PNAS RESEARCH ARTICLE | STATISTICS

Approximate message passing from random initialization with applications to Z_2 synchronization

Gen Li[a], Wei Fan[a], and Yuting Wei[a,1]

Edited by Marco Mondelli, Institute of Science and Technology Austria, Klosterneuburg, Austria; received February 20, 2023; accepted June 24, 2023 by Editorial Board Member David A. Weitz

This paper is concerned with the problem of reconstructing an unknown rank-one matrix with prior structural information from noisy observations. While computing the Bayes optimal estimator is intractable in general due to the requirement of computing high-dimensional integrations/summations, Approximate Message Passing (AMP) emerges as an efficient first-order method to approximate the Bayes optimal estimator. However, the theoretical underpinnings of AMP remain largely unavailable when it starts from random initialization, a scheme of critical practical utility. Focusing on a prototypical model called $\mathbb{Z}_2$ synchronization, we characterize the finite-sample dynamics of AMP from random initialization, uncovering its rapid global convergence. Our theory—which is nonasymptotic in nature—in this model unveils the non-necessity of a careful initialization for the success of AMP.

approximate message passing | random initialization | nonasymptotic analysis | spiked Wigner model | global convergence

The problem of estimating an unknown low-rank matrix, when given access to highly noisy observations, has been the subject of considerable studies, shedding light on a diverse array of contexts including collaborative filtering, synchronization and alignment, localization, and causal panel data, to name just a few (1–8). While low-rank estimators are not in short supply, the quest for algorithms that can work all the way to the information-theoretic limits continues to inspire theoretical and algorithmic development.

1. Motivation and An Informal Overview

In this paper, we focus on how to reconstruct a structured signal $v^\star \in \mathbb{R}^n$ (or equivalently, $v^\star v^{\star\top}$) from noisy data:

$$M = \lambda v^\star v^{\star\top} + W \in \mathbb{R}^{n\times n} \qquad \text{with } \lambda > 0. \tag{1}$$

This classical model is commonly referred to as a deformed Gaussian Wigner model or spiked Gaussian Wigner model when the entries of the noise matrix $W = [W_{ij}]_{1\le i,j\le n}$ are independently drawn from Gaussian distributions—more precisely, $W_{ii} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \frac{2}{n})$ and $W_{ji} = W_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \frac{1}{n})$ for $i \neq j$—which serves as a prototypical model toward understanding the feasibility and fundamental limits of low-rank matrix estimation.

The spectral properties of the observed matrix M have been extensively studied (see, e.g. refs. 9–13), motivating the design of spectral methods when there is no structural information associated with (1, 3, 14–16). In practice, there is no shortage of applications where additional structural information about $v^\star$ is available a priori, examples including finite-group structure (17), cone constraints (18, 19), and sparsity (20, 21), among others. The presence of prior structure further exacerbates the nonconvexity issue when computing the maximum likelihood estimate or Bayes optimal estimate, thereby presenting a pressing need for the search of algorithms that can be executed efficiently.

Remarkably, the approximate message passing (AMP) algorithm emerges as an efficient nonconvex paradigm that rises to the aforementioned challenge (22, 23). Originally proposed in the context of compressed sensing, AMP has served as not only a family of first-order iterative algorithms that enjoy rapid convergence (24–28) but also a powerful statistical machinery that assists in determining the performance limits of other statistical procedures in high-dimensional asymptotics (29–38). Over the past two decades, AMP has also received widespread adoption in a variety of engineering and science applications, including but not limited to imaging, wireless communications, signal processing, and deep learning (see, e.g., refs. 39–43 and references therein).

Significance

Approximate Message Passing (AMP) serves as both a family of efficient first-order algorithms and a powerful theoretical machinery for high-dimensional data analysis, which has found applications in a diverse array of problems such as sparse regression, generalized linear models, and low-rank matrix and tensor estimation. While the existing suite of AMP theory covers a wealth of applications, the theoretical guarantees for AMP remain mostly unavailable when it starts from random initialization, limiting its applicability. To address this issue, our paper delivers a nonasymptotic characterization of AMP when initialized randomly, justifying and advocating the use of random initialization in practice. In other words, a carefully designed initialization is completely unnecessary for the success of AMP.

Author affiliations: [a]Department of Statistics and Data Science, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104

Author contributions: G.L. and Y.W. designed research; G.L., W.F., and Y.W. performed research; G.L. and W.F. analyzed data; and G.L., W.F., and Y.W. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. M.M. is a guest editor invited by the Editorial Board.

[1]To whom correspondence may be addressed. Email: ytwei@wharton.upenn.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2302930120/-/DCSupplemental.

Published July 25, 2023.

Inadequacy of Prior AMP Theory. Nevertheless, while the existing suite of AMP theory covers a wealth of applications, it remains inadequate in at least two aspects. To begin with, a dominant fraction of existing AMP theory is asymptotic in nature, in the sense that it predicts the AMP dynamics in the large-n limit for any fixed iteration t. For this reason, prior AMP theory falls short of describing how AMP evolves after a growing number of iterations (even when it is run for only $\log n$ iterations), which stands in contrast to other optimization-based procedures that often come with nonasymptotic analysis accommodating a large number of iterations (3, 6, 44). Another issue that further complicates matters stems from the requirement of an informative initialization, that is, existing AMP theory for low-rank estimation often requires starting from a point that already enjoys nonvanishing correlation with the true signal (45–47). While an informative initial estimate like spectral initialization is sometimes plausible and analyzable, this requirement presents a hurdle to understanding the effectiveness of other widely adopted alternatives like random initialization. This motivates the following natural questions that remain by and large open:

> *Is a warm start like spectral initialization necessary for the success of AMP? Can we start with a simpler initialization scheme but still work equally well as spectral initialization?*

Thus far, there has been no rigorous evidence precluding one from starting randomly and uninformatively. As shall be made clear shortly, tackling this issue necessitates a different and powerful nonasymptotic framework for AMP, due to the difficulty of tracking the AMP dynamics when the iterates exhibit only extremely weak correlation with the truth.

Inspired by the aforementioned issues, there has been growing interest in understanding the finite-sample performance of AMP. A seminal work by Rush and Venkataramanan (48) [see also its follow-up work (49)], studied AMP for sparse regression and permitted the total number of iterations to be as large as $o\big(\frac{\log n}{\log\log n}\big)$. This order of iteration number, however, is still highly insufficient in understanding randomly initialized AMP, as at least an order of $\log n$ iterations might be required for AMP to achieve nontrivial correlation with the truth. A recent work by Li and Wei (50) developed a nonasymptotic framework for the spiked Gaussian Wigner model, which characterized the AMP behavior for up to $O\big(\frac{n}{\mathsf{poly}(\log n)}\big)$ iterations. Although the theory therein is well suited to the studies of spectrally initialized AMP, it remains largely elusive whether it is capable of accommodating random initialization, a circumstance whose resultant initial stage is far more challenging and subtle to track.

This Paper: Randomly Initialized AMP for $\mathbb{Z}_2$ Synchronization. In this work, we take a step toward addressing the above challenges by studying a concrete model called $\mathbb{Z}_2$ synchronization. To be precise, $\mathbb{Z}_2$ synchronization is a special case of the spiked Gaussian Wigner model when the ground truth is known to have a discrete structure obeying $v^\star \in \{\pm\frac{1}{\sqrt{n}}\}^n$. Here and throughout, we impose a prior distribution on $v^\star = [v_i^\star]_{1\le i\le n}$ such that

$$v_i^\star \overset{\text{i.i.d.}}{\sim} \mathsf{Unif}\Big(\pm\frac{1}{\sqrt{n}}\Big), \qquad 1\le i\le n.$$

The goal is to reconstruct $v^\star$ on the basis of the measurements M (Eq. **1**). This problem can be viewed as a basic example of a more general problem—synchronization over compact groups (1, 2, 17, 51–53)—and has an intimate connection to stochastic block models (35, 54).

The AMP Algorithm. Note that it is in general intractable to calculate the Bayes optimal solution directly due to computational difficulty in computing high-dimensional integrations/summations. A common alternative is to resort to the variational inference approximation, while the computational challenge still remains due to the nonconvexity nature of the variational inference objective. This motivates the search for computationally feasible alternatives, for which AMP emerges as a natural and successful option (46, 50, 54, 55). More concretely, given the initialization $x_0, x_1 \in \mathbb{R}^n$, AMP tailored to $\mathbb{Z}_2$ synchronization adopts the following update rule:

$$x_{t+1} = M\eta_t(x_t) - \langle \eta_t'(x_t)\rangle \eta_{t-1}(x_{t-1}), \qquad t\ge 1, \quad \textbf{[2]}$$

where we denote $\langle x\rangle := \frac{1}{n}\sum_{i=1}^n x_i$ for any vector $x = [x_i]_{1\le i\le n}$, and the denoising function is given by*

$$\begin{aligned} \eta_t(x) &= \gamma_t \tanh(\pi_t x), \qquad \text{for } t\ge 1 \\ \text{with } \pi_t &:= \sqrt{\max\big\{n(\|x_t\|_2^2 - 1), 1\big\}} \\ \text{and } \gamma_t &:= \|\tanh(\pi_t x_t)\|_2^{-1}. \end{aligned} \quad \textbf{[3]}$$

Here, it is understood that the functions $\eta_t(\cdot)$, $\eta_t'(\cdot)$ and $\tanh(\cdot)$ are applied entrywise if the input argument is a vector.

Thus far, there have been two strategies to accommodate a growing number of iterations in the most challenging regime (i.e., when λ is above but very close to the information-theoretic threshold 1). One attempt was made by Celentano et al. (46), which proposed a three-stage hybrid algorithm that runs spectrally initialized AMP followed by natural gradient descent (NGD). It was conjectured therein that the third stage (i.e., NGD) is unnecessary. Recently, Li and Wei (50) put forward another strategy to address this conjecture, showing that a third refinement stage is indeed not needed as long as spectral initialization is adopted. Despite the nonconvex nature of the underlying optimization problem, AMP with spectral initialization is nearly Bayes optimal.

The Effect of Random Initialization. As alluded to previously, all existing AMP theory for this problem (45, 46, 50, 56) requires informative initialization obtained by, for example, spectral methods. By contrast, one initialization strategy that enjoys widespread adoption is to initialize AMP randomly; for instance,

$$\begin{aligned} x_1 \sim \mathcal{N}\Big(0, \frac{1}{n}I_n\Big) \ \ &\text{(independent of } M\text{)} \\ \text{and} \quad \eta_0(x_0) &= 0. \end{aligned} \quad \textbf{[4]}$$

In order to investigate whether a warm start is required for AMP to be effective, let us first conduct a series of numerical experiments using Eq. **4**, as reported in Fig. 1. Encouragingly, AMP with random initialization seems to work surprisingly well: it only takes several tens of iterations to achieve nearly the same performance as spectrally initialized AMP (note that spectral initialization also consists of several tens of power iterations). Such encouraging numerical results motivate us to pursue in-depth theoretical understanding about the effect of random initialization upon AMP convergence, which was previously unavailable in the literature.

*Note that for ease of analysis, we adopt a slightly different scaling from that of ref. 54, but they are equivalent up to global scaling.

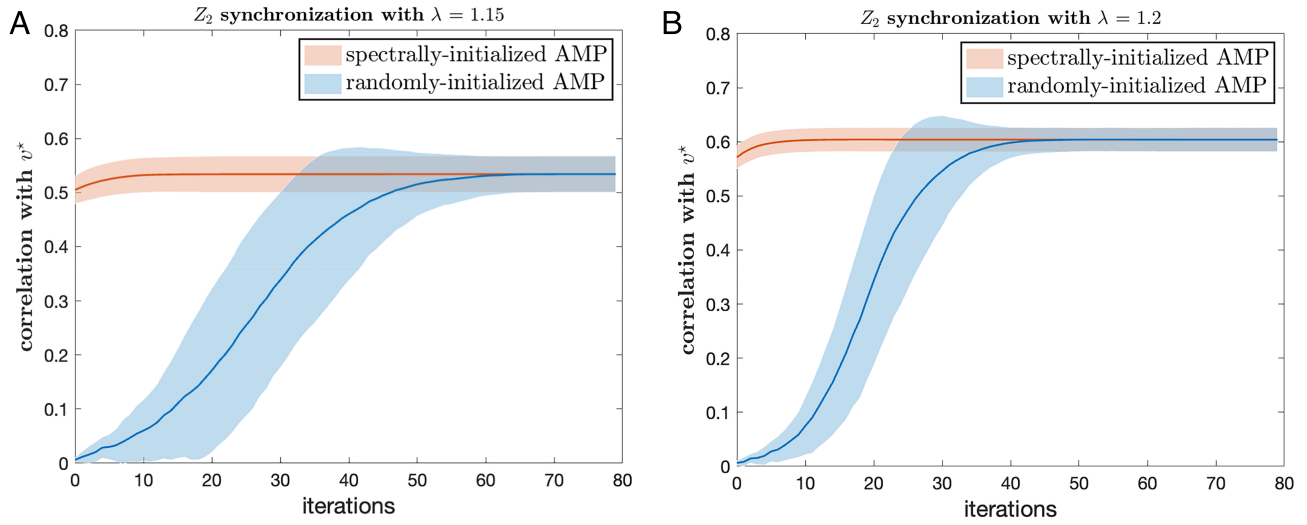

Fig. 1. The correlation of $\eta_t(x_t)$ and $v^\star$ (i.e., $\frac{|\langle \eta_t(x_t), v^\star \rangle|}{\|\eta_t(x_t)\|_2}$) vs. iteration count t for AMP with both random and spectral initialization. Here, n = 10,000 and $v_i^\star \overset{\text{i.i.d.}}{\sim} \text{Unif}(\pm\frac{1}{\sqrt{n}})$ $(1 \le i \le n)$. We generate 20 independent copies of M according to Eq. **1** and report the averaged results, with the width of the shaded region reflecting (twice) the SD. Plots (*A*) and (*B*) correspond to $\lambda = 1.15$ and $\lambda = 1.2$, respectively.

Main Contributions and Technical Challenges. In the present paper, we provide a nonasymptotic analysis that allows one to predict how AMP evolves over time from random initialization, even when the signal strength λ is exceedingly close to the information-theoretic limit. Our theory is able to track the correlation of the AMP iterates and the truth $v^\star$. In particular, we demonstrate in Theorem 1 that the signal component in the AMP iterates increases exponentially fast at the initial stage, taking no more than $O\big(\frac{\log n}{\lambda - 1}\big)$ iterations to grow from $\widetilde{O}(\frac{1}{\sqrt{n}})$ to $O(\sqrt{\lambda^2 - 1})$ (the latter of which coincides with the correlation of spectral initialization and the truth). Furthermore, once the signal component surpasses $O(\sqrt{\lambda^2 - 1})$ in magnitude, the finite-sample AMP dynamics are very well predicted by the asymptotic state evolution recursion derived previously for any fixed t and $n \to \infty$ (even though we are working with the finite-sample regime). Our paper characterizes the performance of AMP when initialized randomly, justifying and advocating the use of random initialization. Put another way, a carefully designed warm start is not necessary at all for this problem.

Built upon the analysis recipe recently developed by Li and Wei (50), the development of our theory requires ideas far beyond this framework in order to track AMP from random initialization. Before continuing, we take a moment to single out the key technical hurdles that need to be overcome.

- Prior theory based on state evolution analysis falls short of offering "fine-grained" understanding about the AMP iterates when they have vanishingly small correlation with the truth. More precisely, past theory fails to measure the progress of AMP during the initial stage when its signal component is of strength $o(1)$ (in fact, as small as $\widetilde{O}(\frac{1}{\sqrt{n}})$ when initialized), but instead treats the signal strength as 0 in the large-n limit.
- Another technical challenge results from the complicated statistical dependency across iterations, which is particularly difficult to cope with when the algorithm starts with random initialization and when the number of iterations grows with the dimension n. While prior literature tackles this issue for other nonconvex optimization methods by resorting to either delicate leave-one-out decoupling arguments (see, e.g. ref. 57) or global landscape analysis (see, e.g. ref. 58), these approaches remain unavailable when analyzing AMP.

Notation. Finally, let us introduce a set of notation that shall be useful throughout. We use $\varphi(\cdot)$ (resp. $\varphi_n(\cdot)$) to denote the probability density function (p.d.f.) of a standard Gaussian random variable (resp. a Gaussian random vector $\mathcal{N}(0, I_n)$). For any matrix M, we let $\|M\|$ and $\|M\|_{\mathsf{F}}$ denote the spectral norm and the Frobenius norm of M, respectively. For any vector $x \in [x_i]_{1 \le i \le n} = \mathbb{R}^n$, we denote by $|x|_{(i)}$ (resp. $x_{(i)}$) the absolute value (resp. value) of the i-th largest entry of x in magnitude. We write $\mathcal{S}^{d-1} = \{x \in \mathbb{R}^d \mid \|x\|_2 = 1\}$ as the unit sphere in ^{d}. Moreover, for any two vectors x, y $\mathbb{R}^n$, we write $x \circ y$ for their Kronecker product, namely, $x \circ y = (x_1 y_1, \ldots, x_n y_n)^\top$ $\mathbb{R}^n$. When a function is applied to a vector, it should be understood as being applied in a component-wise fashion; for instance, for any vector $x = [x_i]_{1 \le i \le n}$, we let $x + 1 := [x_i + 1]_{1 \le i \le n}$.

In addition, given two functions $f(n)$ and $g(n)$, we write $f(n) \lesssim g(n)$ or $f(n) = O(g(n))$ to indicate that $|f(n)| \le c_1 g(n)$ for some universal constant $c_1 > 0$ independent of n, and similarly, $f(n) \gtrsim g(n)$ means that $f(n) \ge c_2 |g(n)|$ for some universal constant $c_2 > 0$. We write $f(n) = \widetilde{O}(g(n))$ if $f(n) = O(g(n))$ up to logarithm factors. We also adopt the notation $f(n) \asymp g(n)$ to indicate that both $f(n) \lesssim g(n)$ and $f(n) \gtrsim g(n)$ hold simultaneously. Moreover, when we write $f(n) \ll g(n)$ or $f(n) = o(g(n))$, it means $f(n)/g(n) \to 0$ as $n \to \infty$; we also write $f(n) \gg g(n)$ if $g(n)/f(n) \to 0$ as $n \to \infty$. We use c, C to denote universal constants that do not depend on n, whose values might change from line to line.

2. Main Results

In this section, we provide precise statements of our main theoretical guarantees for randomly initialized AMP. For notational convenience, let us introduce

$$\alpha_{t+1} := \lambda v^{\star\top} \eta_t(x_t), \qquad [5]$$

which captures the projection of the t-th iterate (after denoising) onto the direction of the truth $v^\star$. In some sense, this quantity captures the size of the signal component carried by the t-th iterate. With this notation in place, we single out a key threshold as follows:

$$\varsigma := \min\left\{t : |\alpha_t| \geq \frac{1}{2}\sqrt{\lambda^2 - 1}\right\}, \qquad [6]$$

which reflects the time taken for the AMP iterate to carry a significant signal component (note that a random initial guess obeys $|v^{\star\top} x_1| \lesssim \widetilde{O}(\frac{1}{\sqrt{n}})$, meaning that the initial signal component is exceedingly small). Additionally, we define the state evolution recursion starting from the ς-th iteration as follows for any $t \geq \varsigma$

$$\alpha^\star_\varsigma = |\alpha_\varsigma| \quad \text{and} \quad \alpha^\star_{t+1} = \lambda\left[\int \tanh\left(\alpha^\star_t\left(\alpha^\star_t + x\right)\right)\varphi(\mathrm{d}x)\right]^{1/2}. \qquad [7]$$

Notably, the asymptotic state evolution recursion (which is concerned with a 1-dimensional sequence in this case) is known to faithfully track the dynamics of AMP for any fixed t in the limit when $n \to \infty$, although its utility in the finite-sample regime was poorly understood in theory.

Equipped with the above definitions, our main results are summarized in the following theorem.

Theorem 1. *Consider the $\mathbb{Z}_2$ synchronization problem with*

$$n^{-1/9}\log n \lesssim \lambda - 1 \leq 0.2.$$

*Suppose we run AMP (cf. Eqs. **2** and **3**) with random initialization* Eq. **4**. *Consider any t obeying $1 \leq t \leq \frac{cn(\lambda-1)^5}{\log^2 n}$, where $c > 0$ is some universal constant. Then, with probability at least $1 - O(n^{-10})$, the following results hold:*

- *(Decomposition and error bound). The AMP iterates admit the decomposition*

$$x_t = \alpha_t v^\star + \sum_{k=1}^{t-1}\beta^k_{t-1}\phi_k + \xi_{t-1}, \qquad [8a]$$

*where α_t is defined in Eq. **5**, the ϕ_k's are i.i.d. Gaussian vectors obeying $\phi_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \frac{1}{n}I_n)$, and*

$$\|\beta_t\|_2 := \left\|(\beta^1_t, \beta^2_t, \ldots, \beta^t_t)\right\|_2 = \|\eta_t(x_t)\|_2 = 1, \qquad [8b]$$

$$\|\xi_t\|_2 \lesssim \sqrt{\frac{t\log n}{n(\lambda-1)^2}} + \sqrt{\frac{\log^4 n}{n(\lambda-1)^3}}; \qquad [8c]$$

- *(Crossing time). The threshold ς defined in Eq. **6** satisfies*

$$\varsigma = O\left(\frac{\log n}{\lambda - 1}\right); \qquad [9]$$

- *(Nonasymptotic state evolution). For any t obeying $\varsigma \leq t \leq \frac{cn(\lambda-1)^5}{\log^2 n}$, we have*

$$\alpha_t^2 = \left(1 + O\left(\sqrt{\frac{\left(t + \frac{\log^3 n}{\lambda-1}\right)\log n}{n(\lambda-1)^5}}\right)\right)\alpha^{\star 2}_t, \qquad [10]$$

where $\{\alpha^\star_t\}$ stand for the asymptotic state evolution parameters defined in Eq. 7.

***Remark 1* (Range of λ):** Theorem 1 only focuses on the regime where λ is larger than but close to 1. In fact, $\lambda = 1$ represents the phase transition point for $\mathbb{Z}_2$ synchronization (54), in the sense that i) when $\lambda < 1$, no estimator performs better than the 0 estimator asymptotically, and ii) when λ is strictly larger than 1, it is possible to achieve nontrivial correlation with $v^\star$. We focus on the feasible regime by considering a more refined yet highly challenging case with $\lambda - 1 \gtrsim n^{-1/9}\log n$ (so that λ can be very close to 1). While it is possible to improve the exponent 1/9, it is beyond the scope of this paper. The upper bound $\lambda \leq 1.2$ is not crucial at all as the problem becomes easier as λ increases. In fact, our result continues to hold when $\lambda > 1.2$, which can be justified via a more refined characterization of the residual term ξ_t as well as κ_t. This paper imposes this assumption $\lambda \leq 1.2$ merely to streamline our presentation and analysis.

Remark 2: We remark that while the iterates x_t are random quantities that depend on the randomnesses in W and $v^\star$, the decomposition Eq. **8a** is purely deterministic. For definitions and properties of $\{\phi_k\}_{k\leq t-1}$ and $\{\beta^k_{t-1}\}_{k\leq t-1}$, we refer the readers to *SI Appendix*, section A.2.2. In order to ensure that each ϕ_k yields a homogeneous Gaussian distribution $\mathcal{N}(0, \frac{1}{n}I_n)$, we have included in ϕ_k additional terms that involve extra randomnesses $\{g^k_i\}_{k\leq t-1}$. These terms are properly subtracted and reflected in the residual ξ_{t-1}. As a result, the right-hand side of expression Eq. **8a** is a function of and therefore measurable with respect to W and $v^\star$.

In the sequel, we provide some interpretations of Theorem 1 and discussions about its implications. It is assumed below that $\lambda > 1$.

Gaussian Approximation. The first result Eq. **8a** in Theorem 1 asserts that each AMP iterate is composed of three components: i) a signal component $\alpha_t v^\star$ that aligns with the true signal $v^\star$, ii) a noise component $\sum_{k=1}^{t-1}\beta^k_{t-1}\phi_k$ that is a linear combination of i.i.d. Gaussian vectors, and iii) a residual component ξ_{t-1}. While this decomposition resembles that of ref. 50, we justify its validity even in the absence of carefully designed spectral initialization. A few remarks are in order.

- Regarding the noise component, Theorem 1 implies that the 1-Wasserstein distance between its distribution (denoted by $\mu\left(\sum_{k=1}^t \beta^k_t\phi_k\right)$) and a Gaussian distribution $\mathcal{N}\left(0, \frac{1}{n}I_n\right)$ is at most

$$W_1\left(\mu\left(\sum_{k=1}^t \beta_k\phi_k\right), \mathcal{N}\left(0, \frac{1}{n}I_n\right)\right) \lesssim \sqrt{\frac{t\log n}{n}}. \qquad [11]$$

  For t not too large, the noise component well approximates a Gaussian vector $\mathcal{N}(0, \frac{1}{n}I_n)$.
- Regarding the signal component $\alpha_t v^\star$, it is self-evident that α_t governs how effective AMP is in recovering the true signal. Importantly, once $|\alpha_t|$ exceeds the threshold $\frac{1}{2}\sqrt{\lambda^2-1}$, it follows a nonasymptotic state evolution that closely resembles the asymptotic counterpart $\alpha^\star_t$ (Eq. **10**), a result that is made possible thanks to the nonasymptotic nature of our analysis.

To summarize, up to a small error term at most $\widetilde{O}(\sqrt{\frac{t}{n(\lambda-1)^2}} + \sqrt{\frac{1}{n(\lambda-1)^3}})$, the AMP iterate is approximately

$$x_t \approx \alpha_t v^\star + \mathcal{N}\left(0, \frac{1}{n}I_n\right), \qquad t < O\left(\frac{n(\lambda-1)^5}{\log^2 n}\right),$$

even when initialized randomly. An asymptotic version of this observation has been made in ref. 45, although the result therein required both informative initialization and a fixed t that does not grow with n.

Dynamics after Random Initialization. The most challenging element of Theorem 1 lies in analyzing the initial stage after random initialization. As shall be made clear from our analysis, we can understand the AMP trajectory by dividing it into three phases.

- Phase #1: escaping from random initialization. When initialized randomly with $x_1 \sim \mathcal{N}(0, \frac{1}{n} I_n)$, AMP starts with an extremely small signal component about the order of $\widetilde{O}(\frac{1}{\sqrt{n}})$, for which the canonical state evolution becomes vacuous. To overcome this technical hurdle, we develop fine-grained characterizations regarding how α_t evolves in this phase (before $|\alpha_t|$ surpasses $\sqrt{\lambda-1}\, n^{-1/4}$), that is,

$$\alpha_{t+1} \approx \lambda \alpha_t + \lambda g_{t-1}, \qquad \text{with } g_{t-1} \sim \mathcal{N}\Big(0, \frac{1}{n}\Big); \qquad [12]$$

  see *SI Appendix*, section B.4 for details. This approximate noisy recursion tells us that while the signal component might be initially buried under the noise term, it takes at most $O(\frac{\log n}{\lambda-1})$ iterations for the signal component to rise above the noise size and reach the order of $\sqrt{\lambda-1}\, n^{-1/4}$ (*SI Appendix*, section A.2.2).
- Phase #2: exponential growth. Once the signal component exceeds $\sqrt{\lambda-1} n^{-1/4}$ in size, the AMP iterate correlates nontrivially with the true signal. Interestingly, the signal strength α_t starts to grow exponentially until reaching the order of $\sqrt{\lambda^2-1}$. As we shall justify in *SI Appendix*, section A.2.2, α_{t+1} obeys

$$|\alpha_{t+1}| \geq \sqrt{1 + \frac{1-o(1)}{3}(\lambda-1)}\, |\alpha_t|, \qquad [13]$$

  in this phase, which accounts for at most $O(\frac{\log n}{\lambda-1})$ iterations.
- Phase #3: local refinement. Upon reaching the order of $\sqrt{\lambda^2-1}$, $|\alpha_t|$ enters a local refinement phase, during which randomly initialized AMP behaves similarly as AMP with spectral or other informative initialization. In this phase, the asymptotic state evolution Eq. **7** also starts to be effective when predicting the evolution of α_t (Eq. **10**). As we shall solidify in *SI Appendix*, section A.2.4, the signal strength α_t satisfies

$$|\alpha_{t+1}^2 - \alpha^{\star 2}| \lesssim \big(1-(\lambda-1)\big)^{t-\varsigma} + \widetilde{O}\left(\sqrt{\frac{t+\frac{1}{\lambda-1}}{n(\lambda-1)^5}}\right), \qquad [14]$$

  where $\alpha^\star$ (determined by λ) denotes the limit of $\alpha_t^\star$ as $t \to \infty$ (cf. Eq. **7**) and is unique solution of

$$\alpha^{\star 2} = \lambda^2 \mathbb{E}\big[\tanh\big(\alpha^\star(\alpha^\star + G)\big)\big], \quad \text{with } G \sim \mathcal{N}(0,1). \qquad [15]$$

Bayes Optimality. As was shown previously [see e.g., (46, Lemma A.7)], we can construct an AMP-based estimator whose risk coincides with that of the Bayes optimal estimator $\widehat{X}^{\text{bayes}} := \mathbb{E}[v^\star v^{\star\top} \mid M]$. More precisely, taking the AMP-based estimator as

$$u_t := \frac{1}{\lambda\sqrt{n(\alpha_t^2+1)}} \tanh(\pi_t x_t), \qquad [16]$$

its asymptotic risk satisfies [*SI Appendix*, section C and (54)]:

$$\lim_{t\to\infty}\lim_{n\to\infty} \mathbb{E}\big[\big\|v^\star v^{\star\top} - u_t u_t^\top\big\|_{\mathsf{F}}^2\big] = \lim_{n\to\infty} \mathbb{E}\big[\|v^\star v^{\star\top} - \widehat{X}^{\text{bayes}}\|_{\mathsf{F}}^2\big] = 1 - \frac{\alpha^{\star 4}}{\lambda^4}, \qquad [17]$$

where $\alpha^\star$ is the fixed point of the limiting state evolution (cf. Eq. **15**). This together with the nonasymptotic results in Theorem 1 leads to a more refined risk characterization, as we shall prove in *SI Appendix*, section C.

Corollary 1. *With probability at least* $1 - O(n^{-10})$, *there exists some* $t = O(\frac{\log n}{\lambda-1})$ *such that*

$$\big\|v^\star v^{\star\top} - u_t u_t^\top\big\|_F^2 = 1 - \frac{\alpha^{\star 2}}{\lambda^4} + O\left(\sqrt{\frac{\log^4 n}{n(\lambda-1)^6}}\right). \qquad [18]$$

In words, it only takes the AMP algorithm at most $O(\frac{\log n}{\lambda-1})$ number of iterations to achieve—up to a discrepancy of $\widetilde{O}(\sqrt{\frac{1}{n(\lambda-1)^6}})$—the Bayes optimal risk.

Roadmap for the Proof of Theorem 1. To provide some intuition underlying Theorem 1, we briefly give an outline of the proof; details can be found in *SI Appendix*.

- First, focusing on the initial stage obeying $1 \leq t \leq \min\big\{\varsigma, \frac{\log n}{c(\lambda-1)}\big\}$ for some constant $c > 0$, we develop an upper bound on $\|\xi_t\|_2$ in *SI Appendix*, section A.2.1 as:

$$\|\xi_t\|_2 \lesssim \sqrt{\frac{t^3 \log n}{n}}; \qquad [19]$$

  here, ς is a threshold defined in Eq. **6**. This step, which is accomplished by means of an inductive argument, helps us justify the validity of the decomposition Eq. **8a** with small residual terms before the crossing time ς.
- Second, with the above decomposition Eq. **8a** in place, we can readily investigate (using the derived Gaussian approximation) how the signal strength α_t evolves during the execution of AMP (*SI Appendix*, section A.2.2). Crucially, recalling that ς reflects the first time t that satisfies $|\alpha_t| \gtrsim \sqrt{\lambda^2-1}$ (cf. Eq. **6**), we can use the dynamics of α_t demonstrate that

$$\varsigma \lesssim \frac{\log n}{\lambda-1}; \qquad [20]$$

  in words, in spite of random (and hence uninformative) initialization, it takes AMP at most $O(\frac{\log n}{\lambda-1})$ iterations to find an informative estimate.
- Third, with the above control of ς in place, we go on to develop a more complete upper bound on $\|\xi_t\|_2$ that covers the iterations after ς, that is,

$$\|\xi_t\|_2 \lesssim \sqrt{\frac{t\,\mathbb{1}(t>\varsigma)\log n}{n(\lambda-1)^2}} + \sqrt{\frac{\min\{t,\varsigma\}^3 \log n}{n}}, \qquad [21]$$

for any $t < \frac{cn(\lambda-1)^5}{\log^2 n}$. In order words, when the number of iterations grows larger that an order of $\frac{\log^3 n}{\lambda-1}$, the size of the residual scales as

$$\|\xi_t\|_2 \lesssim \sqrt{\frac{t\log n}{n(\lambda-1)^2}}.$$

This is the main content of *SI Appendix*, section A.2.3, accomplished again via an inductive argument.

- Finally, after the iteration number exceeds the threshold ς, we demonstrate in *SI Appendix*, section A.2.4 that the asymptotic state evolution (the one characterizing large-system limits) becomes fairly accurate in the finite-sample/finite-time regime. In particular, a connection is established between the nonasymptotic state evolution and its asymptotic analog, namely,

$$\begin{aligned}\frac{|\alpha_{t+1}^2-\alpha_{t+1}^{\star 2}|}{\alpha_{t+1}^{\star 2}} &\le \big(1-c(\lambda-1)\big)\cdot\frac{|\alpha_t^2-\alpha_t^{\star 2}|}{\alpha_t^{\star 2}} \\ &\quad + O\left(\sqrt{\frac{(t+\frac{\log^3 n}{\lambda-1})\log n}{n(\lambda-1)^3}}\right), \quad \text{for some } c>0,\end{aligned}$$

which plays a critical role in characterizing the finite-sample convergence behavior of AMP.

Comparisons to Li and Wei (50). While Li and Wei (50) provided a general decomposition for the AMP iterates $\{x_t\}$, the theory therein is far from sufficient when studying AMP from random initialization. A key reason is that during the initial stage of AMP, the signal component is vanishingly small and asymptotically vanishing compared to the magnitude of the residual. A direct application of ref. 50 leads to a vacuous upper bound on $\|\xi_t\|_2$ and does not reveal the effectiveness of random initialization. In contrast, the current paper focuses on showing that the signal component will undergo a rapid growth phase and reach a level comparable to the noise. A crucial step of our analysis is to prove that $\eta(x_t) \approx \alpha_t v^\star + \phi_{t-1}$ at the initial stage, by demonstrating that $\{\eta_t(x_t)\}$ are almost orthogonal to each other (see *SI Appendix*, section B.4 for more details). Based on this approximation, we then argue that it takes only $O(\log n)$ iterations for the signal strength to reach a nontrivial level. Once the signal strength has reached this level, we then proceed to uncover a new stage in which the signal strength starts to grow exponentially fast. Establishing all these phenomena requires fine-grained analyses about how AMP behaves in different stages, which was not achievable by existing analysis in ref. 50.

3. Discussions

In this paper, we have pinned down the finite-sample convergence behavior of AMP when initialized randomly, focusing on the prototypical $\mathbb{Z}_2$ synchronization problem. This algorithm has been shown to enjoy fast global convergence, as it takes no more than $O(\frac{\log n}{\lambda-1})$ iterations to arrive at a point whose risk is $O(\sqrt{\frac{\log^4 n}{n(\lambda-1)^6}})$ close to Bayes optimal. Our theory offers rigorous evidence supporting the effectiveness of randomly initialized AMP in low-rank matrix estimation. While the present paper concentrates on a specific choice of denoising functions tailored to $\mathbb{Z}_2$ synchronization, we expect our analysis framework to be generalizable to a broader family of separable and Lipschitz-continuous denoising functions.

Moving forward, there is no shortage of research directions worth exploring. One natural extension is concerned with other structural prior about $v^\star$; for instance, it would be interesting to see how randomly initialized AMP performs when $v^\star$ is known to satisfy general cone constraints (see e.g., refs. 59 and 60). Another direction of interest is to go beyond the spiked Gaussian Wigner model. A recent work along this line (61) studied the role of random initialization for power iteration in the problem of tensor decomposition, which leverages upon the AMP-type analysis for analyzing tensor power methods. Can we further extend these to understand (randomly initialized) AMP toward solving more challenging problems like low-rank matrix completion and tensor completion? Moreover, while AMP serves as a versatile machinery for understanding various statistical procedures in high dimensions, there are several alternative analysis frameworks like the convex Gaussian min-max theorem (CGMT) (62–64) and the leave-one-out analysis (2, 65, 66) that also prove effective and enjoy their own benefits. Is there any effective way to combine them so as to exploit all of their advantages at once? Finally, moving beyond $\mathbb{Z}_2$ synchronization, we believe that our nonasymptotic framework and the analysis ideas for understanding random initialization can both be extended to accommodate other important settings such as sparse linear regression and generalized linear models (GLMs). Take generalized approximate message passing (GAMP) for instance (27, 67), which can often be viewed as AMP applied to asymmetric matrix models. More specifically, given an asymmetric design matrix X, GAMP maintains two sequences of updates as follows

$$\begin{aligned}s_t &= XF_t(\beta_t) - \langle F_t'\rangle G_{t-1}(s_{t-1}),\\ \beta_{t+1} &= X^\top G_t(s_t) - \langle G_t'\rangle F_t(\beta_t),\end{aligned}$$

thus resembling the update rule considered in the current paper. One can then employ similar analysis ideas as in ref. 50, while in the meantime keeping track of two sets of orthogonal bases and two sequences of Gaussian random vectors. Once we are equipped with the nonasymptotic decomposition for each sequence, the role of random initialization can be understood via similar yet more complicated arguments as the ones provided in the current paper, given that these two sequences are intertwined and rely heavily on each other. We leave these questions for future investigation.

Data, Materials, and Software Availability. There are no data underlying this work.

ACKNOWLEDGMENTS. This work was partially supported by NSF grants DMS 2147546/2015447, the NSF CAREER award DMS-2143215, and the Google Research Scholar Award.

1. A. Singer, Angular synchronization by eigenvectors and semidefinite programming. *Appl. Comput. Harmonic Anal.* **30**, 20–36 (2011).
2. E. Abbe, J. Fan, K. Wang, Y. Zhong, Entrywise eigenvector analysis of random matrices with low expected rank. *Ann. Stat.* **48**, 1452 (2020).
3. R. Keshavan, A. Montanari, S. Oh, Matrix completion from noisy entries. *Adv. Neural Inf. Process. Syst.* **22** (2009).
4. A. Lemon, A. Man-Cho So, Y. Ye, Low-rank semidefinite programming: Theory and applications. *Found. Trends Opt.* **2**, 1–156 (2016).

5. E. J. Candes, Y. Plan, Matrix completion with noise. *Proc. IEEE* **98**, 925–936 (2010).
6. Y. Chi, Y. M. Lu, Y. Chen, Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Trans. Sig. Process.* **67**, 5239–5269 (2019).
7. A. Javanmard, A. Montanari, Localization from incomplete noisy distance measurements. *Found. Comput. Math.* **13**, 297–345 (2013).
8. S. Athey, M. Bayati, N. Doudchenko, G. Imbens, K. Khosravi, Matrix completion methods for causal panel data models. *J. Am. Stat. Assoc.* **116**, 1716–1730 (2021).
9. S. Péché, The largest eigenvalue of small rank perturbations of Hermitian random matrices. *Probability Theory Related Fields* **134**, 127–173 (2006).
10. J. Baik, G. B. Arous, S. Péché, Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Prob.* **33**, 1643–1697 (2005).
11. D. Féral, S. Péché, The largest eigenvalue of rank one deformation of large Wigner matrices. *Commun. Math. Phys.* **272**, 185–228 (2007).
12. M. Capitaine, C. Donati-Martin, D. Féral, The largest eigenvalues of finite rank deformation of large Wigner matrices: Convergence and nonuniversality of the fluctuations. *Ann. Probab.* **37**, 1–47 (2009).
13. C. Cheng, Y. Wei, Y. Chen, Tackling small eigen-gaps: Fine-grained eigenvector estimation and inference under heteroscedastic noise. *IEEE Trans. Inf. Theory* **67**, 7380–7419 (2021).
14. Y. Chen, Y. Chi, J. Fan, C. Ma, Spectral methods for data science: A statistical perspective. *Found. Trends Mach. Learn.* **14**, 566–806 (2021).
15. T. Tony Cai, A. Zhang, Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *Ann. Stat.* **46**, 60–89 (2018).
16. Y. Yan, Y. Chen, J. Fan, Inference for heteroskedastic PCA with missing data. arXiv [Preprint] (2021). https://arxiv.org/abs/2107.12365
17. A. Perry, A. S. Wein, A. S. Bandeira, A. Moitra, Message-passing algorithms for synchronization problems over compact groups. *Commun. Pure Appl. Math.* **71**, 2275–2322 (2018).
18. Y. Deshpande, A. Montanari, E. Richard, Cone-constrained principal component analysis. *Adv. Neural Inf. Process. Syst.* **27** (2014).
19. T. Lesieur, F. Krzakala, L. Zdeborová, Constrained low-rank matrix estimation: Phase transitions, approximate message passing and applications. *J. Stat. Mech. Theory Exp.* **2017**, 073403 (2017).
20. I. M. Johnstone, A. Y. Lu, On consistency and sparsity for principal components analysis in high dimensions. *J. Am. Stat. Assoc.* **104**, 682–693 (2009).
21. Q. Berthet, P. Rigollet, Optimal detection of sparse principal components in high dimension. *Ann. Stat.* **41**, 1780–1815 (2013).
22. O. Y. Feng, R. Venkataramanan, C. Rush, R. J. Samworth, A unifying tutorial on approximate message passing. *Found. Trends Mach. Learn.* **15**, 335–536 (2022).
23. D. L. Donoho, A. Maleki, A. Montanari, Message-passing algorithms for compressed sensing. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 18914–18919 (2009).
24. D. L. Donoho, A. Maleki, A. Montanari, "Message passing algorithms for compressed sensing: I. motivation and construction." in *2010 IEEE Information Theory Workshop on Information Theory* (ITW 2010, Cairo, IEEE, 2010), pp. 1–5.
25. M. Bayati, A. Montanari, The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Trans. Inf. Theory* **57**, 764–785 (2011).
26. Z. Fan, Approximate message passing algorithms for rotationally invariant matrices. *Ann. Stat.* **50**, 197–224 (2022).
27. S. Rangan, Generalized approximate message passing for estimation with random linear mixing" in *2011 IEEE International Symposium on Information Theory Proceedings (IEEE)* (2011), pp. 2168–2172.
28. M. Celentano, A. Montanari, Fundamental barriers to high-dimensional regression with convex penalties. *Ann. Stat.* **50**, 170–196 (2022).
29. M. Bayati, A. Montanari, The LASSO risk for Gaussian matrices. *IEEE Trans. Inf. Theory* **58**, 1997–2017 (2011).
30. Y. Li, Y. Wei, Minimum ell_1-norm interpolators: Precise asymptotics and multiple descent. arXiv [Preprint] (2021). https://arxiv.org/abs/2110.09502
31. P. Sur, Y. Chen, E. J. Candès, The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *Prob. Theory Related Fields* **175**, 487–558 (2019).
32. Y. Zhang, M. Mondelli, R. Venkataramanan, Precise asymptotics for spectral methods in mixed generalized linear models. arXiv [Preprint] (2022). http://arxiv.org/abs/2211.11368.
33. D. Donoho, A. Montanari, High dimensional robust m-estimation: Asymptotic variance via approximate message passing. *Probab. Theory Relat. Fields* **166**, 935–969 (2016).
34. J. Ma, J. Xu, A. Maleki, Optimization-based AMP for phase retrieval: The impact of initialization and ell_2-regularization. arXiv [Preprint] (2018). https://arxiv.org/abs/1801.01170.
35. M. Lelarge, Léo. Miolane,, Fundamental limits of symmetric low-rank matrix estimation. *Probab. Theory Relat. Fields* **173**, 859–929 (2019).
36. A. Javanmard, A. Montanari, F. Ricci-Tersenghi, Phase transitions in semidefinite relaxations. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E2218–E2223 (2016).
37. B. Zhiqi, J. M. Klusowski, C. Rush, W. J. Su, Algorithmic analysis and statistical estimation of SLOPE via approximate message passing. *IEEE Trans. Inf. Theory* **67**, 506–537 (2020).
38. P. Sur, E. J. Candès, A modern maximum-likelihood theory for high-dimensional logistic regression. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 14516–14525 (2019)
39. A. K. Fletcher, S. Rangan, Scalable inference for neuronal connectivity from calcium imaging. *Adv. Neural Inf. Process Syst.* **27** (2014).
40. C. Jeon, R. Ghods, A. Maleki, C. Studer, "Optimality of large mimo detection via approximate message passing" in *2015 IEEE International Symposium on Information Theory (ISIT) (IEEE)* (2015), pp. 1227–1231.
41. C. Rush, A. Greig, R. Venkataramanan, Capacity-achieving sparse superposition codes via approximate message passing decoding. *IEEE Trans. Inf. Theory* **63**, 1476–1500 (2017).
42. P. Pandit, M. Sahraee, S. Rangan, A. K. Fletcher, "Asymptotics of map inference in deep networks" in *2019 IEEE International Symposium on Information Theory (ISIT) (IEEE)* (2019), pp. 842–846.
43. J. Barbier, F. Krzakala, Approximate message-passing decoder and capacity achieving sparse superposition codes. *IEEE Trans. Inf. Theory* **63**, 4894–4927 (2017).
44. C. Ma, K. Wang, Y. Chi, Y. Chen, Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. *Found. Comut. Math.* **20**, 451–632 (2020).
45. A. Montanari, R. Venkataramanan, Estimation of low-rank matrices via approximate message passing. *Ann. Stat.* **49**, 321–345 (2021).
46. M. Celentano, Z. Fan, S. Mei, Local convexity of the TAP free energy and AMP convergence for Z2-synchronization. arXiv [Preprint] (2021). https://arxiv.org/abs/2106.11428.
47. X. Zhong, T. Wang, Z. Fan, Approximate message passing for orthogonally invariant ensembles: Multivariate non-linearities and spectral initialization. arXiv [Preprint] (2021). https://arxiv.org/abs/2110.02318.
48. C. Rush, R. Venkataramanan, Finite sample analysis of approximate message passing algorithms. *IEEE Trans. Inf. Theory* **64**, 7264–7286 (2018).
49. C. Cademartori, C. Rush, A non-asymptotic analysis of generalized approximate message passing algorithms with right rotationally invariant designs. arXiv [Preprint] (2023). https://arxiv.org/abs/2302.00088.
50. G. Li, Y. Wei, A non-asymptotic framework for approximate message passing in spiked models. arXiv [Preprint] (2022). https://arxiv.org/abs/2208.03313.
51. Y. Chen, E. J. Candès, The projected power method: An efficient algorithm for joint alignment from pairwise differences. *Commun. Pure Appl. Anal.* **71**, 1648–1714 (2018).
52. Y. Zhong, N. Boumal, Near-optimal bounds for phase synchronization. *SIAM J. Optim.* **28**, 989–1016 (2018).
53. C. Gao, A. Y. Zhang, SDP achieves exact minimax optimality in phase synchronization. *IEEE Trans. Inf. Theory* (2022).
54. Y. Deshpande, E. Abbe, A. Montanari, Asymptotic mutual information for the balanced binary stochastic block model. *Inf. Inference: J. IMA* **6**, 125–170 (2017).
55. Z. Fan, S. Mei, A. Montanari, TAP free energy, spin glasses and variational inference. *Ann. Probab.* **49**, 1–45 (2021).
56. M. Mondelli, R. Venkataramanan, PCA initialization for approximate message passing in rotationally invariant models. *Adv. Neural Inf. Process Syst.* **34**, 29616–29629 (2021).
57. Y. Chen, Y. Chi, J. Fan, C. Ma, Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *Math. Program.* **176**, 5–37 (2019).
58. R. Ge, C. Jin, Y. Zheng, "No spurious local minima in nonconvex low rank problems: A unified geometric analysis" in *International Conference on Machine Learning* (2017), pp. 1233–1242.
59. A. S. Bandeira, D. Kunisky, A. S. Wein, Computational hardness of certifying bounds on constrained PCA problems. arXiv [Preprint] (2019). http://arxiv.org/abs/1902.07324.
60. Y. Wei, M. J. Wainwright, A. Guntuboyina, The geometry of hypothesis testing over convex cones: Generalized likelihood ratio tests and minimax radii. *Ann. Stat.* **47**, 994–1024 (2019).
61. W. Yuchen, K. Zhou, Lower bounds for the convergence of tensor power iteration on random overcomplete models. arXiv [Preprint] (2022). http://arxiv.org/abs/2211.03827.
62. M. Celentano, A. Montanari, Y. Wei, The Lasso with general Gaussian designs with applications to hypothesis testing. arXiv [Preprint] (2020). http://arxiv.org/abs/2007.13716.
63. L. Miolane, A. Montanari, The distribution of the lasso: Uniform control over sparse balls and adaptive parameter tuning. *Ann. Stat.* **49**, 2313–2335 (2021).
64. C. Thrampoulidis, E. Abbasi, B. Hassibi, Precise error analysis of regularized *m*-estimators in high dimensions. *IEEE Trans. Inf. Theory* **64**, 5592–5628 (2018).
65. N. El Karoui, On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probab. Theory Relat. Fields* **170**, 95–175 (2018).
66. Y. Chen, Y. Chi, J. Fan, C. Ma, Y. Yan, Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. *SIAM J. Optim.* **30**, 3098–3121 (2020).
67. M. Mondelli, R. Venkataramanan, "Approximate message passing with spectral initialization for generalized linear models" in *International Conference on Artificial Intelligence and Statistics PMLR 2021* (2021), pp. 397–405.

Supporting Information: Approximate message passing from random initialization with applications to synchronization

Gen Li Wei Fan Yuting Wei

Department of Statistics and Data Science, the Wharton School
University of Pennsylvania, Philadelphia, PA

June 29, 2023

Abstract

This document presents the supplementary information of "*Approximate message passing from random initialization with applications to $\mathbb{Z}_2$ synchronization*" in [1]. In Section A, we lay out the main analysis ideas for the proof of Theorem 1, with proofs of corresponding lemmas and claims deferred to Section B. Finally, the proof for the asymptotic optimality of AMP is included in Section C.

Keywords: approximate message passing, random initialization, non-asymptotic analysis, spiked Wigner model, global convergence

Contents

A Proof of Theorem 1

In this section, we present the proof of our main result: Theorem 1. We find it helpful to introduce the following notation that helps streamline the presentation:

$$v_t \coloneqq \alpha_t v^\star + \sum_{k=1}^{t-1} \beta_{t-1}^k \phi_k. \tag{22}$$

A.1 Preliminaries

Before we embark on our proof of the main theorem, we collect a couple of useful results that shall be used frequently throughout this proof.

Concentration results. We first record several useful concentration results from [2]. Here and throughout, we let $|x|_{(i)}$ denote the magnitude of the i-th largest entry (in magnitude) of $x \in \mathbb{R}^n$.

Lemma 1. *Consider a collection of random vectors $\{\phi_k\}_{1\le k<t}$ in $\mathbb{R}^n$. Suppose that for each $1 \le k \le t-1 < n$, $\phi_k = (\phi_{k,1}, \phi_{k,2}, \ldots, \phi_{k,n})$ is i.i.d. drawn from $\mathcal{N}\left(0, \frac{1}{n} I_n\right)$. Consider the following set*

$$\mathcal{E}_s := \left\{ \{\phi_k\}_{k=1}^{t-1} : \max_{1\le k\le t-1} \|\phi_k\|_2 < 1 + C\sqrt{\frac{\log\frac{n}{\delta}}{n}} \right\} \bigcap \left\{ \{\phi_k\}_{k=1}^{t-1} : \sup_{a\in\mathcal{S}^{t-2}} \left\| \sum_{k=1}^{t-1} a_k \phi_k \right\|_2 < 1 + C\sqrt{\frac{t\log\frac{n}{\delta}}{n}} \right\}$$
$$\bigcap \left\{ \{\phi_k\}_{k=1}^{t-1} : \sup_{a\in\mathcal{S}^{t-2}} \sum_{i=1}^{s} \left| \sum_{k=1}^{t-1} a_k \phi_k \right|_{(i)}^2 < \frac{C(t+s)\log\frac{n}{\delta}}{n} \right\} \bigcap \left\{ \{\phi_k\}_{k=1}^{t-1} : \max_{1\le k<t, 1\le i\le n} |\phi_{k,i}| < C\sqrt{\frac{\log\frac{n}{\delta}}{n}} \right\},$$

and denote $\mathcal{E} := \bigcap_{s=1}^n \mathcal{E}_s$. Then there exists some large enough constant $C > 0$ such that, for every $\delta > 0$,

$$\mathbb{P}\big(\{\phi_k\}_{k=1}^{t-1} \in \mathcal{E} \big) \ge 1 - \delta.$$

In particular, by setting $\delta = n^{-11}$, we see that the following event happens with probability at least $1 - O(n^{-11})$:

$$\left| \max_{1\le k\le t-1} \|\phi_k\|_2 - 1 \right| \lesssim \sqrt{\frac{\log n}{n}}, \qquad \text{and} \quad \max_{1\le k\le t, 1\le i\le n} |\phi_{k,i}| \lesssim \sqrt{\frac{\log n}{n}}.$$

This lemma is a consequence of standard concentration of measure for Gaussian random vectors [3]; its proof can be found in [2, Section D.1.1] and is hence omitted for brevity.

Properties of η_t, π_t, and γ_t (cf. (3)). Next, we summarize several basic properties about the three sets of key quantities defined in (3). We begin by gathering several basic properties for our choices of π_t and γ_t defined in (3); the proof is deferred to Section B.1.

Lemma 2. *Suppose the decomposition (8a) is valid with $\|\xi_{t-1}\|_2 \lesssim 1$. With probability at least $1 - O(n^{-10})$, the following properties hold true:*

$$\frac{1}{\sqrt{n}}\pi_t = |\alpha_t| + O\left(\left(\|\xi_{t-1}\|_2 + \sqrt{\frac{t\log n}{n}} \right)^{1/2} \wedge \frac{1}{|\alpha_t|} \left(\|\xi_{t-1}\|_2 + \sqrt{\frac{t\log n}{n}} \right) \right); \tag{23a}$$

$$\gamma_t^{-2} = n \int \tanh^2\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x) \right) \varphi(\mathrm{d}x) + \pi_t^2 O\left(\|\xi_{t-1}\|_2 + \sqrt{\frac{t\log n}{n}} \right). \tag{23b}$$

Additionally, one has

$$\int \tanh^2\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x) \right) \varphi(\mathrm{d}x) = \frac{\pi_t^2}{n}(\alpha_t^2 + 1) + O\left(\frac{\pi_t^4}{n^2} \right), \tag{23c}$$

which in turn implies that

$$\gamma_t^{-2} = \pi_t^2 \left(\alpha_t^2 + 1 + O\left(\frac{\pi_t^2}{n} + \|\xi_{t-1}\|_2 + \sqrt{\frac{t\log n}{n}} \right) \right). \tag{23d}$$

Next, let us single out several properties about the quantity η_t in the lemma below, whose proof is postponed to Section B.2.

Lemma 3. *Consider any $1 \le t \le n$ and suppose that $\|\xi_{t-1}\|_2$ satisfies*

$$\|\xi_{t-1}\|_2 \lesssim \sqrt{\frac{t^3 \log n}{n}} \qquad \text{for all } t \lesssim \frac{\log n}{\lambda - 1}. \tag{24}$$

Then the following properties hold true with probability at least $1 - O(n^{-11})$:

- *If $t \lesssim \frac{n(\lambda-1)^4}{\log^2 n}$ and $\|\xi_{t-1}\|_2 \lesssim \frac{1}{\sqrt{\log n}}$, then any $x \in \mathbb{R}$ obeys*

$$|\eta_t(x)| \lesssim |x|, \qquad |\eta_t'(x)| \lesssim 1 =: \rho, \qquad |\eta_t''(x)| \lesssim \sqrt{n} =: \rho_1, \qquad |\eta_t^{(\prime\prime\prime)}(x)| \lesssim n =: \rho_2; \tag{25a}$$

- *If $t \lesssim \frac{\log n}{\lambda-1}$ and $\alpha_t \lesssim \sqrt{\lambda-1}\, n^{-0.1}$, then one has $\pi_t \lesssim \sqrt{\lambda-1}\, n^{0.4}$; and for any x obeying $|x| \lesssim \sqrt{\log n/n}$, we have*

$$\eta_t'(x) = 1 + O\big((\lambda-1)n^{-0.2}\log n\big) \qquad \text{and} \qquad |\eta_t''(x)| \lesssim (\lambda-1)n^{0.8}|x|; \tag{25b}$$

- *If $t \lesssim \frac{\log n}{\lambda-1}$ and $\alpha_t \lesssim \sqrt{\lambda-1}\, n^{-1/4}$, then one has $\pi_t \lesssim (\lambda-1)^{-3/4} n^{1/4} \log n$; for any x, one can find a quantity $c_0 \lesssim (\log^2 n)/\sqrt{n(\lambda-1)^3}$ independent from x, and another quantity $|c_x| \lesssim n|x|^5(\log^4 n)/(\lambda-1)^3$ depending on x, such that*

$$\eta_t(x) = (1 - c_0)\left(x - \frac{1}{3}\pi_t^2 x^3 + c_x\right). \tag{25c}$$

A.2 Non-asymptotic analysis for the AMP dynamics

We are now in a position to present the proof of our main theorem. Let us recap that the structure of our proof is outlined in what follows.

- Firstly, focusing on the initial stage obeying $t \le \varsigma \wedge \frac{\log n}{c(\lambda-1)}$, we develop an upper bound on $\|\xi_t\|_2$ in Section A.2.1 as follows:

$$\|\xi_t\|_2 \lesssim \sqrt{\frac{t^3 \log n}{n}}; \tag{26}$$

  here, ς is a threshold defined in (6) and $c > 0$ is some constant small enough. This is accomplished by means of an inductive argument.

- Secondly, we investigate in Section A.2.2 how the signal strength α_t evolves during the execution of AMP. Crucially, recalling that ς reflects the first time t that satisfies $|\alpha_t| \gtrsim \sqrt{\lambda^2 - 1}$ (cf. (6)), we demonstrate that

$$\varsigma \lesssim \frac{\log n}{\lambda - 1}; \tag{27}$$

  in words, in spite of random (and hence uninformative) initialization, it takes AMP at most $O\big(\frac{\log n}{\lambda-1}\big)$ iterations to find an informative estimate.

- Thirdly, with the above control of ς in place, we go on to develop a more complete upper bound on $\|\xi_t\|_2$ that covers the iterations after ς, that is,

$$\|\xi_t\|_2 \lesssim \sqrt{\frac{t\,\mathbb{1}(t > \varsigma)\log n}{n(\lambda-1)^2}} + \sqrt{\frac{\min\{t,\varsigma\}^3 \log n}{n}} \tag{28}$$

  for any $t < \frac{cn(\lambda-1)^5}{\log^2 n}$. This is the main content of Section A.2.3, accomplished again via an inductive argument.

- Finally, after the iteration number exceeds the threshold ς, we demonstrate in Section A.2.4 that the asymptotic state evolution (the one characterizing large-system limits) becomes fairly accurate in the finite-sample/finite-time regime. In particular, an intimate connection is established between the non-asymptotic state evolution and its asymptotic analog, which plays a critical role in characterizing the finite-sample convergence behavior of AMP.

These four steps will be explained in detail in the sequel.

A.2.1 Controlling ξ_t when $t \le \varsigma \wedge \frac{\log n}{c(\lambda-1)}$ (Proof of Claim (26))

In this subsection, we establish the claimed bound (26) for $\|\xi_t\|_2$, which leverages on ideas from [2]. To begin with, let us restate [2, Theorem 2] below, with slight simplification tailored to $\mathbb{Z}_2$ synchronization (i.e., through the use of the properties $\|\beta_t\|_2^2 = 1$, $E_t = 0$, and (25a)). For notational convenience, define $\kappa_t > 0$ such that

$$\kappa_t^2 := \max\left\{\left\langle \int \left[x\eta_t'\left(\alpha_t v^\star + \frac{1}{\sqrt{n}}x\right) - \frac{1}{\sqrt{n}}\eta_t''\left(\alpha_t v^\star + \frac{1}{\sqrt{n}}x\right)\right]^2 \varphi_n(\mathrm{d}x)\right\rangle, \left\langle \int \left[\eta_t'\left(\alpha_t v^\star + \frac{1}{\sqrt{n}}x\right)\right]^2 \varphi_n(\mathrm{d}x)\right\rangle\right\}, \tag{29}$$

where we recall that $\varphi_n(\cdot)$ is the p.d.f. of $\mathcal{N}(0, I_n)$ and for any vector $x = [x_i]_{1\le i\le n}$, we denote $\langle x\rangle := \frac{1}{n}\sum_{i=1}^n x_i$ and $x^2 = [x_i^2]_{1\le i\le n}$. We shall work with the following assumptions.

Assumption 1. *For any $1 \le t \le n$, consider arbitrary vectors $\mu_t \in \mathcal{S}^{t-1}$, $\xi_{t-1} \in \mathbb{R}^n$, and coefficients $(\alpha_t, \beta_{t-1}) \in \mathbb{R}\times\mathbb{R}^{t-1}$ that might all be statistically dependent on ϕ_k. Let v_t be defined as in (22). We assume the existence of (possibly random) quantities A_t, B_t, D_t such that with probability at least $1 - O(n^{-11})$, the following inequalities hold:*

$$\left|\sum_{k=1}^{t-1}\mu_t^k\left[\langle\phi_k, \eta_t(v_t)\rangle - \langle\eta_t'(v_t)\rangle\beta_{t-1}^k\right]\right| \le A_t, \tag{30a}$$

$$\left|v^{\star\top}\eta_t(v_t) - v^{\star\top}\int \eta_t\Big(\alpha_t v^\star + \frac{\|\beta_{t-1}\|_2}{\sqrt{n}}x\Big)\varphi_n(\mathrm{d}x)\right| \le B_t, \tag{30b}$$

$$\left\|\sum_{k=1}^{t-1}\mu_t^k\phi_k\circ\eta_t'(v_t) - \frac{1}{n}\sum_{k=1}^{t-1}\mu_t^k\beta_{t-1}^k\eta_t''(v_t)\right\|_2^2 - \kappa_t^2 \le D_t. \tag{30c}$$

Under these assumptions, [2, Theorem 2] developed a general non-asymptotic characterization for AMP iterates as follows.

Proposition 1. *[Adapted from [2, Theorem 2]] Suppose that Assumption 1 holds, and consider any $t \le n$. With probability at least $1 - O(n^{-11})$, the AMP iterates (2) for $\mathbb{Z}_2$-synchronization satisfy the decomposition (8a) with $\|\beta_t\|_2^2 = 1$ and*

$$\alpha_{t+1} = \lambda v^{\star\top}\int \eta_t\left(\alpha_t v^\star + \frac{1}{\sqrt{n}}x\right)\varphi_n(\mathrm{d}x) + \Delta_{\alpha,t} \tag{31}$$

where the residual terms obey

$$|\Delta_{\alpha,t}| \lesssim B_t + \left|v^{\star\top}\eta_t(x_t) - v^{\star\top}\eta_t(v_t)\right| \lesssim B_t + \|\xi_{t-1}\|_2, \tag{32a}$$

$$\|\xi_t\|_2 \le \sqrt{\kappa_t^2 + D_t}\,\|\xi_{t-1}\|_2 + O\left(\sqrt{\frac{t\log n}{n}} + A_t + \sqrt{(1 + t\,\mathbb{1}_{t\le\varsigma})\log n}\,\|\xi_{t-1}\|_2^2 + \sqrt{\frac{t\log n}{n}}\|\xi_{t-1}\|_2\right). \tag{32b}$$

Remark 1. With regards to the above bound (32b) for $\|\xi_t\|_2$, a direct application of [2, Theorem 2] results in a term $\sqrt{t\log n}\,\|\xi_{t-1}\|_2^2$ (as opposed to $\sqrt{(1 + t\,\mathbb{1}_{t\le\varsigma})\log n}\,\|\xi_{t-1}\|_2^2$ in (32b)). We make slight modifications here to make it better-suited for the current setting.

(i) When $t \le \varsigma$, such a term $\sqrt{t\log n}\,\|\xi_{t-1}\|_2^2$ works fine for our purpose;

(ii) When $t > \varsigma$ (so that α_t exceeds the order of $\sqrt{\lambda^2 - 1}$), one can simply invoke [2, display (249)] to improve the factor in front of $\|\xi_{t-1}\|_2^2$ from $\sqrt{t\log n}$ to $\sqrt{\log n}$.

Putting these together leads to the claimed bound (32b). Notably, this seemingly minor change turns out to be essential in order to push the number of iterations to $O(n/\mathsf{poly}(\log n))$ instead of $O(\sqrt{n}/\mathsf{poly}(\log n))$.

With Proposition 1 in mind, in order to control $|\Delta_{\alpha,t}|$ and $\|\xi_t\|_2$, it boils down to determining A_t, B_t, D_t, and κ_t, respectively.

- **Bounding A_t, B_t, D_t.** Repeating the same analysis as in [2, Section D.2], we obtain

$$A_t \lesssim \sqrt{\frac{t\log n}{n}}, \qquad B_t \lesssim \sqrt{\frac{t\log n}{n}}, \qquad D_t \lesssim \sqrt{\frac{t\log^2 n}{n}}. \tag{33}$$

  The only term that needs more discussion is A_t, as [2] only proved that $A_t \lesssim \frac{1}{\alpha_t}\sqrt{\frac{t\log n}{n}}$ (taking $s=1$ therein) for AMP with independent initialization. To get rid of the prefactor $1/\alpha_t$, we rely on an improved control of $\eta_t(x)$ (cf. (25a)). In particular, property (25a) tells us that

$$\|\eta_t(v_t)\|_2 \lesssim \|v_t\|_2 = \Big\|\alpha_t v^\star + \sum_{k=1}^{t-1}\beta_{t-1}^k \phi_k\Big\|_2 \lesssim 1,$$

  where the last inequality can be found in display (90). In turn, this leads to

$$\|\nabla_\Phi f_\theta(\Phi)\|_2 \lesssim \frac{1}{\sqrt{n}} \tag{34}$$

  through the same analyses as detailed around [2, Section D.2.1, inequality (229)]. Here, $\nabla_\Phi f_\theta(\Phi)$ is the key quantity to control A_t in [2, Section D.2.1], and our desired bound for A_t follows immediately. Given that this only consists of very minor and straightforward changes to [2, Section D.2.1], we omit the details for brevity and refer the readers to [2, Section D.2.1] for more details.

- **Bounding κ_t.** The main step then comes down to bounding κ_t. Towards this end, we claim that the following relation holds for κ_t, whose proof is postponed to Section B.3.

  Lemma 4. *With probability at least $1 - O(n^{-10})$, the following results hold true:*

  - *Under the inductive assumption (26) for ξ_{t-1}, one has*

$$\kappa_t \le 1 + o\Big(\frac{\lambda-1}{\log n}\Big) \tag{35a}$$

    provided that $t \le \varsigma \wedge \frac{\log n}{c(\lambda-1)}$;
  - *Under the inductive assumption (28) for ξ_{t-1}, one has*

$$\kappa_t \le 1 - \frac{1}{15}(\lambda-1), \tag{35b}$$

    provided that $t \le \frac{cn(\lambda-1)^5}{\log^2 n}$ and $|\alpha_t| \gtrsim \sqrt{\lambda^2-1}$.

With the above estimates of A_t, B_t, D_t, κ_t in place, we are ready to apply Theorem 1. Under the inductive assumption (26), the recursive formula (32) in Theorem 1 taken together with (33) yields

$$\|\xi_t\|_2 \le \sqrt{\kappa_t + \sqrt{\frac{t\log^2 n}{n}}}\|\xi_{t-1}\|_2 + O\bigg(\sqrt{\frac{t\log n}{n}} + \sqrt{(1+t\mathbb{1}_{t\le\varsigma})\log n}\,\|\xi_{t-1}\|_2^2 + \sqrt{\frac{t\log n}{n}}\|\xi_{t-1}\|_2\bigg), \tag{36}$$

which combined with (35) further implies that

$$\|\xi_t\|_2 \le \bigg(1 + o\Big(\frac{\lambda-1}{\log n}\Big) + O\Big(\sqrt{\frac{t^4\log^2 n}{n}}\Big)\bigg)\|\xi_{t-1}\|_2 + O\Big(\sqrt{\frac{t\log n}{n}}\Big), \tag{37}$$

with the proviso that $t \le \varsigma \wedge \frac{\log n}{c(\lambda-1)}$.

We are now ready to prove relation (26) via induction. To verify its validity for the base case (i.e. $t = 1$), we note that by construction (see, e.g. [2, Step 3, Proof of Theorem 1]), ξ_1 takes the form

$$\xi_1 = \Big(\frac{\sqrt{2}}{2} - 1\Big) z_1 z_1^\top W z_1, \quad \text{where } z_1 = \eta_1(x_1) \text{ is independent of } W.$$

Elementary calculations reveal that, with probability at least $1 - O(n^{-11})$,

$$\|\xi_1\|_2 = \Big|\frac{\sqrt{2}}{2} - 1\Big| \cdot \|z_1\|_2 \cdot |z_1^\top W z_1| \lesssim \sqrt{\frac{\log n}{n}}, \tag{38}$$

given that $\|z_1\|_2 = 1$ and $z_1^\top W z_1 \sim \mathcal{N}(0, \frac{2}{n} I_n)$. This already establishes (26) for the base case with $t = 1$. Next, consider the case where $t \le \varsigma \wedge \frac{\log n}{c(\lambda-1)}$ for some small enough constant $c > 0$. Given that $\sqrt{\frac{t^4 \log^2 n}{n}} = o(\frac{\lambda-1}{\log n})$ under our assumption on $\lambda - 1$, the recursive relation (37) immediately leads to

$$\begin{aligned}
\|\xi_t\|_2 &\le \Big(1 + o\Big(\frac{\lambda-1}{\log n}\Big)\Big) \|\xi_{t-1}\|_2 + O\Big(\sqrt{\frac{t \log n}{n}}\Big) \\
&\le \Big(1 + o\Big(\frac{\lambda-1}{\log n}\Big)\Big)^{t-1} \|\xi_1\|_2 + \sum_{j=0}^{t-2} \Big(1 + o\Big(\frac{\lambda-1}{\log n}\Big)\Big)^{j} O\Big(\sqrt{\frac{(t-j)\log n}{n}}\Big) \\
&\lesssim \sqrt{\frac{t^3 \log n}{n}}
\end{aligned} \tag{39}$$

for all $t \le \varsigma \wedge \frac{\log n}{c(\lambda-1)}$, as claimed.

Remark 2. Careful readers might note that the recursive formula established in (37) for $t \le \varsigma \wedge \frac{\log n}{c(\lambda-1)}$ does not rely on the relation (27) (a relation that shall be established in the next subsection).

A.2.2 Evolution of α_t and a bound on ς (Proof of Claim (27))

We now move on to establish the claim (27) concerning an upper bound on the threshold ς, which requires careful analysis about how the signal strength α_t evolves at the initial stage. Towards this end, we divide into two cases based on the magnitude of α_t, which we shall detail after presenting several preliminary facts.

Preliminary facts. Before proceeding, we first recall some additional preliminary facts already established in [2]. From the analysis of [2, Theorem 1], we know that: by construction,

$$\xi_t \in \mathsf{span}(U_{t-1}) = \mathsf{span}\Big\{\eta_1(x_1), \ldots, \eta_{t-1}(x_{t-1})\Big\}, \tag{40}$$

where $U_{t-1} \in \mathbb{R}^{n \times (t-1)}$ is a matrix whose columns are formed by a set of orthonormal basis $\{z_1, \ldots, z_{t-1}\}$. In fact, we can specify U_t in a more explicit manner. Following [2, Section 4.1], let us define

$$z_1 \coloneqq \frac{\eta_1(x_1)}{\|\eta_1(x_1)\|_2} \in \mathbb{R}^n \qquad \text{and} \qquad W_1 \coloneqq W \in \mathbb{R}^{n \times n}, \tag{41a}$$

which are statistically independent from each other; and then any $2 \le t \le n$, we can define the following objects recursively:

$$U_{t-1} \coloneqq [z_k]_{1 \le k \le t-1} \in \mathbb{R}^{n \times (t-1)}, \tag{41b}$$

and also

$$z_t \coloneqq \frac{\big(I_n - U_{t-1}U_{t-1}^\top\big)\, \eta_t(x_t)}{\big\|\big(I_n - U_{t-1}U_{t-1}^\top\big)\, \eta_t(x_t)\big\|_2}, \tag{41c}$$

$$W_t \coloneqq \big(I_n - z_{t-1}z_{t-1}^\top\big)\, W_{t-1}\, \big(I_n - z_{t-1}z_{t-1}^\top\big), \tag{41d}$$

where $\{x_t\}$ is the sequence generated by the AMP updates (2). This process thus leads to more explicit forms for $\{U_t\}$ and the orthonormal basis $\{z_t\}$ (see [2, Section 4.1] for the orthonormality of $\{z_t\}$). What is more, the orthonormality of $\{z_t\}$ reveals the decomposition

$$\eta_t(x_t) = \sum_{k=1}^{t} \beta_t^k z_k, \qquad \text{with } \beta_t^k := \langle \eta_t(x_t),\, z_k \rangle, \tag{42}$$

which satisfies $\|\eta_t(x_t)\|_2 = \|\beta_t\|_2$ with $\beta_t = [\beta_t^1, \ldots, \beta_t^t]$. Additionally, we find it convenient to generate

$$\phi_k \coloneqq W_k z_k + \zeta_k, \qquad \text{where } \zeta_k \coloneqq \Big(\frac{\sqrt{2}}{2} - 1\Big) z_k^\top W_k z_k \cdot z_k + \sum_{i=1}^{k-1} g_i^k z_i, \qquad 1 \le k \le n, \tag{43}$$

where the g_i^k's are independently drawn from $\mathcal{N}(0, \frac{1}{n})$. The following properties have been shown in [2, Lemma 2], which play a crucial role in our subsequent analysis:

- $\phi_k \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \frac{1}{n} I_n)$, for $1 \le k \le n$;
- The randomness of ϕ_k only comes from W_k, and ϕ_k is independent of x_1 and $\{z_i\}_{i<k}$.
- x_k and z_k are conditionally independent from W_k given $\{z_i\}_{i<k}$ and x_1.
- ϕ_k is independent from $\{x_j\}_{j\le k}$ and $\{z_j\}_{j\le k}$.

Stage I: small correlation ($|\alpha_t| \lesssim \sqrt{\lambda-1}\, n^{-1/4}$). Let us start from the very beginning when the correlation coefficient α_t is reasonably small. Towards this, we define a threshold τ_0 such that

$$\tau_0 := \max\big\{\tau : |\alpha_t| \lesssim \sqrt{\lambda-1}\, n^{-1/4} \text{ for all } t \le \tau\big\}; \tag{44}$$

in words, $\tau_0 + 1$ represents the first term that exceeds the level of $\sqrt{\lambda-1}\, n^{-1/4}$. In the following, we would like to prove that, with probability at least $1 - O(n^{-10})$, this threshold is not too large in the sense that

$$\tau_0 \lesssim \frac{\log n}{\lambda - 1}. \tag{45}$$

Proof of Claim (45). In order to establish this result (45), we first state an important claim: the AMP iterates — when initialized at a random point — satisfy the following recursive relation with high probability:

$$\alpha_{t+1} = \lambda^{t-k+1}\alpha_k + \sum_{i=1}^{t-k+1} \lambda^i g_{t-i} + O\left(\lambda^{t-k} \frac{\log^4 n}{n^{3/4}(\lambda-1)^{1.5}}\right) \tag{46}$$

for any $1 \le k \le t$, where we denote

$$g_k \coloneqq v^{\star\top} \phi_k \quad (1 \le k \le t) \qquad \text{and} \qquad g_0 = 0. \tag{47}$$

This claimed relation lies at the heart of the analysis for Stage I, in which the correlation between the AMP iterate and $v^\star$ keeps growing to a non-trivial value. To streamline the presentation, we defer the proof of this claim to Section B.4.

Equipped with the above recursive formula (46), we now turn to proving the relation (45). Define $t_i := C' i \log n$ for some quantity $C' = \frac{C''}{\lambda-1}$, where C'' is some large enough constant. Observe that

$$\begin{aligned}
\mathbb{P}\Big(|\alpha_k| \lesssim \frac{\sqrt{\lambda-1}}{n^{1/4}},\ \text{for all } k \le 201 C' \log n\Big) &\le \mathbb{P}\Big(|\alpha_{t_i+1}| \lesssim \frac{\sqrt{\lambda-1}}{n^{1/4}},\ \text{for all } 1 \le i \le 200\Big) \\
&= \prod_{i=1}^{200} \mathbb{P}\Big(|\alpha_{t_i+1}| \lesssim \frac{\sqrt{\lambda-1}}{n^{1/4}} \;\Big|\; |\alpha_{t_j+1}| \lesssim \frac{\sqrt{\lambda-1}}{n^{1/4}},\ \forall 1 \le j < i\Big).
\end{aligned}$$

To control the right-hand side of the above relation, consider the following random variable

$$X_i := \sum_{j=1}^{C' \log n} \lambda^j g_{t_i - j} \sim \mathcal{N}\left(0, \frac{\lambda^{2C' \log n + 2} - \lambda^2}{n(\lambda^2 - 1)}\right) \tag{48}$$

for each $1 \le i \le 200$. Armed with this piece of notation, invoking relation (46) gives

$$\alpha_{t_i+1} = \lambda^{t_i - t_{i-1}} \alpha_{t_{i-1}+1} + X_i + O\left(\lambda^{t_i - t_{i-1} - 1} \frac{\log^4 n}{n^{3/4}(\lambda - 1)^{1.5}}\right). \tag{49}$$

As mentioned in the above preliminary facts, each ϕ_j is independent with the AMP iterate x_i for $i \le j$ and therefore α_{i+1}, given that $\alpha_{i+1} = v^{\star\top} \eta_i(x_i)$. As a result, the random variable X_i defined above is independent from α_{t_j+1} for all $j \le i-1$. Taking this together with the relation (46) then leads to

$$\begin{aligned}
&\mathbb{P}\Big(|\alpha_{t_i+1}| \lesssim \frac{\sqrt{\lambda-1}}{n^{1/4}} \;\Big|\; |\alpha_{t_j+1}| \lesssim \frac{\sqrt{\lambda-1}}{n^{1/4}}, 1 \le j < i\Big) \\
&\le \mathbb{P}\Big(|\lambda^{t_i - t_{i-1}} \alpha_{t_{i-1}+1} + X_i| \lesssim \frac{\sqrt{\lambda-1}}{n^{1/4}} + \lambda^{t_i - t_{i-1} - 1} \frac{\log^4 n}{n^{3/4}(\lambda-1)^{1.5}} \;\Big|\; |\alpha_{t_j+1}| \lesssim \frac{\sqrt{\lambda-1}}{n^{1/4}}, 1 \le j < i\Big) \\
&= \mathbb{P}\Big(|\lambda^{t_i - t_{i-1}} \alpha_{t_{i-1}+1} + X_i| \lesssim \frac{\sqrt{\lambda-1}}{n^{1/4}} + \lambda^{C' \log n} \frac{\log^4 n}{n^{3/4}(\lambda-1)^{1.5}} \;\Big|\; |\alpha_{t_{i-1}+1}| \lesssim \frac{\sqrt{\lambda-1}}{n^{1/4}}\Big) \\
&\le \mathbb{P}\Big(|X_i| \lesssim \frac{\sqrt{\lambda-1}}{n^{1/4}} + \lambda^{C' \log n} \frac{\log^4 n}{n^{3/4}(\lambda-1)^{1.5}}\Big).
\end{aligned} \tag{50}$$

Here, the penultimate line follows from the independence relation stated above, whereas the last line follows from the elementary fact that

$$\mathbb{P}_{X \sim \mathcal{N}(\mu, \sigma^2)}(|X| < x) \le \mathbb{P}_{X \sim \mathcal{N}(0, \sigma^2)}(|X| < x), \qquad \forall x > 0.$$

Putting these pieces together, we conclude that

$$\begin{aligned}
\mathbb{P}\Big(|\alpha_k| \lesssim \frac{\sqrt{\lambda-1}}{n^{1/4}}, \text{ for all } k \le 201 C' \log n\Big) &\le \prod_{i=1}^{200} \mathbb{P}\left(|X_i| \lesssim \frac{\sqrt{\lambda-1}}{n^{1/4}} + \frac{\lambda^{C' \log n} \log^4 n}{n^{3/4}(\lambda-1)^{1.5}}\right) \\
&= \prod_{i=1}^{200} \mathbb{P}\left(\sqrt{\frac{n(\lambda^2-1)}{\lambda^{2C' \log n + 2} - \lambda^2}} |X_i| \lesssim \frac{(\lambda-1) n^{1/4}}{\lambda^{C' \log n}} + \frac{\log^4 n}{n^{1/4}(\lambda-1)}\right) \\
&\lesssim \Big(\frac{\log^4 n}{n^{1/4}(\lambda-1)}\Big)^{200} \lesssim n^{-11},
\end{aligned} \tag{51}$$

where we invoke the distribution of X_i in expression (48), and the last inequality results from the assumption that $\lambda - 1 \gtrsim n^{-1/9}$. Therefore, the above inequality guarantees that with probability at least $1 - O(n^{-10})$, there exists some $k \lesssim \frac{\log n}{\lambda - 1}$ such that

$$|\alpha_k| \gtrsim \sqrt{\lambda - 1}\, n^{-1/4}. \tag{52}$$

It thus implies that $\tau_0 \lesssim \frac{\log n}{\lambda - 1}$ (see the definition (44)), as claimed in (45). In other words, after at most $O(\frac{\log n}{\lambda - 1})$ iterations, $|\alpha_t|$ shall surpass the order of $\sqrt{\lambda - 1}\, n^{-1/4}$. □

Stage II: moderate-to-large correlation ($\sqrt{\lambda - 1}\, n^{-1/4} \lesssim |\alpha_t| \le \frac{1}{2}\sqrt{\lambda^2 - 1}$). Next, let us look at the time interval after $|\alpha_t|$ surpasses the level of $\sqrt{\lambda - 1}\, n^{-1/4}$ but before it reaches the level of $\frac{1}{2}\sqrt{\lambda^2 - 1}$. Mathematically, this refers to the interval $(\tau_0, \varsigma]$, where τ_0 and ς are defined in (44) and (6), respectively. In fact, we shall start by examining

$$t \in \Big(\tau_0,\, \varsigma \wedge \frac{c_5 \log n}{\lambda - 1}\Big] \tag{53}$$

for some constant $c_5 > 0$; we shall demonstrate that $\varsigma \lesssim \frac{\log n}{\lambda - 1}$ shortly.

In view of Theorem 1 and the bounds (33), we can write

$$\alpha_{t+1} = \lambda v^{\star\top} \int \eta_t \left(\alpha_t v^\star + \frac{1}{\sqrt{n}} x \right) \varphi_n(\mathrm{d}x) + \Delta_{\alpha,t}, \tag{54}$$

where the residual term obeys

$$|\Delta_{\alpha,t}| \lesssim \sqrt{\frac{t \log n}{n}} + \left| v^{\star\top} \eta_t(x_t) - v^{\star\top} \eta_t(v_t) \right|. \tag{55}$$

We first make a claim concerned with a refined recursive relation for α_{t+1}:

$$|\alpha_{t+1}| \geq \frac{\lambda |\alpha_t|}{\sqrt{\alpha_t^2 + 1}} + o\big((\lambda - 1)|\alpha_t|\big) + O(|\Delta_{\alpha,t}|); \tag{56}$$

the proof of this result is postponed to Section B.5. Observe that whenever $\alpha_t < \frac{1}{2}\sqrt{\lambda^2 - 1}$, it holds that

$$\left(\frac{\lambda}{\sqrt{1 + \alpha_t^2}} \right)^2 \geq \frac{\lambda^2}{\frac{1}{4}\lambda^2 + \frac{3}{4}} > 1 + \frac{1}{3}(\lambda - 1), \quad \text{for } \lambda \in (1, 1.2], \tag{57}$$

which when taken together with expression (56), implies

$$|\alpha_{t+1}| \geq \left(\sqrt{1 + \frac{1}{3}(\lambda - 1)} + o(\lambda - 1) \right) |\alpha_t| + O(|\Delta_{\alpha,t}|). \tag{58}$$

In addition, we claim that for every $t \leq \tau'$ where $\tau' \coloneqq \min\{t : \alpha_t \geq (\lambda - 1)^{-3/4} n^{-1/4}\}$, it satisfies

$$|\Delta_{\alpha,t}| \ll (\lambda - 1)|\alpha_t|, \tag{59}$$

which we shall establish in Section B.6. With the relations (58) and (59) in place, it obeys $|\alpha_{\tau'+1}| \geq |\alpha_{\tau'}|$. Moreover, observe that the bound (55) taken together with (39) ensures that

$$|\Delta_{\alpha,\tau'+1}| \lesssim \sqrt{\frac{(\tau' + 1)^3 \log n}{n}} \ll (\lambda - 1)|\alpha_{\tau'+1}|. \tag{60}$$

Invoking this argument recursively, we thus arrive at

$$|\alpha_{t+1}| \geq \left(\sqrt{1 + \frac{1}{3}(\lambda - 1)} + o(\lambda - 1) \right) |\alpha_t|.$$

Now taking the above recursive relation collectively with the assumption $\lambda - 1 \gtrsim n^{-1/9}$ reveals that $|\alpha_t|$ surpasses $\frac{1}{2}\sqrt{\lambda^2 - 1}$ within at most $O\big(\frac{\log n}{\lambda - 1}\big)$ iterations. Therefore, recalling our definition (6) of ς, we can readily conclude that

$$\varsigma = O\left(\frac{\log n}{\lambda - 1} \right). \tag{61}$$

A.2.3 A more complete bound for ξ_t (Proof of Claim (28))

We now move on to establish claim (28) for any t obeying $\varsigma < t < \frac{cn(\lambda - 1)^5}{\log^2 n}$ (recall from (61) that $\varsigma = O\big(\frac{\log n}{\lambda - 1}\big)$), again via an inductive argument. Along the way, we also need to demonstrate that

$$|\alpha_t| \geq \frac{1}{2}\sqrt{\lambda^2 - 1} \tag{62}$$

within this stage (namely, once $|\alpha_t|$ exceeds $\frac{1}{2}\sqrt{\lambda^2 - 1}$, the signal strength will never fall below this level).

To begin with, the claim (28) for the base case $t = \varsigma$ has already been validated in expression (39); the condition (62) also holds with high probability when $t = \varsigma$. Next, assuming that the claim (28)) holds up till iteration $t-1$, we would like to establish its validity for time t. Towards this end, inequality (36) together with Lemma 4 tells us that

$$\begin{aligned}
\|\xi_t\|_2 &\le \sqrt{\kappa_t + \sqrt{\frac{t\log^2 n}{n}}}\|\xi_{t-1}\|_2 + O\left(\sqrt{\frac{t\log n}{n}} + \sqrt{(1+t\,\mathbb{1}_{t\le\varsigma})\log n}\,\|\xi_{t-1}\|_2^2 + \sqrt{\frac{t\log n}{n}}\|\xi_{t-1}\|_2\right) \\
&\le \sqrt{1 - \frac{1}{15}(\lambda-1) + \sqrt{\frac{t\log^2 n}{n}}}\|\xi_{t-1}\|_2 + O\left(\sqrt{\frac{t\log n}{n}}\right) \\
&\quad + O\left(\sqrt{(1+t\,\mathbb{1}_{t\le\varsigma})\log n}\left(\sqrt{\frac{t\,\mathbb{1}_{t>\varsigma}\log n}{n(\lambda-1)^2}} + \sqrt{\frac{\min\{t,\varsigma\}^3\log n}{n}}\right) + \sqrt{\frac{t\log n}{n}}\right)\|\xi_{t-1}\|_2 \\
&\le \left(1 - \frac{1}{15}(\lambda-1) + O\left(\sqrt{\frac{\left(\frac{t}{(\lambda-1)^2} + \varsigma^3\right)\log^2 n}{n}}\right)\right)\|\xi_{t-1}\|_2 + O\Big(\sqrt{\frac{t\log n}{n}}\Big),
\end{aligned} \tag{63}$$

where the second line comes from (35b) and the induction hypothesis (28) for $t-1$. In addition, the validity of (62) for α_{t+1} in the $(t+1)$-th iteration can be justified as well, which we shall detail in Section B.7.

With the above recursive relation in mind, recognizing $\sqrt{\frac{(t/(\lambda-1)^2+\varsigma^3)\log^2 n}{n}} \le 2c(\lambda-1)$ for some constant $c > 0$ small enough, we can readily derive

$$\begin{aligned}
\|\xi_t\|_2 &\le \left(1 - \frac{1}{20}(\lambda-1)\right)\|\xi_{t-1}\|_2 + O\Big(\sqrt{\frac{t\log n}{n}}\Big) \\
&\le \left(1 - \frac{1}{20}(\lambda-1)\right)^{t-\varsigma}\|\xi_\varsigma\|_2 + \sum_{j=0}^{t-\varsigma-1}\left(1 - \frac{1}{20}(\lambda-1)\right)^j O\Big(\sqrt{\frac{(t-j)\log n}{n}}\Big) \\
&\lesssim \sqrt{\frac{\varsigma^3\log n}{n}} + \sqrt{\frac{t\log n}{n(\lambda-1)^2}}
\end{aligned} \tag{64}$$

for all $\varsigma \le t \le \frac{cn(\lambda-1)^5}{\log^2 n}$. Combining this with the bound (39) (for $t \le \varsigma$) immediately establishes Claim (28) for all $t \le \frac{cn(\lambda-1)^5}{\log^2 n}$.

A.2.4 Analysis for approximate state evolution (Proof of Property (10))

Once the signal strength α_t reaches the order of $\sqrt{\lambda^2-1}$, AMP enters the stage of local refinement. According to (28) and (27), for any $t \le \frac{cn(\lambda-1)^5}{\log^2 n}$, the AMP iterate x_t admits the decomposition (8a) with the error term bounded by

$$\|\xi_t\|_2 \lesssim \sqrt{\frac{(t + \frac{\log^3 n}{\lambda-1})\log n}{n(\lambda-1)^2}}. \tag{65}$$

In the meantime, to describe how α_t evolves, we bound $\Delta_{\alpha,t}$ based on the relation (32a) as follows:

$$|\Delta_{\alpha,t}| \lesssim B_t + \|\xi_{t-1}\|_2 \lesssim \sqrt{\frac{(t + \frac{\log^3 n}{\lambda-1})\log n}{n(\lambda-1)^2}}, \tag{66}$$

where the last inequality arises from (33). Combining this with relation (31) leads to

$$\alpha_{t+1} = \lambda v^{\star\top}\int \eta_t\left(\alpha_t v^\star + \frac{1}{\sqrt{n}}x\right)\varphi_n(\mathrm{d}x) + O\left(\sqrt{\frac{(t + \frac{\log^3 n}{\lambda-1})\log n}{n(\lambda-1)^2}}\right).$$

Next, we shall characterize the distance between α_{t+1} and its asymptotic counterpart to further understand the evolution of α_{t+1}. More specifically, recall that the asymptotic state evolution is defined as

$$\alpha^\star_{t+1} = \lambda \left[\int \tanh\left(\alpha^\star_t \left(\alpha^\star_t + x\right)\right) \varphi(\mathrm{d}x) \right]^{1/2}, \tag{67}$$

assuming we start from $\alpha^\star_\varsigma = |\alpha_\varsigma|$ for some $\varsigma = O\big(\frac{\log n}{\lambda - 1}\big)$. We aim to control the difference between α_{t+1} and $\alpha^\star_{t+1}$. To simplify the presentation, we assume without loss of generality that $\alpha_t > 0$, and employ the notation

$$\tau_t \coloneqq \big(\alpha^\star_t\big)^2.$$

To begin with, the same analysis as in the proof of claim (56) (with different error bound (66) here) gives

$$\begin{aligned}
\alpha_{t+1} &= \lambda \left[\int \tanh\left(\alpha_t^2 + \alpha_t x\right) \varphi(\mathrm{d}x) \right]^{1/2} + O\left(\left| \frac{\pi_t^2}{\alpha_t^2 n} - 1 \right| \alpha_t^3 + \sqrt{\frac{(t + \frac{\log^3 n}{\lambda - 1}) \log n}{n(\lambda-1)^2}} \right) \\
&= \lambda \left[\int \tanh\left(\alpha_t^2 + \alpha_t x\right) \varphi(\mathrm{d}x) \right]^{1/2} + O\left(\sqrt{\frac{(t + \frac{\log^3 n}{\lambda - 1}) \log n}{n(\lambda-1)^2}} \right).
\end{aligned} \tag{68}$$

Here, the last line follows from inequality (23a) which indicates

$$\left(\frac{\pi_t}{\sqrt{n}}\right)^2 = \alpha_t^2 + O\left(\|\xi_{t-1}\|_2 + \sqrt{\frac{t \log n}{n}} \right) = \alpha_t^2 + O\left(\sqrt{\frac{(t + \frac{\log^3 n}{\lambda - 1}) \log n}{n(\lambda-1)^2}} \right).$$

It then follows from relations (68) and (67) that

$$\frac{\alpha_{t+1}^2 - \tau_{t+1}}{\tau_{t+1}} = \frac{\int \left[\tanh\left(\alpha_t^2 + \alpha_t x\right) - \tanh\left(\tau_t + \sqrt{\tau_t} x\right)\right] \varphi(\mathrm{d}x)}{\int \tanh\left(\tau_t + \sqrt{\tau_t} x\right) \varphi(\mathrm{d}x)} + O\left(\sqrt{\frac{(t + \frac{\log^3 n}{\lambda - 1}) \log n}{n(\lambda-1)^3}} \right). \tag{69}$$

Here, we remind the readers that (see also (116) and [4, Appendix B.2])

$$\int \tanh\left(\alpha^2 + \alpha x\right) \varphi(\mathrm{d}x) = \int \tanh^2\left(\alpha^2 + \alpha x\right) \varphi(\mathrm{d}x) \asymp \alpha^2, \qquad \text{for } \alpha \in (0, \lambda],$$

where the last inequality results from relation (116). The recursive formula (69) quantifies how the difference between α_t and $\alpha^\star_t$ changes over time, which plays a key role in our following analysis.

In order to better understand the above recursion, let us define — for every $\tau \in [0, \lambda^2]$ — that

$$h(\tau) \coloneqq \int \tanh\left(\tau + \sqrt{\tau} x\right) \varphi(\mathrm{d}x).$$

Armed with this function, one can write

$$\alpha^{\star 2}_{t+1} = \lambda^2 h(\alpha^{\star 2}_t). \tag{70}$$

Also, direct calculations yield

$$h'(\tau) \coloneqq \int \left(1 + \frac{x}{2\sqrt{\tau}}\right) \left(1 - \tanh^2\left(\tau + \sqrt{\tau} x\right)\right) \varphi(\mathrm{d}x) \in (0, 1),$$

where its range follows from display (263) in [2, Section D.3.3]. We make note of a few direct consequences of the above results.

- Recognizing that $h'(\tau) > 0$, one has $\alpha^\star_{t+1} > \alpha^\star_t \gtrsim \sqrt{\lambda^2 - 1}$ for $t \geq \varsigma$.

- In view of display (264) in [2, Section D.3.3], we have $0 \le \lambda^2 h'(\tau) \le 1-(\lambda-1)$. If we define $\alpha^\star$ to be the limiting point of (67) (as $t \to \infty$)), we can then see that

$$|\alpha_{t+1}^{\star 2} - \alpha^{\star 2}| \le \big(1-(\lambda-1)\big) \cdot |\alpha_t^{\star 2} - \alpha^{\star 2}|, \qquad \text{for } t \ge \varsigma. \tag{71}$$

  In other words, the asymptotic state evolution parameter $\alpha_t^{\star 2}$ converges exponentially to some fixed point $\alpha^{\star 2}$.

- In light of the above notation, we can also write

$$\begin{aligned}\frac{|\alpha_{t+1}^2 - \tau_{t+1}|}{\tau_{t+1}} &= \frac{|h(\alpha_t^2) - h(\tau_t)|}{h(\tau_t)} + O\left(\sqrt{\frac{(t+\frac{\log^3 n}{\lambda-1})\log n}{n(\lambda-1)^3}}\right) \\ &= \frac{h'(\tau)}{h(\tau_t)/\tau_t} \cdot \frac{|\alpha_t^2 - \tau_t|}{\tau_t} + O\left(\sqrt{\frac{(t+\frac{\log^3 n}{\lambda-1})\log n}{n(\lambda-1)^3}}\right)\end{aligned} \tag{72}$$

  for some τ satisfying $\min\{\tau_t, \alpha_t^2\} \le \tau \le \max\{\tau_t, \alpha_t^2\}$.

We first prove that $\alpha_t^2 = (1+o(1))\tau_t$. By definition, $\alpha_\varsigma^\star = \alpha_\varsigma \gtrsim \sqrt{\lambda^2-1}$, and hence this claim holds trivially for $t = \varsigma$. Next, assuming the validity of the inductive assumption $\alpha_t^2 = (1+o(1))\tau_t$, we would like to prove it for the $(t+1)$-th step. Towards this end, we first claim that there exists some universal constant $c > 0$ small enough such that

$$\frac{h'(\tau)}{h(\tau_t)/\tau_t} \le 1 - c(\lambda-1), \tag{73}$$

whose proof is postponed to Section B.9. This further allows us to derive

$$\frac{|\alpha_t^2 - \tau_t|}{\tau_t} \lesssim \sqrt{\frac{(t+\frac{\log^3 n}{\lambda-1})\log n}{n(\lambda-1)^5}}, \tag{74}$$

which we shall demonstrate via an inductive argument. Given that (74) holds trivially when $t = \varsigma$, we intend to establish (74) for the $(t+1)$-th iteration, assuming that it holds for all $s \le t$. To do so, we combine (73) and (72) to show that

$$\begin{aligned}\frac{|\alpha_{t+1}^2 - \tau_{t+1}|}{\tau_{t+1}} &\le \big(1-c(\lambda-1)\big) \cdot \frac{|\alpha_t^2 - \tau_t|}{\tau_t} + O\left(\sqrt{\frac{(t+\frac{\log^3 n}{\lambda-1})\log n}{n(\lambda-1)^3}}\right) \\ &= \big(1-c(\lambda-1)\big)^{t+1-\varsigma} \cdot \frac{|\alpha_\varsigma^2 - \tau_\varsigma|}{\tau_\varsigma} + O\left(\sum_{k=\varsigma}^{t-1} \big(1-c(\lambda-1)\big)^{t-k} \sqrt{\frac{(k+\frac{\log^3 n}{\lambda-1})\log n}{n(\lambda-1)^3}}\right) \\ &\le \sum_{k=\varsigma}^{t-1} \big(1-c(\lambda-1)\big)^{t-k} \cdot O\left(\sqrt{\frac{(t+\frac{\log^3 n}{\lambda-1})\log n}{n(\lambda-1)^3}}\right) \\ &\lesssim \sqrt{\frac{(t+\frac{\log^3 n}{\lambda-1})\log n}{n(\lambda-1)^5}},\end{aligned} \tag{75}$$

where the second line is obtained by applying (74) recursively. Hence, it also leads to $\alpha_{t+1}^2 = (1+o(1))\tau_{t+1}$, which concludes the inductive assumption for the $(t+1)$-th iteration. Putting the above results together with expression (71) also gives

$$\begin{aligned}|\alpha_{t+1}^2 - \alpha^{\star 2}| &= \big(1-(\lambda-1)\big)^{t-\varsigma} \cdot |\alpha_\varsigma^{\star 2} - \alpha^{\star 2}| + \alpha_{t+1}^{\star 2} O\left(\sqrt{\frac{(t+\frac{\log^3 n}{\lambda-1})\log n}{n(\lambda-1)^5}}\right) \\ &\lesssim \big(1-(\lambda-1)\big)^{t-\varsigma} + \sqrt{\frac{(t+\frac{\log^3 n}{\lambda-1})\log n}{n(\lambda-1)^5}}.\end{aligned} \tag{76}$$

B Proof of auxiliary lemmas and claims

B.1 Proof of Lemma 2

Proof of property (23a). Recall that $\pi_t \coloneqq \sqrt{n(\|x_t\|_2^2-1)\vee 1}$. To show property (23a), the first step is to calculate $\|x_t\|_2$. Notice that for independent Gaussian vectors $\phi_k \sim \mathcal{N}\big(0, \frac{1}{n}I_n\big)$, one has

$$v^{\star\top}\left[\phi_1,\ldots,\phi_{t-1}\right] \sim \mathcal{N}\Big(0, \frac{1}{n}I_{t-1}\Big),$$

given that $\|v^\star\|_2 = 1$. Therefore, it is easily seen that

$$\begin{aligned}\left|\left\langle v^\star, \sum_{k=1}^{t-1}\beta_{t-1}^k\phi_k\right\rangle\right| = \left|\left\langle v^{\star\top}\left[\phi_1,\ldots,\phi_{t-1}\right], \left[\beta_{t-1}^1,\ldots,\beta_{t-1}^{t-1}\right]\right\rangle\right| &\le \left\|v^{\star\top}\left[\phi_1,\ldots,\phi_{t-1}\right]\right\|_2 \|\beta_{t-1}\|_2 \\ &\lesssim \sqrt{\frac{t\log n}{n}} \end{aligned} \tag{77}$$

with probability at least $1-O(n^{-11})$. Combining inequalities (77) and (89), we reach

$$\begin{aligned}\|v_t\|_2^2 &= \left\|\alpha_t v^\star + \sum_{k=1}^{t-1}\beta_{t-1}^k\phi_k\right\|_2^2 = \|\alpha_t v^\star\|_2^2 + \left\|\sum_{k=1}^{t-1}\beta_{t-1}^k\phi_k\right\|_2^2 + 2\left\langle \alpha_t v^\star, \sum_{k=1}^{t-1}\beta_{t-1}^k\phi_k\right\rangle \\ &= \alpha_t^2 + 1 + O\left(\sqrt{\frac{t\log n}{n}}\right) \asymp 1, \end{aligned} \tag{78}$$

which in turn leads to

$$\|x_t\|_2^2 = \big(\|v_t\|_2 + O(\|\xi_{t-1}\|_2)\big)^2 = \alpha_t^2 + 1 + O\bigg(\|\xi_{t-1}\|_2 + \sqrt{\frac{t\log n}{n}}\bigg). \tag{79}$$

Based on the above properties, we can also derive that

$$\begin{aligned}\sqrt{n(\|x_t\|_2^2-1)} &= \sqrt{n\left(\alpha_t^2 + O\left(\|\xi_{t-1}\|_2 + \sqrt{\frac{t\log n}{n}}\right)\right)} \le \sqrt{n}\,|\alpha_t| + O\left(\sqrt{n\bigg(\|\xi_{t-1}\|_2 + \sqrt{\frac{t\log n}{n}}\bigg)}\right) \\ &= \sqrt{n}\left(|\alpha_t| + O\bigg(\bigg(\|\xi_{t-1}\|_2 + \sqrt{\frac{t\log n}{n}}\bigg)^{1/2}\bigg)\right). \end{aligned} \tag{80}$$

where the first inequality based on the fact that $\sqrt{a+b} \le \sqrt{a}+\sqrt{b}$ for $a,b\ge 0$. In particular, if $\alpha_t^2 \lesssim \|\xi_{t-1}\|_2 + \sqrt{\frac{t\log n}{n}}$, then the basic relation $\sqrt{a+b} = \sqrt{a} + O(\sqrt{b})$ $(0 \le a \lesssim b)$ enables us to replace "$\le$" in (81) with "$=$" to obtain

$$\sqrt{n(\|x_t\|_2^2-1)} = \sqrt{n}\left(\alpha_t + O\bigg(\bigg(\|\xi_{t-1}\|_2 + \sqrt{\frac{t\log n}{n}}\bigg)^{1/2}\bigg)\right). \tag{81}$$

In contrast, if $\alpha_t^2 \gtrsim \|\xi_{t-1}\|_2 + \sqrt{\frac{t\log n}{n}}$, then the basic relation $\sqrt{a+b} = \sqrt{a} + O(\frac{b}{\sqrt{a}})$ $(0 \le b \lesssim a)$ yields

$$\begin{aligned}\sqrt{n(\|x_t\|_2^2-1)} = \sqrt{n\left(\alpha_t^2 + O\left(\|\xi_{t-1}\|_2 + \sqrt{\frac{t\log n}{n}}\right)\right)} &= \sqrt{n}|\alpha_t|\sqrt{1 + \frac{1}{\alpha_t^2}O\left(\|\xi_{t-1}\|_2 + \sqrt{\frac{t\log n}{n}}\right)} \\ &= \sqrt{n}\left(|\alpha_t| + \frac{1}{|\alpha_t|}O\left(\|\xi_{t-1}\|_2 + \sqrt{\frac{t\log n}{n}}\right)\right). \end{aligned} \tag{82}$$

The preceding two bounds taken collectively demonstrate that

$$\sqrt{n(\|x_t\|_2^2-1)} = \sqrt{n}|\alpha_t| + \sqrt{n}\left\{\frac{1}{|\alpha_t|}O\left(\|\xi_{t-1}\|_2 + \sqrt{\frac{t\log n}{n}}\right) \wedge O\bigg(\bigg(\|\xi_{t-1}\|_2 + \sqrt{\frac{t\log n}{n}}\bigg)^{1/2}\bigg)\right\}.$$

The above bound also leads to the desired form (23a) for π_t when $\sqrt{n(\|x_t\|_2^2-1)} \geq 1$. Consequently, it remains to examine the case where $\sqrt{n(\|x_t\|_2^2-1)} < 1$, which clearly can only happen if

$$c_{10}|\alpha_t| \leq \left(\|\xi_{t-1}\|_2 + \sqrt{\frac{t\log n}{n}}\right)^{1/2}$$

for some sufficiently small constant $c_{10} > 0$. But in this situation, we still have

$$\begin{aligned}\pi_t = 1 &\lesssim \sqrt{n}\bigg(\|\xi_{t-1}\|_2 + \sqrt{\frac{t\log n}{n}}\bigg)^{1/2} \asymp \sqrt{n}\left(|\alpha_t| + O\bigg(\bigg(\|\xi_{t-1}\|_2 + \sqrt{\frac{t\log n}{n}}\bigg)^{1/2}\bigg)\right)\\ &\asymp \sqrt{n}|\alpha_t| + \sqrt{n}\left\{O\bigg(\bigg(\|\xi_{t-1}\|_2 + \sqrt{\frac{t\log n}{n}}\bigg)^{1/2}\bigg) \wedge \frac{1}{|\alpha_t|}O\left(\|\xi_{t-1}\|_2 + \sqrt{\frac{t\log n}{n}}\right)\right\},\end{aligned}$$

and hence the claimed bound is still valid.

Proof of property (23b). First recall the definition $\gamma_t^{-2} \coloneqq \|\tanh(\pi_t x_t)\|_2^2$. Towards establishing property (23b), consider the following difference

$$\begin{aligned}&\left|\|\tanh(\pi_t x_t)\|_2^2 - \int \left\|\tanh\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x)\right)\right\|_2^2 \varphi_n(\mathrm{d}x)\right|\\ &\quad\leq \left|\|\tanh(\pi_t x_t)\|_2^2 - \|\tanh(\pi_t v_t)\|_2^2\right| + \left|\|\tanh(\pi_t v_t)\|_2^2 - \int \left\|\tanh\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x)\right)\right\|_2^2 \varphi_n(\mathrm{d}x)\right|\end{aligned}$$

where $\varphi_n \sim \mathcal{N}(0, I_n)$. To bound the right-hand side above, note that the Lipschitz property of tanh gives

$$\|\tanh(\pi_t x_t) - \tanh(\pi_t v_t)\|_2 \leq \|\tanh(\pi_t(x_t - v_t))\|_2 \leq \pi_t\|\xi_{t-1}\|_2.$$

This in turn yields

$$\begin{aligned}\left|\|\tanh(\pi_t x_t)\|_2^2 - \|\tanh(\pi_t v_t)\|_2^2\right| &= \left|\,\|\tanh(\pi_t x_t)\|_2 + \|\tanh(\pi_t v_t)\|_2\,\right| \cdot \left|\,\|\tanh(\pi_t x_t)\|_2 - \|\tanh(\pi_t v_t)\|_2\,\right|\\ &\leq \big(2\,\|\tanh(\pi_t v_t)\|_2 + \pi_t\|\xi_{t-1}\|_2\big)\pi_t\|\xi_{t-1}\|_2\\ &\lesssim \big(\pi_t\,\|v_t\|_2 + \pi_t\|\xi_{t-1}\|_2\big)\pi_t\|\xi_{t-1}\|_2 \asymp \pi_t^2\|\xi_{t-1}\|_2,\end{aligned}$$

where the last line makes use of the assumption $\|\xi_{t-1}\|_2 \lesssim 1$ and the equation (78). Thus, we arrive at

$$\begin{aligned}&\left|\|\tanh(\pi_t x_t)\|_2^2 - \int \left\|\tanh\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x)\right)\right\|_2^2 \varphi_n(\mathrm{d}x)\right|\\ &\quad\lesssim \pi_t^2\|\xi_{t-1}\|_2 + \left|\|\tanh(\pi_t v_t)\|_2^2 - \int \left\|\tanh\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x)\right)\right\|_2^2 \varphi_n(\mathrm{d}x)\right|.\end{aligned} \tag{83}$$

Next we develop a bound on the second term of (83), which turns out to be a consequence of the uniform concentration property. Specifically, let us consider functions of the following form

$$f_\theta(\Phi) \coloneqq \|\tanh(\pi v)\|_2^2 - \int \left\|\tanh\left(\frac{\pi}{\sqrt{n}}(\alpha + x)\right)\right\|_2^2 \varphi_n(\mathrm{d}x), \qquad \text{where } v = \alpha v^\star + \sum_{k=1}^{t-1}\beta^k\phi_k;$$

here, we define

$$\Phi = \sqrt{n}\left[\phi_1, \phi_2, \ldots, \phi_{t-1}\right], \quad \beta = \left[\beta^1, \beta^2, \ldots, \beta^{t-1}\right], \quad \theta = [\alpha, \beta, \pi] \in \mathbb{R}^{t+1}.$$

Then, it suffices to bound $f_\theta(\Phi)$ uniformly over all θ in the following set

$$\Theta := \left\{\theta = (\alpha, \beta, \pi) \mid \|\beta\|_2 = 1, \alpha \lesssim \sqrt{\lambda^2 - 1}, \pi \lesssim \sqrt{n}\right\}. \tag{84}$$

In order to achieve this, first consider the derivative of f with respect to Φ, which by direct calculations satisfies

$$\begin{aligned}\|\nabla_\Phi f_\theta(\Phi)\|_2 &\leq \frac{2\pi\|\beta\|_2}{\sqrt{n}}\|\tanh(\pi v) \circ \tanh'(\pi v)\|_2 \\ &\leq \frac{2\pi\|\beta\|_2}{\sqrt{n}}\|\tanh(\pi v)\|_2 \lesssim \frac{\pi^2}{\sqrt{n}}\|v\|_2 \lesssim \frac{\pi^2}{\sqrt{n}},\end{aligned} \tag{85}$$

where we note that $\|v\|_2 \leq \alpha + \frac{1}{\sqrt{n}}\|\Phi\| \lesssim 1$. Additionally, since function $f_\theta(\Phi)$ is uniformly bounded by 2, if we take $\delta = n^{-300}$ for the set $\mathcal{E}$ defined in Lemma 1, it satisfies (we refer the readers to [2, Section D.1.1] for the proof of this property)

$$\mathbb{E}\left[|f(\Phi) - f(\mathcal{P}_\mathcal{E}(\Phi))|\right] \lesssim n^{-100}.$$

where $\mathcal{P}_\mathcal{E}(\cdot)$ denotes the Euclidean projection onto the set $\mathcal{E}$. Combining the above inequality with the following properties of function $f_\theta(\Phi)$,

1. $\|\nabla_\theta f_\theta(\Phi)\|_2 \lesssim n^{100}$ for any $\Phi \in \mathcal{E}$,
2. For any fixed θ, one has $\mathbb{E}[f_\theta(\Phi)] = 0$,

we can apply [2, Corollary 3] to reach

$$\sup_{\theta\in\Theta}\left|\frac{1}{\pi^2}f_\theta(\Phi)\right| \lesssim \sqrt{\frac{t\log n}{n}}. \tag{86}$$

Taking everything collectively, we conclude that

$$\left|\|\tanh(\pi_t x_t)\|_2^2 - \int\left\|\tanh\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x)\right)\right\|_2^2\varphi_n(\mathrm{d}x)\right| \lesssim \pi_t^2\left(\|\xi_{t-1}\|_2 + \sqrt{\frac{t\log n}{n}}\right), \tag{87}$$

which leads to the property (23b) by recognizing that

$$\int\left\|\tanh\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x)\right)\right\|_2^2\varphi_n(\mathrm{d}x) = n\int\tanh^2\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x)\right)\varphi(\mathrm{d}x).$$

Proof of property (23c). We are only left with calculating the value of $\int\tanh^2\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x)\right)\varphi(\mathrm{d}x)$ which shall be done as follows. First, by observing that $\tanh(0) = \tanh''(0) = 0$, $\tanh'(0) = 1$, $|\tanh'''(x)| \leq 4$ for $x \in \mathbb{R}$, we find

$$|\tanh(x) - x| \leq \frac{2}{3}|x|^3,$$

as a consequence of the mean value theorem. Combining this relation with the fact that $|\tanh(x) - x| \leq |x|$, we can further obtain

$$|\tanh(x) - x| \leq |x| \wedge |x|^3.$$

As a result, we see that

$$|\tanh(x) + x| = |2x + \tanh(x) - x| = 2|x| + O(|x| \wedge |x|^3) \lesssim |x|,$$

which in turn leads to

$$\tanh^2(x) = x^2 + O(|\tanh(x) - x||\tanh(x) + x|) = x^2 + O(x^4).$$

Now we are ready to compute the value of $\int \tanh^2\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x)\right)\varphi(\mathrm{d}x)$. In view of the expressions obtained above, it follows that

$$\begin{aligned}\int \tanh^2\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x)\right)\varphi(\mathrm{d}x) &= \int \left(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x)\right)^2 \varphi(\mathrm{d}x) + O\left(\int \left(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x)\right)^4 \varphi(\mathrm{d}x)\right) \\ &= \frac{\pi_t^2}{n}(\alpha_t^2 + 1) + O\left(\frac{\pi_t^4}{n^2}\right). \end{aligned} \tag{88}$$

We thus complete the proof of the advertised result.

B.2 Proof of Lemma 3

Proof of property (25a). The property (25a) is concerned with the magnitudes of η_t and its derivatives. In view of the definition of η_t, we proceed to bounding the parameters π_t and γ_t separately. Towards this, recall from Lemma 1 that: with probability at least $1 - O(n^{-11})$,

$$\left\|\sum_{k=1}^{t-1} \beta_{t-1}^k \phi_k\right\|_2 = 1 + O\left(\sqrt{\frac{t\log n}{n}}\right) \tag{89}$$

holds simultaneously for all $\beta_{t-1} = [\beta_{t-1}^k]_{1\le k<t} \in \mathcal{S}^{t-2}$. It then follows from this result and the definition (22) of v_t that

$$\|v_t\|_2 = \left\|\alpha_t v^\star + \sum_{k=1}^{t-1} \beta_{t-1}^k \phi_k\right\|_2 \le \|\alpha_t v^\star\|_2 + \left\|\sum_{k=1}^{t-1} \beta_{t-1}^k \phi_k\right\|_2 = |\alpha_t| + 1 + O\left(\sqrt{\frac{t\log n}{n}}\right) \lesssim 1, \tag{90}$$

given that $t \lesssim \frac{n}{\log n}$ and

$$|\alpha_t| = \left|\lambda {v^\star}^\top \eta_{t-1}(x_{t-1})\right| \le \lambda \lesssim 1. \tag{91}$$

These properties together with the assumption on ξ_t enable us to control the ℓ_2 norm of x_t as follows:

$$\|x_t\|_2 = \|v_t + \xi_{t-1}\|_2 \le \|v_t\|_2 + \|\xi_{t-1}\|_2 \lesssim 1. \tag{92}$$

Given that $\|x_t\|_2 \lesssim 1$, the value of π_t can be controlled as

$$\pi_t \coloneqq \sqrt{n(\|x_t\|_2^2 - 1)} \vee 1 \lesssim \sqrt{n}. \tag{93}$$

Additionally, by observing that $\tanh(0) = 0$, $\tanh'(x) = 1 - \tanh^2(x) \in [0, 1]$ and $|\tanh(x)| \le 1$, one has

$$|\tanh(x)| \asymp |x| \wedge 1. \tag{94}$$

We claim that this leads to the following consequence:

$$\left\|\tanh(\pi_t x_t)\right\|_2 \asymp \pi_t; \tag{95}$$

for the moment, let us first take this as given and we shall come back to its proof after establishing the property (25a). In view of this claim (95), we find that

$$\gamma_t \coloneqq \|\tanh(\pi_t x_t)\|_2^{-1} \asymp \pi_t^{-1}. \tag{96}$$

In order to prove property (25a), it suffices to recall the definition of η_t as in expression (3). For any $x \in \mathbb{R}$, direct calculations give

$$\eta_t(x) = \gamma_t \tanh(\pi_t x) \tag{97a}$$
$$\eta_t'(x) = \gamma_t \pi_t \big(1 - \tanh^2(\pi_t x)\big) \tag{97b}$$
$$\eta_t''(x) = -2\gamma_t \pi_t^2 \tanh(\pi_t x)\big(1 - \tanh^2(\pi_t x)\big) \tag{97c}$$
$$\eta_t'''(x) = -2\gamma_t \pi_t^3 \big(1 - \tanh^2(\pi_t x)\big)\big(1 - 3\tanh^2(\pi_t x)\big). \tag{97d}$$

Combining these identities with (94), (96) and the fact that $|\tanh(x)| \le 1$, one can easily validate expression (25a). It then boils down to justifying the claim (95), which we accomplish below.

Proof of relation (95). Note that from display (94), one has $\big\| \tanh(\pi_t x_t) \big\|_2 \asymp \big\| |\pi_t x_t| \wedge 1 \big\|_2$, where both operators $|\cdot|$ and $\wedge$ are applied in an entrywise manner and we overlad the notation 1 to denote an all-one vector. To establish the relation (95), it is sufficient to prove that $\big\| |\pi_t x_t| \wedge 1 \big\|_2 \asymp \pi_t$. Towards this end, first we invoke (93) to make the observation that

$$\big\| |\pi_t x_t| \wedge 1 \big\|_2 \le \|\pi_t x_t\|_2 \wedge \|1\|_2 = \|\pi_t x_t\|_2 \wedge \sqrt{n} \asymp \pi_t \wedge \sqrt{n} \asymp \pi_t.$$

In addition, let us introduce an index set $\mathcal{I}$ as follows:

$$\mathcal{I} := \left\{ i \in [n] \;\middle|\; |\xi_{t-1,i}| \le 0.9 |v_{t,i}| \text{ and } |v_{t,i}| \lesssim \frac{1}{\sqrt{n}} \right\}, \tag{98}$$

which clearly satisfies

$$0.1|\pi_t v_{t,i}| \le |\pi_t x_{t,i}| \le 1.9 |\pi_t v_{t,i}|, \qquad \forall i \in \mathcal{I}.$$

It then follows that:

$$\big\| |\pi_t x_t| \wedge 1 \big\|_2 \ge \big\| |\pi_t x_t \circ \mathbb{1}_{\mathcal{I}}| \wedge 1 \big\|_2 \asymp \big\| |\pi_t v_t \circ \mathbb{1}_{\mathcal{I}}| \wedge 1 \big\|_2. \tag{99}$$

To further control the right-hand side of display (99), we claim that

$$\big\| |\pi_t v_t \circ \mathbb{1}_{\mathcal{I}}| \wedge 1 \big\|_2 \overset{\text{(i)}}{\asymp} \big\| \pi_t v_t \circ \mathbb{1}_{\mathcal{I}} \big\|_2 \overset{\text{(ii)}}{\asymp} \pi_t. \tag{100}$$

In order to see this, relation (i) can be verified using expression (93) and the definition of (98), as they guarantee that

$$|\pi_t v_{t,i}| \le |\pi_t| \cdot |v_{t,i}| \lesssim 1, \qquad \forall i \in \mathcal{I}.$$

To validate (ii), note that on the index set $\mathcal{I}^c$, one has $|v_{t,i}| \lesssim |\xi_{t-1,i}|$. Therefore, it holds that

$$\|\pi_t v_t \circ \mathbb{1}_{\mathcal{I}^c}\|_2 \lesssim \|\pi_t \xi_{t-1} \circ \mathbb{1}_{\mathcal{I}^c}\|_2 \lesssim \pi_t \|\xi_{t-1}\|_2 \lesssim \pi_t \sqrt{\frac{1}{\log n}}, \tag{101}$$

where we recall our assumption $\|\xi_{t-1}\|_2 \lesssim \sqrt{\frac{1}{\log n}}$. Equipped with the above calculation, we can recall $\|v_t\|_2 \asymp 1$ from expression (90) to obtain

$$\|\pi_t v_t \circ \mathbb{1}_{\mathcal{I}}\|_2^2 = \|\pi_t v_t\|_2^2 - \|\pi_t v_t \circ \mathbb{1}_{\mathcal{I}^c}\|_2^2 \asymp \left(1 - O\left(\frac{1}{\log n}\right)\right) \pi_t^2 \asymp \pi_t^2, \tag{102}$$

as claimed in Part (ii) of (100).

Proof of property (25b). To study the derivatives of $\eta_t(x_t)$, we first consider the parameters π_t and γ_t. Given that $\|\xi_{t-1}\|_2$ satisfies the expression (24), when $t \lesssim \frac{\log n}{\lambda-1}$, it holds true that

$$\|\xi_{t-1}\|_2 \lesssim \sqrt{\frac{t^3 \log n}{n}} \lesssim \sqrt{\frac{\log^4 n}{n(\lambda-1)^3}}. \tag{103}$$

Under the assumption $\lambda - 1 \gtrsim n^{-1/9} \log n$, we see that $\|\xi_{t-1}\|_2$ also satisfies

$$\|\xi_{t-1}\|_2 \lesssim (\lambda-1) \cdot \sqrt{\frac{\log^4 n}{n} \cdot \frac{1}{(\lambda-1)^5}} \lesssim (\lambda-1) \cdot \sqrt{\frac{\log^4 n}{n} \cdot \frac{n^{5/9}}{\log^5 n}} \lesssim (\lambda-1) n^{-0.2}.$$

Similarly, it is also easily seen that

$$\Big(\frac{t \log n}{n}\Big)^{1/2} \lesssim (\lambda-1) n^{-0.2}.$$

Taking these together with the relation (81) in the proof of Lemma 2 and the assumption $\alpha_t \lesssim \sqrt{\lambda-1}\, n^{-0.1}$ yields

$$\pi_t \le \sqrt{n}\alpha_t + O\bigg(\sqrt{n}\Big(\|\xi_{t-1}\|_2 + \sqrt{\frac{t\log n}{n}}\Big)^{1/2}\bigg) \lesssim \sqrt{\lambda-1}\, n^{0.4}. \tag{104}$$

Similarly, the relation (23d) of Lemma 2 combined with the assumption $\alpha_t \le \sqrt{\lambda-1} n^{-0.1}$ and the above bounds leads to

$$(\gamma_t \pi_t)^{-2} = \alpha_t^2 + 1 + O\bigg(\frac{\pi_t^2}{n} + \|\xi_{t-1}\|_2 + \sqrt{\frac{t\log n}{n}}\bigg) = 1 + O\left((\lambda-1)n^{-0.2}\right),$$

which by direct calculation also gives $\gamma_t \pi_t = 1 + O((\lambda-1)n^{-0.2})$ under our assumption on $\lambda - 1$.

Armed with the above properties, some algebra together with (97) further results in

$$\begin{aligned}
\eta_t'(x) &= \gamma_t \pi_t (1 - \tanh^2(\pi_t x)) = 1 + O\big((\lambda-1)n^{-0.2}\log n\big), \\
|\eta_t''(x)| &\lesssim \pi_t \cdot |\pi_t x| \lesssim (\lambda-1) n^{0.8} |x|,
\end{aligned}$$

where we invoke the relation $|1 - \tanh^2(\pi_t x)| \asymp 1$ for any $|x| \lesssim \sqrt{\frac{\log n}{n}}$.

Proof of property (25c). Again when $t \lesssim \frac{\log n}{\lambda-1}$, $\|\xi_{t-1}\|_2$ satisfies inequality (103). Taking this collectively with the assumption $\alpha_t \lesssim \sqrt{\lambda-1} n^{-1/4}$ and the relation (81) in the proof of Lemma 2, we obtain

$$\begin{aligned}
\pi_t &\le \sqrt{n}\alpha_t + O\bigg(\sqrt{n}\Big(\|\xi_{t-1}\|_2 + \sqrt{\frac{t\log n}{n}}\Big)^{1/2}\bigg) \lesssim \sqrt{\lambda-1}\, n^{1/4} + (\lambda-1)^{-3/4} n^{1/4} \log n \\
&\asymp (\lambda-1)^{-3/4} n^{1/4} \log n.
\end{aligned} \tag{105}$$

Similarly, the relation (23d) of Lemma 2 together with our assumption on $\lambda - 1$ yields

$$\gamma_t \pi_t = 1 + O\left(\frac{\log^2 n}{\sqrt{n(\lambda-1)^3}}\right). \tag{106}$$

To derive property (25c), we make note of some simple facts that $\tanh(0) = \tanh''(0) = \tanh''''(0) = 0$, $\tanh'(0) = 1$, $\tanh'''(0) = -2$ and $|\tanh^{(5)}(x)| \le K$ for some constant K. As a result, the mean value theorem ensures that for any x, there exists a quantity c such that

$$\tanh(x) = x - \frac{1}{3}x^3 + cx^5 \qquad \text{for some} \quad 0 \le c \le K',$$

for $K' = K/120$. Based on the calculations above, we can conclude that

$$\eta_t(x) = \gamma_t \tanh(\pi_t x) = (1-c_0)\pi_t^{-1}\tanh(\pi_t x) = (1-c_0)\Big(x - \frac{1}{3}\pi_t^2 x^3 + c_x\Big),$$

where c_0 and c_x are some quantities obeying

$$|c_0| \lesssim \frac{\log^2 n}{\sqrt{n(\lambda-1)^3}} \qquad \text{and} \qquad |c_x| \lesssim \pi_t^4 |x|^5 \lesssim \frac{n|x|^5 \log^4 n}{(\lambda-1)^3}.$$

This completes the proof of the desired property.

B.3 Proof of Lemma 4

Without loss of generality, throughout this proof, we assume $\alpha_t > 0$. Before we begin to bound κ_t, let us simplify the term of interest a little bit. First of all, as each entry follows $v_i^\star \overset{\text{i.i.d.}}{\sim} \mathsf{Unif}\{\frac{1}{\sqrt{n}}, -\frac{1}{\sqrt{n}}\}$, one can easily derive

$$\begin{aligned}
&\left\langle \int \left[x\eta_t'\left(\alpha_t v^\star + \frac{1}{\sqrt{n}}x\right) - \frac{1}{\sqrt{n}}\eta_t''\left(\alpha_t v^\star + \frac{1}{\sqrt{n}}x\right)\right]^2 \varphi_n(\mathrm{d}x)\right\rangle \\
&\quad = \int \left[x\eta_t'\left(\alpha_t v_1^\star + \frac{1}{\sqrt{n}}x\right) - \frac{1}{\sqrt{n}}\eta_t''\left(\alpha_t v_1^\star + \frac{1}{\sqrt{n}}x\right)\right]^2 \varphi(\mathrm{d}x) \\
&\quad = \int \left[x\eta_t'\left(\frac{1}{\sqrt{n}}(\alpha_t + x)\right) - \frac{1}{\sqrt{n}}\eta_t''\left(\frac{1}{\sqrt{n}}(\alpha_t + x)\right)\right]^2 \varphi(\mathrm{d}x),
\end{aligned}$$

where $\varphi(\cdot)$ is the p.d.f. of $\mathcal{N}(0,1)$, and we have used the fact that $\eta_t(\cdot)$ is symmetric about 0 and the integration is over the distribution $\mathcal{N}(0, I_n)$. Similarly, we obtain

$$\left\langle \int \left[\eta_t'\left(\alpha_t v^\star + \frac{1}{\sqrt{n}}x\right)\right]^2 \varphi_n(\mathrm{d}x)\right\rangle = \int \left[\eta_t'\Big(\frac{1}{\sqrt{n}}(\alpha_t + x)\Big)\right]^2 \varphi(\mathrm{d}x).$$

Therefore, it holds that

$$\kappa_t^2 = \max\left\{ \int |I_1(x)|^2 \varphi(\mathrm{d}x),\ \int |I_2(x)|^2 \varphi(\mathrm{d}x)\right\}, \tag{107}$$

where we define

$$\begin{aligned}
I_1(x) &\coloneqq x\eta_t'\Big(\frac{1}{\sqrt{n}}(\alpha_t + x)\Big) - \frac{1}{\sqrt{n}}\eta_t''\Big(\frac{1}{\sqrt{n}}(\alpha_t + x)\Big), \\
I_2(x) &\coloneqq \eta_t'\Big(\frac{1}{\sqrt{n}}(\alpha_t + x)\Big).
\end{aligned}$$

We now proceed to the proof of Lemma 4, and begin by restricting our attention to the range $t < \varsigma \wedge \frac{\log n}{c(\lambda-1)}$. We divide into two cases depending on the value of α_t.

Case I: $\alpha_t \lesssim \sqrt{\lambda^2-1}\, n^{-0.1}$. Let us introduce an event $\mathcal{A} \coloneqq \big\{x : |x| \le \sqrt{24\log n}\big\}$. For $x \sim \mathcal{N}(0,1)$, it is easily verified that $P(\mathcal{A}) = 1 - O(n^{-12})$. Conditional on $\mathcal{A}$ and assuming $\alpha_t \lesssim \sqrt{\lambda^2-1}\, n^{-0.1}$, one has

$$\Big|\frac{1}{\sqrt{n}}(\alpha_t + x)\Big| \lesssim \sqrt{\frac{\log n}{n}}.$$

Meanwhile, recall that $\sqrt{\lambda+1} \asymp 1$, and therefore $\alpha_t \lesssim \sqrt{\lambda^2-1}\, n^{-0.1}$ is equivalent to $\alpha_t \lesssim \sqrt{\lambda-1}\, n^{-0.1}$. As a result, according to the property (25b) established in Lemma 3, we see that: when $|z| \lesssim \sqrt{\frac{\log n}{n}}$, one has

$$\eta_t'(z) = 1 + O\big((\lambda-1)n^{-0.2}\log n\big) \qquad \text{and} \qquad |\eta_t''(z)| \lesssim (\lambda-1)n^{0.8}|z|. \tag{108}$$

Hence, for all x residing within $\mathcal{A}$, we can bound the difference between $I_1(x)$ and x uniformly as follows:

$$\begin{aligned}
|I_1(x)-x| &\le \left|\eta_t'\Big(\frac{1}{\sqrt{n}}(\alpha_t+x)\Big)-1\right||x| + \frac{1}{\sqrt{n}}\left|\eta_t''\Big(\frac{1}{\sqrt{n}}(\alpha_t+x)\Big)\right| \\
&\lesssim \big((\lambda-1)n^{-0.2}\log n\big)\,|x| + (\lambda-1)n^{0.3}\left|\frac{1}{\sqrt{n}}(\alpha_t+x)\right| \\
&\lesssim \big((\lambda-1)n^{-0.2}\log n\big)\,|x| + (\lambda-1)n^{0.3}\left(\frac{1}{\sqrt{n}}+\frac{1}{\sqrt{n}}|x|\right) \\
&\lesssim \big((\lambda-1)n^{-0.2}\log n\big)\,(1+|x|). \qquad (109)
\end{aligned}$$

In addition, consider any $x\in\mathbb{R}$. Recalling the relation (25a) of Lemma 3 — which reveals that $|\eta_t'(x)|\lesssim 1$, $|\eta_t''(x)|\lesssim\sqrt{n}$ — we observe the crude bound that $|I_1(x)|\lesssim |x|+1$. Putting the preceding two bounds together, we arrive at

$$\begin{aligned}
\int |I_1(x)|^2\varphi(\mathrm{d}x) &= \int x^2\varphi(\mathrm{d}x) + \int (I_1(x)-x)(I_1(x)+x)\varphi(\mathrm{d}x) \\
&= 1+\int_{\mathcal{A}}(I_1(x)-x)(I_1(x)+x)\varphi(\mathrm{d}x) + \int_{\mathcal{A}^c}(I_1(x)-x)(I_1(x)+x)\varphi(\mathrm{d}x) \\
&= 1+O\left(\int_{\mathcal{A}}(\lambda-1)n^{-0.2}\log n\cdot(|x|+1)^2\varphi(\mathrm{d}x) + \int_{\mathcal{A}^c}(|x|+1)^2\varphi(\mathrm{d}x)\right) \\
&= 1+O\left((\lambda-1)n^{-0.2}\log n\right), \qquad (110)
\end{aligned}$$

where the last equality utilizes the fact that for $c_n=\sqrt{24\log n}$, one has

$$\int_{\mathcal{A}^c} x^2\varphi(\mathrm{d}x) = 2\int_{c_n}^{\infty} x^2\varphi(\mathrm{d}x) \lesssim \int_{c_n}^{\infty} x^2\exp\left(-\frac{1}{2}x^2\right)\mathrm{d}x \lesssim \frac{\log n}{n^{12}}.$$

Regarding the other term $I_2(x)$, relation (25a) of Lemma 3 implies that $|I_2(x)|\lesssim 1$. Furthermore, if $|x|\le\sqrt{24\log n}$, in view of the relation (25b) we have $|I_2(x)|=1+O((\lambda-1)n^{-0.2}\log n)$. Putting these together, we obtain

$$\begin{aligned}
\int |I_2(x)|^2\varphi(\mathrm{d}x) &= \int_{\mathcal{A}}|I_2(x)|^2\varphi(\mathrm{d}x) + \int_{\mathcal{A}^c}|I_2(x)|^2\varphi(\mathrm{d}x) \\
&= 1+O\big((\lambda-1)n^{-0.2}\log n\big) + O\big(P(\mathcal{A}^c)\big) \\
&= 1+O\big((\lambda-1)n^{-0.2}\log n\big). \qquad (111)
\end{aligned}$$

Combining inequalities (110) and (111) then leads to

$$\kappa_t^2 = 1+O\left((\lambda-1)n^{-0.2}\log n\right) = 1+o\left(\frac{\lambda-1}{\log n}\right).$$

Case II: $\sqrt{\lambda^2-1}\,n^{-0.1}\lesssim\alpha_t\lesssim\sqrt{\lambda^2-1}$. We first note that under the assumption (26), the following relation holds for $\|\xi_{t-1}\|_2$ when $t\lesssim\frac{\log n}{\lambda-1}$:

$$\|\xi_{t-1}\|_2 \lesssim \sqrt{\frac{t^3\log n}{n}} \lesssim \sqrt{\frac{\log^4 n}{n(\lambda-1)^3}}. \qquad (112)$$

We recall the basic facts obtained in property (97), and as a result,

$$\int |I_1(x)|^2\varphi(\mathrm{d}x) = \int\left[\left(\gamma_t\pi_t x + \frac{2}{\sqrt{n}}\gamma_t\pi_t^2\tanh\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t+x)\right)\right)\cdot\left(1-\tanh^2\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t+x)\right)\right)\right]^2\varphi(\mathrm{d}x) \qquad (113)$$

$$\int |I_2(x)|^2\varphi(\mathrm{d}x) = \int\left[\gamma_t\pi_t\left(1-\tanh^2\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t+x)\right)\right)\right]^2\varphi(\mathrm{d}x). \qquad (114)$$

It then comes down to controlling the right-hand side of the above two expressions, under the condition that $\sqrt{\lambda^2-1}n^{-0.1} \lesssim \alpha_t \leq \sqrt{\lambda^2-1}$.

For notational convenience, let us introduce additional shorthand notation as follows:

$$\begin{aligned}
J_1(x) &:= \gamma_t \pi_t x + \frac{2}{\sqrt{n}} \gamma_t \pi_t^2 \tanh\Big(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x)\Big), \\
J_2(x) &:= \mu_t \alpha_t x + 2\mu_t \alpha_t^2 \tanh\Big(\alpha_t(\alpha_t + x)\Big), \\
K_1(x) &:= 1 - \tanh^2\Big(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x)\Big), \\
K_2(x) &:= 1 - \tanh^2\Big(\alpha_t(\alpha_t + x)\Big) \\
\mu_t &:= \bigg[\int \tanh^2\big(\alpha_t(\alpha_t + x)\big)\varphi(\mathrm{d}x)\bigg]^{-\frac{1}{2}}.
\end{aligned}$$

With this set of notation in place, it follows from (113) and a little algebra that

$$\begin{aligned}
\int |I_1(x)|^2 \varphi(\mathrm{d}x) &= \int [J_1(x)K_1(x)]^2 \varphi(\mathrm{d}x) \\
&= \int [J_2(x)K_2(x)]^2 \varphi(\mathrm{d}x) + O\bigg(\int |J_2(x)|^2 \, |K_1(x) - K_2(x)| \, |K_1(x) + K_2(x)| \, \varphi(\mathrm{d}x)\bigg) \\
&\qquad + O\bigg(\int |J_1(x) - J_2(x)||J_1(x) + J_2(x)||K_1(x)|^2 \varphi(\mathrm{d}x)\bigg). \qquad (115)
\end{aligned}$$

To bound $\int |I_1(x)|^2 \varphi(\mathrm{d}x)$, it is thus sufficient to control these three terms on the right-hand side of (115) separately. Before proceeding, we find it helpful to make note of several preliminary properties.

- First, we would like to show that

$$\mu_t \asymp \alpha_t^{-1}. \qquad (116)$$

  In order to see this, note that the elementary fact $\tanh^2(x) \leq x^2$ implies that

$$\int \tanh^2(\alpha_t(\alpha_t + x))\varphi(\mathrm{d}x) \leq \int \alpha_t^2(\alpha_t + x)^2 \varphi(\mathrm{d}x) \asymp \alpha_t^2, \qquad \text{for } \alpha_t \leq \lambda.$$

  On the other hand, when $|x| \leq 1/2$, $\alpha_t \in (0, \lambda]$ and $\lambda \in (1, 1.2]$, one has $\alpha_t(\alpha_t + x) \in [-0.0625, 2.04]$. Clearly, for any $z \in [-0.0625, 2.04]$, we have $\tanh^2(z)/z^2 \geq 0.22$, and as a consequence,

$$\begin{aligned}
\int \tanh^2(\alpha_t(\alpha_t + x))\varphi(\mathrm{d}x) &\geq \int \tanh^2(\alpha_t(\alpha_t + x)) \mathbf{1}(|x| \leq 0.5)\varphi(\mathrm{d}x) \\
&\gtrsim \int \alpha_t^2(\alpha_t + x)^2 \mathbf{1}(|x| \leq 0.5)\varphi(\mathrm{d}x) \asymp \alpha_t^2.
\end{aligned}$$

  The preceding two bounds taken collectively justify the claim (116).

- Additionally, based on equation (23a) of Lemma 2 and the bound (112), we have

$$\pi_t = \sqrt{n}\alpha_t + \frac{\sqrt{n}}{\alpha_t} O\bigg(\|\xi_{t-1}\|_2 + \sqrt{\frac{t \log n}{n}}\bigg) = \alpha_t \sqrt{n} + O\left(\sqrt{\frac{\log^4 n}{\alpha_t^2(\lambda - 1)^3}}\right) \qquad (117)$$

$$= \alpha_t \sqrt{n}\left(1 + O\left(\sqrt{\frac{\log^4 n}{n\alpha_t^4(\lambda - 1)^3}}\right)\right) \asymp \alpha_t \sqrt{n} \qquad (118)$$

as long as $t \lesssim \frac{\log n}{\lambda-1}$. Here, the last inequality holds since

$$\sqrt{\frac{\log^4 n}{(\lambda-1)^3}} \lesssim (\lambda-1)\cdot\frac{\log^2 n}{(\lambda-1)^{5/2}} = o\big((\lambda-1)n^{0.3}\big) = o\big(\alpha_t^2\sqrt{n}\big),$$

provided that $\alpha_t \gtrsim \sqrt{\lambda^2-1}n^{-0.1}$ and $\lambda - 1 \gtrsim n^{-1/9}\log n$.

- Moreover, by combining equation (23b) of Lemma 2 with (117), (118) and (112), we see that

$$\begin{aligned}
\gamma_t^{-2} &= n\int \tanh^2\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t+x)\right)\varphi(\mathrm{d}x) + O\left(\alpha_t^2\sqrt{\frac{n\log^4 n}{(\lambda-1)^3}}\right)\\
&= n\int \tanh^2\left(\alpha_t(\alpha_t+x)\right)\varphi(\mathrm{d}x) + O\left(\sqrt{\frac{n\log^4 n}{(\lambda-1)^3}}\right)\\
&= n\mu_t^{-2} + O\left(\sqrt{\frac{n\log^4 n}{(\lambda-1)^3}}\right).
\end{aligned}$$

Here, the second equality arises from the following fact (see also (149)):

$$\begin{aligned}
&\left|\int \tanh^2\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t+x)\right) - \tanh^2\left(\alpha_t(\alpha_t+x)\right)\varphi(\mathrm{d}x)\right|\\
&\le \left|1-\frac{\alpha_t^2 n}{\pi_t^2}\right|\int \tanh^2\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t+x)\right)\varphi(\mathrm{d}x) + \left|\int \frac{\alpha_t^2 n}{\pi_t^2}\tanh^2\left(\frac{\pi_t}{\sqrt{n}}\left(\alpha_t+x\right)\right) - \tanh^2\left(\alpha_t\left(\alpha_t+x\right)\right)\varphi(\mathrm{d}x)\right|\\
&\lesssim \left|1-\frac{\alpha_t^2 n}{\pi_t^2}\right|\alpha_t^2 + \left|\frac{\pi_t^2}{\alpha_t^2 n}-1\right|\alpha_t^4 = O\left(\sqrt{\frac{\log^4 n}{n(\lambda-1)^3}}\right). \qquad (119)
\end{aligned}$$

As a result, we can express γ_t in term of μ_t as

$$\gamma_t = \left[n\mu_t^{-2} + O\left(\sqrt{\frac{n\log^4 n}{(\lambda-1)^3}}\right)\right]^{-1/2} = \frac{\mu_t}{\sqrt{n}}\left(1+O\left(\mu_t^2\sqrt{\frac{\log^4 n}{n(\lambda-1)^3}}\right)\right) \asymp \frac{\mu_t}{\sqrt{n}}. \qquad (120)$$

- With (118) and (120) in place, a little algebra further leads to

$$|\gamma_t\pi_t - \mu_t\alpha_t| \lesssim \mu_t\alpha_t\sqrt{\frac{\log^4 n}{n(\lambda-1)^3}}\left(\frac{1}{\alpha_t^2}+\mu_t^2\right) \lesssim \frac{1}{\alpha_t^2}\sqrt{\frac{\log^4 n}{n(\lambda-1)^3}}, \qquad (121)$$

$$\left|\frac{1}{\sqrt{n}}\gamma_t\pi_t^2 - \mu_t\alpha_t^2\right| \lesssim \mu_t\alpha_t^2\sqrt{\frac{\log^4 n}{n(\lambda-1)^3}}\left(\frac{1}{\alpha_t^2}+\mu_t^2\right) \lesssim \frac{1}{\alpha_t}\sqrt{\frac{\log^4 n}{n(\lambda-1)^3}}. \qquad (122)$$

In addition, according to the relation (23d) in Lemma 2, we have

$$\gamma_t^2\pi_t^2 = (\alpha_t^2+1)^{-1} + O\left(\frac{\pi_t^2}{n} + \|\xi_{t-1}\|_2 + \sqrt{\frac{t\log n}{n}}\right),$$

and using similar analysis as for (23c) yields

$$\mu_t^{-2} = \int \tanh^2\left(\alpha_t(\alpha_t+x)\right)\varphi(\mathrm{d}x) = \alpha_t^2(\alpha_t^2+1) + O\left(\alpha_t^4\right).$$

These two bounds taken together with a little algebra leads to

$$|\gamma_t\pi_t - \mu_t\alpha_t| \lesssim \frac{\pi_t^2}{n} + \alpha_t^2 + \|\xi_{t-1}\|_2 + \sqrt{\frac{t\log n}{n}} \lesssim \alpha_t^2 + \sqrt{\frac{\log^4 n}{n(\lambda-1)^3}} \lesssim \alpha_t^2, \qquad (123)$$

recognizing the range $\sqrt{\lambda^2-1}\, n^{-0.1} \lesssim \alpha_t \lesssim \sqrt{\lambda^2-1}$ and our assumption $\lambda - 1 \gtrsim n^{-1/9}\log n$. Combining the above bound with (121) gives

$$|\gamma_t \pi_t - \mu_t \alpha_t| \lesssim \alpha_t^2 \wedge \frac{1}{\alpha_t^2}\sqrt{\frac{\log^4 n}{n(\lambda-1)^3}} \lesssim \left(\frac{\log^4 n}{n(\lambda-1)^3}\right)^{1/4}, \tag{124}$$

where the last inequality follows from the elementary fact that $\min\{a,b\} \le \sqrt{ab}$.

- Taking inequalities (118), (121) and (122) collectively with the fact that $\mu_t \alpha_t^2 \asymp \alpha_t$ (cf. (116)) yields

$$|K_1(x) - K_2(x)| \lesssim \left|\frac{\pi_t}{\sqrt{n}} - \alpha_t\right| |\alpha_t + x| \lesssim \frac{1}{\alpha_t}\sqrt{\frac{\log^4 n}{n(\lambda-1)^3}} \cdot (|x| + \alpha_t), \tag{125}$$

and

$$\begin{aligned}
|J_1(x) - J_2(x)| &\le |\gamma_t \pi_t - \mu_t \alpha_t|\,|x| + 2\left|\frac{1}{\sqrt{n}}\gamma_t \pi_t^2 - \mu_t \alpha_t^2\right| \left|\tanh\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x)\right)\right| \\
&\quad + \mu_t \alpha_t^2 \left|\tanh\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x)\right) - \tanh\left(\alpha_t(\alpha_t + x)\right)\right| \\
&\lesssim \left(\frac{\log^4 n}{n(\lambda-1)^3}\right)^{1/4} |x| + \frac{1}{\alpha_t}\sqrt{\frac{\log^4 n}{n(\lambda-1)^3}} + \sqrt{\frac{\log^4 n}{n(\lambda-1)^3}}(|x| + \alpha_t) \\
&\lesssim \left(\frac{\log^4 n}{n(\lambda-1)^3}\right)^{1/4} |x| + \frac{1}{\alpha_t}\sqrt{\frac{\log^4 n}{n(\lambda-1)^3}}.
\end{aligned} \tag{126}$$

Equipped the above relations, we are positioned to control the right-hand side of expression (115). Combining (125) and (126) directly yields

$$\begin{aligned}
\int |I_1(x)|^2 \varphi(\mathrm{d}x) &= \int [J_2(x) K_2(x)]^2 \varphi(\mathrm{d}x) + O\left(\frac{1}{\alpha_t}\sqrt{\frac{\log^4 n}{n(\lambda-1)^3}}\right) + O\left(\left(\frac{\log^4 n}{n(\lambda-1)^3}\right)^{1/4} + \frac{1}{\alpha_t}\sqrt{\frac{\log^4 n}{n(\lambda-1)^3}}\right) \\
&= \int \left[\mu_t \alpha_t x + 2\mu_t \alpha_t^2 \tanh\left(\alpha_t(\alpha_t + x)\right)\right]^2 \left[1 - \tanh^2\left(\alpha_t(\alpha_t + x)\right)\right]^2 \varphi(\mathrm{d}x) \\
&\quad + O\left(\left(\frac{\log^4 n}{n(\lambda-1)^3}\right)^{1/4} + \frac{1}{\alpha_t}\sqrt{\frac{\log^4 n}{n(\lambda-1)^3}}\right),
\end{aligned} \tag{127}$$

where the first line invokes the simple observations that $|J_i(x)| \lesssim |x| + \alpha_t$ and $|K_i(x)| \lesssim 1$ for $i = 1, 2$. Similarly, one can derive in the same manner that

$$\int |I_2(x)|^2 \varphi(\mathrm{d}x) = \mu_t^2 \alpha_t^2 \int \left[1 - \tanh^2\left(\alpha_t(\alpha_t + x)\right)\right]^2 \varphi(\mathrm{d}x) + O\left(\left(\frac{\log^4 n}{n(\lambda-1)^3}\right)^{1/4} + \frac{1}{\alpha_t}\sqrt{\frac{\log^4 n}{n(\lambda-1)^3}}\right). \tag{128}$$

As it turns out, the main terms in the above two identities satisfy

$$\int \left[\mu_t \alpha_t x + 2\mu_t \alpha_t^2 \tanh\left(\alpha_t(\alpha_t + x)\right)\right]^2 \left[1 - \tanh^2\left(\alpha_t(\alpha_t + x)\right)\right]^2 \varphi(\mathrm{d}x) \le 1 - c\alpha_t^2, \tag{129a}$$

$$\mu_t^2 \alpha_t^2 \int \left[1 - \tanh^2\left(\alpha_t(\alpha_t + x)\right)\right]^2 \varphi(\mathrm{d}x) \le 1 - c\alpha_t^2, \tag{129b}$$

which we shall justify momentarily. Combine these results with (107) to conclude that: if $\sqrt{\lambda^2-1}\, n^{-0.1} \lesssim \alpha_t \lesssim \sqrt{\lambda^2-1}$, then

$$\kappa_t^2 \le 1 - c\alpha_t^2 + O\left(\left(\frac{\log^4 n}{n(\lambda-1)^3}\right)^{1/4} + \frac{1}{\alpha_t}\sqrt{\frac{\log^4 n}{n(\lambda-1)^3}}\right) = 1 + o\left(\frac{\lambda-1}{\log n}\right), \tag{130}$$

where the last inequality follows by recognizing that

$$\left(\frac{\log^4 n}{n(\lambda-1)^3}\right)^{1/4} + \frac{1}{\sqrt{\lambda-1}n^{-0.1}}\sqrt{\frac{\log^4 n}{n(\lambda-1)^3}} = o\left(\frac{\lambda-1}{\log n}\right)$$

under our assumption $\lambda - 1 \gtrsim n^{-1/9}\log n$.

Proof of relation (129). To proceed, consider the problem of estimating $v^\star$ (which obeys $v_i^\star \sim \mathsf{Unif}\{\pm\frac{1}{n}\}$) from the noisy observation $Y = \alpha_t v^\star + g$, where $g \sim \mathcal{N}(0, \frac{1}{n}I_n)$. As alluded to previously, the Bayes-optimal estimate (or minimum mean square estimator (MMSE)) is given by

$$\mathbb{E}[v^\star \mid Y] = \tanh(\sqrt{n}\alpha_t Y),$$

which satisfies (due to its optimality)

$$\mathrm{Cor}\left(v^\star, f(Y)\right) \le \mathrm{Cor}\left(v^\star, \mathbb{E}[v^\star \mid Y]\right), \tag{131}$$

for any measurable function f; here $\mathrm{Cor}(\cdot,\cdot)$ denotes the correlation of two random vectors. In particular, the Bayes-optimal estimator outperforms the identity estimator (i.e., $f(Y) = Y$), so that (131) translates to

$$\begin{aligned}\frac{\mathbb{E}[\langle v^\star,\, \alpha_t v^\star + g\rangle]}{\sqrt{\mathbb{E}[\|\alpha_t v^\star + g\|_2^2]}} \le \frac{\mathbb{E}[\langle v^\star,\, \tanh(\sqrt{n}\alpha_t(\alpha_t v^\star + g))\rangle]}{\sqrt{\mathbb{E}[\|\tanh(\sqrt{n}\alpha_t(\alpha_t v^\star + g)\|_2^2]}} &= \frac{\int \tanh(\alpha_t(\alpha_t+x))\varphi(dx)}{\sqrt{\int \tanh^2(\alpha_t(\alpha_t+x))\varphi(\mathrm{d}x)}} \\ &= \sqrt{\int \tanh^2(\alpha_t(\alpha_t+x))\varphi(\mathrm{d}x)},\end{aligned} \tag{132}$$

where the fist equality holds due to the symmetry of $\varphi(\cdot)$, and the second equality holds since $\int \tanh(\alpha^2 + \alpha x)\varphi(\mathrm{d}x) = \int \tanh^2(\alpha^2+\alpha x)\varphi(\mathrm{d}x)$ (see [4, Appendix B.2]). As a consequence, the above relation implies that

$$\frac{\alpha_t}{\sqrt{\alpha_t^2+1}} = \frac{\mathbb{E}[\langle v^\star,\, \alpha_t v^\star + g\rangle]}{\sqrt{\mathbb{E}[\|\alpha_t v^\star + g\|_2^2]}} \le \sqrt{\int \tanh^2(\alpha_t(\alpha_t+x))\varphi(\mathrm{d}x)} = \frac{1}{\mu_t}, \tag{133}$$

which in turns reveals that

$$\mu_t\alpha_t \le \sqrt{\alpha_t^2+1} =: \gamma.$$

Armed with this relation, we can conclude that

$$\begin{aligned}&\max\Bigg\{\int\Big[\mu_t\alpha_t x + 2\mu_t\alpha_t^2\tanh\Big(\alpha_t\big(\alpha_t+x\big)\Big)\Big]\Big[1-\tanh^2\Big(\alpha_t\big(\alpha_t+x\big)\Big)\Big]\varphi(\mathrm{d}x), \\ &\qquad\qquad \mu_t^2\alpha_t^2\int\Big[1-\tanh^2\Big(\alpha_t\big(\alpha_t+x\big)\Big)\Big]^2\varphi(\mathrm{d}x)\Bigg\} \\ &\le \gamma^2\max\Bigg\{\int\Big[x + 2\alpha_t\tanh\Big(\alpha_t\big(\alpha_t+x\big)\Big)\Big]\Big[1-\tanh^2\Big(\alpha_t\big(\alpha_t+x\big)\Big)\Big]\varphi(\mathrm{d}x), \\ &\qquad\qquad \int\Big[1-\tanh^2\Big(\alpha_t\big(\alpha_t+x\big)\Big)\Big]^2\varphi(\mathrm{d}x)\Bigg\} \\ &=: \kappa^2(\gamma,\alpha_t^2).\end{aligned}$$

As it turns out, this function $\kappa^2(\cdot,\cdot)$ has been studied in [2]; more specifically, [2, relation (272)] together with $\gamma := \sqrt{\alpha_t^2+1}$ indicates that $\kappa(\gamma,\alpha_t^2) \le 1 - \frac{\gamma-1}{12}$, and hence

$$\kappa^2(\gamma,\alpha_t^2) \le 1 - \frac{\gamma-1}{12} \le 1 - c\alpha_t^2 \tag{134}$$

for some constant $c > 0$. Here, notice that we view γ and α_t^2 as λ and τ respectively in [2, relation (272)]. Putting everything together completes the proof of the required relation (129).

Case III: $\alpha_t \gtrsim \sqrt{\lambda^2-1}$ and $t < \frac{cn(\lambda-1)^5}{\log^2 n}$. The calculation of κ_t in this case follows from similar arguments as in [2, Section D.3.4.]. The only difference lies in the computing the parameters π_t and γ_t, which was done in [2, Lemma 14] therein but requires a different proof here. Specifically, we aim to show that

$$\pi_t = \big(1+o(\lambda-1)\big)\alpha_t\sqrt{n} \qquad \text{and} \qquad \gamma_t^{-2} = \big(1+o(\lambda-1)\big)n\int \tanh(\alpha_t(\alpha_t+x))\varphi(\mathrm{d}x). \tag{135}$$

If these two relations were valid, then one could follow the argument in [2, Section D.3.4.] verbatim to demonstrate that

$$\kappa_t \le 1-\frac{1}{15}(\lambda-1),$$

as claimed.

We now present how to prove relation (135). In view of the equation (23a) of Lemma 2, we have

$$\begin{aligned}\pi_t &= \alpha_t\sqrt{n}\bigg(1+\frac{1}{\alpha_t^2}O\Big(\|\xi_{t-1}\|_2+\sqrt{\frac{t\log n}{n}}\Big)\bigg) = \alpha_t\sqrt{n}\bigg(1+O\bigg(\sqrt{\frac{(t+\log^3 n/(\lambda-1))\log n}{n(\lambda-1)^2}}\bigg)\bigg)\\ &= \alpha_t\sqrt{n}\bigg(1+O\bigg(\sqrt{\frac{(\lambda-1)^2}{\log n}}\bigg)\bigg) = \alpha_t\sqrt{n}(1+o(\lambda-1))\end{aligned}$$

under the condition $t \lesssim \frac{n(\lambda-1)^5}{\log^2 n}$ and the assumption (28). In addition, with the same analysis as inequality (119), we can guarantee that

$$\begin{aligned}&\int \tanh^2\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t+x)\right)\varphi(\mathrm{d}x) = \int \tanh^2\left(\alpha_t(\alpha_t+x)\right)\varphi(\mathrm{d}x) + O\left(\sqrt{\frac{(t+\log^3 n/(\lambda-1))\log n}{n\alpha_t^2(\lambda-1)^2}}\right)\\ &= \int \tanh\left(\alpha_t(\alpha_t+x)\right)\varphi(\mathrm{d}x) + O\left(\sqrt{\frac{(\lambda-1)^2}{\log n}}\right) = \int \tanh\left(\alpha_t(\alpha_t+x)\right)\varphi(\mathrm{d}x) + o(\lambda-1).\end{aligned}$$

Then according to equation (23b), we can reach

$$\begin{aligned}\gamma_t^{-2} &= n\left(\int \tanh(\alpha_t(\alpha_t+x))\varphi(\mathrm{d}x) + o(\lambda-1)\right) + n\alpha_t^2\big(1+o(\lambda-1)\big)O\left(\sqrt{\frac{(t+\log^3 n/(\lambda-1))\log n}{n}}\right)\\ &= \big(1+o(\lambda-1)\big)n\int \tanh(\alpha_t(\alpha_t+x))\varphi(\mathrm{d}x),\end{aligned}$$

where the last equality follows from the fact that $\int \tanh(\alpha_t(\alpha_t+x))\varphi(\mathrm{d}x) \asymp \alpha_t^2 \asymp 1$ (see relation (116)). This establishes the claim (135).

B.4 Proof of Claim (46)

This subsection aims to establish the advertised decomposition (46). To do so, recall that $\{\eta_i(x_i)\}_{1\le i\le t-1}$ spans the same linear space as $\{z_i\}_{1\le i\le t-1}$ (see (40) and (41b)). It is important to notice that $\{\eta_1(x_1),\dots,\eta_{t-1}(x_{t-1})\}$ are almost orthogonal to each other, thus forming a set of near-orthonormal basis; this property is summarized in the lemma below, whose proof is provided in Section B.8.

Lemma 5. *Suppose that the assumptions of Theorem 1 hold. With probability at least $1-O(n^{-11})$, we have*

$$\Big\|\sum_{i=1}^{t} w_i\eta_i(x_i)\Big\|_2 = \big(1+o(1)\big)\|w\|_2 \tag{136}$$

simultaneously for all $t \le \tau_0$ and all $w=[w_i]_{1\le i\le t}\in\mathbb{R}^t$, where τ_0 is defined in (45).

In view of Lemma 5 and the fact that $\xi_t \in \mathsf{span}(U_{t-1}) = \mathsf{span}\{\eta_1(x_1), \ldots, \eta_{t-1}(x_{t-1})\}$ (cf. (40)), one can write ξ_t as a linear combination of $\{\eta_i(x_i)\}_{1\le i\le t-1}$ as follows:

$$\xi_t = \sum_{k=1}^{t-1} \gamma_t^k \eta_k(x_k), \qquad \text{with } \gamma_t = [\gamma_t^k]_{1\le k<t} \in \mathbb{R}^{t-1} \text{ obeying } \|\gamma_t\|_2 \asymp \|\xi_t\|_2. \tag{137}$$

Armed with this decomposition, we intend to prove that

$$\begin{aligned}\alpha_{t+1} = \lambda v^{\star\top}\eta_t(x_t) &= \lambda v^{\star\top}\eta_t\Big(v_t + \sum_{k=1}^{t-1}\gamma_{t-1}^k\eta_k(x_k)\Big) \\ &= \lambda v^{\star\top}\eta_t(v_t) + O\left(\frac{\log^{2.5} n}{n^{3/4}(\lambda-1)^{1.5}}\right),\end{aligned} \tag{138}$$

which shall be done as follows.

- In order to see this, first note that $\eta_t(\cdot)$ is a Lipschitz function with Lipschitz constant $O(1)$ (see Lemma 3). Therefore, for every $t \lesssim \frac{\log n}{\lambda-1}$ we have

$$\begin{aligned}&\left|v^{\star\top}\eta_t\Big(v_t + \sum_{k=1}^{t-1}\gamma_{t-1}^k\eta_k(x_k)\Big) - v^{\star\top}\eta_t\left(v_t + \sum_{k=1}^{t-1}\gamma_{t-1}^k\eta_k(v_k)\right)\right| \\ &\le \left\|\eta_t\Big(v_t + \sum_{k=1}^{t-1}\gamma_{t-1}^k\eta_k(x_k)\Big) - \eta_t\Big(v_t + \sum_{k=1}^{t-1}\gamma_{t-1}^k\eta_k(v_k)\Big)\right\|_2 \lesssim \sum_{k=1}^{t-1}\left|\gamma_{t-1}^k\right|\|\eta_k(x_k) - \eta_k(v_k)\|_2.\end{aligned}$$

  In view of the decomposition (137) and the Cauchy-Schwarz inequality, we can further obtain

$$\begin{aligned}&\left|v^{\star\top}\eta_t\Big(v_t + \sum_{k=1}^{t-1}\gamma_{t-1}^k\eta_k(x_k)\Big) - v^{\star\top}\eta_t\Big(v_t + \sum_{k=1}^{t-1}\gamma_{t-1}^k\eta_k(v_k)\Big)\right| \lesssim \sum_{k=1}^{t-1}\left|\gamma_{t-1}^k\right|\|\eta_k(x_k) - \eta_k(v_k)\|_2 \\ &\lesssim \sum_{k=1}^{t-1}\left|\gamma_{t-1}^k\right|\|\xi_{k-1}\|_2 \le \|\gamma_{t-1}\|_2\Big(\sum_{k=1}^{t-1}\|\xi_{k-1}\|_2^2\Big)^{1/2} \\ &\asymp \|\xi_{t-1}\|_2\Big(\sum_{k=1}^{t-1}\|\xi_{k-1}\|_2^2\Big)^{1/2} \lesssim \sqrt{\frac{t^3\log n}{n}}\cdot\left(\sum_{k=1}^{t-1}\frac{k^3\log n}{n}\right)^{1/2} \lesssim \frac{\log^{4.5} n}{n(\lambda-1)^{3.5}},\end{aligned}$$

  where the last line invokes $\|\xi_t\|_2 \lesssim \sqrt{\frac{t^3\log n}{n}}$ (cf. (39)) and $t \lesssim \frac{\log n}{\lambda-1}$.

- In addition, when $|\alpha_t| \lesssim \sqrt{\lambda-1}n^{-1/4}$, we know that $\left\|\sum_{k=1}^{t-1}\beta_{t-1}^k\phi_k\right\|_\infty \lesssim \sqrt{\frac{t\log n}{n}}$ conditioned on the event $\{\phi_k\}_{k=1}^{t-1} \in \mathcal{E}$ (defined in Lemma 1 with $\delta = O(n^{-10})$). It therefore guarantees that

$$|v_{t,i}| = \left|\alpha_t v_i^\star + \sum_{k=1}^{t-1}\beta_{t-1}^k\phi_{k,i}\right| \le \frac{|\alpha_t|}{\sqrt{n}} + \left|\sum_{k=1}^{t-1}\beta_{t-1}^k\phi_{k,i}\right| \lesssim \sqrt{\frac{t\log n}{n}}, \tag{139a}$$

$$\begin{aligned}\left|v_{t,i} + \sum_{k=1}^{t-1}\gamma_{t-1}^k\eta_k(v_{k,i})\right| &\lesssim |v_{t,i}| + \sum_{k=1}^{t-1}\left|\gamma_{t-1}^k\right||v_{k,i}| \lesssim |v_{t,i}| + \|\gamma_{t-1}\|_2\|\widetilde{v}_{t-1,i}\|_2 \\ &\lesssim \sqrt{\frac{t\log n}{n}} + \sqrt{\frac{t^3\log n}{n}}\cdot\sqrt{\frac{t^2\log n}{n}} \lesssim \sqrt{\frac{t\log n}{n}},\end{aligned} \tag{139b}$$

  for every $1 \le i < t$, where we denote $\widetilde{v}_{t-1,i} \coloneqq (v_{1,i}, v_{2,i}, \ldots, v_{t-1,i})$. To see why (139b) is valid, we note that the first inequality applies Lemma 3, the second inequality results from the Cauchy-Schwarz inequality, whereas the last line makes use of (139a) and the fact $\|\gamma_{t-1}\|_2 \asymp \|\xi_{t-1}\|_2 \lesssim \sqrt{\frac{t^3\log n}{n}}$

(cf. (137) and (39)). In addition, given that $t \lesssim \log n/(\lambda-1)$ and $\lambda - 1 \gtrsim n^{-1/9}\log n$, we have $t^8 \lesssim n/\log n$. Repeating the argument for inequality (169) in the proof of Lemma 5, one can ensure that for any $2 \le k \le 14$,

$$\sum_{i=1}^{n} |v_{t,i}|^k = \sum_{i=1}^{t} \left|v_{t,(i)}\right|^k + \sum_{i=t+1}^{n} \left|v_{t,(i)}\right|^k \lesssim \left(\frac{\log n}{n}\right)^{k/2-1}, \tag{140a}$$

and

$$\sum_{i=1}^{n} \left| v_{t,i} + \sum_{k=1}^{t-1} \gamma_{t-1}^k \eta_k(v_{k,i}) \right|^k \lesssim \left(1 + t^{2k}\left(\frac{\log n}{n}\right)^{k/2}\right)\left(\frac{\log n}{n}\right)^{k/2-1} \lesssim \left(\frac{\log n}{n}\right)^{k/2-1}. \tag{140b}$$

Here, we have made the observation that

$$\begin{aligned}
&\sum_{i=1}^{n} \left| \sum_{k=1}^{t-1} \gamma_{t-1}^k \eta_k(v_{k,i}) \right|^k \\
&\lesssim \sum_{i=1}^{n} \|\gamma_{t-1}\|_2^k \Big(\sum_{k=1}^{t-1} \eta_k^2(v_{k,i})\Big)^{k/2} \lesssim \sum_{i=1}^{n} \left(\frac{t^3\log n}{n}\right)^{k/2} \Big(\sum_{k=1}^{t-1} v_{k,i}^2\Big)^{k/2} \\
&\lesssim \left(\frac{t^3\log n}{n}\right)^{k/2} \sum_{i=1}^{n} \Big(\sum_{k=1}^{t-1} v_{k,(i)}^2\Big)^{k/2} \\
&\lesssim \left(\frac{t^3\log n}{n}\right)^{k/2} \cdot t\left(\frac{t^2\log n}{n}\right)^{k/2} + \left(\frac{t^3\log n}{n}\right)^{k/2} \Big(\sum_{k=1}^{t-1} v_{k,(t+1)}^2\Big)^{k/2-1} \cdot \sum_{i=t+1}^{n} \Big(\sum_{k=1}^{t-1} v_{k,(i)}^2\Big) \\
&\lesssim t^{\frac{5k}{2}+1}\left(\frac{\log n}{n}\right)^{k} + t^{2k}\left(\frac{\log n}{n}\right)^{k-1} \\
&\lesssim t^{2k}\left(\frac{\log n}{n}\right)^{k-1},
\end{aligned}$$

where the second line uses the fact $\|\gamma_{t-1}\|_2 \asymp \|\xi_{t-1}\|_2 \lesssim \sqrt{\frac{t^3\log n}{n}}$ (cf. (137) and (39)) and Lemma 3, the ante-penultimate line invokes inequality (139); the penultimate line follows from the fact that $\|v_k\|_2 \lesssim 1$ (see e.g. (90)) and conditional on event $\{\phi_k\}_{k=1}^{t-1} \in \mathcal{E}$,

$$|v_{k,(t+1)}| \le \frac{|\alpha_k|}{\sqrt{n}} + \Big|\sum_{i=1}^{k-1} \beta_{k-1}^i \phi_i\Big|_{(t+1)} \lesssim \sqrt{\frac{\log n}{n}};$$

and the last line follows from the fact that $t^{k/2+1} \lesssim t^8 \lesssim n/\log n$. Therefore, combining Lemma 3 with expression (139) gives

$$\begin{aligned}
&\eta_t\left(v_t + \sum_{k=1}^{t-1} \gamma_{t-1}^k \eta_k(v_k)\right) - \eta_t(v_t) \\
&= (1-c_0)\sum_{k=1}^{t-1} \gamma_{t-1}^k \eta_k(v_k) + O\left(\pi_t^2\right) \cdot \left[\left(v_t + \sum_{k=1}^{t-1} \gamma_{t-1}^k \eta_k(v_k)\right)^3 - (v_t)^3\right] + c_x
\end{aligned} \tag{141}$$

for some vectors $c_x \in \mathbb{R}^n$, where the parameters obey $\pi_t \lesssim (\lambda-1)^{-3/4} n^{1/4}\log n$ and $c_0 \lesssim \frac{\log^4 n}{\sqrt{n(\lambda-1)^3}}$. Here, the last equation makes use of the fact that

$$\|c_x\|_2 \lesssim \left\| \frac{n\left|v_t + \sum_{k=1}^{t-1} \gamma_{t-1}^k \eta_k(v_k)\right|^5 \log^4 n}{(\lambda-1)^3} \right\|_2 + \left\| \frac{n|v_t|^5 \log^4 n}{(\lambda-1)^3} \right\|_2$$

$$\lesssim \frac{n\log^4 n}{(\lambda-1)^3}\sqrt{\left(\frac{\log n}{n}\right)^4} \lesssim \frac{\log^6 n}{n(\lambda-1)^3}, \tag{142}$$

where the property (140) is invoked with $k=10$. Next, observe that

$$\begin{aligned}
&\left\| \left(v_t + \sum_{k=1}^{t-1}\gamma_{t-1}^k \eta_k(v_k)\right)^3 - (v_t)^3 \right\|_2 \\
&= \left\| \sum_{k=1}^{t-1}\gamma_{t-1}^k \eta_k(v_k) \circ \left((v_t + \sum_{k=1}^{t-1}\gamma_{t-1}^k \eta_k(v_k))^2 + (v_t)^2 + (v_t + \sum_{k=1}^{t-1}\gamma_{t-1}^k \eta_k(v_k)) \circ v_t \right) \right\|_2 \\
&\lesssim \Big\| \max_{1\le k\le t-1} \eta_k(v_k) \circ \Big((v_t + \sum_{k=1}^{t-1}\gamma_{t-1}^k \eta_k(v_k))^2 + (v_t)^2 \Big) \Big\|_2 \cdot \sqrt{t}\|\xi_{t-1}\|_2 \\
&\lesssim \max_{1\le k\le t-1} \|\eta_k(v_k)\|_\infty \cdot \Big(\big\|(v_t + \sum_{k=1}^{t-1}\gamma_{t-1}^k \eta_k(v_k))^2\big\|_2 + \big\| (v_t)^2 \big\|_2 \Big) \cdot \sqrt{t}\|\xi_{t-1}\|_2 \\
&\lesssim \frac{t\log n}{n} \cdot \|\xi_{t-1}\|_2 \lesssim \frac{\log^4 n}{n^{1.5}(\lambda-1)^{2.5}},
\end{aligned}$$

where the last line can be obtained by invoking property (139) and (140) with $k=4$. Here, we have used the facts that $\|\xi_t\|_2 \lesssim \sqrt{\frac{t^3\log n}{n}}$ (cf. (39)), $t \lesssim \frac{\log n}{\lambda-1}$ and

$$\Big\|v_t + \sum_{k=1}^{t-1}\gamma_{t-1}^k \eta_k(v_k)\Big\|_2 \le \|v_t\|_2 + \Big\| \sum_{k=1}^{t-1}\gamma_{t-1}^k \eta_k(v_k)\Big\|_2 \le 1 + \sqrt{t}\|\gamma_{t-1}\|_2 \le 1 + \sqrt{\frac{t^4\log n}{n}} \lesssim 1.$$

Putting these together, we arrive at

$$\left\| \eta_t\left(v_t + \sum_{k=1}^{t-1}\gamma_{t-1}^k \eta_k(v_k)\right) - \eta_t(v_t) - (1-c_0)\sum_{k=1}^{t-1}\gamma_{t-1}^k \eta_k(v_k)\right\|_2 = O\left(\frac{\log^6 n}{n(\lambda-1)^3}\right).$$

- Finally, it is sufficient for us to consider $v^{\star\top}\sum_{k=1}^{t-1}\gamma_{t-1}^k\eta_k(v_k)$ which shall be controlled as follows:

$$\begin{aligned}
\left|\sum_{k=1}^{t-1}\gamma_{t-1}^k v^{\star\top}\eta_k(v_k)\right| &\le \left|\sum_{k=1}^{t-1}\gamma_{t-1}^k \left[v^{\star\top}\eta_k(x_k) + O(\|\xi_{k-1}\|_2)\right]\right| \\
&= \left|\sum_{k=1}^{t-1}\gamma_{t-1}^k \left(\frac{\alpha_{k+1}}{\lambda} + O\left(\sqrt{\frac{k^3\log n}{n}}\right)\right)\right| \\
&\lesssim \sqrt{t}\|\gamma_{t-1}\|_2 \cdot \frac{\sqrt{\lambda-1}}{n^{1/4}} \asymp \sqrt{t}\|\xi_{t-1}\|_2 \cdot \frac{\sqrt{\lambda-1}}{n^{1/4}} \lesssim \frac{\log^{2.5} n}{n^{3/4}(\lambda-1)^{1.5}}.
\end{aligned}$$

Putting the above three inequalities together yields the desired bound (138).

Built upon expression (138), we now proceed to the proof of claim (46). To begin with, let us recall that $v_t \coloneqq \alpha_t v^\star + \sum_{k=1}^{t-1}\beta_{t-1}^k\phi_k$. If we define $g_{t-1} \coloneqq v^{\star\top}\phi_{t-1} \sim \mathcal{N}\left(0, \frac{1}{n}\right)$ and $\widetilde{\beta}_{t-1} \coloneqq [\beta_{t-1}^1, \ldots, \beta_{t-1}^{t-2}]$, some direct algebra thus leads to

$$\begin{aligned}
\left|v^{\star\top}(v_t - \alpha_t v^\star - \phi_{t-1})\right| &= \left| v^{\star\top}\Big[\sum_{k=1}^{t-2}\beta_{t-1}^k\phi_k - (1-\beta_{t-1}^{t-1})\phi_{t-1}\Big]\right| \\
&\le \left|\sum_{k=1}^{t-2}\beta_{t-1}^k g_k\right| + |1-\beta_{t-1}^{t-1}||g_{t-1}|
\end{aligned}$$

$$
\begin{aligned}
&= \Big|\sum_{k=1}^{t-2}\beta_{t-1}^k g_k\Big| + \frac{1-\big(\beta_{t-1}^{t-1}\big)^2}{1+|\beta_{t-1}^{t-1}|}|\,g_{t-1}| \\
&\lesssim \|\widetilde{\beta}_{t-1}\|_2\sqrt{\sum_{k=1}^{t-2}(g_k)^2} + \|\widetilde{\beta}_{t-1}\|_2^2|g_{t-1}| \lesssim \|\widetilde{\beta}_{t-1}\|_2\sqrt{\frac{t\log n}{n}} \\
&\lesssim \sqrt{\frac{t\log n}{n}}\cdot\left(\frac{t\sqrt{\lambda-1}}{n^{1/4}} + \frac{t\log^4 n}{\sqrt{n(\lambda-1)^3}}\right) \lesssim \frac{\log^2 n}{n^{3/4}(\lambda-1)} + \frac{\log^6 n}{n(\lambda-1)^3},
\end{aligned}
$$

where the last inequality comes from the bound (165) in the proof of Lemma 5 and the condition $t \lesssim \frac{\log n}{\lambda-1}$. By virtue of the above calculations, we can deduce that

$$
v^{\star\top}v_t = \alpha_t + g_{t-1} + O\left(\frac{\log^2 n}{n^{3/4}(\lambda-1)} + \frac{\log^6 n}{n(\lambda-1)^3}\right).
$$

In fact, a direct application of Lemma 3 further leads to the following claim:

$$
\begin{aligned}
\big|v^{\star\top}\big(\eta_t(v_t)-v_t\big)\big| &\lesssim \left|v^{\star\top}\left[c_0 v_t + \pi_t^2(v_t\circ v_t\circ v_t) + O\left(\frac{n\log^4 n}{(\lambda-1)^3}|v_t|^5\right)\right]\right| \\
&\lesssim \frac{\log^4 n}{n^{3/4}(\lambda-1)},
\end{aligned} \tag{143}
$$

whose the proof of the last inequality is postponed to the end of this subsection.

To summarize, taking the above results collectively and using the relation (138), we arrive at

$$
\begin{aligned}
\alpha_{t+1} &= \lambda v^{\star\top}v_t + \lambda v^{\star\top}\big(\eta_t\left(v_t\right)-v_t\big) + O\left(\frac{\log^{2.5} n}{n^{3/4}(\lambda-1)^{1.5}}\right) \\
&= \lambda\alpha_t + \lambda g_{t-1} + O\left(\frac{\log^4 n}{n^{3/4}(\lambda-1)^{1.5}}\right).
\end{aligned} \tag{144}
$$

Therefore, invoking the above relation recursively leads to our desired decomposition:

$$
\alpha_{t+1} = \lambda^{t-k+1}\alpha_k + \sum_{i=1}^{t-k+1}\lambda^i g_{t-i} + O\left(\sum_{i=1}^{t-k+1}\lambda^i\frac{\log^4 n}{n^{3/4}(\lambda-1)^{1.5}}\right)
$$

for any $1\le k\le t$.

Proof of inequality (143). In order to establish inequality (143), let us first make note of the following simple properties: with probability at least $1-O(n^{-11})$,

$$
\begin{aligned}
v^{\star\top}(v^\star\circ v^\star\circ v^\star) &= \frac{1}{n}; \\
v^{\star\top}\Big(v^\star\circ v^\star\circ\sum_{k=1}^{t-1}\beta_{t-1}^k\phi_k\Big) &= \frac{1}{n}v^{\star\top}\Big(\sum_{k=1}^{t-1}\beta_{t-1}^k\phi_k\Big) \lesssim \frac{1}{n}; \\
v^{\star\top}\Big(v^\star\circ\sum_{k=1}^{t-1}\beta_{t-1}^k\phi_k\circ\sum_{k=1}^{t-1}\beta_{t-1}^k\phi_k\Big) &= \frac{1}{n}\Big\|\sum_{k=1}^{t-1}\beta_{t-1}^k\phi_k\Big\|_2^2 \asymp \frac{1}{n}; \\
v^{\star\top}\Big(\sum_{k=1}^{t-1}\beta_{t-1}^k\phi_k\circ\sum_{k=1}^{t-1}\beta_{t-1}^k\phi_k\circ\sum_{k=1}^{t-1}\beta_{t-1}^k\phi_k\Big) &\lesssim \frac{t\log n}{n}v^{\star\top}\Big(\sum_{k=1}^{t-1}\beta_{t-1}^k\phi_k\Big) \lesssim \sqrt{\frac{t^3\log^3 n}{n^3}}.
\end{aligned}
$$

We remind the readers that $v_i^\star \sim \mathsf{Unif}(\pm\frac{1}{\sqrt{n}})$ and we have invoked Lemma 1.

Next, recall $v_t \coloneqq \alpha_t v^\star + \sum_{k=1}^{t-1} \beta_{t-1}^k \phi_k$ to obtain

$$\begin{aligned}
\left|v^{\star\top}(v_t \circ v_t \circ v_t)\right| &= \bigg|\alpha_t^3 v^{\star\top}(v^\star \circ v^\star \circ v^\star) + 3\alpha_t^2 v^{\star\top}\Big(v^\star \circ v^\star \circ \sum_{k=1}^{t-1}\beta_{t-1}^k \phi_k\Big) \\
&\qquad + 3\alpha_t v^{\star\top}\Big(v^\star \circ \sum_{k=1}^{t-1}\beta_{t-1}^k \phi_k \circ \sum_{k=1}^{t-1}\beta_{t-1}^k \phi_k\Big) + v^{\star\top}\Big(\sum_{k=1}^{t-1}\beta_{t-1}^k \phi_k \circ \sum_{k=1}^{t-1}\beta_{t-1}^k \phi_k \circ \sum_{k=1}^{t-1}\beta_{t-1}^k \phi_k\Big)\bigg| \\
&\lesssim \frac{\alpha_t^3}{n} + \frac{\alpha_t^2}{n} + \frac{\alpha_t}{n} + \sqrt{\frac{t^3 \log^3 n}{n^3}} \\
&\lesssim \frac{1}{n^{5/4}},
\end{aligned}$$

where the last line holds as long as $\alpha_t \lesssim \sqrt{\lambda - 1}\, n^{-1/4}$. Consequently, in order to derive (143), it suffices to notice (140), $c_0 \lesssim \frac{\log^4 n}{\sqrt{n(\lambda-1)^3}}$, $\pi_t \lesssim \frac{n^{1/4}}{(\lambda-1)^{3/4}}$ and

$$\left|v^{\star\top} v_t\right| = \left|\alpha_t + v^{\star\top}\Big(\sum_{k=1}^{t-2}\beta_{t-1}^k \phi_k\Big)\right| \lesssim \frac{\sqrt{\lambda-1}}{n^{1/4}} + \sqrt{\frac{t\log n}{n}} \lesssim \frac{\sqrt{\lambda-1}}{n^{1/4}}.$$

B.5 Proof of Claim (56)

For notational simplicity, we assume without loss of generality that $\alpha_t > 0$ throughout this proof. Before delving into the proof of claim (56), let us recall Lemma 2 to obtain

$$\begin{aligned}
\frac{\pi_t}{\alpha_t\sqrt{n}} &= 1 + O\bigg(\frac{1}{\alpha_t^2}\left(\|\xi_{t-1}\|_2 + \sqrt{\frac{t\log n}{n}}\right) \wedge \frac{1}{\alpha_t}\left(\|\xi_{t-1}\|_2 + \sqrt{\frac{t\log n}{n}}\right)^{1/2}\bigg) \\
&= 1 + O\bigg(\frac{\log^2 n}{\alpha_t^2\sqrt{(\lambda-1)^3 n}} \wedge \left(\frac{\log^2 n}{\alpha_t^2\sqrt{(\lambda-1)^3 n}}\right)^{1/2}\bigg),
\end{aligned}$$

where we have used $\|\xi_t\|_2 \le \sqrt{\frac{t^3\log n}{n}}$ (see (39)) and $t \lesssim \frac{\log n}{\lambda - 1}$. In turn, this implies

$$\left|\frac{\pi_t^2}{\alpha_t^2 n} - 1\right| \alpha_t^2 \lesssim \frac{\log^2 n}{\sqrt{(\lambda-1)^3 n}}. \tag{145}$$

Now, let us move on to establish a recursive relation of α_t. Recalling the definition (3) of η_t and Theorem 1, one sees that

$$\begin{aligned}
\alpha_{t+1} &= \lambda v^{\star\top}\int \eta_t\left(\alpha_t v^\star + \frac{1}{\sqrt{n}}x\right)\varphi_n(\mathrm{d}x) + \Delta_{\alpha,t} \\
&= \lambda\gamma_t v^{\star\top}\int \tanh\left(\pi_t\big(\alpha_t v^\star + \frac{1}{\sqrt{n}}x\big)\right)\varphi_n(\mathrm{d}x) + \Delta_{\alpha,t} \\
&= \lambda\gamma_t\sqrt{n}\int \tanh\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x)\right)\varphi(\mathrm{d}x) + \Delta_{\alpha,t},
\end{aligned} \tag{146}$$

where the last equality holds by symmetry of $\varphi(\cdot)$, namely,

$$\frac{1}{\sqrt{n}}\int \tanh\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x)\right)\varphi(\mathrm{d}x) = -\frac{1}{\sqrt{n}}\int \tanh\left(\frac{\pi_t}{\sqrt{n}}(-\alpha_t + x)\right)\varphi(\mathrm{d}x).$$

We note that similar analysis as for relation (116) leads to $\int \tanh^2\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x)\right)\varphi(\mathrm{d}x) \asymp \frac{\pi_t^2}{n}$. Combining this result with Lemma 2 and (39), we arrive at

$$\gamma_t^{-2} = n\left(1 + O\left(\sqrt{\frac{t^3\log n}{n}}\right)\right)\int \tanh^2\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x)\right)\varphi(\mathrm{d}x). \tag{147}$$

Taking (146) and (147) together, we arrive at

$$\alpha_{t+1} = \Big(1 + O\Big(\sqrt{\frac{t^3 \log n}{n}}\Big)\Big) \frac{\lambda \int \tanh\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x)\right) \varphi(\mathrm{d}x)}{\left[\int \tanh^2\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x)\right) \varphi(\mathrm{d}x)\right]^{1/2}} + \Delta_{\alpha,t}. \tag{148}$$

To prove claim (56), it then suffices to control $\int \tanh\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x)\right) \varphi(\mathrm{d}x)$ and $\int \tanh^2\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x)\right) \varphi(\mathrm{d}x)$. Towards this goal, we find it helpful to first make several observations. Define two functions:

$$f(z) := \frac{1}{z}\tanh(zy) - \tanh(y),$$
$$g(z) := \frac{1}{z^2}\tanh^2(zy) - \tanh^2(y).$$

The Taylor expansion of $\tanh(zy)$ gives

$$f'(z) = -\frac{1}{z^2}\big[\tanh(zy) - zy + zy\tanh^2(zy)\big] = -\frac{2}{3}zy^3 + O(z^3y^5),$$
$$g'(z) = -\frac{2\tanh(zy)}{z^3}\big[\tanh(zy) - zy + zy\tanh^2(zy)\big] = \frac{1}{3}zy^4 + O(z^3y^6),$$

which leads the following relation by direct calculation

$$f(z) = \int_1^z f'(t)dt = -\frac{1}{3}(z^2-1)y^3 + O((z^4-1)y^5),$$
$$g(z) = \int_1^z g'(t)dt = \frac{1}{6}(z^2-1)y^4 + O((z^4-1)y^6).$$

By taking $z = \frac{\pi_t}{\alpha_t\sqrt{n}}$, $y = \alpha_t(\alpha_t + x)$, we can see that

$$\frac{\alpha_t\sqrt{n}}{\pi_t}\tanh\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x)\right) - \tanh\left(\alpha_t(\alpha_t + x)\right) = -\frac{1}{3}\left(\frac{\pi_t^2}{\alpha_t^2 n} - 1\right)\alpha_t^3(\alpha_t + x)^3 + O\left(\frac{\pi_t^4}{\alpha_t^4 n^2} - 1\right)\alpha_t^5(\alpha_t + x)^5,$$
$$\frac{\alpha_t^2 n}{\pi_t^2}\tanh^2\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x)\right) - \tanh^2\left(\alpha_t(\alpha_t + x)\right) = \frac{1}{6}\left(\frac{\pi_t^2}{\alpha_t^2 n} - 1\right)\alpha_t^4(\alpha_t + x)^4 + O\left(\frac{\pi_t^4}{\alpha_t^4 n^2} - 1\right)\alpha_t^6(\alpha_t + x)^6.$$

Hence, we can conclude that

$$\begin{aligned}
&\left|\int \frac{\alpha_t\sqrt{n}}{\pi_t}\tanh\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x)\right) - \tanh\left(\alpha_t(\alpha_t + x)\right)\varphi(\mathrm{d}x)\right| \\
&= \left|\int \frac{1}{3}\left(\frac{\pi_t^2}{\alpha_t^2 n} - 1\right)\alpha_t^3(\alpha_t + x)^3\varphi(\mathrm{d}x)\right| + \int O\left(\frac{\pi_t^4}{\alpha_t^4 n^2} - 1\right)\alpha_t^5(\alpha_t + x)^5\varphi(\mathrm{d}x) \\
&\lesssim \left|\frac{\pi_t^2}{\alpha_t^2 n} - 1\right|\alpha_t^4.
\end{aligned}$$

Similarly, we can show that

$$\left|\int \frac{\alpha_t^2 n}{\pi_t^2}\tanh^2\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x)\right) - \tanh^2\left(\alpha_t(\alpha_t + x)\right)\varphi(\mathrm{d}x)\right| \lesssim \left|\frac{\pi_t^2}{\alpha_t^2 n} - 1\right|\alpha_t^4. \tag{149}$$

Substituting these relations into (148), we arrive at

$$\alpha_{t+1} = \lambda\Big(1 + O\Big(\sqrt{\frac{t^3 \log n}{n}}\Big)\Big) \frac{\int \tanh\left(\alpha_t(\alpha_t + x)\right)\varphi(\mathrm{d}x) + O\left(\left|\frac{\pi_t^2}{\alpha_t^2 n} - 1\right|\alpha_t^4\right)}{\left[\int \tanh^2\left(\alpha_t(\alpha_t + x)\right)\varphi(\mathrm{d}x) + O\left(\left|\frac{\pi_t^2}{\alpha_t^2 n} - 1\right|\alpha_t^4\right)\right]^{1/2}} + \Delta_{\alpha,t}$$

$$= \lambda\Big(1+O\Big(\sqrt{\frac{t^3\log n}{n}}\Big)\Big)\frac{\int \tanh\left(\alpha_t\left(\alpha_t+x\right)\right)\varphi(\mathrm{d}x)}{\left[\int \tanh^2\left(\alpha_t\left(\alpha_t+x\right)\right)\varphi(\mathrm{d}x)\right]^{1/2}}\cdot\frac{1+O\left(\left|\frac{\pi_t^2}{\alpha_t^2 n}-1\right|\alpha_t^2\right)}{1+O\left(\left|\frac{\pi_t^2}{\alpha_t^2 n}-1\right|\alpha_t^2\right)}+\Delta_{\alpha,t}$$

$$= \lambda\left[\int \tanh^2\left(\alpha_t^2+\alpha_t x\right)\varphi(\mathrm{d}x)\right]^{-1/2}\int \tanh\left(\alpha_t^2+\alpha_t x\right)\varphi(\mathrm{d}x)+O\left(\left|\frac{\pi_t^2}{\alpha_t^2 n}-1\right|\alpha_t^3+\sqrt{\frac{t^3\log n}{n}}\alpha_t+\Delta_{\alpha,t}\right), \tag{150}$$

where we make use of the fact that (see, (116))

$$\int \tanh\left(\alpha_t^2+\alpha_t x\right)\varphi(\mathrm{d}x)=\int \tanh^2\left(\alpha_t^2+\alpha_t x\right)\varphi(\mathrm{d}x)\asymp\alpha_t^2.$$

In addition, (133) ensures that

$$\frac{\alpha_t}{\sqrt{\alpha_t^2+1}}\le\sqrt{\int\tanh^2(\alpha_t(\alpha_t+x))\varphi(\mathrm{d}x)}=\left[\int \tanh^2\left(\alpha_t^2+\alpha_t x\right)\varphi_n(\mathrm{d}x)\right]^{-1/2}\int \tanh\left(\alpha_t^2+\alpha_t x\right)\varphi(\mathrm{d}x),$$

which in turn gives

$$\alpha_{t+1}\ge\frac{\lambda\alpha_t}{\sqrt{\alpha_t^2+1}}+O\left(\left|\frac{\pi_t^2}{\alpha_t^2 n}-1\right|\alpha_t^3+\sqrt{\frac{t^3\log n}{n}}\alpha_t+|\Delta_{\alpha,t}|\right).$$

To finish up, putting the above results together with (145) leads to

$$\alpha_{t+1}\ge\frac{\lambda\alpha_t}{\sqrt{\alpha_t^2+1}}+o\big((\lambda-1)\alpha_t\big)+O(\Delta_{\alpha,t}) \tag{151}$$

where we again invoke the assumption that $\lambda-1\gtrsim n^{-1/9}\log n$. This concludes the proof of claim (56).

B.6 Proof of Claim (59)

Consider the regime where

$$|\alpha_t|<(\lambda-1)^{-3/4}n^{-1/4}\lesssim\sqrt{\lambda-1}\,n^{-0.1}.$$

First, invoke property (25b) in Lemma 3 to ensure that

$$\left|v^{\star\top}\eta_t(x_t)-v^{\star\top}\eta_t(v_t)\right|\lesssim|v^{\star\top}\xi_{t-1}|+(\lambda-1)n^{-0.2}(\log n)\|\xi_{t-1}\|_2. \tag{152}$$

To further bound (152), note that ξ_{t-1} admits the following decomposition in terms of $\{\eta_k(x_k)\}$:

$$\xi_{t-1}=\sum_{k=1}^{t-2}\gamma_{t-1}^k\eta_k(x_k),\qquad\text{with }\gamma_{t-1}=[\gamma_{t-1}^k]_{1\le k\le t-2}\in\mathbb{R}^{t-2}\text{ obeying }\|\gamma_{t-1}\|_2\lesssim\|\xi_{t-1}\|_2;$$

the proof of this claim can be found in Section B.4 (see Lemma 5 therein and its proof). In view of this relation, we can apply (152) and the Cauchy-Schwarz inequality to reach

$$\begin{aligned}\left|v^{\star\top}\eta_t(x_t)-v^{\star\top}\eta_t(v_t)\right|&\lesssim\left|\sum_{k=1}^{t-2}\gamma_{t-1}^k v^{\star\top}\eta_k(x_k)\right|+(\lambda-1)n^{-0.2}(\log n)\|\xi_{t-1}\|_2\\&\lesssim\|\gamma_{t-1}\|_2\left(\sum_{k=1}^{t-2}\left(v^{\star\top}\eta_k(x_k)\right)^2\right)^{1/2}+(\lambda-1)n^{-0.2}(\log n)\|\xi_{t-1}\|_2\\&\lesssim\sqrt{t}\|\xi_{t-1}\|_2\max_{\tau_0\le s\le t}|\alpha_s|+(\lambda-1)n^{-0.2}(\log n)\|\xi_{t-1}\|_2\lesssim\sqrt{\frac{t\log n}{n}},\end{aligned}$$

provided that $t \lesssim \frac{\log n}{\lambda-1}$ and $\max_{\tau_0 \le s \le t} |\alpha_s| \lesssim \sqrt{\lambda-1}n^{-0.1}$. Here the last inequality invokes $\|\xi_t\|_2 \lesssim \sqrt{\frac{t^3 \log n}{n}}$ (see (39)). Substitution into (55) yields

$$|\Delta_{\alpha,t}| \lesssim \sqrt{\frac{t \log n}{n}} \ll (\lambda-1)|\alpha_t|, \tag{153}$$

given $|\alpha_t| \gtrsim \sqrt{\lambda-1}n^{-1/4}$ and $\lambda - 1 \gtrsim n^{-1/9}\log n$. It thus completes the proof of the relation (59).

B.7 Proof of Claim (62)

Throughout this section, we assume without loss of generality that $\alpha_t > 0$. As computed in Section B.5 for relation (56), applying Lemma 2 reveals that

$$\left|\frac{\pi_t^2}{\alpha_t^2 n} - 1\right| \alpha_t^2 \lesssim \left(\|\xi_{t-1}\|_2 + \sqrt{\frac{t\log n}{n}}\right) \wedge \alpha_t \left(\|\xi_{t-1}\|_2 + \sqrt{\frac{t\log n}{n}}\right)^{1/2} \lesssim \alpha_t \left(\frac{(\lambda-1)^3}{\log n}\right)^{1/4},$$

$$\gamma_t^{-2} = n\left(1 + O\left(\sqrt{\frac{(\lambda-1)^3}{\log n}}\right)\right) \int \tanh^2\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x)\right) \varphi(\mathrm{d}x),$$

$$|\Delta_{\alpha,t}| \lesssim \|\xi_{t-1}\|_2 + \sqrt{\frac{t\log n}{n}} \lesssim \sqrt{\frac{(\lambda-1)^3}{\log n}},$$

given the inductive assumptions $\|\xi_{t-1}\|_2 \lesssim \sqrt{\frac{(\lambda-1)^3}{\log n}}$ when $t \lesssim \frac{n(\lambda-1)^5}{\log^2 n}$ and $\alpha_t \gtrsim \sqrt{\lambda^2-1}$. Now in view of similar calculations for (150), we can deduce that

$$\begin{aligned}
\alpha_{t+1} &= \lambda\left[\int \tanh\left(\alpha_t^2 + \alpha_t x\right)\varphi_n(\mathrm{d}x)\right]^{1/2} + O\left(\left|\frac{\pi_t^2}{\alpha_t^2 n} - 1\right|\alpha_t^3 + \frac{\lambda-1}{\sqrt{\log n}}\alpha_t + \Delta_{\alpha,t}\right) \\
&\ge \frac{\lambda\alpha_t}{\sqrt{1+\alpha_t^2}} + o\left((\lambda-1)^{3/4}\alpha_t^2 + (\lambda-1)\alpha_t\right)
\end{aligned} \tag{154}$$

where we have made use of the fact that $\sqrt{\frac{(\lambda-1)^3}{\log n}} \ll (\lambda-1)\alpha_t$.

We then demonstrate that this relation (154) together with a little algebra indicates that $\alpha_{t+1} \ge \frac{1}{2}\sqrt{\lambda^2-1}$. Specifically, consider the following two cases separately.

- First, consider the case where $\alpha_t \le \frac{2}{3}\sqrt{\lambda^2-1}$. Akin to inequality (57), relation (154) implies the existence of some constant $c > 0$ such that

$$\alpha_{t+1} \ge \big(1 + c(\lambda-1)\big)\alpha_t + o\big((\lambda-1)\alpha_t\big) \ge \alpha_t \ge \frac{1}{2}\sqrt{\lambda^2-1}.$$

- Otherwise, consider the case where $\alpha_t > \frac{2}{3}\sqrt{\lambda^2-1}$. Recognizing the fact that $\frac{\lambda\alpha_t}{\sqrt{1+\alpha_t^2}}$ is monotonically increasing in α_t, we arrive at

$$\alpha_{t+1} \ge \big(1 + c(\lambda-1)\big)\frac{2}{3}\sqrt{\lambda^2-1} + o\big((\lambda-1)^{3/4}\big) \ge \frac{1}{2}\sqrt{\lambda^2-1}.$$

Thus, this completes the proof of our desired bound (28).

B.8 Proof of Lemma 5

Throughout the proof, we work with the event that $\{\phi_k\}_{k=1}^{t-1} \in \mathcal{E}$ (defined in Lemma 1 with $\delta = O(n^{-11})$), which holds true with probability at least $1 - O(n^{-11})$. On this event, one has $\|\phi_t\|_\infty \lesssim \sqrt{\frac{\log n}{n}}$ and

$$\|x_t\|_\infty \le \|\alpha_t v^\star\|_\infty + \Big\|\sum_{k=1}^{t-1}\beta_{t-1}^k\phi_k\Big\|_\infty + \|\xi_{t-1}\|_2 \lesssim \frac{|\alpha_t|}{\sqrt{n}} + \sqrt{\frac{t\log n}{n}} + \sqrt{\frac{t^3\log n}{n}} \lesssim \sqrt{\frac{\log^4 n}{n(\lambda-1)^3}} \tag{155}$$

for any $t \le \tau_0$, where we remind the readers that (see (39), the property $\|\beta_{t-1}\|_2 = 1$, the definition of $\mathcal{E}_1$, and the definition (44) of τ_0)

$$|\alpha_t| \lesssim \sqrt{\lambda - 1}\, n^{-1/4}, \qquad \Big\| \sum_{k=1}^{t-1} \beta_{t-1}^k \phi_k \Big\|_\infty \lesssim \sqrt{\frac{t \log n}{n}} \qquad \text{and} \quad \|\xi_{t-1}\|_2 \lesssim \sqrt{\frac{t^3 \log n}{n}} \tag{156}$$

as long as $t \lesssim \frac{\log n}{\lambda - 1}$ (see (45)).

To show that $\{\eta_1(x_1), \ldots, \eta_t(x_t)\}$ forms a near-orthogonal basis, one strategy is to show that $\eta_i(x_i) \approx \phi_{i-1}$ for each $1 \le i \le t \le \tau_0$, which in turn implies that

$$\Big\| \sum_{i=1}^{t} w_i \eta_i(x_i) \Big\|_2 \approx \Big\| \sum_{i=1}^{t} w_i \phi_{i-1} \Big\|_2 \asymp \|w\|_2, \tag{157}$$

given that $\phi_k \sim \mathcal{N}(0, \frac{1}{n} I_n)$ are independent Gaussian vectors; here, we introduce $\phi_0 := x_1 \sim \mathcal{N}(0, \frac{1}{n} I_n)$ for notational convenience. Guided by this intuition, we first use the triangle inequality to derive that

$$\begin{aligned} &\Big| \Big\| \sum_{i=1}^{t} w_i \eta_i(x_i) \Big\|_2 - \Big\| \sum_{i=1}^{t} w_i \beta_{i-1}^{i-1} \phi_{i-1} \Big\|_2 \Big| \\ &\qquad \le \Big\| \sum_{i=1}^{t} w_i \Big[\eta_i(x_i) - \eta_i(\beta_{i-1}^{i-1} \phi_{i-1}) \Big] \Big\|_2 + \Big\| \sum_{i=1}^{t} w_i \Big[\eta_i(\beta_{i-1}^{i-1} \phi_{i-1}) - \beta_{i-1}^{i-1} \phi_{i-1} \Big] \Big\|_2 \end{aligned} \tag{158}$$

for any vector $w \in \mathbb{R}^t$. In order to bound these terms, we proceed with the following three steps.

- Let us first consider the difference between $\eta_t(x_t)$ and x_t. Invoking property (25c) in Lemma 3 allows us to express

$$\eta_t(x_t) = (1 - c_0)\Big(x_t - \frac{\pi_t^2}{3} x_t \circ x_t \circ x_t + c_{x_t} \Big),$$

where $c_0 \lesssim \frac{\log^2 n}{\sqrt{n(\lambda-1)^3}}$, $\pi_t \lesssim \frac{n^{1/4} \log n}{(\lambda-1)^{3/4}}$, and c_{x_t} is a vector obeying (cf. (155))

$$\|c_{x_t}\|_\infty \lesssim \frac{n \|x_t\|_\infty^5 \log^4 n}{(\lambda - 1)^3} \lesssim \frac{\log^{14} n}{n^{3/2} (\lambda - 1)^{10.5}}. \tag{159}$$

Then one has

$$\|\eta_t(x_t) - x_t\|_2 \lesssim c_0 \|x_t\|_2 + \frac{1}{3} \pi_t 2 \|x_t \circ x_t \circ x_t\|_2 + \sqrt{n}\, \|c_{x_t}\|_\infty. \tag{160}$$

We then claim that

$$\|\eta_t(x_t) - x_t\|_2 \lesssim \frac{\log^3 n}{\sqrt{n(\lambda - 1)^3}}; \tag{161}$$

to streamline the presentation, the proof of this claim is deferred to the end of this section. Applying the same argument once gain also leads to

$$\big\|\eta_t(\beta_{t-1}^{t-1} \phi_{t-1}) - \beta_{t-1}^{t-1} \phi_{t-1}\big\|_2 \lesssim \frac{\log^3 n}{\sqrt{n(\lambda - 1)^3}}. \tag{162}$$

- Combining relations (161) and (162) and invoking the triangle inequality, we obtain

$$\begin{aligned} \big\|\eta_t(x_t) - \eta_t\big(\beta_{t-1}^{t-1} \phi_{t-1}\big)\big\|_2 &\le \|\eta_t(x_t) - x_t\|_2 + \big\|x_t - \beta_{t-1}^{t-1} \phi_{t-1}\big\|_2 + \big\|\eta_t(\beta_{t-1}^{t-1} \phi_{t-1}) - \beta_{t-1}^{t-1} \phi_{t-1}\big\|_2 \\ &= \big\|x_t - \beta_{t-1}^{t-1} \phi_{t-1}\big\|_2 + O\left(\frac{\log^3 n}{\sqrt{n(\lambda - 1)^3}} \right) \end{aligned}$$

$$
\begin{aligned}
&= \Big\|\alpha_t v^\star + \sum_{k=1}^{t-2} \beta_{t-1}^k \phi_k + \xi_t\Big\|_2 + O\left(\frac{\log^3 n}{\sqrt{n(\lambda-1)^3}}\right) \\
&\le \|\alpha_t v^\star\|_2 + \Big\|\sum_{k=1}^{t-2} \beta_{t-1}^k \phi_k\Big\|_2 + \|\xi_t\|_2 + O\left(\frac{\log^3 n}{\sqrt{n(\lambda-1)^3}}\right) \\
&\le \Big\|\sum_{k=1}^{t-2} \beta_{t-1}^k \phi_k\Big\|_2 + |\alpha_t| + O\left(\sqrt{\frac{t^3\log n}{n}}\right) + O\left(\frac{\log^3 n}{\sqrt{n(\lambda-1)^3}}\right), \qquad (163)
\end{aligned}
$$

where the last line makes use of (156). Let us denote $\widetilde{\beta}_{t-1} \coloneqq (\beta_{t-1}^1, \ldots, \beta_{t-1}^{t-2}) \in \mathbb{R}^{t-2}$, obtained by removing the last entry of β_{t-1}. According to Lemma 1, we note that with probability $1 - O(n^{-11})$,

$$
\left|\Big\|\sum_{k=1}^{t-2} \beta_{t-1}^k \phi_k\Big\|_2 - \|\widetilde{\beta}_{t-1}\|_2\right| \le \|\widetilde{\beta}_{t-1}\|_2 \cdot \sup_{a=[a_k]_{1\le k<t-2}\in\mathcal{S}^{t-3}} \left|\Big\|\sum_{k=1}^{t-2} a_k\phi_k\Big\|_2 - 1\right| \lesssim \sqrt{\frac{t\log n}{n}}\|\widetilde{\beta}_{t-1}\|_2,
$$

which in turn implies that

$$
\left(1 - O\Big(\sqrt{\frac{t\log n}{n}}\Big)\right)\|\widetilde{\beta}_{t-1}\|_2 \le \Big\|\sum_{k=1}^{t-2} \beta_{t-1}^k \phi_k\Big\|_2 \le \left(1 + O\Big(\sqrt{\frac{t\log n}{n}}\Big)\right)\|\widetilde{\beta}_{t-1}\|_2.
$$

As a consequence, we can further control the right-hand side of (163) by

$$
\big\|\eta_t(x_t) - \eta_t\big(\beta_{t-1}^{t-1}\phi_{t-1}\big)\big\|_2 \le \left(1 + O\Big(\sqrt{\frac{t\log n}{n}}\Big)\right)\|\widetilde{\beta}_{t-1}\|_2 + O\left(|\alpha_t| + \frac{\log^3 n}{\sqrt{n(\lambda-1)^3}}\right). \qquad (164)
$$

- Our next step is concerned with bounding the term $\|\widetilde{\beta}_{t-1}\|_2$. First, recall that β_t corresponds to the linear coefficients of $\eta_t(x_t)$ when projected to the linear space U_t (see (42) and (41b)). We can thus write

$$
\begin{aligned}
\|\widetilde{\beta}_t\|_2 &= \big\|U_{t-1}^\top \eta_t(x_t)\big\|_2 \\
&\le \big\|U_{t-1}^\top\big[\eta_t(x_t) - \eta_t\big(\beta_{t-1}^{t-1}\phi_{t-1}\big)\big]\big\|_2 + \big\|U_{t-1}^\top\big[\eta_t\big(\beta_{t-1}^{t-1}\phi_{t-1}\big) - \beta_{t-1}^{t-1}\phi_{t-1}\big]\big\|_2 + \big\|U_{t-1}^\top\big(\beta_{t-1}^{t-1}\phi_{t-1}\big)\big\|_2 \\
&\le \left(1 + O\Big(\sqrt{\frac{t\log n}{n}}\Big)\right)\|\widetilde{\beta}_{t-1}\|_2 + O\left(|\alpha_t| + \frac{\log^3 n}{\sqrt{n(\lambda-1)^3}}\right) + O\left(\frac{\log^3 n}{\sqrt{n(\lambda-1)^3}}\right) + \big\|U_{t-1}^\top\big(\beta_{t-1}^{t-1}\phi_{t-1}\big)\big\|_2 \\
&= \left(1 + O\Big(\sqrt{\frac{t\log n}{n}}\Big)\right)\|\widetilde{\beta}_{t-1}\|_2 + O\left(|\alpha_t| + \frac{\log^3 n}{\sqrt{n(\lambda-1)^3}}\right) + O\left(\sqrt{\frac{t\log n}{n}}\right) \\
&= \left(1 + O\Big(\sqrt{\frac{t\log n}{n}}\Big)\right)\|\widetilde{\beta}_{t-1}\|_2 + O\left(|\alpha_t| + \frac{\log^3 n}{\sqrt{n(\lambda-1)^3}}\right).
\end{aligned}
$$

Here, the third line follows from (162) and (164); the penultimate line holds due to the independence between ϕ_{t-1} and U_{t-1} (see the properties below display (43)) and hence $U_{t-1}^\top\phi_{t-1} \sim \mathcal{N}(0, \frac{1}{n}I_t)$; and the last line holds as long as $t \lesssim \frac{\log n}{\lambda-1}$. Recognizing that $\|\widetilde{\beta}_1\|_2 = \sqrt{1 - \|\beta_1\|_2^2} = 0$, we can apply the above relation recursively to yield

$$
\begin{aligned}
\|\widetilde{\beta}_t\|_2 &\le \sum_{\tau=1}^{t} \left(1 + O\Big(\sqrt{\frac{t\log n}{n}}\Big)\right)^{t-1-\tau} \cdot O\left(|\alpha_\tau| + \frac{\log^3 n}{\sqrt{n(\lambda-1)^3}}\right) \\
&\lesssim \sum_{\tau=1}^{t} \left(|\alpha_\tau| + \frac{\log^3 n}{\sqrt{n(\lambda-1)^3}}\right) \lesssim \frac{t\sqrt{\lambda-1}}{n^{1/4}} + \frac{t\log^3 n}{\sqrt{n(\lambda-1)^3}},
\end{aligned}
$$

where the penultimate inequality follows from the fact $\sqrt{\frac{t\log n}{n}} \lesssim \frac{1}{t}$ as $t \lesssim \frac{\log n}{\lambda-1}$, and the last inequality uses $|\alpha_t| \lesssim \sqrt{\lambda-1}\, n^{-1/4}$ for $t \le \tau_0$ (see the definition of τ_0 in (44)). Under the assumption that

$\lambda - 1 \gtrsim n^{-1/9} \log n$, we can further obtain

$$\|\widetilde{\beta}_t\|_2 \lesssim \frac{t\sqrt{\lambda-1}}{n^{1/4}}. \tag{165}$$

Plugging this relation into (164) and using the condition $|\alpha_t| \lesssim \sqrt{\lambda-1}\, n^{-1/4}$ $(\forall t \le \tau_0)$ give

$$\left\|\eta_t\left(x_t\right) - \eta_t\left(\beta_{t-1}^{t-1}\phi_{t-1}\right)\right\|_2 \lesssim \frac{t\sqrt{\lambda-1}}{n^{1/4}}. \tag{166}$$

It is worth noting that the inequality (165) also implies

$$|\beta_t^t| = \sqrt{\|\beta_t\|_2^2 - \|\widetilde{\beta}_t\|_2^2} = \sqrt{1 - \|\widetilde{\beta}_t\|_2^2} = 1 - O\left(\frac{t^2(\lambda-1)}{n^{1/2}}\right).$$

To finish up, putting the above bounds together with expression (158), we conclude that

$$\begin{aligned}
&\left|\left\|\sum_{i=1}^t w_i\eta_i(x_i)\right\|_2 - \left\|\sum_{i=1}^t w_i\phi_{i-1}\right\|_2\right| \\
&\le (1-\beta_{t-1}^{t-1})\left\|\sum_{i=1}^t w_i\phi_{i-1}\right\|_2 + \left\|\sum_{i=1}^t w_i\Big[\eta_i(x_i) - \eta_i(\beta_{i-1}^{i-1}\phi_{i-1})\Big]\right\|_2 + \left\|\sum_{i=1}^t w_i\Big[\eta_i(\beta_{i-1}^{i-1}\phi_{i-1}) - \beta_{i-1}^{i-1}\phi_{i-1}\Big]\right\|_2 \\
&= O\left(\frac{t^2(\lambda-1)}{n^{1/2}}\right)\left\|\sum_{i=1}^t w_i\phi_{i-1}\right\|_2 + O\left(\frac{t\sqrt{\lambda-1}}{n^{1/4}}\right)\cdot\sqrt{t}\|w\|_2 + O\left(\frac{\log^3 n}{\sqrt{n(\lambda-1)^3}}\right)\cdot\sqrt{t}\|w\|_2 \\
&= O\left(\frac{t^{3/2}\sqrt{\lambda-1}}{n^{1/4}} + \sqrt{\frac{t\log^6 n}{n(\lambda-1)^3}}\right)\|w\|_2 = o(\|w\|_2),
\end{aligned}$$

where the last relation results from the facts $t < \tau_0 \lesssim \log n/(\lambda-1)$ and the assumption $\lambda - 1 \ge n^{-1/9}$. Finally, observing $\|\sum_{i=1}^t w_i\phi_{i-1}\|_2 = (1 + O(\sqrt{\frac{t\log n}{n}}))\|w\|_2$, we reach

$$\left\|\sum_{i=1}^t w_i\eta_i(x_i)\right\|_2 = \big(1+o(1)\big)\|w\|_2,$$

which completes the proof of our desired bound.

Proof of inequality (161). For any fixed integer $k \ge 2$ that does not scale with n, we can write

$$\sum_{i=1}^n |x_{t,i}|^k \lesssim \sum_{i=1}^n |\alpha_t v_i^\star|^k + \sum_{i=1}^n |u_{t,i}|^k + \sum_{i=1}^n |\xi_{t-1,i}|^k, \qquad \text{with } \; u_t \coloneqq \sum_{k=1}^{t-1}\beta_{t-1}^k\phi_k.$$

Let us bound each term separately. Firstly, recalling that $\|\xi_{t-1}\|_2 \lesssim \sqrt{\frac{t^3\log n}{n}}$ (cf. (39)) gives

$$\sum_{i=1}^n |\xi_{t-1,i}|^k \le \|\xi_{t-1}\|_2^k \lesssim \left(\frac{t^3\log n}{n}\right)^{k/2}. \tag{167}$$

Secondly, on the event $\{\phi_k\}_{k=1}^{t-1} \in \mathcal{E}$ (see Lemma 1), we see that $\|u_t\|_2 \lesssim 1$ and $\|u_t\|_\infty \le \sqrt{(t\log n)/n}$. This in turn gives

$$\sum_{i=1}^n |u_{t,i}|^k = \sum_{i\le t} |u_{t,(i)}|^k + \sum_{i>t} |u_{t,(i)}|^k \overset{(*)}{\lesssim} t\left(\frac{t\log n}{n}\right)^{k/2} + \left(\frac{\log n}{n}\right)^{k/2-1}\sum_{i>t}|u_{t,(i)}|^2$$

$$\lesssim t\left(\frac{t\log n}{n}\right)^{k/2}+\left(\frac{\log n}{n}\right)^{k/2-1}$$
$$\lesssim \left(1+\frac{t^{k/2+1}\log n}{n}\right)\left(\frac{\log n}{n}\right)^{k/2-1}, \tag{168}$$

with $x_{(i)}$ denoting the i-th largest entry of x (in magnitude). Here, to see why inequality $(*)$ holds, we recall that on the event $\{\phi_k\}_{k=1}^{t-1}\in\mathcal{E}$ (see Lemma 1), one has

$$\sup_{a\in\mathcal{S}^{t-2}}\sum_{i=1}^{t}\left|\sum_{k=1}^{t-1}a_k\phi_k\right|_{(i)}^2\lesssim\frac{t\log n}{n},$$

which also guarantees that $|u_{t,(t+1)}|\lesssim\sqrt{\log n/n}$. Putting the above pieces together, we obtain

$$\sum_{i=1}^{n}|x_{t,i}|^k\lesssim n\left(\frac{1}{n(\lambda-1)}\right)^{3k/4}+\left(1+\frac{t^{k/2+1}\log n}{n}\right)\left(\frac{\log n}{n}\right)^{k/2-1}+\left(\frac{t^3\log n}{n}\right)^{k/2}\lesssim\left(\frac{\log n}{n}\right)^2, \tag{169}$$

where the last inequality is valid if we take $k=6$, $t\lesssim\log n/(\lambda-1)$ and assume $\lambda-1\gtrsim n^{-1/9}\log n$. It therefore leads to

$$\pi_t^2\|x_t\circ x_t\circ x_t\|_2=\pi_t^2\Big(\sum_{i=1}^{n}|x_{t,i}|^6\Big)^{1/2}\lesssim\frac{n^{1/2}\log^2 n}{(\lambda-1)^{3/2}}\cdot\frac{\log n}{n}=\frac{\log^3 n}{\sqrt{n(\lambda-1)^3}},$$

where we have used the bound on π_t in Lemma 3 (the 3rd case). This together with (159), (160) and the fact $c_0\lesssim\frac{\log^2 n}{\sqrt{n(\lambda-1)^3}}$ concludes the proof of inequality (161).

B.9 Proof of inequality (73)

Before proceeding, let us make several observations about $\frac{\tau_t}{h(\tau_t)}$. As discussed around [2, display (254)], the sequence τ_t with $\tau_{t+1}=\lambda^2h(\tau_t)$ is monotonically increasing, which implies that $\frac{\tau_t}{h(\tau_t)}\le\lambda^2$. In addition, the optimality of the Bayes estimator (cf. (133)) implies that $\frac{\tau_t}{h(\tau_t)}\le\tau_t+1$. Combining these two observations, we obtain

$$\frac{\tau_t}{h(\tau_t)}\le(\tau_t+1)\wedge\lambda^2.$$

In view of the inductive assumption, we have $\alpha_t^2=(1+o(1))\tau_t$ for $t\gtrsim\varsigma$. Hence, for every τ obeying $\min\{\tau_t,\alpha_t^2\}\le\tau\le\max\{\tau_t,\alpha_t^2\}$, it holds that $\tau=(1+o(1))\tau_t$ with $\tau_t\gtrsim\lambda^2-1$. Define $\mathcal{T}_2$ as in display (263) of [2] such that

$$\mathcal{T}_2(s,\tau)\coloneqq s^2h'(\tau)=s^2\int\left(1+\frac{x}{2\sqrt{\tau}}\right)\left(1-\tanh^2\left(\tau+\sqrt{\tau}x\right)\right)\varphi(\mathrm{d}x). \tag{170}$$

Armed with this notation, we can bound the target quantity as

$$\frac{\tau_t}{h(\tau_t)}h'(\tau)\le\mathcal{T}_2(\sqrt{\tau_t+1}\wedge\lambda,\tau).$$

Therefore, it suffices to upper bound the right-hand side of the above inequality by $1-c(\lambda-1)$.

Towards this end, direct calculations yield

$$\frac{\mathcal{T}_2(\sqrt{\tau_t+1}\wedge\lambda,\tau)-\mathcal{T}_2(\sqrt{\tau+1}\wedge\lambda,\tau)}{\mathcal{T}_2(\sqrt{\tau+1}\wedge\lambda,\tau)}=(\sqrt{\tau+1}\wedge\lambda)^2-(\sqrt{\tau_t+1}\wedge\lambda)^2=o(\lambda-1). \tag{171}$$

Moreover, it has been proved numerically (see Figure 1 in [2]) that

$$\mathcal{T}_2(\lambda,\tau)\le1-(\lambda-1),\qquad\text{for }\lambda\in(0,1.2]\text{ and }\tau>\sqrt{\lambda^2-1}.$$

Recognizing that $\tau = (1+o(1))\tau_t \gtrsim \lambda^2 - 1$, we can deduce from the relation above that

$$\mathcal{T}_2(\sqrt{\tau+1}\wedge\lambda,\tau) \le 1 - \big((\sqrt{\tau+1}\wedge\lambda) - 1\big) = 1 - c_1(\lambda-1) \tag{172}$$

for some universal constant $c_1 > 0$. Finally, putting relations (171) and (172) together, we arrive at

$$\frac{\tau_t}{h(\tau_t)}h'(\tau) \le \mathcal{T}_2(\sqrt{\tau_t+1}\wedge\lambda,\tau) = (1+o(\lambda-1))\mathcal{T}_2(\sqrt{\tau+1}\wedge\lambda,\tau) \le 1 - c(\lambda-1)$$

for some universal constant $c > 0$. We have thus finished the proof of relation (73).

C Proof of expression (17) and Corollary 1

To begin with, by definition (16) of u_t, one has

$$\|u_t\|_2 = \left\|\frac{1}{\lambda\sqrt{n(\alpha_t^2+1)}}\tanh(\pi_t x_t)\right\|_2 \le \frac{1}{\lambda},$$

where we have used the fact that $|\tanh(\pi_t x_t)| < 1$. Therefore the quantity of interest $\|v^\star v^{\star\top} - u_t u_t^\top\|_{\mathsf{F}}^2$ is uniformly upper bounded by a constant $1 + \frac{1}{\lambda^4}$. In addition, we find it helpful to observe that

$$\begin{aligned}
\|v^\star v^{\star\top} - u_t u_t^\top\|_{\mathsf{F}}^2 &= \|v^\star\|_2^4 - 2(v^{\star\top}u_t)^2 + \|u_t\|_2^4 \\
&= 1 - 2\frac{1}{n\lambda^2(\alpha_t^2+1)}\big(v^{\star\top}\tanh(\pi_t x_t)\big)^2 + \frac{1}{n^2\lambda^4(\alpha_t^2+1)^2}\|\tanh(\pi_t x_t)\|_2^4 \\
&= 1 - \frac{1}{\lambda^4}\left(\frac{2\alpha_{t+1}^2}{(\alpha_t^2+1)n\gamma_t^2} + \frac{1}{(\alpha_t^2+1)^2 n^2\gamma_t^4}\right).
\end{aligned} \tag{173}$$

To validate expression (17), it is sufficient to notice that $\gamma_t^{-2} - n\alpha_t^2(\alpha_t^2+1) = o(1)$ with probability at least $1 - O(n^{-10})$ (according to Lemma 2), which in turn leads to

$$\lim_{t\to\infty}\lim_{n\to\infty}\mathbb{E}\big[\|v^\star v^{\star\top} - u_t u_t^\top\|_{\mathsf{F}}^2\big] = 1 - \frac{\alpha^{\star 4}}{\lambda^4}.$$

To prove Corollary 1, we again invoke Lemma 2 to demonstrate that

$$\begin{aligned}
\frac{1}{n}\gamma_t^{-2} &= \frac{1}{n}\pi_t^2\left(\alpha_t^2 + 1 + O\left(\frac{\pi_t^2}{n} + \|\xi_{t-1}\|_2 + \sqrt{\frac{t\log n}{n}}\right)\right) \\
&= \left(\alpha_t^2 + O\left(\left(\|\xi_{t-1}\|_2 + \sqrt{\frac{t\log n}{n}}\right)^{1/2}\right)\right)\left(\alpha_t^2 + 1 + O\left(\frac{\pi_t^2}{n} + \|\xi_{t-1}\|_2 + \sqrt{\frac{t\log n}{n}}\right)\right) \\
&= (\alpha_t^2 + O(\delta^{1/2}))(\alpha_t^2 + 1 + O(\delta)) \\
&= \alpha_t^2(\alpha_t^2+1) + O(\delta^{1/2})
\end{aligned} \tag{174}$$

with $\delta \coloneqq \sqrt{\frac{t\log n}{n(\lambda-1)^2}} + \sqrt{\frac{\log^4 n}{n(\lambda-1)^3}}$, where we plug in the bound on $\|\xi_t\|_2$ as in expression (8c). Substituting the expression (174) into (173) yields

$$\|v^\star v^{\star\top} - u_t u_t^\top\|_{\mathsf{F}}^2 = 1 - \frac{\alpha_t^2}{\lambda^4}\Big(2\alpha_{t+1}^2 - \alpha_t^2 + O(\delta^{1/2})\Big). \tag{175}$$

After an order of $\frac{\log n}{\lambda-1}$ iterations, property (14) ensures that $\alpha_t^2 - \alpha^{\star 2} = O(\sqrt{\frac{\log^4 n}{n(\lambda-1)^6}})$. Putting everything together, we arrive at

$$\|v^\star v^{\star\top} - u_t u_t^\top\|_{\mathsf{F}}^2 = 1 - \frac{\alpha^{\star 4}}{\lambda^4} + O\left(\sqrt{\frac{\log^4 n}{n(\lambda-1)^6}}\right),$$

which holds true with probability at least $1 - O(n^{-10})$.

References

[1] Gen Li, Wei Fan, and Yuting Wei. Approximate message passing from random initialization with applications to synchronization. *Proceedings of the National Academy of Sciences*.

[2] Gen Li and Yuting Wei. A non-asymptotic framework for approximate message passing in spiked models. *arXiv preprint arXiv:2208.03313*, 2022.

[3] Pascal Massart. *Concentration inequalities and model selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII-2003*. Springer, 2007.

[4] Yash Deshpande, Emmanuel Abbe, and Andrea Montanari. Asymptotic mutual information for the balanced binary stochastic block model. *Information and Inference: A Journal of the IMA*, 6(2):125–170, 2017.