View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Apollo

Finite Sample Analysis of Approximate Message Passing Algorithms

Cynthia Rush* Ramji Venkataramanan†

March 9, 2018

Abstract

Approximate message passing (AMP) refers to a class of efficient algorithms for statistical estimation in high-dimensional problems such as compressed sensing and low-rank matrix estimation. This paper analyzes the performance of AMP in the regime where the problem dimension is large but finite. For concreteness, we consider the setting of high-dimensional regression, where the goal is to estimate a high-dimensional vector β_0 from a noisy measurement $y = A\beta_0 + w$. AMP is a low-complexity, scalable algorithm for this problem. Under suitable assumptions on the measurement matrix A, AMP has the attractive feature that its performance can be accurately characterized in the large system limit by a simple scalar iteration called state evolution. Previous proofs of the validity of state evolution have all been asymptotic convergence results. In this paper, we derive a concentration inequality for AMP with i.i.d. Gaussian measurement matrices with finite size $n \times N$. The result shows that the probability of deviation from the state evolution prediction falls exponentially in n. This provides theoretical support for empirical findings that have demonstrated excellent agreement of AMP performance with state evolution predictions for moderately large dimensions. The concentration inequality also indicates that the number of AMP iterations t can grow no faster than order $\frac{\log n}{\log \log n}$ for the performance to be close to the state evolution predictions with high probability. The analysis can be extended to obtain similar non-asymptotic results for AMP in other settings such as low-rank matrix estimation.

1 Introduction

Consider the 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

$$y = A\beta_0 + w. \tag{1.1}$$

Here A is a known $n \times N$ real-valued measurement matrix, and $w \in \mathbb{R}^n$ is the measurement noise. The sampling ratio $\frac{n}{N} \in (0, \infty)$ is denoted by δ.

Approximate Message Passing (AMP) [1–6] is a class of low-complexity, scalable algorithms to solve the above problem, under suitable assumptions on A and β_0. AMP algorithms are derived as Gaussian or quadratic approximations of loopy belief propagation algorithms (e.g., min-sum, sum-product) on the dense factor graph corresponding to (1.1).

*Department of Statistics, Columbia University. Email: cynthia.rush@columbia.edu.

†Department of Engineering, University of Cambridge. Email: ramji.v@eng.cam.ac.uk. This work was presented in part at the 2016 IEEE International Symposium on Information Theory.

Given the observed vector y, AMP generates successive estimates of the unknown vector, denoted by $\beta^t \in \mathbb{R}^N$ for $t = 1, 2, \ldots$. Set $\beta^0 = 0$, the all-zeros vector. For $t = 0, 1, \ldots$, AMP computes

$$z^t = y - A\beta^t + \frac{z^{t-1}}{n} \sum_{i=1}^{N} \eta'_{t-1}([A^* z^{t-1} + \beta^{t-1}]_i), \tag{1.2}$$

$$\beta^{t+1} = \eta_t(A^* z^t + \beta^t), \tag{1.3}$$

for an appropriately-chosen sequence of functions $\{\eta_t\}_{t\geq 0} : \mathbb{R} \to \mathbb{R}$. In (1.2) and (1.3), A^* denotes the transpose of A, 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 throughout the paper. For a demonstration of how the AMP updates (1.2) and (1.3) are derived from a min-sum-like message passing algorithm, we refer the reader to [1].

For a Gaussian measurement matrix A with entries that are i.i.d. $\sim \mathcal{N}(0, 1/n)$, it was rigorously proven [1, 7] that the performance of AMP can be characterized in the large system limit via a simple scalar iteration called *state evolution*. This result was extended to the class of matrices with i.i.d. sub-Gaussian entries in [8]. In particular, these results imply that performance measures such as the L^2-error $\frac{1}{N}\left\|\beta_0 - \beta^t\right\|^2$ and the L^1-error $\frac{1}{N}\left\|\beta_0 - \beta^t\right\|_1$ converge almost surely to constants that can be computed via the distribution of β_0. (The large system limit is defined as $n, N \to \infty$ such that $\frac{n}{N} = \delta$, a constant.)

AMP has also been applied to a variety of other high-dimensional estimation problems. Some examples are low-rank matrix estimation [9–14], decoding of sparse superposition codes [15–17], matrix factorization [18], and estimation in generalized linear and bilinear models [5, 19, 20].

Main Contributions: In this paper, we obtain a non-asymptotic result for the performance of the AMP iteration in (1.2)–(1.3), when the measurement matrix A has i.i.d. Gaussian entries $\sim \mathcal{N}(0, 1/n)$. We derive a concentration inequality (Theorem 3.1) that implies that the probability of ϵ-deviation between various performance measures (such as $\frac{1}{N}\left\|\beta_0 - \beta^t\right\|^2$) and their limiting constant values fall exponentially in n. Our result provides theoretical support for empirical findings that have demonstrated excellent agreement of AMP performance with state evolution predictions for moderately large dimensions, e.g., n of the order of several hundreds [2].

In addition to refining earlier asymptotic results, the concentration inequality in Theorem 3.1 also clarifies the effect of the iteration number t versus the problem dimension n. One implication is that the actual AMP performance is close to the state evolution prediction with high probability as long as t is of order smaller than $\frac{\log n}{\log \log n}$. This is particularly relevant for settings where the number of AMP iterations and the problem dimension are both large, e.g., solving the LASSO via AMP [6].

We prove the concentration result in Theorem 3.1 by analyzing the following general recursion:

$$\begin{aligned} b^t &= A f_t(h^t, \beta_0) - \lambda_t g_{t-1}(b^{t-1}, w), \\ h^{t+1} &= A^* g_t(b^t, w) - \xi_t f_t(h^t, \beta_0). \end{aligned} \tag{1.4}$$

Here, for $t \geq 0$, the vectors $b^t \in \mathbb{R}^n$, $h^{t+1} \in \mathbb{R}^N$ describe the state of the algorithm, $f_t, g_t : \mathbb{R} \to \mathbb{R}$ are Lipschitz functions that are separable (act component-wise when applied to vectors), and λ_t, ξ_t are scalars that can be computed from the state of the algorithm. The algorithm is initialized with $f_0(h^0 = 0, \beta_0)$. Further details on the recursion in (1.4), including how the AMP in (1.2)–(1.3) can be obtained as a special case, are given in Section 4.1.

For ease of exposition, our analysis will focus on the recursion (1.4) and the problem of high-dimensional regression. However, it can be extended to a number of related problems. A symmetric

version of the above recursion yields AMP algorithms for problems such as solving the TAP equations in statistical physics [21] and symmetric low-rank matrix estimation [10, 12]. This recursion is defined in terms of a symmetric matrix $G \in \mathbb{R}^{N \times N}$ with entries $\{G_{ij}\}_{i<j}$ i.i.d. $\sim \mathcal{N}(0, \frac{1}{N})$, and $\{G_{ii}\}$ i.i.d. $\sim \mathcal{N}(0, \frac{2}{N})$ for $i \in [N]$. (In other words, G can be generated as $(A + A^*)/2$, where $A \in \mathbb{R}^{N \times N}$ has i.i.d. $\mathcal{N}(0, \frac{1}{N})$ entries.) Then, for $t \geq 0$, let

$$m^{t+1} = A\, p_t(m^t) - \mathsf{b}_t\, p_{t-1}(m^{t-1}). \tag{1.5}$$

Here, for $t \geq 0$, the state of the algorithm is represented by a single vector $m^t \in \mathbb{R}^N$, the function $p_t : \mathbb{R} \to \mathbb{R}$ is Lipschitz and separable, and b_t is a constant computed from the state of the algorithm (see [1, Sec. IV] for details). The recursion (1.5) is initialized with a deterministic vector $m^1 \in \mathbb{R}^N$.

Our analysis of the recursion (1.4) can be easily extended to obtain an analogous non-asymptotic result for the symmetric recursion in (1.5). Therefore, for problems of estimating either symmetric or rectangular low-rank matrices in Gaussian noise, our analysis can be used to refine existing asymptotic AMP guarantees (such as those in [9–11]), by providing a concentration result similar to that in Theorem 3.1. We also expect that the non-asymptotic analysis can be generalized to the case where the recursion in (1.4) generates matrices rather than vectors, i.e, $b^t \in \mathbb{R}^{n \times q}$ and $h^{t+1} \in \mathbb{R}^{N \times q}$ (where q remains fixed as n, N grow large; see [7] for details). Extending the analysis to this matrix recursion would yield non-asymptotic guarantees for the generalized AMP [5] and AMP for compressed sensing with spatially coupled measurement matrices [22].

Since the publication of the conference version of this paper, the analysis described here has been used in a couple of recent papers: an error exponent for sparse regression codes with AMP decoding was obtained in [23], and a non-asymptotic result for AMP with non-separable denoisers was given in [24].

1.1 Assumptions

Before proceeding, we state the assumptions on the model (1.1) and the functions used to define the AMP. In what follows, $K, \kappa > 0$ are generic positive constants whose values are not exactly specified but do not depend on n. We use the notation $[n]$ to denote the set $\{1, 2, \ldots, n\}$.

- **Measurement Matrix:** The entries of measurement matrix $A \in \mathbb{R}^{n \times N}$ are i.i.d. $\sim \mathcal{N}(0, 1/n)$.
- **Signal:** The entries of the signal $\beta_0 \in \mathbb{R}^N$ are i.i.d. according to a sub-Gaussian distribution p_β. We recall that a zero-mean random variable X is sub-Gaussian if there exist positive constants K, κ such that $P(|X - \mathbb{E}X| > \epsilon) \leq Ke^{-\kappa\epsilon^2}$, $\forall \epsilon > 0$ [25].
- **Measurement Noise:** The entries of the measurement noise vector w are i.i.d. according to some sub-Gaussian distribution p_w with mean 0 and $\mathbb{E}[w_i^2] = \sigma^2 < \infty$ for $i \in [n]$. The sub-Gaussian assumption implies that, for $\epsilon \in (0, 1)$,

  $$P\left(\left|\frac{1}{n}\|w\|^2 - \sigma^2\right| \geq \epsilon\right) \leq Ke^{-\kappa n\epsilon^2}, \tag{1.6}$$

  for some constants $K, \kappa > 0$ [25].
- **The Functions η_t:** The denoising functions, $\eta_t : \mathbb{R} \to \mathbb{R}$, in (1.3) are Lipschitz continuous for each $t \geq 0$, and are therefore weakly differentiable. The weak derivative, denoted by η_t', is assumed to be differentiable, except possibly at a finite number of points, with bounded derivative everywhere it exists. Allowing η_t' to be non-differentiable at a finite number of points covers denoising functions like soft-thresholding which is used in applications such as the LASSO [6].

Functions defined with scalar inputs are assumed to act component-wise when applied to vectors.

The remainder of the paper is organized as follows. In Section 2 we review state evolution, the formalism predicting the performance of AMP, and discuss how knowledge of the signal distribution p_β and the noise distribution p_w can help choose good denoising functions $\{\eta_t\}$. However, we emphasize that our result holds for the AMP with any choice of $\{\eta_t\}$ satisfying the above condition, even those that do not depend on p_β and p_w. In Section 2.1, we introduce a stopping criterion for termination of the AMP. In Section 3, we give our main result (Theorem 3.1) which proves that the performance of AMP can be characterized accurately via state evolution for large but finite sample size n. Section 4 gives the proof of Theorem 3.1. The proof is based on two technical lemmas: Lemmas 4.3 and 4.5. The proof of Lemma 4.5 is long; we therefore give a brief summary of the main ideas in Section 4.6 and then the full proof in Section 5. In the appendices, we list a number of concentration inequalities that are used in the proof of Lemma 4.5. Some of these, such as the concentration inequality for the sum of pseudo-Lipschitz functions of i.i.d. sub-Gaussian random variables (Lemma B.4), may be of independent interest.

2 State Evolution and the Choice of η_t

In this section, we briefly describe state evolution, the formalism that predicts the behavior of AMP in the large system limit. We only review the main points followed by a few examples; a more detailed treatment can be found in [1, 4].

Given p_β, let $\beta \in \mathbb{R} \sim p_\beta$. Let $\sigma_0^2 = \mathbb{E}[\beta^2]/\delta > 0$, where $\delta = n/N$. Iteratively define the quantities $\{\tau_t^2\}_{t\geq 0}$ and $\{\sigma_t^2\}_{t\geq 1}$ as

$$\tau_t^2 = \sigma^2 + \sigma_t^2, \qquad \sigma_t^2 = \frac{1}{\delta}\mathbb{E}\left[(\eta_{t-1}(\beta + \tau_{t-1}Z) - \beta)^2\right], \tag{2.1}$$

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

The AMP update (1.3) is underpinned by the following key property of the vector $A^*z^t + \beta^t$: *for large n, $A^*z^t + \beta^t$ 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.* In light of this property, a natural way to generate β^{t+1} from the "effective observation" $A^*z^t + \beta^t = s$ is via the conditional expectation:

$$\beta^{t+1}(s) = \mathbb{E}[\,\beta \mid \beta + \tau_t Z = s\,], \tag{2.2}$$

i.e., β^{t+1} is the MMSE 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 (2.2).

In the definition of the "modified residual" z^t, the third term on the RHS of (1.2) is crucial to ensure that the effective observation $A^*z^t + \beta^t$ has the above distributional property. For intuition about the role of this 'Onsager term', the reader is referred to [1, Section I-C].

We review two examples to illustrate how full or partial knowledge of p_β can guide the choice of the denoising function η_t. 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 (2.2) with this p_β, we obtain $\eta_t(s) = \tanh(s/\tau_t^2)$ [1]. The constants τ_t^2 are determined iteratively from the state evolution equations (2.1).

As a second example, consider the compressed sensing problem, where $\delta < 1$, and p_β is such that $P(\beta_0 = 0) = 1 - \xi$. The parameter $\xi \in (0,1)$ determines the sparsity of β_0. For this problem, the authors in [2, 4] suggested the choice $\eta_t(s) = \eta(s;\theta_t)$, where the soft-thresholding function η is

defined as

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

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 (2.1), making the threshold value proportional to the standard deviation of the noise in the effective observation. However, computing τ_t using (2.1) requires knowledge of p_β. In the absence of such knowledge, we can estimate τ_t^2 by $\frac{1}{n}\left\|z^t\right\|^2$: our concentration result (Lemma 4.5(e)) shows that this approximation is increasingly accurate as n grows large. To fix α, one could run the AMP with several different values of α, and choose the one that gives the smallest value of $\frac{1}{n}\left\|z^t\right\|^2$ for large t.

We note that in each of the above examples η_t is Lipschitz, and its derivative satisfies the assumption stated in Section 1.1.

2.1 Stopping Criterion

To obtain a concentration result that clearly highlights the dependence on the iteration t and the dimension n, we include a stopping criterion for the AMP algorithm. The intuition is that the AMP algorithm can be terminated once the expected squared error of the estimates (as predicted by state evolution equations in (2.1)) is either very small or stops improving appreciably.

For Bayes-optimal AMP where the denoising function $\eta_t(\cdot)$ is the conditional expectation given in (2.2), the stopping criterion is as follows. Terminate the algorithm at the first iteration $t > 0$ for which either

$$\sigma_t^2 < \varepsilon_0, \quad \text{or} \quad \frac{\sigma_t^2}{\sigma_{t-1}^2} > 1 - \varepsilon_0', \tag{2.3}$$

where $\varepsilon_0 > 0$ and $\varepsilon_0' \in (0,1)$ are pre-specified constants. Recall from (2.1) that σ_t^2 is expected squared error in the estimate. Therefore, for suitably chosen values of $\varepsilon_0, \varepsilon_0'$, the AMP will terminate when the expected squared error is either small enough, or has not significantly decreased from the previous iteration.

For the general case where $\eta_t(\cdot)$ is not the Bayes-optimal choice, the stopping criterion is: terminate the algorithm at the first iteration $t > 0$ for which at least one of the following is true:

$$\sigma_t^2 < \varepsilon_1, \quad \text{or} \quad (\sigma_t^\perp)^2 < \varepsilon_2, \quad \text{or} \quad (\tau_t^\perp)^2 < \varepsilon_3, \tag{2.4}$$

where $\varepsilon_1, \varepsilon_2, \varepsilon_3 > 0$ are pre-specified constants, and $(\sigma_t^\perp)^2, (\tau_t^\perp)^2$ are defined in (4.19). The precise definitions of the scalars $(\sigma_t^\perp)^2, (\tau_t^\perp)^2$ are postponed to Sec. 4.2 as a few other defintiions are needed first. For now, it suffices to note that $(\sigma_t^\perp)^2, (\tau_t^\perp)^2$ are measures of how close σ_t^2 and τ_t^2 are to σ_{t-1}^2 and τ_{t-1}^2, respectively. Indeed, for the Bayes-optimal case, we show in Sec 4.3 that

$$(\sigma_t^\perp)^2 := \sigma_t^2\left(1 - \frac{\sigma_t^2}{\sigma_{t-1}^2}\right), \quad (\tau_t^\perp)^2 := \tau_t^2\left(1 - \frac{\tau_t^2}{\tau_{t-1}^2}\right).$$

Let $T^* > 0$ be the first value of $t > 0$ for which at least one of the conditions is met. Then the algorithm is run only for $0 \le t < T^*$. It follows that for $0 \le t < T^*$,

$$\sigma_t^2 > \varepsilon_1, \quad \tau_t^2 > \sigma^2 + \varepsilon_1, \quad (\sigma_t^\perp)^2 > \varepsilon_2, \quad (\tau_t^\perp)^2 > \varepsilon_3. \tag{2.5}$$

In the rest of the paper, we will use the stopping criterion to implicitly assume that $\sigma_t^2, \tau_t^2, (\sigma_t^\perp)^2, (\tau_t^\perp)^2$ are bounded below by positive constants.

3 Main Result

Our result, Theorem 3.1, is a concentration inequality for *pseudo-Lipschitz* (PL) loss functions. As defined in [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 $x, y \in \mathbb{R}^m$,$|\phi(x) - \phi(y)| \leq L(1 + \|x\| + \|y\|) \|x - y\|$, where $\|\cdot\|$ denotes the Euclidean norm.

Theorem 3.1. *With the assumptions listed in Section 1.1, the following holds for any (order-2) pseudo-Lipschitz function $\phi : \mathbb{R}^2 \to \mathbb{R}$, $\epsilon \in (0, 1)$ and $0 \leq t < T^*$, where T^* is the first iteration for which the stopping criterion in (2.4) is satisfied.*

$$P\left(\left|\frac{1}{N}\sum_{i=1}^{N}\phi(\beta_i^{t+1}, \beta_{0_i}) - \mathbb{E}\left[\phi\left(\eta_t\left(\beta + \tau_t Z\right), \beta\right)\right]\right| \geq \epsilon\right) \leq K_t e^{-\kappa_t n\epsilon^2}. \tag{3.1}$$

In the expectation in (3.1), $\beta \sim p_\beta$ and $Z \sim \mathcal{N}(0, 1)$ are independent, and τ_t is given by (2.1). The constants K_t, κ_t are given by $K_t = C^{2t}(t!)^{10}, \kappa_t = \frac{1}{c^{2t}(t!)^{22}}$, where $C, c > 0$ are universal constants (not depending on t, n, or ϵ) that are not explicitly specified.

The probability in (3.1) is with respect to the product measure on the space of the measurement matrix A, signal β_0, and the noise w.

Remarks:

1. By considering the pseudo-Lipschitz function $\phi(a, b) = (a-b)^2$, Theorem 3.1 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 $t \geq 0$,

$$P\left(\left|\frac{1}{N}\left\|\beta^{t+1} - \beta_0\right\|^2 - \delta\sigma_{t+1}^2\right| \geq \epsilon\right) \leq K_t e^{-\kappa_t n\epsilon^2}. \tag{3.2}$$

Similarly, taking $\phi(a, b) = |a - b|$ the theorem implies that the normalized L_1-error $\frac{1}{N}\left\|\beta^{t+1} - \beta_0\right\|_1$ is concentrated around $\mathbb{E}\left|\eta_t\left(\beta + \tau_t Z\right) - \beta\right|$.

2. Asymptotic convergence results of the kind given in [1,6] are implied by Theorem 3.1. Indeed, from Theorem 3.1 we have for any fixed $t \geq 0$:

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

Therefore the Borel-Cantelli lemma implies that for any fixed $t \geq 0$:

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

3. Theorem 3.1 also refines the asymptotic convergence result by specifying how large t can be (compared to the dimension n) for the state evolution predictions to be meaningful. Indeed, if we require the bound in (3.1) to go to zero with growing n, we need $\kappa_t n\epsilon^2 \to \infty$ as $n \to \infty$. Using the expression for κ_t from the theorem then yields $t = o\left(\frac{\log n}{\log\log n}\right)$.

Thus, when the AMP is run for a growing number of iterations, the state evolution predictions are guaranteed to be valid until iteration t if the problem dimension grows faster than exponentially in t. Though the constants K_t, κ_t in the bound have not been optimized, we believe that the

dependence of these constants on $t!$ is inevitable in any induction-based proof of the result. An open question is whether this relationship between t and n is fundamental, or a different analysis of the AMP can yield constants which allow t to grow faster with n.

4. As mentioned in the introduction, we expect that non-asymptotic results similar to Theorem 3.1 can be obtained for other estimation problems (with Gaussian matrices) for which rigorous asymptotic results have been proven for AMP. Examples of such problems include low-rank matrix estimation [9–11], robust high-dimensional M-estimation [26], AMP with spatially coupled matrices [22], and generalized AMP [7, 27].

As our proof technique depends heavily on A being i.i.d. Gaussian, extending Theorem 3.1 to AMP with sub-Gaussian matrices [8] and to variants of AMP with structured measurement matrices (e.g., [28–30]) is non-trivial, and an interesting direction for future work.

4 Proof of Theorem 3.1

We first lay down the notation that will be used in the proof, then state two technical lemmas (Lemmas 4.3 and 4.5) and use them to prove Theorem 3.1.

4.1 Notation and Definitions

For consistency and ease of comparison, we use notation similar to [1]. To prove the technical lemmas, we use the general recursion in (1.4), which we write in a slightly different form below. Given $w \in \mathbb{R}^n$, $\beta_0 \in \mathbb{R}^N$, define the column vectors $h^{t+1}, q^{t+1} \in \mathbb{R}^N$ and $b^t, m^t \in \mathbb{R}^n$ for $t \geq 0$ recursively as follows, starting with initial condition $q^0 \in \mathbb{R}^N$:

$$\begin{aligned} b^t &:= Aq^t - \lambda_t m^{t-1}, \qquad & m^t &:= g_t(b^t, w), \\ h^{t+1} &:= A^* m^t - \xi_t q^t, \qquad & q^t &:= f_t(h^t, \beta_0). \end{aligned} \tag{4.1}$$

where the scalars ξ_t and λ_t are defined as

$$\xi_t := \frac{1}{n}\sum_{i=1}^{n} g_t'(b_i^t, w_i), \quad \lambda_t := \frac{1}{\delta N}\sum_{i=1}^{N} f_t'(h_i^t, \beta_{0_i}). \tag{4.2}$$

In (4.2), the derivatives of $g_t : \mathbb{R}^2 \to \mathbb{R}$ and $f_t : \mathbb{R}^2 \to \mathbb{R}$ are with respect to the first argument. The functions f_t, g_t are assumed to be Lipschitz continuous for $t \geq 0$, hence the weak derivatives g_t' and f_t' exist. Further, g_t' and f_t' are each assumed to be differentiable, except possibly at a finite number of points, with bounded derivative everywhere it exists.

Let $\sigma_0^2 := \mathbb{E}\left[f_0^2(0, \beta)\right] > 0$ with $\beta \sim p_\beta$. We let $q^0 = f_0(0, \beta_0)$ and assume that there exist constants $K, \kappa > 0$ such that

$$P\left(\left|\frac{1}{n}\left\|q^0\right\|^2 - \sigma_0^2\right| \geq \epsilon\right) \leq Ke^{-\kappa n\epsilon^2}. \tag{4.3}$$

Define the state evolution scalars $\{\tau_t^2\}_{t\geq 0}$ and $\{\sigma_t^2\}_{t\geq 1}$ for the general recursion as follows.

$$\tau_t^2 := \mathbb{E}\left[\left(g_t(\sigma_t Z, W)\right)^2\right], \qquad \sigma_t^2 := \frac{1}{\delta}\mathbb{E}\left[\left(f_t(\tau_{t-1} Z, \beta)\right)^2\right], \tag{4.4}$$

where $\beta \sim p_\beta, W \sim p_w$, and $Z \sim \mathcal{N}(0, 1)$ are independent random variables. We assume that both σ_0^2 and τ_0^2 are strictly positive.

The AMP algorithm is a special case of the general recursion in (4.1) and (4.2). Indeed, the AMP can be recovered by defining the following vectors recursively for $t \geq 0$, starting with $\beta^0 = 0$ and $z^0 = y$.

$$\begin{aligned} h^{t+1} &= \beta_0 - (A^* z^t + \beta^t), \qquad & q^t &= \beta^t - \beta_0, \\ b^t &= w - z^t, & m^t &= -z^t. \end{aligned} \tag{4.5}$$

It can be verified that these vectors satisfy (4.1) and (4.2) with

$$f_t(a, \beta_0) = \eta_{t-1}(\beta_0 - a) - \beta_0, \quad \text{and} \quad g_t(a, w) = a - w. \tag{4.6}$$

Using this choice of f_t, g_t in (4.4) yields the expressions for σ_t^2, τ_t^2 given in (2.1). Using (4.6) in (4.2), we also see that for AMP,

$$\lambda_t = -\frac{1}{\delta N} \sum_{i=1}^{N} \eta'_{t-1}([A^* \beta^{t-1} + z^{t-1}]_i), \qquad \xi_t = 1. \tag{4.7}$$

Recall that $\beta_0 \in \mathbb{R}^N$ is the vector we would like to recover and $w \in \mathbb{R}^n$ is the measurement noise. The vector h^{t+1} is the noise in the effective observation $A^* z^t + \beta^t$, while q^t is the error in the estimate β^t. The proof will show that h^t and m^t are approximately i.i.d. $\mathcal{N}(0, \tau_t^2)$, while q^t is approximately i.i.d. with zero mean and variance σ_t^2.

For the analysis, we work with the general recursion given by (4.1) and (4.2). Notice from (4.1) that for all t,

$$b^t + \lambda_t m^{t-1} = A q^t, \qquad h^{t+1} + \xi_t q^t = A^* m^t. \tag{4.8}$$

Thus we have the matrix equations $X_t = A^* M_t$ and $Y_t = A Q_t$, where

$$\begin{aligned} X_t &:= [h^1 + \xi_0 q^0 \mid h^2 + \xi_1 q^1 \mid \ldots \mid h^t + \xi_{t-1} q^{t-1}], \quad Q_t := [q^0 \mid \ldots \mid q^{t-1}], \\ Y_t &:= [b^0 \mid b^1 + \lambda_1 m^0 \mid \ldots \mid b^{t-1} + \lambda_{t-1} m^{t-2}], \quad M_t := [m^0 \mid \ldots \mid m^{t-1}]. \end{aligned} \tag{4.9}$$

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. Additionally define the matrices

$$\begin{aligned} H_t &:= [h^1 | \ldots | h^t], \qquad & \Xi_t &:= \operatorname{diag}(\xi_0, \ldots, \xi_{t-1}), \\ B_t &:= [b^0 | \ldots | b^{t-1}], & \Lambda_t &:= \operatorname{diag}(\lambda_0, \ldots, \lambda_{t-1}). \end{aligned} \tag{4.10}$$

Note that B_0, H_0, Λ_0, and Ξ_0 are all-zero vectors. Using the above we see that $Y_t = B_t + [0 | M_{t-1}]\Lambda_t$ and $X_t = H_t + Q_t \Xi_t$.

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

$$\alpha^t := (\alpha_0^t, \ldots, \alpha_{t-1}^t)^*, \qquad \gamma^t := (\gamma_0^t, \ldots, \gamma_{t-1}^t)^* \tag{4.11}$$

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

$$m_\parallel^t := \sum_{i=0}^{t-1} \alpha_i^t m^i, \qquad q_\parallel^t := \sum_{i=0}^{t-1} \gamma_i^t q^i. \tag{4.12}$$

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

$$m_\perp^t := m^t - m_\parallel^t, \qquad q_\perp^t := q^t - q_\parallel^t. \tag{4.13}$$

Lemma 4.5 shows that for large n, the entries of α^t and γ^t are concentrated around constants. We now specify these constants and provide some intuition about their values in the special case where the denoising function in the AMP recursion is the Bayes-optimal choice, as in (2.2).

4.2 Concentrating Values

Let $\{\breve{Z}_t\}_{t\geq 0}$ and $\{\tilde{Z}_t\}_{t\geq 0}$ each be sequences of of zero-mean jointly Gaussian random variables whose covariance is defined recursively as follows. For $r, t \geq 0$,

$$\mathbb{E}[\breve{Z}_r \breve{Z}_t] = \frac{\tilde{E}_{r,t}}{\sigma_r \sigma_t}, \qquad \mathbb{E}[\tilde{Z}_r \tilde{Z}_t] = \frac{\breve{E}_{r,t}}{\tau_r \tau_t}, \tag{4.14}$$

where

$$\tilde{E}_{r,t} := \frac{\mathbb{E}[f_r(\tau_{r-1}\tilde{Z}_{r-1}, \beta) f_t(\tau_{t-1}\tilde{Z}_{t-1}, \beta)]}{\delta}, \quad \breve{E}_{r,t} := \mathbb{E}[g_r(\sigma_r \breve{Z}_r, W) g_t(\sigma_t \breve{Z}_t, W)], \tag{4.15}$$

where $W \sim p_w$, and $Z \sim \mathcal{N}(0,1)$ are independent random variables. In the above, we take $f_0(\cdot, \beta) := f_0(0, \beta)$, the initial condition. Note that $\tilde{E}_{t,t} = \sigma_t^2$ and $\breve{E}_{t,t} = \tau_t^2$, thus $\mathbb{E}[\tilde{Z}_t^2] = \mathbb{E}[\breve{Z}_t^2] = 1$.

Define matrices $\tilde{C}^t, \breve{C}^t \in \mathbb{R}^{t \times t}$ for $t \geq 1$ such that

$$\tilde{C}^t_{i+1,j+1} = \tilde{E}_{i,j}, \quad \text{and} \quad \breve{C}^t_{i+1,j+1} = \breve{E}_{i,j}, \quad 0 \leq i, j \leq t-1. \tag{4.16}$$

With these definitions, the concentrating values for γ^t and α^t (if $\tilde{C}^t$ and $\breve{C}^t$ are invertible) are

$$\hat{\gamma}^t := (\tilde{C}^t)^{-1} \tilde{E}_t, \quad \text{and} \quad \hat{\alpha}^t := (\breve{C}^t)^{-1} \breve{E}_t, \tag{4.17}$$

with

$$\tilde{E}_t := (\tilde{E}_{0,t} \ldots, \tilde{E}_{t-1,t})^*, \quad \text{and} \quad \breve{E}_t := (\breve{E}_{0,t} \ldots, \breve{E}_{t-1,t})^*. \tag{4.18}$$

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

$$\begin{aligned} (\sigma_t^\perp)^2 &:= \sigma_t^2 - (\hat{\gamma}^t)^* \tilde{E}_t = \tilde{E}_{t,t} - \tilde{E}_t^* (\tilde{C}^t)^{-1} \tilde{E}_t, \\ (\tau_t^\perp)^2 &:= \tau_t^2 - (\hat{\alpha}^t)^* \breve{E}_t = \breve{E}_{t,t} - \breve{E}_t^* (\breve{C}^t)^{-1} \breve{E}_t. \end{aligned} \tag{4.19}$$

Finally, we define the concentrating values for λ_t and ξ_t as

$$\hat{\lambda}_t := \frac{1}{\delta} \mathbb{E}[f_t'(\tau_{t-1}\tilde{Z}_{t-1}, \beta)], \quad \text{and} \quad \hat{\xi}_t = \mathbb{E}[g_t'(\sigma_t \breve{Z}_t, W)]. \tag{4.20}$$

Since $\{f_t\}_{t\geq 0}$ and $\{g_t\}_{t\geq 0}$ are assumed to be Lipschitz continuous, the derivatives $\{f_t'\}$ and $\{g_t'\}$ are bounded for $t \geq 0$. Therefore λ_t, ξ_t defined in (4.2) and $\hat{\lambda}_t, \hat{\xi}_t$ defined in (4.20) are also bounded. For the AMP recursion, it follows from (4.6) that

$$\hat{\lambda}_t = -\frac{1}{\delta} \mathbb{E}[\eta_{t-1}'(\beta - \tau_{t-1}\tilde{Z}_{t-1})], \quad \text{and} \quad \hat{\xi}_t = 1. \tag{4.21}$$

Lemma 4.1. *If $(\sigma_k^\perp)^2$ and $(\tau_k^\perp)^2$ are bounded below by some positive constants (say $\tilde{c}$ and $\breve{c}$, respectively) for $1 \leq k \leq t$, then the matrices $\tilde{C}^k$ and $\breve{C}^k$ defined in (4.16) are invertible for $1 \leq k \leq t$.*

Proof. We prove the result using induction. Note that $\tilde{C}^1 = \sigma_0^2$ and $\breve{C}^1 = \tau_0^2$ are both strictly positive by assumption and hence invertible. Assume that for some $k < t$, $\tilde{C}^k$ and $\breve{C}^k$ are invertible. The matrix $\tilde{C}^{k+1}$ can be written as

$$\tilde{C}^{k+1} = \begin{bmatrix} \mathsf{M}_1 & \mathsf{M}_2 \\ \mathsf{M}_3 & \mathsf{M}_4 \end{bmatrix},$$

where $\mathsf{M}_1 = \tilde{C}^k \in \mathbb{R}^{k\times k}$, $\mathsf{M}_4 = \tilde{E}_{k,k} = \sigma_k^2$, and $\mathsf{M}_2 = \mathsf{M}_3^* = \tilde{E}_k \in \mathbb{R}^{k\times 1}$ defined in (4.18). By the block inversion formula, $\tilde{C}^{k+1}$ is invertible if M_1 and the Schur complement $\mathsf{M}_4 - \mathsf{M}_3\mathsf{M}_1^{-1}\mathsf{M}_2$ are both invertible. By the induction hypothesis $\mathsf{M}_1 = \tilde{C}^k$ is invertible, and

$$\mathsf{M}_4 - \mathsf{M}_3\mathsf{M}_1^{-1}\mathsf{M}_2 = \tilde{E}_{k,k} - \tilde{E}_k^*(\tilde{C}^k)^{-1}\tilde{E}_k = (\sigma_k^{\perp})^2 \geq \tilde{c} > 0. \tag{4.22}$$

Hence $\tilde{C}^{t+1}$ is invertible. Showing that $\breve{C}^{t+1}$ is invertible is very similar. □

We note that the stopping criterion ensures that $\tilde{C}^t$ and $\breve{C}^t$ are invertible for all t that are relevant to Theorem 3.1.

4.3 Bayes-optimal AMP

The concentrating constants in (4.14)–(4.19) have simple representations in the special case where the denoising function $\eta_t(\cdot)$ is chosen to be Bayes-optimal, i.e., the conditional expectation of β given the noisy observation $\beta + \tau_t Z$, as in (2.2). In this case:

1. It can be shown that $\tilde{E}_{r,t}$ in (4.15) equals σ_t^2 for $0 \leq r \leq t$. This is done in two steps. First verify that the following Markov property holds for the jointly Gaussian $\tilde{Z}_r, \tilde{Z}_t$ with covariance given by (4.14):

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

   We then use the above in the definition of $\tilde{E}_{r,t}$ (with f_t given by (4.6)), and apply the orthogonality principle to show that $\tilde{E}_{r,t} = \sigma_t^2$ for $r \leq t$.

2. Using $\tilde{E}_{r,t} = \sigma_t^2$ in (4.14) and (4.15), we obtain $\breve{E}_{r,t} = \sigma^2 + \sigma_t^2 = \tau_t^2$.

3. From the orthogonality principle, it also follows that for $0 \leq r \leq t$,

   $$\mathbb{E}\left[\left\|\beta^{t+1}\right\|^2\right] = \mathbb{E}\left[\beta^*\beta^{t+1}\right], \quad \text{and} \quad \mathbb{E}\left[\left\|\beta^{r+1}\right\|^2\right] = \mathbb{E}\left[(\beta^{r+1})^*\beta^{t+1}\right],$$

   where $\beta^{t+1} = \mathbb{E}[\beta \mid \beta + \tau_t\tilde{Z}_t]$.

4. With $\tilde{E}_{r,t} = \sigma_t^2$ and $\breve{E}_{r,t} = \tau_t^2$ for $r \leq t$, the quantities in (4.17)–(4.19) simplify to the following for $t > 0$:

   $$\begin{aligned} &\hat{\gamma}^t = [0, \ldots, 0, \sigma_t^2/\sigma_{t-1}^2], \quad \hat{\alpha}^t = [0, \ldots, 0, \tau_t^2/\tau_{t-1}^2], \\ &(\sigma_t^{\perp})^2 := \sigma_t^2\left(1 - \frac{\sigma_t^2}{\sigma_{t-1}^2}\right), \quad (\tau_t^{\perp})^2 := \tau_t^2\left(1 - \frac{\tau_t^2}{\tau_{t-1}^2}\right), \end{aligned} \tag{4.23}$$

   where $\hat{\gamma}^t, \hat{\alpha}^t \in \mathbb{R}^t$.

For the AMP, $m^t = -z^t$ is the modified residual in iteration t, and $q^t = \beta^t - \beta$ is the error in the estimate β^t. Also recall that γ^t and α^t are the coefficients of the projection of m^t and q^t onto $\{m^0, \ldots, m^{t-1}\}$ and $\{q^0, \ldots, q^{t-1}\}$, respectively. The fact that only the last entry of $\hat{\gamma}^t$ is non-zero in the Bayes-optimal case indicates that residual z^t can be well approximated as a linear combination of z^{t-1} and a vector that is independent of $\{z^0, \ldots, z^{t-1}\}$; a similar interpretation holds for the error $q^t = \beta^t - \beta$.

4.4 Conditional Distribution Lemma

We next characterize the conditional distribution of the vectors h^{t+1} and b^t given the matrices in (4.9) as well as β_0, w. Lemmas 4.3 and 4.4 show that the conditional distributions of h^{t+1} and b^t can each be expressed in terms of a standard normal vector and a deviation vector. Lemma 4.5 shows that the norms of the deviation vectors are small with high probability, and provides concentration inequalities for various inner products and functions involving $\{h^{t+1}, q^t, b^t, m^t\}$.

We use the following notation in the lemmas. Given two random vectors X, Y and a sigma-algebra $\mathscr{S}$, $X|_{\mathscr{S}} \stackrel{d}{=} Y$ denotes 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. We suppress the subscript on the matrix if the dimensions are clear from context. For a matrix A with full column rank, $\mathsf{P}_A^{\|} := A(A^*A)^{-1}A^*$ denotes the orthogonal projection matrix onto the column space of A, and $\mathsf{P}_A^{\perp} := \mathsf{I} - \mathsf{P}_A^{\|}$. If A does not have full column rank, $(A^*A)^{-1}$ is interpreted as the pesudoinverse.

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

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

A key ingredient in the proof is the distribution of A conditioned on the sigma algebra $\mathscr{S}_{t_1,t}$ where t_1 is either $t+1$ or t from which we are able to specify the conditional distributions of b^t and h^{t+1} given $\mathscr{S}_{t,t}$ and $\mathscr{S}_{t+1,t}$, respectively. Observing that conditioning on $\mathscr{S}_{t_1,t}$ is equivalent to conditioning on the linear constraints[1]

$$AQ_{t_1} = Y_{t_1}, \ A^*M_t = X_t,$$

the following lemma from [1] specifies the conditional distribution of $A|_{\mathscr{S}_{t_1,t}}$.

Lemma 4.2. *[1, Lemma 10, Lemma 12] The conditional distributions of the vectors in (4.8) satisfy the following, provided $n > t$ and M_t, Q_t have full column rank.*

$$\begin{aligned} A^*m^t|_{\mathscr{S}_{t+1,t}} &\stackrel{d}{=} X_t(M_t^*M_t)^{-1}M_t^*m_{\|}^t + Q_{t+1}(Q_{t+1}^*Q_{t+1})^{-1}Y_{t+1}^*m_{\perp}^t + \mathsf{P}_{Q_{t+1}}^{\perp}\tilde{A}^*m_{\perp}^t, \\ Aq^t|_{\mathscr{S}_{t,t}} &\stackrel{d}{=} Y_t(Q_t^*Q_t)^{-1}Q_t^*q_{\|}^t + M_t(M_t^*M_t)^{-1}X_t^*q_{\perp}^t + \mathsf{P}_{M_t}^{\perp}\hat{A}q_{\perp}^t, \end{aligned}$$

where $m_{\|}^t, m_{\perp}^t, q_{\|}^t, q_{\perp}^t$ are defined in (4.12) and (4.13). Here $\tilde{A}, \hat{A} \stackrel{d}{=} A$ are random matrices independent of $\mathscr{S}_{t+1,t}$ and $\mathscr{S}_{t,t}$.

Lemma 4.3 (Conditional Distribution Lemma)**.** *For the vectors h^{t+1} and b^t defined in (4.1), the following hold for $t \geq 1$, provided $n > t$ and M_t, Q_t have full column rank.*

$$h^1|_{\mathscr{S}_{1,0}} \stackrel{d}{=} \tau_0 Z_0 + \Delta_{1,0}, \quad \text{and} \quad h^{t+1}|_{\mathscr{S}_{t+1,t}} \stackrel{d}{=} \sum_{r=0}^{t-1} \hat{\alpha}_r^t h^{r+1} + \tau_t^{\perp} Z_t + \Delta_{t+1,t}, \tag{4.24}$$

$$b^0|_{\mathscr{S}_{0,0}} \stackrel{d}{=} \sigma_0 Z_0' + \Delta_{0,0}, \quad \text{and} \quad b^t|_{\mathscr{S}_{t,t}} \stackrel{d}{=} \sum_{r=0}^{t-1} \hat{\gamma}_r^t b^r + \sigma_t^{\perp} Z_t' + \Delta_{t,t}. \tag{4.25}$$

where $Z_0, Z_t \in \mathbb{R}^N$ and $Z_0', Z_t' \in \mathbb{R}^n$ are i.i.d. standard Gaussian random vectors that are independent of the corresponding conditioning sigma algebras. The terms $\hat{\gamma}_i^t$ and $\hat{\alpha}_i^t$ for $i \in [t-1]$ are

[1] While conditioning on the linear constraints, we emphasize that only A is treated as random.

defined in (4.17) *and the terms* $(\tau_t^{\perp})^2$ *and* $(\sigma_t^{\perp})^2$ *in* (4.19). *The deviation terms are*

$$\Delta_{0,0} = \left(\frac{\left\| q^0 \right\|}{\sqrt{n}} - \sigma_0 \right) Z_0', \tag{4.26}$$

$$\Delta_{1,0} = \left[\left(\frac{\left\| m^0 \right\|}{\sqrt{n}} - \tau_0 \right) \mathsf{I}_N - \frac{\left\| m^0 \right\|}{\sqrt{n}} \mathsf{P}^{\parallel}_{q^0} \right] Z_0 + q^0 \left(\frac{\left\| q^0 \right\|^2}{n} \right)^{-1} \left(\frac{(b^0)^* m^0}{n} - \xi_0 \frac{\left\| q^0 \right\|^2}{n} \right), \tag{4.27}$$

and for $t > 0$,

$$\begin{aligned} \Delta_{t,t} = &\sum_{r=0}^{t-1} (\gamma_r^t - \hat{\gamma}_r^t) b^r + \left[\left(\frac{\left\| q_{\perp}^t \right\|}{\sqrt{n}} - \sigma_t^{\perp} \right) \mathsf{I}_n - \frac{\left\| q_{\perp}^t \right\|}{\sqrt{n}} \mathsf{P}^{\parallel}_{M_t} \right] Z_t' \\ &+ M_t \left(\frac{M_t^* M_t}{n} \right)^{-1} \left(\frac{H_t^* q_{\perp}^t}{n} - \frac{{M_t}^*}{n} \left[\lambda_t m^{t-1} - \sum_{r=1}^{t-1} \lambda_r \gamma_r^t m^{r-1} \right] \right), \end{aligned} \tag{4.28}$$

$$\begin{aligned} \Delta_{t+1,t} = &\sum_{r=0}^{t-1} (\alpha_r^t - \hat{\alpha}_r^t) h^{r+1} + \left[\left(\frac{\left\| m_{\perp}^t \right\|}{\sqrt{n}} - \tau_t^{\perp} \right) \mathsf{I}_N - \frac{\left\| m_{\perp}^t \right\|}{\sqrt{n}} \mathsf{P}^{\parallel}_{Q_{t+1}} \right] Z_t \\ &+ Q_{t+1} \left(\frac{Q_{t+1}^* Q_{t+1}}{n} \right)^{-1} \left(\frac{B_{t+1}^* m_{\perp}^t}{n} - \frac{Q_{t+1}^*}{n} \left[\xi_t q^t - \sum_{i=0}^{t-1} \xi_i \alpha_i^t q^i \right] \right). \end{aligned} \tag{4.29}$$

Proof. We begin by demonstrating (4.25). By (4.1) it follows that

$$b^0|_{\mathscr{S}_{0,0}} = A q^0 \stackrel{d}{=} (\left\| q^0 \right\| / \sqrt{n}) Z_0',$$

where $Z_0' \in \mathbb{R}^n$ is an i.i.d. standard Gaussian random vector, independent of $\mathscr{S}_{0,0}$.

Define $\mathbf{Q}_t := Q_t^* Q_t$ and $\mathbf{M}_t := M_t^* M_t$. For the case $t \geq 1$, we use Lemma 4.2 to write

$$\begin{aligned} b^t|_{\mathscr{S}_{t,t}} &= (A q^t - \lambda_t m^{t-1})|_{\mathscr{S}_{t,t}} \stackrel{d}{=} Y_t \mathbf{Q}_t^{-1} Q_t^* q_{\parallel}^t + M_t \mathbf{M}_t^{-1} X_t^* q_{\perp}^t + \mathsf{P}^{\perp}_{M_t} \tilde{A} q_{\perp}^t - \lambda_t m^{t-1} \\ &= B_t \mathbf{Q}_t^{-1} Q_t^* q_{\parallel}^t + [0 | M_{t-1}] \Lambda_t \mathbf{Q}_t^{-1} Q_t^* q_{\parallel}^t + M_t \mathbf{M}_t^{-1} H_t^* q_{\perp}^t + \mathsf{P}^{\perp}_{M_t} \tilde{A} q_{\perp}^t - \lambda_t m^{t-1}. \end{aligned}$$

The last equality above is obtained using $Y_t = B_t + [0|M_{t-1}]\Lambda_t$, and $X_t = H_t + \Xi_t Q_t$. Noticing that $B_t \mathbf{Q}_t^{-1} Q_t^* q_{\parallel}^t = \sum_{i=0}^{t-1} \gamma_i^t b^i$ and $\mathsf{P}^{\perp}_{M_t} \tilde{A} q_{\perp}^t = (\mathsf{I} - \mathsf{P}^{\parallel}_{M_t}) \tilde{A} q_{\perp}^t \stackrel{d}{=} (\mathsf{I} - \mathsf{P}^{\parallel}_{M_t}) \frac{\left\| q_{\perp}^t \right\|}{\sqrt{n}} Z_t'$ where $Z_t' \in \mathbb{R}^n$ is an i.i.d. standard Gaussian random vector, it follows that

$$b^t|_{\mathscr{S}_{t,t}} \stackrel{d}{=} (\mathsf{I} - \mathsf{P}^{\parallel}_{M_t}) \frac{\left\| q_{\perp}^t \right\|}{\sqrt{n}} Z_t' + \sum_{i=0}^{t-1} \gamma_i^t b^i + [0 | M_{t-1}] \Lambda_t \mathbf{Q}_t^{-1} Q_t^* q_{\parallel}^t + M_t \mathbf{M}_t^{-1} H_t^* q_{\perp}^t - \lambda_t m^{t-1}. \tag{4.30}$$

All the quantities in the RHS of (4.30) except Z_t' are in the conditioning sigma-field. We can rewrite (4.30) with the following pair of values:

$$\begin{aligned} b^t|_{\mathscr{S}_{t,t}} &\stackrel{d}{=} \sum_{r=0}^{t-1} \hat{\gamma}_r^t b^r + \sigma_t^{\perp} Z_t' + \Delta_{t,t}, \\ \Delta_{t,t} &= \sum_{r=0}^{t-1} (\gamma_r^t - \hat{\gamma}_r^t) b^r + \left[\left(\frac{\left\| q_{\perp}^t \right\|}{\sqrt{n}} - \sigma_t^{\perp} \right) \mathsf{I} - \frac{\left\| q_{\perp}^t \right\|}{\sqrt{n}} \mathsf{P}^{\parallel}_{M_t} \right] Z_t' \\ &\quad + [0 | M_{t-1}] \Lambda_t \mathbf{Q}_t^{-1} Q_t^* q_{\parallel}^t + M_t \mathbf{M}_t^{-1} H_t^* q_{\perp}^t - \lambda_t m^{t-1}. \end{aligned}$$

The above definition of $\Delta_{t,t}$ equals that given in (4.28) since

$$[0|M_{t-1}]\Lambda_t \mathbf{Q}_t^{-1} Q_t^* q_\|^t + M_t \mathbf{M}_t^{-1} M_t^* \left(\lambda_t m^{t-1} - \sum_{i=0}^{t-2} \lambda_{i+1} \gamma_{i+1}^t m^i \right) - \lambda_t m^{t-1}$$
$$= \sum_{j=0}^{t-2} \lambda_{j+1} \gamma_{j+1}^t m^j + \lambda_t m^{t-1} - \sum_{i=0}^{t-2} \lambda_{i+1} \gamma_{i+1}^t m^i - \lambda_t m^{t-1} = 0.$$

This completes the proof of (4.25). Result (4.24) can be shown similarly. □

The conditional distribution representation in Lemma 4.3 implies that for each $t \geq 0$, h^{t+1} is the sum of an i.i.d. $\mathcal{N}(0, \tau_t^2)$ random vector plus a deviation term. Similarly b^t is the sum of an i.i.d. $\mathcal{N}(0, \sigma_t^2)$ random vector and a deviation term. This is made precise in the following lemma.

Lemma 4.4. *For $t \geq 0$, let $Z_t' \in \mathbb{R}^n$, $Z_t \in \mathbb{R}^N$ be independent standard normal random vectors. Let $b_{\text{pure}}^0 = \sigma_0 Z_0'$, $h_{\text{pure}}^1 = \tau_0 Z_0$, and recursively define for $t \geq 1$:*

$$b_{\text{pure}}^t = \sum_{r=0}^{t-1} \hat{\gamma}_r^t b_{\text{pure}}^r + \sigma_t^\perp Z_t', \qquad h_{\text{pure}}^{t+1} = \sum_{r=0}^{t-1} \hat{\alpha}_r^t h_{\text{pure}}^{r+1} + \tau_t^\perp Z_t. \tag{4.31}$$

Then for $t \geq 0$, the following statements hold.

1. *For $j \in [N]$ and $k \in [n]$,*

$$(b_{\text{pure}_j}^0, \ldots, b_{\text{pure}_j}^t) \overset{d}{=} (\sigma_0 \breve{Z}_0, \ldots, \sigma_t \breve{Z}_t), \quad \text{and} \quad (h_{\text{pure}_k}^1, \ldots, h_{\text{pure}_k}^{t+1}) \overset{d}{=} (\tau_0 \tilde{Z}_0, \ldots, \tau_t \tilde{Z}_t), \tag{4.32}$$

   where $\{\breve{Z}_t\}_{t\geq 0}$ and $\{\tilde{Z}_t\}_{t\geq 0}$ are the jointly Gaussian random variables defined in Sec. 4.2.

2. *For $t \geq 0$,*

$$b_{\text{pure}}^t = \sum_{i=0}^{t} Z_i' \, \sigma_i^\perp \mathsf{c}_i^t, \qquad h_{\text{pure}}^t = \sum_{i=0}^{t} Z_i \, \tau_i^\perp \mathsf{d}_i^t, \tag{4.33}$$

   where the constants $\{\mathsf{c}_i^t\}_{0\leq i\leq t}$ and $\{\mathsf{d}_i^t\}_{0\leq i\leq t}$ are recursively defined as follows, starting with $\mathsf{c}_0^0 = 1$ and $\mathsf{d}_0^0 = 1$. For $t > 0$,

$$\mathsf{c}_t^t = 1, \qquad \mathsf{c}_i^t = \sum_{r=i}^{t-1} \mathsf{c}_i^r \hat{\gamma}_r^t, \quad \text{for } 0 \leq i \leq (t-1), \tag{4.34}$$

$$\mathsf{d}_t^t = 1, \qquad \mathsf{d}_i^t = \sum_{r=i}^{t-1} \mathsf{d}_i^r \hat{\alpha}_r^t, \quad \text{for } 0 \leq i \leq (t-1). \tag{4.35}$$

3. *The conditional distributions in Lemma 4.3 can be expressed as*

$$b^t|_{\mathscr{S}_{t,t}} \overset{d}{=} b_{\text{pure}}^t + \sum_{r=0}^{t} \mathsf{c}_r^t \, \Delta_{r,r}, \qquad h^{t+1}|_{\mathscr{S}_{t+1,t}} \overset{d}{=} h_{\text{pure}}^{t+1} + \sum_{r=0}^{t} \mathsf{d}_r^t \, \Delta_{r+1,r}. \tag{4.36}$$

Proof. We prove (4.32) by induction. We prove the b^t_{pure} result; the proof for h^t_{pure} is very similar. The base case of $t = 0$ holds by the definition of b^0_{pure}. Assume towards induction that (4.32) holds for $(b^0_{\text{pure}}, \ldots, b^{t-1}_{\text{pure}})$. Then using (4.31), b^t_{pure} has the same distribution as $\sum_{r=0}^{t-1} \hat{\gamma}^t_r \sigma_r \breve{Z}_r + \sigma^\perp_t Z$ where $Z \in \mathbb{R}^n$ is a standard Gaussian random vector independent of $\breve{Z}_0, \ldots, \breve{Z}_{t-1}$. We now show that $\sum_{r=0}^{t-1} \hat{\gamma}^t_r \sigma_r \breve{Z}_r + \sigma^\perp_t Z \overset{d}{=} \sigma_t \breve{Z}_t$ by demonstrating that: (i) $\text{var}(\sum_{r=0}^{t-1} \hat{\gamma}^t_r \sigma_r \breve{Z}_r + \sigma^\perp_t Z) = \sigma^2_t$; and (ii) $\mathbb{E}[\sigma_k \breve{Z}_k (\sum_{r=0}^{t-1} \hat{\gamma}^t_r \sigma_r \breve{Z}_r + \sigma^\perp_t Z)] = \sigma_k \sigma_t \mathbb{E}[\breve{Z}_k \breve{Z}_t] = \tilde{E}_{k,t}$, for $0 \le k \le (t-1)$. The variance is

$$\mathbb{E}\Big(\sum_{r=0}^{t-1} \hat{\gamma}^t_r \sigma_r \breve{Z}_r + \sigma^\perp_t Z \Big)^2 = \sum_{r=0}^{t-1} \sum_{k=0}^{t-1} \hat{\gamma}^t_r \hat{\gamma}^t_k \tilde{E}_{k,r} + (\sigma^\perp_t)^2 = \sigma^2_t,$$

where the last equality follows from rewriting the double sum as follows using the definitions in Section 4.1:

$$\sum_{r,k} \hat{\gamma}^t_r \hat{\gamma}^t_k \tilde{E}_{k,r} = (\hat{\gamma}^t)^* \tilde{C}^t \hat{\gamma}^t = [\tilde{E}^*_t (\tilde{C}^t)^{-1}] \tilde{C}^t [(\tilde{C}^t)^{-1} \tilde{E}_t] = \tilde{E}^*_t (\tilde{C}^t)^{-1} \tilde{E}_t = \tilde{E}_{t,t} - (\sigma^\perp_t)^2. \tag{4.37}$$

Next, for any $0 \le k \le t-1$, we have

$$\mathbb{E}[\sigma_k \breve{Z}_k (\sum_{r=0}^{t-1} \hat{\gamma}^t_r \sigma_r \breve{Z}_r + \sigma^\perp_t Z)] \overset{(a)}{=} \sum_{r=0}^{t-1} \tilde{E}_{k,r} \hat{\gamma}^t_r \overset{(b)}{=} [\tilde{C} \hat{\gamma}^t]_{k+1} \overset{(c)}{=} \tilde{E}_{k,t}.$$

In the above, step (a) follows from (4.14); step (b) by recognizing from (4.16) that the required sum is the inner product of $\hat{\gamma}^t$ with row $(k+1)$ of $\tilde{C}^t$; step (c) from the definition of $\hat{\gamma}^t$ in (4.17). This proves (4.32).

Next we show the expression for b^t_{pure} in (4.33) using induction; the proof for h^t_{pure} is similar. The base case of $t = 0$ holds by definition because $\sigma^\perp_1 = \sigma_1$. Using the induction hypothesis that (4.33) holds for $b^0_{\text{pure}}, \ldots, b^{t-1}_{\text{pure}}$, the defintion (4.31) can be written as

$$b^t_{\text{pure}} = \sum_{r=0}^{t-1} \hat{\gamma}^t_r \Big(\sum_{i=0}^{r} Z'_i \sigma^\perp_i \mathsf{c}^r_i \Big) + \sigma^\perp_t Z'_t = \sum_{i=0}^{t-1} Z'_i \sigma^\perp_i \Big(\sum_{r=i}^{t-1} \hat{\gamma}^t_r \mathsf{c}^r_i \Big) + \sigma^\perp_t Z'_t = \sum_{i=0}^{t} Z'_i \sigma^\perp_i \mathsf{c}^t_i, \tag{4.38}$$

where the last inequality follows from the definition of c^t_i for $0 \le i \le t$ in (4.35). This proves (4.33).

The expressions for the conditional distribution of b^t and h^{t+1} in (4.36) can be similarly obtained from (4.25) and (4.24) using an induction argument. □

4.5 Main Concentration Lemma

For $t \ge 0$, let

$$K_t = C^{2t}(t!)^{10}, \quad \kappa_t = \frac{1}{c^{2t}(t!)^{22}}, \quad K'_t = C(t+1)^5 K_t, \quad \kappa'_t = \frac{\kappa_t}{c(t+1)^{11}}, \tag{4.39}$$

where $C, c > 0$ are universal constants (not depending on t, n, or ϵ). To keep the notation compact, we use K, κ, κ' to denote generic positive universal constants whose values may change through the lemma statement and the proof.

Lemma 4.5. *The following statements hold for $1 \le t < T^*$ and $\epsilon \in (0, 1)$.*

(a)

$$P\left(\frac{1}{N}\left\|\Delta_{t+1,t}\right\|^2 \geq \epsilon\right) \leq Kt^2K'_{t-1}\exp\left\{-\frac{\kappa\kappa'_{t-1}n\epsilon}{t^4}\right\}, \tag{4.40}$$

$$P\left(\frac{1}{n}\left\|\Delta_{t,t}\right\|^2 \geq \epsilon\right) \leq Kt^2K_{t-1}\exp\left\{-\frac{\kappa\kappa_{t-1}n\epsilon}{t^4}\right\}. \tag{4.41}$$

(b) i) *Let* $X_n \doteqdot c$ *be shorthand for* $P(|X_n - c| \geq \epsilon) \leq Kt^3K'_{t-1}\exp\left\{-\kappa\kappa'_{t-1}n\epsilon^2/t^7\right\}$. *Then for pseudo-Lipschitz functions* $\phi_h : \mathbb{R}^{t+2} \to \mathbb{R}$

$$\frac{1}{N}\sum_{i=1}^{N}\phi_h\left(h_i^1, \ldots, h_i^{t+1}, \beta_{0_i}\right) \doteqdot \mathbb{E}\,\phi_h\left(\tau_0\tilde{Z}_0, \ldots, \tau_t\tilde{Z}_t, \beta\right). \tag{4.42}$$

The random variables $\tilde{Z}_0, \ldots, \tilde{Z}_t$ *are jointly Gaussian with zero mean and covariance given by* (4.14)*, and are independent of* $\beta \sim p_\beta$.

ii) *Let* $\psi_h : \mathbb{R}^2 \to \mathbb{R}$ *be a bounded function that is differentiable in the first argument except possibly at a finite number of points, with bounded derivative where it exists. Then,*

$$P\left(\left|\frac{1}{N}\sum_{i=1}^{N}\psi_h(h_i^{t+1}, \beta_{0_i}) - \mathbb{E}\,\psi_h(\tau_t\tilde{Z}_t, \beta)\right| \geq \epsilon\right) \leq Kt^2K'_{t-1}\exp\left\{\frac{-\kappa\kappa'_{t-1}n\epsilon^2}{t^4}\right\}. \tag{4.43}$$

As above, $\tilde{Z}_t \sim \mathcal{N}(0,1)$ *and* $\beta \sim p_\beta$ *are independent.*

iii) *Let* $X_n \doteq c$ *be shorthand for* $P(|X_n - c| \geq \epsilon) \leq Kt^3K_{t-1}\exp\left\{-\kappa\kappa_{t-1}n\epsilon^2/t^7\right\}$. *Then for pseudo-Lipschitz functions* $\phi_b : \mathbb{R}^{t+2} \to \mathbb{R}$

$$\frac{1}{n}\sum_{i=1}^{n}\phi_b\left(b_i^0, \ldots, b_i^t, w_i\right) \doteq \mathbb{E}\,\phi_b\left(\sigma_0\breve{Z}_0, \ldots, \sigma_t\breve{Z}_t, W\right). \tag{4.44}$$

The random variables $\breve{Z}_0, \ldots, \breve{Z}_t$ *are jointly Gaussian with zero mean and covariance given by* (4.14)*, and are independent of* $W \sim p_w$.

iv) *Let* $\psi_b : \mathbb{R} \to \mathbb{R}$ *be a bounded function that is differentiable in the first argument except possibly at a finite number of points, with bounded derivative where it exists. Then,*

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\psi_b(b_i^t, w_i) - \mathbb{E}\,\psi_b(\sigma_t\breve{Z}_t, W)\right| \geq \epsilon\right) \leq Kt^2K_{t-1}\exp\left\{\frac{-\kappa\kappa_{t-1}n\epsilon^2}{t^4}\right\}. \tag{4.45}$$

As above, $\breve{Z}_t \sim \mathcal{N}(0,1)$ *and* $W \sim p_w$ *are independent.*

(c)

$$\frac{(h^{t+1})^*q^0}{n} \doteqdot 0, \quad \frac{(h^{t+1})^*\beta_0}{n} \doteqdot 0, \tag{4.46}$$

$$\frac{(b^t)^*w}{n} \doteq 0. \tag{4.47}$$

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

$$\frac{(h^{r+1})^*h^{t+1}}{N} \doteqdot \breve{E}_{r,t}, \tag{4.48}$$

$$\frac{(b^r)^*b^t}{n} \doteq \tilde{E}_{r,t}. \tag{4.49}$$

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

$$\frac{(q^0)^* q^{t+1}}{n} \doteq \tilde{E}_{0,t+1}, \quad \frac{(q^{r+1})^* q^{t+1}}{n} \doteq \tilde{E}_{r+1,t+1}, \tag{4.50}$$

$$\frac{(m^r)^* m^t}{n} \doteq \breve{E}_{r,t}. \tag{4.51}$$

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

$$\lambda_t \doteq \hat{\lambda}_t, \quad \frac{(h^{t+1})^* q^{r+1}}{n} \doteq \hat{\lambda}_{r+1} \breve{E}_{r,t}, \quad \frac{(h^{r+1})^* q^{t+1}}{n} \doteq \hat{\lambda}_{t+1} \breve{E}_{r,t}, \tag{4.52}$$

$$\xi_t \doteq \hat{\xi}_t, \quad \frac{(b^r)^* m^t}{n} \doteq \hat{\xi}_t \tilde{E}_{r,t}, \quad \frac{(b^t)^* m^r}{n} \doteq \hat{\xi}_r \tilde{E}_{r,t} \tag{4.53}$$

(g) Let $\mathbf{Q}_{t+1} := \frac{1}{n} Q_{t+1}^* Q_{t+1}$ *and* $\mathbf{M}_t := \frac{1}{n} M_t^* M_t$. *Then,*

$$P\left(\mathbf{Q}_{t+1} \text{ is singular}\right) \le t K_{t-1} e^{-\kappa_{t-1} \kappa n}, \tag{4.54}$$

$$P\left(\mathbf{M}_t \text{ is singular}\right) \le t K_{t-1} e^{-\kappa_{t-1} \kappa n}. \tag{4.55}$$

When the inverses of $\mathbf{Q}_{t+1}, \mathbf{M}_t$ *exist,*

$$\begin{aligned} &P\left(\left|\left[\mathbf{Q}_{t+1}^{-1} - (\tilde{C}^{t+1})^{-1}\right]_{i+1,j+1}\right| \ge \epsilon\right) \le K K'_{t-1} \exp\left\{-\kappa \kappa'_{t-1} n \epsilon^2\right\}, \\ &P\left(\left|\gamma_i^{t+1} - \hat{\gamma}_i^{t+1}\right| \ge \epsilon\right) \le K t^4 K'_{t-1} \exp\left\{\frac{-\kappa \kappa'_{t-1} n \epsilon^2}{t^9}\right\}, \quad 0 \le i, j \le t. \end{aligned} \tag{4.56}$$

$$\begin{aligned} &P\left(\left|\left[\mathbf{M}_t^{-1} - (\breve{C}^t)^{-1}\right]_{i+1,j+1}\right| \ge \epsilon\right) \le K K_{t-1} \exp\left\{-\kappa \kappa_{t-1} n \epsilon^2\right\}, \\ &P\left(\left|\alpha_i^t - \hat{\alpha}_i^t\right| \ge \epsilon\right) \le K t^4 K_{t-1} \exp\left\{\frac{-\kappa \kappa_{t-1} n \epsilon^2}{t^9}\right\}, \quad 0 \le i, j \le (t-1). \end{aligned} \tag{4.57}$$

where $\hat{\gamma}^{t+1}$ *and* $\hat{\alpha}^t$ *are defined in* (4.17).

(h) With $\sigma_{t+1}^{\perp}, \tau_t^{\perp}$ *defined in* (4.19),

$$P\left(\left|\frac{1}{n}\left\|q_{\perp}^{t+1}\right\|^2 - (\sigma_{t+1}^{\perp})^2\right| \ge \epsilon\right) \le K t^5 K'_{t-1} \exp\left\{\frac{-\kappa \kappa'_{t-1} n \epsilon^2}{t^{11}}\right\}, \tag{4.58}$$

$$P\left(\left|\frac{1}{n}\left\|m_{\perp}^{t}\right\|^2 - (\tau_t^{\perp})^2\right| \ge \epsilon\right) \le K t^5 K_{t-1} \exp\left\{\frac{-\kappa \kappa_{t-1} n \epsilon^2}{t^{11}}\right\}. \tag{4.59}$$

4.6 Remarks on Lemma 4.5

The proof of Theorem 3.1 below only requires the concentration result in part (b).(i) of Lemma 4.5, but the proof of part (b).(i) hinges on the other parts of the lemma. The proof of Lemma 4.5, given in Section 5, uses induction starting at time $t = 0$, sequentially proving the concentration results in parts $(a) - (h)$. The proof is long, but is based on a sequence of a few key steps which we summarize here.

The main result that needs to be proved (part (b).(i), (4.42)) is that within the normalized sum of the pseudo-Lipschitz function ϕ_h, the inputs $h^1, \ldots, h^{t+1}$ can be effectively replaced by

$\tau_0\tilde{Z}_0, \ldots, \tau_t\tilde{Z}_t$, respectively. To prove this, we use the representation for h^{t+1} given by Lemma 4.3, and show that the deviation term given by (4.3) can be effectively dropped. In order to show that the deviation term can be dropped, we need to prove the concentration results in parts (c) – (h) of Lemma 4.5. Parts (b).(ii), (b).(iii), and (b).(iv) of the lemma are used to establish the results in parts (c) – (h).

The concentration constants κ_t, K_t: The concentration results in Lemma 4.5 and Theorem 3.1 for AMP iteration $t \geq 1$ are of the form $K_t e^{-\kappa_t n\epsilon^2}$, where κ_t, K_t are given in (4.39). Due to the inductive nature of the proof, the concentration results for step t depend on those corresponding to all the previous steps — this determines how κ_t, K_t scale with t.

The $t!$ terms in κ_t, K_t can be understood as follows. Suppose that we want prove a concentration result for a quantity that can be expressed as a sum of t terms with step indices $1, \ldots, t$. (A typical example is $\Delta_{t+1,t}$ in (4.3).) For such a term, the deviation from the deterministic concentrating value is less than ϵ if the deviation in each of the terms in the sum is less than ϵ/t. The induction hypothesis (for steps $1, \ldots, t$) is then used to bound the ϵ/t-deviation probability for each term in the sum. This introduces factors of $1/t$ and t multiplying the exponent and pre-factor, respectively, in each step t (see Lemma A.2), which results in the $t!$ terms in K_t and κ_t.

The $(C_2)^t$ and $(c_2)^t$ terms in κ_t, K_t arise due to quantities that can be expressed as the *product* of two terms, for each of which we have a concentration result available (due to the induction hypothesis). This can be used to bound the ϵ-deviation probability of the product, but with a smaller exponent and a larger prefactor (see Lemma A.3). Since this occurs in each step of the induction, the constants K_t, κ_t have terms of the form $(C_2)^t, (c_2)^t$, respectively.

Comparison with earlier work: Lemmas 4.3 and 4.5 are similar to the main technical lemma in [1, Lemma 1], in that they both analyze the behavior of similar functions and inner products arising in the AMP. The key difference is that Lemma 4.5 replaces the asymptotic convergence statements in [1] with concentration inequalities. Other differences from [1, Lemma 1] include:

- Lemma 4.5 gives explicit values for the deterministic limits in parts (c)–(h), which are needed in other parts of our proof.
- Lemma 4.3 characterizes the the conditional distribution of the vectors h^{t+1} and b^t as the sum of an ideal distribution and a deviation term. [1, Lemma 1(a)] is a similar distributional characterization of h^{t+1} and b^t, however it does not use the ideal distribution. We found that working with the ideal distribution throughout Lemma 4.5 simplified our proof.

4.7 Proof of Theorem 3.1

Proof. Applying Part (b).(i) of Lemma 4.5 to a pseudo-Lipschitz (PL) function of the form $\phi_h(h^{t+1}, \beta_0)$, we get

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

where the random variables $Z \sim N(0,1)$ and $\beta \sim p_\beta$ are independent. Now let $\phi_h(h_i^{t+1}, \beta_{0_i}) := \phi(\eta_t(\beta_{0_i} - h_i^{t+1}), \beta_{0_i})$, where ϕ is the PL function in the statement of the theorem. The function $\phi_h(h_i^{t+1}, \beta_{0_i})$ is PL since ϕ is PL and η_t is Lipschitz. We therefore obtain

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

The proof is completed by noting from (1.3) and (4.5) that $\beta^{t+1} = \eta_t(A^* z^t + \beta^t) = \eta_t(\beta_0 - h^{t+1})$. $\square$

5 Proof of Lemma 4.5

5.1 Mathematical Preliminaries

Some of the results below can be found in [1, 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, and let $\tilde{A} \in \mathbb{R}^{n\times N}$ be a matrix with independent $\mathcal{N}(0, 1/n)$ entries. Then:*

(a)

$$\tilde{A}u \overset{d}{=} \frac{1}{\sqrt{n}} \|u\| Z_u \quad \text{and} \quad \tilde{A}^* v \overset{d}{=} \frac{1}{\sqrt{n}} \|v\| Z_v,$$

where $Z_u \in \mathbb{R}^n$ and $Z_v \in \mathbb{R}^N$ are i.i.d. standard Gaussian random vectors.

(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 $\|w_\ell\|^2 = n$ for $\ell \in [d]$, 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 $\mathsf{P}^{\|}_{\mathcal{W}}\tilde{A}u \overset{d}{=} \frac{1}{\sqrt{n}} \|u\| \mathsf{P}^{\|}_{\mathcal{W}} Z_u \overset{d}{=} \frac{1}{\sqrt{n}} \|u\| Dx$ where $x \in \mathbb{R}^d$ is a random vector with i.i.d. $\mathcal{N}(0, 1/n)$ entries.

Fact 2 (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 3. *Let $v_1, \ldots, v_t$ be a sequence of vectors in $\mathbb{R}^n$ such that for $i \in [t]$ $\frac{1}{n}\left\|v_i - \mathsf{P}^{\|}_{i-1}(v_i)\right\|^2 \ge c$, where c is a positive constant that does not depend on n, and $\mathsf{P}^{\|}_{i-1}$ 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 $\lambda_{\min} \ge c'_t$, where c'_t is a positive constant (not depending on n).*

Fact 4. *Let $g : \mathbb{R} \to \mathbb{R}$ be a bounded function. For all $s, \Delta \in \mathbb{R}$ such that g is differentiable in the closed interval between s and $s+\Delta$, there exists a constant $c > 0$ such that $|g(s + \Delta) - g(s)| \le c\,|\Delta|$.*

We also use several concentration results listed in Appendices A and B, with proofs provided for the results that are non-standard. Some of these may be of independent interest, e.g., concentration of sums of a pseudo-Lipschitz function of sub-Gaussians (Lemma B.4).

The proof of Lemma 4.5. proceeds by induction on t. We label as $\mathcal{H}^{t+1}$ the results (4.40), (4.42), (4.43), (4.46), (4.48), (4.50), (4.52), (4.54), (4.56), (4.58) and similarly as $\mathcal{B}^t$ the results (4.41), (4.44), (4.45), (4.47), (4.49), (4.51), (4.53), (4.55), (4.57), (4.59). The proof consists of showing 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 $s \le t$, then $\mathcal{B}_t$ holds.
4. if $\mathcal{B}_r, \mathcal{H}_s$ holds for all $r \le t$ and $s \le t$, then $\mathcal{H}_{t+1}$ holds.

For the proofs of parts (b).(ii) and (b).(iv), for brevity we assume that the functions ψ_h and ψ_b are differentiable everywhere. The case where they are not differentiable at a finite number of points involves additional technical details; see Appendix D.

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

We wish to show results (a)-(h) in (4.41), (4.44), (4.45), (4.47), (4.49), (4.51), (4.53), (4.55), (4.57), (4.59).

(a) We have

$$P\left(\frac{\|\Delta_{0,0}\|^2}{n} \geq \epsilon\right) \overset{(a)}{\leq} P\left(\left|\frac{\|q^0\|}{\sqrt{n}} - \sigma_0^{\perp}\right| \geq \sqrt{\frac{\epsilon}{2}}\right) + P\left(\left|\frac{\|Z_0'\|}{\sqrt{n}} - 1\right| \geq \sqrt{\frac{\epsilon}{2}}\right)$$
$$\overset{(b)}{\leq} K\exp\left\{-\kappa\varepsilon_2 n\epsilon/4\right\} + 2\exp\left\{-n\epsilon/8\right\}.$$

Step (a) is obtained using the definition of $\Delta_{0,0}$ in (4.26), and then applying Lemma A.3. For step (b), we use (4.3), Lemma A.4, and Lemma B.2.

(b).(iii) For $t = 0$, the LHS of (4.44) can be bounded as

$$\begin{aligned}
&P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\phi_b\left(b_i^0, w_i\right) - \mathbb{E}[\phi_b(\sigma_0\breve{Z}_0, W)]\right| \geq \epsilon\right) \\
&\overset{(a)}{=} P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\phi_b\left(\sigma_0 Z_{0_i}' + [\Delta_{0,0}]_i, w_i\right) - \mathbb{E}[\phi_b(\sigma_0\breve{Z}_0, W)]\right| \geq \epsilon\right) \\
&\overset{(b)}{\leq} P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\phi_b\left(\sigma_0 Z_{0_i}', w_i\right) - \mathbb{E}[\phi_b(\sigma_0\breve{Z}_0, W)]\right| \geq \frac{\epsilon}{2}\right) \\
&\quad + P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\phi_b\left(\sigma_0 Z_{0_i}' + [\Delta_{0,0}]_i, w_i\right) - \frac{1}{n}\sum_{i=1}^{n}\phi_b\left(\sigma_0 Z_{0_i}', w_i\right)\right| \geq \frac{\epsilon}{2}\right).
\end{aligned} \tag{5.1}$$

Step (a) uses the conditional distribution of b^0 given in (4.25), and step (b) follows from Lemma A.2. Label the terms on the RHS of (5.1) as T_1 and T_2. Term T_1 can be upper bounded by $Ke^{-\kappa n\epsilon^2}$ using Lemma B.4. We now show a similar upper bound for term T_2.

$$\begin{aligned}
T_2 &\overset{(a)}{\leq} P\left(\frac{1}{n}\sum_{i=1}^{n} L\left(1 + 2\left|\sigma_0 Z_{0_i}'\right| + \left|[\Delta_{0,0}]_i\right| + 2\left|w_i\right|\right)\left|[\Delta_{0,0}]_i\right| \geq \frac{\epsilon}{2}\right) \\
&\overset{(b)}{\leq} P\left(\frac{\|\Delta_{0,0}\|}{\sqrt{n}}\left\|\frac{\mathbf{1}}{\sqrt{n}} + \frac{|\Delta_{0,0}|}{\sqrt{n}} + 2\sigma_0\frac{|Z_0'|}{\sqrt{n}} + 2\frac{|w|}{\sqrt{n}}\right\| \geq \frac{\epsilon}{2L}\right) \\
&\overset{(c)}{\leq} P\left(\frac{\|\Delta_{0,0}\|}{\sqrt{n}} \cdot \left(1 + \frac{\|\Delta_{0,0}\|}{\sqrt{n}} + 2\sigma_0\frac{\|Z_0'\|}{\sqrt{n}} + 2\frac{\|w\|}{\sqrt{n}}\right) \geq \frac{\epsilon}{4L}\right),
\end{aligned} \tag{5.2}$$

where inequality (a) holds because ϕ_b is pseudo-Lipschitz with constant $L > 0$. Inequality (b) follows from Cauchy-Schwarz (with $\mathbf{1}$ denoting the all-ones vector). Inequality (c) is obtained by applying Lemma C.3. From (5.2), we have

$$\begin{aligned}
T_2 &\leq P\left(\frac{\|w\|}{\sqrt{n}} \geq \sigma + 1\right) + P\left(\frac{\|Z_0'\|}{\sqrt{n}} \geq 2\right) + P\left(\frac{\|\Delta_{0,0}\|}{\sqrt{n}} \geq \frac{\epsilon\min\left\{1, \frac{1}{4L}\right\}}{4 + 4\sigma_0 + 2\sigma}\right) \\
&\overset{(a)}{\leq} Ke^{-\kappa n} + e^{-n} + Ke^{-\kappa n\epsilon^2},
\end{aligned} \tag{5.3}$$

where to obtain (a), we use assumption (1.6), Lemma B.2, and $\mathcal{B}_0(a)$ proved above.

(b).(iv) For $t = 0$, the probability in (4.45) can be bounded as

$$\begin{aligned}
&P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\psi_b(b_i^0, w_i) - \mathbb{E}[\psi_b(\sigma_0\breve{Z}_0, W)]\right| \geq \epsilon\right) \\
&\overset{(a)}{=} P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\psi_b(\sigma_0 Z'_{0_i} + [\Delta_{0,0}]_i, w_i) - \mathbb{E}[\psi_b(\sigma_0\breve{Z}_0, W)]\right| \geq \epsilon\right) \\
&\overset{(b)}{\leq} P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\left(\psi_b(\sigma_0 Z'_{0_i} + [\Delta_{0,0}]_i, w_i) - \psi_b(\sigma_0 Z'_{0_i}, w_i)\right)\right| \geq \frac{\epsilon}{2}\right) \\
&\quad + P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\psi_b(\sigma_0 Z'_{0_i}, w_i) - \mathbb{E}[\psi_b(\sigma_0\breve{Z}_0, W)]\right| \geq \frac{\epsilon}{2}\right).
\end{aligned} \tag{5.4}$$

Step (a) uses the conditional distribution of b^0 given in (4.25), and step (b) follows from Lemma A.2. Label the two terms on the RHS of (5.4) as T_1 and T_2, respectively. We now show that each term is bounded by $Ke^{-\kappa n\epsilon^2}$. Since $|\psi_b|$ is bounded (say it takes values in an interval of length B), the term T_2 can be bounded using Hoeffding's inequality (Lemma A.1) by $2e^{-n\epsilon^2/(2B^2)}$.

Next, consider T_1. Let Π_0 be the event under consideration, so that $T_1 = P(\Pi_0)$, and define an event $\mathcal{F}$ as follows.

$$\mathcal{F} := \left\{\left|\frac{1}{\sqrt{n}}\left\|q^0\right\| - \sigma_0\right| \geq \epsilon_0\right\}, \tag{5.5}$$

where $\epsilon_0 > 0$ will be specified later. With this definition, we have

$$T_1 = P(\Pi_0) \leq P(\mathcal{F}) + P(\Pi_0|\mathcal{F}^c) \leq Ke^{-\kappa n\epsilon_0^2} + P(\Pi_0|\mathcal{F}^c). \tag{5.6}$$

The final inequality in (5.6) follows from the concentration of $\left\|q^0\right\|$ in (4.3). To bound the last term $P(\Pi_0|\mathcal{F}^c)$, we write it as

$$\begin{aligned}
P(\Pi_0|\mathcal{F}^c) = \mathbb{E}\left[\mathsf{I}\{\Pi_0\}|\mathcal{F}^c\right] &= \mathbb{E}\left[\mathbb{E}\left[\mathsf{I}\{\Pi_0\}|\mathcal{F}^c, \mathscr{S}_{0,0}\right] \mid \mathcal{F}^c\right] \\
&= \mathbb{E}\left[P\left(\Pi_0|\mathcal{F}^c, \mathscr{S}_{0,0}\right) \mid \mathcal{F}^c\right],
\end{aligned} \tag{5.7}$$

where $\mathsf{I}\{\cdot\}$ denotes the indicator function, and $P\left(\Pi_0|\mathcal{F}^c, \mathscr{S}_{0,0}\right)$ equals

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\left(\psi_b\left(\frac{1}{\sqrt{n}}\left\|q^0\right\| Z'_{0_i}, w_i\right) - \psi_b(\sigma_0 Z'_{0_i}, w_i)\right)\right| \geq \frac{\epsilon}{2} \,\middle|\, \mathcal{F}^c, \mathscr{S}_{0,0}\right). \tag{5.8}$$

To obtain (5.8), we use the fact that $\sigma_0 Z'_{0_i} + [\Delta_{0,0}]_i = \frac{1}{\sqrt{n}}\left\|q^0\right\| Z'_{0_i}$ which follows from the definition of $\Delta_{0,0}$ in Lemma 4.3. Recall from Section 4.4 that $\mathscr{S}_{0,0}$ is the sigma algebra generated by $\{w, \beta_0, q^0\}$; so in (5.8), only Z'_0 is random — all other terms are in $\mathscr{S}_{0,0}$. We now derive a bound for the upper tail of the probability in (5.8); the lower tail bound is similarly obtained.

Define the shorthand $\mathsf{diff}(Z'_{0_i}) := \psi_b(\frac{1}{\sqrt{n}}\left\|q^0\right\| Z'_{0_i}, w_i) - \psi_b(\sigma_0 Z'_{0_i}, w_i)$. Since ψ_b is bounded, so is $\mathsf{diff}(Z'_{0_i})$. Let $|\psi_b| \leq B/2$, so that $\left|\mathsf{diff}(Z'_{0_i})\right| \leq B$ for all i. Then the upper tail of the probability in (5.8) can be written as

$$P\left(\frac{1}{n}\sum_{i=1}^{n}\mathsf{diff}(Z'_{0_i}) - \mathbb{E}[\mathsf{diff}(Z'_{0_i})] \geq \frac{\epsilon}{2} - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\mathsf{diff}(Z'_{0_i})] \,\middle|\, \mathcal{F}^c, \mathscr{S}_{0,0}\right). \tag{5.9}$$

We now show that $\left|\mathbb{E}[\mathsf{diff}(Z'_{0_i})]\right| \le \frac{1}{4}\epsilon$ for all $i \in [n]$. From here on, we suppress the conditioning on $\mathcal{F}^c, \mathscr{S}_{0,0}$ for brevity. Denoting the standard normal density by ϕ, we have

$$\left|\mathbb{E}[\mathsf{diff}(Z'_{0_i})]\right| \le \int_{\mathbb{R}} \phi(z)\ |\mathsf{diff}(z)|\, dz \overset{(a)}{\le} \int_{\mathbb{R}} \phi(z)\, C \left| z\left(\frac{\left\|q^0\right\|}{\sqrt{n}} - \sigma_0\right)\right| dz \overset{(b)}{\le} 2C\epsilon_0.$$

The above is bounded by $\frac{1}{4}\epsilon$ if we choose $\epsilon_0 \le \epsilon/8C$. In the chain above, (a) follows by Fact 4 for a suitable constant $C > 0$ as ψ_b is bounded and assumed to be differentiable. Step (b) follows since $\left|\frac{1}{\sqrt{n}}\left\|q^0\right\| - \sigma_0\right| \le \epsilon_0$ under $\mathcal{F}^c$.

The probability in (5.9) can then be bounded using Hoeffding's inequality (Lemma A.1):

$$P\left(\frac{1}{n}\sum_{i=1}^{n} \mathsf{diff}(Z'_{0_i}) - \mathbb{E}[\mathsf{diff}(Z'_{0_i})] \ge \frac{\epsilon}{4} \,\middle|\, \mathcal{F}^c, \mathscr{S}_{0,0}\right) \le e^{-n\epsilon^2/(8B^2)}.$$

Substituting in (5.8) and using a similar bound for the lower tail, we have shown via (5.7) that $P(\Pi_0 \mid \mathcal{F}^c) \le 2e^{-n\epsilon^2/(8B^2)}$. Using this in (5.6) with $\epsilon_0 \le \epsilon/8C$ proves that the first term in (5.4) is bounded by $Ke^{-n\kappa\epsilon^2}$.

(c) The function $\phi_b(b_i^0, w_i) := b_i^0 w_i \in PL(2)$ by Lemma C.1. By $\mathcal{B}_0(b)$.(iii),

$$P\left(\left|\frac{1}{n}(b^0)^* w - \mathbb{E}[\sigma_0 \breve{Z}_0 W]\right| \ge \epsilon\right) \le Ke^{-\kappa n\epsilon^2}.$$

This result follows since $\mathbb{E}[\sigma_0 \breve{Z}_0 W] = 0$ by the independence of W and $\hat{Z}_0$.

(d) The function $\phi_b(b_i^0, w_i) := (b_i^0)^2 \in PL(2)$ by Lemma C.1. By $\mathcal{B}_0(b)$.(iii),

$$P\left(\left|\frac{1}{n}\left\|b^0\right\|^2 - \mathbb{E}[(\sigma_0 \breve{Z}_0)^2]\right| \ge \epsilon\right) \le Ke^{-\kappa n\epsilon^2}.$$

This result follows since $\mathbb{E}[(\sigma_0 \hat{Z}_0)^2] = \sigma_0^2$.

(e) Since g_0 is Lipschitz, the function $\phi_b(b_i^0, w_i) := (g_0(b_i^0, w_i))^2 \in PL(2)$ by Lemma C.1. By $\mathcal{B}_0(b)$.(iii),

$$P\left(\left|\frac{1}{n}\left\|m^0\right\|^2 - \mathbb{E}[(g_0(\sigma_0 \breve{Z}_0, W))^2]\right| \ge \epsilon\right) \le Ke^{-\kappa n\epsilon^2}.$$

This result follows since $\mathbb{E}[(g_0(\sigma_0 \breve{Z}_0, W))^2] = \tau_0^2$ by (4.4).

(f) The concentration of ξ_0 around $\hat{\xi}_0$ follows from $\mathcal{B}_0(b)$.(iv) applied to the function $\psi_b(b_i^0, w_i) := g_0'(b_i^0, w_i)$. Next, the function $\phi_b(b_i^0, w_i) := b_i^0\, g_0(b_i^0, w_i) \in PL(2)$ by Lemma C.1. Then by $\mathcal{B}_0(b)$.(iii),

$$P\left(\left|\frac{1}{n}(b^0)^* m^0 - \mathbb{E}[\sigma_0 \breve{Z}_0 g_0(\sigma_0 \breve{Z}_0, W)]\right| \ge \epsilon\right) \le Ke^{-\kappa n\epsilon^2}.$$

This result follows since $\mathbb{E}[\sigma_0 \breve{Z}_0 g_0(\sigma_0 \breve{Z}_0, W)] = \sigma_0^2 \mathbb{E}[g_0'(\sigma_0 \breve{Z}_0, W)] = \hat{\xi}_0 \tilde{E}_{0,0}$ by Stein's Lemma given in Fact 2.

(g) Nothing to prove.

(h) The result is equivalent to $\mathcal{B}_0(e)$ since $\left\|m_\perp^0\right\| = \left\|m^0\right\|$ and $(\tau_0^\perp)^2 = \tau_0^2$.

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

We wish to show results (a)–(h) in (4.40), (4.42), (4.43), (4.46), (4.48), (4.50), (4.52), (4.54), (4.56), (4.58).

(a) From the definition of $\Delta_{1,0}$ in (4.27) of Lemma 4.3, we have

$$\Delta_{1,0} \stackrel{d}{=} Z_0 \left(\frac{\left\|m^0\right\|}{\sqrt{n}} - \tau_0^{\perp} \right) - \frac{\left\|m^0\right\| \tilde{q}^0 \bar{Z}_0}{\sqrt{n}} + q^0 \left(\frac{n}{\left\|q^0\right\|^2} \right) \left(\frac{(b^0)^* m^0}{n} - \frac{\xi_0 \left\|q^0\right\|^2}{n} \right). \tag{5.10}$$

where $\tilde{q}^0 = q^0 / \left\|q^0\right\|$, and $\bar{Z}_0 \in \mathbb{R}$ is a standard Gaussian random variable. The equality in (5.10) is obtained using Fact 1 to write $\mathsf{P}^{\parallel}_{q^0} Z_0 \stackrel{d}{=} \tilde{q}^0 \bar{Z}_0$. Then, from (5.10) we have

$$\begin{aligned} P\left(\frac{1}{N} \left\|\Delta_{1,0}\right\|^2 \geq \epsilon \right) \stackrel{(a)}{\leq} & P\left(\left| \frac{\left\|m^0\right\|}{\sqrt{n}} - \tau_0 \right| \frac{\left\|Z_0\right\|}{\sqrt{N}} \geq \sqrt{\frac{\epsilon}{9}} \right) + P\left(\frac{\left\|m^0\right\|}{\sqrt{n}} \cdot \frac{\left|\bar{Z}_0\right|}{\sqrt{N}} \geq \sqrt{\frac{\epsilon}{9}} \right) \\ & + P\left(\left| \frac{(b^0)^* m^0}{\sqrt{n}\left\|q^0\right\|} - \xi_0 \frac{\left\|q^0\right\|}{\sqrt{n}} \right| \geq \sqrt{\frac{\epsilon}{9\delta}} \right). \end{aligned} \tag{5.11}$$

Step (a) follows from Lemma C.3 applied to $\Delta_{1,0}$ in (5.10) and Lemma A.2. Label the terms on the RHS of (5.11) as $T_1 - T_3$. To complete the proof, we show that each term is bounded by $Ke^{-\kappa n\epsilon}$ for generic positive constants K, κ that do not depend on n, ϵ.

Indeed, $T_1 \leq Ke^{-\kappa n\epsilon}$ using Lemma A.3, Lemma A.4, result $\mathcal{B}_0(e)$, and Lemma B.2. Similarly, $T_2 \leq Ke^{-\kappa n\epsilon}$ using Lemma A.3, Lemma A.4, result $\mathcal{B}_0(e)$, and Lemma B.1. Finally,

$$\begin{aligned} T_3 &\stackrel{(a)}{\leq} P\left(\left| \frac{(b^0)^* m^0}{n} \cdot \frac{\sqrt{n}}{\left\|q^0\right\|} - \hat{\xi}_0 \sigma_0 \right| \geq \frac{1}{2}\sqrt{\frac{\epsilon}{9\delta}} \right) + P\left(\left| \xi_0 \frac{\left\|q^0\right\|}{\sqrt{n}} - \hat{\xi}_0 \sigma_0 \right| \geq \frac{1}{2}\sqrt{\frac{\epsilon}{9\delta}} \right) \\ &\stackrel{(b)}{\leq} 2K \exp\left\{ \frac{-\kappa n\epsilon}{4(9^2)\delta \max(1, \hat{\xi}_0^2 \sigma_0^4, \sigma_0^{-2})} \right\} + 2K \exp\left\{ \frac{-\kappa n\epsilon}{4(9^2)\delta \max(1, \hat{\xi}_0^2, \sigma_0^2)} \right\}. \end{aligned}$$

Step (a) follows from Lemma A.2, and step (b) from Lemma A.3, $\mathcal{B}_0(f)$, the concentration of $\left\|q^0\right\|$ given in (4.3), and Lemma A.6.

(b)(i) The proof of (4.42) is similar to analogous $\mathcal{B}_0(b)$(iii) result (4.44).

(b)(ii) First,

$$\begin{aligned} & P\left(\left| \frac{1}{N} \sum_{i=1}^{N} \psi_h(h_i^1, \beta_{0_i}) - \mathbb{E}[\psi_h(\tau_0 \tilde{Z}_0, \beta)] \right| \geq \epsilon \right) \\ & \stackrel{(a)}{=} P\left(\left| \frac{1}{N} \sum_{i=1}^{N} \psi_h(\tau_0 Z_{0_i} + [\Delta_{1,0}]_i, \beta_{0_i}) - \mathbb{E}[\psi_h(\tau_0 \tilde{Z}_0, \beta)] \right| \geq \epsilon \right) \\ & \stackrel{(b)}{\leq} P\left(\left| \frac{1}{N} \sum_{i=1}^{N} (\psi_h(\tau_0 Z_{0_i} + [\Delta_{1,0}]_i, \beta_{0_i}) - \psi_h(\tau_0 Z_{0_i}, \beta_{0_i})) \right| \geq \frac{\epsilon}{2} \right) \\ & \quad + P\left(\left| \frac{1}{N} \sum_{i=1}^{N} \psi_h(\tau_0 Z_{0_i}, \beta_{0_i}) - \mathbb{E}[\psi_h(\tau_0 \tilde{Z}_0, \beta)] \right| \geq \frac{\epsilon}{2} \right). \end{aligned} \tag{5.12}$$

Step (a) follows from the conditional distribution of h^1 stated in (4.24) and step (b) from Lemma A.2. Label the two terms on the RHS as T_1 and T_2. Term T_2 is upper bounded by $Ke^{-\kappa n\epsilon^2}$ by Hoeffding's inequality (Lemma A.1). To complete the proof, we show that T_1 has the same bound.

Consider the first term in (5.12). From the definition of $\Delta_{1,0}$ in Lemma 4.3,

$$\tau_0 Z_{0_i} + [\Delta_{1,0}]_i = \frac{\left\|m^0\right\|}{\sqrt{n}}[(\mathsf{I} - \mathsf{P}^{\parallel}_{q^0})Z_0]_i + u_i, \text{ where } u_i := q^0_i \left(\frac{(b^0)^* m^0}{\left\|q^0\right\|^2} - \xi_0\right). \tag{5.13}$$

For $\epsilon_0 > 0$ to be specified later, define event $\mathcal{F}$ as

$$\mathcal{F} := \left\{\left|\frac{1}{\sqrt{n}}\left\|m^0\right\| - \tau_0\right| \geq \epsilon_0\right\} \cup \left\{\left|\frac{1}{n}(b^0)^* m^0 - \frac{1}{n}\xi_0 \left\|q^0\right\|^2\right| \geq \epsilon_0\right\}. \tag{5.14}$$

Denoting the event we are considering in T_1 by Π_1, so that $T_1 = P(\Pi_1)$, we write

$$T_1 = P(\Pi_1) \leq P(\mathcal{F}) + P(\Pi_1 \mid \mathcal{F}^c) \leq K e^{-\kappa n \epsilon_0^2} + P(\Pi_1 \mid \mathcal{F}^c), \tag{5.15}$$

where the last inequality is by $\mathcal{B}_0(e), \mathcal{B}_0(f)$ and the concentration assumption (4.3) on q^0. Writing $P(\Pi_1|\mathcal{F}^c) = \mathbb{E}[P(\Pi_1|\mathcal{F}^c, \mathscr{S}_{1,0}) \mid \mathcal{F}^c]$, we now bound $P(\Pi_1|\mathcal{F}^c, \mathscr{S}_{1,0})$. In what follows, we drop the explicit conditioning on $\mathcal{F}^c$ and $\mathscr{S}_{1,0}$ for brevity. Then $P(\Pi_1|\mathcal{F}^c, \mathscr{S}_{1,0})$ can be written as

$$\begin{aligned}
&P\left(\left|\frac{1}{N}\sum_{i=1}^N \left(\psi_h\left(\frac{\left\|m^0\right\|}{\sqrt{n}}[(\mathsf{I} - \mathsf{P}^{\parallel}_{q^0})Z_0]_i + u_i, \beta_{0_i}\right) - \psi_h(\tau_0 Z_{0_i}, \beta_{0_i})\right)\right| \geq \frac{\epsilon}{2}\right) \\
&\leq P\left(\left|\frac{1}{N}\sum_{i=1}^N \psi_h\left(\frac{\left\|m^0\right\|}{\sqrt{n}}[(\mathsf{I} - \mathsf{P}^{\parallel}_{q^0})Z_0]_i + u_i, \beta_{0_i}\right) - \psi_h\left(\frac{\left\|m^0\right\|}{\sqrt{n}} Z_{0_i} + u_i, \beta_{0_i}\right)\right| \geq \frac{\epsilon}{4}\right) \\
&\quad + P\left(\left|\frac{1}{N}\sum_{i=1}^N \psi_h\left(\frac{\left\|m^0\right\|}{\sqrt{n}} Z_{0_i} + u_i, \beta_{0_i}\right) - \psi_h(\tau_0 Z_{0_i}, \beta_{0_i})\right| \geq \frac{\epsilon}{4}\right).
\end{aligned} \tag{5.16}$$

The above uses Lemma A.2. Note that in (5.16), only Z_0 is random as the other terms are all in $\mathscr{S}_{1,0}$. Label the two terms on the RHS (5.16) as $T_{1,a}$ and $T_{1,b}$. To complete the proof we show that both are bounded by $Ke^{-\kappa n \epsilon^2}$.

First consider $T_{1,a}$.

$$\begin{aligned}
T_{1,a} &\overset{(a)}{\leq} P\left(\frac{C}{N}\sum_{i=1}^N \left|\frac{\left\|m^0\right\|}{\sqrt{n}}[\mathsf{P}^{\parallel}_{q^0} Z_0]_i\right| \geq \frac{\epsilon}{4}\right) \overset{(b)}{\leq} P\left(\frac{C}{N}\sum_{i=1}^N |\tau_0 + \epsilon_0| \left|[\mathsf{P}^{\parallel}_{q^0} Z_0]_i\right| \geq \frac{\epsilon}{4}\right) \\
&\overset{(c)}{\leq} P\left(\frac{C}{N}\sum_{i=1}^N \frac{|q^0_i|}{\|q^0\|}|Z| \geq \frac{\epsilon}{4|\tau_0 + \epsilon_0|}\right) \overset{(d)}{\leq} P\left(\frac{|Z|}{\sqrt{N}} \geq \frac{\epsilon}{4C|\tau_0 + \epsilon_0|}\right) \overset{(e)}{\leq} e^{-\kappa N \epsilon^2}.
\end{aligned}$$

Step (a) holds by Fact 4 for a suitable constant $C > 0$. Step (b) follows because we are conditioning on $\mathcal{F}^c$ defined in (5.14). Step (c) is obtained by writing out the expression for the vector $\mathsf{P}^{\parallel}_{q^0} Z_0$:

$$\mathsf{P}^{\parallel}_{q^0} Z_0 = \frac{q^0}{\|q^0\|}\sum_{j=1}^N \frac{q^0_j}{\|q^0\|} Z_{0_j} \overset{d}{=} \frac{q^0}{\|q^0\|} Z,$$

where $Z \in \mathbb{R}$ is standard Gaussian (Fact 1). Step (d) follows from Cauchy-Schwarz and step (e) by Lemma B.1.

Considering $T_{1,b}$, the second term of (5.16), and noting that all quantities except Z_0 are in $\mathscr{S}_{1,0}$, define the shorthand $\mathsf{diff}(Z_{0_i}) := \psi_h\left(\frac{1}{\sqrt{n}}\left\|m^0\right\| Z_{0_i} + u_i, \beta_{0_i}\right) - \psi_h(\tau_0 Z_{0_i}, \beta_{0,i})$. Then the upper tail of $T_{1,b}$ can be written as

$$P\left(\frac{1}{N}\sum_{i=1}^N (\mathsf{diff}(Z_{0_i}) - \mathbb{E}[\mathsf{diff}(Z_{0_i})]) \geq \frac{\epsilon}{4} - \frac{1}{N}\sum_{i=1}^N \mathbb{E}[\mathsf{diff}(Z_{0_i})] \,\middle|\, \mathcal{F}^c, \mathscr{S}_{1,0}\right). \tag{5.17}$$

Since ψ_h is bounded, so is $\mathsf{diff}(Z_{0_i})$. Using the conditioning on $\mathcal{F}^c$ and steps similar to those in $\mathcal{B}_0(b)$(iv), we can show that $\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\mathsf{diff}(Z_{0_i})] \leq \frac{1}{8}\epsilon$ for $\epsilon_0 \leq C\tau_0\epsilon$, where $C > 0$ can be explicitly computed. For such ϵ_0, using Hoeffding's inequality the probability in (5.17) can be bounded by $e^{-n\epsilon^2/(128B^2)}$ when ψ_h takes values within an interval of length B. A similar bound holds for the lower tail of $T_{1,b}$. Thus we have now bounded both terms of (5.16) by $Ke^{-n\kappa\epsilon^2}$. The result follows by substituting the value of ϵ_0 (chosen as described above) in (5.15).

(c),(d),(e),(f) These results can be proved by appealing to $\mathcal{H}_1(b)$ in a manner similar to $\mathcal{B}_0(c)(d)(e)(f)$.

(g) From the definitions in Section 4.1 and defining $\mathbf{Q}_1 := \frac{1}{n}\left\|q^0\right\|^2$, we have $\gamma_0^1 = \mathbf{Q}_1^{-1}\frac{1}{n}(q^0)^* q^1$ and $\hat{\gamma}_0^1 = \tilde{E}_{0,1}/\tilde{E}_{0,0} = \tilde{E}_{0,1}\sigma_0^{-2}$. Therefore,

$$P\left(\left|\gamma_0^1 - \hat{\gamma}_0^1\right| \geq \epsilon\right) \overset{(a)}{\leq} P\left(\left|\mathbf{Q}_1^{-1} - \sigma_0^{-2}\right| \geq \tilde{\epsilon}\right) + P\left(\left|\frac{1}{n}(q^0)^* q^1 - \tilde{E}_{0,1}\right| \geq \tilde{\epsilon}\right) \tag{5.18}$$

where (a) follows from Lemma A.3 with $\tilde{\epsilon} := \min\{\sqrt{\epsilon/3},\ \epsilon/(3\tilde{E}_{0,1}),\ \epsilon\sigma_0^2/3\}$. We now show that each of the two terms in (5.18) is bounded by $Ke^{-\kappa n\tilde{\epsilon}^2}$. Since $\sigma_0^2 > 0$, by Lemma A.6 and (4.3), we have $P\left(\left|\mathbf{Q}_1^{-1} - \sigma_0^{-2}\right| \geq \tilde{\epsilon}\right) \leq 2Ke^{-\kappa n\tilde{\epsilon}^2\sigma_0^2\min(1,\sigma_0^2)}$. The concentration bound for $\frac{1}{n}(q^0)^* q^1$ follows from $\mathcal{H}_1(e)$.

(h) From the definitions in Section 4.1, we have $\left\|q_\perp^1\right\|^2 = \left\|q^1\right\|^2 - \left\|q_\parallel^1\right\|^2 = \left\|q^1\right\|^2 - (\gamma_0^1)^2\left\|q^0\right\|^2$, and $(\sigma_1^\perp)^2 = \sigma_1^2 - (\hat{\gamma}_0^1)^2\sigma_0^2$. We therefore have

$$\begin{aligned}
&P\left(\left|\frac{1}{n}\left\|q_\perp^1\right\|^2 - (\sigma_1^\perp)^2\right| \geq \epsilon\right) \\
&\overset{(a)}{\leq} P\left(\left|\frac{1}{n}\left\|q^1\right\|^2 - \sigma_1^2\right| \geq \frac{\epsilon}{2}\right) + P\left(\left|(\gamma_0^1)^2\frac{1}{n}\left\|q^0\right\|^2 - (\hat{\gamma}_0^1)^2\sigma_0^2\right| \geq \frac{\epsilon}{2}\right) \\
&\overset{(b)}{\leq} K\exp\left\{-\kappa n\epsilon^2\right\} + K\exp\left\{\frac{-\kappa n\epsilon^2}{4(9)\max(1,(\hat{\gamma}_0^1)^4,\sigma_0^4)}\right\}
\end{aligned}$$

In the chain above, (a) uses Lemma A.2 and (b) is obtained using $\mathcal{H}_1(e)$ for bounding the first term and by applying Lemma A.3 to the second term along with the concentration of $\left\|q^0\right\|$ in (4.3), $\mathcal{H}_1(g)$, and Lemma A.5 (for concentration of the square).

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

We prove the statements in $\mathcal{B}_t$ assuming that $\mathcal{B}_0, \ldots, \mathcal{B}_{t-1}$, and $\mathcal{H}_1, \ldots, \mathcal{H}_t$ hold due to the induction hypothesis. The induction hypothesis implies that for $0 \leq r \leq (t-1)$, the deviation probabilities $P(\frac{1}{n}\|\Delta_{r,r}\|^2 \geq \epsilon)$ in (4.41) and $P(\frac{1}{n}\|\Delta_{r+1,r}\|^2 \geq \epsilon)$ in (4.40) are each bounded by $K_r e^{-\kappa_r n\epsilon}$. Similarly, the LHS in each of (4.42) – (4.59) is bounded by $K_r e^{-\kappa_r n\epsilon^2}$.

We begin with a lemma that is required to prove $\mathcal{B}_t(a)$. The lemma as well as other parts of $\mathcal{B}_t$ assume the invertibility of $\mathbf{M}_1, \ldots, \mathbf{M}_t$, but for the sake of brevity, we do not explicitly specify the conditioning.

Lemma 5.1. *Let* $v := \frac{1}{n}H_t^* q_\perp^t - \frac{1}{n}M_t^*\left[\lambda_t m^{t-1} - \sum_{i=1}^{t-1}\lambda_i\gamma_i^t m^{i-1}\right]$ *and* $\mathbf{M}_t := \frac{1}{n}M_t^* M_t$. *If* $\mathbf{M}_1, \ldots, \mathbf{M}_t$ *are invertible, we have for* $j \in [t]$,

$$P\left(\left|[\mathbf{M}_t^{-1}v]_j\right| \geq \epsilon\right) \leq Kt^2K_{t-1}\exp\{-n\kappa\kappa_{t-1}\epsilon^2/t^2\}.$$

Proof. We can represent $\mathbf{M}_t$ as

$$\mathbf{M}_t = \frac{1}{n}\begin{pmatrix} n\mathbf{M}_{t-1} & M_{t-1}^* m^{t-1} \\ (M_{t-1}^* m^{t-1})^* & \left\|m^{t-1}\right\|^2 \end{pmatrix},$$

Then, if $\mathbf{M}_{t-1}$ is invertible, by the block inversion formula we have

$$\mathbf{M}_t^{-1} = \begin{pmatrix} \mathbf{M}_{t-1}^{-1} + n\left\|m_\perp^{t-1}\right\|^{-2}\alpha^{t-1}(\alpha^{t-1})^* & -n\left\|m_\perp^{t-1}\right\|^{-2}\alpha^{t-1} \\ -n\left\|m_\perp^{t-1}\right\|^{-2}(\alpha^{t-1})^* & n\left\|m_\perp^{t-1}\right\|^{-2} \end{pmatrix}, \tag{5.19}$$

where we have used $\alpha^{t-1} = \frac{1}{n}\mathbf{M}_{t-1}^{-1}M_{t-1}^* m^{t-1}$ and $(M_{t-1}^* m^{t-1})^*\alpha^{t-1} = (m^{t-1})^* m_\parallel^{t-1}$. Therefore,

$$\mathbf{M}_t^{-1} v = \begin{bmatrix} \mathbf{M}_{t-1}^{-1} v_{[t-1]} + \alpha^{t-1}\left((\alpha^{t-1})^* v_{[t-1]} - v_t\right)\mathsf{a}_{t-1} \\ -\left((\alpha^{t-1})^* v_{[t-1]} - v_t\right)\mathsf{a}_{t-1} \end{bmatrix}, \tag{5.20}$$

where $\mathsf{a}_r := n/\left\|m_\perp^r\right\|^2$ for $r \in [t]$, and $v_{[r]} \in \mathbb{R}^r$ denotes the vector consisting of the first r elements of $v \in \mathbb{R}^t$. Now, using the block inverse formula again to express $\mathbf{M}_{t-1}^{-1}v_{[t-1]}$ and noting that $\alpha^{t-1} = (\alpha_0^{t-1}, \ldots, \alpha_{t-2}^{t-1})$, we obtain

$$\mathbf{M}_t^{-1} v = \begin{bmatrix} \mathbf{M}_{t-2}^{-1} v_{[t-2]} + \alpha^{t-2}\left((\alpha^{t-2})^* v_{[t-2]} - v_{t-1}\right)\mathsf{a}_{t-2} + \alpha_{[t-2]}^{t-1}\left((\alpha^{t-1})^* v_{[t-1]} - v_t\right)\mathsf{a}_{t-1} \\ -\left((\alpha^{t-2})^* v_{[t-2]} - v_{t-1}\right)\mathsf{a}_{t-2} + \alpha_{t-2}^{t-1}\left((\alpha^{t-1})^* v_{[t-1]} - v_t\right)\mathsf{a}_{t-1} \\ -\left((\alpha^{t-1})^* v_{[t-1]} - v_t\right)\mathsf{a}_{t-1} \end{bmatrix}.$$

Continuing in this fashion, we can express each element of $\mathbf{M}_t^{-1}v$ as follows:

$$\left[\mathbf{M}_t^{-1} v\right]_k = \begin{cases} v_1\mathsf{a}_0 + \sum_{j=1}^{t-1}\alpha_0^j\left((\alpha^j)^* v_{[j]} - v_{j+1}\right)\mathsf{a}_j & k = 1, \\ -\left((\alpha^{k-1})^* v_{[k-1]} - v_k\right)\mathsf{a}_{k-1} + \sum_{j=k}^{t-1}\alpha_{k-1}^j\left((\alpha^j)^* v_{[j]} - v_{j+1}\right)\mathsf{a}_j & 2 \le k < t, \\ -\left((\alpha^{t-1})^* v_{[t-1]} - v_t\right)\mathsf{a}_{t-1} & k = t. \end{cases} \tag{5.21}$$

We will prove that each entry of $\mathbf{M}_t^{-1}v$ concentrates around 0 by showing that each entry of v concentrates around zero, and the entries of α^j, a_j concentrate around constants for $j \in [t]$.

For $k \in [t]$, bound $|v_k|$ as follows. Substituting $q_\perp^t = q^t - \sum_{j=0}^{t-1}\gamma_j^t q^j$ in the definition of v and using the triangle inequality, we have

$$|v_k| \le \left|\frac{(h^k)^* q^t}{n} - \lambda_t\frac{(m^{k-1})^* m^{t-1}}{n}\right| + \left|\gamma_0^t\right|\left|\frac{(h^k)^* q^0}{n}\right| + \sum_{i=1}^{t-1}\left|\gamma_i^t\right|\left|\frac{(h^k)^* q^i}{n} - \lambda_i\frac{(m^{k-1})^* m^{i-1}}{n}\right|. \tag{5.22}$$

Therefore,

$$\begin{aligned} P\left(|v_k| \ge \epsilon\right) \le{}& P\left(\left|\frac{1}{n}(h^k)^* q^t - \lambda_t\frac{1}{n}(m^{k-1})^* m^{t-1}\right| \ge \epsilon'\right) + P\left(\left|\gamma_0^t\right|\left|\frac{1}{n}(h^k)^* q^0\right| \ge \epsilon'\right) \\ &+ \sum_{i=1}^{t-1} P\left(\left|\gamma_i^t\right|\left|\frac{1}{n}(h^k)^* q^i - \lambda_i\frac{1}{n}(m^{k-1})^* m^{i-1}\right| \ge \epsilon'\right) \end{aligned} \tag{5.23}$$

where $\epsilon' = \frac{\epsilon}{t+1}$. The first term in (5.23) can be bounded using Lemma A.3 and induction hypotheses $\mathcal{H}_t(f)$ and $\breve{\mathcal{B}}_{t-1}(e)$ as follows.

$$\begin{aligned} &P\left(\left|\frac{(h^k)^* q^t}{n} - \lambda_t\frac{(m^{k-1})^* m^{t-1}}{n}\right| \ge \epsilon'\right) \\ &\le P\left(\left|\frac{(h^k)^* q^t}{n} - \hat{\lambda}_t\breve{E}_{k-1,t-1}\right| \ge \frac{\epsilon'}{2}\right) + P\left(\left|\lambda_t\frac{(m^{k-1})^* m^{t-1}}{n} - \hat{\lambda}_t\breve{E}_{k-1,t-1}\right| \ge \frac{\epsilon'}{2}\right) \\ &\le K_{t-1}\exp\left\{-\kappa\kappa_{t-1}n\epsilon'^2\right\} + 2K_{t-1}\exp\left\{-\frac{\kappa\kappa_{t-1}n\epsilon'^2}{9\max(1, \hat{\lambda}_t^2, \breve{E}_{k-1,t-1}^2)}\right\}. \end{aligned}$$

For $k \in [t]$, the second term in (5.23) can be bounded as

$$\begin{aligned}
&P\left(\left|\gamma_0^t\right|\left|\frac{1}{n}(h^k)^* q^0\right| \geq \epsilon'\right) \leq P\left(\left(\left|\gamma_0^t - \hat{\gamma}_0^t\right| + \left|\hat{\gamma}_0^t\right|\right)\left|\frac{1}{n}(h^k)^* q^0\right| \geq \epsilon'\right) \\
&\leq P\left(\left|\gamma_0^t - \hat{\gamma}_0^t\right| \geq \sqrt{\epsilon'}\right) + P\left(\left|\frac{1}{n}(h^k)^* q^0\right| \geq \frac{\epsilon'}{2}\min\left\{1, \left|\hat{\gamma}_0^t\right|^{-1}\right\}\right) \\
&\leq K_{t-1}\exp\{-\kappa\kappa_{t-1} n\epsilon'\} + K_{t-1}\exp\{-\kappa\kappa_{t-1} n\epsilon'^2\},
\end{aligned}$$

where the last inequality follows from induction hypotheses $\mathcal{H}_t(g)$ and $\mathcal{H}_t(c)$. Similarly, for $k \in [t]$, $i \in [t-1]$, the third term in (5.23) can be bounded as

$$\begin{aligned}
&P\left(\left|\gamma_i^t\right|\left|\frac{(h^k)^* q^i}{n} - \lambda_i \frac{(m^{k-1})^* m^{i-1}}{n}\right| \geq \epsilon'\right) \\
&\leq P\left(\left(\left|\gamma_i^t - \hat{\gamma}_i^t\right| + \left|\hat{\gamma}_i^t\right|\right)\left|\frac{(h^k)^* q^i}{n} - \lambda_i \frac{(m^{k-1})^* m^{i-1}}{n}\right| \geq \epsilon'\right) \\
&\leq P\left(\left|\gamma_i^t - \hat{\gamma}_i^t\right| \geq \sqrt{\epsilon'}\right) + P\left(\left|\frac{(h^k)^* q^i}{n} - \lambda_i \frac{(m^{k-1})^* m^{i-1}}{n}\right| \geq \frac{\epsilon'}{2}\min\left\{1, \frac{1}{\hat{\gamma}_i^t}\right\}\right) \\
&\leq K_{t-1}\exp\left\{-\kappa\kappa_{t-1} n\epsilon'\right\} + 2K_{t-1}\exp\left\{-\kappa\kappa_{t-1} n\epsilon'^2\right\}.
\end{aligned}$$

Substituting $\epsilon' = \frac{\epsilon}{t+1}$ in each of the above bounds and using them in (5.23),

$$P\left(|v_k| \geq \epsilon\right) \leq KtK_{t-1}\exp\left\{-\kappa\kappa_{t-1}\epsilon^2/t^2\right\}. \tag{5.24}$$

Furthermore, from induction hypotheses $\mathcal{B}_0(g) - \mathcal{B}_{t-1}(g)$, for $0 \leq i < j \leq (t-1)$:

$$P\left(\left|\alpha_i^j - \hat{\alpha}_i^j\right| \geq \epsilon\right) \leq K_{t-1}\exp\left\{-n\kappa_{t-1}\epsilon^2\right\}. \tag{5.25}$$

Also, using induction hypotheses $\mathcal{B}_0(h) - \mathcal{B}_{t-1}(h)$ and Lemma A.6, for $0 \leq r \leq (t-1)$:

$$P\left(\left|\mathsf{a}_r - (\tau_t^{\perp})^{-2}\right| \geq \epsilon\right) \leq K_{t-1}\exp\left\{-n\kappa_{t-1}\epsilon^2\right\}. \tag{5.26}$$

Finally, from (5.21), we have for $k \in [t]$,

$$\begin{aligned}
P\left(\left|\left[\mathbf{M}_t^{-1} v\right]_k\right| \geq \epsilon\right) &\overset{(a)}{\leq} P\left(\cup_{k\in[t]}\left\{|v_k| \geq \epsilon\right\} \cup_{0\leq r<t}\left\{\left|\mathsf{a}_r - (\tau_t^{\perp})^{-2}\right| \geq \kappa_1\epsilon/t\right\}\right. \\
&\qquad \left. \cup_{0\leq i<j<t}\left\{\left|\alpha_i^j - \hat{\alpha}_i^j\right| \geq \kappa_2\epsilon/t\right\}\right) \\
&\overset{(b)}{\leq} Kt^2 K_{t-1}\exp\{-n\kappa\kappa_{t-1}\epsilon^2/t^2\}.
\end{aligned}$$

where in step (a), κ_1, κ_2 are appropriately chosen positive constants, and step (b) follows from the bounds in (5.24), (5.25), and (5.26). □

(a) Recall the definition of $\Delta_{t,t}$ from (4.28). Then using Fact 1, it follows $\frac{1}{\sqrt{n}}\left\|q_\perp^t\right\| \mathsf{P}_{M_t}^{\parallel} Z_t' \overset{d}{=} \frac{1}{n}\left\|q_\perp^t\right\| \tilde{M}_t \bar{Z}_t'$, where the columns of $\tilde{M}_t \in \mathbb{R}^{n\times t}$ form an orthogonal basis for the column space of M_t with $\tilde{M}_t^* \tilde{M}_t = n\mathsf{I}_t$, and $\bar{Z}_t' \in \mathbb{R}^t$ is an independent random vector with i.i.d. $\mathcal{N}(0,1)$ entries. Then,

$$\Delta_{t,t} = \sum_{r=0}^{t-1}(\gamma_r^t - \hat{\gamma}_r^t) b^r + Z_t'\left(\frac{1}{\sqrt{n}}\left\|q_\perp^t\right\| - \sigma_t^{\perp}\right) - \frac{1}{n}\left\|q_\perp^t\right\| \tilde{M}_t \bar{Z}_t' + M_t \mathbf{M}_t^{-1} v,$$

where $\mathbf{M}_t \in \mathbb{R}^{t\times t}$ and $v \in \mathbb{R}^t$ are defined in Lemma 5.1. Writing $M_t\mathbf{M}_t^{-1}v = \sum_{j=0}^{t-1} m^j \left[\mathbf{M}_t^{-1}v\right]_{j+1}$ and using Lemma C.3, we have

$$\|\Delta_{t,t}\|^2 \le 2(t+1)\left[\sum_{r=0}^{t-1}(\gamma_r^t - \hat{\gamma}_r^t)^2\|b^r\|^2 + \left\|Z_t'\right\|^2\left(\frac{1}{\sqrt{n}}\left\|q_\perp^t\right\| - \sigma_t^\perp\right)^2 + \frac{1}{n^2}\left\|q_\perp^t\right\|^2\left\|\tilde{M}_t\bar{Z}_t'\right\|^2 + \sum_{j=0}^{t-1}\left\|m^j\right\|^2\left[\mathbf{M}_t^{-1}v\right]_{j+1}^2\right],$$

Applying Lemma A.2,

$$P\left(\frac{\|\Delta_{t,t}\|^2}{n} \ge \epsilon\right) \le \sum_{r=0}^{t-1} P\left(\left|\gamma_r^t - \hat{\gamma}_r^t\right|\frac{\|b^r\|}{\sqrt{n}} \ge \sqrt{\tilde{\epsilon}_t}\right) + P\left(\frac{\left\|q_\perp^t\right\|}{\sqrt{n}}\frac{\left\|\tilde{M}_t\bar{Z}_t'\right\|}{n} \ge \sqrt{\tilde{\epsilon}_t}\right) + P\left(\left|\frac{\left\|q_\perp^t\right\|}{\sqrt{n}} - \sigma_t^\perp\right|\frac{\|Z_t'\|}{\sqrt{n}} \ge \sqrt{\tilde{\epsilon}_t}\right) + \sum_{j=0}^{t-1} P\left(\left|\left[\mathbf{M}_t^{-1}v\right]_{j+1}\right|\frac{\left\|m^j\right\|}{\sqrt{n}} \ge \sqrt{\tilde{\epsilon}_t}\right), \tag{5.27}$$

where $\tilde{\epsilon}_t := \frac{\epsilon}{4(t+1)^2}$. We now bound each of the terms in (5.27).

For $0 \le r \le t-1$, the first term is bounded as

$$\begin{aligned} &P\left(\left|\gamma_r^t - \hat{\gamma}_r^t\right|\frac{1}{\sqrt{n}}\|b^r\| \ge \sqrt{\tilde{\epsilon}_t}\right) \le P\left(\left|\gamma_r^t - \hat{\gamma}_r^t\right|\left(\left|\frac{1}{\sqrt{n}}\|b^r\| - \sigma_r\right| + \sigma_r\right) \ge \sqrt{\tilde{\epsilon}_t}\right) \\ &\le P\left(\left|\gamma_r^t - \hat{\gamma}_r^t\right| \ge \frac{1}{2}\sqrt{\tilde{\epsilon}_t}\min\{1, \sigma_r^{-1}\}\right) + P\left(\left|\frac{1}{\sqrt{n}}\|b^r\| - \sigma_r\right| \ge \sqrt{\epsilon}\right) \\ &\overset{(a)}{\le} K_{t-1}\exp\left\{-\kappa\kappa_{t-1}n\tilde{\epsilon}_t\right\} + K_{t-1}\exp\left\{-\kappa\kappa_{t-1}n\epsilon\right\}, \end{aligned}$$

where step (a) follows from induction hypotheses $\mathcal{H}_t(g)$, $\mathcal{B}_0(d) - \mathcal{B}_{t-1}(d)$, and Lemma A.4. Next, the third term in (5.27) is bounded as

$$\begin{aligned} &P\left(\left|\frac{\left\|q_\perp^t\right\|}{\sqrt{n}} - \sigma_t^\perp\right|\frac{\|Z_t'\|}{\sqrt{n}} \ge \sqrt{\tilde{\epsilon}_t}\right) \le P\left(\left|\frac{\left\|q_\perp^t\right\|}{\sqrt{n}} - \sigma_t^\perp\right| \ge \frac{\sqrt{\tilde{\epsilon}_t}}{\sqrt{2}}\right) + P\left(\frac{\|Z_t'\|}{\sqrt{n}} \ge \sqrt{2}\right) \\ &\overset{(b)}{\le} K_{t-1}\exp\left\{-\kappa\kappa_{t-1}n\tilde{\epsilon}_t\right\} + \exp\left\{-n/8\right\}, \end{aligned}$$

where step (b) is obtained using induction hypothesis $\mathcal{H}_t(h)$, Lemma A.4, and Lemma B.2. Since $\frac{1}{\sqrt{n}}\left\|q_\perp^t\right\|$ concentrates on $\sigma_t^\perp$ by $\mathcal{H}_t(h)$, the second term in (5.27) can be bounded as

$$\begin{aligned} &P\Big(\frac{1}{\sqrt{n}}\left\|q_\perp^t\right\| \cdot \frac{1}{n}\left\|\tilde{M}_t\bar{Z}_t'\right\| \ge \sqrt{\tilde{\epsilon}_t}\Big) \\ &\le P\left(\left|\frac{1}{\sqrt{n}}\left\|q_\perp^t\right\| - \sigma_t^\perp\right| \ge \sqrt{\epsilon}\right) + P\left(\frac{1}{n}\left\|\tilde{M}_t\bar{Z}_t'\right\| \ge \frac{1}{2}\sqrt{\tilde{\epsilon}_t}\min\{1, (\sigma_t^\perp)^{-1}\}\right) \\ &\le K_{t-1}\exp\left\{-\kappa\kappa_{t-1}n\tilde{\epsilon}_t\right\} + tKK_{t-1}\exp\{-\kappa\kappa_{t-1}n\tilde{\epsilon}_t/t\}, \end{aligned} \tag{5.28}$$

where the last inequality is obtained as follows. The concentration for $\left\|q_\perp^t\right\|/\sqrt{n}$ has already been shown above. For the second term, denoting the columns of $\tilde{M}_t$ by $\{\tilde{m}_0, \ldots, \tilde{m}_{t-1}\}$, we have

$\|\tilde{M}_t\bar{Z}'_t\|^2 = \sum_{i=0}^{t-1}\|\tilde{m}_i\|^2(\bar{Z}'_{t_i})^2 = n\sum_{i=0}^{t-1}(\bar{Z}'_{t_i})^2$ since the $\{\tilde{m}_i\}$ are orthogonal, and $\|\tilde{m}_i\|^2 = n$ for $0 \le i \le t-1$. Therefore,

$$P\Big(\frac{1}{n^2}\Big\|\tilde{M}_t\bar{Z}'_t\Big\|^2 \ge \tilde{\epsilon}_t\Big) = P\Big(\sum_{i=0}^{t-1}(\bar{Z}'_{t_i})^2 \ge n\tilde{\epsilon}_t\Big) \overset{(c)}{\le} \sum_{i=0}^{t-1} P\Big(|\bar{Z}'_{t_i}| \ge \sqrt{\frac{n\tilde{\epsilon}_t}{t}}\Big) \overset{(d)}{\le} 2te^{-\frac{n\tilde{\epsilon}_t}{2t}}.$$

Step (c) is obtained from Lemma A.2, and step (d) from Lemma B.1. This yields the second term in (5.28)

Finally, for $0 \le j \le (t-1)$, the last term in (5.27) can be bounded by

$$\begin{aligned}
&P\Big(|[\mathbf{M}_t^{-1}v]_{j+1}|\frac{\|m^j\|}{\sqrt{n}} \ge \sqrt{\tilde{\epsilon}_t}\Big) = P\left(|[\mathbf{M}_t^{-1}v]_{j+1}|\Big(\Big|\frac{\|m^j\|}{\sqrt{n}} - \tau_j\Big| + \tau_j\Big) \ge \sqrt{\tilde{\epsilon}_t}\right)\\
&\le P\left(\Big|\frac{\|m^j\|}{\sqrt{n}} - \tau_j\Big| \ge \sqrt{\epsilon}\right) + P\left(|[\mathbf{M}_t^{-1}v]_{j+1}| \ge \frac{1}{2}\sqrt{\tilde{\epsilon}_t}\min\{1, \tau_j^{-1}\}\right)\\
&\overset{(e)}{\le} K_{t-1}\exp\left\{-\kappa\kappa_{t-1}n\epsilon\right\} + Kt^2K_{t-1}\exp\left\{-\kappa\kappa_{t-1}n\tilde{\epsilon}_t/t^2\right\},
\end{aligned}$$

where step (e) follows from induction hypothesis $\mathcal{B}_{t-1}(e)$, and Lemma 5.1. Substituting $\tilde{\epsilon}_t = \frac{\epsilon}{4(t+1)^2}$, we have bounded each term of (5.27) as desired.

(b).(iii) For brevity, define $\mathbb{E}\phi_b := \mathbb{E}[\phi_b(\sigma_0\breve{Z}_0, ..., \sigma_t\breve{Z}_t, W)]$, and

$$a_i = (b_i^0, ..., b_i^t, w_i), \qquad c_i = (b_{\text{pure}_i}^0, ..., b_{\text{pure}_i}^t, w_i) \tag{5.29}$$

Using Lemma A.2, we have

$$\begin{aligned}
&P\Big(\Big|\frac{1}{n}\sum_{i=1}^n \phi_b\left(b_i^0, ..., b_i^t, w_i\right) - \mathbb{E}\phi_b\Big| \ge \epsilon\Big)\\
&\le P\Big(\Big|\frac{1}{n}\sum_{i=1}^n \phi_b(c_i) - \mathbb{E}\phi_b\Big| \ge \frac{\epsilon}{2}\Big) + P\Big(\Big|\frac{1}{n}\sum_{i=1}^n(\phi_b(a_i) - \phi_b(c_i))\Big| \ge \frac{\epsilon}{2}\Big).
\end{aligned} \tag{5.30}$$

Lemma 4.4 (Eq. (4.32)) shows the joint distribution of $(b_{\text{pure}_i}^0, ..., b_{\text{pure}_i}^t)$ is jointly Gaussian for $i \in [N]$. The first term in (5.30) can therefore be bounded as

$$\begin{aligned}
P\Big(\Big|\frac{1}{n}\sum_{i=1}^n \phi_b(c_i) - \mathbb{E}\phi_b\Big| \ge \frac{\epsilon}{2}\Big) &= P\Big(\Big|\frac{1}{n}\sum_{i=1}^n \phi_b(\sigma_0\breve{Z}_{0,i}, \ldots, \sigma_t\breve{Z}_{t,i}, w_i) - \mathbb{E}\phi_b\Big| \ge \frac{\epsilon}{2}\Big)\\
&\le 2\exp\left(\frac{-\kappa n\epsilon^2}{t^3}\right),
\end{aligned} \tag{5.31}$$

where the last inequality is obtained from Lemma B.4. Here $\kappa > 0$ is a generic absolute constant.

We now bound the second term in (5.30) using the pseudo-Lipschitz property of ϕ_b. Denoting the pseudo-Lipschitz constant by L, we have

$$\begin{aligned}
&\Big|\frac{1}{n}\sum_{i=1}^n(\phi_b(a_i) - \phi_b(c_i))\Big|^2 \le \left[\frac{1}{n}\sum_{i=1}^n|\phi_b(a_i) - \phi_b(c_i)|\right]^2 \le \left[\frac{L}{n}\sum_{i=1}^n(1 + 2\|c_i\| + \|a_i - c_i\|)\|a_i - c_i\|\right]^2\\
&\le 3L^2\left[1 + \frac{4}{n}\sum_{i=1}^n\|c_i\|^2 + \frac{1}{n}\sum_{i=1}^n\|a_i - c_i\|^2\right]\left[\frac{1}{n}\sum_{j=1}^n\|a_j - c_j\|^2\right],
\end{aligned} \tag{5.32}$$

where the last inequality is obtained by first applying Cauchy-Schwarz, and then using Lemma C.3.

For $j \in [N]$, note that $\mathbb{E}\,\|c_j\|^2 = \sigma_1^2 + \ldots + \sigma_t^2 + \sigma^2$. Now using (5.32) we bound the second term in (5.30) as follows.

$$
\begin{aligned}
&P\Big(\Big|\frac{1}{n}\sum_{i=1}^{n}(\phi_b(a_i) - \phi_b(c_i))\Big| \geq \frac{\epsilon}{2}\Big) = P\Big(\Big|\frac{1}{n}\sum_{i=1}^{n}(\phi_b(a_i) - \phi_b(c_i))\Big|^2 \geq \frac{\epsilon^2}{4}\Big) \\
&\leq P\Big(\Big[1 + \frac{4}{n}\sum_{i=1}^{n}\|c_i\|^2 + \frac{1}{n}\sum_{i=1}^{n}\|a_i - c_i\|^2\Big]\Big[\frac{1}{n}\sum_{j=1}^{n}\|a_j - c_j\|^2\Big] \geq \frac{\epsilon^2}{12L^2}\Big) \\
&\leq P\Big(\frac{1}{n}\sum_{i=1}^{n}\|a_i - c_i\|^2 \geq \frac{\epsilon^2 \min\{1, \frac{1}{12L^2}\}}{2 + 8(\sigma_1^2 + \ldots + \sigma_t^2 + \sigma^2)}\Big) \;+\; P\Big(\frac{1}{n}\sum_{j=1}^{n}\|c_j\|^2 \geq 2(\sigma_1^2 + \ldots + \sigma_t^2 + \sigma^2)\Big).
\end{aligned}
\tag{5.33}
$$

Label the two terms above as T_1 and T_2. We bound T_2 as

$$
P\Big(\frac{1}{n}\sum_{j=1}^{n}\|c_j\|^2 \geq 2(\sigma^2 + \sum_{r=1}^{t}\sigma_r^2)\Big) = P\Big(\frac{1}{n}\sum_{j=1}^{n}\big(\,\|c_j\|^2 - \mathbb{E}\|c_j\|^2\big) \;\geq (\sigma^2 + \sum_{r=1}^{t}\sigma_r^2)\Big) \leq e^{-\kappa n/t^3},
\tag{5.34}
$$

for an absolute constant $\kappa > 0$, where the last inequality is obtained by applying the concentration result in Lemma B.4 to the pseudo-Lipschitz function $\phi_b(c_j) = \|c_j\|^2$.

$$
\begin{aligned}
&\sum_{i=1}^{n}\|a_i - c_i\|^2 = \sum_{i=1}^{n}\sum_{k=0}^{t}(b^k_{pure_i} - b^k_i)^2 = \sum_{i=1}^{n}\sum_{k=0}^{t}\left(\sum_{r=0}^{k}\mathsf{c}_r^k\,[\Delta_{r,r}]_i\right)^2 \\
&\leq \sum_{i=1}^{n}\sum_{k=0}^{t}\left(\sum_{r'=0}^{k}(\mathsf{c}_{r'}^k)^2\sum_{r=0}^{k}([\Delta_{r,r}]_i)^2\right) = \sum_{k=0}^{t}\left(\sum_{r'=0}^{k}(\mathsf{c}_{r'}^k)^2\sum_{r=0}^{k}\|\Delta_{r,r}\|^2\right) = \sum_{r=0}^{t}\|\Delta_{r,r}\|^2\sum_{k=r}^{t}\sum_{r'=0}^{k}(\mathsf{c}_{r'}^k)^2,
\end{aligned}
\tag{5.35}
$$

where the inequality is obtained by applying Cauchy-Schwarz.

Comparing (4.32) and (4.33) in Lemma 4.4, we observe that for $k \geq 0$ and $j \in [n]$,

$$
\mathbb{E}\big(b^k_{\mathrm{pure}_j}\big)^2 = \sigma_k^2 = \sum_{i=0}^{t}(\sigma_i^{\perp})^2(\mathsf{c}_i^k)^2.
\tag{5.36}
$$

Therefore,

$$
\sum_{i=0}^{k}(\mathsf{c}_i^k)^2 \leq \frac{\sigma_t^2}{\min_{0\leq i\leq k}(\sigma_i^{\perp})^2} \leq \frac{\sigma_k^2}{\varepsilon_2},
\tag{5.37}
$$

where the last inequality follows from the stopping criterion in (2.5). Using (5.37) and (5.35) we have

$$
\frac{1}{n}\sum_{i=1}^{n}\|a_i - c_i\|^2 \leq \frac{1}{n}\sum_{r=0}^{t}\|\Delta_{r,r}\|^2\sum_{k=r}^{t}\frac{\sigma_k^2}{\varepsilon_2}.
$$

Therefore we can bound the first term T_1 in (5.33) as follows.

$$T_1 = P\Big(\frac{1}{n}\sum_{r=0}^{t}\|\Delta_{r,r}\|^2 \geq \frac{\varepsilon_2}{(\sigma_1^2+\ldots+\sigma_t^2)}\frac{\epsilon^2\min\{1,\frac{1}{12L^2}\}}{(2+8(\sigma_1^2+\ldots+\sigma_t^2+\sigma^2))}\Big)$$
$$\leq \sum_{r=0}^{t} P\left(\frac{1}{n}\|\Delta_{r,r}\|^2 \leq \frac{\kappa\epsilon^2}{t^3}\right) \overset{(a)}{\leq} Kt^3K_{t-1}\exp\left\{-\frac{\kappa\kappa_{t-1}n\epsilon^2}{t^7}\right\}. \tag{5.38}$$

where $K,\kappa>0$ are some absolute constants. The inequality (a) follows from steps $\mathcal{B}_0(a)-\mathcal{B}_t(a)$.

Finally, substituting (5.38) and (5.34) in (5.33), and then combining with (5.31) and (5.30), we obtain

$$P\Big(\Big|\frac{1}{n}\sum_{i=1}^{n}\phi_b\left(b_i^0,...,b_i^t,w_i\right)-\mathbb{E}\phi_b\Big|\geq\epsilon\Big)\leq Kt^3K_{t-1}\exp\left\{-\frac{\kappa\kappa_{t-1}n\epsilon^2}{t^7}\right\}. \tag{5.39}$$

(b).(iv) For brevity, we write $\mathsf{b}_{t,i}:=\sum_{r=0}^{t-1}\hat{\gamma}_r^t b_i^r$. Then using the conditional distribution of b^t in (4.25) and Lemma A.2, we write

$$\begin{aligned}
&P\Big(\Big|\frac{1}{n}\sum_{i=1}^{n}\psi_b(b_i^t,w_i)-\mathbb{E}[\psi_b(\sigma_t\breve{Z}_t,W)]\Big|\geq\epsilon\Big)\\
&=P\Big(\Big|\frac{1}{n}\sum_{i=1}^{n}\psi_b(\mathsf{b}_{t,i}+\sigma_t^{\perp}Z'_{t_i}+[\Delta_{t,t}]_i,w_i)-\mathbb{E}[\psi_b(\sigma_t\breve{Z}_t,W)]\Big|\geq\epsilon\Big)\\
&\leq P\Big(\Big|\frac{1}{n}\sum_{i=1}^{n}\Big(\psi_b(\mathsf{b}_{t,i}+\sigma_t^{\perp}Z'_{t_i}+[\Delta_{t,t}]_i,w_i)-\psi_b(\mathsf{b}_{t,i}+\sigma_t^{\perp}Z'_{t_i},w_i)\Big)\Big|\geq\frac{\epsilon}{3}\Big) \qquad (5.40)\\
&+P\Big(\Big|\frac{1}{n}\sum_{i=1}^{n}\psi_b(\mathsf{b}_{t,i}+\sigma_t^{\perp}Z'_{t_i},w_i)-\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{Z'_t}[\psi_b(\mathsf{b}_{t,i}+\sigma_t^{\perp}Z'_{t_i},w_i)]\Big|\geq\frac{\epsilon}{3}\Big)\\
&+P\Big(\Big|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{Z'_t}[\psi_b(\mathsf{b}_{t,i}+\sigma_t^{\perp}Z'_{t_i},w_i)]-\mathbb{E}[\psi_b(\sigma_t\breve{Z}_t,W)]\Big|\geq\frac{\epsilon}{3}\Big).
\end{aligned}$$

Label the terms of (5.40) as T_1-T_3. First consider T_2. Since ψ_b is bounded, Hoeffding's inequality yields $T_2\leq 2e^{-\kappa n\epsilon^2}$.

To bound T_3, first note that the $\mathbb{R}^2\to\mathbb{R}$ function $\mathbb{E}_Z[\psi_b(x+Z,y)],\ Z\sim\mathcal{N}(0,1)$, is bounded and differentiable in the first argument (due to the smoothness of the Gaussian density). Hence, using induction hypotheses $\mathcal{B}_0(b).(\text{iv})-\mathcal{B}_{t-1}(b).(\text{iv})$, the probability of each of the following events is bounded by $K_{t-1}\exp\left\{-\kappa_{t-1}n\epsilon^2/t^2\right\}$:

$$\begin{aligned}
&\Big|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\,\psi_b(\sum_{r=0}^{t-1}\hat{\gamma}_r^t b_i^r+\sigma_t^{\perp}Z'_{t_i},w_i)-\mathbb{E}\,\psi_b(\sum_{r=0}^{t-2}\hat{\gamma}_r^t b_i^r+\hat{\gamma}_{t-1}^t\sigma_{t-1}\breve{Z}_{t-1}+\sigma_t^{\perp}Z'_{t_i},W)\Big|\geq\frac{\epsilon}{t},\\
&\Big|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\,\psi_b(\sum_{r=0}^{t-2}\hat{\gamma}_r^t b_i^r+\hat{\gamma}_{t-1}^t\sigma_{t-1}\breve{Z}_{t-1}+\sigma_t^{\perp}Z'_{t_i},W)-\mathbb{E}\,\psi_b(\sum_{r=0}^{t-3}\hat{\gamma}_r^t b_i^r+\sum_{r'=t-2}^{t-1}\hat{\gamma}_{r'}^t\sigma_{r'}\breve{Z}_{r'}+\sigma_t^{\perp}Z'_{t_i},W)\Big|\geq\frac{\epsilon}{t},\\
&\quad\vdots\\
&\Big|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\,\psi_b(\hat{\gamma}_0^t b_i^0+\sum_{r'=1}^{t-1}\hat{\gamma}_{t-1}^t\sigma_{t-1}\breve{Z}_{t-1}+\sigma_t^{\perp}Z'_{t_i},W)\}-\mathbb{E}\,\psi_b(\sum_{r'=0}^{t-1}\hat{\gamma}_{r'}^t\sigma_{r'}\breve{Z}_{r'}+\sigma_t^{\perp}Z'_{t_i},W)\Big|\geq\frac{\epsilon}{t}.
\end{aligned} \tag{5.41}$$

In the above, the expectation in each term is over the random variables denoted in upper case. Recall from the proof of Lemma 4.4 above that $\sum_{r'=1}^{t-1} \hat{\gamma}_{t-1}^{t} \sigma_{t-1} \breve{Z}_{t-1} + \sigma_t^{\perp} Z'_{t_i} \stackrel{d}{=} \sigma_t \breve{Z}_t$. Thus T_3, the third term in (5.40), can be bounded by the probability of the union of the events in (5.41), which is no larger than $tK_{t-1} \exp\left\{-\kappa_{t-1} n\epsilon^2/t^2\right\}$.

Finally, consider T_1, the first term of (5.40). From the definition of $\Delta_{t,t}$ in Lemma 4.3, we have $\mathsf{b}_{t,i} + \sigma_t^{\perp} Z'_{t_i} + [\Delta_{t,t}]_i = \mathsf{b}_{t,i} + \frac{1}{n} \left\|q_{\perp}^t\right\| [(\mathsf{I} - \mathsf{P}_{M_t}^{\parallel}) Z'_t]_i + u_i$, where $u = (u_1, \ldots, u_n)$ is defined $u := \sum_{r=0}^{t-1} (\gamma_r^t - \hat{\gamma}_r^t) b^r + \sum_{j=0}^{t-1} m^j [\mathbf{M}_t^{-1} v]_{j+1}$, with v and $\mathbf{M}_t$ defined as in Lemma 5.1. For $\epsilon_0 > 0$ to be specified later, define the event $\mathcal{F}$ as

$$\mathcal{F} := \left\{ \left| \frac{1}{\sqrt{n}} \left\|q_{\perp}^t\right\| - \sigma_t^{\perp} \right| \geq \epsilon_0 \right\} \cup \left\{ \frac{1}{n} \|u\|^2 \geq \epsilon_0 \right\} \cup_{r=0}^{t-1} \left\{ \left| \frac{1}{\sqrt{n}} \|b^r\| - \sigma_r \right| \geq \epsilon_0 \right\}. \tag{5.42}$$

Denoting the event we are considering in T_1 by Π_t and following steps analogous to (5.15)–(5.16) in $\mathcal{H}_1(b)$.(ii), we obtain

$$\begin{aligned} P(T_1) &\leq P(\mathcal{F}) + \mathbb{E}[P(\Pi_t \mid \mathcal{F}^c, \mathscr{S}_{t,t}) \mid \mathcal{F}^c] \\ &\leq K t^2 K_{t-1} \exp\left\{-\kappa \kappa_{t-1} n \epsilon_0^2 / t^4\right\} + \mathbb{E}[P(\Pi_t \mid \mathcal{F}^c, \mathscr{S}_{t,t}) \mid \mathcal{F}^c], \end{aligned} \tag{5.43}$$

where the bound on $P(\mathcal{F})$ is obtained by the induction hypotheses $\mathcal{H}_t(h)$, $\mathcal{B}_0(d) - \mathcal{B}_{t-1}(d)$, Lemma A.4, and steps similar to the proof of $\mathcal{B}_t(a)$ for the concentration of $\|u\|^2/n$ (cf. (5.27)).

For the second term in (5.43), we have

$$\begin{aligned} &P(\Pi_t | \mathcal{F}^c, \mathscr{S}_{t,t}) = \\ &P\left(\left| \frac{1}{n} \sum_{i=1}^{n} \left(\psi_b(\mathsf{b}_{t,i} + \frac{\|q_{\perp}^t\|}{\sqrt{n}} [(\mathsf{I} - \mathsf{P}_{M_t}^{\parallel}) Z'_t]_i + u_i, w_i) - \psi_b(\mathsf{b}_{t,i} + \sigma_t^{\perp} Z'_{t_i}, w_i) \right) \right| \geq \epsilon \right) \\ &\leq P\left(\left| \frac{1}{n} \sum_{i=1}^{n} \left(\psi_b(\mathsf{b}_{t,i} + \frac{\|q_{\perp}^t\|}{\sqrt{n}} Z'_{t_i} + u_i, w_i) - \psi_b(\mathsf{b}_{t,i} + \sigma_t^{\perp} Z'_{t_i}, w_i) \right) \right| \geq \frac{\epsilon}{2} \right) \\ &+ \left(\left| \frac{1}{n} \sum_{i=1}^{n} \left(\psi_b(\mathsf{b}_{t,i} + \frac{\|q_{\perp}^t\|}{\sqrt{n}} [(\mathsf{I} - \mathsf{P}_{M_t}^{\parallel}) Z'_t]_i + u_i, w_i) - \psi_b(\mathsf{b}_{t,i} + \frac{\|q_{\perp}^t\|}{\sqrt{n}} Z'_{t_i} + u_i, w_i) \right) \right| \geq \frac{\epsilon}{2} \right), \end{aligned} \tag{5.44}$$

where we have omitted the conditioning to shorten notation. Label the two terms in (5.44) as $T_{1,a}$ and $T_{1,b}$. To complete the proof we show that both terms are bounded by $Ke^{-\kappa n \epsilon^2/t}$.

First consider $T_{1,b}$. We note that

$$\mathsf{P}_{M_t}^{\parallel} Z'_t = \sum_{r=0}^{t-1} \frac{\tilde{m}^r}{\sqrt{n}} \left[\frac{(\tilde{m}^r)^* Z'_t}{\sqrt{n}} \right] \stackrel{d}{=} \sum_{r=0}^{t-1} \frac{\tilde{m}^r}{\sqrt{n}} U_r, \tag{5.45}$$

where $\tilde{m}^r$, $0 \leq r \leq t-1$, are columns of $\tilde{M}_t$, which form an orthogonal basis for M_t with $\tilde{M}_t^* \tilde{M}_t = n\mathsf{I}_t$, and $U_1, \ldots, U_t$ are i.i.d. $\sim \mathcal{N}(0,1)$. Then,

$$\begin{aligned} T_{1,b} &\stackrel{(a)}{\leq} P\left(\frac{C}{n} \sum_{i=1}^{n} \left| \frac{\left\|q_{\perp}^t\right\|}{\sqrt{n}} [\mathsf{P}_{M_t}^{\parallel} Z'_t]_i \right| \geq \frac{\epsilon}{2} \right) \stackrel{(b)}{\leq} P\left(\frac{C}{n} \sum_{i=1}^{n} \left| (\sigma_t^{\perp} + \epsilon_0) [\mathsf{P}_{M_t}^{\parallel} Z'_t]_i \right| \geq \frac{\epsilon}{2} \right) \\ &= P\left(\left| \frac{C}{n} \sum_{i=1}^{n} \sum_{r=0}^{t-1} \frac{\tilde{m}_i^r U_r}{\sqrt{n}} \right| \geq \frac{\epsilon}{2\left|\sigma_t^{\perp} + \epsilon_0\right|} \right) \stackrel{(c)}{=} P\left(\left| \frac{C}{n} \sum_{i=1}^{n} \left(\sum_{r=0}^{t-1} (\tilde{m}_i^r)^2 \right)^{1/2} \frac{Z}{\sqrt{n}} \right| \geq \frac{\epsilon}{2\left|\sigma_t^{\perp} + \epsilon_0\right|} \right) \\ &\stackrel{(d)}{\leq} P\left(\sqrt{\frac{t}{n}} |Z| \geq \frac{\epsilon}{2C\left|\sigma_t^{\perp} + \epsilon_0\right|} \right) \leq 2e^{-\kappa n \epsilon^2/t}. \end{aligned} \tag{5.46}$$

In the above, (a) follows from Fact 4 for a suitable constant $C > 0$. Step (b) holds since we are conditioning on event $\mathcal{F}^c$ defined in (5.42). In step (c), $Z \sim \mathcal{N}(0,1)$ since $\sum_r \tilde{m}_i^r U_r$ is a zero-mean Gaussian with variance $\sum_r (\tilde{m}_i^r)^2$. Step (d) uses the Cauchy-Schwarz inequality and the fact that $\|\tilde{m}^r\| = \sqrt{n}$ for $0 \le r < t$.

Finally $T_{1,a}$, the first term in (5.44), can be bounded using Hoeffding's inequality. Noting that all quantities except Z'_t are in $\mathscr{S}_{t,t}$, define the shorthand $\mathsf{diff}(Z'_{t_i}) := \psi_b(\sum_{r=0}^{t-1} \hat{\gamma}_r^t b_i^r + \frac{1}{\sqrt{n}} \left\| q_\perp^t \right\| Z'_{t_i} + u_i, w_i) - \psi_b(\sum_{r=0}^{t-1} \hat{\gamma}_r^t b_i^r + \sigma_t^\perp Z'_{t_i}, w_i)$. Then the upper tail of $T_{1,a}$ can be written as

$$P\left(\frac{1}{n} \sum_{i=1}^n \mathsf{diff}(Z'_{t_i}) - \mathbb{E}[\mathsf{diff}(Z'_{t_i})] \ge \frac{\epsilon}{2} - \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\mathsf{diff}(Z'_{t_i})] \,\Big|\, \mathcal{F}^c, \mathscr{S}_{t,t} \right). \tag{5.47}$$

Using the conditioning on $\mathcal{F}^c$ and steps similar to those in $\mathcal{B}_0(b)$.(iv), we can show that $\frac{1}{n}\sum_i \mathbb{E}[\mathsf{diff}(Z'_{t_i})]] \le \frac{1}{4}\epsilon$ for $\epsilon_0 \le C(\sigma_t^\perp)\epsilon$, where the constant $C > 0$ can be explicitly computed. For such ϵ_0, using Hoeffding's inequality the probability in (5.47) can be bounded by $e^{-n\epsilon^2/(32B^2)}$, where B is the upper bound on $|\mathsf{diff}(\cdot)|$. A similar bound holds for the lower tail of $T_{1,a}$. Thus both terms of (5.44) are bounded by $Ke^{-\kappa n \epsilon^2/t}$.

The proof is completed by collecting the above bounds for each of the terms in (5.40), and observing that the overall bound is dominated by $P(T_1)$ in (5.43). Hence the final bound is of the form $Kt^2 K_{t-1} \exp\left\{-\kappa\kappa_{t-1} n\epsilon^2/t^4\right\}$.

(c) The function $\phi_b(b_i^t, w_i) := b_i^t w_i \in PL(2)$ by Lemma C.1. Then by $\mathcal{B}_t(b)$.(iii), $\frac{1}{n}(b^t)^* w \doteq \sigma_t \mathbb{E}[\breve{Z}_t W] = 0$.

(d) The function $\phi_b(b_i^r, b_i^t, w_i) := b_i^r b_i^t \in PL(2)$ by Lemma C.1. The result then follows from $\mathcal{B}_t(b)$.(iii).

(e) The function $\phi_b(b_i^r, b_i^t, w_i) := g_r(b_i^r, w_i) g_t(b_i^t, w_i) \in PL(2)$ since g_t is Lipschitz continuous (by Lemma C.1). Then by $\mathcal{B}_t(b)$.(iii),

$$\frac{1}{n}(m^r)^* m^t \doteq \mathbb{E}[g_r(\sigma_r \breve{Z}_r, W) g_t(\sigma_t \breve{Z}_t, W)] = \breve{E}_{r,t}.$$

where the last equality is due to the definition in (4.15).

(f) The concentration of ξ_t around $\hat{\xi}_t$ follows from $\mathcal{B}_t(b)$.(iv) applied to the function $\psi_b(b_i^t, w_i) := g_t'(b_i^t, w_i)$. Next, for $r \le t$, $\phi_b(b_i^0, \ldots, b_i^t, w_i) := b_i^r g_t(b_i^t, w_i) = b_i^r m_i \in PL(2)$, by Lemma C.1. Thus by $\mathcal{B}_t(b)$.(iii),

$$\begin{aligned} \frac{1}{n}(b^r)^* m^t \doteq \mathbb{E}[\sigma_r \breve{Z}_r \, g_t(\sigma_t \breve{Z}_t, W)] &\overset{(a)}{=} \sigma_r \sigma_t \mathbb{E}[\breve{Z}_r \breve{Z}_t] \mathbb{E}[g_t'(\sigma_t \breve{Z}_t, W)] \\ &= \tilde{E}_{r,t} \mathbb{E}[g_t'(\sigma_t \breve{Z}_t, W)] = \tilde{E}_{r,t} \hat{\xi}_t, \end{aligned}$$

where (a) holds due to Stein's lemma (Fact 2).

(g) For $1 \le r, s \le t$, note that $[\mathbf{M}_t]_{r,s} = \frac{1}{n}(m^{r-1})^* m^{s-1}$. Hence by $\mathcal{B}_{t-1}(e)$, $[\mathbf{M}_t]_{r,s}$ concentrates on $[\breve{C}^t]_{r,s} = \breve{E}_{r-1,s-1}$. We first show (4.55). By Fact 3, if $\frac{1}{n}\left\|m_\perp^r\right\|^2 \ge c > 0$ for all $0 \le r \le t-1$, then $\mathbf{M}_t$ is invertible. Note from $\mathcal{B}_{t-1}(h)$ that $\frac{1}{n}\left\|m_\perp^r\right\|^2$ concentrates on $(\tau_r^\perp)^2$, and $(\tau_r^\perp)^2 > \varepsilon_3$ by the stopping criterion assumption. Choosing $c = \frac{1}{2}\varepsilon_3$, we therefore have

$$\begin{aligned} P\left(\mathbf{M}_t \text{ singular}\right) &\le \sum_{r=0}^{t-1} P\left(\left| \frac{1}{n}\left\|m_\perp^r\right\|^2 - (\tau_r^\perp)^2 \right| \ge \frac{1}{2}\varepsilon_3 \right) \\ &\le \sum_{r=0}^{t-1} K_{r-1} e^{-\kappa_{r-1} n (\varepsilon_3)^2/4} \le t K_{t-1} e^{-\kappa\kappa_{t-1} n (\varepsilon_3)^2}. \end{aligned} \tag{5.48}$$

where the second inequality follows from $\mathcal{B}_0(h) - \mathcal{B}_{t-1}(h)$.

Next, we show (4.57). Recall the expression for $\mathbf{M}_t^{-1}$ from (5.19):

$$\mathbf{M}_t^{-1} = \begin{pmatrix} \mathbf{M}_{t-1}^{-1} + n\left\|m_\perp^{t-1}\right\|^{-2}\alpha^{t-1}(\alpha^{t-1})^* & -n\left\|m_\perp^{t-1}\right\|^{-2}\alpha^{t-1} \\ -n\left\|m_\perp^{t-1}\right\|^{-2}(\alpha^{t-1})^* & n\left\|m_\perp^{t-1}\right\|^{-2} \end{pmatrix}, \tag{5.49}$$

Block inversion can be similarly used to decompose $\breve{C}^t$ in terms of $\breve{C}^{t-1}$, which gives the concentrating values of the elements in (5.49).

Let $\mathcal{F}_r$ denote the event that $\mathbf{M}_r^{-1}$ is invertible, for $r \in [t]$. Then, for $i, j \in [t]$, we have

$$\begin{aligned} &P\left(|[\mathbf{M}_t^{-1}]_{i,j} - [\breve{C}_t^{-1}]_{i,j}| \geq \epsilon \mid \mathcal{F}_t\right) \\ &\leq P(\mathcal{F}_{t-1}^c) + P\left(|[\mathbf{M}_t^{-1}]_{i,j} - [\breve{C}_t^{-1}]_{i,j}| \geq \epsilon \mid \mathcal{F}_t, \mathcal{F}_{t-1}\right) \\ &\leq (t-1)K_{t-2}e^{-\kappa\kappa_{t-2}n} + P\left(|[\mathbf{M}_t^{-1}]_{i,j} - [\breve{C}_t^{-1}]_{i,j}| \geq \epsilon \mid \mathcal{F}_t, \mathcal{F}_{t-1}\right), \end{aligned} \tag{5.50}$$

where the final inequality follows from the inductive hypothesis $\mathcal{B}_{t-1}(g)$. Using the representation in (5.49), we bound the second term in (5.50) for $i, j \in [t]$. In what follows, we drop the conditioning on $\mathcal{F}_t, \mathcal{F}_{t-1}$ for brevity.

First, consider the entry at $i = j = t$. By $\mathcal{B}_{t-1}(h)$ and Lemma A.6,

$$P\left(\left|n\|m_{t-1}^\perp\|^{-2} - (\tau_{t-1}^\perp)^{-2}\right| \geq \epsilon\right) \leq K_{t-1}\exp\left\{-\kappa\kappa_{t-1}n\epsilon^2\right\}.$$

Next, consider the i^{th} element of $-n\left\|m_\perp^{t-1}\right\|^{-2}\alpha^{t-1}$. For $i \in [t-1]$,

$$P\left(\left|n\left\|m_\perp^{t-1}\right\|^{-2}\alpha_{i-1}^{t-1} - (\tau_{t-1}^\perp)^{-2}\hat{\alpha}_{i-1}^{t-1}\right| \geq \epsilon\right) \leq 2K_{t-1}e^{-\kappa\kappa_{t-1}n\epsilon^2}, \tag{5.51}$$

which follows from $\mathcal{B}_{t-1}(g)$, the concentration bound obtained above for $n\left\|m_\perp^{t-1}\right\|^{-2}$, and combining these via Lemma A.3.

Finally consider element (i, j) of $\mathbf{M}_{t-1}^{-1} + n\left\|m_\perp^{t-1}\right\|^{-2}\alpha^{t-1}(\alpha^{t-1})^*$ for $i, j \in [t-1]$. We have

$$\begin{aligned} &P\left(\left|[\mathbf{M}_{t-1}^{-1}]_{i,j} + n\left\|m_\perp^{t-1}\right\|^{-2}\alpha_{i-1}^{t-1}\alpha_{j-1}^{t-1} - [\breve{C}_t^{-1}]_{i,j} - (\tau_{t-1}^\perp)^{-2}\hat{\alpha}_{i-1}^{t-1}\hat{\alpha}_{j-1}^{t-1}\right| \geq \epsilon\right) \\ &\overset{(a)}{\leq} P\left(|[\mathbf{M}_{t-1}^{-1}]_{i,j} - [\breve{C}_t^{-1}]_{i,j}| \geq \frac{\epsilon}{2}\right) + P\left(|\alpha_{j-1}^{t-1} - \hat{\alpha}_{j-1}^{t-1}| \geq \frac{\epsilon'}{2}\right) \\ &\quad + P\left(|n\left\|m_\perp^{t-1}\right\|^{-2}\alpha_{i-1}^{t-1} - (\tau_{t-1}^\perp)^{-2}\hat{\alpha}_{i-1}^{t-1}| \geq \frac{\epsilon'}{2}\right) \\ &\overset{(b)}{\leq} K_{t-1}e^{-\frac{\kappa_{t-1}n\epsilon^2}{4}} + 2K_{t-1}e^{-\frac{\kappa\kappa_{t-1}n\epsilon'^2}{4}} + K_{t-1}e^{-\frac{\kappa_{t-1}n\epsilon'^2}{4}} \leq 4K_{t-1}e^{-\kappa\kappa_{t-1}n\epsilon^2}. \end{aligned}$$

Step (a) follows from Lemma A.2 and Lemma A.3 with $\epsilon' := \min\left(\sqrt{\frac{\epsilon}{3}}, \frac{\epsilon(\tau_{t-1}^\perp)^2}{3\hat{\alpha}_{i-1}^{t-1}}, \frac{\epsilon}{3\hat{\alpha}_{j-1}^{t-1}}\right)$. Step (b) follows from the inductive hypothesis, $\mathcal{H}_t(g)$, and (5.51).

Next, we prove the concentration of α^t around $\hat{\alpha}^t$. Recall from Section 4.1 that $\alpha^t = \frac{1}{n}\mathbf{M}_t^{-1}M_t^*m^t$ where $\mathbf{M}_t := \frac{1}{n}M_t^*M_t$. Thus for $1 \leq i \leq t$, $\alpha_{i-1}^t = \frac{1}{n}\sum_{j=1}^t \left[\mathbf{M}_t^{-1}\right]_{i,j}(m^{j-1})^*m^t$. Then from the

definition of $\hat{\alpha}^t$ in (4.17), for $1 \le i \le t$,

$$\begin{aligned}
P(|\alpha^t_{i-1} - \hat{\alpha}^t_{i-1}| \ge \epsilon) &= P\Big(\Big|\sum_{j=1}^{t}(\frac{1}{n}[\mathbf{M}_t^{-1}]_{i,j}(m^{j-1})^* m^t - [(\breve{C}^t)^{-1}]_{i,j}\breve{E}_{j-1,t})\Big| \ge \epsilon\Big) \\
&\overset{(a)}{\le} \sum_{j=1}^{t}\Big[P\Big(\Big|\frac{(m^{j-1})^* m^t}{n} - \breve{E}_{j-1,t}\Big| \ge \tilde{\epsilon}_j\Big) + P(|[\mathbf{M}_t^{-1}]_{i,j} - [(\breve{C}^t)^{-1}]_{i,j}| \ge \tilde{\epsilon}_j)\Big] \\
&\overset{(b)}{\le} Kt^4 K_{t-1}\exp\left\{-\kappa\kappa_{t-1} n\epsilon^2/t^9\right\} + 4tK_{t-1}\exp\left\{-\kappa\kappa_{t-1}t^{-2}n\epsilon^2\right\}.
\end{aligned}$$

Step (a) uses $\tilde{\epsilon}_j := \min\Big\{\sqrt{\frac{\epsilon}{3t}}, \frac{\epsilon}{3t\breve{E}_{j-1,t}}, \frac{\epsilon}{3t[(\breve{C}^t)^{-1}]_{k,j}}\Big\}$ and follows from Lemma A.2 and Lemma A.3. Step (b) uses $\mathcal{B}_t(e)$ and the work above.

(h) First, note that $\left\|m^t_\perp\right\|^2 = \|m^t\|^2 - \|m^t_\parallel\|^2 = \|m^t\|^2 - \|M_t\alpha^t\|^2$. Using the definition of $\tau_t^\perp$ in (4.19),

$$\begin{aligned}
&P\Big(\Big|\frac{1}{n}\left\|m^t_\perp\right\|^2 - (\tau_t^\perp)^2\Big| \ge \epsilon\Big) = P\Big(\Big|\frac{1}{n}\left\|m^t\right\|^2 - \frac{1}{n}\left\|M_t\alpha^t\right\|^2 - \tau_t^2 + (\hat{\alpha}^t)^*\breve{E}_t\Big| \ge \epsilon\Big) \\
&\le P\Big(\Big|\frac{1}{n}\left\|m^t\right\|^2 - \tau_t^2\Big| \ge \frac{\epsilon}{2}\Big) + P\Big(\Big|\frac{1}{n}\left\|M_t\alpha^t\right\|^2 - (\hat{\alpha}^t)^*\breve{E}_t\Big| \ge \frac{\epsilon}{2}\Big).
\end{aligned} \tag{5.52}$$

The bound for the first term in (5.52) follows by $\mathcal{B}_t(e)$. For the second term,

$$\left\|M_t\alpha^t\right\|^2 = n(\alpha^t)^*\mathbf{M}_t\alpha^t \overset{(a)}{=} (\alpha^t)^*\mathbf{M}_t\mathbf{M}_t^{-1}M_t^* m^t = (\alpha^t)^* M_t^* m^t = \sum_{i=0}^{t-1}\alpha_i^t (m^i)^* m^t,$$

where (a) holds because $\alpha^t = \mathbf{M}_t^{-1}M_t^* m^t/n$. Hence

$$\begin{aligned}
&P\Big(\Big|\frac{1}{n}\left\|M_t\alpha^t\right\|^2 - (\hat{\alpha}^t)^*\breve{E}_t\Big| \ge \frac{\epsilon}{2}\Big) = P\Big(\Big|\sum_{i=0}^{t-1}\Big(\frac{1}{n}\alpha_i^t(m^i)^* m^t - \hat{\alpha}_i^t\breve{E}_{i,t}\Big)\Big| \ge \frac{\epsilon}{2}\Big) \\
&\overset{(a)}{\le} \sum_{i=0}^{t-1} P(|\alpha_i^t - \hat{\alpha}_i^t| \ge \tilde{\epsilon}_i) + \sum_{i=0}^{t-1} P\Big(\Big|\frac{1}{n}(m^i)^* m^t - \breve{E}_{i,t}\Big| \ge \tilde{\epsilon}_i\Big) \\
&\overset{(b)}{\le} Kt^5 K_{t-1}\exp\left\{-\kappa\kappa_{t-1}n\epsilon^2/t^{11}\right\} + Kt^4 K_{t-1}\exp\left\{-\kappa\kappa_{t-1}n\epsilon^2/t^9\right\}.
\end{aligned}$$

Step (a) follows Lemma A.2 and Lemma A.3, using $\tilde{\epsilon}_i := \min\Big\{\sqrt{\frac{\epsilon}{6t}}, \frac{\epsilon}{6t\breve{E}_{i,t}}, \frac{\epsilon}{6t\hat{\alpha}_i^t}\Big\}$, and step (b) using $\mathcal{B}_t(e)$ and the proof of $\mathcal{B}_t(g)$ above.

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

The statements in $\mathcal{H}_{t+1}$ are proved assuming that $\mathcal{B}_t, \mathcal{H}_t$ hold due to the induction hypothesis.

(a) The proof of $\mathcal{H}_{t+1}(a)$ is similar to that of $\mathcal{B}_t(a)$, and uses the following lemma, which is analogous to Lemma 5.1.

Lemma 5.2. *Let* $v := \frac{1}{n}B_{t+1}^* m_t^\perp - \frac{1}{n}Q_{t+1}^*(\xi_t q^t - \sum_{i=0}^{t-1}\alpha_i^t\xi_i q^i)$ *and* $\mathbf{Q}_{t+1} := \frac{1}{n}Q_{t+1}^* Q_{t+1}$. *Then for* $j \in [t+1]$,

$$P(|[\mathbf{Q}_{t+1}^{-1}v]_j| \ge \epsilon) \le Kt^2 K'_{t-1}\exp\left\{-\kappa'_{t-1}n\epsilon^2/t^2\right\}.$$

(b)–**(h)** The proofs of the results in $\mathcal{H}_{t+1}(b) - \mathcal{H}_{t+1}(h)$ are along the same lines as $\mathcal{B}_t(b) - \mathcal{B}_t(h)$. By the end of step $\mathcal{H}_{t+1}(h)$, we will similarly pick up a t^5K term in the pre-factor in front of the exponent, and a κt^{-11} term in the exponent. It then follows that the K_t, κ_t are as given in (4.39).

A Concentration Lemmas

In the following, $\epsilon > 0$ is assumed to be a generic constant, with additional conditions specified whenever needed.

Lemma A.1 (Hoeffding's inequality)**.** *If $X_1, \ldots, X_n$ are bounded random variables such that $a_i \leq X_i \leq b_i$, then for $\nu = 2\left[\sum_i (b_i - a_i)^2\right]^{-1}$*

$$P\Big(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}X_i) \geq \epsilon\Big) \leq e^{-\nu n^2\epsilon^2},\ P\Big(\Big|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}X_i)\Big| \geq \epsilon\Big) \leq 2e^{-\nu n^2\epsilon^2}.$$

Lemma A.2 (Concentration of Sums)**.** *If random variables $X_1, \ldots, X_M$ satisfy $P(|X_i| \geq \epsilon) \leq e^{-n\kappa_i\epsilon^2}$ for $1 \leq i \leq M$, then*

$$P\Big(\Big|\sum_{i=1}^{M} X_i\Big| \geq \epsilon\Big) \leq \sum_{i=1}^{M} P\left(|X_i| \geq \frac{\epsilon}{M}\right) \leq Me^{-n(\min_i \kappa_i)\epsilon^2/M^2}.$$

Lemma A.3 (Concentration of Products)**.** *For random variables X, Y and non-zero constants c_X, c_Y, if*

$$P\left(|X - c_X| \geq \epsilon\right) \leq Ke^{-\kappa n\epsilon^2}, \quad \text{and} \quad P\left(|Y - c_Y| \geq \epsilon\right) \leq Ke^{-\kappa n\epsilon^2},$$

then the probability $P\left(|XY - c_Xc_Y| \geq \epsilon\right)$ is bounded by

$$\begin{aligned} &P\Big(\,|X - c_X| \geq \min\left(\sqrt{\frac{\epsilon}{3}}, \frac{\epsilon}{3c_Y}\right)\Big) + P\Big(\,|Y - c_Y| \geq \min\left(\sqrt{\frac{\epsilon}{3}}, \frac{\epsilon}{3c_X}\right)\Big) \\ &\leq 2K\exp\left\{-\frac{\kappa n\epsilon^2}{9\max(1, c_X^2, c_Y^2)}\right\}. \end{aligned}$$

Proof. The probability of interest, $P\left(|XY - c_Xc_Y| \geq \epsilon\right)$, equals

$$P\left(|(X - c_X)(Y - c_Y) + (X - c_X)c_Y + (Y - c_Y)c_X| \geq \epsilon\right).$$

The result follows by noting that if $|X - c_X| \leq \min(\sqrt{\frac{\epsilon}{3}}, \frac{\epsilon}{3c_Y})$ and $|Y - c_Y| \leq \min(\sqrt{\frac{\epsilon}{3}}, \frac{\epsilon}{3c_X})$, then the following terms are all bounded by $\frac{\epsilon}{3}$:

$$|(X - c_X)c_Y|, |(Y - c_X)c_Y|, \text{ and } |(X - c_X)(Y - c_Y)|. \qquad \square$$

Lemma A.4 (Concentration of Square Roots)**.** *Let $c \neq 0$.*

$$\textit{If } P\left(\left|X_n^2 - c^2\right| \geq \epsilon\right) \leq e^{-\kappa n\epsilon^2}, \textit{ then } P\left(||X_n| - |c|| \geq \epsilon\right) \leq e^{-\kappa n|c|^2\epsilon^2}.$$

Proof. If $\epsilon \leq c^2$, then the event $c^2 - \epsilon \leq X_n^2 \leq c^2 + \epsilon$ implies that $\sqrt{c^2 - \epsilon} \leq |X_n| \leq \sqrt{c^2 + \epsilon}$. On the other hand, if $\epsilon \geq c^2$, then $c^2 - \epsilon \leq X_n^2 \leq c^2 + \epsilon$ implies that $0 \leq |X_n| \leq \sqrt{c^2 + \epsilon}$. Therefore, $|X_n^2 - c^2| \leq \epsilon$ implies

$$||X_n| - |c|| \leq |c|\max(1 - \sqrt{(1 - (\epsilon/c^2))_+}, \sqrt{1 + (\epsilon/c^2)} - 1),$$

where $x_+ := \max\{x, 0\}$. Note, $(1 + x)^{1/2} \leq 1 + \frac{1}{2}x$ for $x \geq 0$, and $(1 - x)^{1/2} \geq 1 - x$ for $x \in (0, 1)$. Using these, we conclude that $|X_n^2 - c^2| \leq \epsilon$ implies

$$\begin{aligned} ||X_n| - |c|| &\leq |c|\max\left(1 - \sqrt{\left(1 - \frac{\epsilon}{c^2}\right)_+}, \sqrt{1 + \frac{\epsilon}{c^2}} - 1\right) \\ &\leq |c|\max\left(\frac{\epsilon}{c^2}, \frac{\epsilon}{2c^2}\right) = \frac{\epsilon}{|c|}. \end{aligned} \qquad \square$$

Lemma A.5 (Concentration of Powers)**.** *Assume $c \neq 0$ and $0 < \epsilon \leq 1$. Then for any integer $k \geq 2$,*

$$\text{if } P\left(|X_n - c| \geq \epsilon\right) \leq e^{-\kappa n \epsilon^2}, \text{ then } P\left(\left|X_n^k - c^k\right| \geq \epsilon\right) \leq e^{-\kappa n \epsilon^2/[(1+|c|)^k - |c|^k]^2}.$$

Proof. Without loss of generality, assume that $c > 0$. First consider the case where $\epsilon < c$. Then $c - \epsilon \leq X_n \leq c + \epsilon$ implies

$$(c-\epsilon)^k - c^k \leq X_n^k - c^k \leq (c+\epsilon)^k - c^k = \sum_{i=1}^{k} \binom{k}{i} c^{k-i} \epsilon^i.$$

Hence, $|X_n - c| \leq \epsilon$ implies $\left|X_n^k - c^k\right| \leq \epsilon c_0$, where

$$c_0 = \sum_{i=1}^{k} \binom{k}{i} c^{k-i} \epsilon^{i-1} < \sum_{i=1}^{k} \binom{k}{i} c^{k-i} = (1+c)^k - c^k.$$

Therefore,

$$P(|X_n^k - c^k| \geq \epsilon) \leq P\left(|X_n - c| \geq \epsilon/c_0\right) \leq e^{-\kappa n \epsilon^2/[(1+c)^k - c^k]^2}. \tag{A.1}$$

For the case where $0 < c < \epsilon < 1$, $X_n \in [c-\epsilon, c+\epsilon]$ implies $(c-\epsilon)^k - c^k \leq X^k - c^k \leq (c+\epsilon)^k - c^k$. Using $\epsilon < 1$, we note that the absolute values of

$$(c-\epsilon)^k - c^k = \sum_{i=1}^{k} \binom{k}{i} c^{k-i} (-\epsilon)^i, \text{ and } (c+\epsilon)^k - c^k = \sum_{i=1}^{k} \binom{k}{i} c^{k-i} \epsilon^i$$

are bounded by $c_1 := (1+c)^k - c^k$. Thus $|X_n - c| \leq \epsilon$ implies $\left|X_n^k - c^k\right| \leq \epsilon c_1$. Therefore the same bound as in (A.1) holds when $0 < c < \epsilon < 1$ (though a tighter bound could be obtained in this case). □

Lemma A.6 (Concentration of Scalar Inverses)**.** *Assume $c \neq 0$ and $0 < \epsilon < 1$.*

$$\text{If } P(|X_n - c| \geq \epsilon) \leq e^{-\kappa n \epsilon^2}, \text{ then } P(|X_n^{-1} - c^{-1}| \geq \epsilon) \leq 2e^{-n\kappa\epsilon^2 c^2 \min\{c^2, 1\}/4}.$$

Proof. Without loss of generality, we can assume that $c > 0$. We have

$$P(|X_n^{-1} - c^{-1}| \leq \epsilon) = P(c^{-1} - \epsilon \leq X_n^{-1} \leq c^{-1} + \epsilon).$$

First consider the case $0 < \epsilon < c^{-1}$. Then, X_n is strictly positive in the interval of interest, and therefore

$$\begin{aligned} P(c^{-1} - \epsilon \leq X_n^{-1} \leq c^{-1} + \epsilon) &= P\Big(\frac{-\epsilon c}{c^{-1} + \epsilon} \leq X_n - c \leq \frac{\epsilon c}{c^{-1} - \epsilon}\Big) \\ &\geq 1 - e^{-n\kappa\epsilon^2 c^2/(\epsilon + c^{-1})^2} \geq 1 - e^{-n\kappa\epsilon^2 c^4/4}. \end{aligned} \tag{A.2}$$

Next consider $0 < c^{-1} < \epsilon < 1$. The probability to be bounded can be written as

$$\begin{aligned} &P(X_n^{-1} \geq c^{-1} + \epsilon) + P(-(\epsilon - c^{-1}) \leq X_n^{-1} < 0) \\ &= P\Big(X_n - c \leq \frac{-\epsilon c}{\epsilon + c^{-1}}\Big) + P\Big(\frac{-\epsilon c}{\epsilon - c^{-1}} \leq X_n - c \leq -c\Big) \\ &\leq e^{-n\kappa\epsilon^2 c^2/(\epsilon + c^{-1})^2} + e^{-n\kappa c^2} \leq e^{-n\kappa c^2/4} + e^{-n\kappa c^2} \leq 2e^{-n\kappa c^2/4}, \end{aligned} \tag{A.3}$$

where the last two inequalities are obtained using $\epsilon > c^{-1}$ and $\epsilon < 1$, respectively. The bounds (A.2) and (A.3) together give the result of the lemma. □

B Gaussian and Sub-Gaussian Concentration

Lemma B.1. *For a random variable $Z \sim \mathcal{N}(0,1)$ and $\epsilon > 0$, $P(|Z| \geq \epsilon) \leq 2e^{-\frac{1}{2}\epsilon^2}$.*

Lemma B.2 (χ^2-concentration)**.** *For Z_i, $i \in [n]$ that are i.i.d. $\sim \mathcal{N}(0,1)$, and $0 \leq \epsilon \leq 1$,*

$$P\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} Z_i^2 - 1\Big| \geq \epsilon\Big) \leq 2e^{-n\epsilon^2/8}.$$

Lemma B.3. *[25] Let X be a centered sub-Gaussian random variable with variance factor ν, i.e., $\ln \mathbb{E}[e^{tX}] \leq \frac{t^2\nu}{2}$, for all $t \in \mathbb{R}$. Then X satisfies:*

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 $k \geq 1$,*
$$\mathbb{E}[X^{2k}] \leq 2(k!)(2\nu)^k \leq (k!)(4\nu)^k. \tag{B.1}$$

Lemma B.4. *Let $Z_1, \ldots, Z_t \in \mathbb{R}^N$ be random vectors such that $(Z_{1,i}, \ldots, Z_{t,i})$ are i.i.d. across $i \in [n]$, with $(Z_{1,i}, \ldots, Z_{t,i})$ being jointly Gaussian with zero mean, unit variance and covariance matrix $K \in \mathbb{R}^{t\times t}$. Let $G \in \mathbb{R}^N$ be a random vector with entries $G_1, \ldots, G_N$ i.i.d. $\sim p_G$, where p_G is sub-Gaussian with variance factor ν. Then for any pseudo-Lipschitz function $f : \mathbb{R}^{t+1} \to \mathbb{R}$, non-negative constants $\sigma_1, \ldots, \sigma_t$, and $0 < \epsilon \leq 1$, we have*

$$\begin{aligned}
&P\Big(\Big|\frac{1}{N}\sum_{i=1}^{N} f(\sigma_1 Z_{1,i}, \ldots, \sigma_t Z_{t,i}, G_i) - \mathbb{E}[f(Z_{1,1}, \ldots, Z_{t,1}, G)]\Big| \geq \epsilon\Big) \\
&\leq 2\exp\left(\frac{-N\epsilon^2}{128L^2(t+1)^2\Big(\nu + 4\nu^2 + \sum_{m=1}^{t}(\sigma_m^2 + 4\sigma_m^4)\Big)}\right),
\end{aligned}$$

where $L > 0$ is an absolute constant. (L can be bounded above by three times the pseudo-Lipschitz constant of f.)

Proof. Without loss of generality, assume $\mathbb{E}[f(\sigma_1 Z_{1,i}, \ldots, \sigma_t Z_{t,i}, G_i)] = 0$ for $i \in [N]$. In what follows we demonstrate the upper-tail bound:

$$P\Big(\frac{1}{N}\sum_{i=1}^{N} f(\sigma_1 Z_{1,i}, \ldots, \sigma_t Z_{t,i}, G_i) \geq \epsilon\Big) \leq \exp\left(\frac{-N\epsilon^2}{4\tilde{\kappa}_t}\right), \tag{B.2}$$

where

$$\tilde{\kappa}_t = 32L^2(t+1)^2\Big(\nu + 4\nu^2 + \sum_{m=1}^{t}(\sigma_m^2 + 4\sigma_m^4)\Big). \tag{B.3}$$

The lower-tail bound follows similarly.

Using the Cramér-Chernoff method, for any $s > 0$ we can write

$$P\Big(\frac{1}{N}\sum_{i=1}^{N} f(\sigma_1 Z_{1,i}, \ldots, \sigma_t Z_{t,i}, G_i) \geq \epsilon\Big) \leq \mathbb{E}\Big[e^{s\sum_{i=1}^{N} f(\sigma_1 Z_{1,i}, \ldots, \sigma_t Z_{t,i}, G_i)}\Big] e^{-sN\epsilon}. \tag{B.4}$$

To prove (B.2), we will show that

$$\mathbb{E}\left[\exp\left(s\sum_{i=1}^{N} f(\sigma_1 Z_{1,i},\ldots,\sigma_t Z_{t,i},G_i)\right)\right] \le \exp\left(N\tilde{\kappa}_t s^2\right) \quad \text{for } 0<s<\sqrt{\frac{1}{\tilde{\kappa}_t}}. \tag{B.5}$$

Then, using (B.5) in (B.4) and taking $s=\epsilon/2\tilde{\kappa}_t$ yields the upper tail bound in (B.2).

We now prove (B.5). For $i\in[N]$, let $(\tilde{Z}_{1,i},\ldots,\tilde{Z}_{t,i},\tilde{G}_i)$ be an independent copy of $(Z_{1,i},\ldots,Z_{t,i},G_i)$. Since $\mathbb{E}[f(\sigma_1\tilde{Z}_{1,i},\ldots,\sigma_t\tilde{Z}_{t,i},\tilde{G}_i)]=0$, using Jensen's inequality we have

$$\mathbb{E}[\exp(-sf(\sigma_1\tilde{Z}_{1,i},\ldots,\sigma_t\tilde{Z}_{t,i},\tilde{G}_i))]\ge\exp(-s\mathbb{E}[f(\sigma_1\tilde{Z}_{1,i},\ldots,\sigma_t\tilde{Z}_{t,i},\tilde{G}_i)])=1.$$

Therefore, using the independence of $\tilde{Z}$ and Z we write

$$\begin{aligned}\mathbb{E}[e^{sf(\sigma_1 Z_{1,i},\ldots,\sigma_t Z_{t,i},G_i)}] &\le \mathbb{E}[e^{sf(\sigma_1 Z_{1,i},\ldots,\sigma_t Z_{t,i},G_i)}]\cdot\mathbb{E}[e^{-sf(\sigma_1\tilde{Z}_{1,i},\ldots,\sigma_t\tilde{Z}_{t,i},\tilde{G}_i)}]\\ &= \mathbb{E}[e^{s(f(\sigma_1 Z_{1,i},\ldots,\sigma_t Z_{t,i},G_i)-f(\sigma_1\tilde{Z}_{1,i},\ldots,\sigma_t\tilde{Z}_{t,i},\tilde{G}_i))}].\end{aligned} \tag{B.6}$$

Using (B.6) we prove (B.5) by demonstrating that for each $i\in[N]$,

$$\mathbb{E}[e^{s(f(\sigma_1 Z_{1,i},\ldots,\sigma_t Z_{t,i},G_i)-f(\sigma_1\tilde{Z}_{1,i},\ldots,\sigma_t\tilde{Z}_{t,i},\tilde{G}_i))}]\le\exp\left(\tilde{\kappa}_t s^2\right) \quad \text{for } 0<s<\sqrt{\frac{1}{\tilde{\kappa}_t}}. \tag{B.7}$$

For $i\in[N]$ we have

$$\begin{aligned}&\mathbb{E}[e^{s(f(\sigma_1 Z_{1,i},\ldots,\sigma_t Z_{t,i},G_i)-f(\sigma_1\tilde{Z}_{1,i},\ldots,\sigma_t\tilde{Z}_{t,i},\tilde{G}_i))}]\\ &=\sum_{q=0}^{\infty}\frac{s^q}{q!}\mathbb{E}\left(f(\sigma_1 Z_{1,i},\ldots,\sigma_t Z_{t,i},G_i)-f(\sigma_1\tilde{Z}_{1,i},\ldots,\sigma_t\tilde{Z}_{t,i},\tilde{G}_i)\right)^q\\ &\overset{(a)}{=}\sum_{k=0}^{\infty}\frac{s^{2k}}{(2k)!}\mathbb{E}\left(f(\sigma_1 Z_{1,i},\ldots,\sigma_t Z_{t,i},G_i)-f(\sigma_1\tilde{Z}_{1,i},\ldots,\sigma_t\tilde{Z}_{t,i},\tilde{G}_i)\right)^{2k},\end{aligned} \tag{B.8}$$

where step (a) holds because the odd moments of the difference equal 0. Next, using the pseudo-Lipschitz property of f, for an absolute constant $L>0$, we have for $k\ge1$:

$$\begin{aligned}&\left(f(\sigma_1 Z_{1,i},\ldots,\sigma_t Z_{t,i},G_i)-f(\sigma_1\tilde{Z}_{1,i},\ldots,\sigma_t\tilde{Z}_{t,i},\tilde{G}_i)\right)^{2k}\\ &\le L^{2k}\left[1+\sum_{m=1}^{t}\sigma_m^2(Z_{m,i}^2+\tilde{Z}_{m,i}^2)+G_i^2+\tilde{G}_i^2\right]^k\left[\sum_{m=1}^{t}\sigma_m^2(Z_{m,i}-\tilde{Z}_{m,i})^2+(G_i-\tilde{G}_i)^2\right]^k\\ &\overset{(a)}{\le} L^{2k}\left[1+\sum_{m=1}^{t}\sigma_m^2(Z_{m,i}^2+\tilde{Z}_{m,i}^2)+G_i^2+\tilde{G}_i^2\right]^k 2^k\left[\sum_{m=1}^{t}\sigma_m^2(Z_{m,i}^2+\tilde{Z}_{m,i}^2)+G_i^2+\tilde{G}_i^2\right]^k\\ &\overset{(b)}{\le}(2L^2)^k\left[\sum_{m=1}^{t}\sigma_m^2(Z_{m,i}^2+\tilde{Z}_{m,i}^2)+G_i^2+\tilde{G}_i^2+(2t+2)\left(\sum_{m=1}^{t}\sigma_m^4(Z_{m,i}^4+\tilde{Z}_{m,i}^4)+G_i^4+\tilde{G}_i^4\right)\right]^k,\\ &\overset{(c)}{\le}\frac{(2L^2(4t+4))^k}{4t+4}\left[\sum_{m=1}^{t}\sigma_m^{2k}(Z_{m,i}^{2k}+\tilde{Z}_{m,i}^{2k})+G_i^{2k}+\tilde{G}_i^{2k}+(2t+2)^k\left(\sum_{m=1}^{t}\sigma_m^{4k}(Z_{m,i}^{4k}+\tilde{Z}_{m,i}^{4k})+G_i^{4k}+\tilde{G}_i^{4k}\right)\right]\\ &\le\frac{(2L(2t+2))^{2k}}{4t+4}\left[\sum_{m=1}^{t}\sigma_m^{2k}(Z_{m,i}^{2k}+\tilde{Z}_{m,i}^{2k})+G_i^{2k}+\tilde{G}_i^{2k}+\sum_{m=1}^{t}\sigma_m^{4k}(Z_{m,i}^{4k}+\tilde{Z}_{m,i}^{4k})+G_i^{4k}+\tilde{G}_i^{4k}\right],\end{aligned} \tag{B.9}$$

where inequalities $(a), (b), (c)$ are all obtained using using Lemma C.3. Using (B.9) in (B.8) and recalling that $\{(Z_{m,i})_{1\le k\le t}, G_i\}$ are identically distributed as $\{(\tilde{Z}_{m,i})_{1\le k\le t}, \tilde{G}_i\}$, we get

$$
\begin{aligned}
&\mathbb{E}[e^{s(f(\sigma_1 Z_{1,i},\ldots,\sigma_t Z_{t,i},G_i)-f(\sigma_1 \tilde{Z}_{1,i},\ldots,\sigma_t \tilde{Z}_{t,i},\tilde{G}_i))}] \\
&\le 1+\sum_{k=1}^{\infty}\frac{(s2L(2t+2))^{2k}}{(2k)!(4t+4)}\ 2\left[\sum_{m=1}^{t}\sigma_m^{2k}\,\mathbb{E}Z_{m,i}^{2k}+\mathbb{E}G_i^{2k}+\sum_{m=1}^{t}\sigma_m^{4k}\,\mathbb{E}Z_{m,i}^{4k}+\mathbb{E}G_i^{4k}\right] \\
&\overset{(a)}{\le} 1+\sum_{k=1}^{\infty}\frac{(s2L(2t+2))^{2k}}{(2k)!(2t+2)}\left[\sum_{m=1}^{t}\sigma_m^{2k}2(k!)2^k+2(k!)(2\nu)^k+\sum_{m=1}^{t}\sigma_m^{4k}\,2(2k!)2^{2k}+2(2k!)(2\nu)^{2k}\right] \\
&\overset{(b)}{\le} 1+\sum_{k=1}^{\infty}\frac{(s2L(2t+2))^{2k}}{(t+1)}\left[\frac{1}{k!}\sum_{m=1}^{t}\sigma_m^{2k}+\frac{\nu^k}{k!}+\sum_{m=1}^{t}(4\sigma_m^4)^k+(4\nu^2)^k\right] \\
&\le 1+\sum_{k=1}^{\infty}(s2L(2t+2))^{2k}\left[\nu+4\nu^2+\sum_{m=1}^{t}(\sigma_m^2+4\sigma_m^4)\right]^k \\
&\overset{(c)}{=}\left(1-s^2 16L^2(t+1)^2\Big[\nu+4\nu^2+\sum_{m=1}^{t}(\sigma_m^2+4\sigma_m^4)\Big]\right)^{-1} \\
&\overset{(d)}{\le} e^{s^2 32L^2(t+1)^2\left[\nu+4\nu^2+\sum_{m=1}^{t}(\sigma_m^2+4\sigma_m^4)\right]}. \qquad \text{(B.10)}
\end{aligned}
$$

In the chain of inequalities above, (a) is obtained using the sub-Gaussian moment bound (B.1); step (b) using the inequality $\frac{(2k)!}{k!}\ge 2^k k!$, which can be seen as follows.

$$\frac{(2k)!}{k!}=\prod_{j=1}^{k}(k+j)=k!\ \prod_{j=1}^{k}\Big(\frac{k}{j}+1\Big)\ge (k!)2^k.$$

The equality (c) holds because s lies in the range specified by (B.5), and (d) holds because $\frac{1}{1-x}\le e^{2x}$ for $x\in[0,\frac{1}{2}]$. This completes the proof of (B.7), and hence the result. □

C Other Useful Lemmas

Lemma C.1 (Product of Lipschitz Functions is PL(2))**.** *Let $f:\mathbb{R}^p\to\mathbb{R}$ and $g:\mathbb{R}^p\to\mathbb{R}$ be Lipschitz continuous. Then the product function $h:\mathbb{R}^p\to\mathbb{R}$ defined as $h(x):=f(x)g(x)$ is pseudo-Lipschitz of order 2.*

Lemma C.2. *Let $\phi:\mathbb{R}^{t+2}\to\mathbb{R}$ be $PL(2)$. For $(c_1,\ldots,c_{t+1})$ constants and $Z\sim\mathcal{N}(0,1)$, the function $\tilde{\phi}:\mathbb{R}^{t+1}\to\mathbb{R}$ defined as $\tilde{\phi}\,(v_1,\ldots,v_t,w)=\mathbb{E}_Z[\phi(v_1,\ldots,v_t,\sum_{r=1}^{t}c_r v_r+c_{t+1}Z,w)]$ is then also PL(2).*

Lemma C.3. *For any scalars $a_1,\ldots,a_t$ and positive integer m, we have $(|a_1|+\ldots+|a_t|)^m\le t^{m-1}\sum_{i=1}^{t}|a_i|^m$. Consequently, for any vectors $\underline{u}_1,\ldots,\underline{u}_t\in\mathbb{R}^N$, $\left\|\sum_{k=1}^{t}\underline{u}_k\right\|^2\le t\sum_{k=1}^{t}\|\underline{u}_k\|^2$.*

Proof. The first result follows from applying Hölder's inequality to the length-t vectors $(|a_1|,\ldots,|a_t|)$ and $(1,\ldots,1)$. The second statement is obtained by applying the result with $m=2$. □

D Supplementary Material: Proof of Lemma 4.5 parts (b).(ii) and (b).(iv)

The supplement available at `http://bit.ly/2iWMgbr` contains the proof of Lemma 4.5 parts (b).(ii) and (b).(iv) for the case where the denoising functions $\{\eta_t(\cdot)\}_{t>0}$ are differentiable in the first argument except at a finite number of points. The proof in Sec. 5 covers the case where the denoising functions $\{\eta_t(\cdot)\}_{t>0}$ are differentiable everywhere. The proof of the general case is longer and somewhat tedious, so we include it in the supplement.

Acknowledgement

We thank Andrew Barron for helpful discussions regarding certain technical aspects of the proof. This work was supported in part by a Marie Curie Career Integration Grant (GA Number 631489).

References

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

[2] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proc. Natl. Acad. Sci. USA*, vol. 106, no. 45, pp. 18914–18919, 2009.

[3] F. Krzakala, M. Mézard, F. Sausset, Y. Sun, and L. Zdeborová, "Probabilistic reconstruction in compressed sensing: algorithms, phase diagrams, and threshold achieving matrices," *J. Stat. Mech. Theory Exp.*, no. 8, 2012.

[4] A. Montanari, "Graphical models concepts in compressed sensing," in *Compressed Sensing* (Y. C. Eldar and G. Kutyniok, eds.), pp. 394–438, Cambridge University Press, 2012.

[5] S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," in *Proc. IEEE Int. Symp. Inf. Theory*, pp. 2168–2172, 2011.

[6] M. Bayati and A. Montanari, "The LASSO Risk for Gaussian Matrices," *IEEE Trans. Inf. Theory*, vol. 58, pp. 1997–2017, April 2012.

[7] A. Javanmard and A. Montanari, "State evolution for general approximate message passing algorithms, with applications to spatial coupling," *Information and Inference: A Journal of the IMA Inference*, vol. 2, no. 2, pp. 115–144, 2013.

[8] M. Bayati, M. Lelarge, and A. Montanari, "Universality in polytope phase transitions and message passing algorithms," *Ann. Appl. Probab.*, vol. 25, pp. 753–822, 04 2015.

[9] S. Rangan and A. K. Fletcher, "Iterative estimation of constrained rank-one matrices in noise," in *Proc. IEEE Int. Symp. Inf. Theory*, pp. 1246–1250, 2012.

[10] Y. Deshpande and A. Montanari, "Information-theoretically optimal sparse PCA," in *Proc. IEEE Int. Symp. Inf. Theory*, pp. 2197–2201, 2014.

[11] A. Montanari and E. Richard, "Non-negative principal component analysis: Message passing algorithms and sharp asymptotics," *IEEE Trans. Inf. Theory*, vol. 62, no. 3, pp. 1458–1484, 2016.

[12] Y. Deshpande, E. Abbe, and A. Montanari, "Asymptotic mutual information for the balanced binary stochastic block model," *Information and Inference: A Journal of the IMA*, vol. 6, pp. 125–170, 2016.

[13] T. Lesieur, F. Krzakala, and L. Zdeborová, "MMSE of probabilistic low-rank matrix estimation: Universality with respect to the output channel," in *53rd Annual Allerton Conference on Communication, Control, and Computing*, pp. 680–687, 2015.

[14] J. Barbier, M. Dia, N. Macris, F. Krzakala, T. Lesieur, and L. Zdeborová, "Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula," in *Advances in Neural Information Processing Systems*, pp. 424–432, 2016.

[15] J. Barbier and F. Krzakala, "Replica analysis and approximate message passing decoder for sparse superposition codes," in *Proc. IEEE Int. Symp. Inf. Theory*, 2014.

[16] C. Rush, A. Greig, and R. Venkataramanan, "Capacity-achieving sparse regression codes via approximate message passing decoding," *IEEE Trans. Inf. Theory*, vol. 63, pp. 1476–1500, March 2017.

[17] J. Barbier and F. Krzakala, "Approximate message-passing decoder and capacity-achieving sparse superposition codes," *IEEE Trans. Inf. Theory*, vol. 63, pp. 4894–4927, August 2017.

[18] Y. Kabashima, F. Krzakala, M. Mézard, A. Sakata, and L. Zdeborová, "Phase transitions and sample complexity in Bayes-optimal matrix factorization," *IEEE Trans. Inf. Theory*, vol. 62, no. 7, pp. 4228–4265, 2016.

[19] J. T. Parker, P. Schniter, and V. Cevher, "Bilinear generalized approximate message passing – part I: Derivation," *IEEE Trans. Signal Processing*, vol. 62, no. 22, pp. 5839–5853, 2014.

[20] J. T. Parker, P. Schniter, and V. Cevher, "Bilinear generalized approximate message passing – part II: Applications," *IEEE Trans. Signal Processing*, vol. 62, no. 22, pp. 5854–5867, 2014.

[21] E. Bolthausen, "An iterative construction of solutions of the TAP equations for the Sherrington–Kirkpatrick model," *Communications in Mathematical Physics*, vol. 325, pp. 333–366, Jan 2014.

[22] D. L. Donoho, A. Javanmard, and A. Montanari, "Information-theoretically optimal compressed sensing via spatial coupling and approximate message passing," *IEEE Trans. Inform. Theory*, vol. 59, pp. 7434–7464, Nov. 2013.

[23] C. Rush and R. Venkataramanan, "The error exponent of sparse regression codes with AMP decoding," in *Proc. IEEE Int. Symp. Inf. Theory*, 2017.

[24] Y. Ma, C. Rush, and D. Baron, "Analysis of approximate message passing with a class of non-separable denoisers," in *Proc. IEEE Int. Symp. Inf. Theory*, 2017.

[25] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

[26] D. Donoho and A. Montanari, "High dimensional robust M-estimation: asymptotic variance via approximate message passing," *Probab. Theory Related Fields*, vol. 166, no. 3, pp. 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. Inf. Theory*, vol. 60, pp. 2969–2985, May 2014.

[28] J. Ma and L. Ping, "Orthogonal AMP for compressed sensing with unitarily-invariant matrices," *IEEE Access*, vol. 5, pp. 2020–2033, 2017.

[29] K. Takeuchi, "Rigorous dynamics of expectation-propagation-based signal recovery from unitarily invariant measurements," in *Proc. IEEE Int. Symp. Inf. Theory*, pp. 501–505, 2017.

[30] S. Rangan, P. Schniter, and A. Fletcher, "Vector approximate message passing," in *Proc. IEEE Int. Symp. Inf. Theory*, 2017.