A non-asymptotic distributional theory of approximate message passing for sparse and robust regression

Gen Li [*] Yuting Wei[†]

January 12, 2024

Abstract

Characterizing the distribution of high-dimensional statistical estimators is a challenging task, due to the breakdown of classical asymptotic theory in high dimension. This paper makes progress towards this by developing non-asymptotic distributional characterizations for approximate message passing (AMP) — a family of iterative algorithms that prove effective as both fast estimators and powerful theoretical machinery — for both sparse and robust regression. Prior AMP theory, which focused on high-dimensional asymptotics for the most part, failed to describe the behavior of AMP when the number of iterations exceeds $o\big(\log n/\log\log n\big)$ (with n the sample size). We establish the first finite-sample non-asymptotic distributional theory of AMP for both sparse and robust regression that accommodates a polynomial number of iterations. Our results derive approximate accuracy of Gaussian approximation of the AMP iterates, which improves upon all prior results and implies enhanced distributional characterizations for both optimally tuned Lasso and robust M-estimator.

Keywords: linear models, approximate message passing, non-asymptotic analysis, sparse regression, robust regression

Contents

[*]Department of Statistics, The Chinese University of Hong Kong, Hong Kong.
[†]Department of Statistics and Data Science, the Wharton School, University of Pennsylvania, Philadelphia, PA.

1 Introduction

Determining the distributions of the estimators of interest plays a pivotal role in addressing fundamental questions in uncertainty quantification, hypothesis testing, and risk prediction, among others. In classical large-sample theory (Fisher, 1922; Le Cam, 2012; Van der Vaart, 2000), this is often achieved by pinning down the limiting distribution, such as asymptotic normality, of the estimators of interest in the limit as the sample size n approaches infinity with the problem dimension held fixed. Nevertheless, such large-sample theory often breaks down in modern high-dimensional settings where the ambient dimension p of the unknowns is large as well (e.g., comparable to the sample size), due to prevalent issues such as non-negligible bias and inflated variance (El Karoui et al., 2013; Donoho and Montanari, 2015, 2016; Sur and Candès, 2019; Sur et al., 2019). These issues have motivated a recent wave of research activities proposing new paradigms and analyses that enable tractable distributional characterizations in high dimension (see e.g. Zhang and Zhang (2014); Van de Geer et al. (2014); Javanmard and Montanari (2014, 2018); Ren et al. (2015); Bellec and Zhang (2022, 2023); Bellec et al. (2022); Chen et al. (2019b); Celentano et al. (2023c); Cai et al. (2022); Xia and Yuan (2021); Celentano and Montanari (2021); Yan et al. (2021) and the references therein). Focusing on linear models, the present paper aims to make progress towards understanding the distribution of a powerful family of statistical estimators, called approximate message passing (AMP) (Donoho et al., 2009; Feng et al., 2022), that are among the most effective when tackling high-dimensional problems.

1.1 Sparse and robust regression in high dimension

The current paper is focused on the prototypical problem of estimating a set of unknown parameters in a linear model. Given a design matrix $X \in \mathbb{R}^{n\times p}$ (with $X_1, \dots, X_n$ denoting the rows of X), the classical linear regression model takes the form of

$$y = X\theta^\star + \varepsilon, \tag{1}$$

where $y = [y_i]_{1\le i\le n} \in \mathbb{R}^n$ stands for the observed data vector, $\theta^\star = [\theta_i^\star]_{1\le i\le p} \in \mathbb{R}^p$ represents some unknown signal of interest, and $\varepsilon = [\varepsilon_i]_{1\le i\le n} \in \mathbb{R}^n$ indicates independent random noise contaminating the observations. The aim is to reconstruct the unknown object $\theta^\star$ based on (y, X). In practice, it is common to encounter situations where either the signal coefficients or the noise distributions exhibit certain structural properties (e.g., sparsity, group sparsity, heavy tails) that are known to scientists *a priori* (e.g., Tibshirani (1996); Chen et al. (2001); Donoho et al. (2001); Mitchell and Beauchamp (1988); Fan and Li (2001); Zou and Hastie (2005); Yuan and Lin (2006); Candes and Tao (2007); Donoho and Montanari (2015); Bogdan et al. (2015); Sun et al. (2020); Bu et al. (2020); Bühlmann and Van De Geer (2011); Hastie et al. (2015); Fan et al. (2020)). While numerous instances of linear regression have been studied across various contexts, we single out two concrete settings that will serve as a guiding thread throughout this paper.

- *Sparse regression.* Imagine that the signal of interest $\theta^\star \in \mathbb{R}^p$ in (1) is sparse, namely,

$$\theta^\star \text{ is } k\text{-sparse} \tag{2}$$

with the sparsity level k much smaller than the ambient dimension p. The widespread applicability of sparse linear regression across diverse data science applications, including but not limited to medical imaging, genomics, geophysics, and signal processing, has inspired substantial research activities dedicated to the design and analysis of sparse statistical estimators (e.g., Tibshirani (1996); Donoho (2006); Candes et al. (2006); Fan and Li (2001); Zou and Hastie (2005); Yuan and Lin (2006)).

- *Robust regression.* While a dominant fraction of linear regression works operates upon commonly encountered noise assumptions like Gaussians, the prevalence of (sparse) outliers in reality has incentivized research into the "robustness" aspect of regression. Originally proposed by Huber (1964, 1973), the gross-errors contamination model assumes that each noise component ε_i is independently drawn from the following distribution:

$$\varepsilon_i \overset{\text{i.i.d.}}{\sim} (1-\epsilon_H)\mathcal{N}(0,\sigma^2) + \epsilon_H H, \tag{3}$$

where H denotes some (unknown) contaminating distribution, and $\epsilon_H \in (0,1)$ represents the contamination fraction. In other words, the observed data might contain a fraction ϵ_H of abnormal data that deviate from situations under Gaussian noise. Statistical performances of robust estimators tailored for this model are developed subsequently in Hampel (1974); Bickel (1975); Maronna et al. (2019); Fan et al. (2014); Loh (2017); Sun et al. (2020); El Karoui et al. (2013); Donoho and Montanari (2016, 2015); El Karoui (2018); Lei et al. (2018), among others.

The current paper concentrates on a particularly challenging scenario called the proportional-growth regime, where the number of observations n and the ambient dimension p are on the same order. In the sparse regression case, our focus is on the linear sparsity regime, where the sparsity level k is on the same order as p and n. A family of algorithm that are well-suited for this challenging scenario is called approximate message passing (AMP), which we shall elaborate on next.

1.2 Approximate message passing (AMP)

Approximate message passing was originally developed in the context of compressed sensing (Donoho et al., 2009; Bayati and Montanari, 2011a) as a family of low-complexity iterative algorithms, and has now been widely recognized as a powerful machinery to assist in understanding the performances of a broad class of statistical procedures, especially in scenarios with low signal-to-noise ratios (SNRs). Its applications span linear and generalized linear models (Bayati and Montanari, 2011b; Rangan, 2011; Schniter and Rangan, 2014; Donoho and Montanari, 2016; Sur et al., 2019; Barbier et al., 2019; Mondelli and Venkataramanan, 2022; Li and Wei, 2021; Fan, 2022; Zhang et al., 2023; Li et al., 2023b; Celentano et al., 2023a), low-rank matrix estimation (Rangan and Fletcher, 2012; Montanari and Venkataramanan, 2021; Deshpande and Montanari, 2014; Mondelli and Venkataramanan, 2021; Zhong et al., 2021; Celentano et al., 2023b; Li and Wei, 2022), community detection (Deshpande et al., 2017; Ma and Nandy, 2021; Wang et al., 2022), and more recently, Bayesian sampling from diffusion processes (Montanari and Wu, 2023). The interested reader is refeerred to Feng et al. (2022) for a comprehensive overview of AMP and its broad applications.

When applied to the linear model (1) with i.i.d. Gaussian design, the AMP procedure typically maintains running estimates $\{\theta_t\}_{t\ge 0} \subset \mathbb{R}^p$ of the signal $\theta^\star$ as well as adjusted residuals $\{r_t\}_{t\ge 0} \subseteq \mathbb{R}^n$. More specifically, letting $f_t : \mathbb{R} \to \mathbb{R}$ and $g_t : \mathbb{R}^n \to \mathbb{R}^n$ be some properly chosen denoising functions, AMP executes the following update rule in the t-th iteration:

$$r_t = y - Xf_t(\theta_t) + \langle f_t'(\theta_t)\rangle \big(\langle g_{t-1}'(r_{t-1})\rangle\big)^{-1} g_{t-1}(r_{t-1}), \tag{4a}$$

$$\theta_{t+1} = \big(\langle g_t'(r_t)\rangle\big)^{-1} X^\top g_t(r_t) + f_t(\theta_t), \tag{4b}$$

where f_t, g_t, and their derivatives (f_t' and g_t') are applied component-wise to the vector argument, and for every integer $m > 0$, we adopt a slightly unconventional piece of notation[1]

$$\langle x\rangle \coloneqq \frac{1}{n}\sum_{i=1}^{m} x_i, \qquad \text{for } x \in \mathbb{R}^m. \tag{5}$$

The algorithm is initialized at

$$f_1(\theta_1) = 0 \in \mathbb{R}^p, \qquad g_0(r_0) = 0,$$

and quantities associated with non-positive iteration numbers are all set to be zero. When instantiated to the two concrete settings described above, the following denoising functions have been suggested in past works.

- *AMP for sparse regression.* When tackling the sparse regression setting, AMP adopts the denoising functions

$$g_t(x) = x \qquad \text{and} \qquad f_t(x) = \mathrm{sign}(x)\big(|x| - \tau_t\big)_+ =: \mathsf{ST}_{\tau_t}(x) \tag{6}$$

  for some properly selected threshold τ_t (to be made precisely in Section 2.2). Notably, f_t is chosen to be the soft-thresholding function in order to promote sparsity. As demonstrated in Bayati and Montanari (2011b), the AMP procedure (4) with the choices (6) serves as a fast algorithm to solve, and help assess the risk of, the Lasso estimator in the most sample-starved regime.

- *AMP for robust regression.* When it comes to the robust regression problem, suppose first that we are given a convex loss function $\rho : \mathbb{R} \to \mathbb{R}_{\geq 0}$. The denoising functions for AMP can then be selected as

$$g_t(x) = \frac{n}{p}\Psi(z, b_t) \qquad \text{and} \qquad f_t(x) = x, \tag{7}$$

  where Ψ is defined such that

$$\Psi(z, b) = \rho_b'(z) \qquad \text{with } \rho_b(z) \coloneqq \min_x \Big\{\rho(x) + \frac{1}{2b}(x - z)^2\Big\}. \tag{8}$$

  Here, ρ_b (with some $b > 0$) can be viewed as a regularized variant of ρ, and the regularization parameter b_t will be made precisely momentarily (see Section 2.3). As we shall elaborate on in Section 2.3, the AMP procedure (4) with the choices (7) is a rapid method for solving the M-estimator with loss function ρ (Donoho and Montanari, 2016).

In addition to their computational efficiency, the aforementioned AMP algorithms often admit exact asymptotic characterizations, in the sense that their risk and dynamics in the high-dimensional asymptotics can often be characterized in a precise manner. Consequently, AMP has now been widely recognized as both a family of standalone fast statistical estimators and a power machineary for analyzing other optimization-based statistical estimators (e.g., the M-estimator and the Lasso).

1.3 From asymptotics to non-asymptotics

Exact asymptotics and state evolution. Recent years have witnessed a flurry of activity in developing theoretical tools towards demystifying the efficacy of AMP. More concretely, existing AMP theory reveals that: the behavior of each iteration of AMP, in the high-dimensional asymptotics (i.e., with n, p approaching infinity and t held fixed), can often be characterized by a low-dimensional recursive formula dubbed as the *state evolution (SE)*. Informally, under i.i.d. Gaussian design (i.e., $X_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1/n)$) as well as some other mild conditions, past theory introduced the following SE recursion

$$\alpha_t^{\star 2} = \mathbb{E}\big[G_t^2(\gamma_t^\star Z, W)\big], \qquad \gamma_t^{\star 2} = \mathbb{E}\big[F_t^2(\alpha_{t-1}^\star Z, V)\big], \qquad t \geq 1, \tag{9}$$

[1] Note that here, instead of using a $1/m$ scaling, we use a $1/n$ normalization instead. Due to this slight different scaling, there is no need for an additional multiplicative factor $1/\delta \coloneqq p/n$ in the last term of (4a) as in Donoho et al. (2009) or Donoho and Montanari (2016).

where G_t (resp. F_t) is some function depending on g_t (resp. f_t) to be specified shortly (see Section 2.1). Here, Z, V, W are independently generated such that (i) $Z \sim \mathcal{N}(0,1)$ and (ii) V (resp. W) is drawn from the empirical distribution of $\{\sqrt{p}\theta_i^\star\}$ (resp. $\{\sqrt{n}\varepsilon_i\}$). With this two-dimensional sequence in place, it has been proven that: for any fixed iteration number t and any pseudo-Lipschitz function $\Phi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, it holds almost surely that

$$\lim_{p\to\infty} \frac{1}{p}\sum_{i=1}^{p} \Phi\Big(\sqrt{p}(\theta_{t+1,i} - \theta_i^\star), \sqrt{p}\theta_i^\star\Big) = \mathbb{E}\big[\Phi(\alpha_t^\star Z, V)\big], \tag{10a}$$

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \Phi\Big(\sqrt{n}(r_{t,i} - \varepsilon_i), \sqrt{n}\varepsilon_i\Big) = \mathbb{E}\big[\Phi(\gamma_t^\star Z, W)\big], \tag{10b}$$

provided that p/n is a fixed constant. For instance, when $\Phi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is chosen to be $\Phi(a,b) = a^2$, the prediction in (10) pins down the asymptotic squared loss of the AMP iterates as follows

$$\lim_{p\to\infty} \big\|\theta_{t+1} - \theta^\star\big\|_2^2 = \mathbb{E}\big[(\alpha_t^\star Z)^2\big] = \big(\alpha_t^\star\big)^2, \tag{11a}$$

$$\lim_{p\to\infty} \big\|r_t - \varepsilon\big\|_2^2 = \mathbb{E}\big[(\gamma_t^\star Z)^2\big] = \big(\gamma_t^\star\big)^2, \tag{11b}$$

making explicit the operational meanings of $\alpha_t^\star$ and $\gamma_t^\star$ constructed in the SE recursion (9).

A non-asymptotic theory? While state evolution has played a pivotal role towards understanding AMP in various applications, it is asymptotic in nature, in the sense that it predicts the AMP dynamics in the presence of asymptotically large dimensions with the number of iterations held fixed. This limits the prediction power of existing AMP theory in at least two aspects:

- Most prior AMP theory fell short in predicting the convergence rate of AMP below a constant error floor (e.g., it did not predict how many iterations are needed in order to achieve a risk that is $o(1)$ away from that of the fixed point).
- Most prior AMP theory did not provide non-asymptotic rates for the statistical estimation error, nor did it offer non-asymptotic distributional guarantees.

In short, when viewed as a fast iterative algorithm, existing theory for AMP provides an incomplete picture of the convergence rate when compared with that of other optimization algorithms; when employed as a theoretical machinery, the AMP theory might sometime lose its benefits as well when compared with other alternative tools (e.g., the Gaussian min-max theorem (Thrampoulidis et al., 2018; Celentano et al., 2023c) and the leave-one-out analysis framework (El Karoui, 2018; Ma et al., 2020; Chen et al., 2019a)).

Developing a finite-sample analysis of AMP is instrumental not only in comprehending AMP's efficacy as an optimization algorithm, but also in extending its utility as a fundamental statistical analysis tool. Consequently, it has been an active research direction over the last couple of years. The seminal work by Rush and Venkataramanan (2018) analyzed AMP for linear models and developed the first result allowing the number of iterations to grow with the problem dimension n — more precisely, the iteration number t can be as large as $o\big(\log n/\log\log n\big)$; this result is further improved to $O\big(\log n/\log\log n\big)$ for symmetric AMP in Bao et al. (2023). Subsequently, Li and Wei (2022) presented a general framework for understanding the non-asymptotic performance of AMP in spiked low-rank matrix estimation, allowing the number of iterations to grow as $O\big(n/\mathsf{poly}(\log n)\big)$ and facilitating a more precise non-asymptotic prediction of AMP's behavior. Of particular interest was the subsequent study by Li et al. (2023a) concerning the problem of Z_2 synchronization — a special case of structured matrix estimation — revealing fast non-asymptotic convergence of AMP even when initialized randomly. This type of results cannot be derived based on previous SE-based asymptotic analysis.

When it comes to sparse and robust regression, however, the non-asymptotic AMP theory remains highly inadequate. On one hand, Rush and Venkataramanan (2018) was only able capable of analyzing AMP up to $o\big(\log n/\log\log n\big)$ iterations, which is typically insufficient to uncover the convergence behavior of AMP for higher precision. On the other hand, the non-asymptotic framework in Li and Wei (2022) is not readily applicable to linear regression. All this gives rise to the following natural questions:

Can we develop a non-asymptotic theory for AMP tailored to sparse and robust regression, allowing the number of iterations to grow polynomially in the problem size?

This question was previously out of reach, and has been posed as an open problem in Cademartori and Rush (2023). Addressing this question is crucial in understanding and unleashing the power of AMP across diverse statistical domains.

1.4 A peek at our main contributions

In this paper, we answer the above-mentioned open problem in the affirmative, through development of a novel non-asymptotic framework that enables faithful prediction of AMP dynamics even when the number of iterations scales with the problem dimension. Based on this framework, we derive finite-sample/finite-time statistical guarantees that substantially strengthen the celebrated Gaussian approximation theory of AMP. In what follows, let us highlight several key findings.

A general analysis recipe. In an attempt to develop non-asymptotic theory for sparse and robust regression, we propose a unified recipe that facilitates fine-grained characterizations of the AMP iterates.

- *A fine-grained Gaussian decomposition of AMP iterates.* For any $1 \leq t \leq \min\{n,p\}$, we rigorize a general decomposition of the AMP updates as follows
$$\theta_{t+1} - \theta^\star = \sum_{k=1}^{t} \alpha_t^k \psi_k + \zeta_t, \tag{12}$$
where $\{\psi_k\}_{k=1}^t$ are independently generated obeying $\psi_k \sim \mathcal{N}(0, \frac{1}{n} I_p)$, $\zeta_t \in \mathbb{R}^p$ stands for a residual vector, and we denote by $\alpha_t = [\alpha_t^k]_{1\leq k\leq t} \in \mathbb{R}^t$ the coefficient vector. See Theorem 1 for details. In particular, for both sparse and robust regression, we can demonstrate that
$$\|\alpha_t\|_2^2 = \sum_{k=1}^{t} \big(\alpha_t^k\big)^2 \approx \big(\alpha_t^\star\big)^2, \tag{13}$$
with $\alpha_t^\star$ obtained in the asymptotic state evolution sequence (9).
- *Finite-sample control of the residual terms.* In light of the above decomposition (12) of AMP, we further prove in Theorem 5 that: under certain conditions, the residual terms in (12) satisfy[2]
$$\|\zeta_t\|_2 \lesssim \left(\frac{t \log^2 n}{n}\right)^{\frac{1}{3}} \tag{14}$$
for every t obeying $t \lesssim n/\log^4 n$. This result in conjunction with (12) and (13) delivers the first finite-sample theory that validates Gaussian approximation of AMP for up to $O(n/\log^4 n)$ iterations.

Non-asymptotic AMP theory for sparse and robust regression. The general recipe described above allows us to derive — in a non-asymptotic fashion — distributional characterizations of the AMP iterates for both sparse and robust regression, detailed below.

- *Sparse regression.* When the unknown signal $\theta^\star$ is k-sparse, we study the dynamics of AMP designed to promote sparsity, which has intimate connection with the optimally tuned Lasso. We demonstrate that the general framework mentioned above is particularly effective in tackling this setting, with the residual term controlled by (14). As a concrete consequence, this reveals that the estimation error $\theta_{t+1} - \theta^\star$ obeys
$$\theta_{t+1} - \theta^\star = v_{t+1} + \zeta_t \qquad \text{with } W_1\left(\mu_{v_{t+1}}, \mathcal{N}\Big(0, \frac{(\alpha_t^\star)^2}{n} I_p\Big)\right) \lesssim \frac{\mathsf{poly}(\log n)}{n^{1/2}} \quad \text{and} \quad \|\zeta_t\|_2 \lesssim \frac{\log n}{n^{1/3}} \tag{15}$$

[2] For two functions $f(n)$ and $g(n)$, we write $f(n) \lesssim g(n)$ (or $f(n) = O(g(n))$) if there exists a universal constant $c_1 > 0$ such that $f(n) \leq c_1 g(n)$; similarly, we write $f(n) \gtrsim g(n)$ if $f(n) \geq c_2 g(n)$ for some universal constant $c_2 > 0$. If both $f(n) \lesssim g(n)$ and $f(n) \gtrsim g(n)$ hold true, we denote $f(n) \asymp g(n)$.

for any $t \lesssim \mathsf{poly}(\log n)$, where μ_{v_t} represents the distribution of v_t, and $W_1(\cdot, \cdot)$ indicates the Wasserstein distance of order 1 between two distributions (to be defined in (18)). Moreover, it can be shown that $\alpha_t^\star$ converges to its limiting point (as $t \to \infty$) exponentially fast. In summary, our result confirms the efficacy of AMP for solving sparse regression, while at the same time improving upon prior theory by providing non-asymptotic distributional guarantees that remain valid up to $n/\mathsf{poly}(\log n)$ iterations. As another implication of our distributional characterization, the distance between the risk of our sparse estimator and the state evolution prediction obeys

$$\left\|\mathsf{ST}_{\tau_t}(\theta_t) - \theta^\star\right\|_2 - \gamma_t^\star = O\Big(\frac{\log n}{n^{1/3}}\Big) \tag{16}$$

after a logarithmic number of iterations; this error estimate improves upon the state-of-the-art theory for the Lasso estimator (which was $O(\frac{\mathsf{poly}(\log n)}{n^{1/4}})$ as derived in Miolane and Montanari (2021); Celentano et al. (2023c)). More details can be found in Section 2.2.

- *Robust regression.* Another contribution of this paper is to establish non-asymptotic distributional guarantees for AMP tailored to robust regression. More specifically, focusing on the Huber loss, we study the dynamics of the AMP designed to solve the robust M-estimation problem (Donoho and Montanari, 2016). In this case, we demonstrate that the AMP iterates also admit the decomposition (12) with the residual term satisfying (14) for all $t \lesssim n/\mathsf{poly}(\log n)$; as a consequence, the non-asymptotic distributional guarantees (15) continue to be valid in robust regression (albeit with a different state evolution prediction). Another implication of our results is the risk estimate

$$\|\theta_{t+1} - \theta^\star\|_2 - \gamma_t^\star = O\Big(\frac{\log n}{n^{1/3}}\Big) \tag{17}$$

for all $t = O(\log n)$. When translated to the risk of robust M-estimator, this exhibits a faster rate compared to prior work (note that the previously known bound in Han and Shen (2023) has an error term on the order of $O(\frac{\mathsf{poly}(\log n)}{n^{1/500}})$). We refer the readers to Section 2.3 for detailed discussions.

1.5 Notation

In this subsection, we introduce a set of notation that will be useful throughout. To begin with, for any integer $n > 0$, we denote $[n] = \{1, \dots, n\}$. An ϵ-cover of a set Θ w.r.t. metric $\rho(\cdot, \cdot)$ refers to a set $\{\theta^1, \theta^2, \dots, \theta^N\} \subseteq \Theta$ such that, for every $\theta \in \Theta$, there exists some $i \in [N]$ such that $\rho(\theta, \theta^i) \le \epsilon$. The ϵ-covering number $N(\epsilon, \rho, \Theta)$ is the cardinality of the smallest ϵ-cover of Θ w.r.t. metric $\rho(\cdot, \cdot)$. For notational convenience, when ρ is taken to be the ℓ_2-norm, we often abbreviate the covering number as $N(\epsilon, \Theta)$. In addition, we denote by $\mathbb{B}^d(r) = \{\theta \in \mathbb{R}^d \mid \|\theta\|_2 \le r\}$ the d-dimensional Euclidean ball of radius r centered at 0, and $\mathcal{S}^{d-1} = \{x \in \mathbb{R}^d \mid \|x\|_2 = 1\}$ the unit sphere in $\mathbb{R}^d$. We often write 0 (resp. 1) to denote the all-zero (resp. all-one) vector, and let I_n (or simply I) indicate the $n \times n$ identity matrix. When a scalar function is applied to a vector, it should be understood that the function is applied in an entry-wise fashion. In addition, for any two functions $f(\cdot)$ and $g(\cdot)$, we write $f(n) \ll g(n)$ or $f(n) = o(g(n))$ if $f(n)/g(n) \to 0$ as $n \to \infty$, and write $f(n) \gg g(n)$ if $g(n)/f(n) \to 0$ as $n \to \infty$. We use $c_1, c_2, \dots, C_1, C_2, \dots$ to denote universal constants that do not change with salient parameters. Note that these universal constants may change from line to line. For any two vectors $a = [a_i]_{1 \le i \le n}$ and $b = [b_i]_{1 \le i \le n}$ of the same dimension, we denote by $a \circ b = [a_i b_i]_{1 \le i \le n}$ the Hadamard product.

Given two probability distributions μ and ν on $\mathbb{R}^n$, the Wasserstein distance of order q between these two distributions is defined as

$$W_q(\mu, \nu) := \Big(\inf_{\gamma \in \mathcal{C}(\mu, \nu)} \mathop{\mathbb{E}}_{(x,y)\sim\gamma}\big[\|x - y\|_2^q\big]\Big)^{1/q}, \tag{18}$$

where $\mathcal{C}(\mu, \nu)$ stands for the set of all couplings of μ and ν (i.e., all joint distributions $\gamma(x, y)$ whose marginal distributions are μ and ν, respectively). We often employ $\mu(X)$ or μ_X to denote the distribution of X.

2 Main results

In this section, we present our non-asymptotic decomposition for AMP when applied to both sparse and robust regression, following the development of a crucial decomposition of the AMP iterates that make explicit their approximate Gaussianity.

2.1 A general decomposition for AMP iterates

In this section, we develop a general recipe that helps decompose each AMP iterate into three components: (i) a signal component, (ii) a superposition of Gaussian vectors that captures the main error component, and (iii) a residual term (which will be shown to be well-controlled for both sparse and robust regression).

Before proceeding, let us first make note of an equivalent form of the original AMP iterations (4) as studied in Bayati and Montanari (2011a). To be precise, by setting

$$\beta_t = \theta_t - \theta^\star \qquad \text{and} \qquad s_t = r_t - \varepsilon, \qquad t = 0, 1, \cdots \tag{19}$$

(namely, β_t (resp. s_t) indicates the error when using θ_t (resp. r_t) to estimate the true signal $\theta^\star$ (resp. the noise ε)), the AMP algorithm (4) can be equivalently expressed as (Bayati and Montanari, 2011a)

$$s_t = XF_t(\beta_t) - \langle F_t'(\beta_t)\rangle\, G_{t-1}(s_{t-1}), \tag{20a}$$
$$\beta_{t+1} = X^\top G_t(s_t) - \langle G_t'(s_t)\rangle\, F_t(\beta_t), \tag{20b}$$

where $\{F_t\}_{t\geq 1}$ and $\{G_t\}_{t\geq 0}$ denote two sequences of properly chosen scalar functions (note that they are applied entrywise to the vector argument here) as

$$G_t(s) = \langle g_t'(s+\varepsilon)\rangle^{-1} g_t(s+\varepsilon) \qquad \text{and} \qquad F_t(\beta) = \theta^\star - f_t(\beta + \theta^\star) \tag{21a}$$

initialized to

$$G_0(s_0) = 0 \qquad \text{and} \qquad F_1(\beta_1) = \theta^\star - f_1(\beta_1) = \theta^\star.$$

As a consequence, in order to understand how θ_t evolves during the execution of the algorithm, it suffices to focus on the dynamics of β_{t+1}. The following theorem introduces a general decomposition of (s_t, β_{t+1}), whose proof is postponed to Section A.1.

Theorem 1. *Consider the linear model (1) under i.i.d. Gaussian design (i.e., $X_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1/n)$). Suppose the functions $\{G_t\}$ and $\{F_t\}$ are differentiable except at a finite number of points. For any $1 \leq t \leq \min\{n, p\}$, the AMP sequence defined in (20) admits the following decomposition:*

$$s_t = \sum_{k=1}^{t} \gamma_t^k \phi_k + \xi_t =: u_t + \xi_t, \tag{22a}$$
$$\beta_{t+1} = \sum_{k=1}^{t} \alpha_t^k \psi_k + \zeta_t =: v_{t+1} + \zeta_t, \tag{22b}$$

where

(i) $\{\phi_k\}_{k=1}^t$ and $\{\psi_k\}_{k=1}^t$ are independent vectors obeying $\phi_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \frac{1}{n}I_n)$ and $\psi_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \frac{1}{n}I_p)$;

(ii) the coefficients $\gamma_t = [\gamma_t^k]_{1\leq k\leq t} \in \mathbb{R}^t$ and $\alpha_t = [\alpha_t^k]_{1\leq k\leq t} \in \mathbb{R}^t$ satisfy

$$\|\gamma_t\|_2 = \|F_t(\beta_t)\|_2 \qquad \textit{and} \qquad \|\alpha_t\|_2 = \|G_t(s_t)\|_2; \tag{23}$$

(iii) the residual vectors obey $\xi_t \in \mathsf{span}\{G_1(s_1), \ldots, G_{t-1}(s_{t-1})\}$ and $\zeta_t \in \mathsf{span}\{F_1(\beta_1), \ldots, F_t(\beta_t)\}$.

Remark 1. *Note that the coefficient vectors γ_t and α_t might be statistically dependent on $\{\phi_k\}_{k=1}^t$ and $\{\psi_k\}_{k=1}^t$.*

In words, Theorem 1 ensures that both s_t and β_{t+1} can be viewed as weighted superpositions of Gaussian vectors in addition to some residual terms. If the residual terms are negligible (which we will demonstrate for both sparse and robust regression), s_t and β_{t+1} are well approximated by $\sum_{k=1}^{t}\gamma_t^k\phi_k$ and $\sum_{k=1}^{t}\alpha_t^k\psi_k$, which are both close to spherical Gaussian distributions (in terms of the 1-Wasserstein distance) in the sense that

$$W_1\left(\mu\left(\frac{1}{\|\alpha_t\|_2}\sum_{k=1}^{t}\alpha_t^k\psi_k\right),\mathcal{N}\Big(0,\frac{1}{n}I_p\Big)\right)\lesssim\sqrt{\frac{t\log n}{n}}.$$

Here, $\mu(\cdot)$ represents the distribution of a random vector; see Li and Wei (2022, Lemma 9) for a proof of this 1-Wasserstein distance result.

In contrast to prior literature, the above decomposition (22) is deterministic and general in nature, requiring very few assumptions (resp. no assumption) on the denoising functions (resp. the underlying signal $\theta^\star$) and making it well-suited for the studies of various models and estimators. The most critical challenge for applying Theorem 1 then boils down to bounding the magnitudes of the residual terms ξ_t and ζ_t, which often require non-trivial treatments. Fortunately, these terms can be very well controlled under both sparse and robust regression, which we shall discuss next in Sections 2.2 and 2.3.

Comparisons with prior approaches. Before continuing, we note that the iterative procedure (20) has also been analyzed in previous works (e.g., Bayati and Montanari (2011a, Section 3.2)) for understanding the high-dimensional asymptotics of AMP for solving various estimators like the Lasso. These prior techniques typically rely on the Gaussian conditioning trick (Bolthausen, 2009; Bayati and Montanari, 2011a; Wu and Zhou, 2024) and the state-evolution type analysis, which are drastically different from our proof strategy (as we shall elucidate momentarily). As another remark, the quantities $\|\gamma_t\|_2$ and $\|\alpha_t\|_2$ in Theorem 1 are often closely connected to the scalars $\gamma_t^\star$ and $\alpha_t^\star$ in the asymptotic state evolution (9), which will be made more clear in the next subsections. For this reason, we will sometimes refer to $\|\gamma_t\|_2$ and $\|\alpha_t\|_2$ as the finite-sample counterpart of the asymptotic state evolution.

2.2 Non-asymptotic AMP theory for sparse regression

With the general decomposition in Theorem 1 in place, we can readily move forward to investigate concrete models, for which we shall begin with sparse regression. Consider the linear model (1) with the underlying signal $\theta^\star\in\mathbb{R}^p$ being k-sparse. The statistical performance of various sparse estimators has been extensively studied, with a primary focus on the regime where k is substantially smaller than p (see, e.g., Donoho (2006); Candes et al. (2006); Candes and Tao (2007); Meinshausen and Bühlmann (2006); Fan and Li (2001); Zou and Hastie (2005); Yuan and Lin (2006); Rudelson and Vershynin (2008); Wainwright (2009); Zhao and Yu (2006); Zhang (2010)). In this work, we focus on the most sample-starved regime with linear sparsity and proportional growth, namely,

$$k\asymp n\asymp p,\tag{24}$$

a regime in which AMP proves extremely powerful (Bayati and Montanari, 2011a,b; Javanmard and Montanari, 2013; Maleki et al., 2013; Donoho et al., 2013; Su et al., 2017; Rush and Venkataramanan, 2018; Fan, 2022).

AMP for sparse regression. Let us first remind the readers of the AMP procedure tailored to sparse regression, which was introduced both as a fast procedure to find a sparse solution and as a theoretical tool for characterizing the risk of the Lasso estimator (Donoho et al., 2009; Bayati and Montanari, 2011b). As mentioned before, the algorithm (4) adopts the following denoising functions:

$$f_t(x)=\mathrm{sign}(x)(|x|-\tau_t)_+=:\mathsf{ST}_{\tau_t}(x)\qquad\text{and}\qquad g_t(x)=x.\tag{25}$$

When it comes to the alternative form (20), we can simply write

$$F_t(\beta)=\theta^\star-\mathsf{ST}_{\tau_t}(\theta^\star+\beta)\qquad\text{and}\qquad G_t(s)=s+\varepsilon.\tag{26}$$

It remains to specify the threshold sequence $\{\tau_t\}$. In this work, we concentrate on a specific choice as follows: by augmenting the notation in (4a) and defining

$$r_t(\tau) := y - Xf_t(\theta_t;\tau) + \big\langle f_t'(\theta_t;\tau)\big\rangle r_{t-1} \qquad \text{with } f_t(x;\tau) := \mathrm{sign}(x)(|x|-\tau)_+, \tag{27}$$

we select the (adaptive) threshold τ_t to be

$$\tau_t := \arg\min_{\tau\geq 0} \|r_t(\tau)\|_2. \tag{28}$$

Remark 2. *As we will demonstrate later (as in Section B.1.3), one can show that τ_t is very close to the quantity $\tau_t^\star$ below:*

$$\tau_t^\star := \inf_{\tau:\tau\geq 0} \mathbb{E}\Big[\big\|\theta^\star - \mathsf{ST}_\tau(\theta^\star + \alpha_t^\star g)\big\|_2^2\Big] \qquad \textit{with } g \sim \mathcal{N}\Big(0, \frac{1}{n}I_p\Big).$$

Informally speaking, τ_t is selected in a data-driven manner aimed at minimizing the mean square estimation error.

State evolution for sparse regression. Next, we find it helpful to recall the (limiting version of) state evolution of AMP described in Donoho et al. (2009); Bayati and Montanari (2011b). Given a fixed sequence of thresholding scalars $\{b_t\}$, for every $t \geq 1$, define a two-dimensional vector $(\alpha_t^\star, \gamma_{t+1}^\star)$ recursively as

$$\alpha_t^{\star 2} = \gamma_t^{\star 2} + \|\varepsilon\|_2^2, \tag{29a}$$

$$\gamma_{t+1}^{\star 2} = \mathbb{E}\Big[\big\|\theta^\star - \mathsf{ST}_{b_t}(\theta^\star + \alpha_t^\star g)\big\|_2^2\Big], \tag{29b}$$

with $g \sim \mathcal{N}\big(0, \frac{1}{n}I_p\big)$ and $\gamma_1^\star = \|\theta^\star\|_2$. In the asymptotic limit (with $p, n \to \infty$), this SE sequence (29) often depends only upon the empirical distribution of $\theta^\star$ and the limit of $\|\varepsilon\|_2^2$, and is independent from the iterates of the AMP procedure (as long as the threshold sequence is given). When the goal is to solve the Lasso estimator for some prescribed regularization parameter $\lambda > 0$

$$\widehat{\theta}^{\mathsf{Lasso}} := \mathrm{argmin}_\theta \frac{1}{2}\|y - X\theta\|_2^2 + \lambda\|\theta\|_1, \tag{30}$$

the thresholding sequence can be selected to be $b_t = a(\lambda)\alpha_t^\star$, with $a(\lambda)$ a function of λ as specified in Bayati and Montanari (2011b).

Given our adaptive threshold (28), this subsection focuses on

$$\alpha_t^{\star 2} = \gamma_t^{\star 2} + \|\varepsilon\|_2^2, \tag{31a}$$

$$\gamma_{t+1}^{\star 2} = \mathbb{E}\Big[\big\|\theta^\star - \mathsf{ST}_{\tau_t^\star}(\theta^\star + \alpha_t^\star g)\big\|_2^2\Big], \qquad \text{with } \tau_t^\star := \inf_{\tau:\tau\geq 0} \mathbb{E}\Big[\big\|\theta^\star - \mathsf{ST}_\tau(\theta^\star + \alpha_t^\star g)\big\|_2^2\Big], \tag{31b}$$

where $g \sim \mathcal{N}\big(0, \frac{1}{n}I_p\big)$ and $\gamma_1^\star = \|\theta^\star\|_2$. With this sequence (31) in place, we shall also defining their limiting values (or fixed points):

$$\alpha^\star := \lim_{t\to\infty} \alpha_t^\star \qquad \text{and} \qquad \gamma^\star := \lim_{t\to\infty} \gamma_t^\star. \tag{32}$$

Combining this with Remark 2, we see that the AMP (4) with the threshold (28) attempts to solve the optimally tuned Lasso (namely, picking the choice of λ that minimizes the asymptotic estimation risk).

Non-asymptotic analysis for sparse regression. With the above setup and notation in place, we are ready to characterize the non-asymptotic performance of AMP below with the assistance of our general decomposition in Theorem 1. The proof of the theorem below is postponed to Section B.1.

Theorem 2. *Suppose that the k-sparse signal $\theta^\star$ and the noise vector satisfy*

$$\|\theta^\star\|_1 \gtrsim \sqrt{k} \qquad \textit{and} \qquad \|\theta^\star\|_2 \asymp \|\varepsilon\|_2 \asymp 1 \tag{33}$$

with probability exceeding $1 - O(n^{-10})$, and assume that $n > 2k\log(p/k)$ and $p > 2.3k$. Then with probability at least $1 - O(n^{-10})$, the AMP iterates (4) with denoising functions (25) and threshold (28) admit the following decomposition

$$\theta_{t+1} - \theta^\star = \sum_{j=1}^{t} \alpha_t^j \psi_j + \zeta_t \qquad \textit{and} \qquad r_t - \varepsilon = \sum_{j=1}^{t} \gamma_t^j \phi_j + \xi_t \tag{34}$$

for every $t \lesssim \frac{n}{\log^4 n}$, where $\{\psi_j\}_{1\le j\le t}$ (resp. $\{\phi_j\}_{1\le j\le t}$) are independent Gaussian vectors drawn from $\mathcal{N}(0, \frac{1}{n}I_p)$ (resp. $\mathcal{N}(0, \frac{1}{n}I_n)$), the coefficient vectors $\alpha_t = [\alpha_t^j]_{1\le j\le t}$ and $\gamma_t = [\gamma_t^j]_{1\le j\le t}$ obey

$$\big|\|\alpha_t\|_2^2 - \alpha_t^{\star 2}\big| \lesssim \Big(\frac{t\log^2 n}{n}\Big)^{1/3} \qquad \textit{and} \qquad \big|\|\gamma_t\|_2^2 - \gamma_t^{\star 2}\big| \lesssim \Big(\frac{t\log^2 n}{n}\Big)^{1/3}, \tag{35a}$$

and the residuals $\{\xi_t\}$ and $\{\zeta_t\}$ satisfy

$$\|\xi_t\|_2, \|\zeta_t\|_2 \lesssim \Big(\frac{t\log^2 n}{n}\Big)^{1/3}. \tag{35b}$$

Remark 3. *In the noiseless case (i.e., $\varepsilon = 0$), the minimum ℓ_1-norm estimator, which corresponds to the $\lambda \to 0$ limit of the Lasso estimator, undergoes a sharp phase transition. As discussed in Amelunxen et al. (2014) and Celentano et al. (2023c, Page 2201), exact recovery by this estimator can only happen when $n \ge 2k(1 + o(k/p))\log(p/k)$, which coincides with the assumption $n > 2k\log(p/k)$ imposed in Theorem 2.*

Remark 4. *It is also worth mentioning that Theorem 2 does not restrict the distribution of the noise vector ε, as long as its ℓ_2-norm is properly controlled to be on the order 1.*

Remark 5. *The exponent in the probability $1 - O(n^{-10})$ can be replaced with $1 - O(n^{-c})$ with an arbitrarily large constant $c > 0$. For simplicity, we have made no efforts to obtain the sharpest possible ones.*

Let us take a moment to highlight several implications of Theorem 2.

- *Non-asymptotic Gaussian approximation.* In a nutshell, Theorem 2 demonstrates the proximity of the AMP update θ_{t+1} and some Gaussian distribution. For instance, taking the number of iterations to be $t = c_t \log n$ for some large enough constant $c_t > 0$, we can guarantee that, with probability at least $1 - O(n^{-10})$,

$$\theta_{t+1} = \theta^\star + v_{t+1} + \zeta_t \qquad \text{with } W_1\left(\mu_{v_{t+1}}, \mathcal{N}\left(0, \frac{(\alpha_t^\star)^2}{n} I_p\right)\right) \lesssim \frac{\log n}{n^{1/2}} \ \text{ and } \ \|\zeta_t\|_2 \lesssim \frac{\log n}{n^{1/3}}, \tag{36}$$

  where $\mu_{v_{t+1}}$ represents the distribution of v_{t+1}. In fact, given that $\alpha_t^\star$ converges to the limiting point $\alpha^\star$ exponentially fast (see discussion in Section B.1.3), we can further conclude that

$$\theta_{t+1} = \theta^\star + v_{t+1} + \zeta_t \qquad \text{with } W_1\left(\mu_{v_{t+1}}, \mathcal{N}\left(0, \frac{(\alpha^\star)^2}{n} I_p\right)\right) \lesssim \frac{\log n}{n^{1/2}} \ \text{ and } \ \|\zeta_t\|_2 \lesssim \frac{\log n}{n^{1/3}} \tag{37}$$

  with probability at least $1 - O(n^{-10})$. As far as we know, this result offers the first non-asymptotic theory of the AMP estimator tailored to sparse regression when $t \gtrsim \frac{\log n}{\log\log n}$, which significantly improves upon the best non-asymptotic prior result Rush and Venkataramanan (2018) that was only able to accommodate $o\big(\frac{\log n}{\log\log n}\big)$ iterations.

- *Improved non-asymptotic risk of the optimally-tuned Lasso.* Given the intimate connection between the aforementioned AMP procedure and Lasso — particularly the one with the regularization parameter carefully chosen to minimize the mean square estimation error — we can immediately see that our result offers a non-asymptotic distributional theory for the optimally-tuned Lasso. Note that the best-known distributional theory for the Lasso has been established by Miolane and Montanari (2021); Celentano et al. (2023c) using the Gaussian min-max theorm; more concretely, Celentano et al. (2023c, Theorem 5) asserts that

$$\Big|\|\widehat{\theta}^{\mathsf{Lasso}} - \theta^\star\|_2 - \gamma^\star\Big| = O\left(\frac{\log n}{n^{1/4}}\right),$$

where $\widehat{\theta}^{\mathsf{Lasso}}$ denotes the solution of the optimally-tuned Lasso. Our result indicates that better rates can be obtained with sparse estimators produced by the AMP algorithm. In particular, taking $t = c_t \log n$ for some large enough constant $c_t > 0$ reveals that

$$\left\|\mathsf{ST}_{\tau_t}(\theta_t) - \theta^\star\right\|_2 = \|F_t(\beta_t)\|_2 = \|\gamma_t\|_2 = \gamma_t^\star + O\Big(\frac{\log n}{n^{1/3}}\Big) = \gamma^\star + O\Big(\frac{\log n}{n^{1/3}}\Big), \tag{38}$$

where we once again invoke the property that $\gamma_t^\star$ converges exponentially fast to $\gamma^\star$ (see discussion in Section B.1.3). It is worth emphasizing that such fine-grained results were unavailable in prior results that used the state-evolution-based analysis of AMP.

2.3 Non-asymptotic AMP theory for robust regression

Next, let us turn to robust regression, which concerns the linear model (1) with the noise being a mixture of Gaussians and some contamination distribution H, i.e.,

$$\varepsilon_i \overset{\text{i.i.d.}}{\sim} (1-\epsilon_H)\mathcal{N}(0,\sigma^2) + \epsilon_H H, \qquad 1 \le i \le n. \tag{39}$$

We shall assume throughout that

$$\sigma^2 \asymp 1/n. \tag{40}$$

Robust regression was originally proposed by Huber (1973) and subsequently developed by Bickel (1975) and many others. The focus therein was on the case where the signal dimension p is much smaller than the sample size n (see, e.g., Hampel (1974); Maronna et al. (2019); Rousseeuw and Leroy (2005); Fan et al. (2014); Loh and Wainwright (2013); Loh (2017); Sun et al. (2020) and the references therein). The modern high-dimensional setting — where the number of variables p is comparable to the sample size n — has been recently explored by El Karoui et al. (2013); Donoho and Montanari (2016, 2015); El Karoui (2018); Lei et al. (2018); Thrampoulidis et al. (2018); Bellec et al. (2022); Bellec and Koriyama (2023); Adomaityte et al. (2023).

AMP for robust regression. A common estimator for robust regression is called the robust M-estimator, which selects a non-negative convex loss function $\rho : \mathbb{R} \to \mathbb{R}_{\ge 0}$ and solves the following optimization problem:

$$\widehat{\theta} \coloneqq \arg\min_{\theta \in \mathbb{R}^p} \mathcal{L}(\theta; y, X), \qquad \text{where } \mathcal{L}(\theta; y, X) \coloneqq \sum_{i=1}^n \rho\big(y_i - \langle X_i, \theta\rangle\big). \tag{41}$$

In this subsection, we focus on the Huber loss as follows

$$\rho(z) = \rho_{\mathsf{huber}}(z,\lambda) \coloneqq \begin{cases} z^2/2, & \text{if } |z| \le \lambda \\ \lambda|z| - \lambda^2/2, & \text{otherwise} \end{cases} \tag{42}$$

for some prescribed threshold $\lambda > 0$ (chosen such that $\lambda \asymp 1/\sqrt{n} \asymp \sigma$), which is arguably the most popular choice to tackle robust regression.

In an attempt to compute and quantify the risk of the above robust M-estimator, one can resort to the AMP algorithm (4) with the following denoising functions (Donoho and Montanari, 2016)

$$f_t(x) = x \qquad \text{and} \qquad g_t(x) = \frac{n}{p}\Psi(z, b_t), \tag{43a}$$

where we remind the reader that

$$\Psi(z,b) = \rho_b'(z) \qquad \text{with} \;\; \rho_b(z) \coloneqq \min_x \Big\{\rho(x) + \frac{1}{2b}(x-z)^2\Big\}$$

for some regularization parameter $b > 0$. When ρ corresponds to the Huber loss (42), it is easily seen that

$$\Psi(z,b) = b\psi\Big(\frac{z}{1+b}, \lambda\Big) \qquad \text{with} \;\; \psi(z;\lambda) = \min\big\{\max\{z, -\lambda\}, \lambda\big\}.$$

Additionally, the algorithm is initialized at $\theta_1 = 0$, $r_0 = 0$ and $r_1 = y$, and the parameter b_t is chosen such that

$$\frac{1}{n}\sum_{i=1}^{n}\Psi'(r_i^t, b_t) = \frac{p}{n}, \qquad \text{or equivalently,} \qquad \big\langle g_t'(r_t)\big\rangle = 1, \tag{44}$$

where $\Psi'(\cdot,\cdot)$ denotes differentiation w.r.t. the first variable.

Remark 6. *When it comes to the equivalent representation (20), one can choose*

$$F_t(\beta) = -\beta \qquad \textit{and} \qquad G_t(s) = g_t(s+\varepsilon) = \frac{n}{p} b_t \psi\left(\frac{s+\varepsilon}{1+b_t};\lambda\right). \tag{45}$$

State evolution for robust regression. In order to predict the dynamics of the AMP algorithm, it is helpful to introduce the following state evolution recursion as introduced in Donoho and Montanari (2016). Specifically, for any $t \geq 1$, Donoho and Montanari (2016) defines a sequence of $(\alpha_t^\star, \gamma_{t+1}^\star) \in \mathbb{R}^2$ recursively as follows:

$$\alpha_t^{\star 2} = \left(\frac{n b_t^\star}{p(1+b_t^\star)}\right)^2 \mathbb{E}\Big[\big\|\psi\big(\varepsilon + \gamma_t^\star g; \lambda(1+b_t^\star)\big)\big\|_2^2\Big], \tag{46a}$$

$$\gamma_{t+1}^{\star 2} = \frac{p}{n}\alpha_t^{\star 2}, \tag{46b}$$

where $\gamma_1^\star = \|\theta^\star\|_2$ and $b_t^\star$ is chosen to satisfy

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big[\Psi'(\varepsilon_i + \gamma_t^\star g_i, b_t^\star)\big] = \frac{p}{n} \qquad \text{with } g_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\Big(0, \frac{1}{n}\Big). \tag{46c}$$

It is worthwhile to remark that the sequence $(\alpha_t^\star, \gamma_{t+1}^\star) \in \mathbb{R}^2$ does not depend on the actual iterates of the AMP procedure.

Non-asymptotic analysis for robust regression. We are now ready to present our non-asymptotic theory for AMP in the context of robust regression. With the aid of our general decomposition in Theorem 1, we can establish the following non-asymptotic theoretical guarantees.

Theorem 3. *Suppose that the signal $\theta^\star$ and the noise satisfy*

$$\|\theta^\star\|_2 \asymp \|\varepsilon\|_2 \asymp 1 \tag{47a}$$

with probability at least $1 - O(n^{-10})$. Assume that

$$n > p \asymp n, \qquad \lambda \asymp 1/\sqrt{n} \qquad \textit{and} \qquad \sigma^2 \asymp 1/n. \tag{47b}$$

Consider the denoising functions chosen as in (45). Then with probability exceeding $1 - O(n^{-10})$, the AMP iterates (4) with denoising functions (43) and threshold (44) admit the decomposition

$$\theta_{t+1} - \theta^\star = \sum_{k=1}^{t}\alpha_t^k \psi_k + \zeta_t \qquad \textit{and} \qquad r_t - \varepsilon = \sum_{k=1}^{t}\gamma_t^k \phi_k + \xi_t \tag{48}$$

for every $t \lesssim \frac{n}{\log^4 n}$, where $\{\psi_j\}_{1\leq j\leq t}$ (resp. $\{\phi_j\}_{1\leq j\leq t}$) are independent Gaussian vectors drawn from $\mathcal{N}\big(0, \frac{1}{n}I_p\big)$ (resp. $\mathcal{N}\big(0, \frac{1}{n}I_n\big)$), the coefficient vectors $\alpha_t = [\alpha_t^j]_{1\leq j\leq t}$ and $\gamma_t = [\gamma_t^j]_{1\leq j\leq t}$ obey

$$\big|\|\alpha_t\|_2^2 - \alpha_t^{\star 2}\big| \lesssim \left(\frac{t\log^2 n}{n}\right)^{1/3} \qquad \textit{and} \qquad \big|\|\gamma_t\|_2^2 - \gamma_t^{\star 2}\big| \lesssim \left(\frac{t\log^2 n}{n}\right)^{1/3}, \tag{49a}$$

and the residuals $\{\xi_t\}$ and $\{\zeta_t\}$ satisfy

$$\|\xi_t\|_2, \|\zeta_t\|_2 \lesssim \left(\frac{t\log^2 n}{n}\right)^{\frac{1}{3}}. \tag{49b}$$

The proof of this result is provided in Section B.2.

In words, Theorem 3 characterizes the distribution of AMP updates with finite-sample guarantees. Akin to the sparse regression case, if the number of iterations is taken to be $t = O(\log n)$, then one can write

$$\theta_{t+1} = \theta^\star + v_{t+1} + \zeta_t \qquad \text{with } W_1\left(\mu_{v_{t+1}}, \mathcal{N}\left(0, \frac{(\alpha_t^\star)^2}{n} I_p\right)\right) \lesssim \frac{\log n}{n^{1/2}} \quad \text{and} \quad \|\zeta_t\|_2 \lesssim \frac{\log n}{n^{1/3}} \tag{50}$$

with high probability. In the meantime, with probability exceeding $1 - O(n^{-10})$ one has

$$\|\theta_{t+1} - \theta^\star\|_2 = \|F_t(\beta_t)\|_2 = \|\gamma_t\|_2 = \gamma_t^\star + O\Big(\frac{\log n}{n^{1/3}}\Big). \tag{51}$$

Evidently, these results recover Donoho and Montanari (2016, Theorem 1.2) in the high-dimensional asymptotics with $n, p \to \infty$ for any fixed t, while at the same time improving upon Donoho and Montanari (2016) by offering non-asymptotic distributional guarantees that account for up to a polynomial number of iterations.

Before concluding this section, we note that the performance of the robust M-estimator in the high dimensional asymptotics has been studied in Donoho and Montanari (2016, 2015) with the aid of the AMP machinery. Certain regularized variants of the robust M-estimator have also been analyzed by means of the leave-one-out analysis (El Karoui, 2013, 2018) and convex Gaussian min-max (CGMT) theorem (Thrampoulidis et al., 2018; Han and Shen, 2023). The only result that offers explicit non-asymptotic guarantees is provided in (Han and Shen, 2023), which leverages the CGMT technique to control the finite-sample error bound to be the order of $O(n^{-1/500})$. In contrast, the AMP analysis in Theorem 3 offers finite-sample error bound on the order of $O\big(\frac{\log n}{n^{1/3}}\big)$. It is worth noting, however, that Han and Shen (2023) is able to extend beyond i.i.d. Gaussian design and unveil interesting universality phenomena.

3 Key technical innovation: controlling the residuals

It is worth noting that the decomposition (22) in Theorem 1, while being fully non-asymptotic and general, has not yet offered quantitative descriptions about the magnitudes of the residual terms ξ_t and ζ_t, making it insufficient to imply any distributional guarantees of the AMP iterates. The key innovation of the current paper thus lies in establishing effective control of ξ_t and ζ_t. In comparison, the approach adopted in our prior work Li and Wei (2022) was insufficient to accommodate the most challenging SNR regime for sparse and robust regression; more discussions on this can be found at the end of this section.

3.1 A fine-grained decomposition for the residuals

A key ingredient of our analysis is to develop a finer representation of ξ_t (resp. ζ_t) by analyzing its corresponding coefficient on the set $\{G_1(s_1), \dots, G_{t-1}(s_{t-1})\}$ (resp. $\{F_1(\beta_1), \dots, F_t(\beta_t)\}$). Towards this end, we find it helpful to first construct a couple of auxiliary sequences, detailed next.

Auxiliary sequences. We now construct recursively a set of auxiliary iterates $\{\widehat{s}_t\} \subseteq \mathbb{R}^n$ and $\{\widehat{\beta}_t\} \subseteq \mathbb{R}^p$, as well as the coefficient vectors $\widehat{\alpha}_t = [\widehat{\alpha}_t^k]_{1 \le k \le t} \in \mathbb{R}^t$ and $\widehat{\gamma}_t = [\widehat{\gamma}_t^k]_{1 \le k \le t} \in \mathbb{R}^t$. Specifically,

- Let us start with $\widehat{\alpha}_0 = 0$ and $\widehat{s}_1 = u_1$ and $\widehat{\beta}_1 = v_1$.
- For each $t \ge 1$ and $\phi_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \frac{1}{n} I_n)$ and $\psi_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \frac{1}{n} I_p)$,
  - construct the vector $\widehat{\gamma}_t = [\widehat{\gamma}_t^k]_{1 \le k \le t} \in \mathbb{R}^t$

$$\widehat{\gamma}_t^k := \begin{cases} \langle G_t'(\widehat{s}_t) - G_t'(s_t)\rangle + \frac{1}{\|\gamma_t\|_2^2}\Big\langle \sum_{k=1}^t \gamma_t^k \phi_k, G_t(s_t) - G_t(\widehat{s}_t)\Big\rangle & \text{for} \quad k = t \\ \widehat{\alpha}_{t-1}^k \left\langle G_t'\left(\widehat{s}_t\right) \circ G_k'\left(u_k\right)\right\rangle & \text{for} \quad k < t. \end{cases} \tag{52a}$$

  - compute

$$\widehat{\beta}_{t+1} := v_{t+1} + \sum_{k=1}^t \widehat{\gamma}_t^k F_k(v_k), \qquad v_{t+1} := \sum_{k=1}^t \alpha_t^k \psi_k. \tag{52b}$$

- construct the vectors $\widehat{\alpha}_t = [\widehat{\alpha}_t^k]_{1\leq k\leq t} \in \mathbb{R}^t$ such that

$$\widehat{\alpha}_t^k := \begin{cases} \langle F'_{t+1}(\widehat{\beta}_{t+1}) - F'_{t+1}(\beta_{t+1})\rangle + \frac{1}{\|\alpha_t\|_2^2}\left\langle \sum_{k=1}^t \alpha_t^k \psi_k, F_{t+1}(\beta_{t+1}) - F_{t+1}(\widehat{\beta}_{t+1})\right\rangle & \text{for} \quad k = t \\ \widehat{\gamma}_t^{k+1}\left\langle F'_{t+1}(\widehat{\beta}_{t+1}) \circ F'_{k+1}(v_{k+1})\right\rangle & \text{for} \quad k < t. \end{cases} \tag{52c}$$

- compute

$$\widehat{s}_{t+1} := u_{t+1} + \sum_{k=1}^{t} \widehat{\alpha}_t^k G_k(u_k), \qquad u_{t+1} := \sum_{k=1}^{t+1} \gamma_{t+1}^k \phi_k. \tag{52d}$$

Crucially, the auxiliary sequence $(\widehat{s}_t, \widehat{\beta}_{t+1})$ constructed above serves as a good proxy of (s_t, β_{t+1}).

Fine-grained representation. Equipped with the above quantities, we are ready to state the following result, whose proof is deferred to Section A.2.

Theorem 4. *Under the assumptions of Theorem 1, the residual terms in the decomposition (22) can be written as*

$$\xi_{t+1} = \sum_{k=1}^{t} \widehat{\alpha}_t^k G_k(s_k) + \widehat{\xi}_{t+1}, \tag{53a}$$

$$\zeta_{t+1} = \sum_{k=1}^{t+1} \widehat{\gamma}_{t+1}^k F_k(\beta_k) + \widehat{\zeta}_{t+1}, \tag{53b}$$

where $\widehat{\xi}_{t+1}$ and $\widehat{\zeta}_{t+1}$ satisfy, with probability at least $1 - O(n^{-10})$, that

$$\begin{aligned} \widehat{\xi}_{t+1} = \sum_{k=1}^{t} a_k \left[\big\langle \psi_k, F_{t+1}(\widehat{\beta}_{t+1})\big\rangle - \big\langle F'_{t+1}(\widehat{\beta}_{t+1})\big\rangle \alpha_t^k - \sum_{j=k+1}^{t} \widehat{\gamma}_t^j \left\langle F'_{t+1}(\widehat{\beta}_{t+1}) \circ F'_j(v_j)\right\rangle \alpha_{j-1}^k \right] \\ + \mathcal{P}^{\perp}_{G_t(s_t)} \sum_{k=1}^{t} a_k \big\langle \psi_k, F_{t+1}(\beta_{t+1}) - F_{t+1}(\widehat{\beta}_{t+1})\big\rangle + \widehat{\xi}_{t+1,\mathsf{res}} \end{aligned} \tag{54a}$$

where

$$\begin{aligned} \widehat{\zeta}_{t+1} = \sum_{k=1}^{t+1} b_k \left[\big\langle \phi_k, G_{t+1}(\widehat{s}_{t+1})\big\rangle - \big\langle G'_{t+1}(\widehat{s}_{t+1})\big\rangle \gamma_{t+1}^k - \sum_{j=k}^{t} \widehat{\alpha}_t^j \left\langle G'_{t+1}(\widehat{s}_{t+1}) \circ G'_j(u_j)\right\rangle \gamma_j^k \right] \\ + \mathcal{P}^{\perp}_{F_{t+1}(\beta_{t+1})} \sum_{k=1}^{t+1} b_k \big\langle \phi_k, G_{t+1}(s_{t+1}) - G_{t+1}(\widehat{s}_{t+1})\big\rangle + \widehat{\zeta}_{t+1,\mathsf{res}} \end{aligned} \tag{54b}$$

with

$$\big\|\widehat{\xi}_{t+1,\mathsf{res}}\big\|_2 \lesssim \sqrt{\frac{t\log n}{n}}\|\gamma_{t+1}\|_2 \qquad \textit{and} \qquad \big\|\widehat{\zeta}_{t+1,\mathsf{res}}\big\|_2 \lesssim \sqrt{\frac{t\log n}{n}}\|\gamma_{t+1}\|_2. \tag{54c}$$

Here, $\mathcal{P}^{\perp}_w$ denotes the linear projection onto the subspace orthogonal to the vector w, and $\{a_k\}_{k=1}^t$ (resp. $\{b_k\}_{k=1}^t$) represents a set of orthogonal basis (which are made precise in expression (62a) (resp. (62b))).

Let us take a moment to provide some technical interpretations about the usefulness of Theorem 4 in controlling the residual terms $\|\xi_t\|_2$ and $\|\zeta_t\|_2$. The readers who are more interested in seeing direct consequences of this result can move directly to Section 3.2.

- Considering the first term in $\widehat{\xi}_{t+1}$. Given that $\psi_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \frac{1}{n} I_p)$, a little algebra reveals that this term takes the following form

$$X^\top f(X) - \mathsf{div} f(X) \qquad \text{with } \mathsf{div} f := \sum_i \frac{\partial f_i}{\partial x_i}$$

for some function f and some Gaussian random vector X. If we pretend that the function f is statistically independent from X, then the celebrated Stein lemma tells us that this term has zero mean, which provides some intuition why one can expect it to be small. Similar messages continue to hold for the first term of $\widehat{\zeta}_{t+1}$.

- Next, let us take a look at the second term in $\widehat{\xi}_{t+1}$. One important component here is the projection operator $\mathcal{P}^{\perp}_{G_t(s_t)}$, which plays a crucial role in achieving the desired bound. To explain this, note that in order to bound the ℓ_2-norm in on the right-hand side of (54a), one strategy is to look at every unit vector $w \perp G_t(s_t)$ and bound

$$\begin{aligned}
\Big\langle w, \mathcal{P}^{\perp}_{G_t(s_t)} \sum_{k=1}^{t} a_k \big\langle \psi_k, F_{t+1}(\beta_{t+1}) - F_{t+1}(\widehat{\beta}_{t+1}) \big\rangle \Big\rangle &= \sum_{k=1}^{t} \langle w, a_k \rangle \big\langle \psi_k, F_{t+1}(\beta_{t+1}) - F_{t+1}(\widehat{\beta}_{t+1}) \big\rangle \\
&\approx \Big\langle \sum_{k=1}^{t} \langle w, a_k \rangle \psi_k,\ F'_{t+1}(v_{t+1}) \circ (\beta_{t+1} - \widehat{\beta}_{t+1}) \Big\rangle \\
&\le \Big\| \sum_{k=1}^{t} \langle w, a_k \rangle \psi_k \circ F'_{t+1}(v_{t+1}) \Big\|_2 \, \big\| \beta_{t+1} - \widehat{\beta}_{t+1} \big\|_2,
\end{aligned}$$

where the second line makes use of the mean value theorem and the fact that $\beta_{t+1} \approx \widehat{\beta}_{t+1} \approx v_{t+1}$ (rigorous derivations are given around inequality (206)). To see why the above bound is useful, we recall two important facts

$$\begin{aligned}
G_t(s_t) &= \sum_{k=1}^{t} \alpha_t^k a_k \qquad \text{for } \alpha_t^k := \big\langle G_t(s_t), a_k \big\rangle \qquad (1 \le k \le t), \\
v_{t+1} &= \sum_{k=1}^{t} \alpha_t^k \psi_k.
\end{aligned}$$

Recalling that $w \perp G_t(s_t)$ and assuming that w and α_t are all statistically independent from $\{\psi_k\}$, one can easily see that

$$\sum_{k=1}^{t} \langle w, a_k \rangle \psi_k \quad \text{is independent from } v_{t+1}.$$

This independence property plays a crucial role in improving the pre-constant on the bound of $\big\| \sum_{k=1}^{t} \langle w, a_k \rangle \psi_k \circ F'_{t+1}(v_{t+1}) \big\|_2$ (compared to the case when no independence is assumed) thus controlling the speed for which $\|\widehat{\xi}_t\|_2$ grows. All this is enabled by considering the projections to $G_t(s_t)$ and its orthogonal space and treat them separately.

- With the decomposition (53) in mind, a natural strategy to bound $\|\xi_{t+1}\|_2$ (resp. $\|\zeta_{t+1}\|_2$) is to control $\|\widehat{\xi}_{t+1}\|_2$ (resp. $\|\widehat{\zeta}_{t+1}\|_2$) and $\sum_{k=1}^{t} \widehat{\alpha}_t^k G_k(s_k)$ (resp. $\sum_{k=1}^{t} \widehat{\gamma}_t^k F_k(\beta_k)$) separately. Let us now take a moment to discuss the term $\sum_{k=1}^{t} \widehat{\alpha}_t^k G_k(s_k)$ — the message for $\sum_{k=1}^{t} \widehat{\gamma}_t^k F_k(\beta_k)$ is similar. As we shall justify in the analysis, the coefficient $|\widehat{\alpha}_t^k|$ decays exponentially in the sense that

$$|\widehat{\alpha}_t^k| \lesssim (1-c)^{t-k} |\widehat{\alpha}_t^t| \tag{56}$$

for some constant $c > 0$ bounded away from 0. Taking this collectively with (23) and the property $\big\|\alpha_t\big\|_2 \approx \alpha_t^\star \lesssim 1$ thus reveals that

$$\Big\| \sum_{k=1}^{t} \widehat{\alpha}_t^k G_k(s_k) \Big\|_2 \le \sum_{k=1}^{t} |\widehat{\alpha}_t^k| \big\| G_k(s_k) \big\|_2 \lesssim |\widehat{\alpha}_t^t| \sum_{k=1}^{t} (1-c)^{t-k} \big\| \alpha_k \big\|_2$$

$$\asymp |\widehat{\alpha}_t^t| \sum_{k=1}^{t} (1-c)^{t-k} \alpha_k^\star \asymp |\widehat{\alpha}_t^t|.$$

Consequently, the analysis will focus on bounding the size of $\widehat{\alpha}_t^t$.

As it turns out, the above observations play an important role in obtaining an effective control of ξ_t and ζ_{t+1} under some mild conditions, to be detailed in the next subsection.

3.2 Bounding the residuals under some key assumptions

With the decomposition in Theorem 4 in place, we further develop upper bounds on the sizes of ξ_t and ζ_{t+1} under some conditions.

To do so, let us begin with some notation. When the function $F : \mathbb{R} \to \mathbb{R}$ is Lipschitz continuous, we define $\rho_F \geq 1$ to be the smallest constant (larger than or equal to 1) such that[3]

$$|F(x_1) - F(x_2)| \leq \rho_F |x_1 - x_2|, \qquad \text{for all } x_1, x_2.$$

Analogously, we define $\rho_G \geq 1$ for the function G. Additionally, suppose that the functions F and G are both differentiable except at a finite number of points. By defining their corresponding derivatives as F' and G' (except at the non-differentiable points), we can introduce the quantities $\rho_{1,F}$ and $\rho_{1,G}$ to represent respectively the maximum local Lipschitz constants of F' and G' over the set of differentiable points. Armed with the above notation, we can introduce the following assumptions regarding the denoising functions F_t and G_t and the AMP updates.

Assumption 1. *Suppose that for any $t \lesssim \frac{n}{\log^4 n}$, the aforementioned Lipschitz constants satisfy*

$$\rho_{F_t}, \rho_{G_t} \asymp 1, \qquad \textit{and} \qquad \rho_{1,F_t}, \rho_{1,G_t} = 0, \tag{57a}$$

In addition, for any $t \lesssim \frac{n}{\log^4 n}$, suppose that conditional on $\|\xi_t\|_2, \|\zeta_t\|_2 \lesssim 1$, one has the coefficients γ_t and α_t in decomposition (22) satisfy

$$\|\gamma_t\|_2, \|\alpha_t\|_2 \asymp 1 \qquad \textit{and} \qquad \|F_t(0)\|_2, \|G_t(0)\|_2 \lesssim 1 \tag{57b}$$

with probability at least $1 - O(n^{-11})$.

Remark 7. *For readers familiar with the AMP literature, $\|\gamma_t\|_2, \|\alpha_t\|_2$ can be viewed as finite-sample counterparts of the asymptotic state evolution (9). Therefore, the assumption (57b) requires the finite-sample state evolution for the corresponding problem to be somewhat regular, with extreme events occuring only with very low probability. The result in our theorem below might not hold if the AMP path degenerates or explode at some point.*

Assumption 2. *Let $\widetilde{u}_t = \|\gamma_t\|_2 g_1$ and $\widetilde{v}_{t+1} = \|\alpha_t\|_2 g_2$, where $g_1 \sim \mathcal{N}(0, \frac{1}{n} I_n)$ and $g_2 \sim \mathcal{N}(0, \frac{1}{n} I_p)$ are independent with $\|\gamma_t\|_2$ and $\|\alpha_t\|_2$. Suppose that there exists some universal constant $0 < c < 1/2$ such that*

$$\frac{1}{n^2} \mathbb{E}\big[\, \|G_t'(\widetilde{u}_t)\|_2^2 \mid \|\gamma_t\|_2 \big] \mathbb{E}\big[\, \big\|F_{t+1}'(\widetilde{v}_{t+1})\big\|_2^2 \mid \|\alpha_t\|_2 \big] < (1-2c)^2, \tag{58}$$

$$\frac{1}{n^2} \mathbb{E}\big[\, \big\|F_{t+1}'(\widetilde{v}_{t+1})\big\|_2^2 \mid \|\alpha_t\|_2 \big] \mathbb{E}\big[\, \big\|G_{t+1}'(\widetilde{u}_{t+1})\big\|_2^2 \mid \|\gamma_{t+1}\|_2 \big] < (1-2c)^2. \tag{59}$$

Under these two assumptions, we can obtain simple bound on the size of the residual terms in decomposition (22) as follows; the proof is deferred to Section A.3.

Theorem 5. *Suppose that the assumptions of Theorem 1 hold. Under Assumptions 1 and 2, the residual terms in decomposition (22) satisfy, with probability at least $1 - O(n^{-10})$,*

$$\|\xi_t\|_2, \|\zeta_t\|_2 \lesssim \left(\frac{t \log^2 n}{n} \right)^{\frac{1}{3}} \tag{60}$$

for every $1 < t \lesssim n / \log^4 n$.

[3] Note that this definition of ρ_F assumes $\rho_F \geq 1$ primarily for notational simplicity; our result would not change if we do not impose this restriction but simply replace ρ_F with $\rho_F \vee 1$.

Remark 8. *Note that in this theorem (and the proof), we allow F_t and G_t to be either deterministic functions, or some random functions. When they are random functions, we assume the existence of a collection of functions $\{F(\cdot;\tau)\}$ (resp. $\{G(\cdot;b)\}$) parameterized by some constant-dimensional τ (resp. b) with $\|\tau\|_2 \lesssim 1$ (resp. $\|b\|_2 \lesssim 1$) such that $F(\cdot;\tau)$ (resp. $G(\cdot;b)$) is $\mathsf{poly}(n)$-Lipschitz in τ (resp. b). It is then assumed that both $\{F(\cdot;\tau)\}$ and $\{G(\cdot;b)\}$ satisfy analogous assumptions as in Assumptions 1 and 2, and that $F_t = F(\cdot;\tau_t)$ (resp. $G_t = G(\cdot;b_t)$) for some random quantity τ_t (resp. b_t). The fact that the parameters τ and b are constant-dimensional makes it feasible to apply a standard covering-based argument.*

Remark 9. *We would like to point out that the exponents in the probability $1 - O(n^{-10})$ and $1 - O(n^{-11})$ in our assumption can be replaced with $1 - O(n^{-c})$ with an arbitrarily large constant $c > 0$.*

Theorem 5 delivers simple upper bounds on the sizes of $\|\xi_t\|_2$ and $\|\zeta_t\|_2$, as long as the required assumptions on F_t, G_t and the AMP updates can be validated. If these assumptions were satisfied, then taking this result collectively with Theorem 1 would ensure that both s_t and β_{t+1} are well approximated by Gaussian distributions with error terms bounded in size by $O((t\log^2 n/n)^{\frac{1}{3}})$.

Consequently, in order to establish our results for sparse and robust regression in Sections 2.2 and 2.3, everything boils down to verifying these assumptions in the two models of interest. It is worth noting that the applicability of Theorem 5 can potentially extend beyond these two important regression problems.

Comparison with Li and Wei (2022). We now pause to emphasize the technical novelty of this paper compared to the prior work Li and Wei (2022). To begin with, while the general decomposition in Theorem 1 shares similarity with the one adopted in Li and Wei (2022) (although we now need to accommodate non-symmetric random matrices), the approach outlined in Li and Wei (2022, Theorem 2) falls short of obtaining effective control the most challenging sample-limited regime, particularly when the denoising functions lack smoothness. For instance, when addressing the case of a sparse $v^\star$, Li and Wei (2022, Theorem 5) requires the number of observations to exceed $n \gtrsim k \log n$, with k the sparsity of the true signal. This requirement arises because in Li and Wei (2022), each direction of the residual terms is treated equivalently and its ℓ_2 norm is then controlled directly. However, if our goal is to handle the most challenging scenario where k is of the same order of n and p, a more fine-grained control over ξ_t and ζ_t along different directions become imperative. More specifically, the residual term ξ_{t+1} turns out to have a larger degree of growth along the direction $G_t(s_t)$, and therefore, it makes sense to single out this direction and control its corresponding size separately as in Theorem 4. In Theorem 5, we single out the quantities $\frac{1}{n^2}\mathbb{E}\big[\|F_t'\|_2^2\big]\mathbb{E}\big[\|G_t'\|_2^2\big]$ to help control how error terms propagate across iterations; in our specific examples, we have demonstrated that this new approach of controlling residuals allows for more effective bounding of this factor.

4 Discussion

In this paper, we have established a general recipe for understanding the non-asymptotic distributions for the celebrated AMP algorithm, tailored to sparse and robust regression. Our framework decomposes the AMP iterates into Gaussian random vectors and residual terms with explicit expressions that are tractable under some mild conditions. For both sparse and robust regression, our results have provided the first finite-sample distributional guarantees for the AMP iterates that can accommodate up to a polynomial number of iterations, which is in sharp contrast to prior theory that cannot go beyond $o\big(\frac{\log n}{\log\log n}\big)$ iterations. Furthermore, our theory has led to to improved distributional guarantees (i.e., improved error rates) for the optimally-tuned Lasso and the robust M-estimator compared to other existing approaches. The insights offered by our non-asymptotic analysis framework have improved upon prior works based on asymptotic state-evolution-type analysis.

Before concluding this paper, let us highlight several possible directions worthy of future investigation.

- Recall that our results have provided improved bounds for the residual terms; for instance, when $t \asymp \log n$, our theory is able to bound the size of the residual terms by $O(\log n/n^{1/3})$ for both sparse and robust regression. A natural question is concerned with the tightness of this error bound. Our current conjecture is that the sharp bound on the residual terms should be $O(\mathsf{poly}(\log n)/n^{1/2})$; establishing or disproving this conjecture require more delicate analyses that go beyond the present analyses.

- Thus far, our framework confirms the validity of Gaussian approximation of AMP up to $O(n/\mathsf{poly}(\log n))$ iterations. It remains to understand the behavior of AMP as the number of iterations further increases beyond this range. Will the state evolution recursion continue to provide reliable predictions as t continues to grow?
- Finally, there has been a recent surge of interest in understanding the performances of AMP beyond the i.i.d. Gaussian design. Certain *universality* phenomena have been empirically observed and theoretically investigated (Bayati et al., 2015; Chen and Lam, 2021; Wang et al., 2022; Dudeja et al., 2022). For instance, the asymptotic theory for AMP has been extended to accommodate a family of rotationally invariant designs (Fan, 2022; Mondelli and Venkataramanan, 2021; Cademartori and Rush, 2023; Venkataramanan et al., 2021). Whether our results can be further generalized beyond Gaussian designs remains an interesting open question for future studies.

APPENDIX

A Proof for our general results

We present the proofs of Theorem 1, 4 and 5 together in this section and defer other technical details and lemmas to the appendices. On the high level, the proof of Theorem 1 resembles the proof of (Li and Wei, 2022, Theorem 1) and the proofs of Theorem 4 and 5 are based on a crucial higher order decomposition and a fine-grained control the residual terms.

A.1 Proof of Theorem 1

Step 1: constructing a key set of auxiliary sequences. Let us first introduce a sequence of auxiliary vectors/matrices $\{a_k, b_k, X_k\}_{1\le t\le \min\{n,p\}}$ in a recursive fashion as below:

(i) With our design matrix X and the initialization $\{s_1, \beta_1\}$ in place, we define

$$a_1 \coloneqq \frac{G_1(s_1)}{\|G_1(s_1)\|_2} \in \mathbb{R}^n, \qquad b_1 \coloneqq \frac{F_1(\beta_1)}{\|F_1(\beta_1)\|_2} \in \mathbb{R}^p, \qquad \text{and} \qquad X_1 \coloneqq X \in \mathbb{R}^{n\times p}; \tag{61}$$

(ii) For every $2 \le t < \min\{p, n\}$, concatenating the a_k's and b_k's into matrices $U_{t-1} = [a_k]_{1\le k\le t-1} \in \mathbb{R}^{n\times(t-1)}$, $V_{t-1} = [b_k]_{1\le k\le t-1} \in \mathbb{R}^{p\times(t-1)}$, we can further define

$$a_t \coloneqq \frac{\left(I - U_{t-1}U_{t-1}^\top\right) G_t(s_t)}{\left\|\left(I - U_{t-1}U_{t-1}^\top\right) G_t(s_t)\right\|_2}, \tag{62a}$$

$$b_t \coloneqq \frac{\left(I - V_{t-1}V_{t-1}^\top\right) F_t(\beta_t)}{\left\|\left(I - V_{t-1}V_{t-1}^\top\right) F_t(\beta_t)\right\|_2}, \tag{62b}$$

$$X_t \coloneqq \left(I_n - a_{t-1}a_{t-1}^\top\right) X_{t-1} \left(I_p - b_{t-1}b_{t-1}^\top\right), \tag{62c}$$

where the pair (s_t, β_t) is generated by iteration (20).

By virtual of these definitions above, it is easily seen that vectors $\{a_k\}_{1\le k\le \min\{n,p\}}$ form an orthonormal basis and so are $\{b_k\}_{1\le k\le \min\{n,p\}}$. By construction, $G_t(s_t)$ lies in the span of $\{a_1, \dots, a_t\}$ and similarly, $F_t(\beta_t) \in \mathsf{span}\{b_1, \dots, b_t\}$. It is therefore legitimate to write

$$G_t(s_t) = \sum_{k=1}^{t} \alpha_t^k a_k, \qquad \text{for } \alpha_t^k \coloneqq \langle G_t(s_t), a_k\rangle \qquad (1 \le k \le t), \tag{63a}$$

$$F_t(\beta_t) = \sum_{k=1}^{t} \gamma_t^k b_k, \qquad \text{for } \gamma_t^k \coloneqq \langle F_t(\beta_t), b_k\rangle \qquad (1 \le k \le t), \tag{63b}$$

which satisfies

$$\|\gamma_t\|_2 = \|F_t(\beta_t)\|_2 \qquad \text{and} \qquad \|\alpha_t\|_2 = \|G_t(s_t)\|_2.$$

Step 2: deriving distributional properties of $X_k b_k$ and $X_k^\top a_k$. Next, we aim to establish some distributional characterizations of $X_k b_k$ and $X_k^\top a_k$. Towards this end, let us first consider another set of auxiliary vectors defined as below

$$\phi_k = X_k b_k + \sum_{i=1}^{k-1} g_k^i a_i, \tag{64a}$$

$$\psi_k = \left(I - b_k b_k^\top\right) X_k^\top a_k + \sum_{i=1}^{k} q_k^i b_i, \tag{64b}$$

where each g_i^k is i.i.d. generated from $\mathcal{N}(0, \frac{1}{n})$. It turns out that ϕ_k and ψ_k admit clean distributional guarantees summarized in the following lemma.

Lemma 1. *With $\{a_k, b_k, X_k\}_{1\le k\le \min\{n,p\}}$ defined in (62), it obeys*

$$\phi_k \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \frac{1}{n} I_n\right), \qquad \textit{and} \quad \psi_k \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \frac{1}{n} I_p\right),$$

for all $1 \le k \le \min\{n, p\}$.

The proof of this result is postponed to Section F.1. We make note here that the covariance matrices of both ϕ_k and ψ_k are identity matrices with normalized constant $1/n$.

Step 3: establishing two key decompositions as in (22). Let us start by showing relation (22a). First, we find it helpful to express X_1 as

$$X_1 = X_t + \sum_{k=1}^{t-1} (X_k - X_{k+1}) = X_t + \sum_{k=1}^{t-1} \Big[X_k b_k b_k^\top + a_k a_k^\top X_k \left(I - b_k b_k^\top\right) \Big]. \tag{65}$$

For every $t \ge 1$, plugging the expansions (as in (63)) that $F_t(\beta_t) = \sum_{k=1}^{t} \gamma_t^k b_k$ and $G_{t-1}(s_{t-1}) = \sum_{k=1}^{t-1} \alpha_{t-1}^k a_k$ leads to

$$\begin{aligned} s_t &= X_1 F_t(\beta_t) - \langle F_t' \rangle \sum_{k=1}^{t-1} \alpha_{t-1}^k a_k \\ &= \sum_{k=1}^{t} \gamma_t^k X_k b_k + \sum_{k=1}^{t-1} a_k \left[\big\langle \left(I - b_k b_k^\top\right) X_k^\top a_k, F_t(\beta_t) \big\rangle - \langle F_t' \rangle \alpha_{t-1}^k \right], \end{aligned} \tag{66}$$

where the last relation invokes the decomposition (65). Substitution of the definition for ϕ_k and reorganizing terms further yield

$$\begin{aligned} s_t &= \sum_{k=1}^{t} \gamma_t^k \Big(\phi_k - \sum_{i=1}^{k-1} g_i^k a_i \Big) + \sum_{k=1}^{t-1} a_k \left[\Big\langle \psi_k - \sum_{i=1}^{k} q_i^k b_i, F_t(\beta_t) \Big\rangle - \langle F_t' \rangle \alpha_{t-1}^k \right] \\ &= \sum_{k=1}^{t} \gamma_t^k \phi_k + \underbrace{\sum_{k=1}^{t-1} a_k \left[\langle \psi_k, F_t(\beta_t) \rangle - \langle F_t' \rangle \alpha_{t-1}^k - \sum_{i=1}^{k} \gamma_t^i q_i^k - \sum_{i=k+1}^{t} \gamma_t^i g_k^i \right]}_{=:\xi_t}. \end{aligned} \tag{67}$$

As a consequence, we have established (22a) with $\xi_t \in \mathsf{span}\{a_1, \dots, a_{t-1}\}$.

As for property (22b), it is useful to write

$$X_1 = X_t \left(I - b_t b_t^\top\right) + X_t b_t b_t^\top + \sum_{k=1}^{t-1} \left[X_k b_k b_k^\top + a_k a_k^\top X_k \left(I - b_k b_k^\top\right) \right]. \tag{68}$$

Again, invoking the expansions $F_t(\beta_t) = \sum_{k=1}^{t} \gamma_t^k b_k$ and $G_t(s_t) = \sum_{k=1}^{t} \alpha_t^k a_k$ gives

$$\begin{aligned}
\beta_{t+1} &= X_1^\top G_t(s_t) - \langle G_t' \rangle \sum_{k=1}^{t} \gamma_t^k b_k \\
&= \sum_{k=1}^{t} \alpha_t^k \left(I - b_k b_k^\top\right) X_k^\top a_k + \sum_{k=1}^{t} b_k \left[\langle X_k^\top b_k, G_t(s_t)\rangle - \langle G_t' \rangle \gamma_t^k\right] && (69) \\
&= \sum_{k=1}^{t} \alpha_t^k \left(\psi_k - \sum_{i=1}^{k} q_i^k b_i\right) + \sum_{k=1}^{t} b_k \left[\left\langle \phi_k - \sum_{i=1}^{k-1} g_i^k a_i, G_t(s_t)\right\rangle - \langle G_t' \rangle \gamma_t^k\right] \\
&= \sum_{k=1}^{t} \alpha_t^k \psi_k + \underbrace{\sum_{k=1}^{t} b_k \left[\langle \phi_k, G_t(s_t)\rangle - \langle G_t' \rangle \gamma_t^k - \sum_{i=1}^{k-1} \alpha_t^i g_i^k - \sum_{i=k}^{t} \alpha_t^i q_k^i\right]}_{=: \zeta_t}, && (70)
\end{aligned}$$

where the penultimate line uses the definitions of ϕ_k and ψ_k (as of (64)). Therefore, inequality (22b) holds with $\zeta_t \in \mathsf{span}\{b_1, \ldots, b_{t-1}\}$.

A.2 Proof of Theorem 4

We move on to the proof of Theorem 4.

Controlling the residual terms ξ_t and ζ_t. In view of the definition of ξ_t (cf. (67)), let us write

$$\xi_t - \sum_{k=1}^{t-1} a_k \left[\langle \psi_k, F_t(\beta_t)\rangle - \langle F_t'(\beta_t)\rangle \alpha_{t-1}^k\right] = \sum_{k=1}^{t-1} a_k \Big[\sum_{i=1}^{k} \gamma_t^i q_i^k - \sum_{i=k+1}^{t} \gamma_t^i g_k^i\Big].$$

We aim to control the magnitude of the right hand side from above. First, we recall that the q_i^k's and g_k^i's are independently drawn from $\mathcal{N}(0, \frac{1}{n})$ — independently of the randomness in the system, as a means to ensure the distributional characterization in Lemma 1. Towards bounding this quantity, recognizing that $\{a_1, \ldots, a_{t-1}\}$ forms an orthonormal basis, there exist a unit vector $\mu_t = [\mu_t^k]_{1 \le k \le t} \in \mathbb{R}^t$ where

$$\Big\|\sum_{k=1}^{t-1} a_k \Big[\sum_{i=1}^{k} \gamma_t^i q_i^k - \sum_{i=k+1}^{t} \gamma_t^i g_k^i\Big]\Big\|_2 = \sum_{k=1}^{t-1} \mu_t^k \Big(\sum_{i=1}^{k} \gamma_t^i q_i^k - \sum_{i=k+1}^{t} \gamma_t^i g_k^i\Big).$$

Therefore, this quantity can be handled via standard Gaussian concentration inequalities as detailed in (Li and Wei, 2022, Lemma 3). We now state the result directly without repeating its proof. With probability at least $1 - O(n^{-11})$, it satisfies

$$\xi_t = \sum_{k=1}^{t-1} a_k \left[\langle \psi_k, F_t(\beta_t)\rangle - \langle F_t'(\beta_t)\rangle \alpha_{t-1}^k\right] + O\left(\sqrt{\frac{t \log n}{n}} \|\gamma_t\|_2\right). \tag{71}$$

By similar analysis, we are also ensured that with probability at least $1 - O(p^{-11})$,

$$\zeta_t = \sum_{k=1}^{t} b_k \left[\langle \phi_k, G_t(s_t)\rangle - \langle G_t'(s_t)\rangle \gamma_t^k\right] + O\left(\sqrt{\frac{t \log n}{n}} \|\alpha_t\|_2\right) \tag{72}$$

holds true.

Establishing the expansions (53). Let us start with the term ξ_t. For every $t \ge 0$, first recall that

$$\xi_{t+1} = \sum_{k=1}^{t} a_k \left[\langle \psi_k, F_{t+1}(\beta_{t+1})\rangle - \langle F_{t+1}'(\beta_{t+1})\rangle \alpha_t^k\right] + O\left(\sqrt{\frac{t \log n}{n}} \|\gamma_{t+1}\|_2\right)$$

$$= \underbrace{\sum_{k=1}^{t} a_k \left[\langle \psi_k, F_{t+1}(\beta_{t+1}) - F_{t+1}(\widehat{\beta}_{t+1})\rangle\right]}_{=:\mathcal{R}_1}$$

$$+ \sum_{k=1}^{t} a_k \left[\langle \psi_k, F_{t+1}(\widehat{\beta}_{t+1})\rangle - \langle F'_{t+1}(\beta_{t+1})\rangle \alpha_t^k\right] + O\left(\sqrt{\frac{t\log n}{n}}\|\gamma_{t+1}\|_2\right).$$

Intuitively, the magnitude of $\mathcal{R}_1$ is determined by the difference between β_{t+1} and $\widehat{\beta}_{t+1}$. If the expansion (53b) were true, the difference between β_{t+1} and $\widehat{\beta}_{t+1}$ arises from $\widehat{\zeta}_t$ as well as the difference between $\sum_{k=1}^{t} \widehat{\gamma}_t^k F_k(\beta_k)$ versus $\sum_{k=1}^{t} \widehat{\gamma}_t^k F_k(v_k)$.

Next, if we project the term $\mathcal{R}_1$ to the direction that aligns with vector $G_t(s_t)$ and its orthogonal linear space, we end up with decomposition

$$\begin{aligned}
&\mathcal{R}_1 \\
&= \frac{1}{\|\alpha_t\|_2^2} \cdot G_t(s_t)^\top \mathcal{R}_1 \cdot G_t(s_t) + \mathcal{P}_{G_t(s_t)}^{\perp} \mathcal{R}_1 \\
&= G_t(s_t) \cdot \frac{1}{\|\alpha_t\|_2^2} \sum_{k=1}^{t} \alpha_t^k \left[\langle \psi_k, F_{t+1}(\beta_{t+1}) - F_{t+1}(\widehat{\beta}_{t+1})\rangle\right] + \mathcal{P}_{G_t(s_t)}^{\perp} \sum_{k=1}^{t} a_k \left[\left\langle \psi_k, F_{t+1}(\beta_{t+1}) - F_{t+1}(\widehat{\beta}_{t+1})\right\rangle\right],
\end{aligned}$$

where the first equality uses property $\|G_t(s_t)\|_2 = \|\alpha_t\|_2$ and the second equality follows from the expansion (63). In addition, due to the property of the orthogonal projection, we also find

$$\begin{aligned}
&\left\langle G_t(s_t), \mathcal{P}_{G_t(s_t)}^{\perp} \sum_{k=1}^{t} a_k \left[\left\langle \psi_k, F_{t+1}(\beta_{t+1}) - F_{t+1}(\widehat{\beta}_{t+1})\right\rangle\right]\right\rangle \\
&= \left\langle \sum_{k=1}^{t} \alpha_t^k a_k, \mathcal{P}_{G_t(s_t)}^{\perp} \sum_{k=1}^{t} a_k \left[\left\langle \psi_k, F_{t+1}(\beta_{t+1}) - F_{t+1}(\widehat{\beta}_{t+1})\right\rangle\right]\right\rangle = 0. \qquad (73)
\end{aligned}$$

Putting the pieces above together, ξ_{t+1} admits the following expression

$$\begin{aligned}
\xi_{t+1} = G_t(s_t) \cdot \frac{1}{\|\alpha_t\|_2^2} \left\langle \sum_{k=1}^{t} \alpha_t^k \psi_k, F_{t+1}(\beta_{t+1}) - F_{t+1}(\widehat{\beta}_{t+1})\right\rangle + \sum_{k=1}^{t} a_k \left[\left\langle \psi_k, F_{t+1}(\widehat{\beta}_{t+1})\right\rangle - \langle F'_{t+1}(\beta_{t+1})\rangle \alpha_t^k\right] \\
+ \mathcal{P}_{G_t(s_t)}^{\perp} \sum_{k=1}^{t} a_k \left\langle \psi_k, F_{t+1}(\beta_{t+1}) - F_{t+1}(\widehat{\beta}_{t+1})\right\rangle + O\left(\sqrt{\frac{t\log n}{n}}\|\gamma_{t+1}\|_2\right), \qquad (74)
\end{aligned}$$

Contrasting what we have shown above with our target, it is sufficient to consider the second term above, which shall be done as follows.

In the following, we establish the claim by decomposing the second term into two parts, corresponding to the influence of $\widehat{\gamma}_t^k F_k(\beta_k)$'s and the randomness of Onsager term, respectively.

To simply the notation, let us define

$$A_t^k := \left\langle \psi_k, F_{t+1}(\widehat{\beta}_{t+1})\right\rangle - \langle F'_{t+1}(\widehat{\beta}_{t+1})\rangle \alpha_t^k - \sum_{j=k+1}^{t} \widehat{\gamma}_t^j \left\langle F'_{t+1}(\widehat{\beta}_{t+1}) \circ F'_j(v_j)\right\rangle \alpha_{j-1}^k.$$

In view of this piece of notation and for each $i \geq 1$, $\sum_{k=1}^{i} a_k \alpha_i^k = G_i(s_i)$, a little algebra leads to

$$\begin{aligned}
&\sum_{k=1}^{t} a_k \left[\left\langle \psi_k, F_{t+1}(\widehat{\beta}_{t+1})\right\rangle - \langle F'_{t+1}(\beta_{t+1})\rangle \alpha_t^k\right] \\
&= \sum_{k=1}^{t} a_k \left[\langle F'_{t+1}(\widehat{\beta}_{t+1}) - F'_{t+1}(\beta_{t+1})\rangle \alpha_t^k + \left\langle \psi_k, F_{t+1}(\widehat{\beta}_{t+1})\right\rangle - \langle F'_{t+1}(\widehat{\beta}_{t+1})\rangle \alpha_t^k\right]
\end{aligned}$$

$$= \sum_{k=1}^{t} a_k \left[\langle F'_{t+1}(\widehat{\beta}_{t+1}) - F'_{t+1}(\beta_{t+1}) \rangle \alpha_t^k + \sum_{j=k+1}^{t} \widehat{\gamma}_t^j \left\langle F'_{t+1}(\widehat{\beta}_{t+1}) \circ F'_j(v_j) \right\rangle \alpha_{j-1}^k + A_t^k \right]$$

$$= G_t(s_t) \cdot \langle F'_{t+1}(\widehat{\beta}_{t+1}) - F'_{t+1}(\beta_{t+1}) \rangle + \sum_{k=1}^{t-1} \widehat{\alpha}_t^k G_k(s_k) + \sum_{k=1}^{t} a_k A_t^k. \tag{75}$$

Here, the last relation follows from

$$\sum_{k=1}^{t} a_k \sum_{j=k+1}^{t} \widehat{\gamma}_t^j \left\langle F'_{t+1}(\widehat{\beta}_{t+1}) \circ F'_j(v_j) \right\rangle \alpha_{j-1}^k = \sum_{k=1}^{t} \sum_{j=k+1}^{t} \widehat{\alpha}_t^{j-1} \alpha_{j-1}^k a_k$$

$$= \sum_{j=2}^{t} \widehat{\alpha}_t^{j-1} \sum_{k=1}^{j-1} \alpha_{j-1}^k a_k = \sum_{k=1}^{t-1} \widehat{\alpha}_t^k G_k(s_k),$$

where we remind the readers that $\widehat{\alpha}_t^k$ is defined as of expression (52c). Taking (75) collectively with (74) and recognizing the definition of $\widehat{\alpha}_t^t$, we end up with

$$\xi_{t+1} = \widehat{\alpha}_t^t G_t(s_t) + \sum_{k=1}^{t-1} \widehat{\alpha}_t^k G_k(s_k)$$

$$+ \sum_{k=1}^{t} a_k A_t^k + \mathcal{P}^{\perp}_{G_t(s_t)} \sum_{k=1}^{t} a_k \left\langle \psi_k, F_{t+1}(\beta_{t+1}) - F_{t+1}(\widehat{\beta}_{t+1}) \right\rangle + O\left(\sqrt{\frac{t \log n}{n}} \|\gamma_{t+1}\|_2 \right),$$

which validates the expansion (53a) for $t+1$. It is also worth noting that $\widehat{\xi}_{t+1}$ is defined exactly as the sum of the last three terms above.

When it comes to the expansion (53b) at $t+1$, repeating a symmetric argument above leads to the required result. We omit its proof for brevity.

A.3 Proof of Theorem 5

In order to prove this result, let us first state a key auxiliary lemma.

Lemma 2. *Under the decomposition (22) with (53) and Assumption 1, the Claim 1, stated below, holds for $t = 1$ with probability at least $1 - O(n^{-10})$. In addition, with probability at least $1 - O(n^{-10})$, for every*

$$1 < t \lesssim \frac{n}{\log^4 n}, \tag{76}$$

if the Claim 1, Assumption 1 and 2 all hold true for t, then Claim 1 holds for $t+1$.

The proof of this result can be found in Section D.

Claim 1. *There exists universal constant $0 < c < 1$, such that the following set of inequalities hold true*

$$\|\widehat{\xi}_t\|_2 \lesssim \sqrt{\frac{t \log^2 n}{n}} \quad \textit{and} \quad \|\widehat{\zeta}_t\|_2 \lesssim \sqrt{\frac{t \log^2 n}{n}}, \tag{77a}$$

$$\widehat{\alpha}_{t-1}^{t-1} \lesssim \left(\frac{t \log^2 n}{n} \right)^{\frac{1}{3}} \quad \textit{and} \quad \widehat{\gamma}_t^t \lesssim \left(\frac{t \log^2 n}{n} \right)^{\frac{1}{3}}, \tag{77b}$$

$$|\widehat{\alpha}_{t-1}^k| \le \begin{cases} (1-c)^{t-k-1} \left| \widehat{\alpha}_{\frac{t+k-1}{2}}^{\frac{t+k-1}{2}} \right| & \textit{if} \quad t-1-k = 2m, \\ (1-c)^{t-k-2} \rho_F^2 \left| \widehat{\gamma}_{\frac{t+k}{2}}^{\frac{t+k}{2}} \right| & \textit{if} \quad t-1-k = 2m+1, \end{cases} \tag{77c}$$

$$|\widehat{\gamma}_t^k| \le \begin{cases} (1-c)^{\frac{t-k}{2}} \left| \widehat{\gamma}_{\frac{t+k}{2}}^{\frac{t+k}{2}} \right| & \textit{if} \quad t-k = 2m, \\ (1-c)^{\frac{t-k}{2}} \rho_G^2 \left| \widehat{\alpha}_{\frac{t+k-1}{2}}^{\frac{t+k-1}{2}} \right| & \textit{if} \quad t-1-k = 2m. \end{cases} \tag{77d}$$

Based on this lemma, we first make the observation that if Claim 1 holds true at iteration t, we arrive at

$$\|\xi_t\|_2 \le \sum_{k=1}^{t-1} |\widehat{\alpha}_{t-1}^k| \|G_k(s_k)\|_2 + \|\widehat{\xi}_t\|_2 \lesssim \sum_{k=1}^{t-1} |\widehat{\alpha}_{t-1}^k| \cdot \|\alpha_k\|_2 + \sqrt{\frac{t\log^2 n}{n}} \lesssim \left(\frac{t\log^2 n}{n}\right)^{\frac{1}{3}}. \tag{78a}$$

Here the first line invokes the relation that $\|G_t(s_t)\|_2 = \|\alpha_t\|_2$, and the second line uses the inductive assumption (77a) and the geometric decay of $\widehat{\alpha}_t^k$ in expression (77c). Similarly, we can deduce

$$\|\zeta_t\|_2 \lesssim \left(\frac{t\log^2 n}{n}\right)^{\frac{1}{3}}. \tag{78b}$$

Now if Assumptions 1 and 2 hold true over the execution of the AMP iterations, Claim 1 is established by induction, since Lemma 2 validates both the initial condition and the inductive argument for Claim 1.

B Proof for sparse and robust regression

B.1 Proof of Theorem 2

The proof of this result is built upon Theorem 5. To show the residuals satisfy relation (34), it is sufficient to validate Assumptions 1 and 2 over the execution of the AMP iterations. We leave the arguments about the state evolution to Section B.1.3.

B.1.1 Validating Assumption 1

First, we make the direct observations that F_t and G_t defined in (26) satisfy $\rho_{F_t}, \rho_{G_t} = 1$ and $\rho_{1,F_t}, \rho_{1,G_t} = 0$ and $\|F_t(0)\|_2, \|G_t(0)\|_2 \lesssim 1$ given $\|\theta^\star\|_2, \|\varepsilon\|_2 \asymp 1$. Then it is sufficient to verify that with high probability,

$$\|\gamma_t\|_2 \asymp \|\alpha_t\|_2 \asymp 1. \tag{79}$$

Towards this goal, recalling the initial choice where

$$\beta_1 = -\theta^\star \qquad \text{and} \qquad s_1 = Y - \varepsilon,$$

and the norm property (23), we find $\|\gamma_1\|_2 = \|F_1(\beta_1)\|_2 = \|\theta^\star\|_2$. In addition, notice that if τ_t is selected to be ∞, we observe

$$\|r_t\|_2 = \|\varepsilon + X^\top\theta^\star\|_2 \lesssim 1.$$

As τ_t is selected as the one that minimizes $\|r_t\|_2$, it implies that

$$\Big\|\varepsilon + \sum_{j=1}^{t} \gamma_t^j \phi_j + \xi_t\Big\|_2 \lesssim 1. \tag{80}$$

Consequently, regarding α_t, we bound

$$\|\alpha_t\|_2 = \|G_t(s_t)\|_2 = \Big\|\varepsilon + \sum_{j=1}^{t} \gamma_t^j \phi_j + \xi_t\Big\|_2 \lesssim 1. \tag{81}$$

To further control the right hand side above, it is helpful to notice that

$$\Big\|\varepsilon + \sum_{j=1}^{t} \gamma_t^j \phi_j + \xi_t\Big\|_2 = \Big\|\varepsilon + \sum_{j=1}^{t} \gamma_t^j \phi_j\Big\|_2 + O(\|\xi_t\|_2),$$

and with probability at least $1 - O(n^{-10})$,

$$\begin{aligned}\Big\|\varepsilon + \sum_{j=1}^{t} \gamma_t^j \phi_j\Big\|_2^2 &= \|\varepsilon\|_2^2 + \Big\|\sum_{j=1}^{t} \gamma_t^j \phi_j\Big\|_2^2 + 2\varepsilon^\top \sum_{j=1}^{t} \gamma_t^j \phi_j \\ &= \|\varepsilon\|_2^2 + \left(1 + O\Big(\sqrt{\frac{t\log n}{n}}\Big)\right) \|\gamma_t\|_2^2 + O\Big(\sqrt{\frac{t\log n}{n}}\, \|\varepsilon\|_2 \, \|\gamma_t\|_2\Big),\end{aligned} \tag{82}$$

where the last inequality invokes the spectral property as in (172a). Putting everything together, we arrive at

$$\|\alpha_t\|_2 = \sqrt{\|\gamma_t\|_2^2 + \|\varepsilon\|_2^2} + O\left(\sqrt{\frac{t\log n}{n}}(\|\gamma_t\|_2 + \|\varepsilon\|_2) + \|\xi_t\|_2\right). \tag{83}$$

Together with the relation (81) and $\|\varepsilon\|_2 \asymp 1$, the relation above leads to

$$\|\alpha_t\|_2 \asymp 1, \qquad \|\gamma_t\|_2 \lesssim 1. \tag{84}$$

Finally, let us establish a proper lower bound for $\|\gamma_{t+1}\|_2$. Again, as a consequence of the norm relation (23) and the Lipschitz property of soft-thresholding function, we write

$$\begin{aligned}\|\gamma_{t+1}\|_2 = \|F_{t+1}(\beta_{t+1})\|_2 &= \Big\|\theta^\star - \mathsf{ST}_{\tau_{t+1}}\Big(\theta^\star + \sum_{k=1}^{t} \alpha_t^k \psi_k + \zeta_t\Big)\Big\|_2 \\ &= \Big\|\theta^\star - \mathsf{ST}_{\tau_{t+1}}\Big(\theta^\star + \sum_{k=1}^{t} \alpha_t^k \psi_k\Big)\Big\|_2 + O\Big(\|\zeta_t\|_2\Big) \\ &= \mathbb{E}\Big[\big\|\theta^\star - \mathsf{ST}_{\tau_{t+1}}(\theta^\star + \|\alpha_t\|_2 g)\big\|_2 \mid \|\alpha_t\|_2\Big] + O\left(\sqrt{\frac{t\log n}{n}} + \|\zeta_t\|_2\right).\end{aligned} \tag{85}$$

Here, we invoke standard concentration inequality for Lipschitz function of Gaussian random variables (Borell, 1975). To accommodate the randomness in $\alpha_t \in \mathbb{R}^t$, we take a union bound over a covering set of $\mathcal{S}^{t-1}$ of accuracy $\frac{1}{n}$. Putting these ideas together, with probability at least $1 - O(n^{-10})$, it is ensured that

$$\Big\|\theta^\star - \mathsf{ST}_{\tau_t}\Big(\theta^\star + \sum_{j=1}^{t} \alpha_t^j \psi_j\Big)\Big\|_2 - \mathbb{E}\left[\Big\|\theta^\star - \mathsf{ST}_{\tau_{t+1}}(\theta^\star + \|\alpha_t\|_2 g)\Big\|_2 \mid \|\alpha_t\|_2\right] = O\Big(\sqrt{\frac{t\log n}{n}}\Big). \tag{86}$$

Hence, it suffices to bound the right hand side of (85) from below which shall be down as follows. Towards this goal, for $\mu \asymp 1$, independent of g, if we define event

$$\mathcal{E} := \left\{ g \sim \mathcal{N}(0, \frac{1}{n} I_p) \mid \big\|\theta^\star - \mathsf{ST}_{\tau_{t+1}}(\theta^\star + \mu g)\big\|_2 = \mathbb{E}\left[\big\|\theta^\star - \mathsf{ST}_{\tau_{t+1}}(\theta^\star + \mu g)\big\|_2 \mid \mu\right] + O\left(\sqrt{\frac{\log n}{n}}\right)\right\}, \tag{87}$$

as discussed above, the event $\mathcal{E}$ happens with probability at least $1 - O(n^{-10})$. In view of this set, we write

$$\begin{aligned}\mathbb{E}\Big[\big\|\theta^\star - \mathsf{ST}_{\tau_{t+1}}(\theta^\star + \mu g)\big\|_2^2\Big] &\overset{\text{(i)}}{=} \mathbb{P}(\mathcal{E})\left(\mathbb{E}\left[\big\|\theta^\star - \mathsf{ST}_{\tau_{t+1}}(\theta^\star + \mu g)\big\|_2\right] + O\left(\sqrt{\frac{\log n}{n}}\right)\right)^2 \\ &\quad + O\left(\mathbb{E}\left[(1 + \|g\|_2^2)\,\mathbb{1}(\mathcal{E}^{\mathrm{c}})\right]\right) \\ &= \left(1 - O\left(n^{-10}\right)\right)\left(\mathbb{E}\left[\big\|\theta^\star - \mathsf{ST}_{\tau_{t+1}}(\theta^\star + \mu g)\big\|_2\right] + O\left(\sqrt{\frac{\log n}{n}}\right)\right)^2 \\ &\quad + O\left(\mathbb{P}(\mathcal{E}^{\mathrm{c}})\log n + \mathbb{E}\left[(1 + \|g\|_2^2)\,\mathbb{1}(\|g\|_2^2 \gtrsim \log n)\right]\right),\end{aligned}$$

$$\overset{\text{(ii)}}{=} \left(\mathbb{E}\left[\left\|\theta^\star - \mathsf{ST}_{\tau_{t+1}}(\theta^\star + \mu g)\right\|_2\right]\right)^2 + O\left(\sqrt{\frac{\log n}{n}}\right), \tag{88}$$

Here (i) uses the fact that

$$\left\|\theta^\star - \mathsf{ST}_{\tau_{t+1}}(\theta^\star + \mu g)\right\|_2^2 \lesssim \|\theta^\star\|^2 + \|\theta^\star + \mu g\|^2 \lesssim 1 + \|g\|_2^2,$$

by recognizing $\|\theta^\star\|_2 \asymp$ and $\mu \asymp 1$; (ii) invokes the basic relation for Gaussian random variable where $\mathbb{E}[\|g\|_2^2 \mathbb{1}(\|g\|_2^2 \gtrsim \log n)] \lesssim \sqrt{\frac{\log n}{n}}$ for $g \sim \mathcal{N}(0, \frac{1}{n} I_p)$.

Let us proceed to controlling the size of $\mathbb{E}[\left\|\theta^\star - \mathsf{ST}_{\tau_{t+1}}(\theta^\star + \|\alpha_t\|_2 g)\right\|_2^2]$ which in turn, provides the control of quantity $\mathbb{E}\left\|\theta^\star - \mathsf{ST}_{\tau_{t+1}}(\theta^\star + \|\alpha_t\|_2 g)\right\|_2$. We claim that

$$\mathbb{E}\Big[\left\|\theta^\star - \mathsf{ST}_{\tau_{t+1}}(\theta^\star + \|\alpha_t\|_2 g)\right\|_2^2\Big] \gtrsim 1. \tag{89}$$

In the following, we prove the above claim by diving into two different cases and considering them separately.

- First consider the case when τ_{t+1} satisfies

$$\tau_{t+1} < \frac{\|\theta^\star\|_1}{4k} \le \frac{\sqrt{k}\|\theta^\star\|_2}{k} \asymp \frac{1}{\sqrt{n}}, \tag{90}$$

  for $\|\theta^\star\|_1$ obeying (33) and $n > 2k\log(p/k)$. In this case, we find

$$\begin{aligned}\mathbb{E}\Big[\left\|\theta^\star - \mathsf{ST}_{\tau_{t+1}}(\theta^\star + \|\alpha_t\|_2 g)\right\|_2^2\Big] &\ge \mathbb{E}\Big[\left\|\mathsf{ST}_{\tau_{t+1}}(\|\alpha_t\|_2 g) \circ \mathbb{1}(\theta^\star = 0)\right\|_2^2\Big] \\ &\gtrsim \mathbb{E}\Big[\Big\||g|\, \mathbb{1}(|g| \gtrsim 2\tau_{t+1})\Big\|_2^2\Big] \gtrsim 1.\end{aligned}$$

- On the other hand, when the relation (90) is violated, we make the observation that

$$\|\theta^\star\|_1 \le 2\tau_{t+1}k + \|\theta^\star\|_2\sqrt{\|\,\mathbb{1}(|\theta^\star| \ge 2\tau_{t+1})\|_0},$$

  which, together with (33), gives

$$\|\,\mathbb{1}(|\theta^\star| \ge 2\tau_{t+1})\|_0 \gtrsim k.$$

  Based on this property, we write

$$\mathbb{E}\Big[\left\|\theta^\star - \mathsf{ST}_{\tau_{t+1}}(\theta^\star + \|\alpha_t\|_2 g)\right\|_2^2\Big] \ge \tau_{t+1}^2 \mathbb{E}\Big[\left\|\mathbb{1}(|\theta^\star| \ge 2\tau_{t+1}) \circ \mathbb{1}(\|\alpha_t\|_2 |g| < \tau_{t+1})\right\|_2^2\Big] \gtrsim 1,$$

  where in the last inequality, we make the observation that $\tau_{t+1} > \|\theta^\star\|_1/4k \gtrsim \frac{1}{\sqrt{k}}$ and $\mathbb{P}(\mathbb{1}(\|\alpha_t\|_2 |g_i| < \tau_{t+1})) = O(1)$.

Combining (85), (88) and (89) leads to

$$\|\gamma_{t+1}\|_2 \gtrsim 1, \qquad \text{for } t \lesssim \frac{n}{\log n}. \tag{91}$$

Taking this together with (84), we have completed the proof of (79) and thus justified Assumption 1.

B.1.2 Validating Assumption 2

It is easily seen that $\frac{1}{n}\mathbb{E}\left\|G_t'(u_t)\right\|_2^2 = 1$. Therefore validating Assumption 2 is equivalent to validating

$$\frac{1}{n}\mathbb{E}\left\|F_{t+1}'(v_{t+1})\right\|_2^2 = \frac{1}{n}\mathbb{E}\left\|\mathbb{1}\left(|\theta^\star + \|\alpha_t\|_2 g| \ge \tau_{t+1}\right)\right\|_0 < (1-2c)^2, \tag{92}$$

for some constant $0 < c < 1/2$. To establish this result, we find it helpful to first make the observation that for some small constant $c' > 0$

$$\mathcal{S}' := \Big\{\tau : -c'\sqrt{k} < \nabla_\tau \mathbb{E}\,\|\theta^\star - \mathsf{ST}_\tau(\theta^\star + \|\alpha_t\|_2 g)\|_2^2 < c'\sqrt{k}\Big\} \subset \mathcal{S} \tag{93}$$

with set

$$\mathcal{S} := \Big\{\tau : \mathbb{E}\Big[\,\|\mathbb{1}\left(|(\theta^\star + \|\alpha_t\|_2 g)| \ge \tau\right)\|_0\Big] < (1-2c)^2 n\Big\}.$$

Let us take the relation (93) as given for the moment, and come back to its proof at the end of this section. Based on this result, we shall prove Claim (92) by showing that

$$\tau_{t+1} \in \mathcal{S}'. \tag{94}$$

Proof of Claim (94). We first prove that, if $\tau_{t+1} \notin \mathcal{S}'$, it must satisfy

$$\inf_{\tau \in \mathcal{S}''} |\tau_{t+1} - \tau| > c''\frac{1}{\sqrt{n}}, \tag{95}$$

for some constant $c'' > 0$. Here let us define an auxiliary set

$$\mathcal{S}'' := \left\{\tau : -\frac{c'}{2}\sqrt{k} < \nabla_\tau \mathbb{E}\,\|\theta^\star - \mathsf{ST}_\tau(\theta^\star + \|\alpha_t\|_2 g)\|_2^2 < \frac{c'}{2}\sqrt{k}\right\}. \tag{96}$$

In order to see this, observe that

$$\begin{aligned}
\nabla_\tau \mathbb{E}\,\|\theta^\star - \mathsf{ST}_\tau(\theta^\star + \|\alpha_t\|_2 g)\|_2^2 &= \mathbb{E}\,\langle \theta^\star - \mathsf{ST}_\tau(\theta^\star + \|\alpha_t\|_2 g), \mathrm{sign}(\theta^\star + \|\alpha_t\|_2 g)\,\mathbb{1}(|\theta^\star + \|\alpha_t\|_2 g| > \tau)\rangle \\
&= \mathbb{E}\Big[\,\langle \tau - \|\alpha_t\|_2 g \mathrm{sign}(\theta^\star + \|\alpha_t\|_2 g), \mathbb{1}(|\theta^\star + \|\alpha_t\|_2 g| > \tau)\rangle\,\Big] \\
&= \sum_{i:\theta_i^\star \neq 0} \mathbb{E}\Big[\,(\tau - \|\alpha_t\|_2 g \mathrm{sign}(\theta_i^\star + \|\alpha_t\|_2 g))\,\mathbb{1}(|\theta_i^\star + \|\alpha_t\|_2 g| > \tau)\Big] \\
&\qquad - (p-k)\mathbb{E}\left[\mathsf{ST}_\tau(|\|\alpha_t\|_2 g|)\right].
\end{aligned} \tag{97}$$

The density function $|p_{\theta_i^\star + \|\alpha_t\|_2 g_i}| \lesssim \sqrt{n}$ for $\|\alpha_t\|_2 \asymp 1$ and $g_i \sim \mathcal{N}(0, 1/n)$, and therefore the right hand is a $O(n)$-Lipschitz function of τ. Consequently, for $\tau \in \mathcal{S}''$, we deduce

$$\frac{c'}{2}\sqrt{k} \le \Big|\nabla_{\tau_{t+1}} \mathbb{E}\,\big\|\theta^\star - \mathsf{ST}_{\tau_{t+1}}(\theta^\star + \|\alpha_t\|_2 g)\big\|_2^2 - \nabla_\tau \mathbb{E}\,\|\theta^\star - \mathsf{ST}_\tau(\theta^\star + \|\alpha_t\|_2 g)\|_2^2\Big| \lesssim n \inf_{\tau \in \mathcal{S}''} |\tau_{t+1} - \tau|,$$

which proves the claimed gap (95) between τ_{t+1} to $\mathcal{S}''$.

Given the relation (95), for $\widehat{\tau} = \inf\{\tau \in \mathcal{S}'', \tau > \tau_{t+1}\}$, this further implies

$$\begin{aligned}
&\mathbb{E}\Big\|\theta^\star - \mathsf{ST}_{\tau_{t+1}}(\theta^\star + \|\alpha_t\|_2 g)\Big\|_2^2 - \inf_\tau \mathbb{E}\Big\|\theta^\star - \mathsf{ST}_\tau(\theta^\star + \|\alpha_t\|_2 g)\Big\|_2^2 \\
&\ge \mathbb{E}\Big\|\theta^\star - \mathsf{ST}_{\tau_{t+1}}(\theta^\star + \|\alpha_t\|_2 g)\Big\|_2^2 - \mathbb{E}\Big\|\theta^\star - \mathsf{ST}_{\widehat{\tau}}(\theta^\star + \|\alpha_t\|_2 g)\Big\|_2^2 \\
&\overset{\text{(i)}}{\gtrsim} c'\sqrt{n}(\widehat{\tau} - \tau_{t+1}) \ge c'\sqrt{n} \cdot \inf_{\tau \in \mathcal{S}''} |\tau_{t+1} - \tau| \gtrsim 1,
\end{aligned} \tag{98}$$

where (i) is a consequence of the mean value theorem. Next we show that since τ_{t+1} is selected to minimize $\|\varepsilon + \sum_{k=1}^{t+1} \gamma_{t+1}^k \phi_k + \xi_t\|_2$, the above relation contradicts with the choice of τ_{t+1}. More specifically, as is shown in (83), (85) and (88), we can write

$$\Big\|\varepsilon + \sum_{k=1}^{t+1} \gamma_{t+1}^k \phi_k + \xi_t\Big\|_2$$

$$
\begin{aligned}
&= \sqrt{\|\gamma_{t+1}\|_2^2 + \|\varepsilon\|_2^2} + O\left(\sqrt{\frac{t\log n}{n}} + \|\xi_{t+1}\|_2\right) \\
&= \sqrt{\mathbb{E}\Big[\big\|\theta^\star - \mathsf{ST}_{\tau_{t+1}}(\theta^\star + \|\alpha_t\|_2 g)\big\|_2^2\Big] + \|\varepsilon\|_2^2 + O\left(\sqrt{\frac{t\log n}{n}} + \|\xi_{t+1}\|_2\right)} + O\left(\sqrt{\frac{t\log n}{n}} + \|\xi_{t+1}\|_2\right).
\end{aligned} \tag{99}
$$

The threshold τ_{t+1} therefore cannot satisfy (98), as otherwise it does not minimize $\|\varepsilon + \sum_{k=1}^{t+1}\gamma_{t+1}^k\phi_k + \xi_t\|_2$, which in turn, validates the claimed relation (92).

Proof of Property (93)**.** It can be seen from numerical calculations (see Figure 1 and the discussions around inequality (122)) that for $G \sim \mathcal{N}(0,1)$, if

$$
\sup_\theta \mathbb{E}\Big[(\omega - G\mathrm{sign}(\theta + G)) \circ \mathbb{1}(|\theta + G| > \tau)\Big] - (p/k - 1)\mathbb{E}\left[\mathsf{ST}_\omega(|G|)\right] \in (-c, c). \tag{100}
$$

we have

$$
1 - \frac{1 + (\frac{p}{k} - 1)\mathbb{P}\left(|G| \geq \omega\right)}{2\log\frac{p}{k}} > 0. \tag{101}
$$

The above result tells us that, for all τ satisfying

$$
\begin{aligned}
\nabla_\tau \mathbb{E}\,\|\theta^\star - \mathsf{ST}_\tau(\theta^\star + \|\alpha_t\|_2 g)\|_2^2 = \sum_{i:\theta_i^\star \neq 0} &\mathbb{E}\left[(\tau - \|\alpha_t\|_2 g\mathrm{sign}(\theta_i^\star + \|\alpha_t\|_2 g))\,\mathbb{1}(|\theta_i^\star + \|\alpha_t\|_2 g| > \tau)\right] \\
&- (p-k)\mathbb{E}\left[\mathsf{ST}_\tau(|\|\alpha_t\|_2 g|)\right] \in (-c'\sqrt{k}, c'\sqrt{k}),
\end{aligned} \tag{102}
$$

which implies that for $\omega := \frac{\sqrt{n}\tau}{\|\alpha_t\|_2}$,

$$
k\sup_\theta \mathbb{E}\left[(\omega - G\mathrm{sign}(\theta + G))\,\mathbb{1}(|\theta + G| > \tau)\right] - (p-k)\mathbb{E}\left[\mathsf{ST}_\omega(|G|)\right] \in (-c'\sqrt{nk}/\|\alpha_t\|_2, c'\sqrt{nk}/\|\alpha_t\|_2),
$$

then we have

$$
\mathbb{P}\left(|\|\alpha_t\|_2 g| \geq \tau\right) = \mathbb{P}\left(|G| \geq \omega\right) \leq \frac{2(1-4c)^2\log\frac{p}{k} - 1}{\frac{p}{k} - 1}, \tag{103}
$$

which establishes the property (93) immediately. In order to see this, plugging in $n > 2k\log\frac{p}{k}$, inequality (103) ensures

$$
\mathbb{P}\left(|\|\alpha_t\|_2 g| \geq \tau\right) < \frac{(1-4c)^2 n - k}{p - k},
$$

and hence,

$$
\mathbb{E}\Big[\left\|\mathbb{1}\left(|(\theta^\star + \|\alpha_t\|_2 g)| \geq \tau\right)\right\|_0\Big] \leq k + (p-k)\mathbb{P}\left(|\|\alpha_t\|_2 g| \geq \tau\right) < (1-4c)^2 n.
$$

Upper bound for τ_{t+1}. Finally, let us establish a property of τ_{t+1}. Specifically, we shall prove that $\tau_{t+1} \lesssim 1/\sqrt{n}$ when $p/k \lesssim 1$. Before proceeding, let us make the following two observations both of which result from direct Gaussian integral. For $G \sim \mathcal{N}(0,1)$, it satisfies

$$
\mathbb{E}\left[\mathsf{ST}_\omega(|G|)\right] = \sqrt{\frac{2}{\pi}}\int_\omega^\infty x\exp\Big(-\frac{x^2}{2}\Big)\mathrm{d}x \lesssim \exp\Big(-\frac{\omega^2}{2}\Big),
$$

and

$$
\mathbb{E}\left[(\omega - G\mathrm{sign}(\theta_i + G)) \circ \mathbb{1}(|\theta_i + G| > \omega)\right]
$$

$$
\begin{aligned}
&= \mathbb{E}\left[(\omega - G) \circ \mathbb{1}(\omega - |\theta_i| < G < \omega + |\theta_i|)\right] + 2\mathbb{E}\left[(\omega - G) \circ \mathbb{1}(G > \omega + |\theta_i|)\right] \\
&\geq \frac{1}{\sqrt{2\pi}} \int_0^{|\theta_i|} x \exp\Big(-\frac{(x-\omega)^2}{2}\Big)\mathrm{d}x - \sqrt{\frac{2}{\pi}} \int_0^{\infty} x \exp\Big(-\frac{(x+\omega)^2}{2}\Big)\mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{\omega^2}{2}\Big)\Big\{\int_0^{|\theta_i|} x \exp(\omega x) \exp\Big(-\frac{x^2}{2}\Big)\mathrm{d}x - 2\int_0^{\infty} x \exp(-\omega x) \exp\Big(-\frac{x^2}{2}\Big)\mathrm{d}x\Big\} \\
&\geq \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{\omega^2}{2}\Big)\Big\{\int_0^{|\theta_i|} \frac{\omega^2 x^3}{2} \exp\Big(-\frac{x^2}{2}\Big)\mathrm{d}x - 2\int_0^{\infty} x \exp\Big(-\frac{x^2}{2}\Big)\mathrm{d}x\Big\} \\
&\geq \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{\omega^2}{2}\Big)\Big[\omega^2\Big(1 - \Big(\frac{\theta_i^2}{2} + 1\Big)\exp\Big(-\frac{\theta_i^2}{2}\Big)\Big) - 2\Big].
\end{aligned}
$$

Based on these two observations, for $\|\theta\|_1 \gtrsim \sqrt{k}$ and ω large enough, it obeys

$$
\frac{1}{k}\sum_{i:\theta_i \neq 0} \mathbb{E}\left[(\omega - G\mathrm{sign}(\theta_i + G)) \circ \mathbb{1}(|\theta_i + G| > \omega)\right] \gtrsim \omega^2 \exp\Big(-\frac{\omega^2}{2}\Big). \tag{104}
$$

Now, considering the transformation $\omega = \sqrt{n}\tau/\|\alpha_t\|_2$ and $\theta_i = \theta_i^\star/\|\alpha_t\|_2$, we obtain

$$
\begin{aligned}
&\nabla_\tau \mathbb{E}\left\|\theta^\star - \mathsf{ST}_\tau(\theta^\star + \|\alpha_t\|_2 g)\right\|_2^2 \\
&\quad= \sum_{i:\theta_i^\star \neq 0} \mathbb{E}\left[(\tau - \|\alpha_t\|_2 g\mathrm{sign}(\theta_i^\star + \|\alpha_t\|_2 g))\, \mathbb{1}(|\theta_i^\star + \|\alpha_t\|_2 g| > \tau)\right] - (p-k)\mathbb{E}\left[\mathsf{ST}_\tau(|\|\alpha_t\|_2 g|)\right] \\
&\quad= \frac{\|\alpha_t\|_2}{\sqrt{n}} \sum_{i:\theta_i \neq 0} \mathbb{E}\left[(\omega - G\mathrm{sign}(\theta_i + G)) \circ \mathbb{1}(|\theta_i + G| > \omega)\right] - \|\alpha_t\|_2 \frac{p-k}{\sqrt{n}} \mathbb{E}\left[\mathsf{ST}_\omega(|G|)\right].
\end{aligned}
$$

Taking the above together with (104), we have that given $\|\alpha_t\|_2 \asymp 1$ and $\tau > C/\sqrt{n}$ for some constant C large enough,

$$
\nabla_\tau \mathbb{E}\left\|\theta^\star - \mathsf{ST}_\tau(\theta^\star + \|\alpha_t\|_2 g)\right\|_2^2 > c\sqrt{k},
$$

for some constant $c > 0$. In view of the property (94), this property ensures that $\tau_{t+1} \lesssim 1/\sqrt{n}$. As a result, it also leads to

$$
\omega := \sqrt{n}\tau_{t+1}/\|\alpha_t\|_2 \lesssim 1, \qquad \mathbb{E}\left[\mathsf{ST}_\omega(|G|)\right] \gtrsim 1. \tag{105}
$$

B.1.3 State evolution

Our final goal is to bound the difference between the non-asymptotic SE (α_t, γ_{t+1}) to the deterministic SE defined in expression (31). In the following, we shall use the induction method to achieve this goal. In particular, it is easily validated that the set of relation (35a) holds true for $t = 1$. Assuming that for some $t \geq 1$,

$$
\left|\|\alpha_{t-1}\|_2^2 - \alpha_{t-1}^{\star 2}\right| \lesssim \Big(\frac{t\log^2 n}{n}\Big)^{1/3} \qquad \text{and} \qquad \left|\|\gamma_t\|_2^2 - \gamma_t^{\star 2}\right| \lesssim \Big(\frac{t\log^2 n}{n}\Big)^{1/3}, \tag{106}
$$

it is thus sufficient to verify them for $t + 1$.

Towards this end, let us first recall the expression (83) that

$$
\|\alpha_t\|_2^2 = \|\gamma_t\|_2^2 + \|\varepsilon\|_2^2 + O\left(\Big(\frac{t\log^2 n}{n}\Big)^{\frac{1}{3}}\right), \tag{107a}
$$

where $\|\gamma_t\|_2, \|\varepsilon\|_2 \asymp 1$. In addition, combining expressions (85) and (88) yields

$$
\|\gamma_{t+1}\|_2 = \mathbb{E}\left[\left\|\theta^\star - \mathsf{ST}_{\tau_{t+1}}(\theta^\star + \|\alpha_t\|_2 g)\right\|_2 \mid \|\alpha_t\|_2\right] + O\left(\Big(\frac{t\log^2 n}{n}\Big)^{\frac{1}{3}}\right)
$$

$$\overset{\text{(i)}}{\leq} \mathbb{E}\Big[\Big\|\theta^\star - \mathsf{ST}_{\tau_{t+1}}(\theta^\star + \sqrt{\|\gamma_t\|_2^2 + \|\varepsilon\|_2^2}g)\Big\|_2 \mid \|\alpha_t\|_2\Big] + O\Big(\Big(\frac{t\log^2 n}{n}\Big)^{\frac{1}{3}}\Big)$$

$$\overset{\text{(ii)}}{\leq} \Big(\mathbb{E}\Big[\big\|\theta^\star - \mathsf{ST}_{\tau_{t+1}}(\theta^\star + \sqrt{\|\gamma_t\|_2^2 + \|\varepsilon\|_2^2}g)\big\|_2^2 \mid \|\alpha_t\|_2\Big]\Big)^{1/2} + O\Big(\Big(\frac{t\log^2 n}{n}\Big)^{\frac{1}{3}}\Big). \tag{107b}$$

Here, for inequality (i), we have plugged in the relationship (107a) and invoked the Lipschitz property

$$\|\theta^\star - \mathsf{ST}_{\tau_{t+1}}(\theta^\star + (\omega + \Delta)g\|_2 \leq \|\theta^\star - \mathsf{ST}_{\tau_{t+1}}(\theta^\star + \omega g)\|_2 + \|\Delta g\|_2.$$

We remind the readers that vector $g \in \mathcal{N}(0, \frac{1}{n}I_p)$, and is independent with the (α_t, γ_{t+1}) sequence. For inequality (ii) to hold, we recall the relation (88). According to the optimality of τ_{t+1} (in (31)), we also find

$$\|\gamma_{t+1}\|_2 = \|F_{t+1}(\beta_{t+1})\|_2 = \Big\|\theta^\star - \mathsf{ST}_{\tau_{t+1}}(\theta_{t+1})\Big\|_2 \tag{108}$$

$$\leq \big\|\theta^\star - \mathsf{ST}_{\tau_{t+1}^\star}(\theta_{t+1})\big\|_2 + O\Big(\Big(\frac{t\log^2 n}{n}\Big)^{\frac{1}{3}}\Big)$$

$$\leq \Big(\mathbb{E}\Big[\big\|\theta^\star - \mathsf{ST}_{\tau_{t+1}^\star}(\theta^\star + \sqrt{\|\gamma_t\|_2^2 + \|\varepsilon\|_2^2}g)\big\|_2^2 \mid \|\alpha_t\|_2\Big]\Big)^{1/2} + O\Big(\Big(\frac{t\log^2 n}{n}\Big)^{\frac{1}{3}}\Big), \tag{109}$$

where the last line is derived again by uniform concentration inequalities, similar to relation (107b).

Armed with the recursive formula (107), controlling the difference between the non-asymptotic state evolution and its asymptotic analogue boils down to considering the growth of function

$$h_\tau(\mu) := \mathbb{E}\Big[\big\|\theta^\star - \mathsf{ST}_\tau(\theta^\star + \sqrt{\mu}g)\big\|_2^2\Big], \tag{110}$$

for every value of $\tau > 0$. Direct computations yield

$$\big|h_\tau'(\mu)\big| = \frac{1}{\sqrt{\mu}}\Big|\mathbb{E}\Big[\big\langle\theta^\star - \mathsf{ST}_\tau(\theta^\star + \sqrt{\mu}g), -\mathbb{1}(|\theta^\star + \sqrt{\mu}g| > \tau) \circ g\big\rangle\Big]\Big|. \tag{111}$$

Considering the new rescaling

$$\theta := \sqrt{n}\theta^\star/\sqrt{\mu}, \qquad \text{and} \qquad \omega := \sqrt{n}\tau/\sqrt{\mu}, \tag{112}$$

we can rewrite

$$h_\tau(\mu) := \frac{\mu}{n}\mathbb{E}\Big[\big\|\theta - \mathsf{ST}_\omega(\theta + G)\big\|_2^2\Big].$$

where $G \sim \mathcal{N}(0, 1)$. In terms of the new scaling, for k-sparse $\theta^\star$, some direct calculations lead to

$$\big|h_\tau'(\mu)\big| = \frac{1}{n}\Big|\mathbb{E}\Big[\big\langle\theta - \mathsf{ST}_\omega(\theta + G), \mathbb{1}(|\theta + G| > \omega) \circ G\big\rangle\Big]\Big|$$

$$= \Big|\frac{k}{n}\mathbb{E}\Big[\big(\mathsf{ST}_\omega(\theta + G) - \theta\,\mathbb{1}(|\theta + G| > \omega)\big)G\Big] + \frac{p-k}{n}\mathbb{E}\big[\mathsf{ST}_\omega(G)G\big]\Big|. \tag{113}$$

We claim that there exists constant $c \in (0, 1)$ that only depends on the ratio p/n and k/p such that

$$\big|h_\tau'(\mu)\big| \leq \Big|\frac{1}{n}\sum_{i:\theta_i \neq 0}\mathbb{E}\Big[\big(\mathsf{ST}_\omega(\theta_i + G) - \theta\,\mathbb{1}(|\theta + G| > \omega)\big)G\Big] + \frac{p-k}{n}\mathbb{E}\big[\mathsf{ST}_\omega(G)G\big]\Big| \leq 1 - c, \tag{114}$$

for both $\tau = \tau_{t+1}$ and $\tau = \tau_{t+1}^\star$. Let us take this result as given for the moment and leave the proof of this claim to the end of this section.

Given this result, we can bound the difference

$$\big|\|\gamma_{t+1}\|_2^2 - \gamma_{t+1}^{\star 2}\big| \leq \max_{\tau = \tau_{t+1}, \tau_{t+1}^\star}\Big|h_\tau\Big(\|\gamma_t\|_2^2 + \|\varepsilon\|_2^2\Big) - h_\tau\big(\gamma_t^{\star 2} + \|\varepsilon\|_2^2\big)\Big| + O\Big(\Big(\frac{t\log^2 n}{n}\Big)^{\frac{1}{3}}\Big)$$

$$\leq (1-c)\big|\|\gamma_t\|_2^2 - (\gamma_t^\star)^2\big| + O\left(\left(\frac{t\log^2 n}{n}\right)^{\frac{1}{3}}\right), \tag{115}$$

as well as

$$|\gamma_{t+1}^{\star 2} - \gamma_t^{\star 2}| \leq \max_{\tau=\tau_t^\star,\tau_{t+1}^\star}\left|h_\tau\big(\gamma_t^{\star 2} + \|\varepsilon\|_2^2\big) - h_\tau\big(\gamma_{t-1}^{\star 2} + \|\varepsilon\|_2^2\big)\right| \leq (1-c)\big|\gamma_t^{\star 2} - \gamma_{t-1}^{\star 2}\big|. \tag{116}$$

The last inequality ensures that sequence $\gamma_t^\star$ converges to some fixed point $\gamma^\star$. Recalling the initialization $\gamma_1 = \gamma_1^\star = \|\theta^\star\|_2$, invoking the relation (115) recursively leads to

$$|\|\gamma_{t+1}\|_2^2 - (\gamma_{t+1}^\star)^2| \lesssim \left(\frac{t\log^2 n}{n}\right)^{\frac{1}{3}}.$$

Taking this together with (107a) ensures that

$$\big|\|\alpha_t\|_2^2 - (\alpha_t^\star)^2\big| \lesssim \left(\frac{t\log^2 n}{n}\right)^{\frac{1}{3}}.$$

In addition, the relation (116) ensures $\gamma_t^\star$ converges to some $\gamma^\star$ exponentially with t.

Proof of Claim (114). We start by proving the claim (114) for τ_{t+1}. Let us recall that τ_{t+1} satisfies inequalities (94). It thus yields

$$\left|\sum_{i:\theta_i^\star\neq 0}\mathbb{E}\Big[(\tau_{t+1} - \|\alpha_t\|_2 g\mathrm{sign}(\theta_i^\star + \|\alpha_t\|_2 g))\,\mathbb{1}(|\theta_i^\star + \|\alpha_t\|_2 g| > \tau_{t+1})\Big] - (p-k)\mathbb{E}\left[\mathsf{ST}_{\tau_{t+1}}(|\|\alpha_t\|_2 g|)\right]\right| < c'\sqrt{k},$$

for some small constant $c' > 0$. It is thus sufficient to consider τ_{t+1} such that the above inequality holds true. Letting $\omega := \frac{\sqrt{n}\tau_{t+1}}{\|\alpha_t\|_2}$, the above relation further leads to

$$\left|\frac{1}{k}\sum_{i:\theta_i\neq 0}\mathbb{E}\Big[(\omega - G\mathrm{sign}(\theta_i + G))\circ\mathbb{1}(|\theta_i + G| > \omega)\Big] - (p/k - 1)\mathbb{E}\left[\mathsf{ST}_\omega(|G|)\right]\right| < c'', \tag{117}$$

for some small constant $c'' > 0$. Here, we remind the readers that in Section B.1.1, we have shown $\|\alpha_t\|_2 \asymp 1$. Recall that we assume $p > 2.3k$ and $n > 2k\log\frac{p}{k}$. To prove Claim (114), it is thus sufficient for us to show that for e_3, c_3 small enough,

$$2.3 < \frac{p}{k} = 1 + e_3 + \frac{\frac{1}{k}\sum_{i:\theta_i\neq 0}\mathbb{E}\left[(\omega - G\mathrm{sign}(\theta_i + G))\circ\mathbb{1}(|\theta_i + G| > \omega)\right]}{\mathbb{E}\left[\mathsf{ST}_\omega(|G|)\right]}, \tag{118}$$

it satisfies

$$\mathrm{LHS} := \left|\frac{1}{2\log\frac{p}{k}}\frac{1}{k}\sum_{i:\theta_i\neq 0}\mathbb{E}\Big[\big(\mathsf{ST}_\omega(\theta_i + G) - \theta\,\mathbb{1}(|\theta_i + G| > \omega)\big)G\Big] + \frac{\frac{p}{k} - 1}{2\log\frac{p}{k}}\mathbb{E}\big[\mathsf{ST}_\omega(G)G\big]\right| < 1 - c_3. \tag{119}$$

It is easily seen that the above relation leads to the advertised bound (114) by recognizing $n > 2k\log\frac{p}{k}$.

In order to prove the required inequality (119), plugging the expression for $\frac{p}{k}$ as in (118) and in view of the the concavity of $\log(\cdot)$, we obtain

$$\begin{aligned}\mathrm{LHS} &\leq \left|\frac{\sum_{i:\theta_i\neq 0}\mathbb{E}\Big[\big(\mathsf{ST}_\omega(\theta_i + G) - \theta\,\mathbb{1}(|\theta_i + G| > \omega)\big)G + \frac{\mathbb{E}\left[\mathsf{ST}_\omega(G)G\right]}{\mathbb{E}[\mathsf{ST}_\omega(|G|)]}\left[(\omega - G\mathrm{sign}(\theta_i + G))\circ\mathbb{1}(|\theta_i + G| > \omega)\right]\Big]}{2\sum_{i:\theta_i\neq 0}\log\left(1 + \frac{\mathbb{E}[(\omega - G\mathrm{sign}(\theta_i + G))\circ\mathbb{1}(|\theta_i + G| > \omega)]}{\mathbb{E}[\mathsf{ST}_\omega(|G|)]}\right)}\right| + e_4 \\ &\leq \sup_\theta\left|\frac{\mathbb{E}\Big[\big(\mathsf{ST}_\omega(\theta + G) - \theta\,\mathbb{1}(|\theta + G| > \omega)\big)G + \frac{\mathbb{E}\left[\mathsf{ST}_\omega(G)G\right]}{\mathbb{E}[\mathsf{ST}_\omega(|G|)]}\left[(\omega - G\mathrm{sign}(\theta + G))\circ\mathbb{1}(|\theta + G| > \omega)\right]\Big]}{2\log\left(1 + \frac{\mathbb{E}[(\omega - G\mathrm{sign}(\theta + G))\circ\mathbb{1}(|\theta + G| > \omega)]}{\mathbb{E}[\mathsf{ST}_\omega(|G|)]}\right)}\right| + e_4.\end{aligned} \tag{120}$$

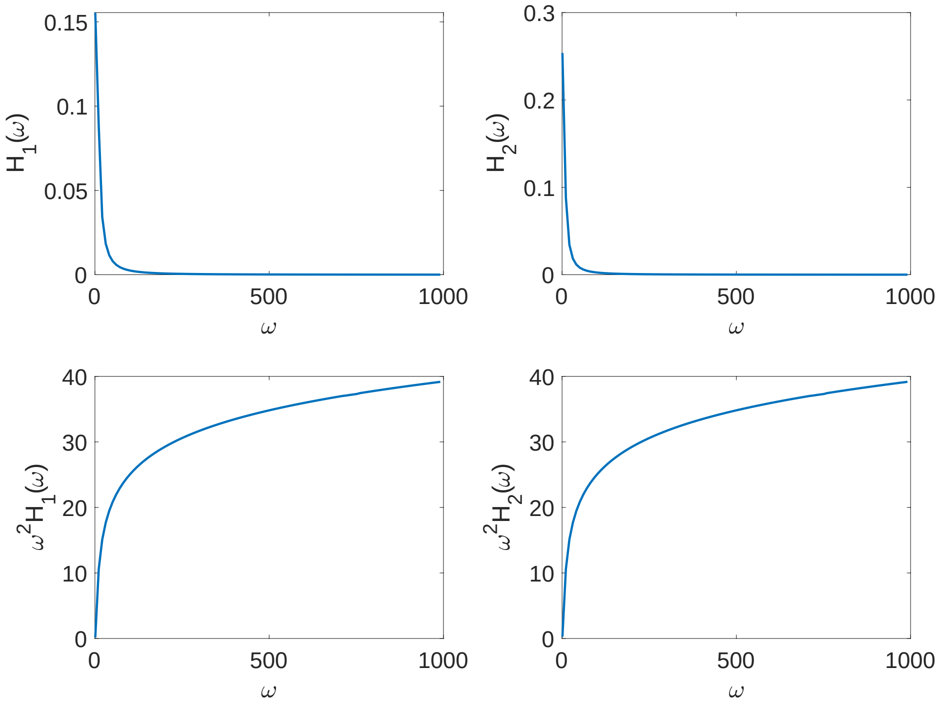

Figure 1: Numerical calculations for $H_1(\omega)$ and $H_2(\omega)$ of (123) and (121) such that $p/k \geq 2.3$.

Let us define

$$H_2(\omega) := 1 - \sup_{\theta} \left| \frac{\mathbb{E}\Big[\big(\mathsf{ST}_\omega(\theta+G) - \theta\,\mathbb{1}(|\theta+G|>\omega)\big)G + \frac{\mathbb{E}\big[\mathsf{ST}_\omega(G)G\big]}{\mathbb{E}[\mathsf{ST}_\omega(|G|)]}\left[(\omega - G\mathrm{sign}(\theta+G))\circ\mathbb{1}(|\theta+G|>\omega)\right]\Big]}{2\log\left(1+\frac{\mathbb{E}[(\omega-G\mathrm{sign}(\theta+G))\circ\mathbb{1}(|\theta+G|>\omega)]}{\mathbb{E}[\mathsf{ST}_\omega(|G|)]}\right)} \right|. \tag{121}$$

Figure 1 demonstrates that for some constant $c_4 > 0$,

$$H_2(\omega) > \min\{c_4, \ \frac{c_4}{\omega^2}\}.$$

Putting this together with the upper bound for ω in (105) establishes (119), and thus the Claim (114) for τ_{t+1}.

In addition, we make observations that $\tau_{t+1}^\star$ satisfies

$$\nabla_\tau \mathbb{E}\left\|\theta^\star - \mathsf{ST}_{\tau_{t+1}^\star}(\theta^\star + \alpha_t^\star g)\right\|_2^2 = 0,$$

and $\alpha_t^\star = \|\alpha_t^\star\|_2 + o(1)$, which follows from (107a) and the induction condition. As a result, $\tau_{t+1}^\star$ also satisfies the relation (94), which by similar argument as above validates the Claim (114) for $\tau_{t+1}^\star$.

Proof of inequality (101). Similarly, we make the observation that

$$\frac{1 + (\frac{p}{k} - 1)\mathbb{P}\left(|G| \geq \omega\right)}{2\log\frac{p}{k}} \leq \sup_\theta \frac{1 + \frac{\mathbb{P}(|G|\geq\omega)\mathbb{E}[(\omega - G\mathrm{sign}(\theta+G))\circ\mathbb{1}(|\theta+G|>\omega)]}{\mathbb{E}[\mathsf{ST}_\omega(|G|)]}}{2\log\left(1+\frac{\mathbb{E}[(\omega-G\mathrm{sign}(\theta+G))\circ\mathbb{1}(|\theta+G|>\omega)]}{\mathbb{E}[\mathsf{ST}_\omega(|G|)]}\right)} + c_4, \tag{122}$$

and define

$$H_1(\omega) := 1 - \sup_\theta \frac{1 + \frac{\mathbb{P}(|G|\geq\omega)\mathbb{E}[(\omega - G\mathrm{sign}(\theta+G))\circ\mathbb{1}(|\theta+G|>\omega)]}{\mathbb{E}[\mathsf{ST}_\omega(|G|)]}}{2\log\left(1 + \frac{\mathbb{E}[(\omega - G\mathrm{sign}(\theta+G))\circ\mathbb{1}(|\theta+G|>\omega)]}{\mathbb{E}[\mathsf{ST}_\omega(|G|)]}\right)}. \tag{123}$$

As a consequence, Figure 1 demonstrates the relation (101) through a similar argument as above.

B.2 Proof of Theorem 3

This result is again a consequence of Theorem 5. To establish the relation (48), we proceed by validating Assumptions 1 and 2 over the execution of the AMP iterations. We derive the state evolution results in Section B.2.3.

B.2.1 Validating Assumption 1

Before proceeding, we make a remark that b_t is chosen according to (44) such that

$$p(1 + 1/b_t) = \big\||r_t| < \lambda(1+b_t)\big\|_0 =: k_t. \tag{124}$$

Such b_t exists since since when $b_t \to \infty$, $k_t = n > p = \lim_{b_t\to\infty} p(1+1/b_t)$, and when $b_t \to 0$, $k_t \leq n < \infty = \lim_{b_t\to 0} p(1+1/b_t)$.

Now we proceed to validate Assumption 1 in this case. For denoising functions defined in (45), requirement (57a) satisfies straightforwardly. In addition, we note that $\|F_t(0)\|_2 = 0$ and

$$\begin{aligned}\|G_t(0)\|_2 = \|g_t(\varepsilon)\|_2 &= \frac{nb_t}{p(1+b_t)}\|\psi\left(\varepsilon;\lambda(1+b_t)\right)\|_2 \\ &\leq \frac{nb_t}{p(1+b_t)}\|\varepsilon\|_2 \lesssim 1,\end{aligned}$$

Therefore, we only need to verify that $\|\gamma_t\|_2 \asymp \|\alpha_t\|_2 \asymp 1$. We claim that this indeed the case.

Proof of $\|\gamma_t\|_2, \|\alpha_t\|_2 \lesssim 1$. Specifically, if $\|\xi_t\|_2, \|\zeta_{t-1}\|_2 = o(1)$, we shall prove that there exists some large enough constant $\overline{\gamma}$ with $\|\varepsilon\|_2/\overline{\gamma} < c$ for some constant $c > 0$ small enough, such that

$$\|\gamma_t\|_2 < \overline{\gamma}, \qquad \text{for every } t \geq 1. \tag{125}$$

It therefore implies $\|\gamma_t\|_2 \lesssim 1$. We begin by noticing that $\|\gamma_1\|_2 = \|\theta^\star\|_2 \lesssim \overline{\gamma}$. In addition, with probability at least $1 - O(n^{-10})$, one has

$$\begin{aligned}\|\alpha_t\|_2 = \|G_t(s_t)\|_2 &= \frac{nb_t}{p(1+b_t)}\|\psi\left(s_t+\varepsilon;\lambda(1+b_t)\right)\|_2 \\ &\leq \frac{n}{p}\|s_t+\varepsilon\|_2 = \frac{n}{p}\Big\|\sum_{k=1}^t \gamma_t^k\phi_k + \xi_t + \varepsilon\Big\|_2 \\ &= \frac{n}{p}\sqrt{\|\gamma_t\|_2^2 + \|\varepsilon\|_2^2} + O\Big(\sqrt{\frac{t\log n}{n}}(\|\gamma_t\|_2 + \|\varepsilon\|_2) + \|\xi_t\|_2\Big),\end{aligned} \tag{126}$$

where the last line follows from similar argument to (83). In addition, invoking the spectral properties as in (172) again gives

$$\|\gamma_{t+1}\|_2 = \|F_{t+1}(\beta_{t+1})\|_2 = \Big\|\sum_{k=1}^t \alpha_t^k\psi_k + \zeta_t\Big\|_2 = \sqrt{\frac{p}{n}}\|\alpha_t\|_2 + O\left(\sqrt{\frac{t\log n}{p}}\|\alpha_t\|_2 + \|\zeta_t\|_2\right). \tag{127}$$

If relation (125) does not hold, then there exists some t such that

$$\|\gamma_{t+1}\|_2 \geq \overline{\gamma} > \|\gamma_t\|_2.$$

then as a consequence of (126) and (127), one has $\|\gamma_t\|_2 \gtrsim \overline{\gamma}$. By definition of $\overline{\gamma}$, this means $\|\varepsilon\|_2/\|\gamma_t\|_2 = o(1)$. In the following, we show this is impossible.

In order to see this, let us first consider quantity $\|\alpha_t\|_2$. Recognizing the Lipschitz property of the function $\|\psi(\cdot \mid \lambda(1+b_t))\|_2$, we write

$$\begin{aligned}\|\alpha_t\|_2 = \|G_t(s_t)\|_2 &= \frac{nb_t}{p(1+b_t)}\left\|\psi\left(s_t+\varepsilon;\lambda(1+b_t)\right)\right\|_2 \\ &= \frac{nb_t}{p(1+b_t)}\Big\|\psi\Big(\sum_{k=1}^{t}\gamma_t^k\phi_k;\lambda(1+b_t)\Big)\Big\|_2 + O(\|\varepsilon+\xi_t\|_2).\end{aligned} \tag{128}$$

Conditional on $\|\gamma_t\|_2$, we invoke the standard concentration result for Lipschitz function of Gaussian random variables (Borell, 1975) to bound $\|\psi(\sum_{k=1}^{t}\gamma_t^k\phi_k;\lambda(1+b_t))\|_2$. To accommodate the randomness in $\gamma_t \in \mathbb{R}^t$, we take a union bound over a covering set of $\mathcal{S}^{t-1}$ of accuracy $\frac{1}{n}$. Combining these ideas together, with probability at least $1-O(n^{-10})$, it satisfies

$$\Big\|\psi\Big(\sum_{k=1}^{t}\gamma_t^k\phi_k;\lambda(1+b_t)\Big)\Big\|_2 - \mathbb{E}\Big[\|\psi(\|\gamma_t\|_2 g;\lambda(1+b_t))\|_2 \mid \|\gamma_t\|_2\Big] = O\Big(\|\varepsilon\|_2 + \sqrt{\frac{t\log n}{n}}\|\gamma_t\|_2\Big). \tag{129}$$

It therefore leads to

$$\|\alpha_t\|_2 = \mathbb{E}\Big[\|\psi(\|\gamma_t\|_2 g;\lambda(1+b_t))\|_2 \mid \|\gamma_t\|_2\Big] + O\Big(\|\varepsilon\|_2 + \sqrt{\frac{t\log n}{n}}\|\gamma_t\|_2\Big). \tag{130}$$

Recall the definition of $\psi(z;\lambda) = \min\{\max\{z,-\lambda\},\lambda\}$ to obtain

$$\begin{aligned}\mathbb{E}\Big[\|\psi(\|\gamma_t\|_2 g;\lambda(1+b_t))\|_2 \mid \|\gamma_t\|_2\Big] &= \|\gamma_t\|_2\mathbb{E}\Big[\Big\|\psi(g;\frac{\lambda(1+b_t)}{\|\gamma_t\|_2})\Big\|_2\Big] \\ &\le \|\gamma_t\|_2\sqrt{n\mathbb{E}\Big[\min\left\{\tilde{g}^2, \lambda^2(1+b_t)^2/\|\gamma_t\|_2^2\right\}\Big]}.\end{aligned} \tag{131}$$

Here, in the last equality, we denote $\tilde{g} \sim \mathcal{N}(0,\frac{1}{n})$. Therefore in order to control $\|\alpha_t\|_2$, it suffices to bound the quantity $\mathbb{E}[\min\{\tilde{g}^2, \lambda^2(1+b_t)^2/\|\gamma_t\|_2^2\}]$. We claim that it satisfies

$$\mathbb{E}\Big[\min\left\{\tilde{g}^2, \lambda^2(1+b_t)^2/\|\gamma_t\|_2^2\right\}\Big] \le (1-2c)^2\frac{p}{n^2}. \tag{132}$$

Putting everything together, we obtain

$$\|\alpha_t\|_2 \le (1-2c)\|\gamma_t\|_2\sqrt{\frac{n}{p}} + O\Big(\|\varepsilon+\xi_t\|_2 + \sqrt{\frac{t\log n}{n}}\|\gamma_t\|_2\Big). \tag{133}$$

Combining the relation (133) with (127), we end up with

$$\|\gamma_{t+1}\|_2 \le (1-c)\|\gamma_t\|_2 + o(1) < \|\gamma_t\|_2 \tag{134}$$

which is contradicted with $\|\gamma_{t+1}\|_2 \ge \overline{\gamma} > \|\gamma_t\|_2$. Hence, we conclude $\|\gamma_t\|_2 \lesssim 1$ for every $t \ge 1$, which in turn leads to $\|\alpha_t\|_2 \lesssim 1$ by virtue of (126).

Proof of inequality (132). First, we claim that the following relation holds true for $\tilde{g} \sim \mathcal{N}(0,\frac{1}{n})$,

$$\mathbb{P}\Big(|\tilde{g}| < \lambda(1+b_t)/\|\gamma_t\|_2\Big) = \frac{p}{n} + o(1). \tag{135}$$

Let us take this relation as give for the moment and come back to its proof later. Based on this, we obtain

$$\frac{n\mathbb{E}\Big[\min\left\{\tilde{g}^2, \lambda^2(1+b_t)^2/\|\gamma_t\|_2^2\right\}\Big]}{\mathbb{P}\Big(|\tilde{g}| < \lambda(1+b_t)/\|\gamma_t\|_2\Big)} \le 1-5c,$$

for some constant c depending on $\frac{p}{n}$. This can be seen from the numerical simulation in Figure 2. Here, we let $\tau := \sqrt{n}\lambda(1+b_t)/\|\gamma_t\|_2$ and

$$H_1(\tau) := \frac{1}{\mathbb{P}\big(|G| > \tau\big)}\Big(1 - \frac{\mathbb{E}\big[\min\big\{G^2, \tau^2\big\}\big]}{\mathbb{P}\big(|G| < \tau\big)}\Big). \tag{136}$$

Putting these two things together finishes the proof of inequality (132).

Finally, we conclude by proving relation (135). In view of the expression $r_t = \sum_{k=1}^t \gamma_t^k \phi_k + \xi_t + \varepsilon$, first we make the observation that

$$\left|\Big\| \mathbb{1}\left(|r_t| < \lambda(1+b_t)\right)\Big\|_0 - \Big\| \mathbb{1}\Big(|\sum_{k=1}^t \gamma_t^k \phi_k| < \lambda(1+b_t)\Big)\Big\|_0\right| \le \left\| \mathbb{1}\left(|r_t| < \lambda(1+b_t)\right) - \mathbb{1}\Big(|\sum_{k=1}^t \gamma_t^k \phi_k| < \lambda(1+b_t)\Big)\right\|_1.$$

Lemma 10 (278b) bounds the ℓ_1 discrepancy of F_t' after pertrubing the input. By similar argument, we can derive perturbation results for G_t'. In particular, we bound

$$\Big\| \mathbb{1}\left(|r_t| < \lambda(1+b_t)\right) - \mathbb{1}\Big(|\sum_{k=1}^t \gamma_t^k \phi_k| < \lambda(1+b_t)\Big)\Big\|_1 \le t\log n + n\Big(\frac{\|\xi_t + \varepsilon\|_2}{\|\gamma_t\|_2}\Big)^{2/3}. \tag{137}$$

Combining the above two inequalities yields

$$\left|\frac{1}{n}\Big\| \mathbb{1}\left(|r_t| < \lambda(1+b_t)\right)\Big\|_0 - \frac{1}{n}\Big\| \mathbb{1}\Big(|\sum_{k=1}^t \gamma_t^k \phi_k| < \lambda(1+b_t)\Big)\Big\|_0\right| \le \frac{t\log n}{n} + o(1), \tag{138}$$

where the last inequality uses the assumption $\|\xi_t\|_2 = o(1)$, $\|\varepsilon\|_2 \asymp 1$, and the relation $\|\gamma_t\|_2 \gtrsim \overline{\gamma}$. Now we move on to control the term $\frac{1}{n}\|\mathbb{1}(|\sum_{k=1}^t \gamma_t^k\phi_k| < \lambda(1+b_t))\|_0$. Some direct algebra leads to

$$\begin{aligned}
\frac{1}{n}\Big\| \mathbb{1}\Big(|\sum_{k=1}^t \gamma_t^k \phi_k| < \lambda(1+b_t)\Big)\Big\|_0 &= \frac{1}{n}\left\| \mathbb{1}\left(\frac{|\sum_{k=1}^t \gamma_t^k \phi_k|}{\|\gamma_t\|_2} < \frac{\lambda(1+b_t)}{\|\gamma_t\|_2}\right)\right\|_0 \\
&\le \sup_{\zeta\in\mathcal{S}^{t-1}} \frac{1}{n}\Big\| \mathbb{1}\Big(|\sum_{k=1}^t \zeta^k \phi_k| < \frac{\lambda(1+b_t)}{\|\gamma_t\|_2}\Big)\Big\|_0 \\
&\le \sup_{\zeta\in\mathcal{N}_\epsilon(\mathcal{S}^{t-1})} \frac{1}{n}\Big\| \mathbb{1}\Big(|\sum_{k=1}^t \zeta^k \phi_k| < \frac{\lambda(1+b_t)}{\|\gamma_t\|_2} + \epsilon\Big)\Big\|_0,
\end{aligned} \tag{139}$$

where $\mathcal{N}_\epsilon(\mathcal{S}^{t-1})$ forms an ϵ-cover of $\mathcal{S}^{t-1}$. Here in the last inequality, we make the observation that for every $\zeta' \in \mathcal{S}^{t-1}$, there exists $\zeta \in \mathcal{N}_\epsilon(\mathcal{S}^{t-1})$ such that $\|\zeta - \zeta'\|_2 \le \epsilon$ and hence

$$\Big||\sum_{k=1}^t \zeta'^k \phi_{kj}| - |\sum_{k=1}^t \zeta^k \phi_{kj}|\Big| \le \Big|\sum_{k=1}^t (\zeta^k - \zeta'^k)\phi_{kj}\Big| \le \|\zeta - \zeta'\|_2\Big(\sum_{j=1}^t \phi_{kj}^2\Big)^{1/2} \le \epsilon, \tag{140}$$

with probability at least $1 - O(n^{-11})$.

Fix each ζ independent of $\{\phi_k\}$, $\|\mathbb{1}(|\sum_{k=1}^t \zeta^k\phi_k| < \frac{\lambda(1+b_t)}{\|\gamma_t\|_2} + \epsilon)\|_0$ is the summation of n independent Bernoulli distribution with parameter $\mathbb{P}(|\tilde{g}| < \frac{\lambda(1+b_t)}{\|\gamma_t\|_2} + \epsilon)$, for $\tilde{g} \sim \mathcal{N}(0, 1/n)$. According to standard concentration result for summation of independent Bernoulli's, we obtain

$$\frac{1}{n}\Big\| \mathbb{1}\Big(|\sum_{k=1}^t \zeta^k \phi_k| < \frac{\lambda(1+b_t)}{\|\gamma_t\|_2} + \epsilon\Big)\Big\|_0 \le \mathbb{P}\Big(|\tilde{g}| < \frac{\lambda(1+b_t)}{\|\gamma_t\|_2} + \epsilon\Big) + O\Big(\sqrt{\frac{\log\frac{1}{\delta}}{n}}\Big),$$

with probability at least $1 - \delta$. If we set $\epsilon = \frac{1}{n}$ and take a union bound over elements in $\mathcal{N}_\epsilon(\mathcal{S}^{t-1})$, it holds that

$$\sup_{\zeta\in\mathcal{N}_\epsilon(\mathcal{S}^{t-1})} \frac{1}{n}\Big\| \mathbb{1}\Big(|\sum_{k=1}^t \zeta^k \phi_k| < \frac{\lambda(1+b_t)}{\|\gamma_t\|_2} + \epsilon\Big)\Big\|_0 \le \mathbb{P}\Big(|\tilde{g}| < \frac{\lambda(1+b_t)}{\|\gamma_t\|_2} + \frac{1}{n}\Big) + O\Big(\sqrt{\frac{t\log n}{n}}\Big)$$

$$\leq \mathbb{P}\Big(|\tilde{g}| < \frac{\lambda(1+b_t)}{\|\gamma_t\|_2}\Big) + O\Big(\sqrt{\frac{t\log n}{n}}\Big), \tag{141}$$

with probability at least $1-O(n^{-10})$. Here the last inequality uses $P\Big(\lambda(1+b_t)/\|\gamma_t\|_2 < |\tilde{g}| < \lambda(1+b_t)/\|\gamma_t\|_2 + 1/n\Big) < \sqrt{2/(\pi n)}$ since the density function of $|\tilde{g}|$ is bounded by $\sqrt{2n/\pi}$. Combining with (139), the above relation leads to

$$\frac{1}{n}\Big\| \mathbb{1}\left(|\sum_{k=1}^{t}\gamma_t^k\phi_k| < \lambda(1+b_t)\right)\Big\|_0 \leq \mathbb{P}\Big(|\tilde{g}| < \frac{\lambda(1+b_t)}{\|\gamma_t\|_2}\Big) + O\Big(\sqrt{\frac{t\log n}{n}}\Big). \tag{142}$$

Similarly, one can deduce

$$\begin{aligned}\frac{1}{n}\Big\| \mathbb{1}\left(|\sum_{k=1}^{t}\gamma_t^k\phi_k| < \lambda(1+b_t)\right)\Big\|_0 &\geq \inf_{\zeta\in\mathcal{S}^{t-1}}\frac{1}{n}\Big\| \mathbb{1}\left(|\sum_{k=1}^{t}\zeta^k\phi_k| < \frac{\lambda(1+b_t)}{\|\gamma_t\|_2}\right)\Big\|_0 \\ &\geq \inf_{\zeta\in\mathcal{N}_\epsilon(\mathcal{S}^{t-1})}\frac{1}{n}\Big\| \mathbb{1}\left(|\sum_{k=1}^{t}\zeta^k\phi_k| < \frac{\lambda(1+b_t)}{\|\gamma_t\|_2} - \epsilon\right)\Big\|_0 \\ &\geq \mathbb{P}\Big(|\tilde{g}| < \frac{\lambda(1+b_t)}{\|\gamma_t\|_2}\Big) + O\Big(\sqrt{\frac{t\log n}{n}}\Big). \end{aligned} \tag{143}$$

Putting these two parts with (138), we conclude

$$\frac{k_t}{n} = \frac{1}{n}\Big\| \mathbb{1}\left(|r_t| < \lambda(1+b_t)\right)\Big\|_0 = \mathbb{P}\Big(|\tilde{g}| < \frac{\lambda(1+b_t)}{\|\gamma_t\|_2}\Big) + o(1). \tag{144}$$

To establish (135), it is sufficient to notice that $k_t \asymp p$. By definition of k_t (cf. (124)), it satisfies straightforwardly that $k_t > p$. According to (144), it then implies

$$\mathbb{P}\Big(|\tilde{g}| < \frac{\lambda(1+b_t)}{\|\gamma_t\|_2}\Big) \geq \frac{p}{n} + o(1). \tag{145}$$

Given $\tilde{g} \sim \mathcal{N}(0, 1/n)$, the above relation ensures $b_t \asymp \overline{\gamma}$ for $\lambda \asymp 1/\sqrt{n}$ and $\lambda(1+b_t)/\|\gamma_t\|_2 \asymp \frac{1}{\sqrt{n}}$, which in turns gives

$$k_t = p(1+1/b_t) = p(1+o(1)).$$

We thus finish the proof of inequality (135).

Proof of $\|\gamma_t\|_2, \|\alpha_t\|_2 \gtrsim 1$. We are only left to show $\|\alpha_t\|_2 \gtrsim 1$. Similar to (128), invoking the Lipschitz property of the function $\|\psi(\cdot \mid \lambda(1+b_t))\|_2$ gives

$$\begin{aligned}\|\alpha_t\|_2 = \|G_t(s_t)\|_2 &= \frac{nb_t}{p(1+b_t)}\|\psi\left(s_t+\varepsilon; \lambda(1+b_t)\right)\|_2 \\ &= \frac{nb_t}{p(1+b_t)}\Big\|\psi\Big(\varepsilon + \sum_{k=1}^{t}\gamma_t^k\phi_k; \lambda(1+b_t)\Big)\Big\|_2 + O(\|\xi_t\|_2) \\ &\geq \frac{nb_t}{p(1+b_t)}\Big\|\psi\Big(\varepsilon + \sum_{k=1}^{t}\gamma_t^k\phi_k; \lambda(1+b_t)\Big)\circ \mathbb{1}_{\mathcal{I}}\Big\|_2 + O(\|\xi_t\|_2)\end{aligned} \tag{146}$$

where we define $\mathcal{I} := \{i : \varepsilon_i \sim \mathcal{N}(0,\sigma^2)\}$. Recall that ε_i is drawn from the mixture distribution of $\mathcal{N}(0,\sigma^2)$ and some other distribution H as in (39).

In order to show $\|\alpha_t\|_2 \gtrsim 1$, it suffices to lower bound $\|\psi(\varepsilon + \sum_{k=1}^{t}\gamma_t^k\phi_k; \lambda(1+b_t)) \circ \mathbb{1}_{\mathcal{I}}\|_2$. Before proceeding, we make two key observations.

- We first make the remark that $\{\phi_k\}$ are independent of ε. In fact, when constructing $\{\phi_k\}$ (cf. (64)), we have viewed ε as a deterministic vector and each ϕ_k admits a fixed distribution $\mathcal{N}(0, \frac{1}{n}I_n)$ no matter what value ε takes. In other words, $\{\phi_k\}$ are independent of ε.
- From our discussions around (145), it satisfies $\lambda(1+b_t) \asymp \frac{\|\gamma_t\|_2}{\sqrt{n}}$, and hence,

$$\Big\|\psi\Big(\varepsilon + \sum_{k=1}^{t}\gamma_t^k\phi_k; \lambda(1+b_t)\Big) \circ \mathbb{1}_{\mathcal{I}}\Big\|_2 \le \|\lambda(1+b_t) \circ \mathbb{1}_{\mathcal{I}}\|_2 \lesssim 1. \tag{147}$$

With these two facts in mind, similar to (143), we obtain

$$\begin{aligned}\Big\|\psi\Big(\varepsilon + \sum_{k=1}^{t}\gamma_t^k\phi_k; \lambda(1+b_t)\Big) \circ \mathbb{1}_{\mathcal{I}}\Big\|_2 &\ge \inf_{\zeta\in\mathcal{N}_\epsilon(S^{t-1})}\Big\|\psi\Big(\varepsilon + \|\gamma_t\|_2\sum_{k=1}^{t}\zeta^k\phi_k - \epsilon; \lambda(1+b_t)\Big) \circ \mathbb{1}_{\mathcal{I}}\Big\|_2 \\ &\ge \mathbb{E}\Big[\|\psi(\varepsilon + \|\gamma_t\|_2 g - 1/n; \lambda(1+b_t)) \circ \mathbb{1}_{\mathcal{I}}\|_2 \mid \|\gamma_t\|_2\Big] + O\left(\sqrt{\frac{t\log n}{n}}\|\gamma_t\|_2\right),\end{aligned} \tag{148}$$

with probability at least $1 - O(n^{-11})$. Here we take $\epsilon = 1/n$. Following the exact same argument for deriving (88), since $\Big\|\psi\Big(\varepsilon + \sum_{k=1}^{t}\gamma_t^k\phi_k; \lambda(1+b_t)\Big) \circ \mathbb{1}_{\mathcal{I}}\Big\|_2$ concentrates tightly around its mean, and is bounded from above, one can deduce

$$\begin{aligned}&\mathbb{E}\Big[\|\psi(\varepsilon + \|\gamma_t\|_2 g - 1/n; \lambda(1+b_t)) \circ \mathbb{1}_{\mathcal{I}}\|_2^2 \mid \|\gamma_t\|_2\Big] \\ &= \Big(\mathbb{E}\Big[\|\psi(\varepsilon + \|\gamma_t\|_2 g - 1/n; \lambda(1+b_t)) \circ \mathbb{1}_{\mathcal{I}}\|_2 \mid \|\gamma_t\|_2\Big]\Big)^2 + O\left(\sqrt{\frac{t\log n}{n}}\right).\end{aligned} \tag{149}$$

When it comes to further bounding quantity $\mathbb{E}[\|\psi(\varepsilon + \|\gamma_t\|_2 g - 1/n; \lambda(1+b_t)) \circ \mathbb{1}_{\mathcal{I}}\|_2^2 \mid \|\gamma_t\|_2]$, some direct algebra gives

$$\begin{aligned}\mathbb{E}\Big[\|\psi(\varepsilon + \|\gamma_t\|_2 g - 1/n; \lambda(1+b_t)) \circ \mathbb{1}_{\mathcal{I}}\|_2^2 \mid \|\gamma_t\|_2\Big] &= \mathbb{E}\Big[\sum_{k\in\mathcal{I}}\min\left\{(\varepsilon_k + \|\gamma_t\|_2 g_k - 1/n)^2,\ \lambda^2(1+b_t)^2\right\}\Big] \\ &\gtrsim \sum_{k\in\mathcal{I}}\frac{1}{2}\lambda^2(1+b_t)^2\mathbb{P}\Big(\varepsilon_k > 0, \|\gamma_t\|_2 g_k - \frac{1}{n} > \frac{1}{2}\lambda(1+b_t), k\in\mathcal{I}\Big) \\ &\gtrsim 1,\end{aligned} \tag{150}$$

where in the last line, we recall the independence between ε and $\{\phi_k\}$ and conclude

$$\mathbb{P}\Big(\varepsilon_k > 0, \|\gamma_t\|_2 g_k - \frac{1}{n} > \frac{1}{2}\lambda(1+b_t), k\in\mathcal{I}\Big) = \mathbb{P}\Big(\varepsilon_k > 0\Big)\mathbb{P}\Big(\|\gamma_t\|_2 g_k - \frac{1}{n} > \frac{1}{2}\lambda(1+b_t), k\in\mathcal{I}\Big) \gtrsim 1.$$

Putting together inequalities (146), (148) and (150) concludes $\|\alpha_t\|_2 \gtrsim 1$ and thus $\|\gamma_t\|_2 \gtrsim 1$.

Combining these two parts together, we have shown that $\|\alpha_t\|_2, \|\gamma_t\|_2 \asymp 1$, thus validating the Assumption 1.

B.2.2 Validating Assumption 2

In view of the definition of (45), it is easily seen that $\frac{1}{n}\mathbb{E}\left\|F'_{t+1}(v_{t+1})\right\|_2^2 = \frac{p}{n}$. Therefore, in order to justify Assumption 2, we are only left with computing $\frac{1}{n}\mathbb{E}[\|G'_{t+1}(u_{t+1})\|_2^2]$. Towards this, according to the choice of b_t (cf. (124)), we first make the observation that

$$\frac{1}{n}\|G'_t(s_t)\|_2^2 = \frac{1}{n}\|g'_t(r_t)\|_2^2 = \Big(\frac{nb_t}{p(1+b_t)}\Big)^2 \cdot \frac{k_t}{n} = \frac{nb_t}{p(1+b_t)}.$$

Next we shall compute the difference between $\|G_t'(u_t)\|_2^2$ and $\|G_t'(s_t)\|_2^2$. Some direct algebra gives

$$\begin{aligned}\Big|\|G_t'(u_t)\|_2^2 - \|G_t'(s_t)\|_2^2\Big| &= \Big(\frac{nb_t}{p(1+b_t)}\Big)^2 \cdot \Big|\big\||\varepsilon+u_t| < \lambda(1+b_t)\big\|_0 - \big\||r_t| < \lambda(1+b_t)\big\|_0\Big| \\ &\leq \Big(\frac{nb_t}{p(1+b_t)}\Big)^2 \cdot \Big|\mathbb{1}(\varepsilon+u_t| < \lambda(1+b_t)) - \mathbb{1}(|r_t| < \lambda(1+b_t))\Big| \\ &\lesssim t\log n + n\Big(\frac{\|\xi_t\|_2}{\|\varepsilon+u_t\|_2}\Big)^{\frac{2}{3}} \ll n.\end{aligned}$$

Here the penultimate inequality results similarly from the relation (137) as a consequence of concentration lemma 4; the last inequality invokes the assumption $\|\xi_t\|_2 = o(1)$ and inequality (82) that

$$\|\varepsilon+u_t\|_2 = \Big\|\varepsilon + \sum_{k=1}^t \gamma_t^k\phi_k\Big\|_2 = \sqrt{\|\varepsilon\|_2^2 + \left(1+O\Big(\sqrt{\frac{t\log n}{n}}\Big)\right)\|\gamma_t\|_2^2 + O\Big(\sqrt{\frac{t\log n}{n}}\,\|\varepsilon\|_2\,\|\gamma_t\|_2\Big)} \gtrsim 1.$$

Combining the derivations above, we arrive at

$$\frac{1}{n}\mathbb{E}\,\|G_t'(u_t)\|_2^2 \leq \frac{n}{p}\Big(1-\frac{1}{1+b_t}\Big) + o(1), \tag{151}$$

provided that $t \lesssim \frac{n}{\log^4 n}$. Putting everything together, we have verified that

$$\frac{1}{n^2}\mathbb{E}\,\big\|F_{t+1}'(v_{t+1})\big\|_2^2\,\mathbb{E}\,\big\|G_{t+1}'(u_{t+1})\big\|_2^2 < 1 - \frac{1}{1+c'\overline{\gamma}}, \tag{152}$$

for some constant c'. Here we recall $b_t \asymp \overline{\gamma}$ as discussed around inequality (145). We have thus validated Assumption 2.

B.2.3 State evolution

Again, our final goal is to bound the difference between the non-asymptotic SE (α_t, γ_{t+1}) to the deterministic SE defined in expression (46). We proceed by using the induction method to achieve this goal. Firstly, it is easily seen that the set of relation (49a) holds true for $t = 1$. Assuming that for some $t \geq 1$,

$$\big|\|\alpha_{t-1}\|_2^2 - \alpha_{t-1}^{\star 2}\big| \lesssim \Big(\frac{t\log^2 n}{n}\Big)^{1/3} \qquad \text{and} \qquad \big|\|\gamma_t\|_2^2 - \gamma_t^{\star 2}\big| \lesssim \Big(\frac{t\log^2 n}{n}\Big)^{1/3}, \tag{153}$$

it is thus sufficient to verify them for $t+1$.

In the derivations above, we have shown that $\|\alpha_t\|_2, \|\gamma_t\|_2 \asymp 1$. Based on these relations, we claim that

$$\|\gamma_{t+1}\|_2^2 = \frac{n}{p}\Big(\frac{b_t}{1+b_t}\Big)^2\mathbb{E}\Big[\big\|\psi\big(\varepsilon + \|\gamma_t\|_2 g; \lambda(1+b_t)\big)\big\|_2^2 \mid \|\gamma_t\|_2, \varepsilon\Big] + O\left(\Big(\frac{t\log^2 n}{n}\Big)^{\frac{1}{3}}\right). \tag{154}$$

Proof of relation (154)**.** First we recall inequality (127) to obtain that

$$\|\gamma_{t+1}\|_2^2 = \frac{p}{n}\|\alpha_t\|_2^2 + O\left(\Big(\frac{t\log^2 n}{n}\Big)^{\frac{1}{3}}\right). \tag{155}$$

The definition of α_t directly yields

$$\|\alpha_t\|_2 = \|G_t(s_t)\|_2 = \frac{nb_t}{p(1+b_t)}\left\|\psi\left(s_t+\varepsilon; \lambda(1+b_t)\right)\right\|_2. \tag{156}$$

Next, in view of the Lipschitz property of the function ψ and the decomposition (22a) of s_t, we can further conclude

$$\|\alpha_t\|_2 = \frac{nb_t}{p(1+b_t)}\Big\|\psi\Big(\varepsilon + \sum_{k=1}^t \gamma_t^k\phi_k; \lambda(1+b_t)\Big)\Big\|_2 + O\left(\Big(\frac{t\log^2 n}{n}\Big)^{\frac{1}{3}}\right)$$

$$= \frac{nb_t}{p(1+b_t)}\mathbb{E}\Big[\big\|\psi\big(\varepsilon + \|\gamma_t\|_2 g; \lambda(1+b_t)\big)\big\|_2 \mid \|\gamma_t\|_2, \varepsilon\Big] + O\left(\left(\frac{t\log^2 n}{n}\right)^{\frac{1}{3}}\right),$$

with probability at least $1 - O(n^{-10})$. Here we invoke the relation $\|\xi_t\|_2 \lesssim O((\frac{t\log^2 n}{n})^{\frac{1}{3}})$ in the first line, and the concentration property of Gaussian vectors as in (130) in the second line. In addition, since $\psi\big(\varepsilon + \|\gamma_t\|_2 g; \lambda(1+b_t)\big)$ concentrates well around its expectation, with similar argument as in (88), one can derive

$$\|\alpha_t\|_2 = \frac{nb_t}{p(1+b_t)}\sqrt{\mathbb{E}\Big[\big\|\psi\big(\varepsilon + \|\gamma_t\|_2 g; \lambda(1+b_t)\big)\big\|_2^2 \mid \|\gamma_t\|_2, \varepsilon\Big]} + O\left(\left(\frac{t\log^2 n}{n}\right)^{\frac{1}{3}}\right),$$

with probability at least $1 - O(n^{-11})$. Putting the above together with (155), we establish the advertised relation (154).

Now based on the recursion (154), in order to study how $\|\gamma_t\|_2$ evolves with t, it is sufficient to study the following function for any value $b > 0$,

$$h_b(\mu) := \mathbb{E}\Big[\big\|\psi\big(\varepsilon + \sqrt{\mu} g; \lambda(1+b)\big)\big\|_2^2 \mid \varepsilon\Big]. \tag{157}$$

With this definition, the limiting state-evolution (46) satisfies

$$\gamma_{t+1}^{\star 2} = \frac{n}{p}\Big(\frac{b_t^\star}{1+b_t^\star}\Big)^2 h_{b_t^\star}(\gamma_t^{\star 2}), \tag{158}$$

while its non-asymptotic analogue satisfies

$$\|\gamma_{t+1}\|_2^2 = \frac{n}{p}\Big(\frac{b_t}{1+b_t}\Big)^2 h_{b_t}(\|\gamma_t\|_2^2) + O\left(\left(\frac{t\log^2 n}{n}\right)^{\frac{1}{3}}\right). \tag{159}$$

Now our goal is to control the difference between $\|\gamma_{t+1}\|_2^2$ and $\gamma_{t+1}^{\star 2}$. Towards this end, we first make note of the following two properties regarding the function h_b.

- First, we claim that

$$\begin{aligned}\big|h_b'(\mu)\big| &= \frac{1}{\sqrt{\mu}}\Big|\mathbb{E}\Big[\big\langle\psi\big(\varepsilon + \sqrt{\mu} g; \lambda(1+b)\big), \mathbb{1}(|\varepsilon + \sqrt{\mu} g| < \lambda(1+b)) \circ g\big\rangle \mid \varepsilon\Big]\Big| \\ &\le \frac{1-2c}{\mu}\mathbb{E}\Big[\big\|\psi\big(\varepsilon + \sqrt{\mu} g; \lambda(1+b)\big)\big\|_2^2 \mid \varepsilon\Big] \\ &= \frac{1-2c}{\mu} h_b(\mu).\end{aligned} \tag{160}$$

  Let us take inequality (160) as given for the moment, and come back to its proof at the end of this section. Note that for $\mu \asymp 1$, $h_b(\mu)/\mu = O(1)$ and the above property guarantees that $\frac{1}{\mu}h(\mu)$ is $O(1)$-Lipschitz continuous function of μ. As a result, one has

$$(1-2c)\frac{h_b(\mu)}{\mu} \le (1-2c)\frac{h_b(\mu')}{\mu'} + O\left(\left(\frac{t\log^2 n}{n}\right)^{\frac{1}{3}}\right) \le (1-c)\frac{h_b(\mu')}{\mu'},$$

  for $|\mu - \mu'| = O((\frac{t\log^2 n}{n})^{\frac{1}{3}})$. Putting the things above together yields

$$\big|h_b'(\mu)\big| \le (1-c)\frac{h_b(\mu')}{\mu'}, \tag{161}$$

  for some constant $c \in (0,1)$, $b = b_t, b_t^\star$ and $|\mu - \mu'| = O((\frac{t\log^2 n}{n})^{\frac{1}{3}})$.

- For any value of μ, regarding $b^2 h_b(\mu)/(1+b)^2$ as a function of b, notice that

$$\frac{\partial\big(b^2\psi(u/(1+b);\lambda)^2\big)}{\partial b} = \frac{2b\psi(u/(1+b);\lambda)^2}{1+b} \ge 0. \tag{162}$$

  Therefore, $b^2 h_b(\mu)/(1+b)^2$ is a non-decreasing function of b.

With these properties in place, putting relations (158) and (159) together gives

$$\begin{aligned}
&\big|\|\gamma_{t+1}\|_2^2-(\gamma_{t+1}^{\star})^2\big|\\
&\leq\frac{n}{p}\max\Big\{\Big(\frac{b_t^{\star}}{1+b_t^{\star}}\Big)^2\big|h_{b_t^{\star}}\big(\|\gamma_t\|_2^2\big)-h_{b_t^{\star}}\big((\gamma_t^{\star})^2\big)\big|,\Big(\frac{b_t}{1+b_t}\Big)^2\big|h_{b_t}\big(\|\gamma_t\|_2^2\big)-h_{b_t}\big((\gamma_t^{\star})^2\big)\big|\Big\}+O\bigg(\Big(\frac{t\log^2 n}{n}\Big)^{\frac{1}{3}}\bigg)\\
&\lesssim\big|\|\gamma_t\|_2^2-(\gamma_t^{\star})^2\big|+O\bigg(\Big(\frac{t\log^2 n}{n}\Big)^{\frac{1}{3}}\bigg)=o(1), \qquad (163)
\end{aligned}$$

where the last relation invokes our inductive assumption. Furthermore, we can write

$$\begin{aligned}
&(\gamma_{t+1}^{\star})^{-2}\big|\|\gamma_{t+1}\|_2^2-(\gamma_{t+1}^{\star})^2\big|\\
&\leq\max\big\{\big(h_{b_t^{\star}}(\gamma_t^{\star 2})\big)^{-1}\big|h_{b_t^{\star}}\big(\|\gamma_t\|_2^2\big)-h_{b_t^{\star}}\big((\gamma_t^{\star})^2\big)\big|,(1+o(1))\big(h_{b_t}(\|\gamma_t\|_2^2)\big)^{-1}\big|h_{b_t}\big(\|\gamma_t\|_2^2\big)-h_{b_t}\big((\gamma_t^{\star})^2\big)\big|\big\}+O\bigg(\Big(\frac{t\log^2 n}{n}\Big)^{\frac{1}{3}}\bigg)\\
&\lesssim(1-c)(\gamma_t^{\star})^{-2}\big|\|\gamma_t\|_2^2-(\gamma_t^{\star})^2\big|+O\bigg(\Big(\frac{t\log^2 n}{n}\Big)^{\frac{1}{3}}\bigg), \qquad (164)
\end{aligned}$$

where the second line holds since for $b_t^{\star}\geq b_t$,

$$h_{b_t}\big(\|\gamma_t\|_2^2\big)-h_{b_t}\big((\gamma_t^{\star})^2\big)\leq h_{b_t}\big(\|\gamma_t\|_2^2\big)-h_{b_t^{\star}}\big((\gamma_t^{\star})^2\big)\leq h_{b_t^{\star}}\big(\|\gamma_t\|_2^2\big)-h_{b_t^{\star}}\big((\gamma_t^{\star})^2\big)$$

and the last line makes use of (161) with $\mu=c'\|\gamma_t\|_2^2+(1-c')(\gamma_t^{\star})^2$ and $\mu'=(\gamma_t^{\star})^2$ for some $0\leq c'\leq 1$.

In view of the initialization $\gamma_1=\gamma_1^{\star}=\|\theta^{\star}\|_2$ and $\|\gamma_t\|_2\asymp 1$, invoking the above relation (164) recursively leads to

$$\big|\|\gamma_{t+1}\|_2^2-(\gamma_{t+1}^{\star})^2\big|\lesssim\Big(\frac{t\log^2 n}{n}\Big)^{\frac{1}{3}}.$$

which in turns gives that

$$\big|\|\alpha_t\|_2^2-(\alpha_t^{\star})^2\big|\lesssim\Big(\frac{t\log^2 n}{n}\Big)^{\frac{1}{3}}.$$

Proof of Claim (160). To establish the expression (160) for $b=b_t$ and $b_t^{\star}$, consider a change of variable $\tau:=\frac{\sqrt{n}\lambda(1+b)}{\sqrt{\mu}}$. It is then sufficient to prove that for any $\tau>0$,

$$H_2(\tau):=1-\sup_{\varepsilon}\frac{\Big|\mathbb{E}\Big[G(\varepsilon+G)\,\mathbb{1}(|\varepsilon+G|<\tau)\Big]\Big|}{\mathbb{E}\big[(\varepsilon+G)^2\wedge\tau^2\big]}>c, \qquad (165)$$

for some constant $c\in(0,1)$.

- For $\tau\in(0,3)$, the required inequality (165) can be directly obtained from the numerical simulation in Figure 2. Here the value $H_2(\tau)\geq 0.02$ for $\tau\in(0,3)$.

- For $\tau\geq 3$, due to symmetry, it is enough to consider the case $\varepsilon\geq 0$. When $\varepsilon\geq\tau$, some direct calculations yield

$$\begin{aligned}
\Big|\mathbb{E}\Big[G(\varepsilon+G)\,\mathbb{1}(|\varepsilon+G|<\tau)\Big]\Big|&\leq\max\Big\{\mathbb{E}\Big[G(\varepsilon+G)\,\mathbb{1}(-\varepsilon-\tau<G<-\varepsilon)\Big],-\mathbb{E}\Big[G(\varepsilon+G)\,\mathbb{1}(-\varepsilon<G<\tau-\varepsilon)\Big]\Big\}\\
&\leq\max\Big\{\mathbb{E}\Big[G^2\,\mathbb{1}(G<-\tau)\Big],-\tau\mathbb{E}\Big[G\,\mathbb{1}(G<0)\Big]\Big\}\leq\tau,
\end{aligned}$$

and

$$\mathbb{E}\big[(\varepsilon+G)^2\wedge\tau^2\big]>\tau^2\mathbb{E}\big[\,\mathbb{1}(G>0)\big]=\frac{\tau^2}{2}.$$

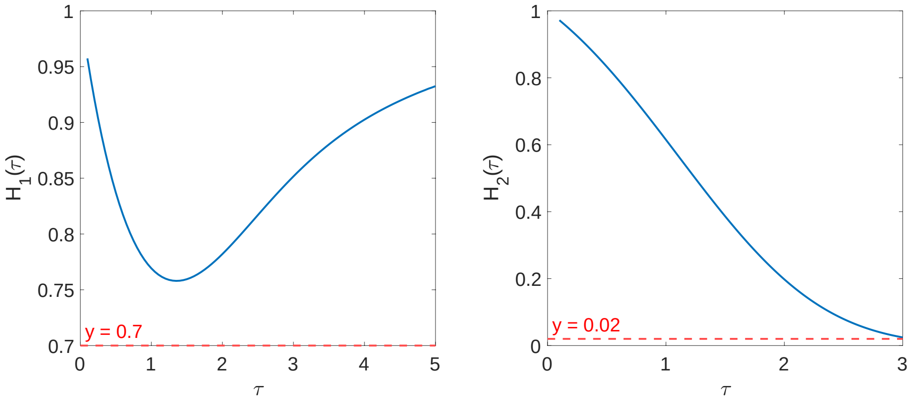

Figure 2: Numerical calculations for $H_1(\tau)$ of (136) with $\tau \in (0, 5)$ and $H_2(\tau)$ of (165) with $\tau \in (0, 3)$.

In this case, $H_2(\tau) \geq 1 - \frac{2}{\tau} \geq 1/3$. In the other case, for $0 \leq \varepsilon \leq \tau$, we obtain

$$\begin{aligned}
\Big|\mathbb{E}\Big[G(\varepsilon+G)\,\mathbb{1}(|\varepsilon+G|<\tau)\Big]\Big| &\leq \Big|\mathbb{E}\Big[G(\varepsilon+G)\,\mathbb{1}(-\varepsilon-\tau<G<\varepsilon-\tau)\Big]\Big| + 2\mathbb{E}\Big[G^2\,\mathbb{1}(0<G<\tau-\varepsilon)\Big] \\
&< \varepsilon\Big|\mathbb{E}\Big[(\varepsilon+G)\,\mathbb{1}(-\varepsilon-\tau<G<\varepsilon-\tau)\Big]\Big| + \mathbb{E}\Big[(\varepsilon+G)^2\,\mathbb{1}(-\varepsilon-\tau<G<\varepsilon-\tau)\Big] \\
&\quad + 2\mathbb{E}\Big[G^2\,\mathbb{1}(0<G<\tau-\varepsilon)\Big] \\
&< \tau^2\mathbb{E}\Big[\mathbb{1}(G<\varepsilon-\tau)\Big] + \mathbb{E}\Big[(\varepsilon+G)^2\,\mathbb{1}(-\varepsilon-\tau<G<\varepsilon-\tau)\Big] \\
&\quad + 2\mathbb{E}\Big[G^2\,\mathbb{1}(0<G<\tau-\varepsilon)\Big],
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}\big[(\varepsilon+G)^2 \wedge \tau^2\big] &> \tau^2\mathbb{E}\Big[\mathbb{1}(G>\tau-\varepsilon)\Big] + \mathbb{E}\Big[(\varepsilon+G)^2\,\mathbb{1}(-\varepsilon-\tau<G<\varepsilon-\tau)\Big] \\
&\quad + 2\mathbb{E}\Big[(\varepsilon^2+G^2)\,\mathbb{1}(0<G<\tau-\varepsilon)\Big].
\end{aligned}$$

As a result, $H_2(\tau) > 0$. Since we only need to consider those τ such that $\tau \asymp 1$ which forms a compact set, therefore it implies there exists some constant c such that $H_2(\tau) > c$.

We have thus completed the proof of Claim (160).

C Auxiliary concentration lemmas and their proofs

In this section, we collect a few concentration results for functions of random vectors that shall be used multiple times throughout this paper.

C.1 Lemma statements

The first result below considers the summation of independent sub-exponential random variables and develops a Bernstein-like concentration bound.

Lemma 3. *Suppose that Z_i's are independent random variables satisfying*

$$\mathbb{E}[Z_i] = 0 \qquad \textit{and} \qquad \mathbb{P}\Big(|Z_i| \geq B\log\frac{1}{\delta}\Big) \leq \delta, \qquad \textit{for every } 0 < \delta \leq \frac{1}{\mathsf{poly}(n)}, \tag{166}$$

for some $B \geq 0$. Then with probability at least $1-\delta$, one has

$$\Big|\sum_{i=1}^{n} Z_i\Big| \lesssim \sqrt{\sum_{i=1}^{n}\Big(\mathsf{Var}(Z_i) + \big(\frac{B\log n}{n}\big)^2\Big)\log\frac{1}{\delta}} + B\log n\log\frac{1}{\delta}. \tag{167}$$

The proof of this result can be found in Section C.2.

It is worth pointing out that a direct application of Bernstein's inequality — in view of the boundedness condition (166) for each Z_i — adds an additional $\log\frac{1}{\delta}$ to the second term of (167) (see e.g. (Wainwright, 2019, Section 2.1.3)). In that case, this additional term when combined with a covering argument shall result in an inferior dependence on the parameter t.

Next, we derive a useful concentration bound associated with indicator functions. In particular, we count the number of times that a Lipschitz function crosses a certain threshold over multiple independent realizations. This concentration result turns out to be useful when dealing with discontinuous denoising functions with its proof postponed to Section C.3.

Lemma 4. *Consider independent random vectors $\{X_i\}_{i=1}^n$. Suppose for each $i \in [n]$, $h_i(x;\theta)$ is a Lipschitz function w.r.t. $\theta \in \Theta$, with Lipschitz constant equals to L. Additionally, assume that for any fixed θ, there exists some $\sigma > 0$ such that*

$$\mathbb{P}\left(|h_i(X_i;\theta)| < \frac{s\sigma}{n}\right) < \frac{s}{n}, \qquad \forall s \in [n]. \tag{168}$$

Then for every $\varepsilon \in \mathbb{R}^n$, with probability at least $1 - O(n^{-10})$, it obeys

$$\sup_{\theta\in\Theta}\sum_{i=1}^{n}\mathbb{1}(|h_i(X_i;\theta)| < \varepsilon_i) \lesssim \log N\Big(\frac{\sigma}{n^2},\Theta\Big)\log n + \left(\frac{n\|\varepsilon\|_2}{\sigma}\right)^{\frac{2}{3}}. \tag{169}$$

Finally, we conclude this section by summarizing some standard concentration results of independent Gaussian random vectors.

Again, denote a collection of independent Gaussian vectors $\{\phi_k\}_{1\le k\le t}$ and $\{\psi_k\}_{1\le k\le t}$, with $\phi_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\frac{1}{n}I_n)$ and $\psi_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\frac{1}{n}I_p)$. If we concatenate $\{\phi_k\}_{k=1}^t$ into a matrix $\Phi \in \mathbb{R}^{n\times t}$ as

$$\Phi \coloneqq \big[\phi_1,\dots,\phi_t\big] \in \mathbb{R}^{n\times t}, \qquad \text{where } \phi_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\frac{1}{n}I_n), \tag{170}$$

its maximum singular value satisfies

$$\mathbb{P}\left(\|\Phi\|_{\text{op}} \ge 1 + \sqrt{\frac{t}{n}} + \frac{\delta}{\sqrt{n}}\right) \le e^{-\delta^2/2}. \tag{171}$$

In addition, for Wishart matrices, invoking the above result together with a union bound tells us that

$$\big\|(\phi_1,\dots,\phi_{t-1})^\top(\phi_1,\dots,\phi_{t-1}) - I_{t-1}\big\|_{\mathsf{op}} \lesssim \sqrt{\frac{t\log\frac{n}{\delta}}{n}}, \qquad \text{for every } 1 < t \le n \tag{172a}$$

with probability at least $1-\delta$ (see, also, Wainwright (2019, Example 6.2)). Similarly, for random vectors $\{\psi_k\}_{k=1}^{t-1}$, independently drawn from $\mathcal{N}(0,\frac{1}{n}I_p)$, one has

$$\left\|\frac{n}{p}(\psi_1,\dots,\psi_{t-1})^\top(\psi_1,\dots,\psi_{t-1}) - I_{t-1}\right\|_{\mathsf{op}} \lesssim \sqrt{\frac{t\log\frac{p}{\delta}}{p}}, \qquad \text{for every } 1 < t \le n. \tag{172b}$$

For our convenience, we also recall the following lemma from Li and Wei (2022). Here for every vector $x \in \mathbb{R}^n$, we follow the convention and write $|x|_{(i)}$ as its i-th largest entry in magnitude.

Lemma 5. *(Li and Wei, 2022, Lemma 8) With probability at least $1-\delta$, it holds that*

$$\Big|\max_{1\le k\le t-1}\|\phi_k\|_2 - 1\Big| \lesssim \sqrt{\frac{\log\frac{n}{\delta}}{n}}, \tag{173a}$$

$$\sup_{a=[a_k]_{1\le k<t}\in\mathcal{S}^{t-2}}\Big|\Big\|\sum_{k=1}^{t-1}a_k\phi_k\Big\|_2 - 1\Big| \lesssim \sqrt{\frac{t\log\frac{n}{\delta}}{n}}, \tag{173b}$$

$$\sup_{a=[a_k]_{1\le k<t}\in\mathcal{S}^{t-2}}\sum_{i=1}^{s}\Big|\sum_{k=1}^{t-1}a_k\phi_k\Big|_{(i)}^2 \lesssim \frac{(t+s)\log\frac{n}{\delta}}{n}, \qquad \forall 1 \le s \le n. \tag{173c}$$

For a set of random vectors $\{\phi_k\}_{k=1}^t$ independently drawn from $\mathcal{N}(0,\frac{1}{n}I_n)$,

C.2 Proof of Lemma 3

For every integer $k \ge 2$, let us consider the k-moment of random variable Z_i. For notational convenience, define $Y_i \coloneqq \frac{Z_i}{B\log n}$ and direct calculations yield

$$\begin{aligned}\mathbb{E}\left[|Z_i|^k\right] &= \mathbb{E}\left[|Z_i|^k\,\mathbb{1}\left(|Z_i| \le Bk\log n\right)\right] + \mathbb{E}\left[|Z_i|^k\,\mathbb{1}\left(|Z_i| > Bk\log n\right)\right] \\ &\overset{\text{(i)}}{\lesssim} (Bk\log n)^{k-2}\mathsf{Var}(Z_i) + (Bk\log n)^k\exp(-k\log n) \\ &\le (Bk\log n)^k\left(\mathsf{Var}(Y_i) + \frac{1}{n^2}\right),\end{aligned} \tag{174}$$

where the last step uses the definition of Y_i and $k \ge 2$. To verify the relation (i), let us use $\mu_{|Z_i|}$ to denote the density of $|Z_i|$. By direct calculations, one has

$$\begin{aligned}\mathbb{E}\left[|Z_i|^k\,\mathbb{1}\left(|Z_i| > Bk\log n\right)\right] &= \int_{Bk\log n}^{\infty} x^k\mu_{|Z_i|}(\mathrm{d}x) \\ &\le (Bk\log n)^k\exp(-k\log n) + \int_{Bk\log n}^{\infty}\exp\Big(-\frac{x}{B}\Big)\mathrm{d}x^k \\ &= (Bk\log n)^k\exp(-k\log n) + kB^k\int_{k\log n}^{\infty} x^{k-1}\exp(-x)\mathrm{d}x\end{aligned} \tag{175}$$

where the first inequality invokes the condition (166). To continue, invoking the rule of integration by part, the right hand side of (175) equals to

$$\begin{aligned}(175) &= (Bk\log n)^k\exp(-k\log n) + kB^k\exp(-k\log n)\sum_{m=1}^{k-1}\frac{(k-1)!}{(k-m)!}(k\log n)^{k-m} \\ &\le (Bk\log n)^k\exp(-k\log n) + kB^k\exp(-k\log n)\sum_{m=1}^{k-1}\frac{1}{\log^{m-1}n}(k\log n)^k \\ &\lesssim (Bk\log n)^k\exp(-k\log n),\end{aligned} \tag{176}$$

where the first inequality is proved by upper bounding the ratio between consecutive terms by $\frac{1}{\log n}$. Putting the pieces together establishes the relation (i) and thus the Bernstein condition stated in (174).

Given that each Z_i satisfies the Bernstein-type condition (174), by the power series expansion, we obtain,

$$\begin{aligned}\mathbb{E}\left[\exp\left(\lambda Z_i\right)\right] = \mathbb{E}\left[\sum_{k=0}^{\infty}\frac{\lambda^k Z_i^k}{k!}\right] = 1 + \mathbb{E}\left[\sum_{k=2}^{\infty}\frac{\lambda^k Z_i^k}{k!}\right] &\le \exp\left(\sum_{k=2}^{\infty}\frac{\mathbb{E}\left[\lambda^k Z_i^k\right]}{k!}\right) \\ &\le \exp\left(\sum_{k=2}^{\infty}\frac{\lambda^k\left(Bk\log n\right)^k}{k!}\left(\mathsf{Var}(Y_i) + \frac{1}{n^2}\right)\right) \\ &\le \exp\left(\frac{e^2\lambda^2}{2}\left(B\log n\right)^2\left(\mathsf{Var}(Y_i) + \frac{1}{n^2}\right)\right),\end{aligned}$$

for any $0 < \lambda < \frac{1}{2eB\log n}$. Here in the last step, we use the fact that for $x = \lambda B\log n \le \frac{1}{2e}$,

$$\sum_{k=2}^{\infty}\frac{(kx)^k}{k!} \overset{\text{(i)}}{\le} \sum_{k=2}^{\infty}\frac{1}{\sqrt{2\pi k}}\left(ex\right)^k \le e^2x^2\sum_{k=0}^{\infty}\frac{1}{2^k} = \frac{e^2x^2}{2},$$

where (i) follows from Stirling's formula where $\sqrt{2\pi}k^{k+\frac{1}{2}}e^{-k} \le k!$. The above bound of the moment generating function leads naturally to a high probability control where one can apply Markov's inequality to arrive

$$\mathbb{P}\left(\sum_{i=1}^{n} Z_i > t\right) \le \min_{0<\lambda<\frac{1}{2eB\log n}}\left\{\exp(-\lambda t)\cdot\mathbb{E}\Big[\exp\Big(\lambda\sum_{i=1}^{n} Z_i\Big)\Big]\right\}$$

$$\leq \min_{0<\lambda<\frac{1}{2eB\log n}} \left\{ \exp(-\lambda t) \cdot \exp\left(\frac{1}{2} \sum_{i=1}^{n} e^2 \lambda^2 \left(B \log n\right)^2 \left(\mathsf{Var}(Y_i) + \frac{1}{n^2}\right) \right) \right\}.$$

Selecting λ according to

$$\lambda \asymp \min \left\{ \frac{1}{B \log n}, \frac{t}{\sum_{i=1}^{n} (B \log n)^2 \left(\mathsf{Var}(Y_i) + \frac{1}{n^2}\right)} \right\},$$

it is ensured that with probability at least $1 - \delta/2$,

$$\sum_{i=1}^{n} Z_i \leq t \lesssim \max \left\{ B \log n \log \frac{1}{\delta}, \sqrt{\sum_{i=1}^{n} \left(\mathsf{Var}(Z_i) + \left(\frac{B \log n}{n}\right)^2\right) \log \frac{1}{\delta}} \right\}. \tag{177}$$

Repeating the same argument above for $-Z_i$ shows that the above inequality holds for $\sum_{i=1}^{n} -Z_i$. Putting these together completes the proof of relation (167).

C.3 Proof of Lemma 4

Given any fixed $\theta \in \Theta$, independent of $\{X_i\}$, let us consider random variables $\mathbb{1}(|h_i(X_i;\theta)| < \frac{s\sigma}{n})$ for $i \in [n]$. In view of inequality (168), $\mathbb{1}(|h_i(X_i;\theta)| < \frac{s\sigma}{n})$ forms a set of independent Bernoulli random variables with parameter smaller than $\frac{s}{n}$. Invoking Bernstein's inequality (see, e.g. (Wainwright, 2019, Chapter 2)) ensures that

$$\sum_{i=1}^{n} \mathbb{1}\left(|h_i(X_i;\theta)| < \frac{s\sigma}{n}\right) \leq s + \sqrt{2s \log \frac{1}{\delta}} + 2 \log \frac{1}{\delta}, \tag{178}$$

with probability at least $1 - \delta$. In order to deal with random $\widehat{\theta}$, consider an ϵ-cover of Θ of ℓ_2-norm and denote it by $\mathcal{N}_\epsilon$. By virtue of the Lipschitz property of h, there exists some $\theta \in \mathcal{N}_\epsilon$ such that

$$|h_i(X_i;\widehat{\theta}) - h_i(X_i;\theta)| \leq L\epsilon,$$

which implies

$$\mathbb{1}(|h_i(X_i;\widehat{\theta})| < x_i) \leq \mathbb{1}(|h_i(X_i;\theta)| < x_i + L\epsilon), \qquad \text{for every } x_i$$

and hence,

$$\sup_{\widehat{\theta}\in\Theta} \sum_{i=1}^{n} \mathbb{1}(|h_i(X_i;\widehat{\theta})| < x_i) \leq \sup_{\theta\in\mathcal{N}_\epsilon} \sum_{i=1}^{n} \mathbb{1}(|h_i(X_i;\theta)| < x_i + L\epsilon). \tag{179}$$

Before diving into our main proof, let us state a key property regarding $h_i(X_i;\widehat{\theta})$'s based on the above observation. In particular, select parameters

$$x_i = \frac{\sigma s}{4n}, \quad L\epsilon = \frac{\sigma}{100n^2}, \quad \text{and} \quad \delta = \frac{1}{n^{11} N(\epsilon, \Theta)}.$$

Taking a union bound of (178) over the ϵ-cover $\mathcal{N}_\epsilon$, we arrive at

$$\sup_{\widehat{\theta}\in\Theta} \sum_{i=1}^{n} \mathbb{1}\left(|h_i(X_i;\widehat{\theta})| < \frac{s\sigma}{4n}\right) \leq \sup_{\theta\in\mathcal{N}_\epsilon} \sum_{i=1}^{n} \mathbb{1}\left(|h_i(X_i;\theta)| < \frac{s\sigma}{4n} + L\epsilon\right) \leq \frac{s}{3} + O\Big(\log N(\frac{\sigma}{100n^2}, \Theta) \log n\Big), \tag{180}$$

with probability at least $1 - O(n^{-11})$. In words, for every $\widehat{\theta} \in \Theta$ and $s \gtrsim \log N(\frac{\sigma}{100n^2}, \Theta) \log n$, the total number of index i, such that $|h_i(X_i;\widehat{\theta})| < s\sigma/(4n)$ is with high probability smaller than $s/2$. It implies that, if we rank the magnitude of $|h_i(X_i;\widehat{\theta})|$ from the smallest to the largest, then regardless of the value of $\widehat{\theta}$, it always holds true with probability at least $1-O(n^{-11})$ that

$$\sum_{i=1}^{s} h_{(i)}^2(X_i,\widehat{\theta}) \geq \frac{s}{2}\cdot(\frac{s\sigma}{4n})^2 = \frac{s^3\sigma^2}{n^2}. \tag{181}$$

We emphasize that (181) holds true for every $\widehat{\theta}$ and s, as long as $s \gtrsim \log N(\frac{\sigma}{100n^2},\Theta)\log n$.

We are now ready to establish the proof of (169). For every $\varepsilon\in\mathbb{R}^n$ and $\theta\in\mathcal{N}_\epsilon$, let us define set

$$\mathcal{I}_\theta := \left\{i : |h_i(X_i;\theta)| < \varepsilon_i + L\epsilon\right\}.$$

Then according to this definition, one naturally has

$$\sum_{i=1}^{n}\mathbb{1}(|h_i(X_i;\theta)| < \varepsilon_i + L\epsilon) = |\mathcal{I}_\theta|, \tag{182}$$

and as well as

$$\sum_{i=1}^{|\mathcal{I}_\theta|} h_{(i)}^2 \leq \sum_{i=1}^{|\mathcal{I}_\theta|}(\varepsilon_i + L\epsilon)^2 \lesssim \|\varepsilon\|_2^2 + \frac{\sigma^2}{n^4}, \tag{183}$$

where the last steps invokes the choice of $L\epsilon = \frac{\sigma}{100n^2}$. If $|\mathcal{I}_\theta| \lesssim \log N(\frac{\sigma}{100n^2},\Theta)\log n$ for every $\theta\in\mathcal{N}_\epsilon$, then it is straightforward to see that

$$\sup_{\widehat{\theta}\in\mathbb{B}(r)}\sum_{i=1}^{n}\mathbb{1}\left(|h_i(X_i;\widehat{\theta})| < \varepsilon_i\right) \leq \sup_{\theta\in\mathcal{N}_\epsilon}\sum_{i=1}^{n}\mathbb{1}\left(|h_i(X_i;\theta)| < \varepsilon_i + L\epsilon\right) = \sup_{\theta\in\mathcal{N}_\epsilon}|\mathcal{I}_\theta| \lesssim \log N(\frac{\sigma}{100n^2},\Theta)\log n.$$

Otherwise, taking collectively inequality (183) with inequality (181), one has

$$\forall\theta\in\mathcal{N}_\epsilon, \qquad \frac{|\mathcal{I}_\theta|^3\sigma^2}{n^2} \lesssim \|\varepsilon\|_2^2 + \frac{\sigma^2}{n^4},,$$

which further implies

$$\sum_{i=1}^{n}\mathbb{1}(|h_i(X_i;\theta)| < \varepsilon_i + L\epsilon) = |\mathcal{I}_\theta| \lesssim \left(\frac{n\|\varepsilon\|_2}{\sigma}\right)^{\frac{2}{3}}. \tag{184}$$

Thus, in this case, we are guaranteed that

$$\sup_{\widehat{\theta}\in\mathbb{B}(r)}\sum_{i=1}^{n}\mathbb{1}\left(|h_i(X_i;\widehat{\theta})| < \varepsilon_i\right) \leq \sup_{\theta\in\mathcal{N}_\epsilon}\sum_{i=1}^{n}\mathbb{1}\left(|h_i(X_i;\theta)| < \varepsilon_i + L\epsilon\right) \lesssim \left(\frac{n\|\varepsilon\|_2}{\sigma}\right)^{\frac{2}{3}}. \tag{185}$$

Putting these two cases together completes the proof of property (169).

D Proof of Lemma 2

D.1 A general statement

In this section, we prove a more general version of Lemma 2 without imposing the Assumption 1. This general result reduces to Lemma 2 in the special case.

To simplify our statement, we start by introducing some auxiliary notation. Specifically, let us define

$$\overline{\alpha}_t := \max_{k\leq t}\left\{\left\|G_k(0)\right\|_2, \|\alpha_k\|_2, \frac{1}{\mathsf{poly}(n)}\right\}, \tag{186a}$$

$$\overline{\gamma}_t := \max_{k\le t}\Big\{ \left\|F_k(0)\right\|_2, \|\gamma_k\|_2, \frac{1}{\mathsf{poly}(n)}\Big\}, \tag{186b}$$

and we write

$$\mu_t := \frac{\rho_F\rho_G\overline{\alpha}_{t-1} + \rho_G\overline{\gamma}_{t-1}}{\|\alpha_{t-1}\|_2}, \qquad \nu_t := \frac{\rho_F\rho_G\overline{\gamma}_t + \rho_F\overline{\alpha}_{t-1}}{\|\gamma_t\|_2}, \tag{186c}$$

and

$$\overline{\xi}_t \coloneqq \overline{\alpha}_{t-1}\left(\widehat{\alpha}_{t-1}^{t-1} + \rho_F^2\widehat{\gamma}_{t-1}^{t-1} + \rho_F\sqrt{\frac{t\log^2 n}{n}}\right) + \sqrt{\frac{t\log^2 n}{n}}\rho_F\rho_G\overline{\gamma}_t, \tag{186d}$$

$$\overline{\zeta}_t \coloneqq \overline{\gamma}_t\left(\widehat{\gamma}_t^{t} + \rho_G^2\widehat{\alpha}_{t-1}^{t-1} + \rho_G\sqrt{\frac{t\log^2 n}{n}}\right) + \sqrt{\frac{t\log^2 n}{n}}\rho_F\rho_G\overline{\alpha}_t. \tag{186e}$$

Our goal is to establish the following claim in order to control the sizes of ξ_t and ζ_t. This claim is of the same form as in Claim 1.

Claim 2. *There exists universal constant $0 < c < 1$, such that the following set of inequalities hold true*

$$\|\widehat{\xi}_t\|_2 \lesssim \sqrt{\frac{t\log^2 n}{n}}\left(\rho_F\rho_G\overline{\gamma}_t + \rho_F\overline{\alpha}_{t-1}\right), \tag{187a}$$

$$\|\widehat{\zeta}_t\|_2 \lesssim \sqrt{\frac{t\log^2 n}{n}}\left(\rho_F\rho_G\overline{\alpha}_t + \rho_G\overline{\gamma}_t\right), \tag{187b}$$

$$|\widehat{\alpha}_{t-1}^{t-1}| \lesssim \sqrt{\frac{t\log^2 n}{n^2}}\rho_{1,F}\left(\rho_F\rho_G\overline{\alpha}_{t-1} + \rho_G\overline{\gamma}_{t-1}\right) + \rho_F\left(\frac{\mu_t^2 t\log^2 n}{n}\right)^{\frac{1}{3}}, \tag{187c}$$

$$|\widehat{\gamma}_t^{t}| \lesssim \sqrt{\frac{t\log^2 n}{n^2}}\rho_{1,G}\left(\rho_F\rho_G\overline{\gamma}_t + \rho_F\overline{\alpha}_{t-1}\right) + \rho_G\left(\frac{\nu_t^2 t\log^2 n}{n}\right)^{\frac{1}{3}}, \tag{187d}$$

$$|\widehat{\alpha}_{t-1}^{k}| \le \begin{cases} (1-c)^{t-k-1}\left|\widehat{\alpha}_{\frac{t+k-1}{2}}^{\frac{t+k-1}{2}}\right| & \textit{if} \quad t-1-k = 2m, \\ (1-c)^{t-k-2}\rho_F^2\left|\widehat{\gamma}_{\frac{t+k}{2}}^{\frac{t+k}{2}}\right| & \textit{if} \quad t-1-k = 2m+1, \end{cases} \tag{187e}$$

$$|\widehat{\gamma}_t^{k}| \le \begin{cases} (1-c)^{\frac{t-k}{2}}|\widehat{\gamma}_{\frac{t+k}{2}}^{\frac{t+k}{2}}| & \textit{if} \quad t-k = 2m, \\ (1-c)^{\frac{t-k}{2}}\rho_G^2|\widehat{\alpha}_{\frac{t+k-1}{2}}^{\frac{t+k-1}{2}}| & \textit{if} \quad t-1-k = 2m. \end{cases} \tag{187f}$$

Let us state an inductive results regarding the above Claim 2. The proof of this result is provided in Section D.2.

Lemma 6. *Under the decomposition (22) with (53), the bound (187) holds for $t = 1$ with probability at least $1 - n^{-10}$. In addition, with probability at least $1 - O(n^{-10})$, for every t satisfying*

$$t \ll \frac{n}{\rho_F^4\rho_G^4\log^4 n}, \tag{188}$$

if the bound (187) and Assumption 3 below hold for t, then the bound (187) holds for $t + 1$.

Assumption 3. *For the decomposition (22) with (53), assume the following conditions hold:*

- *there exists some universal constant $0 < c < 1/2$, such that*

$$\frac{1}{n^2}\mathbb{E}\big[\,\|G_t'(\widetilde{u}_t)\|_2^2 \mid \|\gamma_t\|_2\big]\mathbb{E}\big[\,\big\|F_{t+1}'(\widetilde{v}_{t+1})\big\|_2^2 \mid \|\alpha_t\|_2\big] < (1-2c)^2, \tag{189}$$

$$\frac{1}{n^2}\mathbb{E}\big[\,\big\|F_{t+1}'(\widetilde{v}_{t+1})\big\|_2^2 \mid \|\alpha_t\|_2\big]\mathbb{E}\big[\,\big\|G_{t+1}'(\widetilde{u}_{t+1})\big\|_2^2 \mid \|\gamma_{t+1}\|_2\big] < (1-2c)^2. \tag{190}$$

- *for some universal constant $0 \le c < 1$, such that*

$$|\widehat{\alpha}_{t-1}^k|, |\widehat{\gamma}_t^k| \le c^{t-k}\mathsf{poly}(n), \tag{191a}$$

$$\frac{1}{\mathsf{poly}(n)}\rho_{1,F}\overline{\alpha}_t \lesssim \rho_F \lesssim \mathsf{poly}(n) \qquad \textit{and} \qquad \frac{1}{\mathsf{poly}(n)}\rho_{1,G}\overline{\gamma}_t \lesssim \rho_G \lesssim \mathsf{poly}(n), \tag{191b}$$

$$\rho_G \sum_{k=1}^{t} |\widehat{\alpha}_{t-1}^k| \overline{\gamma}_t \ll \|\gamma_t\|_2 \qquad \textit{and} \qquad \rho_F \sum_{k=1}^{t} |\widehat{\gamma}_t^k| \overline{\alpha}_t \ll \|\alpha_t\|_2; \tag{191c}$$

- *in addition, we assume*

$$\rho_F \rho_G^2 \sum_{k=1}^{t-1} |\widehat{\alpha}_{t-1}^k| \overline{\xi}_{k-1} \lesssim \sqrt{\frac{t\log^2 n}{n}} \left(\rho_F \overline{\alpha}_{t-1} + \overline{\gamma}_t\right), \tag{192a}$$

$$\rho_F^2 \rho_G \sum_{k=1}^{t} |\widehat{\gamma}_t^k| \overline{\zeta}_{k-1} \lesssim \sqrt{\frac{t\log^2 n}{n}} \left(\rho_G \overline{\gamma}_t + \overline{\alpha}_t\right), \tag{192b}$$

$$\rho_{1,G} \sqrt{\frac{t\log n}{n}} \overline{\xi}_t + \rho_G \sqrt{\log n} \left(\frac{\overline{\xi}_t}{\|\gamma_t\|_2}\right)^{\frac{1}{3}} \ll \frac{1}{\rho_F}, \tag{192c}$$

$$\rho_{1,F} \sqrt{\frac{t\log n}{n}} \overline{\zeta}_t + \rho_F \sqrt{\log n} \left(\frac{\overline{\zeta}_t}{\|\alpha_t\|_2}\right)^{\frac{1}{3}} \ll \frac{1}{\rho_G}, \tag{192d}$$

$$\frac{1}{n}\rho_G \left[\sqrt{n}\rho_{1,G}\overline{\alpha}_{t-1}(\widehat{\alpha}_{t-1}^{t-1} + \rho_F^2 \widehat{\gamma}_{t-1}^{t-1}) + \rho_G n \left(\frac{\overline{\alpha}_{t-1}(\widehat{\alpha}_{t-1}^{t-1} + \rho_F^2 \widehat{\gamma}_{t-1}^{t-1})}{\|\gamma_t\|_2}\right)^{\frac{2}{3}}\right] \ll \frac{1}{\rho_F^2}, \tag{192e}$$

$$\frac{1}{n}\rho_F \left[\sqrt{n}\rho_{1,F}\overline{\gamma}_t(\widehat{\gamma}_t^t + \rho_G^2 \widehat{\alpha}_{t-1}^{t-1}) + \rho_F n \left(\frac{\overline{\gamma}_t(\widehat{\gamma}_t^t + \rho_G^2 \widehat{\alpha}_{t-1}^{t-1})}{\|\alpha_t\|_2}\right)^{\frac{2}{3}}\right] \ll \frac{1}{\rho_G^2}. \tag{192f}$$

Remark 10. *Under the Assumptions 1 and 2, and further assuming that the Claim 1 holds true at t, it is straightforward to verify Assumption 3. Crucially, taking $\rho_{1,F}, \rho_{1,G} = 0$, $\rho_F, \rho_G \asymp 1$ and $\|\gamma_t\|_2, \|\alpha_t\|_2 \asymp 1$ helps simplify the presentations of the above formulas to a large extent.*

D.2 Proof of Lemma 6

Before diving into the proof of Lemma 6, let us first state a few preliminaries. Throughout this proof, we condition on the event

$$\begin{aligned} &\left\|(\phi_1, \dots, \phi_{t-1})^\top (\phi_1, \dots, \phi_{t-1}) - I_{t-1}\right\|_{\mathsf{op}} \lesssim \sqrt{\frac{t\log\frac{n}{\delta}}{n}}, &&\text{for every } 1 < t \le n, \\ \text{and} \quad &\left\|\frac{n}{p}(\psi_1, \dots, \psi_{t-1})^\top (\psi_1, \dots, \psi_{t-1}) - I_{t-1}\right\|_{\mathsf{op}} \lesssim \sqrt{\frac{t\log\frac{p}{\delta}}{p}}, &&\text{for every } 1 < t \le n, \end{aligned} \tag{193}$$

both of which hold with probability at least $1-\delta$ according to inequalities (171) and (172b); in this proof, we shall take $\delta = n^{-11}$. In addition, the following results turn out to be essential for our analysis, whose proof is deferred to Section E.

Lemma 7. *Under the assumptions (191b) – (188), the following two inequalities hold true with probability at least $1 - O(n^{-11})$:*

$$\left\|\sum_{k=1}^{t} a_k \Big[\Big\langle \psi_k, F_{t+1}(\widehat{\beta}_{t+1}) \Big\rangle - \langle F_{t+1}'(\widehat{\beta}_{t+1})\rangle \alpha_t^k - \sum_{j=k+1}^{t} \widehat{\gamma}_t^j \Big\langle F_{t+1}'(\widehat{\beta}_{t+1}) \circ F_j'(v_j) \Big\rangle \alpha_{j-1}^k \Big]\right\|_2$$

$$\lesssim \sqrt{\frac{t\log^2 n}{n}}\left(\overline{\gamma}_{t+1} + \rho_F \overline{\alpha}_t\right), \tag{194}$$

and

$$\left\|\sum_{k=1}^{t+1} b_k\Big[\langle \phi_k, G_{t+1}(\widehat{s}_{t+1})\rangle - \langle G'_{t+1}(\widehat{s}_{t+1})\rangle \gamma^k_{t+1} - \sum_{j=k}^{t} \widehat{\alpha}^j_t \left\langle G'_{t+1}(\widehat{s}_{t+1}) \circ G'_j\left(u_j\right)\right\rangle \gamma^k_j\Big]\right\|_2$$
$$\lesssim \sqrt{\frac{t\log^2 n}{n}}\left(\overline{\alpha}_{t+1} + \rho_G \overline{\gamma}_{t+1}\right). \tag{195}$$

Initial case for $t = 1$. In order to prove inequalities (187) for $t = 1$, first recall the initialization condition that

$$s_1 = r_1 - \epsilon = X\theta^\star.$$

By construction of (61) and (64a), $\phi_1 = \theta^\star/\|\theta^\star\|_2$ and $\gamma^1_1 = \|\theta^\star\|_2$ and hence $\xi_1 = \widehat{\xi}_1 = 0$. In addition, since $\widehat{s}_1 = s_1$, it is guaranteed that

$$\widehat{\gamma}^1_1 := \langle G'_1(\widehat{s}_1) - G'_1(s_1)\rangle + \frac{1}{\|\gamma_1\|_2}\langle \phi_1, G_1(s_1) - G_1(\widehat{s}_1)\rangle = 0.$$

As a result, inequalities (187a) and (187d) hold for $t = 1$. The requirements for coefficient $|\widehat{\alpha}_t|$ naturally holds as they equal to zero when $t = 0$. It suffices to establish the required result for $\|\widehat{\zeta}_1\|_2$. Towards this, by construction (63) and definition (70), we observe that

$$\begin{aligned}
\alpha^1_1 &= \langle G_1(s_1),\, a_1\rangle = \|G_1(s_1)\|_2 \\
\zeta_1 &= \widehat{\zeta}_1 = b_1\left[\langle \phi_1, G_1(s_1)\rangle - \langle G'_1\rangle \gamma^1_1 - \alpha^1_1 q^1_1\right].
\end{aligned}$$

To obtain a control of the right hand side of $\widehat{\zeta}_1$, expression (195) of Lemma 7 — whose assumptions satisfy naturally as both $\widehat{\alpha}_0$ and $\widehat{\gamma}^1_1$ vanish — ensures that

$$\|\widehat{\zeta}_1\|_2 = \left|\langle \phi_1, G_1(s_1)\rangle - \langle G'_1\rangle \gamma^1_1 - \alpha^1_1 q^1_1\right| \lesssim \sqrt{\frac{\log n}{n}}\left(\overline{\alpha}_1 + \rho_G \overline{\gamma}_1\right), \tag{196}$$

from which, we complete the proof of (187b).

Inductive relation. Suppose both Assumption 3 and the target conclusion (187) hold at the t-th iteration. We shall prove the inequality set (187) at the $t+1$-th iteration. First, we remark that given (187) holds at iteration t, decomposition (53a) leads to

$$\begin{aligned}
\|\xi_t\|_2 &\le \sum_{k=1}^{t-1} |\widehat{\alpha}^k_{t-1}| \|G_k(s_k)\|_2 + \|\widehat{\xi}_t\|_2 \\
&\overset{\text{(i)}}{\lesssim} \sum_{k=1}^{t-1} |\widehat{\alpha}^k_{t-1}| \cdot \|\alpha_k\|_2 + \sqrt{\frac{t\log^2 n}{n}}\left(\rho_F\rho_G \overline{\gamma}_t + \rho_F \overline{\alpha}_{t-1}\right) \\
&\overset{\text{(ii)}}{\lesssim} \overline{\alpha}_{t-1}\left(\widehat{\alpha}^{t-1}_{t-1} + \rho_F^2 \widehat{\gamma}^{t-1}_{t-1} + \rho_F\sqrt{\frac{t\log^2 n}{n}}\right) + \sqrt{\frac{t\log^2 n}{n}}\rho_F\rho_G\overline{\gamma}_t =: \overline{\xi}_t.
\end{aligned} \tag{197}$$

Here (i) invokes the relation that $\|G_t(s_t)\|_2 = \|\alpha_t\|_2$ and the inductive assumption (187a); (ii) follows from the geometric decay of $\widehat{\alpha}^k_t$ in expression (187e). Similarly, in view of expression (53b), one has

$$\|\zeta_t\|_2 \le \sum_{k=1}^{t} |\widehat{\gamma}^k_t| \|F_k(\beta_k)\|_2 + \|\widehat{\zeta}_t\|_2$$

$$\lesssim \sum_{k=1}^{t} |\widehat{\gamma}_t^k| \cdot \|\gamma_k\|_2 + \sqrt{\frac{t\log^2 n}{n}}\, (\rho_F\rho_G\overline{\alpha}_t + \rho_G\overline{\gamma}_t)$$

$$\lesssim \overline{\gamma}_t \left(\widehat{\gamma}_t^t + \rho_G^2\widehat{\alpha}_{t-1}^{t-1} + \rho_G\sqrt{\frac{t\log^2 n}{n}}\right) + \sqrt{\frac{t\log^2 n}{n}}\rho_F\rho_G\overline{\alpha}_t =: \overline{\zeta}_t \tag{198}$$

In addition, we obtain in the similar fashion that

$$\|\widehat{s}_t - u_t\|_2 = \Big\|\sum_{k=1}^{t-1}\widehat{\alpha}_{t-1}^k G_k(u_k)\Big\|_2 \lesssim \overline{\alpha}_{t-1}(\widehat{\alpha}_{t-1}^{t-1} + \rho_F^2\widehat{\gamma}_{t-1}^{t-1}) \le \overline{\xi}_t, \tag{199}$$

$$\Big\|\widehat{\beta}_{t+1} - v_{t+1}\Big\|_2 = \Big\|\sum_{k=1}^{t}\widehat{\gamma}_t^k F_k(v_k)\Big\|_2 \lesssim \overline{\gamma}_t(\widehat{\gamma}_t^t + \rho_G^2\widehat{\alpha}_{t-1}^{t-1}) \le \overline{\zeta}_t. \tag{200}$$

With these control in place, let us verify the induction results for the next iteration based on Assumption 3.

D.2.1 Induction step for quantities $\|\widehat{\xi}_{t+1}\|_2$ and $\|\widehat{\zeta}_{t+1}\|_2$

Let us start by showing that expression (187a) holds at the $t+1$-th iteration. In view of expression (54a), it directly satisfies that

$$\begin{aligned}\|\widehat{\xi}_{t+1}\|_2 \le \Big\|\sum_{k=1}^{t} a_k\Big[\Big\langle\psi_k, F_{t+1}(\widehat{\beta}_{t+1})\Big\rangle - \langle F_{t+1}'(\widehat{\beta}_{t+1})\rangle\alpha_t^k - \sum_{j=k+1}^{t}\widehat{\gamma}_t^j\Big\langle F_{t+1}'(\widehat{\beta}_{t+1})\circ F_j'(v_j)\Big\rangle\alpha_{j-1}^k\Big]\Big\|_2 \\ + \Big\|\mathcal{P}_{G_t(s_t)}^{\perp}\sum_{k=1}^{t} a_k\Big\langle\psi_k, F_{t+1}(\beta_{t+1}) - F_{t+1}(\widehat{\beta}_{t+1})\Big\rangle\Big\|_2 + O\left(\sqrt{\frac{t\log n}{n}}\|\gamma_{t+1}\|_2\right).\end{aligned} \tag{201}$$

It is then sufficient to control the above two terms on the right accordingly. Recall that Lemma 7 ensures that

$$\left\|\sum_{k=1}^{t} a_k\Big[\Big\langle\psi_k, F_{t+1}(\widehat{\beta}_{t+1})\Big\rangle - \langle F_{t+1}'(\widehat{\beta}_{t+1})\rangle\alpha_t^k - \sum_{j=k+1}^{t}\widehat{\gamma}_t^j\Big\langle F_{t+1}'(\widehat{\beta}_{t+1})\circ F_j'(v_j)\Big\rangle\alpha_{j-1}^k\Big]\right\|_2 \lesssim \sqrt{\frac{t\log^2 n}{n}}\left(\overline{\gamma}_{t+1} + \rho_F\overline{\alpha}_t\right), \tag{202}$$

$$\left\|\sum_{k=1}^{t+1} b_k\Big[\langle\phi_k, G_{t+1}(\widehat{s}_{t+1})\rangle - \langle G_{t+1}'(\widehat{s}_{t+1})\rangle\gamma_{t+1}^k - \sum_{j=k}^{t}\widehat{\alpha}_t^j\langle G_{t+1}'(\widehat{s}_{t+1})\circ G_j'(u_j)\rangle\gamma_j^k\Big]\right\|_2 \lesssim \sqrt{\frac{t\log^2 n}{n}}\left(\overline{\alpha}_{t+1} + \rho_G\overline{\gamma}_{t+1}\right). \tag{203}$$

which completes the control of the first term.

Regarding the second term, recall that $\{a_k\}$ (defined in expression (62)) forms a set of orthonormal basis. There exists a unit vector $w \in G_t(s_t)^{\perp} \cap \mathsf{span}\{a_k\}$ such that

$$\begin{aligned}\left\|\mathcal{P}_{G_t(s_t)}^{\perp}\sum_{k=1}^{t} a_k\Big\langle\psi_k, F_{t+1}(\beta_{t+1}) - F_{t+1}(\widehat{\beta}_{t+1})\Big\rangle\right\|_2 &= w^{\top}\mathcal{P}_{G_t(s_t)}^{\perp}\sum_{k=1}^{t} a_k\Big\langle\psi_k, F_{t+1}(\beta_{t+1}) - F_{t+1}(\widehat{\beta}_{t+1})\Big\rangle \\ &= \left\langle\sum_{k=1}^{t}\omega_k a_k, \sum_{k=1}^{t} a_k\Big\langle\psi_k, F_{t+1}(\beta_{t+1}) - F_{t+1}(\widehat{\beta}_{t+1})\Big\rangle\right\rangle,\end{aligned} \tag{204}$$

where $\omega \in \mathcal{S}^{t-1}$ for $\omega_k = w^{\top}a_k$. In view of this decomposition, we remark that as stated in (73), one has

$$0 = \left\langle\sum_{k=1}^{t}\omega_k a_k, \sum_{k=1}^{t}\alpha_t^k a_k\right\rangle = \langle\omega,\, \alpha_t\rangle. \tag{205}$$

In addition, there exists some $\overline{\zeta} \in \mathbb{R}^p$ such that

$$\begin{aligned}
\Big\| \mathcal{P}^{\perp}_{G_t(s_t)} \sum_{k=1}^{t} a_k \Big\langle \psi_k, F_{t+1}(\beta_{t+1}) - F_{t+1}(\widehat{\beta}_{t+1}) \Big\rangle \Big\|_2 &= \Big\langle \sum_{k=1}^{t} \omega_k \psi_k, F_{t+1}(\beta_{t+1}) - F_{t+1}(\widehat{\beta}_{t+1}) \Big\rangle \\
&= \Big\langle \sum_{k=1}^{t} \omega_k \psi_k, F'_{t+1}(v_{t+1} + \overline{\zeta}) \circ \Big(\beta_{t+1} - \widehat{\beta}_{t+1}\Big) \Big\rangle \\
&\le \Big\| \sum_{k=1}^{t} \omega_k \psi_k \circ F'_{t+1}(v_{t+1} + \overline{\zeta}) \Big\|_2 \Big\| \beta_{t+1} - \widehat{\beta}_{t+1} \Big\|_2 . \qquad (206)
\end{aligned}$$

Therefore, it is enough to control the two terms on the right hand side above, which is what shall be done in the following.

- **Control of $\|\beta_{t+1} - \widehat{\beta}_{t+1}\|_2$.** Here, the size of $\overline{\zeta}$ can be bounded by

$$\|\overline{\zeta}\|_2 \le \|\widehat{\beta}_{t+1} - v_{t+1}\|_2 + \|\beta_{t+1} - v_{t+1}\|_2 \le \|\widehat{\beta}_{t+1} - v_{t+1}\|_2 + \|\zeta_t\|_2 \lesssim \overline{\zeta}_t, \qquad (207)$$

where the last relation follows from the derivations in (198) and (200). Putting the pieces above together, we arrive at

$$\|\widehat{\xi}_{t+1}\|_2 \le \Big\| \sum_{k=1}^{t} \omega_k \psi_k \circ F'_{t+1}(v_{t+1} + \overline{\zeta}) \Big\|_2 \Big\| \beta_{t+1} - \widehat{\beta}_{t+1} \Big\|_2 + O\Big(\sqrt{\frac{t \log^2 n}{n}} \left(\overline{\gamma}_{t+1} + \rho_F \overline{\alpha}_t\right) \Big). \qquad (208)$$

To further control the right hand side, recall the definitions that $\widehat{\beta}_{t+1} = v_{t+1} + \sum_{k=1}^{t} \widehat{\gamma}_t^k F_k(v_k)$ and $\beta_{t+1} = v_{t+1} + \zeta_t = v_{t+1} + \sum_{k=1}^{t} \widehat{\gamma}_t^k F_k(\beta_k) + \widehat{\zeta}_t$. Therefore, we have

$$\begin{aligned}
\|\beta_{t+1} - \widehat{\beta}_{t+1}\|_2 &\le \|\widehat{\zeta}_t\|_2 + \Big\| \sum_{k=1}^{t} \widehat{\gamma}_t^k \left(F_k(\beta_k) - F_k(v_k)\right) \Big\|_2 \\
&\le \|\widehat{\zeta}_t\|_2 + O\Big(\frac{1}{\rho_F} \sqrt{\frac{t \log^2 n}{n}} \left(\overline{\gamma}_{t+1} + \overline{\alpha}_t\right) \Big). \qquad (209)
\end{aligned}$$

Here, in the last inequality, we make the observation that

$$\Big\| \sum_{k=1}^{t} \widehat{\gamma}_t^k \left(F_k(\beta_k) - F_k(v_k)\right) \Big\|_2 \le \sum_{k=1}^{t} |\widehat{\gamma}_t^k| \rho_F \|\zeta_{k-1}\|_2 \le \rho_F \sum_{k=1}^{t} |\widehat{\gamma}_t^k| \overline{\zeta}_{k-1} \lesssim \frac{1}{\rho_F} \sqrt{\frac{t \log^2 n}{n}} \left(\overline{\gamma}_t + \overline{\alpha}_t\right),$$

where the last inequality follows from the assumptions (192b) and $\rho_G \ge 1$. Now invoking our inductive assumption (187b) and the assumption that $\rho_F, \rho_G \ge 1$, we arrive at

$$\|\beta_{t+1} - \widehat{\beta}_{t+1}\|_2 \lesssim \sqrt{\frac{t \log^2 n}{n}} \left(\rho_F \rho_G \overline{\alpha}_t + \rho_G \overline{\gamma}_t\right). \qquad (210)$$

As a result, it can be concluded that

$$\begin{aligned}
&\|\widehat{\xi}_{t+1}\|_2 \\
&\le \Big\| \sum_{k=1}^{t} \omega_k \psi_k \circ F'_{t+1}(v_{t+1} + \overline{\zeta}) \Big\|_2 \Big(\|\widehat{\zeta}_t\|_2 + O\Big(\frac{1}{\rho_F} \sqrt{\frac{t \log^2 n}{n}} \left(\overline{\gamma}_{t+1} + \overline{\alpha}_t\right) \Big) \Big) + O\Big(\sqrt{\frac{t \log^2 n}{n}} \left(\overline{\gamma}_{t+1} + \rho_F \overline{\alpha}_t\right) \Big) \\
&\le \Big\| \sum_{k=1}^{t} \omega_k \psi_k \circ F'_{t+1}(v_{t+1} + \overline{\zeta}) \Big\|_2 \|\widehat{\zeta}_t\|_2 + \rho_F \Big\| \sum_{k=1}^{t} \omega_k \psi_k \Big\|_2 O\Big(\frac{1}{\rho_F} \sqrt{\frac{t \log^2 n}{n}} \left(\overline{\gamma}_t + \overline{\alpha}_t\right) \Big) + O\Big(\sqrt{\frac{t \log^2 n}{n}} \left(\overline{\gamma}_{t+1} + \rho_F \overline{\alpha}_t\right) \Big) \\
&\le \Big\| \sum_{k=1}^{t} \omega_k \psi_k \circ F'_{t+1}(v_{t+1} + \overline{\zeta}) \Big\|_2 \|\widehat{\zeta}_t\|_2 + O\Bigg(\sqrt{\frac{t \log^2 n}{n}} \left(\overline{\gamma}_{t+1} + \rho_F \overline{\alpha}_t\right) \Bigg), \qquad (211)
\end{aligned}$$

where the last line follows since we condition on event (193).

- **Control of $\|\sum_{k=1}^{t}\omega_k\psi_k\circ F'_{t+1}(v_{t+1}+\overline{\zeta})\|_2$.** Next, we shall focus our attention on quantity $\|\sum_{k=1}^{t}\omega_k\psi_k\circ F'_{t+1}(v_{t+1}+\overline{\zeta})\|_2$ with the size of $\overline{\zeta}$ bounded by (207). To begin with, the property (278a) summarized in Lemma 10 ensures

  $$\begin{aligned}&\left\|\sum_{k=1}^{t}\omega_k\psi_k\circ\left[F'_{t+1}(v_{t+1}+\overline{\zeta})-F'_{t+1}(v_{t+1})\right]\right\|_2\\&\leq\rho_{1,F}\sqrt{\frac{t\log n}{n}}\|\overline{\zeta}\|_2+\rho_F\left(\sqrt{\frac{t\log^2 n}{n}}+\sqrt{\log n}\left(\frac{\|\overline{\zeta}\|_2}{\|\alpha_t\|_2}\right)^{\frac{1}{3}}\right)\ll\frac{1}{\rho_G}.\end{aligned}\tag{212}$$

  Here in the last inequality, we invoke the assumptions (188) and (192d).

  Now in view of triangle's inequality, in order to bound the size of $\|\sum_{k=1}^{t}\omega_k\psi_k\circ F'_{t+1}(v_{t+1}+\overline{\zeta})\|$, it is sufficient to control the size of $\|\sum_{k=1}^{t}\omega_k\psi_k\circ F'_{t+1}(v_{t+1})\|_2$. Towards this goal, we introduce the following lemma.

  Lemma 8. *With probability at least $1-O(n^{-10})$, for any $\omega\in\mathcal{S}^{t-1}$, it holds that*

  $$\left\|\sum_{k=1}^{t}\omega_k\psi_k\circ F'_{t+1}(v_{t+1})\right\|_2^2-\frac{1}{n}\mathbb{E}\big[\big\|F'_{t+1}(\widetilde{v}_{t+1})\big\|_2^2\mid\|\alpha_t\|_2\big]\lesssim\sqrt{\frac{t\log^2 n}{n}}\rho_F^2,\tag{213a}$$

  and

  $$\frac{1}{n}\big\|F'_{t+1}(v_{t+1})\big\|_2^2-\frac{1}{n}\mathbb{E}\big[\big\|F'_{t+1}(\widetilde{v}_{t+1})\big\|_2^2\mid\|\alpha_t\|_2\big]\lesssim\sqrt{\frac{t\log^2 n}{n}}\rho_F^2.\tag{213b}$$

  The proof of this lemma is postponed to Section F.2.

  Combining (213a) with the assumption (188), which suggests $\rho_F^2\sqrt{\frac{t\log^2 n}{n}}\ll 1/\rho_G^2$, we arrive at

  $$\left\|\sum_{k=1}^{t}\omega_k\psi_k\circ F'_{t+1}(v_{t+1})\right\|_2^2-\frac{1}{n}\mathbb{E}\big[\big\|F'_{t+1}(\widetilde{v}_{t+1})\big\|_2^2\mid\|\alpha_t\|_2\big]\ll\frac{1}{\rho_G^2}.$$

  Taking this collectively with (212), it is ensured that

  $$\begin{aligned}\Big\|\sum_{k=1}^{t}\omega_k\psi_k\circ F'_{t+1}(v_{t+1}+\overline{\zeta})\Big\|_2&\leq o\Big(\frac{1}{\rho_G}\Big)+\sqrt{\frac{1}{n}\mathbb{E}\big[\big\|F'_{t+1}(\widetilde{v}_{t+1})\big\|_2^2\mid\|\alpha_t\|_2\big]+o\Big(\frac{1}{\rho_G^2}\Big)}\\&\leq\sqrt{\frac{1}{n}\mathbb{E}\big[\big\|F'_{t+1}(\widetilde{v}_{t+1})\big\|_2^2\mid\|\alpha_t\|_2\big]}+o\Big(\frac{1}{\rho_G}\Big).\end{aligned}\tag{214}$$

  Similarly, one can also conclude that

  $$\Big\|\sum_{k=1}^{t}\mu_k\phi_k\circ G'_t(u_t+\overline{\xi})\Big\|_2\leq\sqrt{\frac{1}{n}\mathbb{E}\big[\|G'_t(\widetilde{u}_t)\|_2^2\mid\|\gamma_t\|_2\big]}+o\Big(\frac{1}{\rho_F}\Big).\tag{215}$$

In summary. With these properties in place, we are ready to bound $\|\widehat{\xi}_{t+1}\|_2$. Recall expression (211) to obtain

$$\|\widehat{\xi}_{t+1}\|_2\leq\left(\sqrt{\frac{1}{n}\mathbb{E}\big[\big\|F'_{t+1}(\widetilde{v}_{t+1})\big\|_2^2\mid\|\alpha_t\|_2\big]}+o\Big(\frac{1}{\rho_G}\Big)\right)\|\widehat{\zeta}_t\|_2+O\Big(\sqrt{\frac{t\log^2 n}{n}}\left(\overline{\gamma}_{t+1}+\rho_F\overline{\alpha}_t\right)\Big),\tag{216}$$

where we make use of the relation (214). Similar to (216), one can establish the recursive relation between $\|\widehat{\zeta}_t\|_2$ and $\|\widehat{\xi}_t\|_2$ as in

$$\|\widehat{\zeta}_t\|_2 \le \left(\sqrt{\frac{1}{n}\mathbb{E}\big[\,\|G_t'(\widetilde{u}_t)\|_2^2 \mid \|\gamma_t\|_2\big]} + o\Big(\frac{1}{\rho_F}\Big)\right)\|\widehat{\xi}_t\|_2 + O\Big(\sqrt{\frac{t\log^2 n}{n}}\,(\overline{\alpha}_t + \rho_G\overline{\gamma}_t)\Big).$$

Consequently, the above two relations combined together yields

$$\begin{aligned}
\|\widehat{\xi}_{t+1}\|_2 \le{}& \left(\sqrt{\frac{1}{n}\mathbb{E}\big[\,\big\|F_{t+1}'(\widetilde{v}_{t+1})\big\|_2^2 \mid \|\alpha_t\|_2\big]} + o\Big(\frac{1}{\rho_G}\Big)\right)\left(\sqrt{\frac{1}{n}\mathbb{E}\big[\,\|G_t'(\widetilde{u}_t)\|_2^2 \mid \|\gamma_t\|_2\big]} + o\Big(\frac{1}{\rho_F}\Big)\right)\cdot\|\widehat{\xi}_t\|_2 \\
&+ O\Big(\sqrt{\frac{t\log^2 n}{n}}\,(\rho_F\rho_G\overline{\gamma}_{t+1} + \rho_F\overline{\alpha}_t)\Big) \\
\le{}& \left(\sqrt{\frac{1}{n^2}\mathbb{E}\big[\,\|G_t'(\widetilde{u}_t)\|_2^2 \mid \|\gamma_t\|_2\big]\mathbb{E}\big[\,\big\|F_{t+1}'(\widetilde{v}_{t+1})\big\|_2^2 \mid \|\alpha_t\|_2\big]} + o(1)\right)\cdot\|\widehat{\xi}_t\|_2 + O\Big(\sqrt{\frac{t\log^2 n}{n}}\,(\rho_F\rho_G\overline{\gamma}_{t+1} + \rho_F\overline{\alpha}_t)\Big) \\
\le{}& \Big(1-c\Big)\|\widehat{\xi}_t\|_2 + O\Big(\sqrt{\frac{t\log^2 n}{n}}\,(\rho_F\rho_G\overline{\gamma}_{t+1} + \rho_F\overline{\alpha}_t)\Big), \qquad (217)
\end{aligned}$$

where the last inequality follows from the assumption (189).

Putting inequality (217) and the inductive assumption (187a) that $\|\widehat{\xi}_t\|_2 \lesssim \sqrt{\frac{t\log^2 n}{n}}\,(\rho_F\rho_G\overline{\gamma}_t + \rho_F\overline{\alpha}_{t-1})$, we complete the proof of (187a) at $t+1$. Additionally, the control of $\widehat{\zeta}_{t+1}$ can be derived in a similar way, which we omit here for brevity.

D.2.2 Induction step for quantity $|\widehat{\alpha}_t^t|$ and $|\widehat{\gamma}_{t+1}^{t+1}|$

Let us recall our definition of $\widehat{\alpha}_t^t$ in expression (52c) where

$$\widehat{\alpha}_t^t := \langle F_{t+1}'(\widehat{\beta}_{t+1}) - F_{t+1}'(\beta_{t+1})\rangle + \frac{1}{\|\alpha_t\|_2^2}\left\langle\sum_{k=1}^t \alpha_t^k\psi_k, F_{t+1}(\beta_{t+1}) - F_{t+1}(\widehat{\beta}_{t+1})\right\rangle.$$

To establish (187c) at the $t+1$-th iteration, it is sufficient to bound the two terms on right of the above expression respectively.

To begin with, according to inequality (278c) in Lemma 10, we are ensured that

$$\begin{aligned}
\langle F_{t+1}'(\widehat{\beta}_{t+1}) - F_{t+1}'(\beta_{t+1})\rangle &\le \frac{1}{\sqrt{n}}\rho_{1,F}\left\|\beta_{t+1} - \widehat{\beta}_{t+1}\right\|_2 + \rho_F\left(\frac{t\log n}{n} + \left(\frac{\|\beta_{t+1} - \widehat{\beta}_{t+1}\|_2}{\|\alpha_t\|_2}\right)^{\frac{2}{3}}\right) \\
&\lesssim \sqrt{\frac{t\log^2 n}{n^2}}\rho_{1,F}\left(\rho_F\rho_G\overline{\alpha}_t + \rho_G\overline{\gamma}_t\right) + \rho_F\left(\frac{\mu_{t+1}^2 t\log^2 n}{n}\right)^{\frac{1}{3}}.
\end{aligned}$$

where we invoke the relation (210) and recall that $\mu_{t+1} := \frac{\rho_F\rho_G\overline{\alpha}_t + \rho_G\overline{\gamma}_t}{\|\alpha_t\|_2}$ as in expression (186c). In addition, conditioning on event (193), it is easily seen that

$$\begin{aligned}
\frac{1}{\|\alpha_t\|_2^2}\left\langle\sum_k \alpha_t^k\psi_k, F_{t+1}(\beta_{t+1}) - F_{t+1}(\widehat{\beta}_{t+1})\right\rangle &\le \frac{1}{\|\alpha_t\|_2^2}\Big\|\sum_k \alpha_t^k\psi_k\Big\|_2\cdot\rho_F\left\|\beta_{t+1} - \widehat{\beta}_{t+1}\right\|_2 \\
&\lesssim \rho_F\frac{\left\|\beta_{t+1} - \widehat{\beta}_{t+1}\right\|_2}{\|\alpha_t\|_2} \lesssim \rho_F\sqrt{\frac{\mu_{t+1}^2 t\log^2 n}{n}}.
\end{aligned}$$

Combining these two bounds establishes (187c) for $\widehat{\alpha}_t^t$. Moreover, the upper bound of $\widehat{\gamma}_{t+1}^{t+1}$ as in (187d) can be derived in a similar manner, thus is omitted here.

D.2.3 Induction step for quantities $|\widehat{\alpha}_{t-1}^{k}|$

For $k < t$, recall the definitions in (52c) and (52a) that

$$\begin{aligned}\widehat{\alpha}_t^k &:= \widehat{\gamma}_t^{k+1}\left\langle F_{t+1}'(\widehat{\beta}_{t+1})\circ F_{k+1}'(v_{k+1})\right\rangle\\ \widehat{\gamma}_{t+1}^{k} &:= \widehat{\alpha}_t^k\left\langle G_{t+1}'(\widehat{s}_{t+1})\circ G_k'(u_k)\right\rangle\end{aligned}$$

When $k = t-1$, it is easily seen that $|\widehat{\alpha}_t^{t-1}| := |\widehat{\gamma}_t^t\langle F_{t+1}'(\widehat{\beta}_{t+1})\circ F_t'(v_t)\rangle| \le \rho_F^2|\widehat{\gamma}_t^t|$. So we only need to prove for $k \le t-2$ where

$$\widehat{\alpha}_t^k := \widehat{\gamma}_t^{k+1}\left\langle F_{t+1}'(\widehat{\beta}_{t+1})\circ F_{k+1}'(v_{k+1})\right\rangle = \widehat{\alpha}_{t-1}^{k+1}\left\langle G_t'(\widehat{s}_t)\circ G_{k+1}'(u_{k+1})\right\rangle\left\langle F_{t+1}'(\widehat{\beta}_{t+1})\circ F_{k+1}'(v_{k+1})\right\rangle. \tag{218}$$

Before proceeding, let us make note of the simple relation that

$$\left\langle F_{t+1}'(\widehat{\beta}_{t+1})\circ F_{k+1}'(v_{k+1})\right\rangle \le \left\langle F_{t+1}'(v_{t+1})\circ F_{k+1}'(v_{k+1})\right\rangle + \frac{1}{n}\rho_F\left\|F_{t+1}'(\widehat{\beta}_{t+1}) - F_{t+1}'(v_{t+1})\right\|_1. \tag{219}$$

We shall bound the two parts on the right respectively as below.

- In view of Lemma 10 inequality (278b), it satisfies that

$$\begin{aligned}\frac{1}{n}\rho_F\left\|F_{t+1}'(\widehat{\beta}_{t+1}) - F_{t+1}'(v_{t+1})\right\|_1 &\le \frac{1}{n}\rho_F\left[\sqrt{n}\rho_{1,F}\|\widehat{\beta}_{t+1} - v_{t+1}\|_2 + \rho_F\left(t\log n + n\left(\frac{\|\widehat{\beta}_{t+1} - v_{t+1}\|_2}{\|\alpha_t\|_2}\right)^{\frac{2}{3}}\right)\right]\\ &\lesssim \frac{1}{n}\rho_F\left[\sqrt{n}\rho_{1,F}\overline{\gamma}_t(\widehat{\gamma}_t^t + \rho_G^2\widehat{\alpha}_{t-1}^{t-1}) + \rho_F n\left(\frac{\overline{\gamma}_t(\widehat{\gamma}_t^t + \rho_G^2\widehat{\alpha}_{t-1}^{t-1})}{\|\alpha_t\|_2}\right)^{\frac{2}{3}}\right] + \frac{t\log n\rho_F^2}{n}\end{aligned} \tag{220}$$

  where the last step follows from the inequality (200) where we proved $\|\widehat{\beta}_{t+1} - v_{t+1}\|_2 \le \overline{\gamma}_t(\widehat{\gamma}_t^t + \rho_G^2\widehat{\alpha}_{t-1}^{t-1})$. Next, by virtue of assumptions (192f) and (188), the right hand side of inequality (220) further satisfies

$$\frac{1}{n}\rho_F\left\|F_{t+1}'(\widehat{\beta}_{t+1}) - F_{t+1}'(v_{t+1})\right\|_1 \ll \frac{1}{\rho_G^2}. \tag{221}$$

- It is then sufficient to consider the quantity $\langle F_{t+1}'(v_{t+1})\circ F_{k+1}'(v_{k+1})\rangle$. Towards this, it is easily seen that

$$\left\langle F_{t+1}'(v_{t+1})\circ F_{k+1}'(v_{k+1})\right\rangle \le \frac{1}{n}\left\|F_{t+1}'(v_{t+1})\right\|_2\left\|F_{k+1}'(v_{k+1})\right\|_2.$$

  To bound the right hand side, notice that according to Lemma 8, for every k, it obeys

$$\frac{1}{n}\left\|F_{k+1}'(v_{k+1})\right\|_2^2 - \frac{1}{n}\mathbb{E}\left[\left\|F_{k+1}'(\widetilde{v}_{k+1})\right\|_2^2 \mid \|\alpha_k\|_2\right] \lesssim \rho_F^2\sqrt{\frac{t\log^2 n}{n}} \ll \frac{1}{\rho_G^2}.$$

  Therefore we arrive at

$$\begin{aligned}\left\langle F_{t+1}'(v_{t+1})\circ F_{k+1}'(v_{k+1})\right\rangle &\le \sqrt{\frac{1}{n^2}\left\|F_{t+1}'(v_{t+1})\right\|_2^2\left\|F_{k+1}'(v_{k+1})\right\|_2^2}\\ &\le \sqrt{\frac{1}{n}\mathbb{E}\left[\left\|F_{t+1}'(\widetilde{v}_{t+1})\right\|_2^2 \mid \|\alpha_t\|_2\right] + o(\frac{1}{\rho_G^2})}\sqrt{\frac{1}{n}\mathbb{E}\left[\left\|F_{k+1}'(\widetilde{v}_{k+1})\right\|_2^2 \mid \|\alpha_k\|_2\right] + o(\frac{1}{\rho_G^2})}\end{aligned}$$

Taking the above inequality collectively with displays (219) and (221) gives us

$$\left\langle F_{t+1}'(\widehat{\beta}_{t+1})\circ F_{k+1}'(v_{k+1})\right\rangle \le \sqrt{\frac{1}{n^2}\mathbb{E}\left[\left\|F_{t+1}'(\widetilde{v}_{t+1})\right\|_2^2 \mid \|\alpha_t\|_2\right]\mathbb{E}\left[\left\|F_{k+1}'(\widetilde{v}_{k+1})\right\|_2^2 \mid \|\alpha_k\|_2\right]} + o(\frac{1}{\rho_G^2}). \tag{222}$$

Similarly, one can develop a symmetric bound for G' as

$$\big\langle G_t'(\widehat{s}_t) \circ G_{k+1}'(u_{k+1})\big\rangle \leq \sqrt{\frac{1}{n}\mathbb{E}\big[\, \|G_t'(\widetilde{u}_t)\|_2^2 \mid \|\gamma_t\|_2\big]\mathbb{E}\big[\, \big\|G_{k+1}'(\widetilde{u}_{k+1})\big\|_2^2 \mid \|\gamma_{k+1}\|_2\big]} + o(\frac{1}{\rho_F^2}). \tag{223}$$

Therefore, under assumption (189), it holds that

$$\Big\langle F_{t+1}'(\widehat{\beta}_{t+1}) \circ F_{k+1}'(v_{k+1})\Big\rangle \big\langle G_t'\,(\widehat{s}_t) \circ G_{k+1}'\,(u_{k+1})\big\rangle < (1-c)^2.$$

As a consequence, we can establish the inductive relationship that

$$|\widehat{\alpha}_t^k| \leq (1-c)^2 |\widehat{\alpha}_{t-1}^{k+1}|.$$

Invoking the above inequality recursively validates the relation (187e) at $t+1$-th step. In addition, one can prove for inequality (187f) in a similar manner.

E Proof of Lemma 7

Let us present the proof of inequality (194) and inequality (195) can be established in the same fashion. Throughout this proof, let us condition on the event where both expressions (172a) and (172b) hold true (with δ chosen as $\max(n,p)^{-11}$)

$$\big\|(\phi_1, \ldots, \phi_{t-1})^\top(\phi_1, \ldots, \phi_{t-1}) - I_{t-1}\big\|_{\mathrm{op}} \lesssim \sqrt{\frac{t \log n}{n}}, \qquad \text{for every } 1 < t \leq n, \tag{224a}$$

$$\Big\|\frac{n}{p}(\psi_1, \ldots, \psi_{t-1})^\top(\psi_1, \ldots, \psi_{t-1}) - I_{t-1}\Big\|_{\mathrm{op}} \lesssim \sqrt{\frac{t \log p}{p}}, \qquad \text{for every } 1 < t \leq n, \tag{224b}$$

with probability at least $1 - O(n^{-11})$.

We are ready to control the norm on the left hand side of inequality (194). First, recalling that $\{a_k\}$ forms an orthogonal basis, therefore there exists a unit vector

$$\omega := \sum_{k=1}^t \omega_k a_k \in \mathcal{S}^{t-1}, \tag{225}$$

— depending on the randomness of $\{\psi_k\}$ — such that

$$\begin{aligned}
&\Bigg\|\sum_{k=1}^t a_k\Big[\Big\langle\psi_k, F_{t+1}(\widehat{\beta}_{t+1})\Big\rangle - \langle F_{t+1}'(\widehat{\beta}_{t+1})\rangle\alpha_t^k - \sum_{j=k+1}^t \widehat{\gamma}_t^j \Big\langle F_{t+1}'(\widehat{\beta}_{t+1}) \circ F_j'\,(v_j)\Big\rangle\, \alpha_{j-1}^k\Big]\Bigg\|_2 \\
&= \omega^\top \sum_{k=1}^t a_k\Big[\Big\langle\psi_k, F_{t+1}(\widehat{\beta}_{t+1})\Big\rangle - \langle F_{t+1}'(\widehat{\beta}_{t+1})\rangle\alpha_t^k - \sum_{j=k+1}^t \widehat{\gamma}_t^j \Big\langle F_{t+1}'(\widehat{\beta}_{t+1}) \circ F_j'\,(v_j)\Big\rangle\, \alpha_{j-1}^k\Big] \\
&= \Bigg\langle \sum_{k=1}^t \omega_k\psi_k, F_{t+1}(\widehat{\beta}_{t+1})\Bigg\rangle - \langle F_{t+1}'(\widehat{\beta}_{t+1})\rangle \cdot \omega^\top\alpha_t - \sum_{j=1}^{t-1}\widehat{\gamma}_t^{j+1}\Big\langle F_{t+1}'(\widehat{\beta}_{t+1}) \circ F_{j+1}'\,(v_{j+1})\Big\rangle \sum_{k=1}^j \omega_k\alpha_j^k \\
&=: \sum_{i=1}^p X_i^0 + \sum_{i=1}^p X_i - \sum_{i=1}^p Y_i - \sum_{i=1}^p Z_i,
\end{aligned} \tag{226}$$

where to simplify our presentation, we introduce the following short-hand notation

$$X_i^0 := \Bigg[\sum_{k=1}^t \omega_k\psi_k \circ F_{t+1}\Big(\sum_{k=1}^t \widehat{\gamma}_t^k F_k(0)\Big)\Bigg]_i \tag{227a}$$

$$X_i := \Bigg[\sum_{k=1}^{t} \omega_k \psi_k \circ \Bigg(F_{t+1}(\widehat{\beta}_{t+1}) - F_{t+1}\Big(\sum_{k=1}^{t} \widehat{\gamma}_t^k F_k(0)\Big)\Bigg)\Bigg]_i \tag{227b}$$

$$Y_i := \frac{1}{n}\Big[F'_{t+1}(\widehat{\beta}_{t+1})\Big]_i \cdot \omega^\top \alpha_t \tag{227c}$$

$$Z_i := \frac{1}{n}\sum_{j=1}^{t-1} \widehat{\gamma}_t^{j+1} \Big[F'_{t+1}(\widehat{\beta}_{t+1}) \circ F'_{j+1}\left(v_{j+1}\right)\Big]_i \sum_{k=1}^{j} \omega_k \alpha_j^k. \tag{227d}$$

Here, in view of definitions in (52), we remind the readers that, in these expressions above, the parameters concerned are

$$\widehat{\theta} := \Big\{\omega, \{\alpha_k\}_{k\le t}, \widehat{\gamma}_t, \{\tau_k\}_{k\le t+1}\Big\}, \quad \text{where } \omega \in \mathcal{S}^{t-1}, \alpha_k \in \mathbb{R}^k, \widehat{\gamma}_t \in \mathbb{R}^t, \text{ and } \tau_k \in \mathbb{R}^s. \tag{228}$$

Let us point out a few properties for the above parameters:

- τ_k corresponds to the parameter used for defining function F_k. By assumption, τ_k is of finite and low dimension and $\|\tau_k\|_2 \le c$ for some universal constant;
- according to the assumption (191b), $\|\alpha_k\|_2 \le \overline{\alpha}_t \le \mathsf{poly}(n)$, for every $k \le t$;
- in view of the assumption (191a), $\|\widehat{\gamma}_t\|_2^2 = \sum_{k=1}^{t}(\widehat{\gamma}_t^k)^2 \le \mathsf{poly}(n)/(1-c^2)$.

The value of $\widehat{\theta}$ depends on the randomness in $\{\psi_k\}$; we collect all the possible values of $\widehat{\theta}$ to be space Θ, namely,

$$\Theta := \Big\{(\omega, \{\alpha_k\}_{k\le t}, \widehat{\gamma}_t, \{\tau_k\}_{k\le t+1}) \;\big|\; \omega \in \mathcal{S}^{t-1}, \|\tau_k\|_2 \le 1, \|\alpha_k\|_2 \le \mathsf{poly}(n), \|\widehat{\gamma}_t\|_2 \le \frac{\mathsf{poly}(n)}{1-c^2}\Big\}. \tag{229}$$

Next, let us control the right hand side of inequality (226), which shall be done by bounding each term in the summation separately.

E.1 Controlling (227a)

Regarding the first term $\sum_{i=1}^{p} X_i^0$, since $(\omega, \widehat{\gamma})$ has complicated statistical dependence on the randomness of $\{\psi_k\}$, we find it helpful to construct an ϵ-covering set $\mathcal{N}_\epsilon$ of the space $\mathcal{S}^{t-1}(1) \times \mathcal{S}^{t-1}(\mathsf{poly}(n))$ for $\epsilon = \frac{1}{\mathsf{poly}(n)}$, with its cardinality satisfying

$$|\mathcal{N}_\epsilon| \lesssim \Big(\frac{\mathsf{poly}(n)}{\epsilon}\Big)^t = \mathsf{poly}(n)^t.$$

Before diving into the main proof, we make note of the following property

$$\begin{aligned}\Big\|F_{t+1}\Big(\sum_{k=1}^{t} \widehat{\gamma}_t^k F_k(0)\Big)\Big\|_2 &\le \|F_{t+1}(0)\|_2 + \rho_F \Big\|\sum_{k=1}^{t} \widehat{\gamma}_t^k F_k(0)\Big\|_2 \\ &\le \|F_{t+1}(0)\|_2 + \rho_F \sum_{k=1}^{t} |\widehat{\gamma}_t^k| \cdot \|F_k(0)\|_2 \lesssim \overline{\gamma}_{t+1},\end{aligned} \tag{230}$$

where in the last inequality, we invoke assumption (191c) which implies $\rho_F \sum_{k=1}^{t} |\widehat{\gamma}_t^k| \ll 1$.

Given every $(\omega, \widehat{\gamma}_t) \in \mathcal{S}^{t-1}(1) \times \mathcal{S}^{t-1}(\mathsf{poly}(n))$, there exists $(\widetilde{\omega}, \widetilde{\gamma}_t) \in \mathcal{N}_\epsilon$ satisfying $\|\widetilde{\omega} - \omega\|_2 + \|\widetilde{\gamma}_t - \widehat{\gamma}_t\|_2 \le \epsilon = \frac{1}{\mathsf{poly}(n)}$. This fact together with the Lipschitz property of function F_t gives

$$\begin{aligned}&\Big|\Big\langle \sum_{k=1}^{t} \omega_k \psi_k, F_{t+1}\Big(\sum_{k=1}^{t} \widehat{\gamma}_t^k F_k(0)\Big)\Big\rangle - \Big\langle \sum_{k=1}^{t} \widetilde{\omega}_k \psi_k, F_{t+1}\Big(\sum_{k=1}^{t} \widetilde{\gamma}_t^k F_k(0)\Big)\Big\rangle\Big| \\ &\le \Big\|\sum_{k=1}^{t} \omega_k \psi_k\Big\|_2 \Big\|F_{t+1}\Big(\sum_{k=1}^{t} \widehat{\gamma}_t^k F_k(0)\Big) - F_{t+1}\Big(\sum_{k=1}^{t} \widetilde{\gamma}_t^k F_k(0)\Big)\Big\|_2 + \Big\|F_{t+1}\Big(\sum_{k=1}^{t} \widehat{\gamma}_t^k F_k(0)\Big)\Big\|_2 \Big\|\sum_{k=1}^{t} \omega_k \psi_k - \sum_{k=1}^{t} \widetilde{\omega}_k \psi_k\Big\|_2\end{aligned}$$

$$\overset{\text{(i)}}{\lesssim} \sqrt{\frac{p}{n}}\Big(1+\sqrt{\frac{t\log p}{p}}\Big)\cdot\Big(\rho_F\|\widehat{\gamma}_t-\widetilde{\gamma}_t\|_2\overline{\gamma}_t+\|\omega-\widetilde{\omega}\|_2\overline{\gamma}_{t+1}\Big)$$

$$\lesssim \sqrt{\frac{p}{n}}\Big(1+\sqrt{\frac{t\log p}{p}}\Big)\cdot\epsilon\cdot\rho_F\overline{\gamma}_{t+1}\lesssim\frac{1}{\mathsf{poly}(n)}\rho_F\overline{\gamma}_{t+1}, \tag{231}$$

with probability at least $1-O(n^{-11})$. Here, (i) follows from the fact that we condition on the event (224). Putting things together, we arrive at

$$\sum_{i=1}^{p} X_i^0 = \Big\langle \sum_{k=1}^{t}\omega_k\psi_k, F_{t+1}\Big(\sum_{k=1}^{t}\widehat{\gamma}_t^k F_k(0)\Big)\Big\rangle$$

$$\leq \sup_{(\widetilde{\omega},\widetilde{\gamma})\in\mathcal{N}_\epsilon}\Big\langle \sum_{k=1}^{t}\widetilde{\omega}_k\psi_k, F_{t+1}\Big(\sum_{k=1}^{t}\widetilde{\gamma}_t^k F_k(0)\Big)\Big\rangle+\frac{1}{\mathsf{poly}(n)}\rho_F\overline{\gamma}_{t+1}.$$

To further control the right hand side above, let us make note of the following two observations that (i) $\Big\langle\sum_{k=1}^{t}\widetilde{\omega}_k\psi_k, F_{t+1}\Big(\sum_{k=1}^{t}\widetilde{\gamma}_t^k F_k(0)\Big)\Big\rangle$ is stochastically dominated by $\mathcal{N}(0,\frac{\overline{\gamma}_{t+1}^2}{n})$ and (ii) the standard concentration result (see, e.g. (Wainwright, 2019, Exercise 2.12)).

$$\mathbb{P}\Bigg(\sup_{i\in[k]} X_i-\sqrt{2\sigma^2\log k}\geq t\Bigg)\leq 2e^{-\frac{t^2}{2\sigma^2}}, \tag{232}$$

for $X_i \overset{\text{i.i.d}}{\sim}\mathcal{N}(0,\sigma^2)$. Consequently, we conclude that

$$\sum_{i=1}^{n} X_i^0 \lesssim \sqrt{\frac{t\log(n)}{n}}\overline{\gamma}_{t+1}+\frac{1}{\mathsf{poly}(n)}\rho_F\overline{\gamma}_{t+1}, \tag{233}$$

with probability at least $1-O(n^{-11})$.

E.2 Controlling the remaining terms

For notational simplicity, if we concatenate $\{\psi_k\}_{k=1}^t$ into matrix Ψ, namely,

$$\Psi \coloneqq \big[\psi_1,\dots,\psi_t\big]\in\mathbb{R}^{p\times t}, \qquad \text{where } \psi_k \overset{\text{i.i.d.}}{\sim}\mathcal{N}(0,\frac{1}{n}I_p), \tag{234}$$

it suffices to control the summation of the remaining three terms as

$$H(\Psi;\widehat{\theta}) \coloneqq \sum_{i=1}^{p}X_i-\sum_{i=1}^{p}Y_i-\sum_{i=1}^{p}Z_i. \tag{235}$$

For any fixed parameter $\theta\in\Theta$ and fixed $i\in[p]$, it is easily seen from definitions (227), random vector (X_i,Y_i,Z_i) only depends on the i-th row of Ψ matrix, namely, $[\psi_{k,i}]_{k=1}^t$, which implies that $\{(X_i,Y_i,Z_i)\}_{i=1}^p$ are independent for different i. In addition, in view of Stein's lemma of Gaussian random vectors — which ensures $\mathbb{E}_{X\sim\mathcal{N}(0,1)}\big[Xf(X)\big]=\mathbb{E}_{X\sim\mathcal{N}(0,1)}\big[f'(X)\big]$ — one can verify

$$\mathbb{E}\big[H(\Psi;\theta)\big]=0. \tag{236}$$

However, the above property holds only when θ is held as a fixed vector, independent of $\{\psi_k\}$. When a random $\widehat{\theta}$ is concerned, due to the statistical dependence between $\widehat{\theta}$ and (X_i,Y_i,Z_i)'s, the mean zero property does not hold anymore.

On the high level, to control $H(\Psi;\widehat{\theta})$, the idea is to invoke Lemma 3 to bound $H(\Psi;\theta)$ for each fixed $\theta\in\Theta$, which is achieved via Step 1-3 below. In Step 4, we develop a uniform control of $H(\Psi;\theta)$ over the space Θ in order to deal with the statistical dependence involved in $\widehat{\theta}$.

In order to apply Lemma 3, it boils down to computing the variance of $H(\Psi;\theta)$ and validating the property (166), as shall be done as follows.

Step 1: variances control. We claim that for every fixed $\theta \in \Theta$ (independent of $\{\psi_k\}$), the variance terms satisfy the following relations respectively

$$\sum_{i=1}^{p} \mathsf{Var}(X_i) \lesssim \frac{\log n}{n}(\overline{\gamma}_{t+1}^2 + \rho_F^2 \overline{\alpha}_t^2) \tag{237a}$$

$$\sum_{i=1}^{p} \mathsf{Var}(Y_i) \lesssim \frac{\rho_F^2}{n^2} \left\|\alpha_t\right\|_2^2 \tag{237b}$$

$$\sum_{i=1}^{p} \mathsf{Var}(Z_i) \lesssim \frac{\rho_F^2}{n^2} \left\|\alpha_t\right\|_2^2 . \tag{237c}$$

Given the above relations, the variance of $H(\Psi;\theta)$ satisfies

$$\mathrm{var}(H(\Psi;\theta)) \lesssim \frac{\log n}{n}(\overline{\gamma}_{t+1}^2 + \rho_F^2 \overline{\alpha}_t^2).$$

Let us establish these three claims respectively. We remark that throughout this step, θ should always be viewed as a fixed constant that does not dependent on any randomness of the problem. With a slight abuse of notation, we still write parameters such as α_t, $\widehat{\beta}_t$, $\widehat{\gamma}_t$, but here they should be understood as fixed parameters.

- First, by noticing $|F'_{t+1}(\widehat{\beta}_{t+1})| \le \rho_F$ and $|\omega^\top \alpha_t| \le \|\alpha_t\|_2$, we obtain

$$\sum_{i=1}^{p} \mathsf{Var}(Y_i) \le \frac{1}{n^2}\sum_{i=1}^{p} \mathbb{E}[F'_{t+1}(\widehat{\beta}_{t+1})]_i^2 \left\|\alpha_t\right\|_2^2 \lesssim \frac{\rho_F^2}{n} \left\|\alpha_t\right\|_2^2 ,$$

  which establishes relation (237b).

- In terms of the relation (237c), for each fixed $\theta \in \Theta$, we remind the readers that $[F'_{t+1}(\widehat{\beta}_{t+1}) \circ F'_{j+1}\left(v_{j+1}\right)]_i$ are independent of each other. This fact leads to

$$\begin{aligned}
\sqrt{\sum_{i=1}^{p} \mathsf{Var}(Z_i)} &= \frac{1}{n}\sqrt{\sum_{i=1}^{p} \mathsf{Var}\left(\sum_{j=1}^{t-1}\Big(\sum_{k=1}^{j}\omega_k\alpha_j^k\Big)\cdot\widehat{\gamma}_t^{j+1}\left[F'_{t+1}(\widehat{\beta}_{t+1}) \circ F'_{j+1}\left(v_{j+1}\right)\right]_i\right)} \\
&= \frac{1}{n}\sqrt{\mathsf{Var}\left(\sum_{j=1}^{t-1}\Big(\sum_{k=1}^{j}\omega_k\alpha_j^k\Big)\cdot\widehat{\gamma}_t^{j+1}\sum_{i=1}^{p}\left[F'_{t+1}(\widehat{\beta}_{t+1}) \circ F'_{j+1}\left(v_{j+1}\right)\right]_i\right)} \\
&= \frac{1}{n}\sqrt{\mathbb{E}\left[\sum_{j=1}^{t-1}\Big(\sum_{k=1}^{j}\omega_k\alpha_j^k\Big)\cdot\widehat{\gamma}_t^{j+1}\sum_{i=1}^{p}\left[F'_{t+1}(\widehat{\beta}_{t+1}) \circ F'_{j+1}\left(v_{j+1}\right)\right]_i - \mathbb{E}\left[F'_{t+1}(\widehat{\beta}_{t+1}) \circ F'_{j+1}\left(v_{j+1}\right)\right]_i\right]^2} \\
&\le \frac{1}{n}\sum_{j=1}^{t-1}\Big|\sum_{k=1}^{j}\omega_k\alpha_j^k\cdot\widehat{\gamma}_t^{j+1}\Big|\sqrt{\mathbb{E}\left[\sum_{i=1}^{p}\left[F'_{t+1}(\widehat{\beta}_{t+1}) \circ F'_{j+1}\left(v_{j+1}\right)\right]_i - \mathbb{E}\left[F'_{t+1}(\widehat{\beta}_{t+1}) \circ F'_{j+1}\left(v_{j+1}\right)\right]_i\right]^2},
\end{aligned}$$

  where the last used the basic property that $\sqrt{\mathbb{E}[\sum_i X_i]^2} \le \sum_i \sqrt{\mathbb{E}[X_i^2]}$. To avoid confusion, we remind the readers that the parameters here are treated as fixed and independent of the randomness in the problem. Now, in view of the basic inequality $|\sum_{k=1}^{j}\omega_k\alpha_j^k| \le \|\alpha_j\|_2$, we can further bound

$$\begin{aligned}
\sqrt{\sum_{i=1}^{p} \mathsf{Var}(Z_i)} &\le \frac{1}{n}\sum_{j=1}^{t-1}|\widehat{\gamma}_t^{j+1}|\cdot\|\alpha_j\|_2\sqrt{\mathbb{E}\left\|F'_{t+1}(\widehat{\beta}_{t+1}) \circ F'_{j+1}\left(v_{j+1}\right)\right\|_2^2} \\
&\le \frac{1}{n}\sum_{j=1}^{t-1}|\widehat{\gamma}_t^{j+1}|\|\alpha_j\|_2\rho_F^2 \lesssim \frac{\rho_F}{n}\left\|\alpha_t\right\|_2 ,
\end{aligned} \tag{238}$$

  where the last inequality invokes the assumption (191c). This completes the proof of inequality (237c).

- When it comes to the inequality (237a), basic inequality ensures that

$$\sum_{i=1}^{p} \mathsf{Var}(X_i) \lesssim \sum_{i=1}^{p} \mathsf{Var}(X_i^0 + X_i) + \sum_{i=1}^{p} \mathsf{Var}(X_i^0).$$

  We shall compute these two terms on the right respectively. Note that for fixed $\omega \in \mathcal{S}^{t-1}$ independent of $\{\psi_k\}$, $\sum_k \omega_k \psi_k \sim \mathcal{N}(0, \frac{1}{n} I_p)$, therefore

$$\sum_{i=1}^{p} \mathsf{Var}(X_i^0) \le \mathbb{E} \left\| \sum_{k=1}^{t} \omega_k \psi_k \circ F_{t+1}\Big(\sum_{k=1}^{t} \widehat{\gamma}_t^k F_k(0) \Big) \right\|_2^2 = \frac{1}{n} \left\| F_{t+1}\Big(\sum_{k=1}^{t} \widehat{\gamma}_t^k F_k(0) \Big) \right\|_2^2 \lesssim \frac{1}{n} \overline{\gamma}_{t+1}^2,$$

  where the last inequality makes use of (230).

  Lemma 9. *Under the assumptions of Lemma 7, it satisfies*

$$\sum_{i=1}^{p} \mathsf{Var}(X_i^0 + X_i) \lesssim \frac{\log n}{n} (\overline{\gamma}_{t+1}^2 + \rho_F^2 \overline{\alpha}_t^2). \tag{239}$$

  The proof of this result is postponed to Section F.3.

  Putting these pieces together proves the claimed result in inequality (237a).

Step 2: sizes control. To apply Lemma 3, one needs to check condition (166). Since the zero mean condition holds as in (236), it only requires us to provide a high probability control on the size of $H(\Psi, \theta)$ for each $\theta \in \Theta$. In the following, we bound the sizes of X_i, Y_i and Z_i respectively.

- By definition of X_i in expression (227b), we claim that

$$\begin{aligned} |X_i| &\le \left| \sum_{k=1}^{t} \omega_k \psi_{k,i} \right| \cdot \left| F_{t+1}(\widehat{\beta}_{t+1}) - F_{t+1}\left(\sum_{k=1}^{t} \widehat{\gamma}_t^k F_k(0) \right) \right|_i \\ &\lesssim \sqrt{\frac{\log \frac{1}{\delta}}{n}} \cdot \rho_F \sqrt{\frac{\log \frac{1}{\delta}}{n}} \|\alpha_t\|_2 \lesssim \frac{\log \frac{1}{\delta}}{n} \rho_F \overline{\alpha}_t, \end{aligned} \tag{240}$$

  with probability at least $1 - \delta$, for every $\delta \lesssim \frac{1}{t}$.

  In order to see this, first notice that for every $k \ge 1$ and $t \le n$, $v_k \sim \mathcal{N}(0, \frac{\|\alpha_{k-1}\|_2^2}{n})$ and $\sum_{k=1}^{t} \omega_k \psi_{k,i} \sim \mathcal{N}(0, \frac{1}{n})$. Standard concentration inequality ensures that

$$\left| \sum_{k=1}^{t} \omega_k \psi_{k,i} \right| \lesssim \sqrt{\frac{\log \frac{1}{\delta}}{n}} \qquad \text{and } \|v_k\|_\infty \lesssim \|\alpha_{k-1}\|_2 \sqrt{\frac{\log \frac{1}{\delta}}{n}},$$

  with probability at least $1 - \delta$. In addition, taking the Lipschitz property of F_{t+1} together with the above concentration results ensures

$$\begin{aligned} \left| F_{t+1}(\widehat{\beta}_{t+1}) - F_{t+1}\left(\sum_{k=1}^{t} \widehat{\gamma}_t^k F_k(0) \right) \right|_i &\le \rho_F \left| \widehat{\beta}_{t+1} - \sum_{k=1}^{t} \widehat{\gamma}_t^k F_k(0) \right|_i \\ &= \rho_F \left| v_{t+1} + \sum_{k=1}^{t} \widehat{\gamma}_t^k F_k(v_k) - \sum_{k=1}^{t} \widehat{\gamma}_t^k F_k(0) \right|_i \\ &\lesssim \rho_F \sqrt{\frac{\log \frac{1}{\delta}}{n}} \left[\|\alpha_t\|_2 + \rho_F \sum_{k=1}^{t} \widehat{\gamma}_t^k \|\alpha_{k-1}\|_2 \right] \end{aligned}$$

$$\lesssim \rho_F \sqrt{\frac{\log \frac{1}{\delta}}{n}} \overline{\alpha}_t, \tag{241}$$

with probability at least $1-\delta$. Here the last inequality uses the definition (186a) and the condition (191c), Putting everything together completes the proof of (240).

- As for Y_i, in view of the definition (227c), some direct algebra leads to

$$|Y_i| = \left| \frac{1}{n} \left[F'_{t+1}(\widehat{\beta}_{t+1}) \right]_i \cdot \omega^\top \alpha_t \right| \leq \frac{1}{n} \left\| F'_{t+1}(\widehat{\beta}_{t+1}) \right\|_\infty \|\alpha_t\|_2 \leq \frac{\rho_F}{n} \|\alpha_t\|_2.$$

- Regarding Z_i, it is easily seen that

$$\begin{aligned} |Z_i| &= \left| \frac{1}{n} \sum_{j=1}^{t-1} \widehat{\gamma}_t^{j+1} \left[F'_{t+1}(\widehat{\beta}_{t+1}) \circ F'_{j+1}\left(v_{j+1}\right) \right]_i \sum_{k=1}^{j} \omega_k \alpha_j^k \right| \\ &\leq \frac{1}{n} \sum_{j=1}^{t-1} |\widehat{\gamma}_t^{j+1}| \cdot \left\| F'_{t+1}(\widehat{\beta}_{t+1}) \circ F'_{j+1}\left(v_{j+1}\right) \right\|_\infty \|\alpha_j\|_2 \leq \frac{\rho_F^2}{n} \sum_{j=1}^{t-1} |\widehat{\gamma}_t^{j+1}| \|\alpha_j\|_2 \lesssim \frac{\rho_F}{n} \left\|\alpha_t\right\|_2, \end{aligned}$$

where the last inequality follows from the condition $\rho_F \sum_{k=1}^{t} |\widehat{\gamma}_t^k| \ll 1$.

Combining the pieces above, we are ensured that

$$\mathbb{P}\left(|X_i| + |Y_i| + |Z_i| \lesssim \frac{\rho_F \overline{\alpha}_t}{n} \log \frac{1}{\delta} \right) \geq 1 - \delta. \tag{242}$$

Step 3: Putting everything together. Equipped with the above variances control and sizes control, a direct application of Lemma 3 yields that for each $\theta \in \Theta$ independent of the problem randomnesses,

$$|H(\Psi;\theta)| = \left| \sum_{i=1}^{p} X_i - \sum_{i=1}^{p} Y_i - \sum_{i=1}^{p} Z_i \right| \lesssim \sqrt{\frac{\log n \log \frac{1}{\delta}}{n}} (\overline{\gamma}_{t+1} + \rho_F \overline{\alpha}_t) \tag{243}$$

with probability at least $1-\delta$.

Step 4: a covering argument. Thus far, we have established an upper bound for $|H(\Psi;\theta)|$ regarding any fixed $\theta \in \Theta$. In the following, we aim to develop a control of $|H(\Psi;\theta)|$ uniformly over the space Θ via a standard covering argument in addition to a uniform bound. A direct covering of space Θ requires a covering number of order $O\Big((\frac{1}{\epsilon})^{t^2} \Big)$, which leads to a squared dependence of t in expression (194). Next, we show that it is sufficient to construct a set of cardinality $O\Big((\frac{1}{\epsilon})^{t \log n} \Big)$, where function $H(\Psi;\cdot)$ lies at most ϵ apart from each other. Implementing this idea leads to the right dependence of t in expression (194).

Before proceeding, let us denote $\mathcal{M}_\epsilon$ as an ϵ-net for a subset of Θ (referred to as Θ_0) where

$$\Theta_0 := \left\{ \Big(\omega, \{\alpha_k\}_{k\leq t}, \widetilde{\gamma}_t, \{\tau_k\}_{k\leq t+1} \Big) \mid \text{for every } \Big(\omega, \{\alpha_k\}_{k\leq t}, \widehat{\gamma}_t, \{\tau_k\}_{k\leq t+1} \Big) \in \Theta \right\}, \tag{244}$$

where for each $k \leq t$,

$$\widetilde{\gamma}_t^k = \begin{cases} 0, & \text{for } k \leq t - O(\log n), \\ \widehat{\gamma}_t^k & \text{o.w.} \end{cases} \tag{245}$$

In words, Θ_0 stands for the subset of Θ where the corresponding $\widehat{\gamma}_t$ is restricted to have zero entries except for the last $O(\log n)$ coordinates, namely, $\widetilde{\gamma}_t^k = 0$ for $k \leq t - O(\log n)$. For every $\widehat{\theta} = \Big(\omega, \{\alpha_k\}_{k\leq t}, \widehat{\gamma}_t, \{\tau_k\}_{k\leq t+1} \Big) \in \Theta$, assumption (191a) ensures that $|\widehat{\gamma}_t^k| \leq \epsilon$ for $k \leq t - O(\log n)$. Therefore, there exists some $\widetilde{\theta} \in \mathcal{M}_\epsilon$ with

$$\widetilde{\theta} := \Big(\widetilde{\omega}, \{\widetilde{\alpha}_k\}_{k\leq t}, \widetilde{\gamma}_t, \{\widetilde{\tau}_k\}_{k\leq t+1} \Big),$$

such that $\|\omega - \widetilde{\omega}\|_2 \le \epsilon$, and for every k, $\|\alpha_k - \widetilde{\alpha}_k\|_2 \le \epsilon$, $\|\tau_k - \widetilde{\tau}_k\|_2 \le \epsilon$ and $|\widehat{\gamma}_t^k - \widetilde{\gamma}_t^k| \le \epsilon$. We claim that

$$|H(\Psi;\widehat{\theta}) - H(\Psi;\widetilde{\theta})| \lesssim \frac{t\log^2 n}{n}(\overline{\gamma}_{t+1} + \rho_F\overline{\alpha}_t). \tag{246}$$

Let us take inequality (246) as given for the moment and come back to its proof after establishing our main result. It turns out that given the structure of Θ_0, it is sufficient to consider a subset $\Theta_0' \subset \Theta_0$ where the corresponding $\alpha_k = 0 \in \mathbb{R}^k$ for $k \le t - O(\log n)$. More specifically, define set

$$\Theta_0' := \left\{ \widetilde{\theta}' = (\widetilde{\omega}', \{\widetilde{\alpha}_k'\}_{k\le t}, \widetilde{\gamma}_t', \{\widetilde{\tau}_k'\}_{k\le t+1}) \in \Theta_0 \;\middle|\; \text{for every } \widetilde{\theta} = \Big(\widetilde{\omega}, \{\widetilde{\alpha}_k\}_{k\le t}, \widetilde{\gamma}_t, \{\widetilde{\tau}_k\}_{k\le t+1}\Big) \in \mathcal{M}_\epsilon \right\}, \tag{247}$$

where

$$\widetilde{\omega}' = \widetilde{\omega},\; \widetilde{\gamma}_t' = \widetilde{\gamma}_t,\; \widetilde{\tau}_k' = \widetilde{\tau}_k,\; \forall k \le t+1,$$
$$\widetilde{\alpha}_k' = 0 \in \mathbb{R}^k \text{ for } k \le t - O(\log n), \quad \widetilde{\alpha}_k' = \widetilde{\alpha}_k \text{ for } t - O(\log n) \le k \le t.$$

Given a $\widetilde{\theta}$ and $\widetilde{\theta}'$ pair, since the corresponding $\widetilde{\gamma}_t^k = \widetilde{\gamma}_t'^k = 0$, we are guaranteed that

$$\widetilde{\beta}_{t+1} := v_{t+1} + \sum_{k=1}^t \widetilde{\gamma}_t^k F_k(v_k) = v_{t+1} + \sum_{k=t-O(\log n)}^t \widetilde{\gamma}_t'^k F_k(v_k) = \widetilde{\beta}_{t+1}', \tag{248}$$

as v_{k+1} is determined by $\alpha_k \in \mathbb{R}^k$. In addition, in view of the definition of H function (cf. (235)), the corresponding H functions have the same value as

$$H(\Psi;\widetilde{\theta}) = H(\Psi;\widetilde{\theta}'). \tag{249}$$

These observations imply that it is enough to restrict to set Θ_0' when consider a covering for function $H(\Psi;\theta)$. By construction, the cardinality of Θ_0' equals to

$$|\Theta_0'| \lesssim \left(\frac{1}{\epsilon}\right)^{t\log n}, \tag{250}$$

which yields a much small size compared to $|\mathcal{M}_\epsilon|$.

Armed with the properties above, we are ready to control $\sup_{\theta\in\Theta} H(\Psi;\theta)$. Specifically, as a consequence of relation (246), we find that

$$\begin{aligned}\sup_{\theta\in\Theta} H(\Psi;\theta) &\le \sup_{\widetilde{\theta}\in\mathcal{M}_\epsilon} H(\Psi;\widetilde{\theta}) + C\frac{t\log^2 n}{n}(\overline{\gamma}_{t+1} + \rho_F\overline{\alpha}_t)\\ &= \sup_{\widetilde{\theta}'\in\Theta_0'} H(\Psi;\widetilde{\theta}') + C\frac{t\log^2 n}{n}(\overline{\gamma}_{t+1} + \rho_F\overline{\alpha}_t),\end{aligned}$$

for some universal constant C. Here the last equality follows from property (249). Now in order to control quantity $\sup_{\widetilde{\theta}'\in\Theta_0'} H(\Psi;\widetilde{\theta}')$, recall that we have shown that for every fixed $\theta \in \Theta$, (243) holds true with probability at least $1-\delta$. Now setting $\delta = n^{-11}\epsilon^{t\log n}$ and taking a uniform bound over $|\Theta_0'|$ ensure

$$\sup_{\widetilde{\theta}'\in\Theta_0'} H(\Psi;\widetilde{\theta}') \lesssim \sqrt{\frac{t\log^2 n}{n}}(\overline{\gamma}_{t+1} + \rho_F\overline{\alpha}_t), \tag{251}$$

and hence,

$$\sup_{\theta\in\Theta} H(\Psi;\theta) \le \sup_{\widetilde{\theta}\in\mathcal{M}_\epsilon} H(\Psi;\widetilde{\theta}) + C\frac{t\log^2 n}{n}(\overline{\gamma}_{t+1} + \rho_F\overline{\alpha}_t) \lesssim \sqrt{\frac{t\log^2 n}{n}}(\overline{\gamma}_{t+1} + \rho_F\overline{\alpha}_t), \tag{252}$$

with probability at least $1 - O(n^{-11})$.

In summary. Putting together inequality (226) with (233) and (252), we conclude that

$$
\begin{aligned}
&\sup_{\theta\in\Theta}\left\|\sum_{k=1}^{t} a_k\Big[\Big\langle \psi_k, F_{t+1}(\widehat{\beta}_{t+1})\Big\rangle - \langle F'_{t+1}(\widehat{\beta}_{t+1})\rangle\alpha_t^k - \sum_{j=k+1}^{t}\widehat{\gamma}_t^j\Big\langle F'_{t+1}(\widehat{\beta}_{t+1})\circ F'_j(v_j)\Big\rangle\alpha_{j-1}^k\Big]\right\|_2 \\
&\le \sup_{\theta\in\Theta}\sum_{i=1}^{p} X_i^0 + \sum_{i=1}^{p} X_i - \sum_{i=1}^{p} Y_i - \sum_{i=1}^{p} Z_i \\
&\le \sup_{\theta\in\Theta}\sum_{i=1}^{p} X_i^0 + \sup_{\theta\in\Theta} H(\Psi;\theta) \\
&\lesssim \sqrt{\frac{t\log^2 n}{n}}(\overline{\gamma}_{t+1} + \rho_F\overline{\alpha}_t),
\end{aligned}
$$

with probability at least $1 - O(n^{-11})$. Thus, we complete the proof of the targeted bound (194).

E.3 Other auxiliary details for Lemma 7

Proof of inequality (246). Let us begin by considering quantity $\sum_{i=1}^{p}(X_i - \widetilde{X}_i)$ with X_i and $\widetilde{X}_i$ associated with $\widehat{\theta}$ and $\widetilde{\theta}$ respectively. Here $\widehat{\theta} = \Big(\omega, \{\alpha_k\}_{k\le t}, \widehat{\gamma}_t, \{\tau_k\}_{k\le t+1}\Big)$ and $\widetilde{\theta} = \Big(\widetilde{\omega}, \{\widetilde{\alpha}_k\}_{k\le t}, \widetilde{\gamma}_t, \{\widetilde{\tau}_k\}_{k\le t+1}\Big)$ satisfy

$$\|\omega - \widetilde{\omega}\|_2 \le \epsilon,$$

and for every k,

$$\|\alpha_k - \widetilde{\alpha}_k\|_2 \le \epsilon, \quad \|\tau_k - \widetilde{\tau}_k\|_2 \le \epsilon, \quad |\widehat{\gamma}_t^k - \widetilde{\gamma}_t^k| \le \epsilon.$$

We aim to show that

$$\sum_{i=1}^{p}(X_i - \widetilde{X}_i) \lesssim \frac{1}{\mathsf{poly}(n)}\rho_F(\overline{\alpha}_t + \overline{\gamma}_{t+1}) \asymp \frac{1}{\mathsf{poly}(n)}(\rho_F\overline{\alpha}_t + \overline{\gamma}_{t+1}). \tag{253}$$

As already shown by inequality (231), for every $(\omega, \widehat{\gamma}_t)$ and $(\widetilde{\omega}, \widetilde{\gamma}_t)$ pairs satisfying $\|\widetilde{\omega} - \omega\|_2 + \|\widetilde{\gamma}_t - \widehat{\gamma}_t\|_2 \le \epsilon = \frac{1}{\mathsf{poly}(n)}$, one has

$$\left|\Big\langle\sum_{k=1}^{t}\omega_k\psi_k, F_{t+1}\Big(\sum_{k=1}^{t}\widehat{\gamma}_t^k F_k(0)\Big)\Big\rangle - \Big\langle\sum_{k=1}^{t}\widetilde{\omega}_k\psi_k, F_{t+1}\Big(\sum_{k=1}^{t}\widetilde{\gamma}_t^k F_k(0)\Big)\Big\rangle\right| \lesssim \frac{1}{\mathsf{poly}(n)}\rho_F\overline{\gamma}_{t+1}. \tag{254}$$

Similarly, the Lipschitz property of function F_t ensures that

$$
\begin{aligned}
&\left|\Big\langle\sum_{k=1}^{t}\omega_k\psi_k, F_{t+1}\Big(\widehat{\beta}_{t+1}\Big)\Big\rangle - \Big\langle\sum_{k=1}^{t}\widetilde{\omega}_k\psi_k, F_{t+1}\Big(\widetilde{\beta}_{t+1}\Big)\Big\rangle\right| \\
&\le \Big\|\sum_{k=1}^{t}\omega_k\psi_k\Big\|_2\Big\|F_{t+1}\Big(\widehat{\beta}_{t+1}\Big) - F_{t+1}\Big(\widetilde{\beta}_{t+1}\Big)\Big\|_2 + \Big\|F_{t+1}\Big(\widetilde{\beta}_{t+1}\Big)\Big\|_2\Big\|\sum_{k=1}^{t}\omega_k\psi_k - \sum_{k=1}^{t}\widetilde{\omega}_k\psi_k\Big\|_2 \\
&\overset{\text{(i)}}{\lesssim} \frac{p}{n}\Big(1 + \sqrt{\frac{t\log p}{p}}\Big)\Big(\rho_F\|\widehat{\beta}_{t+1} - \widetilde{\beta}_{t+1}\|_2 + \Big\|F_{t+1}\Big(\widetilde{\beta}_{t+1}\Big)\Big\|_2\|\omega - \widetilde{\omega}\|_2\Big) && (255) \\
&\lesssim \frac{1}{\mathsf{poly}(n)}\rho_F\overline{\alpha}_t. && (256)
\end{aligned}
$$

Here (i) follows from the event (224) and we leave the proof of (256) to the end of this step. Putting these two relations together yields the claimed bound in (253).

Regarding quantity $\sum_{i=1}^{p}(Y_i - \widetilde{Y}_i)$, some direct algebra shows that

$$\begin{aligned}
\Big|\sum_{i=1}^{p}(Y_i - \widetilde{Y}_i)\Big| &= \Big|\langle F'_{t+1}(\widehat{\beta}_{t+1})\rangle \cdot \widehat{\omega}^\top \widehat{\alpha}_t - \langle F'_{t+1}(\widetilde{\beta}_{t+1})\rangle \cdot \widetilde{\omega}^\top \widetilde{\alpha}_t\Big| \\
&\le \Big|\langle F'_{t+1}(\widehat{\beta}_{t+1})\rangle\Big| \cdot |\widehat{\omega}^\top \widehat{\alpha}_t - \widetilde{\omega}^\top \widetilde{\alpha}_t| + \Big|\langle F'_{t+1}(\widehat{\beta}_{t+1})\rangle - \langle F'_{t+1}(\widetilde{\beta}_{t+1})\rangle\Big| \cdot \Big|\widetilde{\omega}^\top \widetilde{\alpha}_t\Big| \\
&\le \rho_F \cdot (\|\widehat{\alpha}_t - \widetilde{\alpha}_t\|_2 + \|\widehat{\omega} - \widetilde{\omega}\|_2) + \Big|\langle F'_{t+1}(\widehat{\beta}_{t+1})\rangle - \langle F'_{t+1}(\widetilde{\beta}_{t+1})\rangle\Big| \cdot \|\widetilde{\alpha}_t\|_2.
\end{aligned}$$

In order to bound the right hand side above, Lemma 10 provides an upper bound of quantity $\Big|\langle F'_{t+1}(\widehat{\beta}_{t+1})\rangle - \langle F'_{t+1}(\widetilde{\beta}_{t+1})\rangle\Big|$ in expression (277b). Taking this upper bound together with the fact that $\|\widehat{\alpha}_t - \widetilde{\alpha}_t\|_2, \|\widehat{\omega} - \widetilde{\omega}\|_2 \le \epsilon$, we obtain

$$\begin{aligned}
\Big|\sum_{i=1}^{p}(Y_i - \widetilde{Y}_i)\Big| &\lesssim \rho_F \epsilon + \Big(\frac{t\log^2 n}{n}\rho_F + \frac{1}{\mathsf{poly}(n)}\rho_{1,F}\|\alpha_t\|_2\Big) \cdot \|\widetilde{\alpha}_t\|_2 \\
&\lesssim \frac{t\log^2 n}{n}\rho_F \overline{\alpha}_t,
\end{aligned} \tag{257}$$

where the last step invokes assumption (191b) and $\epsilon = 1/\mathsf{poly}(n)$.

It remains to consider quantity $\sum_{i=1}^{p}(Z_i - \widetilde{Z}_i)$. Recalling the definition of Z_i (cf. (227d)), we can write

$$\begin{aligned}
&\Big|\sum_{i=1}^{p}(Z_i - \widetilde{Z}_i)\Big| \\
=&\Big|\sum_{j=1}^{t-1}\widehat{\gamma}_t^{j+1}\Big\langle F'_{t+1}(\widehat{\beta}_{t+1}) \circ F'_{j+1}(v_{j+1})\Big\rangle \sum_{k=1}^{j}\widehat{\omega}_k\widehat{\alpha}_j^k - \sum_{j=1}^{t-1}\widetilde{\gamma}_t^{j+1}\Big\langle F'_{t+1}(\widetilde{\beta}_{t+1}) \circ F'_{j+1}(v_{j+1})\Big\rangle \sum_{k=1}^{j}(\widetilde{\omega}_k\widetilde{\alpha}_j^k\Big| \\
=&\Big|\sum_{j=1}^{t-1}\widehat{\gamma}_t^{j+1}\sum_{k=1}^{j}\widehat{\omega}_k\widehat{\alpha}_j^k\Big(\Big\langle F'_{t+1}(\widehat{\beta}_{t+1}) \circ F'_{j+1}(v_{j+1})\Big\rangle - \Big\langle F'_{t+1}(\widetilde{\beta}_{t+1}) \circ F'_{j+1}(v_{j+1})\Big\rangle\Big) \\
&\quad + \sum_{j=1}^{t-1}\Big(\widehat{\gamma}_t^{j+1}\sum_{k=1}^{j}\widehat{\omega}_k\widehat{\alpha}_j^k - \widetilde{\gamma}_t^{j+1}\sum_{k=1}^{j}\widetilde{\omega}_k\widetilde{\alpha}_j^k\Big)\Big\langle F'_{t+1}(\widetilde{\beta}_{t+1}) \circ F'_{j+1}(v_{j+1})\Big\rangle\Big| \\
\le&\sum_{j=1}^{t-1}|\widehat{\gamma}_t^{j+1}|\|\widehat{\alpha}_j\|_2 \cdot \Big|\Big\langle F'_{t+1}(\widehat{\beta}_{t+1}) \circ F'_{j+1}(v_{j+1})\Big\rangle - \Big\langle F'_{t+1}(\widetilde{\beta}_{t+1}) \circ F'_{j+1}(v_{j+1})\Big\rangle\Big| \\
&\quad + \sum_{j=1}^{t-1}\Big|\widehat{\gamma}_t^{j+1}\sum_{k=1}^{j}\widehat{\omega}_k\widehat{\alpha}_j^k - \widetilde{\gamma}_t^{j+1}\sum_{k=1}^{j}\widetilde{\omega}_k\widetilde{\alpha}_j^k\Big| \cdot \Big|\Big\langle F'_{t+1}(\widetilde{\beta}_{t+1}) \circ F'_{j+1}(v_{j+1})\Big\rangle\Big|.
\end{aligned} \tag{258}$$

It is then sufficient to bound the two terms above respectively. First, note that

$$\begin{aligned}
\Big|\widehat{\gamma}_t^{j+1}\sum_{k=1}^{j}\widehat{\omega}_k\widehat{\alpha}_j^k - \widetilde{\gamma}_t^{j+1}\sum_{k=1}^{j}\widetilde{\omega}_k\widetilde{\alpha}_j^k\Big| &= |\widehat{\gamma}_t^{j+1}| \cdot \Big|\sum_{k=1}^{j}\widehat{\omega}_k\widehat{\alpha}_j^k - \sum_{k=1}^{j}\widetilde{\omega}_k\widetilde{\alpha}_j^k\Big| + |\sum_{k=1}^{j}\widetilde{\omega}_k\widetilde{\alpha}_j^k| \cdot \Big|\widehat{\gamma}_t^{j+1} - \widetilde{\gamma}_t^{j+1}\Big| \\
&\le |\widehat{\gamma}_t^{j+1}|(\|\widehat{\alpha}_j - \widetilde{\alpha}_j\|_2 + \|\widehat{\omega} - \widetilde{\omega}\|_2) + \|\widetilde{\alpha}_j\|_2 \cdot \Big|\widehat{\gamma}_t^{j+1} - \widetilde{\gamma}_t^{j+1}\Big| \\
&\lesssim \epsilon \cdot (|\widehat{\gamma}_t^{j+1}| + \|\widetilde{\alpha}_j\|_2),
\end{aligned}$$

where the last step uses the relation between $\widehat{\theta}$ and $\widetilde{\theta}$. Therefore, the second term of (258) satisfies

$$\sum_{j=1}^{t-1}\Big|\widehat{\gamma}_t^{j+1}\sum_{k=1}^{j}\widehat{\omega}_k\widehat{\alpha}_j^k - \widetilde{\gamma}_t^{j+1}\sum_{k=1}^{j}\widetilde{\omega}_k\widetilde{\alpha}_j^k\Big| \cdot \Big|\Big\langle F'_{t+1}(\widetilde{\beta}_{t+1}) \circ F'_{j+1}(v_{j+1})\Big\rangle\Big| \le \frac{\rho_F^2}{\mathsf{poly}(n)}\sum_{j=1}^{t-1}(|\widehat{\gamma}_t^{j+1}| + \|\widetilde{\alpha}_j\|_2)$$

$$\leq \frac{\rho_F}{\mathsf{poly}(n)} \overline{\alpha}_t, \tag{259}$$

recognizing the uniform bound $|F_k'| \leq \rho_F$.

When it comes to the first term in expression (258), Lemma 10 controls the difference $|\langle F'_{t+1}(\widehat{\beta}_{t+1}) \circ F'_{j+1}(v_{j+1})\rangle - \langle F'_{t+1}(\widetilde{\beta}_{t+1}) \circ F'_{j+1}(v_{j+1})\rangle|$ in expression (277a); invoking this bound, we arrive at

$$\begin{aligned}
&\sum_{j=1}^{t-1} |\widehat{\gamma}_t^{j+1}| \|\widehat{\alpha}_j\|_2 \cdot \Big| \Big\langle F'_{t+1}(\widehat{\beta}_{t+1}) \circ F'_{j+1}(v_{j+1}) \Big\rangle - \Big\langle F'_{t+1}(\widetilde{\beta}_{t+1}) \circ F'_{j+1}(v_{j+1}) \Big\rangle \Big| \\
&\lesssim \sum_{j=1}^{t-1} |\widehat{\gamma}_t^{j+1}| \|\widehat{\alpha}_j\|_2 \Big(\frac{t \log^2 n}{n} \rho_F^2 + \frac{1}{\mathsf{poly}(n)} \rho_F \rho_{1,F} \|\alpha_t\|_2 \Big) \\
&\lesssim \frac{t \log^2 n}{n} \rho_F \overline{\alpha}_t + \frac{1}{\mathsf{poly}(n)} \rho_{1,F} \overline{\alpha}_t^2 \lesssim \frac{t \log^2 n}{n} \rho_F \overline{\alpha}_t,
\end{aligned}$$

where we plug in the assumptions (191b) and (191c). Combining pieces together, we conclude that

$$\Big| \sum_{i=1}^{p} (Z_i - \widetilde{Z}_i) \Big| \lesssim \frac{t \log^2 n}{n} \rho_F \overline{\alpha}_t. \tag{260}$$

To conclude, by combining the three parts above in expressions (253), (257) and (260) together, we end up with

$$|H(\Psi; \widehat{\theta}) - H(\Psi; \widetilde{\theta})| \leq \Big| \sum_{i=1}^{p} (X_i - \widetilde{X}_i) \Big| + \Big| \sum_{i=1}^{p} (Y_i - \widetilde{Y}_i) \Big| + \Big| \sum_{i=1}^{p} (Z_i - \widetilde{Z}_i) \Big| \lesssim \frac{t \log^2 n}{n} (\overline{\gamma}_{t+1} + \rho_F \overline{\alpha}_t),$$

thus completing the inequality (246).

Proof of inequality (256). First, conditioning on event (224), we make note of the following two relations where

$$\|F_k(v_k)\|_2 \leq \|F_k(0)\|_2 + \rho_F \|v_k\|_2 \leq \overline{\alpha}_k + \rho_F \frac{p}{n} \Big(1 + \sqrt{\frac{t \log p}{p}} \Big) \|\alpha_{k-1}\|_2 \lesssim \rho_F \overline{\alpha}_k, \tag{261}$$

and

$$\begin{aligned}
\|\widehat{\beta}_{t+1}\|_2 \leq \|v_{t+1}\|_2 + \sum_{k=1}^{t} \|\widehat{\gamma}_t^k F_k(v_k)\|_2 &\lesssim \frac{p}{n} \Big(1 + \sqrt{\frac{t \log p}{p}} \Big) \|\alpha_t\| + \sum_{k=1}^{t} |\widehat{\gamma}_t^k| \|F_k(v_k)\|_2 \\
&\lesssim \frac{p}{n} \Big(1 + \sqrt{\frac{t \log p}{p}} \Big) \|\alpha_t\| + \sum_{k=1}^{t} |\widehat{\gamma}_t^k| \rho_F \overline{\alpha}_t \lesssim \overline{\alpha}_t.
\end{aligned} \tag{262}$$

Here inequality (262) holds for $\|\widetilde{\beta}_{t+1}\|_2$ similarly. Combining the above two relations, we are guaranteed that

$$\Big\| F_{t+1}\Big(\widetilde{\beta}_{t+1} \Big) \Big\|_2 \leq \|F_{t+1}(0)\|_2 + \rho_F \|\widetilde{\beta}_{t+1}\|_2 \leq \rho_F \overline{\alpha}_k.$$

In addition, some direct algebra together with the Lipschitz property of function F_k leads to

$$\begin{aligned}
\|\widehat{\beta}_{t+1} - \widetilde{\beta}_{t+1}\|_2 &\leq \|v_{t+1} - \widetilde{v}_{t+1}\|_2 + \sum_{k=1}^{t} \|\widehat{\gamma}_t^k F_k(v_k) - \widetilde{\gamma}_t^k F_k(\widetilde{v}_k)\|_2 \\
&\lesssim \frac{p}{n} \Big(1 + \sqrt{\frac{t \log p}{p}} \Big) \|\alpha_t - \widetilde{\alpha}_t\|_2 + \sum_{k=1}^{t} |\widehat{\gamma}_t^k - \widetilde{\gamma}_t^k| \|F_k(\widetilde{v}_k)\|_2 + \sum_{k=1}^{t} |\widehat{\gamma}_t^k| \|F_k(v_k) - F_k(\widetilde{v}_k)\|_2
\end{aligned}$$

$$\lesssim \epsilon + \epsilon \sum_{k=1}^{t} \rho_F \overline{\alpha}_k + \sum_{k=1}^{t} |\widehat{\gamma}_t^k| \rho_F \|v_k - \widetilde{v}_k\|_2. \tag{263}$$

Conditioning on event (224), for each $1 \leq k \leq t$, one has

$$\|v_k - \widetilde{v}_k\|_2 \lesssim \frac{p}{n}\Big(1 + \sqrt{\frac{t \log p}{p}}\Big) \|\alpha_{k-1} - \widetilde{\alpha}_{k-1}\|_2 \lesssim \epsilon, \tag{264}$$

which implies that (263) can be further bounded as

$$\|\widehat{\beta}_{t+1} - \widetilde{\beta}_{t+1}\|_2 \leq \epsilon + \epsilon \sum_{k=1}^{t} \rho_F \overline{\alpha}_k + \epsilon \sum_{k=1}^{t} |\widehat{\gamma}_t^k| \rho_F \lesssim \frac{1}{\mathsf{poly}(n)} \rho_F \overline{\alpha}_t. \tag{265}$$

Substituting the above relations into (255) establishes relation (256).

F Proof of auxiliary lemmas

F.1 Proof of Lemma 1

To facilitate our analysis, let us first introduce some definitions and basic properties. Recall that we define two sets of orthonormal basis $\{a_k\}_{1 \leq k \leq \min\{n,p\}}$ and $\{b_k\}_{1 \leq k \leq \min\{n,p\}}$ and for each t, concatenate them into orthonormal matrices

$$U_t = [a_k]_{1 \leq k \leq t} \in \mathbb{R}^{n \times t}, \qquad V_t = [b_k]_{1 \leq k \leq t} \in \mathbb{R}^{p \times t}.$$

For every $1 \leq k \leq \min\{n,p\}$, we write the orthogonal complement of U_k as $U_k^{\perp} \in \mathbb{R}^{n \times (n-k)}$ which satisfies $U_k^{\top} U_k^{\perp} = 0$ and $U_k^{\perp\top} U_k^{\perp} = I_{n-k}$. Similarly, we write the orthogonal complement of V_k as $V_k^{\perp} \in \mathbb{R}^{p \times (p-k)}$. Additionally, we find it helpful to consider the projection where the rows of X are projected to the $p-k$-dimensional space $V_k^{\perp}$, and the columns to the $n-k$-dimensional space $U_k^{\perp}$, which is denoted by

$$\widetilde{X}_{k+1} \coloneqq U_k^{\perp\top} X V_k^{\perp} \in \mathbb{R}^{(n-k) \times (p-k)}. \tag{266}$$

With these notation in place, X_{k+1} (defined as in (62)) obeys

$$\begin{aligned} X_{k+1} &= \left(I_n - a_k a_k^{\top}\right) X_k \left(I_p - b_k b_k^{\top}\right) = \cdots = \left(I_n - U_k U_k^{\top}\right) X \left(I_p - V_k V_k^{\top}\right) \\ &= U_k^{\perp} U_k^{\perp\top} X V_k^{\perp} V_k^{\perp\top} = U_k^{\perp} \widetilde{X}_{k+1} V_k^{\perp\top}. \end{aligned}$$

Claim 3. *For every $1 \leq k \leq \min\{n,p\}$, conditional on $\{a_i, b_i\}_{1 \leq i \leq k}$ and (s_0, β_1), the following properties hold true:*

- *$\widetilde{X}_{k+1}$ is a rescaled Wigner matrix in $\mathbb{R}^{(n-k)\times(p-k)}$, with $(\widetilde{X}_{k+1})_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, \frac{1}{n})$;*
- *X_{k+1} is conditional independent of $\{X_i b_i, X_i^{\top} a_i\}_{1 \leq i \leq k}$;*
- *the randomness of (s_k, β_{k+1}) and $\{a_{k+1}, b_{k+1}\}$ comes purely from $\{X_i b_i, X_i^{\top} a_i\}_{1 \leq i \leq k}$, and hence (s_k, β_{k+1}) and $\{a_{k+1}, b_{k+1}\}$ are conditionally independent of W_{k+1}.*

In view of relations (66) and (69), the proof of Claim 3 proceeds by the same induction method as in the proof of (Li and Wei, 2022, Claim 1). We thus omit its details here.

Equipped with the results above, let us characterize the distribution of $W_k b_k$ for $2 \leq k \leq \min\{n,p\}$. To begin with, conditional on $\{a_i, b_i\}_{1 \leq i \leq k-1}$ and (s_0, β_1), we have

$$\begin{aligned} a_i^{\top} X_k b_k &= a_i^{\top} U_{k-1}^{\perp} \widetilde{X}_k V_{k-1}^{\perp\top} b_k = 0 \qquad \text{for } i \leq k-1; \\ U_{k-1}^{\perp\top} X_k b_k &= (U_{k-1}^{\perp\top} U_{k-1}^{\perp}) \widetilde{X}_k (V_{k-1}^{\perp\top} b_k) \sim \mathcal{N}\left(0, \frac{1}{n} I_{n-k+1}\right), \end{aligned}$$

where we make use of the fact that $\widetilde{X}_k$ is a rescaled Wigner matrix in $\mathbb{R}^{(n-k+1)\times(p-k+1)}$ conditionally independent of b_k. As a consequence, if one generates i.i.d. Gaussian random variables $g_k^i \sim \mathcal{N}(0, \frac{1}{n})$ for all $1 \le i < k$, then conditional on $\{a_i, b_i\}_{1\le i\le k-1}$ and (s_0, β_1), it follows that

$$\phi_k = X_k b_k + \sum_{i=1}^{k-1} g_k^i a_i \sim \mathcal{N}\left(0, \frac{1}{n} I_n\right). \tag{267}$$

Similarly, one can characterize the distribution of $a_k^\top X_k (I_p - b_k b_k^\top)$ by noticing

$$\begin{aligned}
a_k^\top X_k (I_n - b_k b_k^\top) b_k &= 0;\\
a_k^\top X_k (I_p - b_k b_k^\top) b_i &= a_k^\top X_k b_i = a_k^\top U_{k-1}^\perp \widetilde{X}_k V_{k-1}^{\perp\top} b_i = 0 \qquad \text{for } i \le k-1;\\
a_k^\top X_k (I_p - b_k b_k^\top) V_k^\perp &= a_k^\top U_{k-1}^\perp \widetilde{X}_k V_{k-1}^{\perp\top} (I_p - b_k b_k^\top) V_k^\perp = (a_k^\top U_{k-1}^\perp) \widetilde{X}_k (V_{k-1}^{\perp\top} V_k^\perp) \sim \mathcal{N}\left(0, \frac{1}{n} I_{n-k}\right).
\end{aligned}$$

Here the last relation follows since conditioning on $\{a_i, b_i\}_{1\le i\le k-1}$ and (s_0, β_1), $\widetilde{X}_k$ is a rescaled Wigner matrix in $\mathbb{R}^{(n-k+1)\times(p-k+1)}$ independent of (a_k, b_k). It thus obeys that

$$\psi_k = \left(I - b_k b_k^\top\right) X_k^\top a_k + \sum_{i=1}^{k} q_k^i b_i \sim \mathcal{N}\left(0, \frac{1}{n} I_p\right).$$

Finally, we make the observation that $\{\phi_i\}_{1\le i\le k}$ are independent, so as $\{\psi_i\}_{1\le i\le k}$. In order to see this, first note that each ϕ_k is independent of $\{a_i, b_i\}_{1\le i\le k-1}$ and (s_0, β_1) which follows immediately from the conditional distributional guarantee established in (267). Next, putting together Claim 3 with the definition of ϕ_k implies that conditional on $\{a_i, b_i\}_{1\le i\le k-1}$ and (s_0, β_1), ϕ_k — whose randomness comes purely from $X_k b_k$ and g_k^i — is statistically independent of $\phi_1, \dots, \phi_{k-1}$. Again, as the distribution of ϕ_k does not relies on $\{a_i, b_i\}_{1\le i\le k-1}$ and (s_0, β_1), therefore, we conclude $\{\phi_i\}_{1\le i\le k}$ are statistically independent. Similarly, one can also validate $\{\psi_i\}_{1\le i\le k}$ are statistically independent.

F.2 Proof of Lemma 8

To control quantity $\|\sum_{k=1}^t \omega_k \psi_k \circ F'_{t+1}(v_{t+1})\|_2^2$ with $\omega \perp \alpha_t$ (see (205)), the idea is to invoke Lemma 3 for any fixed $\omega \in \mathcal{S}^{t-1}$ — independent of v_{t+1} and apply a standard covering argument. Given any fixed α_t and $\omega \perp \alpha_t \in \mathcal{S}^{t-1}$, $\sum_{k=1}^t \omega_k \psi_k$ follows $\mathcal{N}(0, \frac{1}{n} I_n)$, which is independent with $v_{t+1} := \sum_{k=1}^t \alpha_t^k \psi_k$. This implies that

$$\mathbb{E}\left[\left\|\sum_{k=1}^t \omega_k \psi_k \circ F'_{t+1}(v_{t+1})\right\|_2^2\right] = \frac{1}{n}\mathbb{E}\Big[\big\|F'_{t+1}(v_{t+1})\big\|_2^2\Big].$$

Recognizing that $|F'_{t+1}| \le \rho_F$, it can be easily verified via properties for Gaussian distribution that

$$\mathbb{E}\left[\left(\sum_{k=1}^t \omega_k \psi_k \circ F'_{t+1}(v_{t+1})\right)_i^4\right] \lesssim \frac{1}{n^2}\rho_F^4,$$

and

$$\mathbb{P}\left(\max_{i\in[n]}\left(\sum_{k=1}^t \omega_k \psi_k \circ F'_{t+1}\left(v_{t+1}\right)\right)_i^2 \lesssim \frac{\rho_F^2}{n}\log\frac{n}{\delta}\right) \ge 1-\delta.$$

In view of Lemma 3, we obtain

$$\left\|\sum_{k=1}^t \omega_k \psi_k \circ F'_{t+1}(v_{t+1})\right\|_2^2 - \frac{1}{n}\mathbb{E}\Big[\big\|F'_{t+1}(v_{t+1})\big\|_2^2\Big] \lesssim \sqrt{\frac{\log\frac{1}{\delta}}{n}}\rho_F^2 + \frac{\rho_F^2}{n}\log^2 n \log\frac{1}{\delta}, \tag{268}$$

which holds with probability at least $1-\delta$. To take care of the statistical dependence between ω, α_t and ψ_k, let us consider an ϵ-cover of $\mathcal{S}^{t-1}$ in terms of the ℓ_2-norm, denoted by $\mathcal{N}_\epsilon$. With this definition, we can write

$$\begin{aligned}
&\sup_{\omega\perp\alpha\in\mathcal{S}^{t-1}} \left\{\Big\|\sum_{k=1}^{t}\omega_k\psi_k\circ F'_{t+1}(v_{t+1})\Big\|_2^2 - \frac{1}{n}\mathbb{E}\Big[\big\|F'_{t+1}(v_{t+1})\big\|_2^2\Big]\right\}\\
&\leq \sup_{\omega\perp\alpha\in\mathcal{N}_\epsilon} \left\{\Big\|\sum_{k=1}^{t}\omega_k\psi_k\circ F'_{t+1}(v_{t+1})\Big\|_2^2 - \frac{1}{n}\mathbb{E}\Big[\big\|F'_{t+1}(v_{t+1})\big\|_2^2\Big]\right\} + \mathsf{poly}(n)\cdot\epsilon\\
&\leq \sup_{\omega\perp\alpha\in\mathcal{N}_\epsilon} \left\{\Big\|\sum_{k=1}^{t}\omega_k\psi_k\circ F'_{t+1}(v_{t+1})\Big\|_2^2 - \frac{1}{n}\mathbb{E}\Big[\big\|F'_{t+1}(v_{t+1})\big\|_2^2\Big]\right\} + \mathsf{poly}(n)\cdot\epsilon\\
&\lesssim \sqrt{\frac{\log\left(\frac{N(\epsilon,\mathcal{S}^{t-1})}{\delta}\right)}{n}}\rho_F^2 + \frac{\rho_F^2}{n}\log^2 n\log\left(\frac{N(\epsilon,\mathcal{S}^{t-1})}{\delta}\right) + \mathsf{poly}(n)\cdot\epsilon,
\end{aligned}$$

where the last inequality holds with probability $1-\delta$ and the second inequality follows from that conditioning on the event in (172b), $\|v_{t+1}-\widetilde{v}_{t+1}\|_2 \leq \mathsf{poly}(n)\cdot\epsilon$ and $\big\|\sum_{k=1}^t (w_k-\widetilde{w}_k)\psi_k\big\|_2 \leq \mathsf{poly}(n)\cdot\epsilon$. Selecting parameters

$$\delta = \frac{1}{n^{10}} \qquad \text{and} \quad \epsilon = \frac{1}{\mathsf{poly}(n)},$$

gives

$$\sup_{\omega\in\mathcal{S}^{t-1}} \left\{\Big\|\sum_{k=1}^{t}\omega_k\psi_k\circ F'_{t+1}(v_{t+1})\Big\|_2^2 - \frac{1}{n}\mathbb{E}\Big[\big\|F'_{t+1}(v_{t+1})\big\|_2^2\Big]\right\} \lesssim \sqrt{\frac{t\log^2 n}{n}}\rho_F^2,$$

which completes the proof of the targeted bound (213a). Following similar argument, one can also derive inequality (213b).

F.3 Proof of Lemma 9

Recalling the definition in expression (227), we begin by directly decomposing the quantity of interest as

$$\begin{aligned}
&\sum_{i=1}^{n}\mathsf{Var}(X_i^0+X_i)\\
&\leq \mathbb{E}\left\|\sum_{k=1}^{t}\omega_k\psi_k\circ F_{t+1}(\widehat{\beta}_{t+1})\right\|_2^2\\
&\lesssim \mathbb{E}\left\|\sum_{k=1}^{t}\omega_k\psi_k\circ \mathbb{1}\left(\sum_{k=1}^{t}\omega_k\psi_k \lesssim \sqrt{\frac{\log n}{n}}\right)\circ F_{t+1}(\widehat{\beta}_{t+1})\right\|_2^2 + \mathbb{E}\left\|\sum_{k=1}^{t}\omega_k\psi_k\circ \mathbb{1}\left(\sum_{k=1}^{t}\omega_k\psi_k \gtrsim \sqrt{\frac{\log n}{n}}\right)\circ F_{t+1}(\widehat{\beta}_{t+1})\right\|_2^2\\
&\lesssim \frac{\log n}{n}\mathbb{E}\left\|F_{t+1}(\widehat{\beta}_{t+1})\right\|_2^2 + \mathbb{E}\left\|\sum_{k=1}^{t}\omega_k\psi_k\circ \mathbb{1}\left(\sum_{k=1}^{t}\omega_k\psi_k \gtrsim \sqrt{\frac{\log n}{n}}\right)\circ F_{t+1}(\widehat{\beta}_{t+1})\right\|_2^2.
\end{aligned} \tag{269}$$

Next, we control these two parts above separately.

- Regarding the first part, following by the Lipschitz property of F_{t+1} and relation (230), it satisfies

$$\begin{aligned}
\left\|F_{t+1}(\widehat{\beta}_{t+1})\right\|_2^2 &\lesssim \left\|F_{t+1}\Big(\sum_{k=1}^{t}\widehat{\gamma}_t^k F_k(0)\Big)\right\|_2^2 + \rho_F^2\left\|\widehat{\beta}_{t+1}-\sum_{k=1}^{t}\widehat{\gamma}_t^k F_k(0)\right\|_2^2\\
&\lesssim \overline{\gamma}_{t+1}^2 + \rho_F^2\left\|\widehat{\beta}_{t+1}-\sum_{k=1}^{t}\widehat{\gamma}_t^k F_k(0)\right\|_2^2.
\end{aligned} \tag{270}$$

Here, recall $\widehat{\beta}_{t+1} \coloneqq v_{t+1} + \sum_{k=1}^{t} \widehat{\gamma}_t^k F_k(v_k)$ and $v_{k+1} = \sum_{k=1}^{t} \alpha_t^k \psi_k$ to obtain

$$\Big\|\widehat{\beta}_{t+1} - \sum_{k=1}^{t} \widehat{\gamma}_t^k F_k(0)\Big\|_2^2 = \Big\|v_{t+1} + \sum_{k=1}^{t} \widehat{\gamma}_t^k F_k(v_k) - \sum_{k=1}^{t} \widehat{\gamma}_t^k F_k(0)\Big\|_2^2 \lesssim \|v_{t+1}\|_2^2 + \rho_F^2 \Big\|\sum_{k=1}^{t} |\widehat{\gamma}_t^k v_k|\Big\|_2^2, \tag{271}$$

where the last inequality again invokes the Lipschitz property of F_k. Taking the expectation on both sides, we arrive at

$$\mathbb{E}\Big\|\widehat{\beta}_{t+1} - \sum_{k=1}^{t} \widehat{\gamma}_t^k F_k(0)\Big\|_2^2 \lesssim \|\alpha_t\|_2^2 + \rho_F^2 \sum_{i,j=1}^{t} |\widehat{\gamma}_t^i \widehat{\gamma}_t^j| \cdot \mathbb{E}[\|v_i\|_2 \, \|v_j\|_2]. \tag{272}$$

Here, again, we remind the readers that $\widehat{\gamma}_t$ is regarded as a fixed parameter. Now in order to bound the right hand side of (272), since $v_{i+1} \sim \mathcal{N}(0, \frac{\|\alpha_i\|_2^2}{n} I_p)$ for every fixed α_i, it obeys that

$$\begin{aligned} \mathbb{E}\left[\frac{\|v_{i+1}\|_2}{\|\alpha_i\|_2} \frac{\|v_{j+1}\|_2}{\|\alpha_j\|_2}\right] &\le \mathbb{E}\Big[\max\{\|X\|_2^2, \|Y\|_2^2\}\Big] \qquad \text{where } X, Y \sim \mathcal{N}\Big(0, \frac{1}{n} I_p\Big) \\ &\le \mathbb{E}\|X\|_2^2 + \mathbb{E}\|Y\|_2^2 \lesssim \frac{p}{n}. \end{aligned}$$

Therefore, the right hand side of (272) further satisfies

$$\begin{aligned} \mathbb{E}\Big\|\widehat{\beta}_{t+1} - \sum_{k=1}^{t} \widehat{\gamma}_t^k F_k(0)\Big\|_2^2 &\lesssim \|\alpha_t\|_2^2 + \frac{p\rho_F^2}{n} \sum_{i,j=0}^{t-1} |\widehat{\gamma}_t^i \widehat{\gamma}_t^j| \|\alpha_i\|_2 \|\alpha_j\|_2 \\ &\lesssim \left(\|\alpha_t\|_2 + \sqrt{\frac{p}{n}} \rho_F \sum_{k=1}^{t} \widehat{\gamma}_t^k \|\alpha_{k-1}\|_2\right)^2 \lesssim \overline{\alpha}_t^2. \end{aligned}$$

Combining with (270) ensures

$$\mathbb{E}\left\|F_{t+1}(\widehat{\beta}_{t+1})\right\|_2^2 \lesssim \overline{\gamma}_{t+1}^2 + \rho_F^2 \overline{\alpha}_t^2. \tag{273}$$

Thus, we complete the control of the first term in (269).

- It then suffices to control the second term, which shall again be done by means of concentration of measure. We claim that

$$\mathbb{E}\left\|\sum_{k=1}^{t} \omega_k \psi_k \circ \mathbb{1}\left(\sum_{k=1}^{t} \omega_k \psi_k \gtrsim \sqrt{\frac{\log n}{n}}\right) \circ F_{t+1}(\widehat{\beta}_{t+1})\right\|_2^2 \lesssim \frac{1}{\mathsf{poly}(n)} (\overline{\gamma}_{t+1}^2 + \overline{\alpha}_t^2). \tag{274}$$

In order to see this, first, by putting together inequalities (270), (271) and (230), we have

$$\begin{aligned} \left\|F_{t+1}(\widehat{\beta}_{t+1})\right\|_2 &\lesssim \overline{\gamma}_{t+1} + \|v_{t+1}\|_2 + \rho_F \Big\|\sum_{k=1}^{t} |\widehat{\gamma}_t^k v_k|\Big\|_2 \\ &\le \overline{\gamma}_{t+1} + \|v_{t+1}\|_2 + \rho_F \sum_{k=1}^{t} |\widehat{\gamma}_t^k| \|v_k\|_2 \le \overline{\gamma}_{t+1} + 2\|\Psi\|_{\mathrm{op}} \overline{\alpha}_t, \end{aligned}$$

where the last inequality uses the fact that for each k, $v_{k+1} = \sum_{k=1}^{t} \alpha_t^k \psi_k$ and $\rho_F \sum_{k=1}^{t} |\widehat{\gamma}_t^k| \ll 1$. In view of this relation, we can decompose the quantity of interest as

$$\mathbb{E}\left[\Big\|\sum_{k=1}^{t} \omega_k \psi_k \circ \mathbb{1}\Big(\sum_{k=1}^{t} \omega_k \psi_k \gtrsim \sqrt{\frac{\log n}{n}}\Big) \circ F_{t+1}(\widehat{\beta}_{t+1})\Big\|_2^2\right]$$

$$
\begin{aligned}
&\lesssim \overline{\gamma}_{t+1}^2 \mathbb{E}\left[\Big\|\sum_{k=1}^t \omega_k\psi_k \circ \mathbb{1}\Big(\sum_{k=1}^t \omega_k\psi_k \gtrsim \sqrt{\frac{\log n}{n}}\Big)\Big\|_2^2\right] + \overline{\alpha}_t^2 \mathbb{E}\left[\Big\|\sum_{k=1}^t \omega_k\psi_k \circ \mathbb{1}\Big(\sum_{k=1}^t \omega_k\psi_k \gtrsim \sqrt{\frac{\log n}{n}}\Big)\Big\|_2^2 \cdot \|\Psi\|_{\rm op}^2\right] \\
&\lesssim \frac{\overline{\gamma}_{t+1}^2}{\mathsf{poly}(n)} + \overline{\alpha}_t^2 \mathbb{E}\left[\Big\|\sum_{k=1}^t \omega_k\psi_k \circ \mathbb{1}\Big(\sum_{k=1}^t \omega_k\psi_k \gtrsim \sqrt{\frac{\log n}{n}}\Big)\Big\|_2^2 \cdot \|\Psi\|_{\rm op}^2\right]. \qquad (275)
\end{aligned}
$$

Here, the last inequality uses the property that

$$
\mathbb{E}\left[X_i^2\, \mathbb{1}\Big(X_i \gtrsim \sqrt{\frac{\log n}{n}}\Big)\right] \le \frac{1}{\mathsf{poly}(n)}, \qquad \text{for } X_i \sim \mathcal{N}\Big(0, \frac{1}{n}\Big), \qquad (276)
$$

and given a fixed vector $\omega \in \mathcal{S}^t$, $\sum_{k=1}^t \omega_k\psi_k \sim \mathcal{N}(0, \frac{1}{n}I_p)$. We now turn to the upper bound of the second term on the right of expression (275).

$$
\begin{aligned}
&\mathbb{E}\left[\Big\|\sum_{k=1}^t \omega_k\psi_k \circ \mathbb{1}\Big(\sum_{k=1}^t \omega_k\psi_k \gtrsim \sqrt{\frac{\log n}{n}}\Big)\Big\|_2^2 \cdot \|\Psi\|_{\rm op}^2\right] \\
&= \mathbb{E}\left[\Big\|\sum_{k=1}^t \omega_k\psi_k \circ \mathbb{1}\Big(\sum_{k=1}^t \omega_k\psi_k \gtrsim \sqrt{\frac{\log n}{n}}\Big)\Big\|_2^2 \cdot \|\Psi\|_{\rm op}^2\, \mathbb{1}(\|\Psi\|_{\rm op} - 1 \lesssim \sqrt{\frac{\log n}{n}})\right] \\
&\quad + \mathbb{E}\left[\Big\|\sum_{k=1}^t \omega_k\psi_k \circ \mathbb{1}\Big(\sum_{k=1}^t \omega_k\psi_k \gtrsim \sqrt{\frac{t\log n}{n}}\Big)\Big\|_2^2 \cdot \|\Psi\|_{\rm op}^2\, \mathbb{1}(\|\Psi\|_{\rm op} - 1 \gtrsim \sqrt{\frac{t\log n}{n}})\right] \\
&\overset{\text{(i)}}{\lesssim} \frac{1}{\mathsf{poly}(n)} + \mathbb{E}\left[\|\Psi\|_{\rm op}^4\, \mathbb{1}(\|\Psi\|_{\rm op} - 1 \gtrsim \sqrt{\frac{t\log n}{n}})\right] \\
&\overset{\text{(ii)}}{\lesssim} \frac{1}{\mathsf{poly}(n)},
\end{aligned}
$$

where (i) results from relation (276) and (ii) follows from the concentration result for $\|\Psi\|_{\rm op}$ (cf. (171)).

Finally, combining relations (273) and (274) leads to our target bound.

F.4 Covering lemmas

As defined around display (244), for every $\widehat{\theta} = \Big(\omega, \{\alpha_k\}_{k\le t}, \widehat{\gamma}_t, \{\tau_k\}_{k\le t+1}\Big) \in \Theta$, $\widetilde{\theta} := \Big(\widetilde{\omega}, \{\widetilde{\alpha}_k\}_{k\le t}, \widetilde{\gamma}_t, \{\widetilde{\tau}_k\}_{k\le t+1}\Big)$ is a point that lies in the ϵ-cover $\mathcal{M}_\epsilon$ of Θ_0 which satisfies

$$
\|\omega - \widetilde{\omega}\|_2 \le \epsilon, \quad \|\alpha_k - \widetilde{\alpha}_k\|_2 \le \epsilon, \quad \|\tau_k - \widetilde{\tau}_k\|_2 \le \epsilon, \quad |\widehat{\gamma}_t^k - \widetilde{\gamma}_t^k| \le \epsilon,
$$

for every k and $\epsilon = 1/\mathsf{poly}(n)$. We also record that

$$
\widetilde{\gamma}_t^k = \begin{cases} 0, & \text{for } k \le t - O(\log n), \\ \widehat{\gamma}_t^k & \text{o.w.} \end{cases}
$$

Lemma 10. *Under the assumptions* (191b) – (188)*, the following set of relations holds with probability at least* $1 - O(n^{-10})$

-

$$
\Big|\Big\langle F'_{t+1}(\widehat{\beta}_{t+1}) \circ F'_j(v_j)\Big\rangle - \Big\langle F'_{t+1}(\widetilde{\beta}_{t+1}) \circ F'_j(\widetilde{v}_j)\Big\rangle\Big| \lesssim \frac{t\log^3 n}{n}\rho_F^2 + \frac{1}{\mathsf{poly}(n)}\rho_F\rho_{1,F}\overline{\alpha}_t, \qquad (277a)
$$

$$
\Big|\langle F'_{t+1}(\widehat{\beta}_{t+1})\rangle - \langle F'_{t+1}(\widetilde{\beta}_{t+1})\rangle\Big| \lesssim \frac{t\log^3 n}{n}\rho_F + \frac{1}{\mathsf{poly}(n)}\rho_{1,F}\overline{\alpha}_t; \qquad (277b)
$$

-
$$\left\|\sum_{k=1}^{t}\omega_k\psi_k\circ\left[F'_{t+1}(v_{t+1}+\varepsilon)-F'_{t+1}(v_{t+1})\right]\right\|_2 \lesssim \rho_{1,F}\sqrt{\frac{t\log n}{n}}\|\varepsilon\|_2+\rho_F\left(\sqrt{\frac{t\log^3 n}{n}}+\sqrt{\log n}\Big(\frac{\|\varepsilon\|_2}{\|\alpha_t\|_2}\Big)^{\frac{1}{3}}\right), \tag{278a}$$
$$\left\|F'_{t+1}(\widehat{\beta}_{t+1})-F'_{t+1}(v_{t+1})\right\|_1 \lesssim \sqrt{n}\rho_{1,F}\left\|\widehat{\beta}_{t+1}-v_{t+1}\right\|_2+\rho_F\left(t\log n+n\Big(\frac{\|\widehat{\beta}_{t+1}-v_{t+1}\|_2}{\|\alpha_t\|_2}\Big)^{\frac{2}{3}}\right), \tag{278b}$$
$$\left|\langle F'_{t+1}(\widehat{\beta}_{t+1})-F'_{t+1}(\beta_{t+1})\rangle\right| \lesssim \frac{1}{\sqrt{n}}\rho_{1,F}\left\|\widehat{\beta}_{t+1}-\beta_{t+1}\right\|_2+\frac{1}{n}\rho_F\left(t\log^2 n+n\Big(\frac{\|\widehat{\beta}_{t+1}-\beta_{t+1}\|_2}{\|\alpha_t\|_2}\Big)^{\frac{2}{3}}\right); \tag{278c}$$

-
$$\left|\left\|\sum_{k=1}^{t}\omega_k\psi_k\circ F'_{t+1}(v_{t+1})\right\|_2^2-\left\|\sum_{k=1}^{t}\widetilde{\omega}_k\psi_k\circ F'_{t+1}(\widetilde{v}_{t+1})\right\|_2^2\right| \lesssim \frac{t\log^3 n}{n}\rho_F^2+\frac{1}{\mathsf{poly}(n)}\rho_{1,F}\rho_F, \tag{279a}$$
$$\left|\left\|F'_{t+1}(v_{t+1})\right\|_2^2-\left\|F'_{t+1}(\widetilde{v}_{t+1})\right\|_2^2\right| \lesssim \rho_F^2 t\log^2 n+\frac{1}{\mathsf{poly}(n)}\rho_{1,F}\rho_F. \tag{279b}$$

The proof of this lemma is provided in Section F.5.

F.5 Proof of Lemma 10

Before diving into details, let us first describe a general framework for bounding the fluctuation of a function when its input is perturbed slightly. Validating each inequality of Lemma 10 then boils down to computing specific parameters in the general framework. Throughout this proof, we condition on the event where both (172a) and (172b) satisfy with δ selected as $\max(n,p)^{-11}$.

F.5.1 A general framework

Let us first set up the stage. The multivariate mapping and its perturbation that we are interested in are of the form

$$H(x,\theta)=\Big[c_i(x_i,\theta)h_i(u_i(x_i,\theta))\Big]_{i=1}^{n} \qquad \text{and} \qquad H_\varepsilon(x,\theta)=\Big[c_i(x_i,\theta)h_i(u_i(x_i,\theta)+\varepsilon_i)\Big]_{i=1}^{n},$$

for perturbation vector $\varepsilon\in\mathbb{R}^n$ and parameter $\theta\in\mathbb{R}^d$. Here c_i and u_i denote $\mathsf{poly}(n)$-Lipschitz continuous functions of θ and h_i stands for functions with finite jump points. Specifically, consider functions h_i that can be decomposed into a continuous component and a discontinuous component

$$h_i(u)=h_i^{\mathsf{cont}}(u)+h_i^{\mathsf{dis}}(u). \tag{280}$$

We assume the continuous part of h_i is L-Lipschitz and the discontinuous component takes the form

$$h_i^{\mathsf{dis}}(u):=\sum_{k=1}^{M_i}s_i^k\,\mathbb{1}(u>\tau_i^k).$$

Here for each $i\in[n]$, we denote the discontinuous points of h_i as $\{\tau_i^k\}_{k=1}^{M_i}$, and the size of their jumps as $\{s_i^k\}_{k=1}^{M_i}$.

Given every x and ε, in order to compute the difference between $H(x,\theta)$ and $H_\varepsilon(x,\theta)$, it is critical to track where $h_i^{\mathsf{dis}}(u)$ and $h_i^{\mathsf{dis}}(u+\varepsilon)$ differ. For this purpose, let us define the index set

$$\mathcal{I}:=\Big\{i:\mathbb{1}(u_i(x_i,\theta)>\tau_i^k)\neq\mathbb{1}(u_i(x_i,\theta)+\varepsilon_i>\tau_i^k)\text{ for some }k\Big\}.$$

In words, $h_i^{\mathsf{dis}}(u_i) = h_i^{\mathsf{dis}}(u_i + \varepsilon_i)$ for all $i \in [n]$, on set $\mathcal{I}$. In terms of this notation, the Lipschitz property of h_i^{cont} ensures that

$$\|H(x,\theta) - H_\varepsilon(X,\theta)\|_1 \lesssim \sum_{i\in\mathcal{I}} B|c_i(x_i,\theta)| + \sum_{i\notin\mathcal{I}} L|c_i(x_i,\theta)\varepsilon_i|, \tag{281a}$$

$$\|H(x,\theta) - H_\varepsilon(X,\theta)\|_2^2 \lesssim \sum_{i\in\mathcal{I}} B^2|c_i(x_i,\theta)|^2 + \sum_{i\notin\mathcal{I}} L^2|c_i(x_i,\theta)\varepsilon_i|^2, \tag{281b}$$

provided that $|h_i(x_i,\theta)| \lesssim B$ for every i.

Additionally, consider mappings

$$H^j(x,\theta) = \Big[h_i^j(u_i^j(x_i,\theta))\Big]_{i=1}^n \qquad \text{and} \qquad H_\varepsilon^j(x,\theta) = \Big[h_i^j(u_i^j(x_i,\theta) + \varepsilon_i^j)\Big]_{i=1}^n,$$

for $j = 1$ or 2. Under the assumption $|h_i^j| \lesssim B$, in view of the Lipschitz property for the continuous part of h_i^j, we can conclude similarly that

$$\|H^1(x,\theta) \circ H^2(x,\theta) - H_{\varepsilon^1}^1(x,\theta) \circ H_{\varepsilon^2}^2(x,\theta)\|_1 \lesssim \sum_{i\in\widetilde{\mathcal{I}}} B^2 + \sum_{i\notin\widetilde{\mathcal{I}}} LB(|\varepsilon_i^1| + |\varepsilon_i^2|), \tag{281c}$$

for the index set

$$\widetilde{\mathcal{I}} := \Big\{i : \mathbb{1}(u_i^j(x_i,\theta) > \tau_i^{j,k}) \neq \mathbb{1}(u_i^j(x_i,\theta) + \varepsilon_i^j > \tau_i^{j,k}) \text{ for some } j,k\Big\}.$$

We shall employ these three relations above to establish Lemma 10, which boils down to compute the right hand side of each inequality in (281). Towards this goal, the idea is to apply the concentration results developed in Section C. Below, we state two key observations and then turn to the calculations of each inequality individually.

Consider a random vector $X \in \mathbb{R}^n$. For any fixed $\theta \in \Theta$, suppose there exists some $\sigma > 0$ such that

$$\mathbb{P}\left(|u_i^j(X_i;\theta)| < \frac{s\sigma}{n}\right) < \frac{s}{n}, \tag{282}$$

for every $s \in [n]$ and $j = 1, 2$ if there are two sets of u_i^j concerned. In view of Lemma 4 and (282), we have

$$|\mathcal{I}| \lesssim \log N\Big(\frac{\sigma}{100n^2}, \Theta\Big) \log n + \left(\frac{n\|\varepsilon\|_2}{\sigma}\right)^{\frac{2}{3}}, \tag{283}$$

where $N(\frac{\sigma}{100n^2}, \Theta)$ denotes the covering number of Θ.

In addition, suppose that for every fixed θ and $i \in [n]$, $c_i(x_i;\theta)$ is sub-exponential with

$$\mathbb{P}\Big(|c_i(x_i;\theta)| \leq M \log\frac{1}{\delta}\Big) \geq 1 - \delta, \tag{284}$$

for some $M > 0$. It is easily seen that $\mathbb{E}\big[|c_i(x_i,\theta)|\big] \lesssim M$ and $\mathsf{Var}\big(|c_i(x_i,\theta)|\big) \lesssim M^2$. Conditioning on the cardinality of $\mathcal{I}$, let us consider the quantity $\sum_{i=1}^n w_i c_i(x_i,\theta)$ where $w \in \{0, \pm 1\}^n$ and $\|w\|_1 = |\mathcal{I}|$. For every fixed w and $\theta \in \Theta$, by virtue of Lemma 3, it holds true that

$$\sum_{i=1}^n w_i c_i(x_i,\theta) \leq M|\mathcal{I}| + \sum_{i=1}^n w_i\big(c_i(x_i,\theta) - \mathbb{E}\big[w_i(x_i,\theta)\big]\big) \lesssim M|\mathcal{I}| + M\sqrt{|\mathcal{I}|\log\frac{1}{\delta}} + M\log n \log\frac{1}{\delta} \tag{285}$$

with probability at least $1 - \delta$. Here we make use of the following relations

$$\mathbb{E}\big[w_i(x_i,\theta)\big] \lesssim M \qquad \text{and} \quad \sum_{i=1}^n \mathsf{Var}\big(s_i c_i(x_i,\theta)\big) \lesssim M^2|\mathcal{I}|.$$

Note that (285) holds true for every fixed $w \in \{0, \pm 1\}^n$ and $\theta \in \Theta$. In order to accommodate the possible statistical dependences, we consider an ϵ-cover of Θ. Selecting parameters

$$\delta = \frac{1}{n^{11} N(\epsilon, \Theta) 2^{|\mathcal{I}|} \binom{n}{|\mathcal{I}|}}, \qquad \epsilon = \frac{1}{\mathsf{poly}(n)}$$

and taking union bound of (285) over possible choices of w and $\theta \in \Theta$ give

$$\begin{aligned} \sup_{\widehat{\theta}\in\Theta} \sum_{i\in\mathcal{I}} |c_i(x_i, \widehat{\theta})| &\le \sup_{\substack{\widehat{\theta}\in\Theta, w\in\{0,\pm1\}^m \\ \|w\|_1=|\mathcal{I}|}} \sum_{i=1}^n s_i c_i(x_i, \theta) \\ &\overset{\text{(i)}}{\le} \sup_{\substack{\theta\in\mathcal{N}_\epsilon, w\in\{0,\pm1\}^m \\ \|w\|_1=|\mathcal{I}|}} \sum_{i=1}^n w_i c_i(x_i, \theta) + \frac{1}{\mathsf{poly}(n)} \\ &\lesssim M|\mathcal{I}| \log n + M \log N\Big(\frac{1}{\mathsf{poly}(n)}, \Theta\Big) \log^2 n, \end{aligned} \tag{286}$$

where (i) follows from the choice of ϵ and the Lipschitz property of each c_i.

F.5.2 Validating inequalities of Lemma 10

To validate Lemma 10, we follow the general recipe provided above for specific choices of functions h_i, u_i and c_i. In particular, we shall select h_i as either $F'_{t+1,i}$ or $(F'_{t+1,i})^2$, u_i as either $\widehat{\beta}_{t+1,i}$ or $v_{t+1,i}$, and $c_i(x_i, \theta)$ as 1, $\sum_{k=1}^t \omega_k \psi_{k,i}$, or $(\sum_{k=1}^t \omega_k \psi_{k,i})^2$.

As discussed above, we make note of the following observations:

- Inequality set (281) requires a uniform bound B for function h_i, in which case, we can take $B = \rho_F$ when $h_i = F'_{t+1,i}$ and $B = \rho_F^2$ when $(F'_{t+1,i})^2$.
- For assumption (284), M can be set as 1, $\frac{1}{\sqrt{n}}$, and $\frac{1}{n}$, respectively;
- Regarding assumption (282), it suffices to select σ parameter as $\frac{\overline{\alpha}_t}{\sqrt{n}}$ for both $\widehat{\beta}_{t+1}$ and v_{t+1}. We leave the proof of this fact to the end of this section.

Proof of inequality (277a). The idea is to apply inequality (281c) for proper choices of H^1 and H^2. Specifically, set

$$H^1(\Psi, \theta) \coloneqq F'_{t+1}(\widetilde{\beta}_{t+1}) \quad \text{and } H^2(\Psi, \theta) \coloneqq F'_j\left(\widetilde{v}_j\right),$$

and

$$H^1_\varepsilon(\Psi, \theta) \coloneqq F'_{t+1}(\widehat{\beta}_{t+1}) \quad \text{and } H^2_\varepsilon(\Psi, \theta) \coloneqq F'_j\left(v_j\right).$$

With these choices in mind, $\varepsilon^1 = \widehat{\beta}_{t+1} - \widetilde{\beta}_{t+1}$, $\varepsilon^2 = v_j - \widetilde{v}_j$, and they satisfy

$$\|\varepsilon^1\|_2 \lesssim \frac{1}{\mathsf{poly}(n)} \rho_F \overline{\alpha}_t, \qquad \|\varepsilon^2\|_2 \lesssim \frac{1}{\mathsf{poly}(n)}, \tag{287}$$

in view of relations (264) and (265). Then according to (281c), we have

$$\begin{aligned} \Big|\Big\langle F'_{t+1}(\widehat{\beta}_{t+1}) \circ F'_j\left(v_j\right)\Big\rangle - \Big\langle F'_{t+1}(\widetilde{\beta}_{t+1}) \circ F'_j\left(\widetilde{v}_j\right)\Big\rangle\Big| &\le \frac{1}{n} \Big\|F'_{t+1}(\widehat{\beta}_{t+1}) \circ F'_j\left(v_j\right) - F'_{t+1}(\widetilde{\beta}_{t+1}) \circ F'_j\left(\widetilde{v}_j\right)\Big\|_1 \\ &\lesssim \frac{1}{n} \rho_F^2 |\widetilde{\mathcal{I}}| + \frac{1}{\mathsf{poly}(n)} \rho_F \rho_{1,F} \overline{\alpha}_t. \end{aligned}$$

To establish inequality (277a), it suffices to bound the cardinality of $\widetilde{\mathcal{I}}$ which shall be done via inequality (283). Specifically, inequality (283) requires bounding the covering number of the space $\{(\widetilde{\beta}_{t+1}, \widetilde{v}_j)\}$.

Recall the definitions

$$\widetilde{v}_j = \sum_{k=1}^{j-1} \widetilde{a}^i_{j-1}\psi_k \quad \text{and} \quad \widetilde{\beta}_{t+1} \coloneqq \sum_{k=1}^{t} \widetilde{\alpha}^k_t \psi_k + \sum_{k=1}^{t} \widetilde{\gamma}^k_t F_k(\widetilde{v}_k).$$

For every $\widetilde{\theta} = (\widetilde{\omega}, \{\widetilde{\alpha}_k\}_{k=1}^t, \widetilde{\gamma}_t, \{\widetilde{\tau}_j\}_{j=1}^{t+1})$, let

$$\widetilde{\theta}' = (\widetilde{\omega}', \{\widetilde{\alpha}'_k\}_{k=1}^t, \widetilde{\gamma}'_t, \{\widetilde{\tau}'_j\}_{j=1}^{t+1}),$$

where $\widetilde{\alpha}'_k = \widetilde{\alpha}_k$ for $k > t - O(\log n)$, $\widetilde{\alpha}'_k = 0$ for $k \le t - O(\log n)$. It is proved in (248) that $\widetilde{\beta}_{t+1} = \widetilde{\beta}'_{t+1}$. Therefore to construct a ϵ-cover for space $\{(\widetilde{\beta}_{t+1}, \widetilde{v}_j)\}$, it is sufficient to consider a ϵ cover for $\widetilde{a}_{j-1}$ together with $\widetilde{\theta}'$. The total dimension is of order $t\log n$. As a result, inequality (283) gives

$$|\widetilde{\mathcal{I}}| \lesssim t\log^3 n + \left(\frac{n\|\varepsilon\|_2}{\sigma}\right)^{\frac{2}{3}} \lesssim t\log^3 n. \tag{288}$$

Here $\varepsilon = (\varepsilon_1, \varepsilon_2)$ satisfies inequality (287), and $\sigma = \sqrt{\frac{2}{n\pi}}\overline{\alpha}_t$. The last relation follows from assumption (191b) that $\overline{\alpha}_t \le \mathsf{poly}(n)$.

Putting things together completes the proof of inequality (277a). Inequality (277b) is a direct consequence of inequality (277a) by directly setting $H^2(\Psi, \theta) = H^2_\varepsilon(\Psi, \theta) = 1$.

Proof of inequality (278a)**.** In order to prove inequality (278a), let us take $c_i = (\sum_{k=1}^t \omega_k\psi_k)_i$,

$$\begin{aligned} H(\Psi, \theta) &\coloneqq \sum_{k=1}^{t} \omega_k\psi_k \circ F'_{t+1}(v_{t+1}) \\ \text{and } H_\varepsilon(\Psi, \theta) &\coloneqq \sum_{k=1}^{t} \omega_k\psi_k \circ F'_{t+1}(v_{t+1} + \varepsilon). \end{aligned}$$

Then according to (281b), we have

$$\left\|\sum_{k=1}^{t} \omega_k\psi_k \circ \left[F'_{t+1}(v_{t+1} + \varepsilon) - F'_{t+1}(v_{t+1})\right]\right\|_2^2 \lesssim \sum_{i\in\mathcal{I}} \rho_F^2 \Big|\sum_{k=1}^{t} \omega_k\psi_{k,i}\Big|^2 + \sum_{i\notin\mathcal{I}} \rho_{1,F}^2 \Big|\sum_{k=1}^{t} \omega_k\psi_{k,i}\varepsilon_i\Big|^2. \tag{289}$$

We control each term of the right hand side of (289) respectively. To begin with, the parameter that we shall build a ϵ-cover with is v_{t+1} which is determined by $\widehat{\alpha}_t \in \mathbb{R}^t$. In view of inequality (283), we have

$$|\mathcal{I}| \lesssim \log N\Big(\frac{\sigma}{100n^2}, \Theta\Big)\log n + \left(\frac{n\|\varepsilon\|_2}{\sigma}\right)^{\frac{2}{3}} \lesssim t\log^2 n + \left(\frac{n\|\varepsilon\|_2}{\|\widehat{\alpha}_t\|/\sqrt{n}}\right)^{\frac{2}{3}}, \tag{290}$$

where we recall the σ parameter for v_{t+1} equals to $\sqrt{\frac{2}{n\pi}}\overline{\alpha}_t$. In view of the relation (173c) in Lemma 5, one has

$$\Big\|\sum_{k=1}^{t} \omega_k\psi_k\Big\|_\infty \lesssim \frac{t\log n}{n}, \quad \text{and} \quad \sum_{i\in\mathcal{I}}\Big|\sum_{k=1}^{t} \omega_k\psi_{k,i}\Big|^2 \lesssim \frac{(t + |\mathcal{I}|)\log n}{n}, \tag{291}$$

with probability at least $1 - O(n^{-10})$. Taking everything collectively, we arrive at

$$\left\|\sum_{k=1}^{t} \omega_k\psi_k \circ \left[F'_{t+1}(v_{t+1} + \varepsilon) - F'_{t+1}(v_{t+1})\right]\right\|_2^2 \lesssim \rho_F^2\left(\frac{t\log^3 n}{n} + \log n\left(\frac{\|\varepsilon\|_2}{\|\widehat{\alpha}_t\|_2}\right)^{\frac{2}{3}}\right) + \rho_{1,F}^2\frac{t\log n}{n}\|\varepsilon\|_2^2,$$

from which the advertised claim in (278a) follows. The proofs of (278b) and (278c) can be established in the same manner, by invoking relation (281a) with $c_i = 1$ and $h_i = F'_{t+1,i}$. Here $\widehat{\beta}_{t+1} - v_{t+1}$ and $\widehat{\beta}_{t+1} - \beta_{t+1}$ play the role of ε in these cases.

Proof of inequality (279a). To establish inequality (279a), consider $c_i = (\sum_{k=1}^t \omega_k \psi_k)_i^2$ and

$$H(\Psi, \theta) := \left(\sum_{k=1}^t \omega_k \psi_k \circ F'_{t+1}(v_{t+1})\right)^2 \in \mathbb{R}^p,$$
$$H_\varepsilon(\Psi, \theta) := \left(\sum_{k=1}^t \omega_k \psi_k \circ F'_{t+1}(\widetilde{v}_{t+1})\right)^2 \in \mathbb{R}^p.$$

By virtue of (264), $\|\varepsilon\|_2 = \|\widetilde{v}_{t+1} - v_{t+1}\|_2 \lesssim \epsilon = \frac{1}{\mathsf{poly}(n)}$. Some basic algebra leads to

$$\left|\left\|\sum_{k=1}^t \omega_k \psi_k \circ F'_{t+1}(v_{t+1})\right\|_2^2 - \left\|\sum_{k=1}^t \omega_k \psi_k \circ F'_{t+1}(\widetilde{v}_{t+1})\right\|_2^2\right| \leq \left\|\left(\sum_{k=1}^t \omega_k \psi_k \circ F'_{t+1}(v_{t+1})\right)^2 - \left(\sum_{k=1}^t \omega_k \psi_k \circ F'_{t+1}(\widetilde{v}_{t+1})\right)^2\right\|_1$$
$$\lesssim \sum_{i \in \mathcal{I}} \rho_F^2 \left|\sum_{k=1}^t \omega_k \psi_{k,i}\right|^2 + \sum_{i \notin \mathcal{I}} \rho_{1,F}\rho_F \left|\sum_{k=1}^t \omega_k \psi_{k,i}\right|^2 |\varepsilon_i|$$

where the last line follows from relation (281a). Similar to the discussions around display (290), inequality (283) gives

$$|\mathcal{I}| \lesssim \log N\Big(\frac{\sigma}{100n^2}, \Theta\Big) \log n + \left(\frac{n\|\varepsilon\|_2}{\sigma}\right)^{\frac{2}{3}} \lesssim t \log^2 n + \left(\frac{n\|\varepsilon\|_2}{\|\widehat{\alpha}_t\|/\sqrt{n}}\right)^{\frac{2}{3}} \lesssim t \log^2 n,$$

where the last inequality invokes $\|\varepsilon\| \lesssim \epsilon = \frac{1}{\mathsf{poly}(n)}$. Taking this together with concentration bounds in (291) further leads to

$$\left|\left\|\sum_{k=1}^t \omega_k \psi_k \circ F'_{t+1}(v_{t+1})\right\|_2^2 - \left\|\sum_{k=1}^t \omega_k \psi_k \circ F'_{t+1}(\widetilde{v}_{t+1})\right\|_2^2\right| \lesssim \frac{t \log^3 n}{n}\rho_F^2 + \frac{t^2 \log^2 n}{n^2}\rho_{1,F}\rho_F\|\varepsilon\|_1$$
$$\lesssim \frac{t \log^3 n}{n}\rho_F^2 + \frac{1}{\mathsf{poly}(n)}\rho_{1,F}\rho_F. \tag{292}$$

Contrasting the above to our target bound (279a), we are only left to consider replacing ω_k to $\widetilde{\omega}_k$ in the second term of the left hand side. Specifically, note that

$$\left|\left\|\sum_{k=1}^t \omega_k \psi_k \circ F'_{t+1}(\widetilde{v}_{t+1})\right\|_2^2 - \left\|\sum_{k=1}^t \widetilde{\omega}_k \psi_k \circ F'_{t+1}(\widetilde{v}_{t+1})\right\|_2^2\right|$$
$$\left|\Big(\sum_{k=1}^t \omega_k \psi_k \circ F'_{t+1}(\widetilde{v}_{t+1}) - \sum_{k=1}^t \widetilde{\omega}_k \psi_k \circ F'_{t+1}(\widetilde{v}_{t+1})\Big)^\top \Big(\sum_{k=1}^t \omega_k \psi_k \circ F'_{t+1}(\widetilde{v}_{t+1}) + \sum_{k=1}^t \widetilde{\omega}_k \psi_k \circ F'_{t+1}(\widetilde{v}_{t+1})\Big)\right|$$
$$\leq \sqrt{p}\rho_F \left\|\sum_{k=1}^t \omega_k \psi_k\right\|_\infty \left\|\sum_{k=1}^t (\omega_k - \widetilde{\omega}_k)\psi_k \circ F'_{t+1}(\widetilde{v}_{t+1})\right\|_2$$
$$\lesssim \sqrt{p}\rho_F \frac{t \log n}{n} \cdot \rho_F \frac{p}{n}\Big(1 + \sqrt{\frac{t \log n}{p}}\Big)\|\omega - \widetilde{\omega}\|_2$$
$$\lesssim \frac{1}{\mathsf{poly}(n)}\rho_F^2, \tag{293}$$

where for the penultimate line, recall that we condition on the event (172b) with probability at least $1 - O(n^{-11})$; the last line invokes the assumption that $\|\omega - \widetilde{\omega}\|_2 \leq \epsilon$. Combined with inequality (292), the above relation implies that

$$\left|\left\|\sum_{k=1}^t \omega_k \psi_k \circ F'_{t+1}(v_{t+1})\right\|_2^2 - \left\|\sum_{k=1}^t \widetilde{\omega}_k \psi_k \circ F'_{t+1}(\widetilde{v}_{t+1})\right\|_2^2\right|$$

$$\leq \left| \Big\| \sum_{k=1}^{t} \omega_k \psi_k \circ F'_{t+1}(v_{t+1}) \Big\|_2^2 - \Big\| \sum_{k=1}^{t} \omega_k \psi_k \circ F'_{t+1}(\widetilde{v}_{t+1}) \Big\|_2^2 \right| + \frac{1}{\mathsf{poly}(n)} \rho_F^2$$
$$\lesssim \frac{1}{\mathsf{poly}(n)} \rho_{1,F} \rho_F + \frac{t \log^2 n}{n} \rho_F^2 .$$

We thus complete the proof of inequality (279a). Similarly, by taking $c_i = 1$ for every $i \in [n]$, inequality (279b) follows by the same argument above immediately.

The remaining terms can be proved in a similar way, which is omitted here for simplicity.

F.5.3 Other auxiliary details

σ-parameter for $\widehat{\beta}_{t+1}$ and v_{j+1}. Recall that $v_{j+1} = \sum_{k=1}^{j} \widehat{\alpha}_j^i \psi_k$ and $\widehat{\beta}_{t+1} \coloneqq \sum_{k=1}^{t} \widehat{\alpha}_t^k \psi_k + \sum_{k=1}^{t} \widehat{\gamma}_t^k F_k(v_k)$. Given every fixed $\theta \in \Theta$ and $i \in [n]$, by definition, each $v_{j+1,i}$ follows $\mathcal{N}(0, \frac{\|\widehat{\alpha}_j\|_2^2}{n})$, and hence, the density function of $|v_{j+1,i}|$ is uniformly bounded by $\sqrt{\frac{2}{n\pi}} \|\widehat{\alpha}_j\|_2$. Therefore, it is sufficient to set $\sigma = \sqrt{\frac{2}{n\pi}} \|\widehat{\alpha}_j\|_2$ for assumption (282).

Additionally, the quantity of interest $\widehat{\beta}_{t+1}$ yields the following decomposition

$$\widehat{\beta}_{t+1} = v_{t+1} + \sum_{k=1}^{t} \widehat{\gamma}_t^k F_k(v_k) = v_{t+1} + \sum_{k=1}^{t} \widehat{\gamma}_t^k F_k(v_k^{\parallel} + v_k^{\perp}) \tag{294}$$
$$\text{where} \qquad v_k^{\parallel} = \frac{v_k^\top v_{t+1}}{\|v_{t+1}\|_2^2} v_{t+1}, \quad v_k^{\perp} = v_k - v_k^{\parallel},$$

where $v_k^{\parallel}$ denotes the component that aligns with v_{t+1} while $v_k^{\perp}$ denotes the component that is orthogonal to v_{t+1}. As discussed previously, given every fixed θ, each v_k follows a Gaussian distribution $\mathcal{N}(0, \frac{\|\widehat{\alpha}_j\|_2^2}{n} I_p)$, therefore $\widehat{\beta}_{t+1}$ is a function of Gaussian vectors. In addition, we make the following observation that

$$\begin{aligned}
\frac{|v_k^\top v_{t+1}|}{\|v_{t+1}\|_2^2} &= \left| \Big(\sum_{j=1}^{k-1} \widehat{\alpha}_{k-1}^j \psi_j \Big)^\top \sum_{j=1}^{t} \widehat{\alpha}_t^j \psi_j \right| \cdot \Big\| \sum_{j=1}^{t} \widehat{\alpha}_t^j \psi_j \Big\|_2^{-2} \\
&= \left| (\widehat{\alpha}_{k-1}^1, \dots, \widehat{\alpha}_{k-1}^{k-1}, 0, \dots, 0) \Psi^\top \Psi \widehat{\alpha}_t \right| \cdot (\widehat{\alpha}_t^\top \Psi^\top \Psi \widehat{\alpha}_t)^{-1} \\
&\leq \frac{p}{n} \Big(1 + \sqrt{\frac{t \log \frac{p}{\delta}}{p}} \Big) \|\widehat{\alpha}_{k-1}\|_2 \|\widehat{\alpha}_t\|_2 \cdot \frac{n}{p} \Big(1 - \sqrt{\frac{t \log \frac{p}{\delta}}{p}} \Big)^{-1} \|\widehat{\alpha}_t\|_2^{-2} \lesssim \frac{\|\widehat{\alpha}_{k-1}\|_2}{\|\widehat{\alpha}_t\|_2} \leq \frac{\overline{\alpha}_{k-1}}{\|\widehat{\alpha}_t\|_2}.
\end{aligned}$$

Recalling the assumption (191c) that $\rho_F \sum_{k=1}^{t} \widehat{\gamma}_t^k \overline{\alpha}_t \ll \|\widehat{\alpha}_t\|_2$, it therefore implies that conditioning on any value of $v_{t+1}^{\perp}$, $\widehat{\beta}_{t+1}$ is a Lipschitz function of v_{t+1} with Lipschitz constant of order 1. As a result, for every $i \in [n]$ and interval $\mathcal{I}_i$ of length ε, it holds that

$$\mathbb{P}_{v_{t+1}} \Big(\frac{1}{\|\alpha_t\|_2 / \sqrt{n}} \widehat{\beta}_{t+1,i} \in \mathcal{I}_i \;\Big|\; v_{t+1}^{\perp} \Big) \lesssim \varepsilon,$$

by noticing that $\widehat{\beta}_{t+1,i}$ is a $\Theta(1)$-Lipschitz function of $v_{t+1,i}$. This implies that assumption (282) holds with $\sigma = \|\alpha_t\|_2 / \sqrt{n}$.

Acknowledgment

This work was partially supported by the NSF grant DMS 2147546/2015447, the NSF CAREER award DMS-2143215, and the Google Research Scholar Award. Part of this work was done while G. Li and Y. Wei were visiting the Simons Institute for the Theory of Computing.

References

Adomaityte, U., Defilippis, L., Loureiro, B., and Sicuro, G. (2023). High-dimensional robust regression under heavy-tailed data: Asymptotics and universality. *arXiv preprint arXiv:2309.16476*.

Amelunxen, D., Lotz, M., McCoy, M. B., and Tropp, J. A. (2014). Living on the edge: Phase transitions in convex programs with random data. *Information and Inference: A Journal of the IMA*, 3(3):224–294.

Bao, Z., Han, Q., and Xu, X. (2023). A leave-one-out approach to approximate message passing. *arXiv preprint arXiv:2312.05911*.

Barbier, J., Krzakala, F., Macris, N., Miolane, L., and Zdeborová, L. (2019). Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460.

Bayati, M., Lelarge, M., and Montanari, A. (2015). Universality in polytope phase transitions and message passing algorithms. *The Annals of Applied Probability*, 25(2):753–822.

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

Bayati, M. and Montanari, A. (2011b). The LASSO risk for gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017.

Bellec, P. C. and Koriyama, T. (2023). Error estimation and adaptive tuning for unregularized robust M-estimator. *arXiv preprint arXiv:2312.13257*.

Bellec, P. C., Shen, Y., and Zhang, C.-H. (2022). Asymptotic normality of robust m-estimators with convex penalty. *Electronic Journal of Statistics*, 16(2):5591–5622.

Bellec, P. C. and Zhang, C.-H. (2022). De-biasing the lasso with degrees-of-freedom adjustment. *Bernoulli*, 28(2):713–743.

Bellec, P. C. and Zhang, C.-H. (2023). De-biasing convex regularized estimators and interval estimation in linear models. *The Annals of Statistics*, 51(2):391–436.

Bickel, P. J. (1975). One-step huber estimates in the linear model. *Journal of the American Statistical Association*, 70(350):428–434.

Bogdan, M., Van Den Berg, E., Sabatti, C., Su, W., and Candès, E. J. (2015). Slope—adaptive variable selection via convex optimization. *The annals of applied statistics*, 9(3):1103.

Bolthausen, E. (2009). On the high-temperature phase of the sherrington-kirkpatrick model. In *Seminar at EURANDOM, Eindhoven*.

Borell, C. (1975). The brunn-minkowski inequality in gauss space. *Inventiones mathematicae*, 30(2):207–216.

Bu, Z., Klusowski, J. M., Rush, C., and Su, W. J. (2020). Algorithmic analysis and statistical estimation of SLOPE via approximate message passing. *IEEE Transactions on Information Theory*, 67(1):506–537.

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Cademartori, C. and Rush, C. (2023). A non-asymptotic analysis of generalized approximate message passing algorithms with right rotationally invariant designs. *arXiv preprint arXiv:2302.00088*.

Cai, C., Poor, H. V., and Chen, Y. (2022). Uncertainty quantification for nonconvex tensor completion: Confidence intervals, heteroscedasticity and optimality. *IEEE Transactions on Information Theory*, 69(1):407–452.

Candes, E. and Tao, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n.

Candes, E. J., Romberg, J. K., and Tao, T. (2006). Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 59(8):1207–1223.

Celentano, M., Fan, Z., Lin, L., and Mei, S. (2023a). Mean-field variational inference with the TAP free energy: Geometric and statistical properties in linear models. *arXiv preprint arXiv:2311.08442*.

Celentano, M., Fan, Z., and Mei, S. (2023b). Local convexity of the TAP free energy and AMP convergence for z_2-synchronization. *The Annals of Statistics*, 51(2):519–546.

Celentano, M. and Montanari, A. (2021). Cad: Debiasing the lasso with inaccurate covariate model. *arXiv preprint arXiv:2107.14172*.

Celentano, M., Montanari, A., and Wei, Y. (2023c). The Lasso with general Gaussian designs with applications to hypothesis testing. *The Annals of Statistics*, 51(5):2194–2220.

Chen, S. S., Donoho, D. L., and Saunders, M. A. (2001). Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159.

Chen, W.-K. and Lam, W.-K. (2021). Universality of approximate message passing algorithms. *Electronic Journal of Probability*, 26:1–44.

Chen, Y., Chi, Y., Fan, J., and Ma, C. (2019a). Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *Mathematical Programming*, 176(1):5–37.

Chen, Y., Fan, J., Ma, C., and Yan, Y. (2019b). Inference and uncertainty quantification for noisy matrix completion. *Proceedings of the National Academy of Sciences*, 116(46):22931–22937.

Deshpande, Y., Abbe, E., and Montanari, A. (2017). Asymptotic mutual information for the balanced binary stochastic block model. *Information and Inference: A Journal of the IMA*, 6(2):125–170.

Deshpande, Y. and Montanari, A. (2014). Information-theoretically optimal sparse PCA. In *2014 IEEE International Symposium on Information Theory*, pages 2197–2201. IEEE.

Donoho, D. and Montanari, A. (2016). High dimensional robust m-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166(3):935–969.

Donoho, D. L. (2006). Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306.

Donoho, D. L., Huo, X., et al. (2001). Uncertainty principles and ideal atomic decomposition. *IEEE transactions on information theory*, 47(7):2845–2862.

Donoho, D. L., Javanmard, A., and Montanari, A. (2013). Information-theoretically optimal compressed sensing via spatial coupling and approximate message passing. *IEEE transactions on information theory*, 59(11):7434–7464.

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

Donoho, D. L. and Montanari, A. (2015). Variance breakdown of huber (m)-estimators: $n/p \to m$ in $(1, \infty)$. *arXiv preprint arXiv:1503.02106*.

Dudeja, R., Lu, Y. M., and Sen, S. (2022). Universality of approximate message passing with semi-random matrices. *arXiv preprint arXiv:2204.04281*.

El Karoui, N. (2013). Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results. *arXiv preprint arXiv:1311.2445*.

El Karoui, N. (2018). On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields*, 170(1):95–175.

El Karoui, N., Bean, D., Bickel, P. J., Lim, C., and Yu, B. (2013). On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, 110(36):14557–14562.

Fan, J., Li, Q., and Wang, Y. (2014). Robust estimation of high-dimensional mean regression. *arXiv preprint arXiv:1410.2150*.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Fan, J., Li, R., Zhang, C.-H., and Zou, H. (2020). *Statistical foundations of data science*. CRC press.

Fan, Z. (2022). Approximate message passing algorithms for rotationally invariant matrices. *The Annals of Statistics*, 50(1):197–224.

Feng, O. Y., Venkataramanan, R., Rush, C., and Samworth, R. J. (2022). A unifying tutorial on approximate message passing. *Foundations and Trends® in Machine Learning*, 15(4):335–536.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, 222(594-604):309–368.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the american statistical association*, pages 383–393.

Han, Q. and Shen, Y. (2023). Universality of regularized regression estimators in high dimensions. *The Annals of Statistics*, 51(4):1799–1823.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC press.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, pages 73–101.

Huber, P. J. (1973). Robust regression: asymptotics, conjectures and monte carlo. *The Annals of Statistics*, pages 799–821.

Javanmard, A. and Montanari, A. (2013). State evolution for general approximate message passing algorithms, with applications to spatial coupling. *Information and Inference: A Journal of the IMA*, 2(2):115–144.

Javanmard, A. and Montanari, A. (2014). Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *IEEE Transactions on Information Theory*, 60(10):6522–6554.

Javanmard, A. and Montanari, A. (2018). Debiasing the lasso: Optimal sample size for gaussian designs. *The Annals of Statistics*, 46(6A):2593–2622.

Le Cam, L. (2012). *Asymptotic methods in statistical decision theory*. Springer Science & Business Media.

Lei, L., Bickel, P. J., and El Karoui, N. (2018). Asymptotics for high dimensional regression M-estimates: fixed design results. *Probability Theory and Related Fields*, 172:983–1079.

Li, G., Fan, W., and Wei, Y. (2023a). Approximate message passing from random initialization with applications to Z_2 synchronization. *Proceedings of the National Academy of Sciences*, 120(31):e2302930120.

Li, G. and Wei, Y. (2022). A non-asymptotic framework for approximate message passing in spiked models. *arXiv preprint arXiv:2208.03313*.

Li, Y., Fan, Z., Sen, S., and Wu, Y. (2023b). Random linear estimation with rotationally-invariant designs: Asymptotics at high temperature. *IEEE Transactions on Information Theory*.

Li, Y. and Wei, Y. (2021). Minimum ℓ_1-norm interpolators: Precise asymptotics and multiple descent. *arXiv preprint arXiv:2110.09502*.

Loh, P.-L. (2017). Statistical consistency and asymptotic normality for high-dimensional robust m-estimators. *The Annals of Statistics*, 45(2):866 – 896.

Loh, P.-L. and Wainwright, M. J. (2013). Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Advances in Neural Information Processing Systems*.

Ma, C., Wang, K., Chi, Y., and Chen, Y. (2020). Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. *Foundations of Computational Mathematics*, 20(3):451–632.

Ma, Z. and Nandy, S. (2021). Community detection with contextual multilayer networks. *arXiv preprint arXiv:2104.02960*.

Maleki, A., Anitori, L., Yang, Z., and Baraniuk, R. G. (2013). Asymptotic analysis of complex LASSO via complex approximate message passing (CAMP). *IEEE Transactions on Information Theory*, 59(7):4290–4308.

Maronna, R. A., Martin, R. D., Yohai, V. J., and Salibián-Barrera, M. (2019). *Robust statistics: theory and methods (with R)*. John Wiley & Sons.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462.

Miolane, L. and Montanari, A. (2021). The distribution of the lasso: Uniform control over sparse balls and adaptive parameter tuning. *The Annals of Statistics*, 49(4):2313–2335.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the american statistical association*, 83(404):1023–1032.

Mondelli, M. and Venkataramanan, R. (2021). PCA initialization for approximate message passing in rotationally invariant models. *Advances in Neural Information Processing Systems*, 34:29616–29629.

Mondelli, M. and Venkataramanan, R. (2022). Approximate message passing with spectral initialization for generalized linear models. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(11):114003.

Montanari, A. and Venkataramanan, R. (2021). Estimation of low-rank matrices via approximate message passing. *The Annals of Statistics*, 49(1):321–345.

Montanari, A. and Wu, Y. (2023). Posterior sampling from the spiked models via diffusion processes. *arXiv preprint arXiv:2304.11449*.

Rangan, S. (2011). Generalized approximate message passing for estimation with random linear mixing. In *2011 IEEE International Symposium on Information Theory Proceedings*, pages 2168–2172. IEEE.

Rangan, S. and Fletcher, A. K. (2012). Iterative estimation of constrained rank-one matrices in noise. In *2012 IEEE International Symposium on Information Theory Proceedings*, pages 1246–1250. IEEE.

Ren, Z., Sun, T., Zhang, C.-H., and Zhou, H. H. (2015). Asymptotic normality and optimalities in estimation of large gaussian graphical models.

Rousseeuw, P. J. and Leroy, A. M. (2005). *Robust regression and outlier detection*. John wiley & sons.

Rudelson, M. and Vershynin, R. (2008). On sparse reconstruction from fourier and gaussian measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 61(8):1025–1045.

Rush, C. and Venkataramanan, R. (2018). Finite sample analysis of approximate message passing algorithms. *IEEE Transactions on Information Theory*, 64(11):7264–7286.

Schniter, P. and Rangan, S. (2014). Compressive phase retrieval via generalized approximate message passing. *IEEE Transactions on Signal Processing*, 63(4):1043–1055.

Su, W., Bogdan, M., and Candes, E. (2017). False discoveries occur early on the lasso path. *The Annals of statistics*, pages 2133–2150.

Sun, Q., Zhou, W.-X., and Fan, J. (2020). Adaptive huber regression. *Journal of the American Statistical Association*, 115(529):254–265.

Sur, P. and Candès, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525.

Sur, P., Chen, Y., and Candès, E. J. (2019). The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *Probability Theory and Related Fields*, 175(1-2):487–558.

Thrampoulidis, C., Abbasi, E., and Hassibi, B. (2018). Precise error analysis of regularized m-estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Venkataramanan, R., Kögler, K., and Mondelli, M. (2021). Estimation in rotationally invariant generalized linear models via approximate message passing. *arXiv preprint arXiv:2112.04330*.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5):2183–2202.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.

Wang, T., Zhong, X., and Fan, Z. (2022). Universality of approximate message passing algorithms and tensor networks. *arXiv preprint arXiv:2206.13037*.

Wu, Y. and Zhou, K. (2024). Sharp analysis of power iteration for tensor pca. *arXiv preprint arXiv:2401.01047*.

Xia, D. and Yuan, M. (2021). Statistical inferences of linear forms for noisy matrix completion. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(1):58–77.

Yan, Y., Chen, Y., and Fan, J. (2021). Inference for heteroskedastic PCA with missing data. *arXiv preprint arXiv:2107.12365*.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(1):49–67.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):217–242.

Zhang, Y., Ji, H. C., Venkataramanan, R., and Mondelli, M. (2023). Spectral estimators for structured generalized linear models via approximate message passing. *arXiv preprint arXiv:2308.14507*.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563.

Zhong, X., Wang, T., and Fan, Z. (2021). Approximate message passing for orthogonally invariant ensembles: Multivariate non-linearities and spectral initialization. *arXiv preprint arXiv:2110.02318*.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320.