'\|}{\sqrt{n}} - 1\right| \geq \sqrt{\frac{\Delta}{2}}\right) \\ &\overset{(b)}{\leq} e^{-\kappa n \Delta} + e^{-\kappa n \Delta}.\end{aligned}$$

Step (a) follows from the definition of $\Delta_{0,0}$ in Lemma 4 (3.23) and Lemma 24, and step (b) from the sub-Gaussian assumption on p_β, Lemma 25, and Lemma 37.

(b) We first show concentration of $(b^0)^*\epsilon/n$.

$$\begin{aligned}\mathbf{Pr}\left(\left|\frac{(b^0)^*\epsilon}{n}\right| \geq \Delta\right) &= \mathbf{Pr}\left(\sigma_0^{\perp}\frac{|\epsilon^* Z_0'|}{n} + \frac{|w^*\Delta_{0,0}|}{n} \geq \Delta\right) \\ &\leq \mathbf{Pr}\left(\frac{|\epsilon^* Z_0'|}{n} \geq \frac{\Delta}{2\sigma_0^{\perp}}\right) + \mathbf{Pr}\left(\frac{|\epsilon^*\Delta_{0,0}|}{n} \geq \frac{\Delta}{2}\right).\end{aligned}$$

The above follows from the conditional distribution of b^0 stated in Lemma 4 (B.10) and Lemma 23. Label the terms on the right side of the above as T_1 and T_2. To complete the proof we show that each is upper bounded by $e^{-\kappa n \Delta^2}$. For Z independent standard

normal,

$$T_1 \overset{(a)}{=} \mathbf{Pr}\left(\left|\frac{\|\epsilon\|}{\sqrt{n}} \cdot \frac{Z}{\sqrt{n}}\right| \geq \frac{\Delta}{2\sigma_0^{\perp}}\right) \overset{(b)}{\leq} \mathbf{Pr}\left(\frac{|Z|}{\sqrt{n}} \geq \frac{\Delta}{4\sigma\sigma_0^{\perp}}\right) + \mathbf{Pr}\left(\left|\frac{\|\epsilon\|}{\sqrt{n}} - \sigma\right| \geq \frac{\Delta}{4\sigma\sigma_0^{\perp}}\right)$$
$$\overset{(c)}{\leq} 2e^{-\frac{n\Delta^2}{32\sigma^2(\sigma_0^{\perp})^2}} + e^{-\kappa n\Delta^2}.$$

Step (a) follows since ϵ is independent of Z_0', step (b) from Lemma 24, and step (c) from Lemma 36, assumed concentration of the noise, and Lemma 25. Next,

$$T_2 \overset{(a)}{=} \mathbf{Pr}\left(\frac{\|\epsilon\|}{\sqrt{n}} \cdot \frac{\|\Delta_{0,0}\|}{\sqrt{n}} \geq \frac{\Delta}{2}\right) \overset{(b)}{\leq} \mathbf{Pr}\left(\frac{\|\Delta_{0,0}\|}{\sqrt{n}} \geq \frac{\Delta}{4\sigma}\right) + \mathbf{Pr}\left(\left|\frac{\|\epsilon\|}{\sqrt{n}} - \sigma\right| \geq \frac{\Delta}{4\sigma}\right)$$
$$\overset{(c)}{\leq} Ke^{-\kappa n\Delta^2} + e^{-\kappa n\Delta^2}.$$

Step (a) follows by Cauchy-Schwarz, step (b) from Lemma 24, and step (c) from $\mathcal{B}_0(a)$, assumed concentration of the noise, and Lemma 25.

Next we show concentration of $(m^0)^*\epsilon/n$.

$$\mathbf{Pr}\left(\left|\frac{(m^0)^*\epsilon}{n} + \sigma^2\right| \geq \Delta\right) \overset{(a)}{\leq} \mathbf{Pr}\left(\left|\frac{(b^0)^*\epsilon}{n}\right| \geq \frac{\Delta}{2}\right) + \mathbf{Pr}\left(\left|\frac{\|\epsilon\|^2}{n} - \sigma^2\right| \geq \frac{\Delta}{2}\right)$$
$$\overset{(b)}{\leq} Ke^{-\kappa n\Delta^2} + e^{-\kappa n\Delta^2}.$$

Step (a) follows since $m^0 = b^0 - \epsilon$ and from Lemma 23, and step (b) from the work above above and assumed concentration of the noise.

(c)

$$\mathbf{Pr}\left(\left|\frac{\|b^0\|^2}{n} - \sigma_0^2\right| \geq \epsilon\right)$$
$$\overset{(a)}{\leq} \mathbf{Pr}\left(\left|\frac{\|Z_0'\|^2}{n} - 1\right| \geq \frac{\Delta}{3\sigma_0^2}\right) + \mathbf{Pr}\left(\frac{\|\Delta_{0,0}\|^2}{n} \geq \frac{\epsilon}{3}\right) + \mathbf{Pr}\left(\frac{|(Z_0')^*\Delta_{0,0}|}{n} \geq \frac{\Delta}{3\sigma_0}\right)$$
$$\overset{(b)}{\leq} e^{-\kappa n\Delta^2} + Ke^{-\kappa n\Delta^2} + Ke^{-\kappa n\Delta^2}.$$

Step (a) follows from the conditional distribution of b^0 stated in Lemma 4 (B.10) and

Lemma 23, and step (b) from Lemma 37, $\mathcal{B}_0(a)$, and the following:

$$\mathbf{Pr}\left(\frac{|(Z_0')^*\Delta_{0,0}|}{n} \geq \frac{\Delta}{3\sigma_0}\right) \overset{(a)}{\leq} \mathbf{Pr}\left(\left|\frac{\|Z_0'\|}{\sqrt{n}} - 1\right| \geq \frac{\Delta}{6\sigma_0}\right) + \mathbf{Pr}\left(\frac{\|\Delta_{0,0}\|}{\sqrt{n}} \geq \frac{\Delta}{6\sigma_0}\right)$$
$$\overset{(b)}{\leq} e^{-\kappa n\Delta^2} + Ke^{-\kappa n\Delta^2}.$$

Step (a) follows from Cauchy-Schwarz and Lemma 24, and step (b) from Lemma 37, Lemma 25, and $\mathcal{B}_0(a)$.

(d) Nothing for $\mathcal{B}$ steps.

(e)

$$\mathbf{Pr}\left(\left|\frac{(b^0)^*m^0}{n} - \sigma_0^2\right| \geq \Delta\right) \overset{(a)}{\leq} \mathbf{Pr}\left(\left|\frac{\|b^0\|^2}{n} - \sigma_0^2\right| \geq \frac{\Delta}{2}\right) + \mathbf{Pr}\left(\left|\frac{(b^0)^*\epsilon}{n}\right| \geq \frac{\Delta}{2}\right)$$
$$\overset{(b)}{\leq} Ke^{-\kappa n\Delta^2} + Ke^{-\kappa n\Delta^2}.$$

Step (a) follows since $m^0 = b^0 - \epsilon$ and from Lemma 23, and step (b) from $\mathcal{B}_0(c)$ and $\mathcal{B}_0(b)$.

(f)

$$\mathbf{Pr}\left(\left|\frac{\|m^0\|^2}{n} - \tau_0^2\right| \geq \Delta\right) \overset{(a)}{\leq} \mathbf{Pr}\left(\left|\frac{(m^0)^*b^0}{n} - \sigma_0^2\right| \geq \frac{\Delta}{2}\right) + \mathbf{Pr}\left(\left|\frac{(m^0)^*\epsilon}{n} + \sigma^2\right| \geq \frac{\Delta}{2}\right)$$
$$\overset{(b)}{\leq} Ke^{-\kappa n\Delta^2} + Ke^{-\kappa n\Delta^2}.$$

Step (a) follows since $m^0 = b^0 - \epsilon$ and from Lemma 23 and step (b) from $\mathcal{B}_0(e)$ and $\mathcal{B}_0(b)$.

(g) Nothing to prove.

(h) Since $\|m_\perp^0\| = \|m^0\|$ and $(\tau_0^\perp)^2 = \tau_0^2$, this result is equivalent to $\mathcal{B}_0(f)$.

Step 1: Showing $\mathcal{H}_1$ holds

We wish to show results (a) - (h) in (B.9), (B.11), (B.13), (B.15), (B.16), (B.17), (B.19), (B.21), (B.23).

(a)

$$
\begin{aligned}
&\mathbf{Pr}\left(\frac{\|\Delta_{1,0}\|^2}{N} \geq \Delta\right) \\
&\overset{(a)}{\leq} \mathbf{Pr}\left(3\left(\frac{\|m^0\|}{\sqrt{n}} - \tau_0\right)^2 \frac{\|Z_0\|^2}{N} + 3\delta \cdot \frac{\|m^0\|^2}{n} \cdot \frac{(\bar{Z}_0')^2}{n} + \frac{3}{N} \cdot \left(\frac{(b^0)^* m^0}{\|q^0\|} - \|q^0\|\right)^2 \geq \Delta\right) \\
&\overset{(b)}{\leq} \mathbf{Pr}\left(\left|\frac{\|m^0\|}{\sqrt{n}} - \tau_0\right| \frac{\|Z_0\|}{\sqrt{N}} \geq \sqrt{\frac{\Delta}{9}}\right) + \mathbf{Pr}\left(\frac{\|m^0\|}{\sqrt{n}} \cdot \frac{|\bar{Z}_0'|}{\sqrt{n}} \geq \sqrt{\frac{\Delta}{9\delta}}\right) \\
&\qquad + \mathbf{Pr}\left(\left|\frac{(b^0)^* m^0}{\sqrt{n}\|q^0\|} - \frac{\|q^0\|}{\sqrt{n}}\right| \geq \sqrt{\frac{\Delta}{9\delta}}\right).
\end{aligned}
$$

Step (a) follows from the definition of $\Delta_{0,0}$ in Lemma 4 (3.24) and Lemma 38 and step (b) from Lemma 23. Label the terms in (b) as $T_1 - T_3$. To complete the proof, we show that each is upper bounded by $Ke^{-\kappa n\Delta}$. Term $T_1 \leq Ke^{-\kappa n\Delta}$ using Lemma 24, Lemma 25, result $\mathcal{B}_0(f)$, and Lemma 37. Next, $T_2 \leq Ke^{-\kappa n\Delta}$ using Lemma 24, Lemma 25, result $\mathcal{B}_0(f)$, and Lemma 36. Finally,

$$
\begin{aligned}
T_3 &\overset{(a)}{\leq} \mathbf{Pr}\left(\left|\frac{(b^0)^* m^0}{n} \cdot \frac{\sqrt{n}}{\|q^0\|} - \sigma_0\right| \geq \frac{1}{2}\sqrt{\frac{\Delta}{9\delta}}\right) + \mathbf{Pr}\left(\left|\frac{\|q^0\|}{\sqrt{n}} - \sigma_0\right| \geq \frac{1}{2}\sqrt{\frac{\Delta}{9\delta}}\right) \\
&\overset{(b)}{\leq} \mathbf{Pr}\left(\left|\frac{(b^0)^* m^0}{n} - \sigma_0^2\right| \geq \frac{1}{4(\sigma_0^2 + 1/\sigma_0)}\sqrt{\frac{\Delta}{9\delta}}\right) \\
&\qquad + \mathbf{Pr}\left(\left|\frac{\sqrt{n}}{\|q^0\|} - \frac{1}{\sigma_0}\right| \geq \frac{1}{4(\sigma_0^2 + 1/\sigma_0)}\sqrt{\frac{\Delta}{9\delta}}\right) + \mathbf{Pr}\left(\left|\frac{\|q^0\|}{\sqrt{n}} - \sigma_0\right| \geq \frac{1}{2}\sqrt{\frac{\Delta}{9\delta}}\right) \\
&\overset{(c)}{\leq} Ke^{-\kappa n\Delta} + e^{-\kappa n\Delta} + e^{-\kappa n\Delta}.
\end{aligned}
$$

Step (a) follows from Lemma 23, step (b) from Lemma 24, and step (c) from $\mathcal{B}_0(e)$, the sub-Gaussian assumption on p_β, and Lemma 29.

(b)

$$
\begin{aligned}
&\mathbf{Pr}\left(\left|\frac{(h^1)^* q^0}{n}\right| \geq \Delta\right) \\
&\quad = \mathbf{Pr}\left(\left|\frac{\tau_0 Z_0^* q^0}{n} + \frac{\Delta_{1,0}^* q^0}{n}\right| \geq \Delta\right) \leq \mathbf{Pr}\left(\left|\frac{\tau_0 Z_0^* q^0}{n}\right| \geq \frac{\Delta}{2}\right) + \mathbf{Pr}\left(\left|\frac{\Delta_{1,0}^* q^0}{n}\right| \geq \frac{\Delta n}{2}\right).
\end{aligned}
$$

The above follows from the conditional distribution of h^1 stated in Lemma 4 (B.9) and Lemma 23. Label the two terms on the right side of the above as T_1 and T_2. To complete the proof we show each is upper bounded by $e^{-\kappa n\Delta^2}$. Let Z be an independent standard Gaussian random variable.

$$T_1 \overset{(a)}{=} \mathbf{Pr}\left(\frac{\|q^0\|}{\sqrt{n}}\frac{|Z|}{\sqrt{n}} \geq \frac{\Delta}{2\tau_0}\right) \overset{(b)}{\leq} \mathbf{Pr}\left(\left|\frac{\|q^0\|}{\sqrt{n}} - \sigma_0\right| \geq \frac{\Delta}{4\tau_0\sigma_0}\right) + \mathbf{Pr}\left(\frac{|Z|}{\sqrt{n}} \geq \frac{\Delta}{4\tau_0\sigma_0}\right)$$
$$\overset{(c)}{\leq} e^{-\kappa n\Delta^2} + 2e^{-\frac{n\Delta^2}{32\tau_0^2\sigma_0^2}}.$$

Step (a) follows since q^0 is independent of Z_0, step (b) follows from Lemma 24, and step (c) from the sub-Gaussian assumption on p_β, Lemma 25 and Lemma 36. Finally,

$$T_2 \overset{(a)}{\leq} \mathbf{Pr}\left(\frac{\|q^0\|}{\sqrt{n}} \cdot \frac{\|\Delta_{1,0}\|}{\sqrt{n}} \geq \frac{\Delta}{2}\right) \overset{(b)}{\leq} \mathbf{Pr}\left(\left|\frac{\|q^0\|}{\sqrt{n}} - \sigma_0\right| \geq \frac{\Delta}{4\sigma_0}\right) + \mathbf{Pr}\left(\frac{\|\Delta_{1,0}\|}{\sqrt{n}} \geq \frac{\Delta}{4\sigma_0}\right)$$
$$\overset{(c)}{\leq} e^{-\kappa n\Delta^2} + Ke^{-\kappa n\Delta^2}.$$

Step (a) using Cauchy-Schwarz, step (b) follows from Lemma 24, and step (c) from the sub-Gaussian assumption on p_β, Lemma 25, and $\mathcal{H}_1(a)$.

(c) Using the conditional distribution of h^1 from Lemma 4 (B.9), the proof is similar to that of $\mathcal{B}_0(c)$.

(d) We first demonstrate (B.15).

$$\begin{aligned}
&\mathbf{Pr}\left(\left|\frac{1}{N}\sum_{i=1}^{N}\phi_h\left(h_i^1, \beta_{0_i}\right) - \mathbb{E}\left[\phi_h\left(\tau_0\tilde{Z}_0, \beta\right)\right]\right| \geq \Delta\right) \\
&\overset{(a)}{=} \mathbf{Pr}\left(\left|\frac{1}{N}\sum_{i=1}^{N}\phi_h\left(\tau_0 Z_{0_i} + [\Delta_{1,0}]_i, \beta_{0_i}\right) - \mathbb{E}\left[\phi_h\left(\tau_0\tilde{Z}_0, \beta\right)\right]\right| \geq \Delta\right) \\
&\overset{(b)}{=} \mathbf{Pr}\left(\left|\frac{1}{N}\sum_{i=1}^{N}\phi_h\left(\tau_0 Z_{0_i}, \beta_{0_i}\right) - \mathbb{E}\left[\phi_h\left(\tau_0\tilde{Z}_0, \beta\right)\right]\right| \geq \frac{\Delta}{2}\right) \\
&\quad + \mathbf{Pr}\left(\left|\frac{1}{N}\sum_{i=1}^{N}\phi_h\left(\tau_0 Z_{0_i} + [\Delta_{1,0}]_i, \beta_{0_i}\right) - \frac{1}{N}\sum_{i=1}^{N}\phi_h\left(\tau_0 Z_{0_i}, \beta_{0_i}\right)\right| \geq \frac{\Delta}{2}\right).
\end{aligned}$$

Step (a) uses the conditional distribution of h^1 given in Lemma 4 (B.9) and step (b) follows from Lemma 23. Label the terms in (b) as T_1 and T_2. To complete the proof

we show that both terms are upper bounded by $Ke^{-\kappa n\Delta^2}$.

First consider T_1.

$$\begin{aligned} T_1 &\overset{(a)}{\leq} \mathbf{Pr}\left(\left|\frac{1}{N}\sum_{i=1}^{N}\phi_h\left(\tau_0 Z_{0_i},\beta_{0_i}\right) - \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{Z_{0_i}}\left[\phi_h\left(\tau_0 Z_{0_i},\beta_{0_i}\right)\right]\right| \geq \frac{\Delta}{4}\right) \\ &\quad + \mathbf{Pr}\left(\left|\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{Z_{0_i}}\left[\phi_h\left(\tau_0 Z_{0_i},\beta_{0_i}\right)\right] - \mathbb{E}_{\tilde{Z}_0,\beta}\left[\phi_h\left(\tau_0\tilde{Z}_0,\beta\right)\right]\right| \geq \frac{\Delta}{4}\right) \\ &\overset{(b)}{\leq} Ke^{-\kappa n\Delta^2} + Ke^{-\kappa n\Delta^2}. \end{aligned}$$

Step (a) follows from Lemma 23 and step (b) from Lemma 34 and Lemma 35 since by Lemma 32, $\tilde{\phi}_h : \mathbb{R} \to \mathbb{R}$ defined as

$$\tilde{\phi}_h(s) := \mathbb{E}_{Z_0}\left[\phi_h\left(\tau_0 Z_0, s\right)\right] \in PL(2).$$

Next consider T_2.

$$\begin{aligned} T_2 &\overset{(a)}{\leq} \mathbf{Pr}\left(\frac{1}{N}\sum_{i=1}^{N}\left|\phi_h\left(\tau_0 Z_{0_i} + [\Delta_{1,0}]_i,\beta_{0_i}\right) - \phi_h\left(\tau_0 Z_{0_i},\beta_{0_i}\right)\right| \geq \frac{\Delta}{2}\right) \\ &\overset{(b)}{\leq} \mathbf{Pr}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{L}\left(1 + \left|\tau_0 Z_{0_i} + [\Delta_{1,0}]_i\right| + \left|\tau_0 Z_{0_i}\right|\right)\left|[\Delta_{1,0}]_i\right| \geq \frac{\Delta}{2}\right) \\ &\overset{(c)}{\leq} \mathbf{Pr}\left(\frac{\|\Delta_{1,0}\|}{\sqrt{N}}\cdot\left(1 + \frac{\|\Delta_{1,0}\|}{\sqrt{N}} + 2\tau_0\frac{\|Z_0\|}{\sqrt{N}}\right) \geq \frac{\Delta}{2\sqrt{3}\mathbf{L}}\right). \end{aligned} \tag{B.25}$$

Step (a) follows from the Triangle Inequality, step (b) from the fact that conditional on $\mathscr{S}_{1,0}$, the functions $\tilde{\phi}_{h,i} : \mathbb{R} \to \mathbb{R}$, for each $i \in [N]$, defined as $\tilde{\phi}_{h,i}(s) := \phi_h(s, \beta_{0_i}) \in PL(2)$, and step (c) from Cauchy-Schwarz and the following application of Lemma 38:

$$\begin{aligned} \sum_{i=1}^{N}\frac{\left(1 + \left|[\Delta_{1,0}]_i\right| + 2\left|\tau_0 Z_{0_i}\right|\right)^2}{N} &\leq 3\left(1 + \frac{\|\Delta_{1,0}\|^2}{N} + 4\tau_0^2\frac{\|Z_0\|^2}{N}\right) \\ &\leq 3\left(1 + \frac{\|\Delta_{1,0}\|}{\sqrt{N}} + 2\tau_0\frac{\|Z_0\|}{\sqrt{N}}\right)^2. \end{aligned}$$

Next,

$$T_2 \overset{(a)}{\leq} \mathbf{Pr}\left(\frac{\|\Delta_{1,0}\|}{\sqrt{N}} \geq \frac{\Delta}{6\sqrt{3}\mathbf{L}}\right) + \mathbf{Pr}\left(\frac{\|\Delta_{1,0}\|^2}{N} \geq \frac{\Delta}{6\sqrt{3}\mathbf{L}}\right) + \mathbf{Pr}\left(\frac{\|Z_0\|}{\sqrt{N}} \cdot \frac{\|\Delta_{1,0}\|}{\sqrt{N}} \geq \frac{\Delta}{12\sqrt{3}\tau_0\mathbf{L}}\right)$$
$$\overset{(b)}{\leq} Ke^{-\kappa n\Delta^2} + Ke^{-\kappa n\Delta^2} + Ke^{-\kappa n\Delta^2}.$$

Step (a) follows from (B.25) and Lemma 23, and step (b) from $\mathcal{H}_1(a)$ and the following fact:

$$\mathbf{Pr}\left(\frac{\|Z_0\|}{\sqrt{N}} \cdot \frac{\|\Delta_{1,0}\|}{\sqrt{N}} \geq \Delta\right) \overset{(a)}{\leq} \mathbf{Pr}\left(\frac{\|\Delta_{1,0}\|}{\sqrt{N}} \geq \frac{\Delta}{2}\right) + \mathbf{Pr}\left(\left|\frac{\|Z_0\|}{\sqrt{N}} - 1\right| \geq \frac{\Delta}{2}\right)$$
$$\overset{(b)}{\leq} Ke^{-\kappa n\Delta^2} + Ke^{-\kappa n\Delta^2}.$$

Step (a) follows from Lemma 24 and step (b) from $\mathcal{H}_1(a)$, Lemma 37, and Lemma 25. This completes the proof of (B.15).

The proof of result (B.16) can be found in [25].

(e) We first show concentration for $\|q^1\|^2/n$. Note,

$$\mathbf{Pr}\left(\left|\frac{\|q^1\|^2}{n} - \sigma_1^2\right| \geq \Delta\right)$$
$$= \mathbf{Pr}\left(\left|\frac{1}{N}\sum_{i=1}^{N}\left(\eta_0(\beta_{0_i} - h_i^1) - \beta_{0_i}\right)^2 - \mathbb{E}[(\eta_0(\beta + \tau_0\tilde{Z}_0) - \beta)^2]\right| \geq \delta\Delta\right).$$

The result follows by $\mathcal{H}_1(d)$ since Lemma 31 implies

$$\phi_h(h_i^1, \beta_{0_i}) := \left(\eta_0(\beta_{0_i} - h_i^1) - \beta_{0_i}\right)^2 \in PL(2).$$

The proof of concentration for $(q^0)^* q^1/n$ is similar.

(f) We will show

$$\mathbb{E}\left\{\tau_0 Z_0[\eta_0(\beta - \tau_0 Z_0) - \beta]\right\} = \delta\tau_0^2\hat{\lambda}_1. \tag{B.26}$$

It follows,

$$\begin{aligned}&\mathbf{Pr}\left(\left|\frac{(h^1)^* q^1}{n} - \tau_0^2 \hat{\lambda}_1\right| \geq \Delta\right) \\ &\quad = \mathbf{Pr}\left(\left|\frac{1}{N}\sum_{i=1}^{N} h_i^1 \left(\eta_0(\beta_{0_i} - h_i^1) - \beta_{0_i}\right) - \mathbb{E}\left\{\tau_0 Z_0[\eta_0(\beta - \tau_0 Z_0) - \beta]\right\}\right| \geq \delta\Delta\right).\end{aligned}$$

The result follows by $\mathcal{H}_1(d)$ since Lemma 31 implies

$$\phi_h(h_i^1, \beta_{0_i}) := h_i^1 \left(\eta_0(\beta_{0_i} - h_i^1) - \beta_{0_i}\right) \in PL(2).$$

Now we show (B.26).

$$\mathbb{E}\left\{\tau_0 Z_0[\eta_0(\beta - \tau_0 Z_0) - \beta]\right\} \overset{(a)}{=} \tau_0 \mathbb{E}\left\{\frac{\partial}{\partial Z_0}\left[\eta_0(\beta - \tau_0 Z_0) - \beta\right]\right\} = -\tau_0^2 \mathbb{E}\left\{\eta_0'(\beta - \tau_0 Z_0)\right\}.$$

Step (a) follows by Stein's Lemma, Fact 10.

(g)

$$\begin{aligned}\mathbf{Pr}\left(\left|\gamma_0^1 - \hat{\gamma}_0^1\right| \geq \Delta\right) &\overset{(a)}{=} \mathbf{Pr}\left(\left|\frac{(q^0)^* q^1}{\|\beta\|^2} - \frac{E_{0,1}}{\sigma_0^2}\right| \geq \Delta\right) \\ &\overset{(b)}{\leq} \mathbf{Pr}\left(\left|\frac{n}{\|\beta\|^2} - \frac{1}{\sigma_0^2}\right| \geq \tilde{\Delta}\right) + \mathbf{Pr}\left(\left|\frac{(q^0)^* q^1}{n} - E_{0,1}\right| \geq \tilde{\Delta}\right) \\ &\overset{(c)}{=} e^{-\kappa n \Delta^2} + e^{-\kappa n \Delta^2}.\end{aligned}$$

Step (a) follows since $\gamma_0^1 = \frac{(q^0)^* q^1}{\|\beta\|^2}$, step (b) follows from Lemma 24 with $\tilde{\Delta} = \Delta/[2(1/\sigma_0^2 + |E_{0,1}|)]$, and step (c) from the sub-Gaussian assumption on p_β, Lemma 29, and $\mathcal{H}_1(e)$.

(h)

$$\begin{aligned}&\mathbf{Pr}\left(\left|\frac{\|q_\perp^1\|^2}{n} - (\sigma_1^\perp)^2\right| \geq \Delta\right) \\ &\overset{(a)}{\leq} \mathbf{Pr}\left(\left|\frac{\|q^1\|^2}{n} - \sigma_1^2\right| \geq \frac{\Delta}{2}\right) + \mathbf{Pr}\left(\left|(\gamma_0^1)^2 \frac{\|q^0\|^2}{n} - \hat{\gamma}_0^1 E_{0,1}\right| \geq \frac{\Delta}{2}\right) \\ &\overset{(b)}{\leq} e^{-\kappa n \Delta^2} + 2e^{-\kappa n \Delta^2}\end{aligned}$$

Step (a) follows from the concentration of sums, Lemma 23, and step (B) from $\mathcal{H}_1(e)$ and the following.

$$\begin{aligned}
&\mathbf{Pr}\left(\left|(\gamma_0^1)^2\frac{\|q^0\|^2}{n} - \hat{\gamma}_0^1 E_{0,1}\right| \geq \frac{\Delta}{2}\right) \\
&\overset{(a)}{\leq} \mathbf{Pr}\left(\left|(\gamma_0^1)^2 - \left(\hat{\gamma}_0^1\right)^2\right| \geq \tilde{\Delta}\right) + \mathbf{Pr}\left(\left|\frac{\|q^0\|^2}{n} - E_{0,0}\right| \geq \tilde{\Delta}\right) \\
&\overset{(b)}{\leq} e^{-\kappa n\Delta^2} + e^{-\kappa n\Delta^2}.
\end{aligned}$$

Step (a) follows from Lemma 24, for $\tilde{\Delta} = \Delta/[4(\left(\hat{\gamma}_0^1\right)^2 + E_{0,0})]$, and step (b) from $\mathcal{H}_1(g)$, Lemma 26, and the sub-Gaussian assumption on p_β.

Lemma 20 (Matrix Inverses)**.** *Symmetric matrices* $\boldsymbol{Q}_{t+1} := \frac{Q_{t+1}^* Q_{t+1}}{n}$ *and* $\boldsymbol{M}_t := \frac{M_t^* M_t}{n}$ *concentrate element-wise to the invertible matrices* C^{t+1} *and* $\sigma^2 + C^t$*, respectively, and are invertible with high probability, meaning:*

$$\boldsymbol{Pr}\left(\boldsymbol{Q}_{t+1} \text{ not invertible}\right) \leq Ke^{-\kappa n\Delta^2} \quad \text{and} \quad \boldsymbol{Pr}(\boldsymbol{M}_t \text{ not invertible}) \leq Ke^{-\kappa n\Delta^2}. \tag{B.27}$$

When they exist, the inverses also concentrate element-wise for all $0 \leq i, j \leq t+1$ *and* $0 \leq i', j' \leq t$ *as follows:*

$$\boldsymbol{Pr}\left(\left|\left[\boldsymbol{Q}_{t+1}^{-1}\right]_{i,j} - [(C^{t+1})^{-1}]_{i,j}\right| \geq \Delta\right) \leq Ke^{-\kappa n\Delta^2}, \tag{B.28}$$

$$\boldsymbol{Pr}\left(\left|\left[\boldsymbol{M}_t^{-1}\right]_{i',j'} - [(\sigma^2 + C^t)^{-1}]_{i',j'}\right| \geq \Delta\right) \leq Ke^{-\kappa n\Delta^2}. \tag{B.29}$$

Proof. We show the $\mathbf{M}_t$ results and those of $\mathbf{Q}_{t+1}$ follow similarly. We first show that $\mathbf{M}_t$ is invertible with high probability. By Fact 11, if $\|m_\perp^r\|^2/n \geq c_r$ for all $0 \leq r \leq t-1$ where c_r are positive constants, then $\mathbf{M}_t$ is invertible. Let $c_r = (\tau_r^\perp)^2 - \Delta_r$ for $\Delta_r \leq (\tau_r^\perp)^2$ which can be done since $(\tau_r^\perp)^2 > 0$ by Lemma 3. Then it follows,

$$\mathbf{Pr}\left(\mathbf{M}_t \text{ not invertible}\right) \leq \sum_{r=0}^{t-1} \mathbf{Pr}\left(\left|\frac{\|m_\perp^r\|^2}{n} - (\tau_r^\perp)^2\right| \geq \Delta_r\right) \leq K\epsilon^{-\kappa n\Delta^2},$$

where the last line follows from inductive hypotheses $\mathcal{B}_0 - \mathcal{B}_{t-1}$ (B.22).

Matrix $(\sigma^2 + C^{t-1})$ is invertible by Lemma 3, and the concentration result, (B.29),

follows from Lemma 30 since $\mathbf{M}_t$ concentrates on σ^2+C^t entry-wise by inductive hypotheses $\mathcal{B}_{t-1}(f)$. □

Step 3: Showing $\mathcal{B}_t$ holds

We wish to show results (a) - (h) in (B.10), (B.12), (B.14), (B.18), (B.20), (B.22), (B.24).

(a)

Lemma 21. *Let* $v := \frac{H_t^* q_\perp^t}{n} - \frac{M_t}{n}^* \left[\lambda_t m^{t-1} - \sum_{i=0}^{t-2} \lambda_{i+1}\gamma_{i+1}^t m^i\right]$, *a* $t-$*length vector and* $\boldsymbol{M}_t := \frac{M_t^* M_t}{n}$, *a* $t \times t$ *symmetric matrix. For* $j \in [t]$,

$$\boldsymbol{Pr}\left(\left|\left[\boldsymbol{M}_t^{-1} v\right]_j\right| \geq \Delta\right) \leq te^{-\kappa n\Delta^2}.$$

Proof. Define $\hat{\phi}_1, \ldots, \hat{\phi}_t$ to be the eigenvalues of $\mathbf{M}_t$. Let $j = 1$.

$$\mathbf{Pr}\left(\left|\left[\mathbf{M}_t^{-1} v\right]_1\right| \geq \Delta\right) \overset{(a)}{\leq} \mathbf{Pr}\left(\|v\| \max_{k\in[t]} \frac{1}{|\hat{\phi}_k|} \geq \Delta\right) \overset{(b)}{\leq} \mathbf{Pr}\left(\|v\| \geq \tilde{\kappa}\Delta\right) + \mathbf{Pr}\left(\max_{k\in[t]} \frac{1}{|\hat{\phi}_k|} \geq \frac{1}{\tilde{\kappa}}\right).$$

Step (a) follows from Lemma 39 and step (b) for $\tilde{\kappa} > 0$ constant. Label the two terms on the right side of the above T_1 and T_2. To complete the proof we will show that both terms are upper bounded by $te^{-\kappa n\Delta^2}$.

First consider term T_2. Define $\hat{\phi}_{min}$ to be the minimum eigenvalue of $\mathbf{M}_t$. By Fact 11, if $\|m_\perp^i\|^2/n \geq c_i$ for all $0 \leq i \leq t-1$, and for some positive constants $c_i > 0$, then $\hat{\phi}_{min} \geq \tilde{\kappa}$ where $\tilde{\kappa}$ is a strictly positive constant depending only on c_i and t. This implies that

$$\mathbf{Pr}\left(\hat{\phi}_{min} \geq \tilde{\kappa}\right) \geq \mathbf{Pr}\left(\cap_{i=0}^{t-1}\left\{\frac{\|m_\perp^i\|^2}{n} \geq c_i\right\}\right).$$

Let $c_i = (\tau_i^\perp)^2 - \Delta_i$, then it follows from the above,

$$\mathbf{Pr}\left(\hat{\phi}_{min} \leq \tilde{\kappa}\right) \leq \sum_{i=0}^{t-1} \mathbf{Pr}\left(\left|\frac{\|m_\perp^i\|^2}{n} - (\tau_i^\perp)^2\right| \geq \Delta_i\right) \leq te^{-\kappa n\Delta^2},$$

where the last line follows from inductive hypotheses $\mathcal{B}_0(g)$ - $\mathcal{B}_{t-1}(g)$. The upper bound for T_2 follows.

Next consider T_1. For $1 \leq k \leq t$,

$$|v_k| \leq \left| \frac{(h^k)^* q^t}{n} - \lambda_t \frac{(m^{k-1})^* m^{t-1}}{n} \right| + |\gamma_0^t| \left| \frac{(h^k)^* q^0}{n} \right| + \sum_{i=1}^{t-1} |\gamma_i^t| \left| \frac{(h^k)^* q^i}{n} - \lambda_i \frac{(m^{k-1})^* m^{i-1}}{n} \right|. \tag{B.30}$$

The above follows from the fact that $q_\perp^t = q^t - q_\parallel^t = q^t - \sum_{j=0}^{t-1} \gamma_j^t q^j$ and the Triangle Inequality. Therefore,

$$\begin{aligned}
&\mathbf{Pr}\left(\|v\|^2 \geq \tilde{\kappa}^2 \Delta^2\right) \\
&\quad \leq \sum_{k=1}^{t} \mathbf{Pr}\left(v_k^2 \geq \frac{\tilde{\kappa}^2 \Delta^2}{t} \right) \\
&\quad \leq \sum_{k=1}^{t} \mathbf{Pr}\left(\left| \frac{(h^k)^* q^t}{n} - \lambda_t \frac{(m^{k-1})^* m^{t-1}}{n} \right| \geq \Delta' \right) + \sum_{k=1}^{t} \mathbf{Pr}\left(|\gamma_0^t| \left| \frac{(h^k)^* q^0}{n} \right| \geq \Delta' \right) \\
&\qquad + \sum_{k=1}^{t} \sum_{i=1}^{t-1} \mathbf{Pr}\left(|\gamma_i^t| \left| \frac{(h^k)^* q^i}{n} - \lambda_i \frac{(m^{k-1})^* m^{i-1}}{n} \right| \geq \Delta' \right).
\end{aligned}$$

In the above $\Delta' = \frac{\tilde{\kappa}\Delta}{(t+1)\sqrt{t}}$, and both inequalities follow from Lemma 23. Label the terms of the above as $T_{a,k}, T_{b,k}$, and $T_{c,k,i}$ for $1 \leq k \leq t$ and $1 \leq i \leq t-1$. We show that each term is upper bounded by $Ke^{-\kappa n \Delta^2}$ to prove the desired bound on T_1. First for $1 \leq k \leq t$,

$$\begin{aligned}
T_{a,k} &\overset{(a)}{\leq} \mathbf{Pr}\left(\left| \frac{(h^k)^* q^t}{n} - \hat{\lambda}_t \left(\sigma^2 + E_{k-1,t-1} \right) \right| \geq \frac{\Delta'}{2} \right) \\
&\qquad + \mathbf{Pr}\left(\left| \lambda_t \frac{(m^{k-1})^* m^{t-1}}{n} - \hat{\lambda}_t \left(\sigma^2 + E_{k-1,t-1} \right) \right| \geq \frac{\Delta'}{2} \right) \\
&\overset{(b)}{\leq} Ke^{-\kappa n \Delta^2} + Ke^{-\kappa n \Delta^2}.
\end{aligned}$$

Step (a) follows from Lemma 23 and step (b) from Lemma 24 and inductive hypotheses $\mathcal{B}_{t-1}(f)$, $\mathcal{H}_t(f)$, and $\mathcal{H}_t$ (B.16). Next consider $T_{b,k}$ for $1 \leq k \leq t$,

$$T_{b,k} \overset{(a)}{\leq} \mathbf{Pr}\left(\left| |\gamma_0^t| - |\hat{\gamma}_0^t| \right| \geq \frac{\Delta'}{2|\hat{\gamma}_0^t|} \right) + \mathbf{Pr}\left(\left| \frac{(h^k)^* q^0}{n} \right| \geq \frac{\Delta'}{2|\hat{\gamma}_0^t|} \right) \overset{(b)}{\leq} Ke^{-\kappa n \Delta^2} + Ke^{-\kappa n \Delta^2}.$$

Step (a) follows from Lemma 24 and step (b) from inductive hypotheses $\mathcal{H}_t(b)$ and

$\mathcal{H}_t(g)$along with Lemma 28. Finally consider $T_{c,k,i}$ for $1 \le k \le t$ and $1 \le i \le t-1$,

$$\begin{aligned} T_{c,k,i} &\overset{(a)}{\le} \mathbf{Pr}\left(\left||\gamma_i^t| - |\hat{\gamma}_i^t|\right| \ge \frac{\Delta'}{2|\hat{\gamma}_i^t|}\right) + \mathbf{Pr}\left(\left|\frac{(h^k)^* q^i}{n} - \lambda_i \frac{(m^{k-1})^* m^{i-1}}{n}\right| \ge \frac{\Delta'}{2|\hat{\gamma}_i^t|}\right) \\ &\overset{(b)}{\le} Ke^{-\kappa n \Delta^2} + Ke^{-\kappa n \Delta^2}. \end{aligned}$$

Step (a) follows from Lemma 24 and step (b) from Lemma 28, inductive hypothesis $\mathcal{H}_t(g)$, and a method similar to that used to prove the bound for $T_{a,k}$ above. This completes the bound for T_1. □

Now we prove (a). Recall $\Delta_{t,t}$ is defined in Lemma 4 (3.25).

$$\begin{aligned} \|\Delta_{t,t}\|^2 \le 2(t+1)\sum_{r=0}^{t-1}(\gamma_r^t - \hat{\gamma}_r^t)^2\|b^r\|^2 + 2(t+1)\|Z_t'\|^2\left(\frac{\|q_\perp^t\|}{\sqrt{n}} - \sigma_t^\perp\right)^2 \\ + \frac{\|q_\perp^t\|^2\|\tilde{M}_t\bar{Z}_t'\|^2}{n^2} + 2(t+1)\sum_{j=0}^{t-1}\|m^j\|^2\left[\mathbf{M}_t^{-1}v\right]_{j+1}^2. \end{aligned}$$

The above follows form Lemma 38 and the fact $M_t\mathbf{M}_t^{-1}v = \sum_{j=0}^{t-1} m^j\left[\mathbf{M}_t^{-1}v\right]_{j+1}$. It follows,

$$\begin{aligned} &\mathbf{Pr}\left(\frac{\|\Delta_{t,t}\|^2}{n} \ge \Delta\right) \\ &\le \sum_{r=0}^{t-1}\mathbf{Pr}\left(\left|\gamma_r^t - \hat{\gamma}_r^t\right|\frac{\|b^r\|}{\sqrt{n}} \ge \frac{\sqrt{\Delta}}{2(t+1)}\right) + \mathbf{Pr}\left(\left|\frac{\|q_\perp^t\|}{\sqrt{n}} - \sigma_t^\perp\right|\frac{\|Z_t'\|}{\sqrt{n}} \ge \frac{\sqrt{\Delta}}{2(t+1)}\right) \\ &\quad + \mathbf{Pr}\left(\frac{\|q_\perp^t\|}{\sqrt{n}}\cdot\frac{\|\tilde{M}_t\bar{Z}_t'\|}{n} \ge \frac{\sqrt{\Delta}}{2(t+1)}\right) + \sum_{j=0}^{t-1}\mathbf{Pr}\left(\left|\left[\mathbf{M}_t^{-1}v\right]_{j+1}\right|\frac{\|m^j\|}{\sqrt{n}} \ge \frac{\sqrt{\Delta}}{2(t+1)}\right). \end{aligned}$$

The above follows from Lemma 23. Label the terms of the above as $T_{1,r}$, T_2, T_3, and $T_{4,j}$ for $0 \le r, j \le t-1$. In what follows we show that each term is upper bounded by $Ke^{-\kappa n\Delta}$ to prove the result. For each $0 \le r \le t-1$,

$$\begin{aligned} T_{1,r} &\overset{(a)}{\le} \mathbf{Pr}\left(\left|\gamma_r^t - \hat{\gamma}_r^t\right| \ge \frac{\sqrt{\Delta}}{4(t+1)\sigma_r}\right) + \mathbf{Pr}\left(\left|\frac{\|b^r\|}{\sqrt{n}} - \sigma_r\right| \ge \frac{\sqrt{\Delta}}{4(t+1)\sigma_r}\right) \\ &\overset{(b)}{\le} Ke^{-\kappa n\Delta} + Ke^{-\kappa n\Delta}. \end{aligned}$$

Step (a) follows from Lemma 24 and step (b) from inductive hypotheses $\mathcal{H}_t(g)$, $\mathcal{B}_0(c)$ - $\mathcal{B}_{t-1}(c)$, and Lemma 25. Next,

$$\begin{aligned} T_2 &\overset{(a)}{\leq} \mathbf{Pr}\left(\left|\frac{\|q_\perp^t\|}{\sqrt{n}} - \sigma_t^\perp\right| \geq \frac{\sqrt{\Delta}}{4(t+1)}\right) + \mathbf{Pr}\left(\left|\frac{\|Z_t'\|}{\sqrt{n}} - 1\right| \geq \frac{\sqrt{\Delta}}{4(t+1)}\right) \\ &\overset{(b)}{\leq} Ke^{-\kappa n\Delta} + e^{-\kappa n\Delta}. \end{aligned}$$

Step (a) follows from Lemma 24 and step (b) from inductive hypothesis $\mathcal{H}_t(h)$, Lemma 25, and Lemma 37. For each $0 \leq j \leq t-1$,

$$\begin{aligned} T_{4,j} &\overset{(a)}{\leq} \mathbf{Pr}\left(\left|\frac{\|m^j\|}{\sqrt{n}} - \tau_j\right| \geq \frac{\sqrt{\Delta}}{4(t+1)\tau_j}\right) + \mathbf{Pr}\left(\left|\left[\mathbf{M}_t^{-1}v\right]_{j+1}\right| \geq \frac{\sqrt{\Delta}}{4(t+1)\tau_j}\right) \\ &\overset{(b)}{\leq} Ke^{-\kappa n\Delta} + te^{-\kappa n\Delta}. \end{aligned}$$

Step (a) follows from Lemma 24 and step (b) from inductive hypothesis $\mathcal{B}_0(f)$ - $\mathcal{B}_{t-1}(f)$, Lemma 25, and Lemma 21. Finally,

$$\begin{aligned} T_3 &\overset{(a)}{\leq} \mathbf{Pr}\left(\left|\frac{\|q_\perp^t\|}{\sqrt{n}} - \sigma_t^\perp\right| \geq \frac{\sqrt{\Delta}}{4(t+1)\sigma_t^\perp}\right) + \mathbf{Pr}\left(\frac{\|\tilde{M}_t\bar{Z}_t'\|}{n} \geq \frac{\sqrt{\Delta}}{4(t+1)\sigma_t^\perp}\right) \\ &\overset{(b)}{\leq} Ke^{-\kappa n\Delta} + te^{-\kappa n\Delta}. \end{aligned}$$

Step (a) follows from Lemma 24 and step (b) from inductive hypothesis $\mathcal{H}_t(h)$, Lemma 25, and the fact that

$$\mathbf{Pr}\left(\frac{\|\tilde{M}_t\bar{Z}_t'\|}{n} \geq \sqrt{\Delta}\right) \leq te^{-\kappa n\Delta}$$

which we show in what follows. First,

$$\|\tilde{M}_t\bar{Z}_t'\|^2 = \|\sum_{i=0}^{t-1}\tilde{m}_i\bar{Z}_{t_i}'\|^2 \overset{(a)}{\leq} t\sum_{i=0}^{t-1}\|\tilde{m}_i\|^2(\bar{Z}_{t_i}')^2 \overset{(b)}{=} nt\sum_{i=0}^{t-1}(\bar{Z}_{t_i}')^2. \tag{B.31}$$

Step (a) follows from Lemma 38 and step (b) uses the fact that $\|\tilde{m}_i\|^2 = n$ for all

$0 \le i \le t-1$. Finally,

$$\mathbf{Pr}\left(\frac{\|\tilde{M}_t \bar{Z}'_t\|^2}{n^2} \ge \Delta\right) \overset{(a)}{\le} \mathbf{Pr}\left(\sum_{i=0}^{t-1} \frac{(\bar{Z}'_{t_i})^2}{n} \ge \frac{\Delta}{t}\right) \overset{(b)}{\le} \sum_{i=0}^{t-1} \mathbf{Pr}\left(|\bar{Z}'_{t_i}| \ge \frac{\sqrt{n\Delta}}{t}\right)$$
$$\overset{(c)}{\le} \sum_{i=0}^{t-1} 2\exp\left\{-\frac{n\Delta}{2t^2}\right\}.$$

Step (a) follows from (B.31), step (b) form Lemma 23, and step (c) from Lemma 36.

(b) We first show concentration of $(b^t)^*\epsilon/n$. First,

$$\mathbf{Pr}\left(\left|\frac{(b^t)^*\epsilon}{n}\right| \ge \Delta\right)$$
$$\le \sum_{r=0}^{t-1} \mathbf{Pr}\left(\left|\frac{(b^r)^*\epsilon}{n}\right| \ge \frac{\tilde{\Delta}}{\hat{\gamma}_r^t}\right) + \mathbf{Pr}\left(\left|\frac{(Z'_t)^*\epsilon}{n}\right| \ge \frac{\tilde{\Delta}}{\sigma_t^\perp}\right) + \mathbf{Pr}\left(\left|\frac{\Delta_{t,t}^*\epsilon}{n}\right| \ge \tilde{\Delta}\right).$$

In the above $\tilde{\Delta} := \Delta/(t+2)$. The above uses the conditional representation of b^t given in Lemma 4 (B.10) and Lemma 23. Label the three terms on the right side of the above as $T_{1,r}$, T_2 and T_3 for $0 \le r \le t-1$. To complete the proof we show that each is upper bounded by $Ke^{-\kappa n\Delta^2}$. First, $T_{1,r}$ is upper bounded by $Ke^{-\kappa n\Delta^2}$ using inductive hypothesis $\mathcal{B}_r(b)$. Next consider T_2 and let Z be an independent standard Gaussian random variable. Then,

$$T_2 \overset{(a)}{=} \mathbf{Pr}\left(\left|\frac{\|\epsilon\|}{\sqrt{n}} \cdot \frac{Z}{n}\right| \ge \frac{\tilde{\Delta}}{\sigma_t^\perp}\right) \overset{(b)}{\le} \mathbf{Pr}\left(\left|\frac{\|\epsilon\|}{\sqrt{n}} - \sigma\right| \ge \frac{\tilde{\Delta}}{2\sigma\sigma_t^\perp}\right) + \mathbf{Pr}\left(\frac{|Z|}{n} \ge \frac{\tilde{\Delta}}{2\sigma\sigma_t^\perp}\right)$$
$$\overset{(c)}{\le} e^{-\kappa n\Delta^2} + 2e^{-\frac{n\tilde{\Delta}^2}{8\sigma^2(\sigma_t^\perp)^2}}.$$

Step (a) follows since ϵ is independent of Z'_t, step (b) from Lemma 24, and step (c) from concentration of the noise, Lemma 25, and Lemma 36. Finally,

$$T_3 \overset{(a)}{\le} \mathbf{Pr}\left(\frac{\|\Delta_{t,t}\|\|\epsilon\|}{n} \ge \tilde{\Delta}\right) \overset{(b)}{\le} \mathbf{Pr}\left(\frac{\|\Delta_{t,t}\|}{\sqrt{n}} \ge \frac{\tilde{\Delta}}{2\sigma}\right) + \mathbf{Pr}\left(\left|\frac{\|\epsilon\|}{\sqrt{n}} - \sigma\right| \ge \frac{\tilde{\Delta}}{2\sigma}\right)$$
$$\overset{(c)}{\le} Ke^{-\kappa n\Delta^2} + e^{-\kappa n\Delta^2}.$$

Step (a) follows by Cauchy-Schwartz, step (b) from Lemma 24, and step (b) from $\mathcal{B}_t(a)$,

Lemma 25, and concentration of the noise. This completes the proof for concentration of $(b^t)^*\epsilon/n$.

Next we show concentration for $(m^t)^*\epsilon/n$.

$$\begin{aligned}\mathbf{Pr}\left(\left|\frac{(m^t)^*\epsilon}{n}+\sigma^2\right|\geq\Delta\right) &\overset{(a)}{=} \mathbf{Pr}\left(\left|\frac{(b^t)^*\epsilon}{n}-\frac{\|\epsilon\|^2}{n}+\sigma^2\right|\geq\Delta\right)\\ &\overset{(b)}{\leq} \mathbf{Pr}\left(\left|\frac{(b^t)^*\epsilon}{n}\right|\geq\frac{\Delta}{2}\right)+\mathbf{Pr}\left(\left|\frac{\|\epsilon\|^2}{n}-\sigma^2\right|\geq\frac{\Delta}{2}\right)\\ &\overset{(c)}{\leq} Ke^{-\kappa n\Delta^2}+e^{-\kappa n\Delta^2}.\end{aligned}$$

Step (a) uses $m^t = b^t - \epsilon$, step (b) from Lemma 23, and step (c) from the work above and concentration of the noise.

(c) We first demonstrate concentration for $(b^t)^*b^r/n$ for $0 \leq r \leq t-1$ and then for $\|b^t\|^2/n$.

$$\begin{aligned}&\mathbf{Pr}\left(\left|\frac{(b^t)^*b^r}{n}-E_{r,t}\right|\geq\Delta\right)\\ &\quad\leq\sum_{i=0}^{t-1}\mathbf{Pr}\left(\left|\frac{(b^i)^*b^r}{n}-E_{r,i}\right|\geq\frac{\tilde{\Delta}}{\hat{\gamma}_i^t}\right)+\mathbf{Pr}\left(\left|\frac{(Z_t')^*b^r}{n}\right|\geq\frac{\tilde{\Delta}}{\sigma_t^\perp}\right)+\mathbf{Pr}\left(\left|\frac{\Delta_{t,t}^*b^r}{n}\right|\geq\tilde{\Delta}\right).\end{aligned}$$

We define $\tilde{\Delta} := \Delta/(t+2)$. The above uses the conditional representation of b^t given in Lemma 4 (B.10), Lemma 23, and the fact that $\sum_{i=0}^{t-1}\hat{\gamma}_i^t E_{r,i} = E_{r,t}$. Label the terms of the above as $T_{1,i}$, T_2, and T_3 for $0 \leq i \leq t-1$. To complete the proof we show each term is upper bounded by $Ke^{-\kappa n\Delta^2}$. Term T_1 is upper bounded by $Ke^{-\kappa n\Delta^2}$ using inductive hypotheses $\mathcal{B}_0(c)$ - $\mathcal{B}_{t-1}(c)$. Consider T_2 and let Z be a standard Gaussian random variable.

$$\begin{aligned}T_2 \overset{(a)}{=} \mathbf{Pr}\left(\frac{\|b^r\|}{\sqrt{n}}\frac{|Z|}{\sqrt{n}}\geq\frac{\tilde{\Delta}}{\sigma_t^\perp}\right) &\overset{(b)}{=} \mathbf{Pr}\left(\left|\frac{\|b^r\|}{\sqrt{n}}-\sigma_r\right|\geq\frac{\tilde{\Delta}}{2\sigma_t^\perp\sigma_r}\right)+\mathbf{Pr}\left(|Z|\geq\frac{\sqrt{n}\tilde{\Delta}}{2\sigma_t^\perp\sigma_r}\right)\\ &\overset{(c)}{=} Ke^{-\kappa n\Delta^2}+2e^{-\frac{n\tilde{\Delta}^2}{8(\sigma_t^\perp)^2\sigma_r^2}}\end{aligned}$$

Step (a) follows since b^r is independent of Z_t', step (b) from Lemma 24, and step (c) from Lemma 25, inductive hypothesis $\mathcal{B}_0(c)$ - $\mathcal{B}_{t-1}(c)$ since $0 \leq r \leq t-1$, and Lemma

36. Finally, for T_3 note that by $|(b^r)^*\Delta_{t,t}| \leq \|b^r\|\|\Delta_{t,t}\|$ and therefore,

$$T_3 \overset{(a)}{\leq} \mathbf{Pr}\left(\frac{\|b^r\|}{\sqrt{n}}\frac{\|\Delta_{t,t}\|}{\sqrt{n}} \geq \tilde{\Delta}\right) \overset{(b)}{\leq} \mathbf{Pr}\left(\left|\frac{\|b^r\|}{\sqrt{n}} - \sigma_r\right| \geq \frac{\tilde{\Delta}}{2\sigma_r}\right) + \mathbf{Pr}\left(\frac{\|\Delta_{t,t}\|}{\sqrt{n}} \geq \frac{\tilde{\Delta}}{2\sigma_r}\right)$$
$$\overset{(c)}{\leq} Ke^{-\kappa n\Delta^2} + Ke^{-\kappa n\Delta^2}.$$

Step (a) follows by Cauchy-Schwartz, step (b) from Lemma 24, and step (c) from $\mathcal{B}_t(a)$, Lemma 25, and inductive hypotheses $\mathcal{B}_0(c)$ - $\mathcal{B}_{t-1}(c)$ since $0 \leq r \leq t-1$. We now have demonstrated concentration for $(b^t)^*b^r/n$ when $0 \leq r \leq t-1$.

Now we show that $\|b^t\|^2/n$ concentrates.

$$\begin{aligned}
&\mathbf{Pr}\left(\left|\frac{\|b^t\|^2}{n} - \sigma_t^2\right| \geq \Delta\right) \\
&\quad\leq \sum_{i=0}^{t-1}\sum_{j=0}^{t-1}\mathbf{Pr}\left(\left|\frac{(b^i)^*b^j}{n} - E_{i,j}\right| \geq \frac{\tilde{\Delta}}{|\hat{\gamma}_i^t\hat{\gamma}_j^t|}\right) + \mathbf{Pr}\left(\left|\frac{\|Z_t'\|^2}{n} - 1\right| \geq \frac{\tilde{\Delta}}{(\sigma_t^\perp)^2}\right) \\
&\quad+ \mathbf{Pr}\left(\frac{\|\Delta_{t,t}\|^2}{n} \geq \tilde{\Delta}\right) + \sum_{i=0}^{t-1}\mathbf{Pr}\left(\left|\frac{(b^i)^*Z_t'}{n}\right| \geq \frac{\tilde{\Delta}}{2\sigma_t^\perp|\hat{\gamma}_i^t|}\right) \\
&\quad+ \sum_{i=0}^{t-1}\mathbf{Pr}\left(\left|\frac{(b^i)^*\Delta_{t,t}}{n}\right| \geq \frac{\tilde{\Delta}}{2|\hat{\gamma}_i^t|}\right) + \mathbf{Pr}\left(\left|\frac{(Z_t')^*\Delta_{t,t}}{n}\right| \geq \frac{\tilde{\Delta}}{2\sigma_t^\perp}\right)
\end{aligned}$$

We have defined $\tilde{\Delta} := \Delta/(t^2+2t+3)$. The above uses the conditional distribution of b^t from Lemma 4 (B.10), Lemma 23, and the fact that $(\sigma_t^\perp)^2 + \sum_{i=0}^{t-1}\sum_{j=0}^{t-1}\hat{\gamma}_i^t\hat{\gamma}_j^t E_{i,j} = \sigma_t^2$. Label the terms of the above as $T_{1,i,j}$, T_2, T_3, $T_{4,i}$, $T_{5,i}$, and T_6 for $0 \leq i,j \leq t-1$. To complete the proof we show that each term is upper bounded by $Ke^{-\kappa n\Delta^2}$. Term $T_{1,i,j}$ has the desired upper bound using inductive hypotheses $\mathcal{B}_0(c)$ - $\mathcal{B}_{t-1}(c)$. Next, $T_2 \leq Ke^{-\kappa n\Delta^2}$ Lemma 37. By $\mathcal{B}_t(a)$, term T_3 can be upper bounded by $Ke^{-\kappa n\Delta}$. Next, let Z be a standard Gaussian random variable,

$$\begin{aligned}
T_{4,i} &\overset{(a)}{=} \mathbf{Pr}\left(\left|\frac{\|b^i\|}{\sqrt{n}}\frac{Z}{\sqrt{n}}\right| \geq \frac{\tilde{\Delta}}{2\sigma_t^\perp|\hat{\gamma}_i^t|}\right) \\
&\overset{(b)}{\leq} \mathbf{Pr}\left(\left|\frac{\|b^i\|}{\sqrt{n}} - \sigma_i\right| \geq \frac{\tilde{\Delta}}{4\sigma_i\sigma_t^\perp|\hat{\gamma}_i^t|}\right) + \mathbf{Pr}\left(\frac{|Z|}{\sqrt{n}} \geq \frac{\tilde{\Delta}}{4\sigma_i\sigma_t^\perp|\hat{\gamma}_i^t|}\right) \\
&\overset{(c)}{\leq} Ke^{-\kappa n\Delta^2} + 2e^{-\frac{n\tilde{\Delta}^2}{32\sigma_i^2(\sigma_t^\perp)^2(\hat{\gamma}_i^t)^2}}.
\end{aligned}$$

Step (a) follows since b^i is independent of Z'_t, step (b) from Lemma 24, and step (c) from Lemma 25, inductive hypotheses $\mathcal{B}_0(c)$ - $\mathcal{B}_{t-1}(c)$, and Lemma 36. Now consider $T_{5,i}$.

$$\begin{aligned} T_{5,i} &\overset{(a)}{\leq} \mathbf{Pr}\left(\left|\frac{\|b^i\|}{\sqrt{n}}\frac{\|\Delta_{t,t}\|}{\sqrt{n}}\right| \geq \frac{\tilde{\Delta}}{2|\hat{\gamma}_i^t|}\right) \\ &\overset{(b)}{\leq} \mathbf{Pr}\left(\left|\frac{\|b^i\|}{\sqrt{n}} - \sigma_i\right| \geq \frac{\tilde{\Delta}}{2\sigma_i|\hat{\gamma}_i^t|}\right) + \mathbf{Pr}\left(\frac{\|\Delta_{t,t}\|}{\sqrt{n}} \geq \frac{\tilde{\Delta}}{4\sigma_i|\hat{\gamma}_i^t|}\right) \\ &\overset{(c)}{\leq} Ke^{-\kappa n\Delta^2} + Ke^{-\kappa n\Delta^2}. \end{aligned}$$

Step (a) follows by Cauchy-Schwartz, step (b) from Lemma 24, and step (c) from Lemma 25, inductive hypotheses $\mathcal{B}_0(c)$ - $\mathcal{B}_{t-1}(c)$, and $\mathcal{B}_t(a)$. Finally,

$$\begin{aligned} T_6 &\overset{(a)}{\leq} \mathbf{Pr}\left(\frac{\|\Delta_{t,t}\|}{\sqrt{n}}\frac{\|Z'_t\|}{\sqrt{n}} \geq \frac{\tilde{\Delta}}{2\sigma_t^\perp}\right) \overset{(b)}{\leq} \mathbf{Pr}\left(\frac{\|\Delta_{t,t}\|}{\sqrt{n}} \geq \sqrt{\frac{\tilde{\Delta}}{4\sigma_t^\perp}}\right) + \mathbf{Pr}\left(\left|\frac{\|Z'_t\|}{\sqrt{n}} - 1\right| \geq \sqrt{\frac{\tilde{\Delta}}{4\sigma_t^\perp}}\right) \\ &\overset{(c)}{\leq} Ke^{-\kappa n\Delta} + Ke^{-\kappa n\Delta}. \end{aligned}$$

Step (a) follows by Cauchy-Schwartz, step (b) from Lemma 24, and step (c) from $\mathcal{B}_t(a)$, Lemma 37, and Lemma 25.

(d) Nothing for $\mathcal{B}$ steps.

(e) We show concentration of $\frac{(b^r)^* m^s}{n}$ when either $r = t$, $s = t$, or both $r = s = t$. The other cases are assumed in the inductive hypothesis.

$$\begin{aligned} \mathbf{Pr}\left(\left|\frac{(b^r)^* m^s}{n} - E_{r,s}\right| \geq \Delta\right) &\overset{(a)}{\leq} \mathbf{Pr}\left(\left|\frac{(b^r)^* b^s}{n} - E_{r,s}\right| \geq \frac{\Delta}{2}\right) + \mathbf{Pr}\left(\left|\frac{(b^r)^* \epsilon}{n}\right| \geq \frac{\Delta}{2}\right) \\ &\overset{(b)}{\leq} Ke^{-\kappa n\Delta^2} + Ke^{-\kappa n\Delta^2}. \end{aligned}$$

Step (a) follows since $m^s = b^s - \epsilon$ and from Lemma 23, and step (b) from $\mathcal{B}_t(c)$ and $\mathcal{B}_0(b)$ - $\mathcal{B}_t(b)$.

(f)

$$\begin{aligned}
&\mathbf{Pr}\left(\left|\frac{(m^r)^* m^t}{n} - \left(\sigma^2 + E_{r,t}\right)\right| \geq \Delta\right) \\
&\quad \overset{(a)}{\leq} \mathbf{Pr}\left(\left|\frac{(m^r)^* b^t}{n} - E_{r,t}\right| \geq \frac{\Delta}{2}\right) + \mathbf{Pr}\left(\left|\frac{(m^r)^* \epsilon}{n} + \sigma^2\right| \geq \frac{\Delta}{2}\right) \\
&\quad \overset{(b)}{\leq} Ke^{-\kappa n \Delta^2} + Ke^{-\kappa n \Delta^2}.
\end{aligned}$$

Step (a) follows since $m^t = b^t - \epsilon$ and from Lemma 23, and step (b) from $\mathcal{B}_t(e)$ and $\mathcal{B}_0(b)$ - $\mathcal{B}_t(b)$.

(g) For each $0 \leq k \leq t-1$,

$$\mathbf{Pr}\left(\left|\alpha_k^t - \hat{\alpha}_k^t\right| \geq \Delta\right) \tag{B.32}$$

$$\overset{(a)}{=} \mathbf{Pr}\left(\left|\sum_{i=0}^{t-1}\left(\left[\mathbf{M}_t^{-1}\right]_{k+1,i+1}\frac{(m^i)^* m^t}{n} - \left[(\sigma^2 + C^t)^{-1}\right]_{k+1,i+1}\left(\sigma^2 + E_{i,t}\right)\right)\right| \geq \Delta\right) \tag{B.33}$$

$$\overset{(b)}{\leq} \sum_{i=0}^{t-1}\mathbf{Pr}\left(\left|\left[\mathbf{M}_t^{-1}\right]_{k+1,i+1}\frac{(m^i)^* m^t}{n} - \left[(\sigma^2 + C^t)^{-1}\right]_{k+1,i+1}\left(\sigma^2 + E_{i,t}\right)\right| \geq \frac{\Delta}{t}\right) \tag{B.34}$$

$$\overset{(c)}{\leq} \sum_{i=0}^{t-1}\mathbf{Pr}\left(\left|\frac{(m^i)^* m^t}{n} - \left(\sigma^2 + E_{i,t}\right)\right| \geq \tilde{\Delta}_{k,i}\right) \tag{B.35}$$

$$+ \sum_{i=0}^{t-1}\mathbf{Pr}\left(\left|\left[\mathbf{M}_t^{-1}\right]_{k+1,i+1} - \left[(\sigma^2 + C^t)^{-1}\right]_{k+1,i+1}\right| \geq \tilde{\Delta}_{k,i}\right) \tag{B.36}$$

$$\overset{(d)}{\leq} tKe^{-\kappa n \Delta^2} + te^{-\kappa n \Delta^2}. \tag{B.37}$$

We have $\tilde{\Delta}_{k,i} = \Delta/t([(\sigma^2 + C^t)^{-1}]_{k+1,i+1} + \sigma^2 + E_{i,t})$. Step (a) follows from (3.14) and (3.20), step (b) from Lemma 23, step (c) from Lemma 24, and step (d) using $\mathcal{B}_t(f)$ and Lemma 20.

(h)

$$\begin{aligned}
&\mathbf{Pr}\left(\left|\frac{\|m_{\perp}^{t}\|^2}{n}-(\tau_t^{\perp})^2\right|\geq\Delta\right)\\
&\overset{(a)}{=}\mathbf{Pr}\left(\left|\frac{\|m^t\|^2}{n}-\frac{\|M_t\alpha^t\|^2}{n}-\tau_t^2+\sum_{j=0}^{t-1}\hat{\alpha}_j^t\left(\sigma^2+E_{j,t}\right)\right|\geq\Delta\right)\\
&\overset{(b)}{\leq}\mathbf{Pr}\left(\left|\frac{\|m^t\|^2}{n}-\tau_t^2\right|\geq\frac{\Delta}{2}\right)+\mathbf{Pr}\left(\left|\frac{\|M_t\alpha^t\|^2}{n}-\sum_{j=0}^{t-1}\hat{\alpha}_j^t\left(\sigma^2+E_{j,t}\right)\right|\geq\frac{\Delta}{2}\right).
\end{aligned}$$

Step (a) follows from the fact that $\|m_{\perp}^{t}\|^2=\|m^t\|^2-\|M_t\alpha^t\|^2$ and step (b) from Lemma 23. Label the terms of the above as T_1 and T_2. We show that both are upper bounded by $Ke^{-\kappa n\Delta^2}$ to get the desired result. Term T_1 has the desired upper bound by $\mathcal{B}_t(f)$. Consider T_2.

$$\begin{aligned}
T_2&\overset{(a)}{=}\mathbf{Pr}\left(\left|\sum_{j=0}^{t-1}\left(\alpha_j^t\frac{(m^j)^*m^t}{n}-\hat{\alpha}_j^t\left(\sigma^2+E_{j,t}\right)\right)\right|\geq\frac{\Delta}{2}\right)\\
&\overset{(b)}{\leq}\sum_{j=0}^{t-1}\mathbf{Pr}\left(\left|\alpha_j^t\frac{(m^j)^*m^t}{n}-\hat{\alpha}_j^t\left(\sigma^2+E_{j,t}\right)\right|\geq\frac{\Delta}{2t}\right)\\
&\overset{(c)}{\leq}\sum_{j=0}^{t-1}\mathbf{Pr}\left(\left|\alpha_j^t-\hat{\alpha}_j^t\right|\geq\tilde{\Delta}_j\right)+\sum_{j=0}^{t-1}\mathbf{Pr}\left(\left|\frac{(m^j)^*m^t}{n}-\left(\sigma^2+E_{j,t}\right)\right|\geq\tilde{\Delta}_j\right)\\
&\overset{(d)}{\leq}tKe^{-\kappa n\Delta^2}+tKe^{-\kappa n\Delta^2}.
\end{aligned}$$

We define $\tilde{\Delta}_j:=\Delta/4t(|\hat{\alpha}_j^t|+\sigma^2+|E_{j,t}|)$. Step (a) follows since $\frac{\|M_t\alpha^t\|^2}{n}=\sum_{i=0}^{t-1}\alpha_i^t\frac{(m^i)^*m^t}{n}$ using the definition (3.14), step (b) follows from the Triangle Inequality and Lemma 23, step (c) from Lemma 24, and step (d) using $\mathcal{B}_t(g)$ and $\mathcal{B}_t(f)$.

Step 4: Showing $\mathcal{H}_{t+1}$ holds

We wish to show results (a) - (h) in (B.9), (B.11), (B.13), (B.15), (B.16), (B.17), (B.19), (B.21), (B.23).

(a) The proof of $\mathcal{H}_{t+1}(a)$ is similar to that shown to prove $\mathcal{B}_t(a)$, including the use and proof of Lemma 22 stated below.

Lemma 22. *Let* $\boldsymbol{Q}_{t+1} := \frac{Q^*_{t+1}Q_{t+1}}{n}$ *and* $v := \frac{B^*_{t+1}m^{\perp}_t}{n} - \frac{Q^*_{t+1}(q^t - \sum_{i=0}^{t-1}\alpha^t_i q^i)}{n}$. *For each element,* $j \in [t+1]$,

$$\boldsymbol{Pr}\left(\left|\left[\boldsymbol{Q}^{-1}_{t+1}v\right]_j\right| \geq \Delta\right) \leq e^{-\kappa n\Delta^2}.$$

(b) The proof of $\mathcal{H}_{t+1}(b)$ is similar to that shown to prove $\mathcal{B}_t(b)$.

(c) The proof of $\mathcal{H}_{t+1}(c)$ is similar to that shown to prove $\mathcal{B}_t(c)$.

(d) We first show (B.15). Label

$$a_i = \left(h^1_i, ..., h^t_i, \sum_{r=0}^{t-1}\hat{\alpha}^t_r h^{r+1}_i + \tau^{\perp}_t Z_{t_i} + [\Delta_{t+1,t}]_i, \beta_{0_i}\right) \tag{B.38}$$

$$c_i = \left(h^1_i, ..., h^t_i, \sum_{r=0}^{t-1}\hat{\alpha}^t_r h^{r+1}_i + \tau^{\perp}_t Z_{t_i}, \beta_{0_i}\right). \tag{B.39}$$

Using (B.39) it follows,

$$\begin{aligned}
&\mathbf{Pr}\left(\left|\frac{1}{N}\sum_{i=1}^{N}\phi_h\left(h^1_i, ..., h^{t+1}_i, \beta_{0_i}\right) - \mathbb{E}\left[\phi_h\left(\tau_0\tilde{Z}_0, ..., \tau_t\tilde{Z}_t, \beta\right)\right]\right| \geq \Delta\right) \\
&= \mathbf{Pr}\left(\left|\frac{1}{N}\sum_{i=1}^{N}\phi_h\left(a_i\right) - \mathbb{E}\left[\phi_h\left(\tau_0\tilde{Z}_0, ..., \tau_t\tilde{Z}_t, \beta\right)\right]\right| \geq \Delta\right) \\
&\leq \mathbf{Pr}\left(\left|\frac{1}{N}\sum_{i=1}^{N}\phi_h\left(c_i\right) - \mathbb{E}\left[\phi_h\left(\tau_0\tilde{Z}_0, ..., \tau_t\tilde{Z}_t, \beta\right)\right]\right| \geq \frac{\Delta}{2}\right) \\
&\quad + \mathbf{Pr}\left(\left|\frac{1}{N}\sum_{i=1}^{N}\phi_h\left(a_i\right) - \frac{1}{N}\sum_{i=1}^{N}\phi_h\left(c_i\right)\right| \geq \frac{\Delta}{2}\right).
\end{aligned} \tag{B.40}$$

Bound (B.40) follows from Lemma 23. We will show at the end,

$$\mathbf{Pr}\left(\left|\frac{1}{N}\sum_{i=1}^{N}\phi_h\left(a_i\right) - \frac{1}{N}\sum_{i=1}^{N}\phi_h\left(c_i\right)\right| \geq \frac{\Delta}{2}\right) \leq e^{-\kappa n\Delta^2}. \tag{B.41}$$

Next bound the first term in (B.40). Define the function $\tilde{\phi}_{h_i} : \mathbb{R} \to \mathbb{R}$ as

$$\tilde{\phi}_{h_i}(s_i) := \phi_h\left(h^1_i, ..., h^t_i, \sum_{r=0}^{t-1}\hat{\alpha}^t_r h^{r+1}_i + \tau^{\perp}_t s_i, \beta_{0_i}\right).$$

From Lemma 32 it follows $\tilde{\phi}_{h_i} \in PL(2)$ for each $i \in [N]$ when conditioning on $\mathscr{S}_{t+1,t}$

(the sigma-field containing, among other things, $(h^1, \ldots, h^t, \beta_0)$). Then we can bound the first term of (B.40) as follows:

$$\begin{aligned}
&\mathbf{Pr}\left(\left|\frac{1}{N}\sum_{i=1}^{N}\phi_h\left(c_i\right) - \mathbb{E}\left[\phi_h\left(\tau_0\tilde{Z}_0, ..., \tau_t\tilde{Z}_t, \beta\right)\right]\right| \geq \frac{\Delta}{2}\right) \\
&= \mathbf{Pr}\left(\left|\frac{1}{N}\sum_{i=1}^{N}\tilde{\phi}_{h_i}(Z_{t_i}) - \mathbb{E}\left[\phi_h\left(\tau_0\tilde{Z}_0, ..., \tau_t\tilde{Z}_t, \beta\right)\right]\right| \geq \frac{\Delta}{2}\right) \\
&\leq \mathbf{Pr}\left(\left|\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{Z_t}\left[\tilde{\phi}_{h_i}(Z_{t_i})\right] - \mathbb{E}\left[\phi_h\left(\tau_0\tilde{Z}_0, ..., \tau_t\tilde{Z}_t, \beta\right)\right]\right| \geq \frac{\Delta}{4}\right) \\
&\quad + \mathbf{Pr}\left(\left|\frac{1}{N}\sum_{i=1}^{N}\tilde{\phi}_{h_i}(Z_{t_i}) - \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{Z_t}\left[\tilde{\phi}_{h_i}(Z_{t_i})\right]\right| \geq \frac{\Delta}{4}\right). \qquad \text{(B.42)}
\end{aligned}$$

Bound (B.42) follows from Lemma 23. The second term of (B.42) is upper bounded by $e^{-\kappa N\Delta^2}$ using Lemma 34. We show the same bound for the first term of (B.42). Recall from the definition of $\tilde{\phi}_{h_i}$ above,

$$\mathbb{E}_{Z_t}\left[\tilde{\phi}_{h_i}(Z_{t_i})\right] = \mathbb{E}_{Z_t}\left[\phi_h\left(h_i^1, ..., h_i^t, \sum_{r=0}^{t-1}\hat{\alpha}_r^t h_i^{r+1} + \tau_t^{\perp}Z_{t_i}, \beta_{0_i}\right)\right].$$

Now considering the above define the function $\phi_h' : \mathbb{R}^{t+1} \to \mathbb{R}$ as

$$\phi_h'\left(h_i^1, \ldots, h_i^t, \beta_{0_i}\right) = \mathbb{E}_{Z_t}\left[\phi_h\left(h_i^1, \ldots, h_i^t, \sum_{r=0}^{t-1}\hat{\alpha}_r^t h_i^{r+1} + \tau_t^{\perp}Z_{t_i}, \beta_{0_i}\right)\right].$$

This function is $PL(2)$ by Lemma 32. Then the first term of (B.42) equals:

$$\begin{aligned}
&\mathbf{Pr}\left(\left|\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{Z_t}\left[\tilde{\phi}_{h_i}(Z_{t_i})\right] - \mathbb{E}\left[\phi_h\left(\tau_0\tilde{Z}_0, ..., \tau_t\tilde{Z}_t, \beta\right)\right]\right| \geq \frac{\Delta}{4}\right) \\
&\quad = \mathbf{Pr}\left(\left|\frac{1}{N}\sum_{i=1}^{N}\phi_h'\left(h_i^1, \ldots, h_i^t, \beta_{0_i}\right) - \mathbb{E}\left[\phi_h\left(\tau_0\tilde{Z}_0, ..., \tau_t\tilde{Z}_t, \beta\right)\right]\right| \geq \frac{\Delta}{4}\right). \qquad \text{(B.43)}
\end{aligned}$$

We will show

$$\mathbb{E}\left[\phi_h\left(\tau_0\tilde{Z}_0, ..., \tau_t\tilde{Z}_t, \beta\right)\right] = \mathbb{E}\left[\phi_h'\left(\tau_0\tilde{Z}_0, ..., \tau_{t-1}\tilde{Z}_{t-1}, \beta\right)\right], \qquad \text{(B.44)}$$

and then (B.43) can be upper bounded by $e^{-\kappa n\Delta^2}$ using the inductive hypothesis $\mathcal{H}_t$

(B.15). Finally to complete the proof we must show that (B.41) and (B.44) hold.

- First we show result (B.41).

$$\begin{aligned}
&\mathbf{Pr}\left(\left|\frac{1}{N}\sum_{i=1}^{N}\phi_h\left(a_i\right)-\frac{1}{N}\sum_{i=1}^{N}\phi_h\left(c_i\right)\right|\geq\frac{\Delta}{2}\right)\\
&\leq\mathbf{Pr}\left(\frac{1}{N}\sum_{i=1}^{N}\left|\phi_h\left(a_i\right)-\phi_h\left(c_i\right)\right|\geq\frac{\Delta}{2}\right)\\
&\overset{(a)}{\leq}\mathbf{Pr}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{L}\left(1+|a_i|+|c_i|\right)|a_i-c_i|\geq\frac{\Delta}{2}\right)\\
&\overset{(b)}{\leq}\mathbf{Pr}\left(\frac{\|a-c\|}{\sqrt{N}}\cdot\left(1+\frac{\|a\|}{\sqrt{N}}+\frac{\|c\|}{\sqrt{N}}\right)\geq\frac{\Delta}{2\sqrt{3}\mathbf{L}}\right).
\end{aligned}$$

Step (a) follows from the fact that $\phi_h \in PL(2)$ and step (b) from Cauchy-Schwarz and the following application of Lemma 38:

$$\sum_{i=1}^{N}\frac{\left(1+|a_i|+|c_i|\right)^2}{N}\leq 3\left(1+\frac{\|a\|^2}{N}+\frac{\|c\|^2}{N}\right)\leq 3\left(1+\frac{\|a\|}{\sqrt{N}}+\frac{\|c\|}{\sqrt{N}}\right)^2. \quad \text{(B.45)}$$

The term after step (b) above can be upper bounded as follows using Lemma 23.

$$\begin{aligned}
\mathbf{Pr}\left(\frac{\|a-c\|}{\sqrt{N}}\geq\frac{\Delta}{6\sqrt{3}\mathbf{L}}\right)+\mathbf{Pr}\left(\frac{\|a-c\|}{\sqrt{N}}\cdot\frac{\|a\|}{\sqrt{N}}\geq\frac{\Delta}{6\sqrt{3}\mathbf{L}}\right)\\
+\mathbf{Pr}\left(\frac{\|a-c\|}{\sqrt{N}}\cdot\frac{\|c\|}{\sqrt{N}}\geq\frac{\Delta}{6\sqrt{3}\mathbf{L}}\right).
\end{aligned}$$

Label the terms ofthe above as $T_1 - T_3$. To complete the proof of (B.41) we will show that each can be upper bounded by $e^{-\kappa n\Delta^2}$. We demonstrate the bound for T_2. The bounds for terms T_1 and T_3 follows similarly noting that it follows from the Triangle Inequality $\|c\| \leq \|a\| + \|\Delta_{t+1,t}\|$. By Lemma 24,

$$T_2\leq\mathbf{Pr}\left(\left|\frac{\|a\|}{\sqrt{N}}-\sqrt{\mathbb{E}_a}\right|\geq\tilde{\Delta}\right)+\mathbf{Pr}\left(\frac{\|a-c\|}{\sqrt{N}}\geq\tilde{\Delta}\right), \quad \text{(B.46)}$$

where $\mathbb{E}_a = \delta\sigma_0^2 + \sum_{r=0}^{t-1}\sum_{r'=0}^{t-1}(\sigma^2 + E_{r,r'})$ and $\tilde{\Delta} = \Delta/(12\mathbf{L}\sqrt{3\mathbb{E}_a})$. Label the two terms of (B.46) as $T_{2,a}$ and $T_{2,b}$. We show both can be upper bounded by $e^{-\kappa n\Delta^2}$.

First notice,

$$T_{2,b} = \mathbf{Pr}\left(\frac{\|\Delta_{t+1,t}\|}{\sqrt{N}} \geq \tilde{\Delta}\right) \overset{(a)}{\leq} e^{-\kappa n \Delta^2},$$

Step (a) follows from $\mathcal{H}_{t+1}$ (B.11). Now we bound $T_{2,a}$. First notice,

$$\|a\|^2 = \|\beta_0\|^2 + \sum_{r=0}^{t}\sum_{r'=0}^{t}(h^{r+1})^* h^{r'+1} + 2\sum_{r=0}^{t}(h^{r+1})^*\beta_0.$$

It follows,

$$\begin{aligned}
&\mathbf{Pr}\left(\left|\frac{\|a\|^2}{N} - \mathbb{E}_a\right| \geq \Delta\right) \\
&= \mathbf{Pr}\left(\left|\frac{\|\beta_0\|^2}{N} + \sum_{r=0}^{t}\sum_{r'=0}^{t}\frac{(h^{r+1})^* h^{r'+1}}{N} + 2\sum_{r=0}^{t}\frac{(h^{r+1})^*\beta_0}{N} - \mathbb{E}_a\right| \geq \Delta\right) \\
&\overset{(a)}{\leq} \mathbf{Pr}\left(\left|\frac{\|\beta_0\|^2}{N} - \delta\sigma_0^2\right| \geq c\Delta\right) + \sum_{r=0}^{t}\sum_{r'=0}^{t}\mathbf{Pr}\left(\left|\frac{(h^{r+1})^* h^{r'+1}}{N} - (\sigma^2 + E_{r',r})\right| \geq c\Delta\right) \\
&\quad + \sum_{r=0}^{t}\mathbf{Pr}\left(\left|\frac{(h^{r+1})^*\beta_0}{N}\right| \geq \frac{c\Delta}{2}\right).
\end{aligned}$$

Step (a) follows by Lemma 23 for $c = (t+1)^2 + t + 2$. Label the terms after step (a) as $T_A, T_{B,r,r'}$, and $T_{C,r}$ for $0 \leq r, r' \leq t$. We show that each is upper bounded by $e^{-\kappa n \Delta^2}$. The bound for $T_{2,a}$ then follows from Lemma 25. Term T_A has the desired bound by assumption. Next, term $T_{B,r,r'}$ has the desired upper bound by $\mathcal{H}_1 - \mathcal{H}_{t+1}$ (B.13). Finally,

$$T_{C,r} = \mathbf{Pr}\left(\left|\frac{(h^{r+1})^* q^0}{n}\right| \geq \frac{c\Delta}{2\delta}\right) \overset{(a)}{\leq} e^{-\kappa n \Delta^2}.$$

Step (a) follows from $\mathcal{H}_1 - \mathcal{H}_{t+1}$ (B.11). This completes the proof of (B.41).

- Finally we show (B.44). Recall the definition of ϕ'_h.

$$\begin{aligned}
&\mathbb{E}\left[\phi'_h\left(\tau_0\tilde{Z}_0,...,\tau_{t-1}\tilde{Z}_{t-1},\beta\right)\right] \\
&\quad= \mathbb{E}_{\tilde{Z}_0,...,\tilde{Z}_{t-1},\beta}\left[\mathbb{E}_{Z_t}\left[\phi_h\left(\tau_0\tilde{Z}_0,\ldots,\tau_{t-1}\tilde{Z}_{t-1},\sum_{r=0}^{t-1}\hat{\alpha}_r^t\tau_r\tilde{Z}_r+\tau_t^{\perp}Z_t,\beta\right)\right]\right] \\
&\quad= \mathbb{E}_{Z_t,\tilde{Z}_0,...,\tilde{Z}_{t-1},\beta}\left[\phi_h\left(\tau_0\tilde{Z}_0,\ldots,\tau_{t-1}\tilde{Z}_{t-1},\sum_{r=0}^{t-1}\hat{\alpha}_r^t\tau_r\tilde{Z}_r+\tau_t^{\perp}Z_t,\beta\right)\right].
\end{aligned}$$

To complete the proof of (B.44) we show that

$$\sum_{r=0}^{t-1}\hat{\alpha}_r^t\tau_r\tilde{Z}_r+\tau_t^{\perp}Z_t\overset{d}{=}\tau_t\tilde{Z}_t, \tag{B.47}$$

and for $0\le k\le t-1$

$$\tau_k\tau_t\mathbb{E}[\tilde{Z}_t\tilde{Z}_k]=\sigma^2+E_{k,t}. \tag{B.48}$$

First for (B.47). Since the left side is a sum of Gaussians we just show that the variance of the left side equals τ_t^2.

$$\begin{aligned}
\mathbb{E}\left[\left(\sum_{r=0}^{t-1}\hat{\alpha}_r^t\tau_r\tilde{Z}_r+\tau_t^{\perp}Z_t\right)^2\right] &\overset{(a)}{=} \sum_{r'=0}^{t-1}\sum_{r=0}^{t-1}\hat{\alpha}_{r'}^t\hat{\alpha}_r^t\tau_{r'}\tau_r\mathbb{E}\left[\tilde{Z}_{r'}\tilde{Z}_r\right]+(\tau_t^{\perp})^2 \\
&\overset{(b)}{=} \sum_{r'=0}^{t-1}\hat{\alpha}_{r'}^t\sum_{r=0}^{t-1}\hat{\alpha}_r^t(\sigma^2+E_{r',r})+\tau_t^2-\sum_{i=0}^{t-1}\hat{\alpha}_i^t(\sigma^2+E_{i,t}) \\
&\overset{(c)}{=} \tau_t^2
\end{aligned}$$

Step (a) follows since Z_t is independent of $(\tilde{Z}_0,\ldots,\tilde{Z}_{t-1})$, step (b) from the fact $\tau_{r'}\tau_r\mathbb{E}\left[\tilde{Z}_{r'}\tilde{Z}_r\right]=\sigma^2+E_{r',r}$ for $0\le r,r'\le t-1$ and step (c) from definition (3.20). Finally we show that (B.48) is true.

$$\begin{aligned}
\tau_k\tau_t\mathbb{E}[\tilde{Z}_t\tilde{Z}_k]\overset{(a)}{=}\tau_k\sum_{r=0}^{t-1}\hat{\alpha}_r^t\tau_r\mathbb{E}[\tilde{Z}_k\tilde{Z}_r]+\hat{\tau}_t\mathbb{E}[\tilde{Z}_kZ_t] &\overset{(b)}{=} \sum_{r=0}^{t-1}\hat{\alpha}_r^t(\sigma^2+E_{r,k}) \\
&\overset{(c)}{=} \sigma^2+E_{k,t}.
\end{aligned}$$

Step (a) follows from (B.47), step (b) from the fact that Z_t is independent of $\tilde{Z}_k$

for $0 \leq k \leq t-1$ and the fact $\tau_k \tau_r \mathbb{E}\left[\tilde{Z}_k \tilde{Z}_r\right] = \sigma^2 + E_{k,r}$ for $0 \leq r, k \leq t-1$. Step (c) follows from definitions definition (3.20). This completes the proof of (B.44).

The proof of result (B.16) can be found in [25].

(e) We demonstrate concentration for $(q^r)^* q^{t+1}/n$ when $0 \leq r \leq t+1$. Note,

$$\begin{aligned}
&\mathbf{Pr}\left(\left|\frac{(q^r)^* q^{t+1}}{n} - E_{r,t+1}\right| \geq \Delta\right) \\
&\quad = \mathbf{Pr}\left(\left|\frac{1}{N}\sum_{i=1}^{N} (\eta_{r-1}(\beta_{0_i} - h_i^r) - \beta_{0_i})\left(\eta_t(\beta_{0_i} - h_i^{t+1}) - \beta_{0_i}\right) - \delta E_{r,t+1}\right| \geq \delta\Delta\right).
\end{aligned}$$

The result follows by $\mathcal{H}_1(d)$ since Lemma 31 implies

$$\phi_h(h_i^1, \ldots, h_i^{t+1}, \beta_{0_i}) := (\eta_{r-1}(\beta_{0_i} - h_i^r) - \beta_{0_i})\left(\eta_t(\beta_{0_i} - h_i^{t+1}) - \beta_{0_i}\right) \in PL(2),$$

and

$$\delta E_{r,t+1} = \mathbb{E}[(\eta_{r-1}(\beta + \tau_{r-1}\tilde{Z}_{r-1}) - \beta)(\eta_t(\beta + \tau_t \tilde{Z}_t) - \beta)].$$

(f) We show that $(h^{r+1})^* q^{s+1}/n$ concentrates where either $r = t$ or $s = t$ or both $r = s = t$. The other cases are assumed in the inductive hypothesis. We will prove

$$\mathbb{E}\left\{\tau_r \tilde{Z}_r(\eta_s(\beta - \tau_s \tilde{Z}_s) - \beta)\right\} = \delta \hat{\lambda}_{s+1}(\sigma^2 + E_{r,s}), \tag{B.49}$$

It follows,

$$\begin{aligned}
&\mathbf{Pr}\left(\left|\frac{(h^{r+1})^* q^{s+1}}{n} - \hat{\lambda}_{s+1}(\sigma^2 + E_{r,s})\right| \geq \Delta\right) \\
&\quad = \mathbf{Pr}\left(\left|\frac{1}{N}\sum_{i=1}^{N} h_i^{r+1}(\eta_s(\beta_{0_i} - h_i^{s+1}) - \beta_{0_i}) - \mathbb{E}\left\{\tau_r \tilde{Z}_r(\eta_s(\beta - \tau_s \tilde{Z}_s)\right\}\right| \geq \delta\Delta\right).
\end{aligned}$$

The result follows by $\mathcal{H}_1(d)$ since Lemma 31 implies

$$\phi_h(h_i^1, \ldots, h_i^{t+1}, \beta_{0_i}) := h_i^{r+1}(\eta_s(\beta_{0_i} - h_i^{s+1}) - \beta_{0_i}) \in PL(2).$$

Now we show (B.49).

$$\begin{aligned}\mathbb{E}\left\{\tau_r\tilde{Z}_r[\eta_s(\beta-\tau_s\tilde{Z}_s)-\beta]\right\} &\overset{(a)}{=} \tau_r\mathbb{E}[\tilde{Z}_r\tilde{Z}_s]\mathbb{E}\left\{\frac{\partial}{\partial\tilde{Z}_s}[\eta_s(\beta-\tau_s\tilde{Z}_s)-\beta]\right\}\\ &= -\tau_r\tau_s\mathbb{E}[\tilde{Z}_r\tilde{Z}_s]\mathbb{E}\left\{\eta_s'(\beta-\tau_s\tilde{Z}_s)\right\}\\ &\overset{(b)}{=} -(\sigma^2+E_{r,s})\mathbb{E}\left\{\eta_s'(\beta-\tau_s\tilde{Z}_s)\right\}.\end{aligned}$$

Step (a) follows by Stein's Lemma, Fact 10 and step (b) follows from the definition of the covariance of $\tilde{Z}_r$ and $\tilde{Z}_s$ given in (3.17).

(g) The proof of $\mathcal{H}_{t+1}(g)$ is similar to that shown to prove $\mathcal{B}_t(g)$.

(h) The proof of $\mathcal{H}_{t+1}(h)$ is similar to that shown to prove $\mathcal{B}_t(h)$.

B.5.1 Concentration Lemmas

Lemma 23 (Concentration of Sums)**.** *For a sequence of random variables $X_1,\ldots,X_n$,*

$$\boldsymbol{Pr}\left(|\sum_{i=0}^{n}X_i|\geq\Delta\right)\leq\sum_{i=0}^{n}\boldsymbol{Pr}\left(|X_i|\geq\frac{\Delta}{n}\right).$$

Proof. Notice that if $|X_i|<\frac{\Delta}{n}$ for all $i\in[n]$ then $|\sum_{i=0}^{n}X_i|<\Delta$ and therefore

$$\mathbf{Pr}\left(|\sum_{i=0}^{n}X_i|\geq\Delta\right)\leq\mathbf{Pr}\left(|X_i|\geq\frac{\Delta}{n}\text{ for some }i\right)\leq\sum_{i=0}^{n}\mathbf{Pr}\left(|X_i|\geq\frac{\Delta}{n}\right).$$

□

Lemma 24 (Concentration of Products)**.** *Let either $C_X\neq0$ or $C_Y\neq0$. I Define $\mathcal{M}^2:=\max(1,2(|C_X|+|C_Y|)^2)$. If $\Delta\leq\mathcal{M}^2$,*

$$\boldsymbol{Pr}\left(|X_nY_n-C_XC_Y|\geq\Delta\right)\leq\boldsymbol{Pr}\left(|X_n-C_X|\geq\frac{\Delta}{2\mathcal{M}}\right)\boldsymbol{Pr}\left(|Y_n-C_Y|\geq\frac{\Delta}{2\mathcal{M}}\right).$$

Proof. We will argue that whenever

$$|X_n-C_X|\leq\frac{\Delta}{2\mathcal{M}}\text{ and }|Y_n-C_Y|\leq\frac{\Delta}{2\mathcal{M}}\tag{B.50}$$

then $|X_nY_n - C_XC_Y| \leq \Delta$, and the probability statement follows.

Define $\tilde{\Delta} := \frac{\Delta}{2\mathcal{M}}$. If $C_X - \tilde{\Delta} \leq X_n \leq C_X + \tilde{\Delta}$ and $C_Y - \tilde{\Delta} \leq Y_n \leq C_Y + \tilde{\Delta}$, then

$$|X_nY_n - C_XC_Y| \tag{B.51}$$

$$\leq \max(|\tilde{\Delta}(C_X + C_Y) + \tilde{\Delta}^2|, |-\tilde{\Delta}(C_X + C_Y) + \tilde{\Delta}^2|, |\tilde{\Delta}(C_X - C_Y) - \tilde{\Delta}^2|, |\tilde{\Delta}(C_Y - C_X) - \tilde{\Delta}^2|)$$

$$\leq \tilde{\epsilon}(|C_X| + |C_Y|) + \tilde{\Delta}^2. \tag{B.52}$$

Now assume $\mathcal{M} = 1$ meaning $\Delta \leq 1$ and $|C_X| + |C_Y| \leq 1/\sqrt{2}$. Then from (B.52),

$$|X_nY_n - C_XC_Y| \leq \frac{\Delta}{2}\left[(|C_X| + |C_Y|) + \frac{\Delta}{2}\right] \leq \frac{\Delta}{2}\left[\frac{1}{\sqrt{2}} + \frac{1}{2}\right] \leq \Delta. \tag{B.53}$$

Next assume $\mathcal{M} = \sqrt{2}(|C_X| + |C_Y|)$ meaning $\Delta \leq 2(|C_X| + |C_Y|)^2$. Then from (B.52),

$$|X_nY_n - C_XC_Y| \leq \frac{\Delta}{2\sqrt{2}}\left[1 + \frac{\Delta}{2\sqrt{2}(|C_X| + |C_Y|)^2}\right] \leq \frac{\Delta}{2\sqrt{2}}\left[1 + \frac{1}{\sqrt{2}}\right] \leq \Delta. \tag{B.54}$$

□

Lemma 25 (Concentration of Square Roots)**.** *Let $C_X \neq 0$.*

$$\textit{If } \boldsymbol{Pr}\left(\left|X_n^2 - C_X^2\right| \geq \Delta\right) \leq e^{-\kappa n\Delta^2}, \textit{ then } \boldsymbol{Pr}(||X_n| - |C_X|| \geq \Delta) \leq e^{-\kappa n|C_X|^2\Delta^2}.$$

Proof. First assume $\Delta \leq C_X^2$. If $C_X^2 - \Delta \leq X_n^2 \leq C_X^2 + \Delta$ then $\sqrt{C_X^2 - \Delta} \leq |X_n| \leq \sqrt{C_X^2 + \Delta}$. Assume $\Delta \geq C_X^2$. If $C_X^2 - \Delta \leq X_n^2 \leq C_X^2 + \Delta$ then $0 \leq |X_n| \leq \sqrt{C_X^2 + \Delta}$. Therefore,

$$||X_n| - |C_X|| \leq |C_X| \max\left(1 - \sqrt{1 - \frac{\Delta}{C_X^2}}, \sqrt{1 + \frac{\Delta}{C_X^2}} - 1\right).$$

Note that $(1+x)^{1/2} \leq 1 + \frac{1}{2}x$ and for $x \leq 1$, then $(1-x)^{1/2} \geq 1 - x$. Putting these together,

$$||X_n| - |C_X|| \leq |C_X| \max\left(1 - \sqrt{1 - \frac{\Delta}{C_X^2}}, \sqrt{1 + \frac{\Delta}{C_X^2}} - 1\right) \leq |C_X| \max\left(\frac{\Delta}{C_X^2}, \frac{\Delta}{2C_X^2}\right) = \frac{\Delta}{|C_X|}.$$

□

Lemma 26 (Concentration of Squares)**.** *Let $C_X \neq 0$ and $\Delta \leq 1$.*

$$\textit{If } \boldsymbol{Pr}(|X_n - C_X| \geq \Delta) \leq e^{-\kappa n \Delta^2}, \textit{ then } \boldsymbol{Pr}\left(\left|X_n^2 - C_X^2\right| \geq \Delta\right) \leq e^{\frac{-\kappa n \Delta^2}{(2|C_X|+1)^2}}.$$

Proof. Assume without loss of generality $C_X > 0$. If $C_X - \Delta \leq X_n \leq C_X + \Delta$ then $(C_X - \Delta)^2 \leq X_n^2 \leq (C_X + \Delta)^2$ meaning,

$$\begin{aligned}
\left|X_n^2 - C_X^2\right| &\leq C_X^2 \max\left(\left|1 - \left(1 - \frac{\Delta}{|C_X|}\right)^2\right|, \left|\left(1 + \frac{\Delta}{|C_X|}\right)^2 - 1\right|\right) \\
&\leq \Delta|C_X| \max\left(\left|2 - \frac{\Delta}{|C_X|}\right|, \left|2 + \frac{\Delta}{|C_X|}\right|\right) \\
&\leq \Delta(2|C_X| + \Delta).
\end{aligned}$$

It follows form the above, when $\Delta \leq 1$,

$$\mathbf{Pr}\left(\left|X_n^2 - C_X^2\right| \geq \Delta(2|C_X| + \Delta)\right) \leq e^{-\kappa n \Delta^2} \implies \mathbf{Pr}\left(\left|X_n^2 - C_X^2\right| \geq \Delta\right) \leq \exp\left\{\frac{-\kappa n \Delta^2}{(2|C_X|+1)^2}\right\}.$$

□

Lemma 27 (Concentration of Powers)**.** *Assume $C_X \neq 0$ and $\Delta \leq 1$. For each integer $k \geq 0$,*

$$\textit{if } \boldsymbol{Pr}(|X_n - C_X| \geq \Delta) \leq e^{-\kappa n \Delta^2}, \textit{ then } \boldsymbol{Pr}\left(\left|X_n^k - C_X^k\right| \geq \Delta\right) \leq e^{-\kappa' n \Delta^2}.$$

Proof. First note that the cases $k = 0$ and $k = 1$ are trivial so we prove the result for integers $k \geq 2$. If $C_X - \Delta \leq X_n \leq C_X + \Delta$ then $C_X^k \left(1 - \frac{\Delta}{|C_X|}\right)^k \leq X_n^k \leq C_X^k \left(1 + \frac{\Delta}{|C_X|}\right)^k$

meaning,

$$\begin{aligned}\left|X_n^k - C_X^k\right| &\leq |C_X|^k \max\left(\left|1-\left(1-\frac{\Delta}{|C_X|}\right)^k\right|, \left|\left(1+\frac{\Delta}{|C_X|}\right)^k - 1\right|\right) \\ &= |C_X|^k \max\left(\left|\sum_{j=1}^{k}\binom{k}{j}(-1)^j\left(\frac{\Delta}{|C_X|}\right)^j\right|, \left|\sum_{i=1}^{k}\binom{k}{i}\left(\frac{\Delta}{|C_X|}\right)^i\right|\right) \\ &= \max\left(\left|\sum_{j=1}^{k}\binom{k}{j}(-\Delta)^j|C_X|^{k-j}\right|, \left|\sum_{i=1}^{k}\binom{k}{i}\Delta^i|C_X|^{k-i}\right|\right) \\ &\leq \epsilon\sum_{i=1}^{k}\binom{k}{i}\Delta^{i-1}|C_X|^{k-i}.\end{aligned} \tag{B.55}$$

This means,

$$\mathbf{Pr}\left(\left|X_n^k - C_X^k\right| \geq \Delta\sum_{i=1}^{k}\binom{k}{i}\Delta^{i-1}|C_X|^{k-i}\right) \leq e^{-\kappa n\Delta^2} \tag{B.56}$$

and so, for some constant $c > 0$,

$$\mathbf{Pr}\left(\left|X_n^k - C_X^k\right| \geq \Delta\right) \leq \mathbf{Pr}\left(\left|X_n^k - C_X^k\right| \geq c\Delta\sum_{i=1}^{k}\binom{k}{i}\epsilon^{i-1}|C_X|^{k-i}\right) \leq e^{-\kappa' n\Delta^2}, \tag{B.57}$$

where for $\Delta \leq 1$ it follows $\Delta \geq c\Delta\sum_{i=1}^{k}\binom{k}{i}\Delta^{i-1}|C_X|^{k-i}$ when

$$c = \frac{1}{\sum_{i=1}^{k}\binom{k}{i}|C_X|^{k-i}} = \frac{1}{(1+|C_X|)^k - |C_X|^k}. \tag{B.58}$$

□

Lemma 28 (Concentration of Absolute Values)**.** *Assume* $C_X \neq 0$.

$$\textit{If } \boldsymbol{Pr}(|X_n - C_X| \geq \Delta) \leq e^{-\kappa n\Delta^2}, \textit{ then } \boldsymbol{Pr}(||X_n| - |C_X|| \geq \Delta) \leq e^{-\kappa n\Delta^2},$$

Proof. If $C_X - \Delta \leq X_n \leq C_X + \Delta$ then, $||C_X| - \Delta| \leq |X_n| \leq |C_X| + \Delta$. Therefore,

$$||X_n| - |C_X|| \leq \max(|C_X| - ||C_X| - \Delta|, \Delta) \leq \Delta.$$

□

Lemma 29 (Concentration of Inverses)**.** *Assume* $\Delta < \frac{1}{2}|C_X|$.

$$\textit{If } \boldsymbol{Pr}(|X_n - C_X| \geq \Delta) \leq e^{-\kappa n \Delta^2}, \textit{ then } \boldsymbol{Pr}\left(\left|\frac{1}{X_n} - \frac{1}{C_X}\right| \geq \Delta\right) \leq \exp\left\{-\frac{\kappa n C_X^2 \Delta^2}{2}\right\}.$$

Proof. If $C_X - \Delta \leq X_n \leq C_X + \Delta$ then $\frac{1}{C_X}\left(\frac{1}{1+\Delta/C_X}\right) \leq \frac{1}{X_n} \leq \frac{1}{C_X}\left(\frac{1}{1-\Delta/C_X}\right)$.

$$\left|\frac{1}{X_n} - \frac{1}{C_X}\right| \leq \frac{1}{|C_X|}\max\left(1 - \frac{1}{1+\Delta/|C_X|}, \frac{1}{1-\Delta/|C_X|} - 1\right) \overset{(a)}{\leq} \frac{1}{|C_X|}\max\left(\frac{\Delta}{|C_X|}, \frac{2\Delta}{|C_X|}\right) \leq \frac{2\Delta}{C_X^2}.$$

Step (a) uses the fact $\frac{1}{1-x} \leq 1 + 2x$ when $0 \leq x \leq \frac{1}{2}$ and $\frac{1}{1+x} \geq 1 - x$. □

Lemma 30 (Inverse Matrix Concentration)**.** *Suppose we have a sequence of symmetric, invertible* $t \times t$ *matrices indexed by* $n \geq 1$*:*

$$\{A_n\}_{n\geq 1} = \{A_1, A_2, \ldots\} \textit{ with } A_i \in \mathbb{R}^{t\times t},$$

such that

$$\boldsymbol{Pr}(|[A_n]_{i,j} - A_{i,j}| \geq \Delta) \leq e^{-\kappa n \Delta^2}, \tag{B.59}$$

where $A = [a_{i,j}]_{1\leq i,j\leq t}$ *is invertible. Then*

$$\boldsymbol{Pr}\left(\left|[A_n^{-1}]_{i,j} - [A^{-1}]_{i,j}\right| \geq \Delta\right) \leq e^{-\kappa n \Delta^2}.$$

Proof. Recall the Cayley-Hamilton theorem which allows us to represent the inverse of a matrix in terms of its determinate, traces, and powers. We apply this to A and A_n as follows.

$$A^{-1} = \frac{1}{det(A)}\sum_{r=0}^{t-1} A^r C(A, r), \text{ and } A_n^{-1} = \frac{1}{det(A_n)}\sum_{r=0}^{t-1} A_n^r C(A_n, r), \tag{B.60}$$

where

$$C(A,r) = \sum_{k_1,\ldots,k_{t-1}} \prod_{s=1}^{t-1} \frac{(-1)^{k_s+1}}{s^{k_s} k_s!} tr(A^s)^{k_s}, \text{ and } C(A_n, r) = \sum_{k_1,\ldots,k_{t-1}} \prod_{s=1}^{t-1} \frac{(-1)^{k_s+1}}{s^{k_s} k_s!} tr(A_n^s)^{k_s}, \tag{B.61}$$

with the sum in the above taken over all sets of integer $k_s \geq 0$ satisfying

$$r + \sum_{s=1}^{t-1} sk_s = t - 1. \tag{B.62}$$

Now we can use (B.60) bound the concentration probability.

$$\begin{aligned}
&\mathbf{Pr}\left(|[A_n]_{i,j} - A_{i,j}| \geq \Delta\right) \\
&\quad = \mathbf{Pr}\left(\left|\frac{1}{det(A_n)}\sum_{r=0}^{t-1}[A_n^r]_{i,j}C(A_n,r) - \frac{1}{det(A)}\sum_{r=0}^{t-1}[A^r]_{i,j}C(A,r)\right| \geq \Delta\right) \\
&\quad \overset{(a)}{\leq} \sum_{r=0}^{t-1}\mathbf{Pr}\left(\left|\frac{[A_n^r]_{i,j}C(A_n,r)}{det(A_n)} - \frac{[A^r]_{i,j}C(A,r)}{det(A)}\right| \geq \frac{\Delta}{t}\right) \\
&\quad \overset{(b)}{\leq} \sum_{r=0}^{t-1}\mathbf{Pr}\left(|[A_n^r]_{i,j} - [A^r]_{i,j}| \geq \epsilon_{i,j}\right) + \sum_{r=0}^{t-1}\mathbf{Pr}\left(|C(A_n,r) - C(A,r)| \geq \Delta_{i,j}\right) \\
&\qquad + \sum_{r=0}^{t-1}\mathbf{Pr}\left(\left|\frac{1}{det(A_n)} - \frac{1}{det(A)}\right| \geq \Delta_{i,j}\right)
\end{aligned}$$

Step (a) follows from Lemma 23 and step (b) from repeated applications of Lemma 24, with

$$\Delta_{i,j} = \frac{\Delta}{4t(1/|det(A)| + |[A^r]_{i,j}||C(A,r)|)(|[A^r]_{i,j}| + |C(A,r)|)} \tag{B.63}$$

Label the terms of step(b) as $T_{1,r} - T_{3,r}$ for $0 \leq r \leq t-1$. To complete the proof we will show each is upper bounded by $e^{-\kappa n\Delta^2}$.

We begin with term $T_{1,r}$. Note for $r = 0$, $[A_n^0]_{i,j} = [A^0]_{i,j} = 1$ for all $1 \leq i,j \leq t$ and so $T_{1,0} = 0$. The bound for $T_{1,1}$ follows from the assumption (B.59). So we show the bound holds for $r \geq 2$. Let $[A_n]_i$ and $[A]_i$ be the i^{th} columns of A_n and A, respectively. Then due

to the symmetry of A_n and A it follows:

$$
\begin{aligned}
&T_{1,r} \\
&= \mathbf{Pr}\left(\left|\sum_{u_1=1}^{t}\sum_{u_2=1}^{t}\dots\sum_{u_{r-1}=1}^{t}\left([A_n]_{i,u_1}[A_n]_{u_1,u_2}\dots[A_n]_{u_{r-1},j} - A_{i,u_1}A_{u_1,u_2}\dots A_{u_{r-1},j}\right)\right| \geq \Delta_{i,j}\right) \\
&\overset{(a)}{\leq} \sum_{u_1=1}^{t}\sum_{u_2=1}^{t}\dots\sum_{u_{r-1}=1}^{t}\mathbf{Pr}\left(\left|[A_n]_{i,u_1}[A_n]_{u_1,u_2}\dots[A_n]_{u_{r-1},j} - A_{i,u_1}A_{u_1,u_2}\dots A_{u_{r-1},j}\right| \geq \frac{\Delta_{i,j}}{t(r-1)}\right) \\
&\overset{(b)}{\leq} \sum_{u_1=1}^{t}\sum_{u_2=1}^{t}\dots\sum_{u_{r-1}=1}^{t}\left[\mathbf{Pr}\left(\left|[A_n]_{i,u_1} - A_{i,u_1}\right| \geq \tilde{\Delta}_{i,j}\right) + \dots + \mathbf{Pr}\left(\left|[A_n]_{u_{r-1},j} - A_{u_{r-1},j}\right| \geq \tilde{\Delta}_{i,j}\right)\right] \\
&\overset{(c)}{\leq} \sum_{u_1=1}^{t}\sum_{u_2=1}^{t}\dots\sum_{u_{r-1}=1}^{t}\left[e^{-\kappa n\tilde{\Delta}_{i,j}^2} + \dots + e^{-\kappa n\tilde{\Delta}_{i,j}^2}\right] \\
&= (r-1)\cdot t\cdot r\cdot e^{-\kappa n\tilde{\Delta}_{i,j}^2}.
\end{aligned}
$$

Step (a) follows from Lemma 23, step (b) from repeated use of Lemma 24, step (c) from assumption (B.59).

Next we bound term $T_{2,r}$. Using the definition in (B.61), it follows:

$$T_{2,r} = \mathbf{Pr}\left(\left|\sum_{k_1,\dots,k_{t-1}}\left(\prod_{s=1}^{t-1}\frac{(-1)^{k_s+1}}{s^{k_s}k_s!}tr(A_n^s)^{k_s} - \prod_{s=1}^{t-1}\frac{(-1)^{k_s+1}}{s^{k_s}k_s!}tr(A^s)^{k_s}\right)\right| \geq \Delta_{i,j}\right), \quad \text{(B.64)}$$

where the $k_1,\dots,k_{t-1}$ values are determined by (B.62). Note the number of sets of $\{k_1,\dots,k_{t-1}\}$ summed over in (B.64) is some constant value not depending on n. Therefore using (B.64) and the concentration of sums, Lemma 23, it follows:

$$T_{2,r} \leq \sum_{k_1,\dots,k_{t-1}}\mathbf{Pr}\left(\left|\prod_{s=1}^{t-1}tr(A_n^s)^{k_s} - \prod_{s=1}^{t-1}tr(A^s)^{k_s}\right| \geq \frac{\Delta_{i,j}}{c|c(k)|}\right), \quad \text{(B.65)}$$

where $c>0$ is the number of sets $\{k_1,\dots,k_{t-1}\}$ determined by (B.62) and $c(k)=\prod_{s=1}^{t-1}\frac{1}{s^{k_s}k_s!}$ is a constant depending on $\{k_1,\dots,k_{t-1}\}$. Now using (B.65) and repeated use of the concentration of products, Lemma 24, we find:

$$T_{2,r} \leq \sum_{k_1,\dots,k_{t-1}}\prod_{s=1}^{t-1}\mathbf{Pr}\left(\left|tr(A_n^s)^{k_s} - tr(A^s)^{k_s}\right| \geq \Delta'\right), \quad \text{(B.66)}$$

and we show for each $s \in [t-1]$ for some integer $k_s \geq 0$,

$$\mathbf{Pr}\left(\left|tr(A_n^s)^{k_s} - tr(A^s)^{k_s}\right| \geq \Delta\right) \leq e^{-\kappa n \Delta^2} \tag{B.67}$$

from which the desired bound on $T_{2,r}$ follows since we can use the concentration of products, Lemma 24, to bound upper bound (B.66).

$$T_{2,r} \leq \sum_{k_1,\ldots,k_{t-1}} \prod_{s=1}^{t-1} \mathbf{Pr}\left(\left|tr(A_n^s)^{k_s} - tr(A^s)^{k_s}\right| \geq \Delta'\right) \overset{(a)}{\leq} \sum_{k_1,\ldots,k_{t-1}} \prod_{s=1}^{t-1} e^{-\kappa n \Delta^2} \leq e^{-\kappa n \Delta^2}.$$

Step (a) follows from (B.67). We now prove (B.67). Note that we will prove for each $s \in [t-1]$,

$$\mathbf{Pr}\left(|tr(A_n^s) - tr(A^s)| \geq \Delta\right) \leq e^{-\kappa n \Delta^2},$$

and then result (B.67) follows via Lemma 27 since k_s is an integer. Now,

$$\begin{aligned}
\mathbf{Pr}\left(|tr(A_n^s) - tr(A^s)| \geq \Delta\right) &= \mathbf{Pr}\left(\left|\sum_{i=1}^{t}[A_n^s]_{i,i} - \sum_{i=1}^{t}[A^s]_{i,i}\right| \geq \Delta\right) \\
&\overset{(a)}{\leq} \sum_{i=1}^{t} \mathbf{Pr}\left(|[A_n^s]_{i,i} - [A^s]_{i,i}| \geq \frac{\Delta}{t}\right) \\
&\overset{(b)}{\leq} \sum_{i=1}^{t} e^{-\kappa n \Delta^2}.
\end{aligned}$$

Step (a) follows from Lemma 23, and step (b) by using a method similar to that used to bound term $T_{1,r}$ above since $s \in [t-1]$.

Finally to show the desired bound for $T_{3,2}$ we prove that

$$\mathbf{Pr}\left(|det(A_n) - det(A)| \geq \Delta_{i,j}\right) \leq e^{-\kappa n \Delta^2},$$

from which the desired bound for $T_{3,r}$ follows via the concentration of inverses, Lemma 29. By the Leibniz formula, the determinate of the matrix A can be represented as follows.

$$det(A) = \sum_{\sigma \in S_t} sgn(\sigma) \prod_{u=1}^{t} A_{u,\sigma_u}, \tag{B.68}$$

where the sum is taken over all permutations σ of the set $\{1, \ldots, t\}$ (the set of which we call S_t), the value σ_u is the value in the u^{th} position after the permutation, and $sgn(\sigma)$ equals either $+1$ or -1 and is the signature of the permutation. Using (B.68),

$$\begin{aligned}
\mathbf{Pr}\left(|det(A_n) - det(A)| \geq \Delta_{i,j}\right) &= \mathbf{Pr}\left(\left|\sum_{\sigma \in S_t} sgn(\sigma)\left(\prod_{u=1}^{t}[A_n]_{u,\sigma_u} - \prod_{u=1}^{t} A_{u,\sigma_u}\right)\right| \geq \Delta_{i,j}\right) \\
&\overset{(a)}{\leq} \sum_{\sigma \in S_t} \mathbf{Pr}\left(\left|\prod_{u=1}^{t}[A_n]_{u,\sigma_u} - \prod_{u=1}^{t} A_{u,\sigma_u}\right| \geq \frac{\Delta_{i,j}}{t!}\right) \\
&\overset{(b)}{\leq} \sum_{\sigma \in S_t} \prod_{u=1}^{t} \mathbf{Pr}\left(|[A_n]_{u,\sigma_u} - A_{u,\sigma_u}| \geq \tilde{\Delta}_{i,j}\right) \\
&\overset{(c)}{\leq} \sum_{\sigma \in S_t} \prod_{u=1}^{t} e^{-\kappa n \Delta^2} = t! \cdot t \cdot e^{-\kappa n \Delta^2}.
\end{aligned}$$

Step (a) follows from Lemma 23, step (b) from repeated application of Lemma 24, and step (c) from assumption (B.59).

□

B.5.2 Lipschitz Lemmas

Lemma 31 (Products of Lipschitz Functions)**.** *Let $f : \mathbb{R}^p \to \mathbb{R}$ and $g : \mathbb{R}^p \to \mathbb{R}$ be Lipschitz continuous. Let $\vec{s} = (s_1, \ldots, s_p)$ and $\vec{r} = (r_1, \ldots, r_p)$. Then $f \cdot g$ is pseudo-Lipschitz of order 2.*

Proof Lemma 31. Let f and g have Lipschitz constants $\mathbf{L}_f$ and $\mathbf{L}_g$, respectively. Then for some constants $\mathbf{L}_{f,0}$ and $\mathbf{L}_{g,0}$,

$$|f(\vec{s})| \leq \mathbf{L}_{f,0} + \mathbf{L}_f\|\vec{s}\|, \quad \text{and} \quad |g(\vec{s})| \leq \mathbf{L}_{g,0} + \mathbf{L}_g\|\vec{s}\|. \tag{B.69}$$

To see that this is true, note that it follows from the Lipschitz property of f that $\left|f(\vec{s}) - f(\vec{0})\right| \leq \mathbf{L}_f\|\vec{s}\|$, and therefore $|f(\vec{s})| \leq |f(\vec{0})| + \mathbf{L}_f\|\vec{s}\|$. The above result follows letting $\mathbf{L}_{f,0} = |f(\vec{0})|$.

The same reasoning gives the bound on $|g(\vec{s})|$. Therefore,

$$\begin{aligned}
|f(\vec{s})g(\vec{s}) - f(\vec{r})g(\vec{r})| &= |f(\vec{s})g(\vec{s}) - f(\vec{s})g(\vec{r}) + f(\vec{s})g(\vec{r}) - f(\vec{r})g(\vec{r})| \\
&\overset{(a)}{\leq} |f(\vec{s})|\,|g(\vec{s}) - g(\vec{r})| + |g(\vec{r})|\,|f(\vec{s}) - f(\vec{r})| \\
&\overset{(b)}{\leq} (\mathbf{L}_{f,0} + \mathbf{L}_f\|\vec{s}\|)\,\mathbf{L}_g\|\vec{s} - \vec{r}\| + (\mathbf{L}_{g,0} + \mathbf{L}_g\|\vec{s}\|)\,\mathbf{L}_f\|\vec{s} - \vec{r}\| \\
&= (\mathbf{L}_{g,0}\mathbf{L}_f + \mathbf{L}_g\mathbf{L}_{f,0} + \mathbf{L}_f\mathbf{L}_g\|\vec{s}\| + \mathbf{L}_f\mathbf{L}_g\|\vec{r}\|)\,\|\vec{s} - \vec{r}\| \\
&\overset{(c)}{\leq} \mathbf{L}\,(1 + \|\vec{s}\| + \|\vec{r}\|)\,\|\vec{s} - \vec{r}\|.
\end{aligned}$$

Step (a) follows from the Triangle Inequality, step (b) from the Lipschitz property of f and g along with (B.69), and step (c) by choosing, for example, $\mathbf{L} \geq \max(\mathbf{L}_{g,0}\mathbf{L}_f + \mathbf{L}_g\mathbf{L}_{f,0}, \mathbf{L}_g\mathbf{L}_f)$.

□

Lemma 32. *Let $\phi_h : \mathbb{R}^{t+2} \to \mathbb{R}$ be $PL(2)$ and $(c_0, c_1, \ldots, c_t)$ be constants. Then both of the following functions are also $PL(2)$.*

- *For $0 \leq i \leq N$, treating $\{h^1, \ldots, h^t, \beta_0\}$ as constants, $\phi^1_{h_i} : \mathbb{R} \to \mathbb{R}$ defined as*

$$\phi^1_{h_i}(s_i) := \phi_h\left(h_i^1, ..., h_i^t, \sum_{r=0}^{t-1} c_r h_i^{r+1} + c_t s_i, \beta_{0_i}\right). \tag{B.70}$$

- *$\phi^2_h : \mathbb{R}^{t+1} \to \mathbb{R}$ defined as*

$$\phi^2_h\left(h_i^1, \ldots, h_i^t, \beta_{0_i}\right) = \mathbb{E}_{Z_t}\left[\phi_h\left(h_i^1, \ldots, h_i^t, \sum_{r=0}^{t-1} c_r h_i^{r+1} + c_t Z_{t_i}, \beta_{0_i}\right)\right]. \tag{B.71}$$

Proof Lemma 32. First we show that the function $\phi^1_{h_i}$ defined in (B.70) is $PL(2)$ for each $i \in [N]$.

$$\begin{aligned}
\left|\phi^1_{h_i}(s) - \phi^1_{h_i}(s')\right| &= \left|\phi_h\left(h_i^1, ..., h_i^t, \sum_{r=0}^{t-1} c_r h_i^{r+1} + c_t s, \beta_{0_i}\right) - \phi_h\left(h_i^1, ..., h_i^t, \sum_{r=0}^{t-1} c_r h_i^{r+1} + c_t s', \beta_{0_i}\right)\right| \\
&\overset{(a)}{\leq} \mathbf{L}\left[1 + 2\left(|h_i^1| + \ldots + |h_i^t| + |\beta_{0_i}|\right) + 2|\sum_{r=0}^{t-1} c_r h_i^{r+1}| + c_t|s| + c_t|s'|\right] c_t|s - s'| \\
&\overset{(b)}{\leq} \mathbf{L'}\left(1 + |s| + |s'|\right)|s - s'|.
\end{aligned}$$

Step (a) follows from the fact that $\phi_h \in PL(2)$ and the Triangle Inequality and step (b) follows since all terms $\{h^1, \ldots, h^t, \beta_0\}$ are treated as constants. We have therefore shown that $\phi^1_{h_i}$ is $PL(2)$.

Next we show that the function $\phi^2_{h_i}$ defined in (B.71) is $PL(2)$.

$$
\begin{aligned}
&\left|\phi^2_h\left(h^1_i, \ldots, h^t_i, \beta_{0_i}\right) - \phi^2_h\left(\tilde{h}^1_i, \ldots, \tilde{h}^t_i, \tilde{\beta}_{0_i}\right)\right| \\
&\overset{(a)}{\leq} \mathbb{E}_{Z_t}\left|\phi_h\left(h^1_i, \ldots, h^t_i, \sum_{r=0}^{t-1} c_r h_i^{r+1} + c_t Z_{t_i}, \beta_{0_i}\right) - \phi_h\left(\tilde{h}^1_i, \ldots, \tilde{h}^t_i, \sum_{r=0}^{t-1} c_r \tilde{h}_i^{r+1} + c_t Z_{t_i}, \tilde{\beta}_{0_i}\right)\right| \\
&\overset{(b)}{\leq} \mathbf{L}\left[1 + \mathbb{E}_{Z_t}\left|\left|\left(h^1_i, \ldots, h^t_i, \sum_{r=0}^{t-1} c_r h_i^{r+1} + c_t Z_{t_i}, \beta_{0_i}\right)\right|\right| + \mathbb{E}_{Z_t}\left|\left|\left(\tilde{h}^1_i, \ldots, \tilde{h}^t_i, \sum_{r=0}^{t-1} c_r \tilde{h}_i^{r+1} + c_t Z_{t_i}, \tilde{\beta}_{0_i}\right)\right|\right|\right] \\
&\quad \times \left|\left|\left(h^1_i - \tilde{h}^1_i, \ldots, h^t_i - \tilde{h}^t_i, \sum_{r=0}^{t-1} c_r (h_i^{r+1} - \tilde{h}_i^{r+1}), \beta_{0_i} - \tilde{\beta}_{0_i}\right)\right|\right|
\end{aligned}
$$

Step (a) follows from Jensen's Inequality and step (b) from the fact that $\phi_h \in PL(2)$. To complete the proof, we will show for some constants $\kappa_1, \kappa_2 > 0$,

$$
\mathbb{E}_{Z_t}\left|\left|\left(h^1_i, \ldots, h^t_i, \sum_{r=0}^{t-1} c_r h_i^{r+1} + c_t Z_{t_i}, \beta_{0_i}\right)\right|\right| \leq \kappa_1 \left|\left|\left(h^1_i, \ldots, h^t_i, \beta_{0_i}\right)\right|\right| + \kappa_2 \tag{B.72}
$$

$$
\mathbb{E}_{Z_t}\left|\left|\left(\tilde{h}^1_i, \ldots, \tilde{h}^t_i, \sum_{r=0}^{t-1} c_r \tilde{h}_i^{r+1} + c_t Z_{t_i}, \tilde{\beta}_{0_i}\right)\right|\right| \leq \kappa_1 \left|\left|\left(\tilde{h}^1_i, \ldots, \tilde{h}^t_i, \tilde{\beta}_{0_i}\right)\right|\right| + \kappa_2 \tag{B.73}
$$

$$
\left|\left|\left(h^1_i - \tilde{h}^1_i, \ldots, h^t_i - \tilde{h}^t_i, \sum_{r=0}^{t-1} c_r (h_i^{r+1} - \tilde{h}_i^{r+1}), \beta_{0_i} - \tilde{\beta}_{0_i}\right)\right|\right| \leq \sqrt{\kappa}\left|\left|\left(h^1_i - \tilde{h}^1_i, \ldots, h^t_i - \tilde{h}^t_i, \beta_{0_i} - \tilde{\beta}_{0_i}\right)\right|\right|. \tag{B.74}
$$

First we show (B.72). Note that

$$
\begin{aligned}
\left|\left|\left(h^1_i, \ldots, h^t_i, \sum_{r=0}^{t-1} c_r h_i^{r+1} + c_t Z_{t_i}, \beta_{0_i}\right)\right|\right|^2 &= \sum_{r=0}^{t-1} (h_i^{r+1})^2 + \left(\sum_{r=0}^{t-1} c_r h_i^{r+1} + c_t Z_{t_i}\right)^2 + \beta^2_{0_i} \\
&\overset{(a)}{\leq} \sum_{r=0}^{t-1} (h_i^{r+1})^2 (1 + (t+1)(c_r)^2) + (t+1)(c_t)^2 Z^2_{t_i} + \beta^2_{0_i} \\
&\leq \kappa \left|\left|\left(h^1_i, \ldots, h^t_i, Z_{t_i}, \beta_{0_i}\right)\right|\right|^2.
\end{aligned}
$$

We have defined κ a constant such that $\kappa \leq 1+(t+1)\max_{0\leq r\leq t}\left(c_r^2\right)$. Step (a) follows from Lemma 38. Using the above, it follows:

$$\mathbb{E}_{Z_t}\left|\left|\left(h_i^1,\ldots,h_i^t,\sum_{r=0}^{t-1}c_r h_i^{r+1}+\tau_t^{\perp}Z_{t_i},\beta_{0_i}\right)\right|\right| \leq \sqrt{\kappa}\mathbb{E}_{Z_t}\left|\left|\left(h_i^1,\ldots,h_i^t,Z_{t_i},\beta_{0_i}\right)\right|\right| \overset{(a)}{\leq} \sqrt{\kappa}\left|\left|\left(h_i^1,\ldots,h_i^t,\beta_{0_i}\right)\right|\right|+\sqrt{\kappa}\mathbb{E}_{Z_t}|Z_{t_i}|.$$

Step (a) follows from the Triangle Inequality. This proves (B.72) for $c_1=\sqrt{\kappa}$ and $c_2=\sqrt{2\kappa/\pi}$. Upper bounds (B.73) and (B.74) can be shown similarly. □

Lemma 33 (Gradient of Lipschitz and Pseudo-Lipschitz Functions are Bounded)**.** *Let $f_L:\mathbb{R}^N\to\mathbb{R}$ be Lipschitz continuous with Lipschitz constant $\boldsymbol{L}$ and everywhere differentiable. Let $f_{PL}:\mathbb{R}^N\to\mathbb{R}$ be pseudo-Lipschitz constant $\boldsymbol{L}$ and everywhere differentiable. Then for any vector $x:=(x_1,\ldots,x_N)\in\mathbb{R}^N$,*

$$\|\nabla f_L(x)\|\leq \boldsymbol{L},$$
$$\|\nabla f_{PL}(x)\|\leq \boldsymbol{L}(1+2\|x\|).$$

Proof Lemma 33. From Taylor's Theorem, for any $x,\Delta\in\mathbb{R}^N$ and any function $f:\mathbb{R}^N\to\mathbb{R}$,

$$f(x+\Delta)=f(x)+[\nabla f(x+\xi\Delta)]^*\Delta, \tag{B.75}$$

for some $\xi\in(0,1)$. From (B.75) it follows,

$$\left|[\nabla f(x+\xi\Delta)]^*\frac{\Delta}{\|\Delta\|}\right|=\frac{|f(x+\Delta)-f(x)|}{\|\Delta\|}.$$

Using the above,

$$\left|[\nabla f_L(x+\xi\Delta)]^*\frac{\Delta}{\|\Delta\|}\right|\leq \mathbf{L},$$
$$\left|[\nabla f_{PL}(x+\xi\Delta)]^*\frac{\Delta}{\|\Delta\|}\right|\leq \mathbf{L}(1+\|x+\Delta\|+\|x\|).$$

The result follows by letting $\Delta=\epsilon\nabla f(x)$ for $\epsilon>0$ and taking $\epsilon\to 0$. □

B.5.3 Gaussian Concentration Lemmas

Lemma 34 (Pseudo-Lipschitz Functions of Gaussians Concentrate)**.** *Let $Z \in \mathbb{R}^N$ be a random vector with entries that are i.i.d. standard Gaussian and let $f_i : \mathbb{R} \to \mathbb{R}$ such that $f_i \in PL(2)$ for each $i \in [N]$. Then it follows,*

$$\boldsymbol{Pr}\left(\left|\frac{1}{N}\sum_{i=1}^{N} f_i\left(Z_i\right) - \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[f_i\left(Z_i\right)\right]\right| \geq \Delta\right) \leq e^{-\kappa N\Delta^2}.$$

Proof of Lemma 34. We use a proof method of Maury and Pisier generalized for pseudo-Lipschitz functions f_i for each $i \in [N]$. Without loss of generality, assume $\mathbb{E}[f_i(Z_i)] = 0$. (Otherwise subtract a constant from f_i). In what follows we demonstrate the upper-tail case:

$$\mathbf{Pr}\left(\frac{1}{N}\sum_{i=1}^{N} f_i\left(Z_i\right) \geq \Delta\right) \leq e^{-\kappa N\Delta^2}, \tag{B.76}$$

and the lower-tail bound follows by symmetry to give the desired result. Let $\mathbf{L}_i$ be the pseudo-Lipschitz constant associated with function f_i for each $i \in [N]$. To show (B.76) we will show that the following is true for all $0 < t < 1/\pi\mathbf{L}\sqrt{2}$ where $\mathbf{L} = \max_i \mathbf{L}_i$ and some $\kappa' > 0$,

$$\mathbb{E}\left[\exp\left(t\sum_{i=1}^{N} f_i\left(Z_i\right)\right)\right] \leq \exp(\kappa' N t^2). \tag{B.77}$$

Using (B.77), result (B.76) follows via the Cramer-Chernoff method:

$$\begin{aligned}
\mathbf{Pr}\left(\frac{1}{N}\sum_{i=1}^{N} f_i\left(Z_i\right) \geq \Delta\right) &= \mathbf{Pr}\left(\exp\left(t\sum_{i=1}^{N} f_i\left(Z_i\right)\right) \geq \exp(tN\Delta)\right) \\
&\leq \mathbb{E}\left[\exp\left(t\sum_{i=1}^{N} f_i\left(Z_i\right)\right)\right]\exp(-tN\Delta) \\
&\overset{(a)}{\leq} e^{\kappa' N t^2 - tN\Delta}.
\end{aligned}$$

Step (a) follows from (B.77). Result (B.76) follows from the above work by minimizing over $0 < t < 1/\pi\mathbf{L}\sqrt{2}$ at choice $t_{min} = \Delta/2\kappa'$. Note that $\Delta > 0$ is small, so $t_{min} < 1/\pi\mathbf{L}\sqrt{2}$.

We now prove (B.77). For $i \in [N]$, let $\tilde{Z}_i$ be an independent copy of Z_i. Using Jensen's

Inequality and the fact that $\mathbb{E}f_i(\tilde{Z}_i) = 0$,

$$\mathbb{E}\exp\left(-tf_i(\tilde{Z}_i)\right) \geq \exp\left(-t\mathbb{E}f_i(\tilde{Z}_i)\right) = 1. \tag{B.78}$$

Since $\tilde{Z}$ and Z are independent, using (B.78),

$$\mathbb{E}\exp\left(tf_i(Z_i)\right) \leq \mathbb{E}\exp\left(tf_i(Z_i)\right) \times \mathbb{E}\exp\left(-tf_i(\tilde{Z}_i)\right) = \mathbb{E}\exp\left(t[f_i(Z_i) - f_i(\tilde{Z}_i)]\right). \tag{B.79}$$

Note that we can represent the difference using the following integral:

$$f_i(Z_i) - f_i(\tilde{Z}_i) = \int_0^{\pi/2} \frac{\partial}{\partial\theta} f_i(\tilde{Z}_i\cos\theta + Z_i\sin\theta)d\theta. \tag{B.80}$$

Keep in mind the following three facts:

- The random variable $U_{i,\theta} := \tilde{Z}_i\cos\theta + Z_i\sin\theta$ has the same distribution as Z_i.

  This is true because $U_{i,\theta}$ is the sum of independent Gaussians and therefore is also Gaussian. It has zero mean, since $\mathbb{E}Z_i = \mathbb{E}\tilde{Z}_i = 0$ and variance equal to

$$\begin{aligned} VAR[U_{i,\theta}] = \mathbb{E}\left[\left(\tilde{Z}_i\cos\theta + Z_i\sin\theta\right)^2\right] &= (\cos\theta)^2\mathbb{E}\left[\tilde{Z}_i^2\right] + (\sin\theta)^2\mathbb{E}\left[Z_i^2\right] \\ &= (\cos\theta)^2 + (\sin\theta)^2 = 1. \end{aligned}$$

- The random variable $\frac{\partial}{\partial\theta}U_{i,\theta} = V_{i,\theta} = -\tilde{Z}_i\sin\theta + Z_i\cos\theta$ has the same distribution as Z_i. The justification is very similar to the above.

- The random variables $U_{i,\theta}$ and $V_{i,\theta}$ are independent. To see this, note that their covariance equals 0:

$$\begin{aligned} COV[U_{i,\theta}V_{i,\theta}] = \mathbb{E}\left[U_{i,\theta}V_{i,\theta}\right] = \mathbb{E}\left[\left(\tilde{Z}_i\cos\theta + Z_i\sin\theta\right)\left(-\tilde{Z}_i\sin\theta + Z_i\cos\theta\right)\right] \\ = -(\cos\theta)(\sin\theta)\mathbb{E}\left[\tilde{Z}_i^2\right] + (\cos\theta)(\sin\theta)\mathbb{E}\left[Z_i^2\right] \\ = -(\cos\theta)(\sin\theta) + (\cos\theta)(\sin\theta) = 0. \end{aligned}$$

Next, using (B.80),

$$\begin{aligned}
\exp\left(t[f_i(Z_i) - f_i(\tilde{Z}_i)]\right) &= \exp\left(t\int_0^{\pi/2} \frac{\partial}{\partial\theta} f_i(U_{i,\theta})d\theta\right) \\
&= \exp\left(\frac{t\pi}{2}\int_0^{\pi/2} \frac{2}{\pi}\frac{\partial}{\partial\theta} f_i(U_{i,\theta})d\theta\right) \\
&\overset{(a)}{\leq} \int_0^{\pi/2} \frac{2}{\pi}\exp\left(\frac{t\pi}{2}\frac{\partial}{\partial\theta} f_i(U_{i,\theta})\right)d\theta. \\
&= \int_0^{\pi/2} \frac{2}{\pi}\exp\left(\frac{t\pi}{2} f_i'(U_{i,\theta})V_{i,\theta}\right)d\theta. \qquad \text{(B.81)}
\end{aligned}$$

Step (a) follows from Jensen's Inequality and (B.81) from the chain rule.

Now we take expectation on both sides with respect to the product measure on $(Z_i, \tilde{Z}_i)$, which are i.i.d. $\mathcal{N}(0,1)$. For a fixed θ, the pair $(U_{i,\theta}, V_{i,\theta})$ given by

$$\begin{bmatrix} U_{i,\theta} \\ V_{i,\theta} \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} Z_i \\ \tilde{Z}_i \end{bmatrix}$$

is also i.i.d. $\mathcal{N}(0,1)$. Therefore taking expectation with respect to $(Z_i, \tilde{Z}_i)$ on the RHS of (B.81) is the same as taking expectation with respect to $(U_{i,\theta}, V_{i,\theta})$. Therefore we obtain

$$\mathbb{E}_{Z,\tilde{Z}} \exp\left(t[f_i(Z_i) - f_i(\tilde{Z}_i)]\right) \leq \int_0^{\pi/2} \frac{2}{\pi}\mathbb{E}_{U_{i,\theta},V_{i,\theta}} \exp\left(\frac{t\pi}{2} f_i'(U_{i,\theta})V_{i,\theta}\right)d\theta. \qquad \text{(B.82)}$$

Considering just the expectation on the right side of (B.82),

$$\begin{aligned}
\mathbb{E}_{U_{i,\theta},V_{i,\theta}} \exp\left(\frac{t\pi}{2} f_i'(U_{i,\theta})V_{i,\theta}\right) &= \mathbb{E}_{U_{i,\theta},V_{i,\theta}} \exp\left(\frac{t\pi}{2} f_i'(U_{i,\theta})V_{i,\theta}\right) \\
&\overset{(a)}{\leq} \mathbb{E}_{U_{i,\theta}}\mathbb{E}_{V_{i,\theta}|U_{i,\theta}} \exp\left(\frac{t\pi}{2}\mathbf{L}_i(1+2|U_{i,\theta}|)V_{i,\theta}\right) \\
&\overset{(b)}{=} \mathbb{E}_{U_{i,\theta}} \exp\left(\frac{1}{2}\left(\frac{t\pi\mathbf{L}_i(1+2|U_{i,\theta}|)}{2}\right)^2\right) \\
&\leq \exp\left(\frac{(t\pi\mathbf{L})^2}{4}\right)\mathbb{E}_{U_{i,\theta}} \exp\left((t\pi\mathbf{L}_i)^2|U_{i,\theta}|^2)\right). \qquad \text{(B.83)}
\end{aligned}$$

Step (a) follows from Lemma 33, step (b) follows by the moment-generating function of the standard Gaussian and the fact that $U_{i,\theta}$ and $V_{i,\theta}$ are independent, and (B.83) follows from

Lemma 38. Finally we will show for $t \leq 1/\pi \mathbf{L}_i \sqrt{2}$,

$$\mathbb{E}_{U_{i,\theta}} \exp\left((t\pi \mathbf{L}_i)^2 |U_{i,\theta}|^2\right) \leq e^{\kappa t^2}, \tag{B.84}$$

and then from (B.79), (B.82), (B.83), and (B.84) it follows:

$$\mathbb{E} \exp\left(t \sum_{i=1}^{N} f_i(Z_i)\right) \leq e^{\kappa N t^2},$$

which is result (B.77). We now demonstrate (B.84). First note, for $t \leq 1/\pi \mathbf{L}_i \sqrt{2}$,

$$\mathbb{E}_{U_{i,\theta}} \exp\left((t\pi \mathbf{L}_i)^2 |U_{i,\theta}|^2\right) = \left(\frac{1}{1-(t\pi \mathbf{L}_i)^2}\right)^{\frac{1}{2}}. \tag{B.85}$$

Note that for $0 < x \leq 1/2$, we have the following bound: $1 - x \geq e^{-2x}$. Applying this to (B.85) we find

$$\mathbb{E}_{U_{i,\theta}} \exp\left((t\pi \mathbf{L}_i)^2 |U_{i,\theta}|^2\right) = \left(\frac{1}{1-(t\pi \mathbf{L}_i)^2}\right)^{\frac{1}{2}} \leq e^{t^2(\pi \mathbf{L}_i)^2},$$

when $t \leq 1/\pi \mathbf{L}_i \sqrt{2}$. When optimizing over t above, we use $t = c\Delta$ for some constant c, so t is sufficiently small. □

Fact 15 (Sub-Gaussian RV [Boucheron-Lugosi-Massart pp. 24–27])**.** *A zero-mean random variable is said to be sub-Gaussian with variance factor ν if $\mathbb{E}[e^{tX}] \leq \frac{t^2\nu}{2}$ for all $t \in \mathbf{R}$. A sub-Gaussian rv X with variance factor ν satisfies the following:*

1. *For all $x > 0$, $P(X > x) \vee P(X < -x) \leq e^{-\frac{x^2}{2\nu}}$, for all $x > 0$.*
2. *For every integer $k \geq 1$,*

$$\mathbb{E}[X^{2k}] \leq 2(k!)(2\nu)^k. \tag{B.86}$$

Lemma 35 (Pseudo-Lipschitz Functions of Sub-Gaussians Concentrate)**.** *Let $f : \mathbb{R} \to \mathbb{R}$ be a function $\in PL(2)$ with PL constant $\boldsymbol{L}$. Let $Z \in \mathbb{R}^{\mathbb{N}}$ be a random vector with entries which are i.i.d. with distribution p_Z which is zero mean and sub-Gaussian. Let Z' be a random*

variable such that $Z' \sim p_Z$. *Then,*

$$\boldsymbol{Pr}\left(\left|\frac{1}{N}\sum_{i=1}^{N} f\left(Z_i\right) - \mathbb{E}\left[f\left(Z'\right)\right]\right| \geq \Delta\right) \leq e^{-\kappa N \Delta^2}.$$

Proof of Lemma 35. Without loss of generality, assume $\mathbb{E}[f(Z')] = 0$. In what follows we demonstrate the upper-tail case:

$$\mathbf{Pr}\left(\frac{1}{N}\sum_{i=1}^{N} f\left(Z_i\right) \geq \Delta\right) \leq e^{-\kappa N \Delta^2}, \tag{B.87}$$

and the lower-tail bound follows similarly. To show (B.87) we will show that the following is true for all $0 < t < 1/(2\sqrt{2}\nu\mathbf{L})$ and some $\kappa' > 0$,

$$\mathbb{E}\left[\exp\left(t\sum_{i=1}^{N} f\left(Z_i\right)\right)\right] \leq \exp(\kappa' N t^2). \tag{B.88}$$

Using (B.88), result (B.87) follows via the Cramer-Chernoff method:

$$\begin{aligned}
\mathbf{Pr}\left(\frac{1}{N}\sum_{i=1}^{N} f\left(Z_i\right) \geq \Delta\right) &= \mathbf{Pr}\left(\exp\left(t\sum_{i=1}^{N} f\left(Z_i\right)\right) \geq \exp(tN\Delta)\right) \\
&\leq \mathbb{E}\left[\exp\left(t\sum_{i=1}^{N} f\left(Z_i\right)\right)\right]\exp(-tN\Delta) \\
&\overset{(a)}{\leq} e^{\kappa' N t^2 - tN\Delta}.
\end{aligned}$$

Step (a) follows from (B.88) and then minimizing over $0 < t < 1/(2\sqrt{2}\nu\mathbf{L})$ gives result (B.87) for the choice $t_{min} = \frac{\Delta}{2\kappa'}$. Note that $\Delta > 0$ is small, so $t_{min} < 1/(2\sqrt{2}\nu\mathbf{L})$.

We now prove (B.88). For $i \in [N]$, let $\tilde{Z}_i$ be an independent copy of Z_i. Using Jensen's Inequality and the fact that $\mathbb{E}f(\tilde{Z}_i) = 0$,

$$\mathbb{E}\exp\left(-tf(\tilde{Z}_i)\right) \geq \exp\left(-t\mathbb{E}f(\tilde{Z}_i)\right) = 1.$$

Since $\tilde{Z}$ and Z are independent, using the above,

$$\mathbb{E}\exp\left(tf(Z_i)\right) \leq \mathbb{E}\exp\left(tf(Z_i)\right) \times \mathbb{E}\exp\left(-tf(\tilde{Z}_i)\right) = \mathbb{E}\exp\left(t[f(Z_i) - f(\tilde{Z}_i)]\right). \quad \text{(B.89)}$$

We prove (B.88) by demonstrating that for each $i \in [N]$,

$$\mathbb{E}\left[\exp\left(t\left(f(Z_i) - f(\tilde{Z}_i)\right)\right)\right] \leq \exp(\kappa' t^2). \quad \text{(B.90)}$$

Then (B.88) follows by (B.89) and (B.90) since

$$\begin{aligned}\mathbb{E}\exp\left(t\sum_{i=1}^{N}(Z_i)\right) = \prod_{i=1}^{N}\mathbb{E}\exp\left(t(Z_i)\right) &\leq \prod_{i=1}^{N}\mathbb{E}\exp\left(t[f(Z_i) - f(\tilde{Z}_i)]\right) \\ &\leq \prod_{i=1}^{N}\exp(\kappa' t^2) = \exp(\kappa' N t^2).\end{aligned}$$

So we show (B.90). For each $i \in [N]$,

$$\begin{aligned}\mathbb{E}\left[\exp\left(t\left(f(Z_i) - f(\tilde{Z}_i)\right)\right)\right] &= \sum_{q=0}^{\infty}\frac{t^q}{q!}\cdot\mathbb{E}\left(f(Z_i) - f(\tilde{Z}_i)\right)^q \\ &\overset{(a)}{=} \sum_{k=0}^{\infty}\frac{t^{2k}}{(2k)!}\cdot\mathbb{E}\left(f(Z_i) - f(\tilde{Z}_i)\right)^{2k}.\end{aligned}$$

Step (a) follows since the odd moments of the difference $f(Z_i) - f(\tilde{Z}_i)$ equal 0. Now using the above we find:

$$\begin{aligned}\mathbb{E}\left[\exp\left(t\left(f(Z_i) - f(\tilde{Z}_i)\right)\right)\right] &= \sum_{k=0}^{\infty}\frac{t^{2k}}{(2k)!}\cdot\mathbb{E}\left(f(Z_i) - f(\tilde{Z}_i)\right)^{2k} \\ &\leq 2c'\sum_{k=0}^{\infty}\frac{(t\mathbf{L})^{2k}}{(2k)!}\cdot\mathbb{E}|Z_i|^{2k} + 2c'\sum_{k=0}^{\infty}\frac{(t\mathbf{L})^{2k}}{(2k)!}\cdot\mathbb{E}|Z_i|^{4k}. \quad \text{(B.91)}\end{aligned}$$

Result (B.91) is obtained using the pseudo-Lipschitz property of f. Since the Z_i's are sub-Gaussian, the first term in (B.91) is

$$\sum_{k=0}^{\infty}\frac{(t\mathbf{L})^{2k}}{(2k)!}\cdot\mathbb{E}|Z_i|^{2k} \leq \sum_{k=0}^{\infty}\frac{(t\mathbf{L})^{2k}}{(2k)!}2(k!)(2\nu)^k \overset{(a)}{\leq} 2\sum_{k=0}^{\infty}\frac{(t^2\mathbf{L}^2\nu)^k}{k!} = 2e^{t^2\mathbf{L}^2\nu}.$$

Step (a) above is obtained using the inequality $\frac{(2k)!}{k!} \geq 2^k k!$, which is obtained as follows.

$$\frac{(2k)!}{k!} = \prod_{j=1}^{k}(k+j) = k! \prod_{j=1}^{k}\left(\frac{k}{j}+1\right) \geq (k!)2^k.$$

The second term in (B.91) is

$$\sum_{k=0}^{\infty} \frac{(t\mathbf{L})^{2k}}{(2k)!} \cdot \mathbb{E}|Z_i|^{4k} \leq \sum_{k=0}^{\infty} \frac{(t\mathbf{L})^{2k}}{(2k)!} 2(2k!)(2\nu)^{2k} = 2\sum_{k=0}^{\infty}(2\nu t\mathbf{L})^{2k} = \frac{2}{1-(2\nu t\mathbf{L})^2}$$
$$\leq 2e^{8\nu^2\mathbf{L}^2t^2}$$

for $t < 1/(2\sqrt{2}\nu\mathbf{L})$.

□

Lemma 36 (Normal Random Variables)**.** *Let Z be a standard Gaussian random variable. Then it follows, for all $\Delta > 0$,*

$$\boldsymbol{Pr}\left(|Z| \geq \Delta\right) \leq 2e^{-\frac{1}{2}\Delta^2}.$$

Proof of Lemma 36. First note,

$$\mathbf{Pr}\left(|Z| \geq \Delta\right) = \mathbf{Pr}\left(Z \geq \Delta \text{ or } Z \leq -\Delta\right) \leq \mathbf{Pr}\left(Z \geq \Delta\right) + \mathbf{Pr}\left(Z \leq -\Delta\right).$$

We will show

$$\mathbf{Pr}\left(Z \geq \Delta\right) \leq \exp\left(-\frac{1}{2}\Delta^2\right).$$

The lower tail result follows similarly to give the desired result. A Cramer-Chernoff bound gives

$$\mathbf{Pr}\left(Z \geq \Delta\right) \leq \exp\left(\inf_{\lambda>0}\left[-\lambda\Delta + \log \mathbb{E}e^{\lambda Z}\right]\right). \tag{B.92}$$

For a standard Gaussian random variable, $\mathbb{E}\left[\exp(\lambda Z)\right] = \exp\left(\frac{\lambda^2}{2}\right)$. Plugging this into (B.92) we find,

$$\mathbf{Pr}\left(Z \geq \Delta\right) \leq \exp\left(\inf_{\lambda>0}\left[-\lambda\Delta + \frac{\lambda^2}{2}\right]\right) \overset{(a)}{=} \exp\left(-\frac{\Delta^2}{2}\right). \tag{B.93}$$

Step (a) follows from setting $\lambda = \Delta$, it's minimizing value. $\square$

Lemma 37 (Sum of Gaussian Squares)**.** *For i.i.d. standard Gaussian random variables, Z_i, with $i = 1, 2, \ldots, n$ and for $\epsilon \leq 1$*

$$\boldsymbol{Pr}\left(\left|\frac{\sum_{i=1}^{n} Z_i^2}{n} - 1\right| \geq \Delta\right) \leq e^{-\kappa n \Delta^2}.$$

Proof. First note that

$$\begin{aligned}\mathbf{Pr}\left(\left|\frac{\sum_{i=1}^{n} Z_i^2}{n} - 1\right| \geq \Delta\right) &= \mathbf{Pr}\left(\frac{\sum_{i=1}^{n} Z_i^2}{n} - 1 \geq \epsilon \text{ or } \frac{\sum_{i=1}^{n} Z_i^2}{n} - 1 \leq -\Delta\right) \\ &\leq \mathbf{Pr}\left(\sum_{i=1}^{n} Z_i^2 - n \geq n\Delta\right) + \mathbf{Pr}\left(-\left[\sum_{i=1}^{n} Z_i^2 - n\right] \geq n\Delta\right).\end{aligned} \tag{B.94}$$

We bound both terms of (B.94) by $e^{-\kappa n \Delta^2}$ to complete the proof.

Consider the first term of (B.94). A Cramer-Chernoff bound gives

$$\mathbf{Pr}\left(\sum_{i=1}^{n} Z_i^2 - n \geq n\Delta\right) \leq \exp\left(\inf_{\lambda > 0}\left[-\lambda n(1+\Delta) + \log \mathbb{E} e^{\lambda \sum_{i=1}^{n} Z_i^2}\right]\right). \tag{B.95}$$

For a chi-square random variable, $\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n} Z_i^2\right)\right] = (1-2\lambda)^{-n/2}$ for $0 \leq \lambda \leq 1/2$. Plugging this into (B.95) we find,

$$\begin{aligned}\mathbf{Pr}\left(\sum_{i=1}^{n} Z_i^2 - n \geq n\Delta\right) &\leq \exp\left(\inf_{1/2 > \lambda > 0}\left[-\lambda n\Delta + \frac{n}{2}\left(-2\lambda - \log\left(1-2\lambda\right)\right)\right]\right) \\ &\overset{(a)}{\leq} \exp\left(\inf_{1/2 > \lambda > 0}\left[-\lambda n\Delta + \frac{n\lambda^2}{1-2\lambda}\right]\right) \\ &\overset{(b)}{\leq} \exp\left(-\frac{n\Delta^2}{4}\left(1 - \frac{1}{2}\cdot\frac{1}{2-\Delta}\right)\right).\end{aligned}$$

Step (a) follows from the fact that $-u - \log(1-u) \leq \frac{u^2}{2(1-u)}$ when $0 \leq u \leq 1$. Step (b) follows by choosing $\lambda = \Delta/4$.

Consider the second term of (B.94). A Cramer-Chernoff bound gives

$$\mathbf{Pr}\left(-\left[\sum_{i=1}^{n} Z_i^2 - n\right] \geq n\Delta\right) \leq \exp\left(\inf_{\lambda > 0}\left[\lambda n(1-\Delta) + \log \mathbb{E} e^{-\lambda \sum_{i=1}^{n} Z_i^2}\right]\right). \tag{B.96}$$

For the sum of squared Gaussians, $\mathbb{E}\left[\exp\left(-\lambda\sum_{i=1}^{n} Z_i^2\right)\right] = (1+2\lambda)^{-n/2}$ for $0 \leq \lambda$. Plugging this into (B.96) we find,

$$\begin{aligned}\mathbf{Pr}\left(-\left[\sum_{i=1}^{n} Z_i^2 - n\right] \geq n\Delta\right) &\leq \exp\left(\inf_{\lambda>0}\left[-\lambda n\Delta + \frac{n}{2}\left(2\lambda - \log\left(1+2\lambda\right)\right)\right]\right) \\ &\overset{(a)}{\leq} \exp\left(\inf_{\lambda>0}\left[-\lambda n\Delta + n\lambda^2\right]\right) \\ &\overset{(b)}{\leq} \exp\left(-\frac{n\Delta^2}{4}\right).\end{aligned}$$

Step (a) follows from the fact that $u - \log(1+u) \leq \frac{u^2}{2}$ when $u \geq 0$. Step (b) follows form setting $\lambda = \frac{\Delta}{2}$, it's minimizing value.

□

B.5.4 Other useful Lemmas

Lemma 38 (Squared Sums)**.** *For any* $a_1, \ldots, a_t$, $(a_1 + \ldots + a_t)^2 \leq t\sum_{i=1}^{t} a_i^2$.

Lemma 39. *For an* $n \times n$ *symmetric matrix* A *with eigenvalues* $\lambda_1, \ldots, \lambda_n$ *and a vector* $x \in \mathbb{R}^n$*, for each element* $0 \leq i \leq n$*,*

$$|[A^{-1}x]_i| \leq \|x\| \sum_{k=1}^{n} \left|\frac{1}{\lambda_k}\right|.$$

Proof. We can represent the symmetric matrix A^{-1} as UDU^* where D is an $n \times n$ diagonal matrix with the eigenvalues $1/\lambda_1, \ldots, 1/\lambda_n$ along the diagonal and the columns of the $n \times n$ matrix U form an orthonormal basis for the column space of A^{-1}. Then,

$$A^{-1}x = \sum_{k=1}^{n} \frac{1}{\lambda_k}\, \underline{u}_k\, \underline{u}_k^* x,$$

where $\underline{u}_i$ is the i^{th} column of U and is therefore orthonormal. Then we can represent each element as follows.

$$[A^{-1}x]_i = \sum_{k=1}^{n} \left(\frac{\underline{u}_k^* x}{\lambda_k}\right) \underline{u}_{k_i},$$

and so

$$|[A^{-1}x]_i| \leq \sum_{k=1}^{n} \left|\frac{1}{\lambda_k}\right| |\underline{u}_k^* x| \, |\underline{u}_{k_i}| \overset{a}{\leq} \|x\| \sum_{k=1}^{n} \left|\frac{1}{\lambda_k}\right|.$$

where step (a) uses Cauchy-Schwarz and the facts that $\|\underline{u}_k\| = 1$ and $|\underline{u}_{k_i}| \leq 1$ for each $i \in [n]$. $\square$

Appendix C

Chapter 4 Appendix

C.1 Proof of Lemma 7

Recall the definition of Shannon entropy in (4.9). We want to consider the expectation under the true joint distribution of the log ratio of the distributions,

$$\begin{aligned} D(\mathbb{Q}_L||\mathbb{Q}_L^a) &= \mathbb{E}_{\mathbb{Q}_L}\left[\log \frac{q_L(X_1,...,X_L,Y)}{q_L^a(X_1,...,X_L,Y)}\right] \\ &= \mathbb{E}_{\mathbb{Q}_L}\left[\log \frac{\phi_\epsilon(Y-\sum_{\ell=1}^L \sqrt{nP_\ell}X_\ell)\prod_{\ell=1}^L \left(\frac{1}{2}\exp\{aY\sqrt{P_\ell}\}+\frac{1}{2}\exp\{-aY\sqrt{P_\ell}\}\right)}{p_Y(Y)\exp\{aY\sum_{\ell=1}^L X_\ell\sqrt{nP_\ell}\}}\right]. \end{aligned} \tag{C.1}$$

In (C.1) we use the definitions of the joint distributions given in (4.10) and (4.11). We use the following Lemma.

Lemma 40. *For any value $x \in \mathbb{R}$,*

$$\frac{1}{2}e^x + \frac{1}{2}e^{-x} \leq e^{\frac{x^2}{2}}.$$

Proof. Using the MacLaurin expansion $e^x = \sum_{k=0}^\infty \frac{1}{k!}x^k$,

$$\frac{1}{2}e^x + \frac{1}{2}e^{-x} = \frac{1}{2}\sum_{k=0}^\infty \frac{1}{k!}(x^k + (-x)^k) = \sum_{k'=0}^\infty \frac{1}{(2k')!}x^{2k'} \overset{(a)}{\leq} \sum_{k'=0}^\infty \frac{1}{k'!}\left(\frac{x^2}{2}\right)^{k'} = e^{\frac{x^2}{2}}.$$

Step (a) follows from the fact that $(2k)! \geq 2^k k!$. □

Recall (C.1).

$$D(\mathbb{Q}_L||\mathbb{Q}_L^a) = \mathbb{E}_{\mathbb{Q}_L}\left[\log \frac{\phi_\epsilon(Y - \sum_{\ell=1}^L \sqrt{nP_\ell}X_\ell)\prod_{\ell=1}^L \left(\frac{1}{2}\exp\{aY\sqrt{P_\ell}\} + \frac{1}{2}\exp\{-aY\sqrt{P_\ell}\}\right)}{p_Y(Y)\exp\{aY\sum_{\ell=1}^L X_\ell\sqrt{nP_\ell}\}}\right] \tag{C.2}$$

$$\leq \mathbb{E}_{\mathbb{Q}_L}\left[\log \frac{\phi_\epsilon(Y - \sum_{\ell=1}^L \sqrt{nP_\ell}X_\ell)\exp\{\frac{(aY)^2P}{2}\}}{p_Y(Y)\exp\{aY\sum_{\ell=1}^L X_\ell\sqrt{nP_\ell}\}}\right] \tag{C.3}$$

$$= -\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\mathbb{E}_{\mathbb{Q}_L}\left[\left(Y - \sum_{\ell=1}^L \sqrt{nP_\ell}X_\ell\right)^2\right] + \frac{a^2P}{2}\mathbb{E}_{\mathbb{Q}_L}[Y^2]$$

$$- a\mathbb{E}_{\mathbb{Q}_L}\left[Y\sum_{\ell=1}^L X_\ell\sqrt{nP_\ell}\right] - \mathbb{E}_{\mathbb{Q}_L}\left[\log p_Y(Y)\right]. \tag{C.4}$$

First (C.2) follows from (C.1) and (C.3) uses Lemma 40. Finally, we know that $-\mathbb{E}[\log p_Y(Y)]$ is the entropy of Y which is upper bounded by the entropy of a normal random variable with the same variance (see, for example Thomas and Cover [36]). This means that

$$-\mathbb{E}_{\mathbb{Q}_L}[\log p_Y(Y)] \leq \frac{1}{2}\log(2\pi(\sigma^2 + P)) + \frac{1}{2}. \tag{C.5}$$

Applying (C.5) to upper bound (C.4) and taking the expectation of the remaining terms gives the desired upper bound:

$$D(\mathbb{Q}_L||\mathbb{Q}_L^a) \leq -\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\mathbb{E}_{\mathbb{Q}_L}\left[\left(Y - \sum_{\ell=1}^L \sqrt{nP_\ell}X_\ell\right)^2\right] + \frac{a^2P}{2}\mathbb{E}_{\mathbb{Q}_L}[Y^2]$$

$$- a\mathbb{E}_{\mathbb{Q}_L}\left[Y\sum_{\ell=1}^L X_\ell\sqrt{nP_\ell}\right] - \mathbb{E}_{\mathbb{Q}_L}\left[\log p_Y(Y)\right]$$

$$\leq -\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2} + \frac{a^2(\sigma^2 + P)P}{2} - aP + \frac{1}{2}\log(2\pi(\sigma^2 + P)) + \frac{1}{2}. \tag{C.6}$$

In (C.6) we use the following

$$\mathbb{E}_{\mathbb{Q}_L}\left[\left(Y-\sum_{\ell=1}^{L}\sqrt{nP_\ell}X_\ell\right)^2\right]=\mathbb{E}_{\mathbb{Q}_L}\left[\epsilon^2\right]=\sigma^2,$$

$$\mathbb{E}_{\mathbb{Q}_L}[Y^2]=\sigma^2+P,$$

$$\mathbb{E}_{\mathbb{Q}_L}\left[Y\sum_{\ell=1}^{L}X_\ell\sqrt{nP_\ell}\right]=P.$$

Calculating the minimizing value of a is straightforward.

C.2 Proof of Lemma 8

We want to consider the log of the expectation under the true joint distribution of the ratio of the distributions,

$$\begin{aligned}D_2(\mathbb{Q}_L||\mathbb{Q}_L^a)&=\log\mathbb{E}_{\mathbb{Q}_L}\left[\frac{q_L(X_1,...,X_L,Y)}{q_L^a(X_1,...,X_L,Y)}\right]\\&=\log\mathbb{E}_{\mathbb{Q}_L}\left[\frac{\phi_\epsilon(Y-\sum_{\ell=1}^{L}\sqrt{nP_\ell}X_\ell)\prod_{\ell=1}^{L}\left(\frac{1}{2}\exp\{aY\sqrt{P_\ell}\}+\frac{1}{2}\exp\{-aY\sqrt{P_\ell}\}\right)}{p_Y(Y)\exp\{aY\sum_{\ell=1}^{L}X_\ell\sqrt{nP_\ell}\}}\right]\end{aligned}\tag{C.7}$$

$$\leq\log\mathbb{E}_{\mathbb{Q}_L}\left[\frac{\phi_\epsilon(Y-\sum_{\ell=1}^{L}\sqrt{nP_\ell}X_\ell)\exp\{\frac{(aY)^2P}{2}\}}{p_Y(Y)\exp\{aY\sum_{\ell=1}^{L}X_\ell\sqrt{nP_\ell}\}}\right].\tag{C.8}$$

In (C.7) we use the definitions of the joint distributions given in (4.10) and (4.11) and upper bound (C.8) follows from Lemma 40. We make use of the following Lemma, which is a generalization of a result given by Brown [37], to upper bound the probability density function of Y.

Lemma 41. *Let Y be defined as $Y=\sum_{\ell=1}^{L}\sqrt{nP_\ell}X_\ell+\epsilon$ where $\epsilon\sim N(0,\sigma^2)$, then for any $\gamma>0$*

$$P_Y(Y)\geq\frac{3}{4}\frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{-(1+\gamma)}{2\sigma^2}Y^2}e^{-(1+\frac{1}{\gamma})2\mathsf{snr}}.$$

Proof. We first supply a quick proof of the inequality $(A+B)^2\leq(1+\gamma)A^2+(1+\frac{1}{\gamma})B^2$, for $\gamma>0$. Notice that $(A+B)^2=A^2+B^2+2AB$, so it suffices to show that $2AB\leq\gamma A^2+\frac{1}{\gamma}B^2$.

We know that this is true since $0 \leq (\sqrt{\gamma}A - \frac{1}{\sqrt{\gamma}B})^2 = \gamma A^2 + \frac{1}{\gamma}B^2 - 2AB$.

In what follows we assign W to be the codeword $\sum_{\ell=1}^{L}\sqrt{nP_\ell}X_\ell$. Note that $\mathbb{E}[W] = 0$ and $\mathbb{E}[W^2] = P$. Then by Chebyshev's Inequality,

$$\mathbf{Pr}\left(|W| \leq 2\sqrt{P}\right) \geq \frac{3}{4}. \tag{C.9}$$

Then,

$$\begin{aligned}
P_Y(Y) = \int_{-\infty}^{\infty} p_W(s)\phi_\epsilon(Y-s)ds &\geq \int_{-2\sqrt{P}}^{2\sqrt{P}} p_W(s)\phi_\epsilon(Y-s)ds, \\
&\overset{(a)}{\geq} \frac{3}{4}\min_{s:|s|\leq 2\sqrt{P}}\phi_\epsilon(Y-s) \\
&\overset{(b)}{=} \frac{3}{4}\phi_\epsilon(|Y|+2\sqrt{P}) \\
&= \frac{3}{4}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{\frac{-1}{2\sigma^2}(|Y|+2\sqrt{P})^2\right\} \\
&\overset{(c)}{\geq} \frac{3}{4}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{\frac{-1}{2\sigma^2}((1+\gamma)Y^2 + (1+\frac{1}{\gamma})4P)\right\}.
\end{aligned}$$

Step (a) comes from (C.9) and step (b) follows since ϕ_ϵ is minimized when its input takes its largest value, which occurs at $|Y| + 2\sqrt{P}$. Finally step (c) follows form the work int he first paragraph of the proof. □

Applying Lemma 41 to (C.8) and assigning $W = \sum_{\ell=1}^{L}\sqrt{nP_\ell}X_\ell$, it follows:

$$\begin{aligned}
D_2(\mathbb{Q}_L||\mathbb{Q}_L^a) &\leq \log \mathbb{E}_{\mathbb{Q}_L}\left[\frac{\phi_\epsilon(Y-W)\exp\{\frac{(aY)^2P}{2}\}}{p_Y(Y)\exp\{aYW\}}\right] \\
&\leq \log\left[\frac{4}{3}\exp\left\{\left(1+\frac{1}{\gamma}\right)2\mathsf{snr})\right\}\right] + \log \mathbb{E}_{\mathbb{Q}_L}\left[\frac{\exp\left\{\frac{-1}{2\sigma^2}(Y-W)^2\right\}\exp\{\frac{(aY)^2P}{2}\}}{\exp\left\{\frac{-1}{2\sigma^2}(1+\gamma)Y^2\right\}\exp\{aYW\}}\right].
\end{aligned} \tag{C.10}$$

Using the fact that the expectation of the true distribution equals the expectation taken

first over $Y|W$ and then over W, the expectation in (C.10) equals

$$\mathbb{E}_{\mathbb{Q}_L}\left[\frac{\exp\left\{\frac{-1}{2\sigma^2}(Y-W)^2\right\}\exp\{\frac{(aY)^2P}{2}\}}{\exp\left\{\frac{-1}{2\sigma^2}(1+\gamma)Y^2\right\}\exp\{aYW\}}\right]$$
$$=\mathbb{E}_W\exp\left\{\frac{-W^2}{2\sigma^2}\right\}\mathbb{E}_{Y|W}\left[\exp\left\{\frac{Y^2}{2}\left(\frac{\gamma}{\sigma^2}+a^2P\right)+YW\left(\frac{1}{\sigma^2}-a\right)\right\}\right]. \tag{C.11}$$

Note that in the above $Y|W\sim\mathcal{N}(W,\sigma^2)$ and so the above is equal to the following.

$$\mathbb{E}_W\exp\left\{\frac{-W^2}{\sigma^2}\right\}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\sigma^2}}\left[\exp\left\{-\frac{1}{2\sigma^2}\left[y^2\left(1-\gamma-a^2\sigma^2P\right)-2yW\left(2-a\sigma^2\right)\right]\right\}\right]$$
$$=\frac{1}{\sqrt{1-\gamma-a^2\sigma^2P}}\mathbb{E}_W\left[\exp\left\{W^2\left[\frac{1}{2\sigma^2}\left(\frac{(2-a\sigma^2)^2}{1-\gamma-a^2\sigma^2P}\right)-\frac{1}{\sigma^2}\right]\right\}\right] \tag{C.12}$$

The above follows only if $1-a^2\sigma^2P>\gamma>0$. Now using (C.10) and (C.12) in upper bound (C.10) we find:

$$D_2(\mathbb{Q}_L||\mathbb{Q}_L^a)\leq\log\left[\frac{4\exp\left\{\left(1+\frac{1}{\gamma}\right)2\mathsf{snr}\right)\right\}}{3\sqrt{1-\gamma-a^2\sigma^2P}}\right]+\log\mathbb{E}_W\left[\exp\left\{W^2\left[\frac{1}{2\sigma^2}\left(\frac{(2-a\sigma^2)^2}{1-\gamma-a^2\sigma^2P}\right)-\frac{1}{\sigma^2}\right]\right\}\right]. \tag{C.13}$$

Now if we let $a=\frac{1}{\sigma^2+P}$ and setting

$$c^*=\frac{1}{2\sigma^2}\left(\frac{(2-\frac{1}{1+snr})^2}{1-\gamma-\frac{snr}{(1+snr)^2}}\right)-\frac{1}{\sigma^2}, \tag{C.14}$$

from (C.13) it is clear that we we wish to upper bound $\mathbb{E}_W e^{c^*W^2}$. The following Lemma, a result from Pollard [38], is used to supply an upper bound for this expectation.

Lemma 42. *For a random variable Z, if $\mathbb{E}[\exp\{\lambda Z\}]\leq\exp\left\{\frac{c^2\lambda^2}{2}\right\}$ for some constant c and for all real λ, then for all $\tilde{c}\geq c$,*

$$\mathbb{E}\left[\exp\left\{\frac{Z^2}{4\tilde{c}^2}\right\}\right]\leq 5. \tag{C.15}$$

Proof.

$$\begin{aligned}\mathbb{E}\left[\exp\left\{\frac{Z^2}{4\tilde{c}^2}\right\}\right]-1=\mathbb{E}\int_0^\infty \mathsf{1}\left\{0\le t\le\frac{Z^2}{4\tilde{c}^2}\right\}e^*\,dt &\overset{(a)}{\le}\int_0^\infty\mathbb{E}\left[\exp\left\{\frac{|Z|\sqrt{t}}{\tilde{c}}-t\right\}\right]dt\\ &\le\int_0^\infty e^{-t}\mathbb{E}\left[e^{\frac{Z\sqrt{t}}{\tilde{c}}}+e^{\frac{-Z\sqrt{t}}{\tilde{c}}}\right]dt,\\ &\overset{(b)}{\le}2\int_0^\infty e^{\frac{-t}{2}}dt,\\ &=4.\end{aligned}$$

Step (a) follows by Markov's inequality and step (b) from the fact that $\mathbb{E}e^{\lambda Z}\le e^{\frac{c^2\lambda^2}{2}}$ for all λ. □

Recall $W=\sum_{\ell=1}^L\sqrt{nP_\ell}X_\ell$. By Lemma 40 it follows for any real λ,

$$\begin{aligned}\mathbb{E}[\exp\{\lambda W\}]&=\prod_{\ell=1}^L\mathbb{E}_{X_\ell}[\exp\{\lambda\sqrt{nP_\ell}X_\ell\}] &\text{(C.16)}\\ &=\prod_{\ell=1}^L\left(\frac{1}{2}\exp\{\lambda\sqrt{P_\ell}\}+\frac{1}{2}\exp\{-\lambda\sqrt{P_\ell}\}\right) &\text{(C.17)}\\ &\le\exp\left\{\frac{\lambda^2P}{2}\right\}. &\text{(C.18)}\end{aligned}$$

Therefore by Lemma 42 we find $\mathbb{E}e^{\frac{W^2}{4P}}\le 5$, and whenever $c^*\le\frac{1}{4P}$ the expectation in expression (C.13) is upper bounded by 5 giving

$$D_2(\mathbb{Q}_L||\mathbb{Q}_L^a)\le\log\left[\frac{20\exp\left\{\left(1+\frac{1}{\gamma}\right)2\mathsf{snr})\right\}}{3\sqrt{1-\gamma-\frac{\mathsf{snr}}{(1+\mathsf{snr})^2}}}\right].$$

We will show that for any $\mathsf{snr}<.58$, there exists a range of γ values in the interval $0<\gamma<1-\frac{\mathsf{snr}}{(1+\mathsf{snr})^2}$, which will make $c^*<\frac{1}{4P}$. To see this, recall the definition of c^* from (C.14) and notice that $c^*<\frac{1}{4P}$ whenever

$$c^*P=\frac{\mathsf{snr}}{2}\left(\frac{(2-\frac{1}{1+\mathsf{snr}})^2}{1-\gamma-\frac{\mathsf{snr}}{(1+\mathsf{snr})^2}}\right)-\mathsf{snr}=\frac{\mathsf{snr}}{2}\left(\frac{(1+2\mathsf{snr})^2}{(1-\gamma)(1+2\mathsf{snr}+\mathsf{snr}^2)-\mathsf{snr}}\right)-\mathsf{snr}<\frac{1}{4}. \tag{C.19}$$

For $\gamma=0$, the left-hand side of (C.19) equals $\frac{1}{4}$ when $\mathsf{snr}\approx.58$. Therefore, for snr values

strictly less than .58, there exists a range of γ values close to 0 making the inequality hold.

C.3 Proof of Lemma 9

Let $\delta = \alpha - 1$ so we would like to bound the following.

$$\begin{aligned} D_\alpha(\mathbb{Q}_L||\mathbb{Q}_L^a) &= \frac{1}{\delta}\log \mathbb{E}_{\mathbb{Q}_L}\left[\left(\frac{q_L(X_1,...,X_L,Y)}{q_L^a(X_1,...,X_L,Y)}\right)^\delta\right] \\ &= \frac{1}{\delta}\log \mathbb{E}_{\mathbb{Q}_L}\left[\left(\frac{q(X_1)...q(X_L)\phi_\epsilon(Y-\sum_{l=1}^L\sqrt{P_l}X_l)}{q^a_{X_1|Y}(X_1)...q^a_{X_L|Y}(X_L)p_Y(Y)}\right)^\delta\right] \qquad \text{(C.20)} \\ &\leq \frac{1}{\delta}\log \mathbb{E}_{\mathbb{Q}_L}\left[\left(\frac{\phi_\epsilon(Y-\sum_{\ell=1}^L\sqrt{nP_\ell}X_\ell)\exp\{\frac{(aY)^2P}{2}\}}{p_Y(Y)\exp\{aY\sum_{\ell=1}^L X_\ell\sqrt{nP_\ell}\}}\right)^\delta\right]. \qquad \text{(C.21)} \end{aligned}$$

In (C.20) we use the definitions of the joint distributions given in (4.10) and (4.11) and upper bound (C.21) follows from Lemma 40. Now using Lemma 41 and assigning $W = \sum_{\ell=1}^L X_\ell\sqrt{nP_\ell}$, we can upper bound (C.21) as follows.

$$D_\alpha(\mathbb{Q}_L||\mathbb{Q}_L^a) \leq \log\left[\frac{4}{3}\right] + 2\mathsf{snr}\left(1+\frac{1}{\gamma}\right) + \frac{1}{\delta}\log \mathbb{E}_{\mathbb{Q}_L}\left[\frac{\exp\{-\frac{\delta}{2\sigma^2}(Y-W)^2\}\exp\{\frac{(aY)^2\delta P}{2}\}}{\exp\{\frac{-\delta(1+\gamma)}{2\sigma^2}Y^2\}\exp\{a\delta YW\}}\right]. \qquad \text{(C.22)}$$

Using the fact that the expectation of the true distribution equals the expectation taken first over $Y|W$ and then over W, as in the proof of Theorem 8 in Section C.2, expression (C.22) can be simplified to

$$D_\alpha(\mathbb{Q}_L||\mathbb{Q}_L^a) \leq \log\left[\frac{4}{3}\right] + 2\mathsf{snr}\left(1+\frac{1}{\gamma}\right) + \frac{1}{\delta}\log\frac{\mathbb{E}_W\left[\exp\{c_\delta^* W^2\}\right]}{\sqrt{\delta(1-\gamma-a^2\sigma^2P)}}, \qquad \text{(C.23)}$$

where

$$c_\delta^* = \frac{\delta}{2\sigma^2}\left(\frac{(2-a\sigma^2)^2}{1-\gamma-a^2\sigma^2P}\right) - \frac{\delta}{\sigma^2}.$$

We again must restrict $0 < \gamma < 1 - a^2\sigma^2P$. Another appeal to Lemma 42 is made in order to obtain an upper bound for $\mathbb{E}_W\left[\exp\{c_\delta^* W^2\}\right]$. As before, whenever $c_\delta^* \leq \frac{1}{4P}$ the expectation

in (C.23) is bounded by 5 and so we find the bound

$$D_\alpha(\mathbb{Q}_L||\mathbb{Q}_L^a) \leq \frac{1}{\delta}\log\left[5\left(\frac{4}{3}\right)^\delta\right] + 2\mathsf{snr}\left(1+\frac{1}{\gamma}\right) - \frac{1}{2\delta}\log[\delta(1-\gamma-a^2\sigma^2P)], \qquad \text{(C.24)}$$

The bound $c_\delta^* \leq \frac{1}{4P}$ occurs whenever

$$c_\delta^* P = \delta\left(\frac{1}{2}\mathsf{snr}\left(\frac{(2-a\sigma^2)^2}{1-\gamma-a^2\sigma^2P}\right) - \mathsf{snr}\right) \leq \frac{1}{4}.$$

Since we can take δ arbitrarily close to 0, it is obvious that there is a small enough δ for this inequality to hold for any γ and snr pair.

C.4 Proof of Lemma 10

We prove the result

$$\mathbf{Pr}\left(S \geq \tau\right) \leq \kappa\left(1 - \Phi_{0,\sigma_S^2}(\tau)\right)$$

for $\tau \geq 0$. The symmetric result can be proved similarly. Writing the probability as an iterated expectation it follows that

$$\mathbf{Pr}\left(S \geq \tau\right) = \mathbb{E}_Z\left[\mathbf{Pr}\left(S \geq \tau | Z = z\right)\right]. \qquad \text{(C.25)}$$

Here, the outer expectation integrates over $Z \sim \mathcal{N}(0, \sigma_Z^2)$. Remembering that $S = \sum_{i=1}^n a_i X_i + Z$, the right-hand side of (C.25) equals

$$\mathbb{E}_Z\left[\mathbf{Pr}\left(\sum_{i=1}^n a_i X_i \geq \tau - z\right)\right],$$

where $(X_1, \ldots, X_n) \in \{-1, +1\}^n$ equiprobable. Let $a = (a_1, \ldots, a_n)$. By the tail bound (4.18), this is less than

$$2\kappa\mathbb{E}_Z\left[\mathbf{Pr}\left(Z' + z > \tau\right)\right], \qquad \text{(C.26)}$$

where $Z' \sim \mathcal{N}(0, \|a\|^2)$ where $\|a\|^2 = \sum_{i=1}^n a_i^2$. Since the convolution of two normals is again normal, it follows that (C.26) equals

$$2\kappa \left(1 - \Phi_{0,\sigma_S^2}(\tau)\right),$$

which is what we wanted to show with $\sigma_S^2 = \|a\|^2 + \sigma_Z^2$.

C.5 Proof of Lemma 11

Recall that for each $j \in J$, the X_j are independent random vectors of Bernoulli $\pm\frac{1}{\sqrt{n}}$ random variables. In what follows we write P_j to be the power allocation P_ℓ for $j \in sec(\ell)$. Then

$$\beta_j = \begin{cases} \sqrt{nP_j} & \text{if } j \in sent, \\ 0 & \text{if } j \in other. \end{cases}$$

It follows that

$$Y = \sum_{j=1}^N \beta_j X_j + \epsilon = \sum_{j \in sent} \sqrt{nP_j} X_j + \epsilon, \tag{C.27}$$

where $\epsilon \sim N_n(0, \sigma^2 I)$ is a random vector. We wish to explore the marginal distributions of each test statistic $\mathcal{Z}_{1,j}$. Using representation (C.27) and the fact that $X_j^* X_j = \|X_j\|^2 = 1$ for all j, the inner product of $X_j^* Y$ is expanded as

$$X_j^* Y = \beta_j + \sum_{\substack{j' \in sent \\ j' \neq j}} \beta_{j'} X_j^* X_{j'} + X_j^* \epsilon. \tag{C.28}$$

Notice that the terms in the sum of column inner products are independent. To see this, consider two such terms, $X_j^* X_l$ and $X_j^* X_k$ where $k \neq l$ making X_k is independent of X_l. We show the independence of the two terms $X_j^* X_l$ and $X_j^* X_k$ by conditioning on the random vector X_j. Because of the independence of random vectors X_k and X_l, the conditional distributions $X_j^* X_k | X_j$ and $X_j^* X_l | X_j$ are also independent. Moreover, no matter what values the random Bernoulli vector X_j takes, $X_j^* X_k | X_j$ is always equal in distribution to the sum of n independent Bernoulli $\pm\frac{1}{n}$ random variables. This is because the Bernoulli distribu-

tion is symmetric, so for any element $X_{i,k}$ of the random vector X_k, the random variables $\left(-\frac{1}{\sqrt{n}}\right) X_{i,k}$ and $\left(\frac{1}{\sqrt{n}}\right) X_{i,k}$ are equal in distribution. Therefore, since the conditional joint distribution is the same for all values of X_j, specifically $X_j^* X_l | X_j$ and $X_j^* X_k | X_j$ are independent of each other and each is equal in distribution to the sum of n independent Bernoulli $\pm\frac{1}{n}$ random variables, this is also the unconditional distribution.

By the same reasoning, the inner product $X_j^* X_k$ is independent of $X_j^* \epsilon$. Again, we have conditional independence since X_k is independent of ϵ. Notice also, that conditioned on X_j, by the symmetry of ϵ, the conditional random variable $X_j^* \epsilon | X_j$ has the same distribution as the sum of n independent $N(0, \sigma^2/n)$ random variables, irrespective of the value that X_j takes. Then conditionally on X_j, the inner products $X_j^* X_i$ and $X_j^* \epsilon$ are independent, with $X_j^* X_i | X_j$ having the same distribution as the sum of n independent Bernoulli $\pm\frac{1}{n}$ random variables and $X_j^* \epsilon | X_j$ having the same distribution as the sum of n independent $N(0, \sigma^2/n)$ random variables, both irrespective of the values taken by the random vector X_j. This is then the unconditional joint distribution as well.

For each $j \in J$, it follows that $X_j^* Y$ can be marginally represented as

$$X_j^* Y = \beta_j + \sum_{\substack{j' \in sent \\ j' \neq j}} \beta_{j'} \sum_{i=1}^{n} X_{i,j} X_{i,j'} + \sum_{i=1}^{n} X_{i,j} \epsilon_{i,j} = \beta_j + \sum_{\substack{j' \in sent \\ j' \neq j}} \frac{\beta_{j'}}{n} \sum_{i=1}^{n} B_{i,j'} + \sigma Z. \quad \text{(C.29)}$$

where $B_{i,j'}$ are i.i.d. equiprobable $\{+1, -1\}$ for $i \in [n]$ and $j' \in sent$ with $j' \neq j$ and Z is independent standard normal. Normalizing by σ_Y we get result (4.20). To analyze the tail bounds we proceed with the cases of $j \in other$ and $j \in sent$ separately.

First assume $j \in other$, then (C.29) is normalized to give the marginal representation of $\mathcal{Z}_{1,j}$ as

$$\mathcal{Z}_{1,j} = \sum_{j' \in sent} \frac{\beta_{j'}}{n \sigma_Y} \sum_{i=1}^{n} B_{i,j'} + \frac{\sigma}{\sigma_Y} Z. \quad \text{(C.30)}$$

Notice that (C.30) has unit variance and is a weighted sum of $L \times n$ i.i.d. Bernoulli ± 1 random variables and an independent mean-zero normal. By Lemma 1, it follows that $\mathcal{Z}_{1,j}$ has distribution induced by the convolution measure which has the following property

$$\mathbf{Pr}(\mathcal{Z}_{1,j} \geq \tau) \leq \kappa(1 - \Phi(\tau)). \quad \text{(C.31)}$$

Now consider $j \in sent$, so then (C.29) is normalized to give the marginal representation of $\mathcal{Z}_{1,j}$ as

$$\mathcal{Z}_{1,j} = \frac{\beta_j}{\sigma_Y} + \sum_{\substack{j' \in sent \\ j' \neq j}} \frac{\beta_{j'}}{n\sigma_Y} \sum_{i=1}^{n} B_{i,j'} + \frac{\sigma}{\sigma_Y} Z, \tag{C.32}$$

The variance of (C.32) equals $\sigma^2_{sent} = 1 - \frac{P_j}{\sigma^2_Y} = 1 - \frac{\beta_j^2}{nP}\left(\frac{1}{1+\mathsf{snr}^{-1}}\right)$ and the test statistic is equal in distribution to the shifted sum of $(L-1) \times n$ i.i.d. Bernoulli ± 1 random variables and an independent mean-zero normal. The shift equals,

$$shift_1 = \frac{\beta_j}{\sqrt{\sigma^2 + P}} = \sqrt{\frac{n(P_j/P)}{1+\mathsf{snr}^{-1}}}.$$

Therefore we can choose a $\tau > 0$ such that $\mathbb{E}[\mathcal{Z}_{1,j}] \geq \tau$. Again by Lemma 1, it follows that $\mathcal{Z}_{1,j}$ has distribution induced by the convolution measure which has the following property:

$$\mathbf{Pr}(\mathcal{Z}_{1,j} \leq \tau) \leq \kappa \Phi_{0,\sigma^2_{sent}}(\tau - shift_1). \tag{C.33}$$

C.6 Proof of Lemma 12

Let p_{S^n} be the true density function of the sum meaning p_{S^n} takes as input vectors in $\mathbb{R}^L$. Let $p_{S^n_\ell}$ be the density function of the sum $S^n_\ell = \sum_{i=1}^n U_{i,\ell} a_i$ meaning $p_{S^n_\ell}$ takes as input values in $\mathbb{R}$. Then $\mathbb{P}_{S^n}$ is the measure associated with the joint mass function $p_{S^n}(\bar{s}^n)$ for $\bar{s}^n \in \mathbb{R}^n$ and $\mathbb{Q}_{S^n}$ is the measure associated with the joint mass function $\prod_{\ell=1}^L p_{S^n_\ell}(s^n_\ell)$ where $\bar{s}^n = (s_1, \ldots, s_L)$.

First note,

$$p_{S^n}(\bar{s}^n) = \mathbb{E}_{S^{n-1}}\left[p_{S^1}(\bar{s}^n - S^{n-1})\right],$$

and for each $\ell \in L$,

$$p_{S^n_\ell}(s^n_\ell) = \mathbb{E}_{S^{n-1}_\ell}\left[p_{S^1_\ell}(s^n_\ell - S^{n-1}_\ell)\right].$$

Then we can represent the Rényi divergence as

$$D_\alpha\left(\mathbb{P}_{S^n} || \mathbb{Q}_{S^n}\right) = D_\alpha\left(\mathbb{E}_{S^{n-1}}\left[p_{S^1}(\cdot - S^{n-1})\right] || \prod_{\ell=1}^{L} \mathbb{E}_{S^{n-1}_\ell}\left[p_{S^1_\ell}(\cdot - S^{n-1}_\ell)\right]\right). \tag{C.34}$$

Recall from (4.7),

$$D_\alpha(\mathbb{P}||\mathbb{Q}) = \frac{1}{\alpha-1}\log \mathbb{E}_{\mathbb{P}}\left[\left(\frac{p(X)}{q(X)}\right)^{\alpha-1}\right] = \frac{1}{\alpha-1}\log \mathbb{E}_{\mathbb{Q}}\left[\left(\frac{p(X)}{q(X)}\right)^{\alpha}\right], \tag{C.35}$$

and notice that the term inside the log, $\mathbb{E}_{\mathbb{Q}}\left[\left(\frac{p(X)}{q(X)}\right)^{\alpha}\right]$, is a Csiszar f-divergence when $\alpha > 1$ since $f(r) = r^\alpha$ is convex. Therefore,

$$\int \lambda q_1(x)+(1-\lambda)q_2(x)\left(\frac{\lambda p_1(x)+(1-\lambda)p_2(x)}{\lambda q_1(x)+(1-\lambda)q_2(x)}\right)^{\alpha} < \lambda\int q_1(x)\left(\frac{p_1(x)}{q_1(x)}\right)^{\alpha}+(1-\lambda)\int q_2(x)\left(\frac{p_2(x)}{q_2(x)}\right)^{\alpha}. \tag{C.36}$$

By (C.34) and (C.35), we find

$$D_\alpha\left(\mathbb{P}_{S^n}||\mathbb{Q}_{S^n}\right) = \frac{1}{\alpha-1}\log\left(\sum_{\bar{s}^n}\prod_{\ell=1}^{L}\mathbb{E}_{S_\ell^{n-1}}\left[p_{S_\ell^1}(s_\ell^n - S_\ell^{n-1})\right]\left(\frac{\mathbb{E}_{S^{n-1}}\left[p_{S^1}(\bar{s}^n - S^{n-1})\right]}{\prod_{\ell=1}^{L}\mathbb{E}_{S_\ell^{n-1}}\left[p_{S_\ell^1}(s_\ell^n - S_\ell^{n-1})\right]}\right)^{\alpha}\right), \tag{C.37}$$

which by (C.36) is upper bounded by

$$\frac{1}{\alpha-1}\log\sum_{\bar{s}^{n-1}} p_{S^{n-1}}(\bar{s}^{n-1})\sum_{\bar{s}^n}\frac{\prod_{\ell=1}^{L} p_{S_\ell^{n-1}}(s_\ell^{n-1})p_{S_\ell^1}(s_\ell^n - s_\ell^{n-1})}{p_{S^{n-1}}(\bar{s}^{n-1})}\left(\frac{p_{S^1}(\bar{s}^n - \bar{s}^{n-1})}{\frac{\prod_{\ell=1}^{L} p_{S_\ell^{n-1}}(s_\ell^{n-1})p_{S_\ell^1}(s_\ell^n - s_\ell^{n-1})}{p_{S^{n-1}}(\bar{s}^{n-1})}}\right)^{\alpha}. \tag{C.38}$$

In the above we used the fact that the expectation taken over S^{n-1} is just a sum over terms which we treat as our weights λ that sum to one. The above simplifies to

$$\begin{aligned}&\frac{1}{\alpha-1}\log\sum_{\bar{s}^{n-1}}\prod_{\ell=1}^{L} p_{S_\ell^{n-1}}(s_\ell^{n-1})\left(\frac{p_{S^{n-1}}(\bar{s}^{n-1})}{\prod_{\ell=1}^{L} p_{S_\ell^{n-1}}(s_\ell^{n-1})}\right)^{\alpha}\sum_{\bar{s}^n}\prod_{\ell=1}^{L} p_{S_\ell^1}(s_\ell^n - s_\ell^{n-1})\left(\frac{p_{S^1}(\bar{s}^n - \bar{s}^{n-1})}{\prod_{\ell=1}^{L} p_{S_\ell^1}(s_\ell^n - s_\ell^{n-1})}\right)^{\alpha}\\ &= \frac{1}{\alpha-1}\log\sum_{\bar{s}^{n-1}} p_{S^{n-1}}(\bar{s}^{n-1})\left(\frac{p_{S^{n-1}}(\bar{s}^{n-1})}{\prod_{\ell=1}^{L} p_{S_\ell^{n-1}}(s_\ell^{n-1})}\right)^{\alpha-1}\sum_{\bar{s}^n} p_{S^1}(\bar{s}^n - \bar{s}^{n-1})\left(\frac{p_{S^1}(\bar{s}^n - \bar{s}^{n-1})}{\prod_{\ell=1}^{L} p_{S_\ell^1}(s_\ell^n - s_\ell^{n-1})}\right)^{\alpha-1}.\end{aligned} \tag{C.39}$$

We will argue for each value of $\bar{s}^{n-1}$,

$$\sum_{\bar{s}^n} p_{S^1}(\bar{s}^n - \bar{s}^{n-1}) \left(\frac{p_{S^1}(\bar{s}^n - \bar{s}^{n-1})}{\prod_{\ell=1}^{L} p_{S_\ell^1}(s_\ell^n - s_\ell^{n-1})} \right)^{\alpha-1} = \sum_{\bar{s}^1} p_{S^1}(\bar{s}^1) \left(\frac{p_{S^1}(\bar{s}^1)}{\prod_{\ell=1}^{L} p_{S_\ell^1}(s_\ell^1)} \right)^{\alpha-1}. \quad \text{(C.40)}$$

The reasoning is the following: the sum on the left side is taken over $\bar{s}^n = \bar{s}^1 + \bar{s}^{n-1}$ and for each fixed $\bar{s}^{n-1}$ the probability $p_{S^1}(\bar{s}^n - \bar{s}^{n-1})$ is positive only for the 2^L terms corresponding to $\bar{s}^n = \bar{s}^1 + \bar{s}^{n-1}$ for some $\bar{s}^1$ value (there are 2^L possible values). This follows from the independence of the vectors $U^1, \ldots, U^n$ From (C.40) it follows that (C.39) equals

$$\frac{1}{\alpha - 1} \log \sum_{\bar{s}^{n-1}} p_{S^{n-1}}(\bar{s}^{n-1}) \left(\frac{p_{S^{n-1}}(\bar{s}^{n-1})}{\prod_{\ell=1}^{L} p_{S_\ell^{n-1}}(s_\ell^{n-1})} \right)^{\alpha-1} \sum_{\bar{s}^1} p_{S^1}(\bar{s}^1) \left(\frac{p_{S^1}(\bar{s}^1)}{\prod_{\ell=1}^{L} p_{S_\ell^1}(s_\ell^1)} \right)^{\alpha-1}$$

$$= \frac{1}{\alpha - 1} \log \sum_{\bar{s}^{n-1}} p_{S^{n-1}}(\bar{s}^{n-1}) \left(\frac{p_{S^{n-1}}(\bar{s}^{n-1})}{\prod_{\ell=1}^{L} p_{S_\ell^{n-1}}(s_\ell^{n-1})} \right)^{\alpha-1} + \frac{1}{\alpha - 1} \log \sum_{\bar{s}^1} p_{S^1}(\bar{s}^1) \left(\frac{p_{S^1}(\bar{s}^1)}{\prod_{\ell=1}^{L} p_{S_\ell^1}(s_\ell^1)} \right)^{\alpha-1}$$

$$= D_\alpha \left(\mathbb{P}_{S^{n-1}} || \mathbb{Q}_{S^{n-1}} \right) + D_\alpha \left(\mathbb{P}_{S^1} || \mathbb{Q}_{S^1} \right)$$

Bibliography

[1] A. Barron and A. Joseph. Least squares superposition codes of moderate dictionary size are reliable at rates up to capacity. *IEEE Trans. on Inf. Theory*, 58(5):2541–2557, Feb 2012.

[2] A. Joseph and A. R. Barron. Fast sparse superposition codes have near exponential error probability for $R < \mathcal{C}$. *IEEE Trans. Information Theory*, 60(2):919–942, Feb. 2014.

[3] A. R. Barron and S. Cho. High-rate sparse superposition codes with iteratively optimal estimates. In *Proc. IEEE Int. Symp. Information Theory*, 2012.

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

[5] A. Montanari. Graphical models concepts in compressed sensing. In Y. C. Eldar and G. Kutyniok, editors, *Compressed Sensing*, pages 394–438. Cambridge University Press, 2012.

[6] M. Bayati and A. Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Trans. Information Theory*, pages 764–785, 2011.

[7] F. Krzakala, M. Mézard, F. Sausset, Y. Sun, and L. Zdeborová. Probabilistic reconstruction in compressed sensing: algorithms, phase diagrams, and threshold achieving matrices. *Journal of Statistical Mechanics: Theory and Experiment*, (8), 2012.

[8] S. Rangan. Generalized approximate message passing for estimation with random linear mixing. In *Proc. IEEE Int. Symp. Information Theory*, pages 2168–2172, 2011.

[9] C. Rush, A. Greig, and R. Venkataramanan. Capacity-achieving sparse superposition codes via approximate message passing decoding. *arXiv:1501.05892*, 2015.

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

[11] T. Cover and J. Thomas. *Elements of information theory*. Wiley-interscience, 2006.

[12] C. E. Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

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

[14] Y. Takeishi, M. Kawakita, and J. Takeuchi. Least squares superposition codes with bernoulli dictionary are still reliable at rates up to capacity. *IEEE Trans. Information Theory*, pages 2737–2750, 2014.

[15] S. Cho. *High-dimensional regression with random design, including sparse superposition codes*. PhD thesis, Yale University, 2014.

[16] J. Barbier and F. Krzakala. Replica analysis and approximate message passing decoder for sparse superposition codes. In *ISIT*, 2014.

[17] J. Barbier and F. Krzakala. Approximate message-passing decoder and capacity-achieving sparse superposition codes. *arXiv:1503.08040*, 2015.

[18] C. Condo and W. J. Gross. Sparse superposition codes: A practical approach. In *Signal Processing Systems (SiPS), 2015 IEEE Workshop on*, pages 1–6, 2015.

[19] D. L. Donoho, A. Javanmard, and A. Montanari. Information-theoretically optimal compressed sensing via spatial coupling and approximate message passing. *IEEE Trans. Information Theory*, (11):7434–7464, Nov. 2013.

[20] M. Bayati and A. Montanari. The LASSO Risk for Gaussian Matrices. *IEEE Trans. Inf. Theory*, 58(4):1997–2017, April 2012.

[21] E. Bolthausen. An iterative construction of solutions of the TAP equations for the Sherrington-Kirkpatrick model. *arXiv:1201.2891*, 2012.

[22] A. Javanmard and A. Montanari. State evolution for general approximate message passing algorithms, with applications to spatial coupling. *Information and Inference*, 2013.

[23] J. Shanks. Computation of the Fast Walsh-Fourier transform. *IEEE Trans. Computers*, 18:457–459, 1969.

[24] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. OUP Oxford, 2013.

[25] C. Rush and R. Venkataramanan. Main technical lemma in the finite sample analysis of AMP. Online: `http://sigproc.eng.cam.ac.uk/foswiki/pub/Main/RV285/main_lemma_finite_sample_analysis.pdf`.

[26] D. Donoho and A. Montanari. High dimensional robust M-estimation: asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, pages 1–35, 2015.

[27] U. Kamilov, S. Rangan, A. K. Fletcher, and M. Unser. Approximate message passing with consistent parameter estimation and applications to sparse learning. *IEEE Trans. on Inf. Theory*, 60:2969–2985, May 2014.

[28] C. Rush, A. Greig, and R. Venkataramanan. Capacity-achieving sparse regression codes via approximate message passing decoding. In *ISIT*, 2015.

[29] J. Van Campenhout and T. Cover. Maximum entropy and conditional probability. *Information Theory, IEEE Transactions on*, 27(4):483–489, 1981.

[30] I. Csiszár. Sanov property, generalized I-projection and a conditional limit theorem. *The Annals of Probability*, pages 768–793, 1984.

[31] O. Lanford. Entropy and equilibrium states in classical statistical mechanics. In *Statistical mechanics and mathematical problems*, pages 1–113. Springer, 1973.

[32] I. Pinelis. Extremal probabilistic problems and hotelling's t2 test under a symmetry condition. *The Annals of Statistics*, 22(1):357–368, 1994.

[33] S. Bobkov, F. Götze, and C. Houdré. On gaussian and bernoulli covariance representations. *Bernoulli*, 7(3):439–451, 2001.

[34] I. Pinelis. Toward the best constant factor for the rademacher-gaussian tail comparison. *arXiv preprint math/0605340*, 2006.

[35] Extended proof of steps 2(b) and 4(b). *Online*. `http://sigproc.eng.cam.ac.uk/foswiki/pub/Main/RV285/Steps_2b4b.pdf`.

[36] T. Cover, J. Thomas, and J. Kieffer. Elements of information theory. *SIAM Review*, 36(3):509–510, 1994.

[37] L. Brown. A proof of the central limit theorem motivated by the Cramér-Rao inequality. *Statistics and probability: essays in honor of CR Rao*, pages 141–148, 1982.

[38] D. Pollard. Asymptopia, book in progress, 2000.