Analysis of Approximate Message Passing
with Non-Separable Denoisers and Markov Random Field Priors

Yanting Ma
North Carolina State University
yma7@ncsu.edu

Cynthia Rush
Columbia University
cynthia.rush@columbia.edu

Dror Baron
North Carolina State University
barondror@ncsu.edu

May 13, 2019

Abstract

Approximate message passing (AMP) is a class of low-complexity, scalable algorithms for
solving high-dimensional linear regression tasks where one wishes to recover an unknown signal
from noisy, linear measurements. AMP is an iterative algorithm that performs estimation by
updating an estimate of the unknown signal at each iteration and the performance of AMP
(quantified, for example, by the mean squared error of its estimates) depends on the choice of
a “denoiser” function that is used to produce these signal estimates at each iteration.

An attractive feature of AMP is that its performance can be tracked by a scalar recursion
referred to as state evolution. Previous theoretical analysis of the accuracy of the state
evolution predictions has been limited to the use of only separable denoisers or block-separable
denoisers, a class of denoisers that underperform when sophisticated dependencies exist between
signal entries. Since signals with entrywise dependencies are common in image/video-processing
applications, in this work we study the high-dimensional linear regression task when the depen-
dence structure of the input signal is modeled by a Markov random field prior distribution. We
provide a rigorous analysis of the performance of AMP, demonstrating the accuracy of the state
evolution predictions, when a class of non-separable sliding-window denoisers is applied. More-
over, we provide numerical examples where AMP with sliding-window denoisers can successfully
capture local dependencies in images.

1 Introduction

In this work, we study the problem of estimating an unknown signal \( \beta := (\beta_i)_{i \in \Gamma} \) from noisy, linear measurements as in the following model:

\[
y = AV(\beta) + w, \tag{1}
\]

where for some integer \( p \in \{1, 2, 3\}, \Gamma \subset \mathbb{Z}^p \) is an index set with cardinality \(|\Gamma|\), \( y \in \mathbb{R}^n \) is the output, \( A \in \mathbb{R}^{n \times |\Gamma|} \) is a known measurement matrix, \( w \in \mathbb{R}^n \) is zero-mean noise with finite variance \( \sigma^2 \), and \( V \) (script \( V \) stands for “vectorization”) is an invertible operator that rearranges elements of
an array into a vector, hence \( V(\beta) \) is a length-\(|\Gamma|\) vector. We assume that the ratio of the dimensions
of the measurement matrix is a constant value, \( \delta := n/|\Gamma| \), with \( \delta \in (0, \infty) \).

Approximate message passing (AMP) \([2–6]\) is a class of low-complexity, scalable algorithms
studied to solve the high-dimensional regression task of (1). The performance of AMP depends

*Portions of the work appeared at the IEEE International Symposium on Information Theory (ISIT), Aachen,
Germany, June 2017 [1].
on a sequence of functions $\{\eta_t\}_{t \geq 0}$ used to generate a sequence of estimates $\{\beta^t\}_{t \geq 0}$ from effective observations computed in every iteration of the algorithm. A nice property of AMP is that under some technical conditions these observations can be approximated as the input signal $\beta$ plus independent and identically distributed (i.i.d.) Gaussian noise. For this reason, the functions $\{\eta_t\}_{t \geq 0}$ are referred to as “denoisers.”

Previous analysis of the performance of AMP only considers denoisers $\{\eta_t\}_{t \geq 0}$ that act coordinate-wise when applied to a vector; such denoisers are referred to as separable. If the unknown signal $\beta$ has a prior distribution with i.i.d. entries, restricting consideration to only separable denoisers causes no loss in performance. However, in many real-world applications, the unknown signal $\beta$ contains dependencies between entries, and therefore a coordinate-wise independence structure does not approximate the prior for $\beta$ well. Instead of using a separable denoiser, non-separable denoisers can improve reconstruction quality for signals with such dependencies among entries. For example, when the signals are images [7,8] or sound clips [9], non-separable denoisers outperform reconstruction techniques based on over-simplified i.i.d. models. In such cases, a more appropriate model might be a finite memory model, well-approximated with a Markov random field (MRF) prior. In this paper, we extend the previous performance guarantees for AMP to a class of non-separable sliding-window denoisers when the unknown signal is a realization of an MRF. Sliding-window schemes have been studied for denoising signals with dependencies among entries by, for example, Sivaramakrishnan and Weissman [10,11]. MRFs are appropriate models for many types of images, especially texture images, which have an inherently random component [12,13].

When the measurement matrix $A$ has i.i.d. Gaussian entries and the empirical distribution function\(^1\) of the unknown signal $\beta$ converges to some distribution function on $\mathbb{R}$, Bayati and Montanari [4] proved that at each iteration the performance of AMP can be accurately predicted by a simple, scalar iteration referred to as state evolution in the large system limit ($n/|\Gamma| \to \infty$ such that $(n/|\Gamma|) \to \delta$ is a constant). For example, if $\beta^t$ is the estimate produced by AMP at iteration $t$, the result by Bayati and Montanari [4] implies that the normalized squared error, $\frac{1}{n^2} \| \beta^t - \beta \|^2$, and other performance measures converge to deterministic values predicted by state evolution, which is a deterministic recursion calculated using the prior distribution of $\beta$.\(^2\) Rush and Venkataraman [14] provided a concentration version of the asymptotic result when the prior distribution of $\beta$ is i.i.d. sub-Gaussian. The result in Rush and Venkataraman [14] implies that the probability of $\epsilon$-deviation between various performance measures and their limiting constant values decay exponentially in $|\Gamma|$.

Extensions of AMP performance guarantees beyond separable denoisers have been considered in special cases [15,16] for certain classes of block-separable denoisers that allow dependencies within blocks of the signal $\beta$ with independence across blocks. A preliminary version [1] of this work has analyzed the performance of AMP with sliding-window denoisers applied to the setting where the unknown signal has a Markov chain prior. In this paper, we generalize the previous result with the applications of compressive imaging [7,8] and compressive hyperspectral imaging [17] in mind. We consider 2D/3D MRF priors for the input signal $\beta$, and provide performance guarantees for AMP with 2D/3D sliding-window denoisers under some technical conditions.

While we were concluding this manuscript, we became aware of recent work of Berthier et al. [18]. The authors prove that the loss of the estimates generated by AMP (for a class of loss functions) with general non-separable denoisers converges to the state evolution predictions asymptotically.

\(^1\)For a vector $(x_1, \ldots, x_N) \in \mathbb{R}^N$, the empirical distribution function $F_N : \mathbb{R} \to [0,1]$ is defined as $F_N(x) := \frac{1}{N} \sum_{i=1}^{N} I(x_i \leq x)$, where $I$ denotes the indicator function. The empirical distribution function $F_N$ is said to converge to some distribution function $F : \mathbb{R} \to [0,1]$ if for all $x \in \mathbb{R}$ such that $F(x)$ is continuous, we have $\lim_{N \to \infty} F_N(x) = F(x)$.

\(^2\)Throughout the paper, $\|v\|^2$ denotes the sum of squares of all the entries in $v$, where $v$ could be, for example, in $\mathbb{R}^n$, $\mathbb{R}^{n \times n}$, or $\mathbb{R}^{n \times n \times n}$. 

Similarly, let $\Lambda$ be an initial initialization $\beta$ where $2$ for an integer $N$ of the state evolution sequence is with respect to (w.r.t.) the signal $\beta$, the matrix $A$, and the noise $w$, whereas in [18], the signal $\beta$ is deterministic and unknown, hence the expectation is only w.r.t. the matrix $A$ and the noise $w$.

1.1 Sliding-Window Denoisers and AMP Algorithm

**Notation:** Before introducing the algorithm, we provide some notation that is used to define the sliding window in the sliding-window denoiser. Without loss of generality, we let the index set $\Gamma \subset \mathbb{Z}^p$, on which the input signal $\beta$ in (1) is defined, be

$$\Gamma = \begin{cases} \{N\}, & \text{if } p = 1, \\ \{N\} \times \{N\}, & \text{if } p = 2, \\ \{N\} \times \{N\} \times \{N\}, & \text{if } p = 3, \end{cases} \quad (2)$$

where for an integer $N$, the notation $\{N\}$ represents the set of integers $\{1, \ldots, N\}$, hence, $|\Gamma| = N^p$. Similarly, let $\Lambda$ be a $p$-dimensional cube in $\mathbb{Z}^p$ with length $(2k+1)$ in each dimension, namely,

$$\Lambda := \begin{cases} [2k+1], & \text{if } p = 1, \\ [2k+1] \times [2k+1], & \text{if } p = 2, \\ [2k+1] \times [2k+1] \times [2k+1], & \text{if } p = 3, \end{cases} \quad (3)$$

where $2k+1 \leq N$. We call $k$ the half-window size.

**AMP with sliding-window denoisers:** The AMP algorithm for estimating $\beta$ from $y$ and $A$ in (1) generates a sequence of estimates $\{\beta^t\}_{t \geq 0}$, where $\beta^t \in \mathbb{R}^\Gamma$, $t$ is the iteration index, and the initialization $\beta^0 := 0$ is an all-zero array with the same dimension as the input signal $\beta$. For $t \geq 0$, the algorithm proceeds as follows:

$$z^t = y - A \mathcal{V}(\beta^t) + \frac{1}{n} \sum_{i \in \Gamma} \eta'_{t-1} \left( \left[ \mathcal{V}^{-1}(A^*z^{t-1}) + \beta^{t-1} \right]_{\Lambda_i} \right), \quad (4)$$

$$\beta^{t+1}_i = \eta_t \left( \left[ \mathcal{V}^{-1}(A^*z^t) + \beta^t \right]_{\Lambda_i} \right), \quad \text{for all } i \in \Gamma, \quad (5)$$

where the function $\{\eta_t\}_{t \geq 0} : \mathbb{R}^\Lambda \rightarrow \mathbb{R}$ is a sequence of denoisers, $\eta'_{t-1}$ is the partial derivative w.r.t. the center coordinate of the argument, $A^*$ is the transpose of $A$, and $\Lambda_i \subset \mathbb{Z}^p$ for each $i \in \Gamma \subset \mathbb{Z}^p$ is the $p$-dimensional cube $\Lambda$ translated to be centered at location $i$. The translated $p$-dimensional cubes $\{\Lambda_i\}_{i \in \Gamma}$ are referred to as “sliding windows,” which will be used to subset elements of a $p$-dimensional array. The effective observation at iteration $t$ is $\mathcal{V}^{-1}(A^*z^t) + \beta^t \in \mathbb{R}^\Gamma$, which can be approximated as the true signal $\beta$ plus i.i.d. Gaussian noise (in a sense that will be made clear in the statement of our main result, Theorem 1). Note that the sliding-windows $\{\Lambda_i\}_{i \in \Gamma}$ and the sliding-window denoiser $\eta_t$ are defined on multidimensional signals, hence we use the inverse of the vectorization operator, $\mathcal{V}^{-1}$, to rearrange elements of vectors into arrays before applying the sliding-window denoiser $\eta_t$. It should also be noted that the denoiser $\eta_t$ may only process part of the signal elements in $\Lambda$. For example, in the 2D case, if $\Lambda$ is defined as a $3 \times 3$ window, then $\eta_t$ may only process the center and the four adjacent pixel values in the window (see Figure 1) and
ignore the four corners. To simplify notation, we will write \( \eta: \mathbb{R}^\Lambda \to \mathbb{R} \) throughout the paper, and interpret this notation to mean that any processing of neighboring signal values is allowed, including the possibility of ignoring some of their values.

**Edge cases:** Notice that when the center coordinate \( i \) is near the edge of \( \Gamma \), some of the elements in \( \Lambda_i \) may fall outside \( \Gamma \), meaning that \( \Lambda_i \cap \Gamma^c \neq \emptyset \) where \( \Gamma^c \) is the complement of \( \Gamma \) w.r.t. \( \mathbb{Z}^p \). In the definition of the AMP algorithm with sliding-window denoisers and the subsequent analysis, these “edge cases” must be handled carefully. The following definitions provide a framework for the special treatment of the edge cases.

Based on whether \( \Lambda_i \) has elements outside \( \Gamma \), we partition the index set \( \Gamma \) into two sets \( \Gamma^{\text{mid}} \) and \( \Gamma^{\text{edge}} \) defined as:

\[
\Gamma^{\text{mid}} := \{ i \in \Gamma \mid \Lambda_i \cap \Gamma^c = \emptyset \} \quad \text{and} \quad \Gamma^{\text{edge}} := \{ i \in \Gamma \mid \Lambda_i \cap \Gamma^c \neq \emptyset \}.
\]

(6)

That is, for \( i \in \Gamma^{\text{mid}} \), all elements in \( \Lambda_i \) lie inside \( \Gamma \), whereas for \( i \in \Gamma^{\text{edge}} \), some of the elements in \( \Lambda_i \) fall outside \( \Gamma \). The size of each set will depend on the half-window size \( k \) and the dimension \( p \).

For any \( v \in \mathbb{R}^\Gamma \), let \( v_{\Lambda_i} \) be a subset of the elements of \( v \) with indices in \( \Lambda_i \) and for any \( j \in \Lambda \), let \( [\Lambda_i]_j \) be the \( j \)th index of \( \Lambda_i \) so that \( v_{[\Lambda_i]_j} \) returns a single element of \( v_{\Lambda_i} \). Notice that for \( i \in \Gamma^{\text{mid}} \), all entries of \( v_{\Lambda_i} \) are well-defined. However, for \( i \in \Gamma^{\text{edge}} \), the subset \( v_{\Lambda_i} \) has undefined entries, namely, for all \( j \in \Lambda \) such that \( [\Lambda_i]_j \in \Gamma^c \), the entry \( v_{[\Lambda_i]_j} \) is undefined. We now define the value of those “missing” entries to be the average of the entries of \( v_{\Lambda_i} \) having indices in \( \Gamma \). Formally,

\[
v_{[\Lambda_i]_j} := \frac{1}{|\Lambda_i \cap \Gamma|} \sum_{\ell \in \Lambda_i \cap \Gamma} v_{\ell}, \quad \text{for all} \quad j \in \Lambda \quad \text{such that} \quad [\Lambda_i]_j \in \Gamma^c.
\]

(7)

It may improve signal recovery quality to use other schemes for “missing” entries, like interpolation. We leave the study of these improved schemes for future work, but the effect of improved processing around edges should become minor as \( N \) increases. Notice that \( v_{\Lambda_i} \) for all \( i \in \Gamma \) are now defined using only the entries in the original \( v \in \mathbb{R}^\Gamma \). It will be useful to emphasize this point in the proof of our main result, so we define a set of functions \( \{ \mathcal{T}_i \}_{i \in \Gamma} \) with \( \mathcal{T}_i: \mathbb{R}^{\Lambda_i \cap \Gamma} \to \mathbb{R}^\Lambda \) as

\[
\mathcal{T}_i(v_{\Lambda_i \cap \Gamma}) := v_{\Lambda_i}, \quad \text{for all} \quad i \in \Gamma,
\]

(8)

where \( v_{\Lambda_i} \) follows our definition above. That is, \( \mathcal{T}_i \) is identity for \( i \in \Gamma^{\text{mid}} \), whereas for \( i \in \Gamma^{\text{edge}} \), \( \mathcal{T}_i \) extends a smaller array \( v_{\Lambda_i \cap \Gamma} \) to a larger one \( v_{\Lambda_i} \) with the extended entries defined by (7).

**Examples for defining “missing” entries:** To illustrate the notations defined above, we present an example for the \( p = 1 \) case (hence \( v \in \mathbb{R}^N \) is a vector). As defined above in (2) and (3), we have \( \Gamma = \{1, \ldots, N\}, \Lambda = \{1, \ldots, 2k+1\} \), and \( \Lambda_i = (i-k, \ldots, i-1, i, i+1, \ldots, i+k) \) for each \( i \in [N] \).
Moreover, $\Gamma_{\text{mid}} = \{k + 1, k + 2, \ldots, N - k\}$ and $\Gamma_{\text{edge}} = \{1, 2, \ldots, k\} \cup \{N - k + 1, N - k + 2, \ldots, N\}$ as defined in (6). Therefore, for $i \in \Gamma_{\text{mid}}$,

$$v_{\Lambda_i} = \mathcal{T}_i(v_{i-k}, v_{i-k+1}, \ldots, v_{i+k}) := (v_{i-k}, v_{i-k+1}, \ldots, v_{i+k}) \in \mathbb{R}^{2k+1}.$$  

For $i \in \Gamma_{\text{edge}}$, the vector $v_{\Lambda_i}$ is still length-$(2k + 1)$, and we set the values of the non-positive indices, i.e., $1 - k, 2 - k, \ldots, -1, 0$, or indices above $N$, i.e., $N + 1, N + 2, \ldots, N + k$, to be the average of values in the vector $v_{\Lambda_i}$ with indices in $\Lambda_i \cap [N]$. For example, let $i = 3$ and $k = 5$ giving $\Lambda_3 = (-2, -1, 0, 1, \ldots, 8)$ so that for $j \in \{1, 2, 3\}$ we have $[\Lambda_3]_j \in \Gamma$. Following (7), define

$$\bar{v} = \frac{1}{8} \sum_{j=1}^{8} v_j, \quad \text{and set} \quad v_{[\Lambda_3]_1} = v_{[\Lambda_3]_2} = v_{[\Lambda_3]_3} = \bar{v},$$

and so $v_{\Lambda_3} = \mathcal{T}_3(v_1, \ldots, v_8) := (\bar{v}, \bar{v}, \bar{v}, v_1, \ldots, v_8) \in \mathbb{R}^{11}$. An example for the $p = 2$ case (hence $v \in \mathbb{R}^{N \times N}$ is a matrix), is shown in Figure 2.

### 1.2 Contributions and Outline

Our main result proves concentration for (order-2) pseudo-Lipschitz (PL(2)) loss functions acting on the AMP estimate given in (5) at any iteration $t$ of the algorithm to constant values predicted by the state evolution equations that will be introduced in the following. This work covers the case where the unknown signal $\beta$ has an MRF prior on $\mathbb{Z}^p$. For example, when $p = 2$, $\beta$ can be thought of as an image, whereas when $p = 3$, $\beta$ can be thought of as a hyperspectral image cube. Moreover, we use numerical examples to demonstrate the effectiveness of AMP with sliding-window denoisers when used to reconstruct images from noisy linear measurements.

The rest of the paper is organized as follows. Section 2 provides model assumptions, state evolution formulas, the main performance guarantee, and numerical examples illustrating the effectiveness of the algorithm for compressive image reconstruction. Our main performance guarantee

---

\footnote{A function $f : \mathbb{R}^m \rightarrow \mathbb{R}$ is (order-2) pseudo-Lipschitz if there exists a constant $L > 0$ such that for all $x, y \in \mathbb{R}^m$, $|f(x) - f(y)| \leq L(1 + \|x\| + \|y\|)\|x - y\|$.}
Note that if $E$ is countable, then

$$\{ i \in E \}. Note that if $E$ is countable, then

$$\{ i \in E \}. Note that if $E$ is countable, then

$$\{ i \in E \}. Note that if $E$ is countable, then

$$\{ i \in E \}. Note that if $E$ is countable, then

$$\{ i \in E \}. Note that if $E$ is countable, then

Then the random field is said to be stationary if, for all $i \in \Gamma$ such that $i + \Lambda \subset \Gamma$, it is true that $u_{\Lambda} = u_{i+\Lambda}$.

Next we introduce the Dobrushin uniqueness condition, under which the random field admits a unique stationary distribution. Define the Dobrushin interdependence matrix $(C_{i,j})_{i,j \in \Gamma}$ for the measure $\mu$ of the random field $X$ to be

$$C_{i,j} := \sup_{\xi,\xi' \in E^c \atop \xi_j = \xi'_j} ||\mu_i(\cdot|\xi) - \mu_i(\cdot|\xi')||_{tv}. \quad (9)$$

In the above, the index set $j^c := \Gamma \setminus \{ j \}$ and the total variation distance $|| \cdot ||_{tv}$ between two probability measures $\rho_1$ and $\rho_2$ on $(E, \mathcal{E})$ is defined as

$$||\rho_1(\cdot) - \rho_2(\cdot)||_{tv} := \max_{B \in \mathcal{E}} |\rho_1(B) - \rho_2(B)|. \quad (10)$$

Note that if $E$ is countable, then

$$||\rho_1(\cdot) - \rho_2(\cdot)||_{tv} = \frac{1}{2} \sum_{x \in E} |\rho_1(x) - \rho_2(x)|. \quad (10)$$

The measure $\mu$ is said to satisfy the Dobrushin uniqueness condition if

$$c := \sup_{i \in \Gamma} \sum_{j \in \Gamma} C_{i,j} < 1.$$
The ** Dobrushin contraction coefficient**, $c$, is a quantity that estimates the magnitude of change of the single site conditional expectations, as they appear in (9), when the field values at the other sites vary. Similarly, we define the **transposed Dobrushin contraction condition** as

$$c^* := \sup_{j \in \Gamma} \sum_{i} C_{i,j} < 1.$$ 

**Assumptions:** We can now state our assumptions on the signal $\beta$, the matrix $A$, and the noise $w$ in the linear system (1), as well as the denoiser function $\eta_t$ used in the algorithm (4) and (5).

**Signal:** Let $E \subset \mathbb{R}$ be a bounded state space (countable or uncountable). Let $\beta = (\beta_i)_{i \in \Gamma}$ be a **stationary** MRF with Gibbs distribution measure $\mu$ on $E^\Gamma$, where $\Gamma \subset \mathbb{Z}^p$ is a finite and nonempty rectangular lattice. We assume that $\mu$ satisfies the Dobrushin uniqueness condition and the transposed Dobrushin uniqueness condition. These two conditions together are needed for the results in Lemma C.1 and Lemma C.2, which demonstrate concentration of sums of pseudo-Lipschitz functions when the input to the functions are MRFs with distribution measure $\mu$. Roughly, the conditions ensure that the dependencies between the terms in the sums are sufficiently weak for the desired concentration to hold. The class of finite state space stationary MRFs, which is widely used for image analysis [21], is one example that satisfies our assumption.

**Denoiser functions:** The denoiser functions $\eta_t : \mathbb{R}^\Lambda \to \mathbb{R}$ used in (5) are assumed to be **Lipschitz**

for each $t > 0$ and are, therefore, also weakly differentiable with bounded (weak) partial derivatives. We further assume that the partial derivative w.r.t. the center coordinate of $\Lambda$, which is denoted by $\eta'_t : \mathbb{R}^\Lambda \to \mathbb{R}$, is itself differentiable with bounded partial derivatives. Note that this implies $\eta'_t$ is Lipschitz. (It is possible to weaken this condition to allow $\eta'_t$ to have a finite number of discontinuities, if needed, as in [14].)

**Matrix:** The entries of the matrix $A$ are i.i.d. $\sim \mathcal{N}(0, 1/n)$.

**Noise:** The entries of the measurement noise vector $w$ are i.i.d. according to some **sub-Gaussian** distribution $p_w$ with mean 0 and finite variance $\sigma^2$. The sub-Gaussian assumption implies [22] that for all $\epsilon \in (0, 1)$ and for some constants $K, \kappa > 0$,

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

### 2.2 Performance Guarantee

As noted in Section 1, the behavior of the AMP algorithm is predicted by a deterministic scalar recursion referred to as state evolution, which we now introduce. More specifically, the state evolution sequences $\{\tau^2_t\}_{t \geq 0}$ and $\{\sigma^2_t\}_{t \geq 0}$ defined below in (11) will be used in Theorem 1 to characterize the estimation error of the estimates produced by AMP. Let the joint distribution $\mu$ define the (stationary) prior distribution for the unknown signal $\beta$ in (1). Following our assumption of stationarity, $\beta_i \sim \mu_1$ for all $i \in \Gamma$ and $\beta_{\Lambda_i} \sim \mu_\Lambda$ for all $i \in \Gamma$ defined in (6), where $\mu_1$ and $\mu_\Lambda$ denote the one-dimensional marginal and $\Lambda$-dimensional marginal of $\mu$, respectively. Define $\sigma^2_\beta = \mathbb{E}[\beta^2_\beta] > 0$, and $\sigma^2_0 = \sigma^2_\beta / \delta$. Iteratively define the state evolution sequences $\{\tau^2_t\}_{t \geq 0}$ and $\{\sigma^2_t\}_{t \geq 1}$ as follows:

$$\tau^2_t = \sigma^2 + \sigma^2_t \quad \text{and} \quad \sigma^2_t = \frac{1}{\delta|\Gamma|} \sum_{i \in \Gamma} \mathbb{E} \left[ (\eta_{t-1}(\beta + \tau_{t-1}Z_{\Lambda_i}) - \beta_i)^2 \right],$$

\[ \text{(11)} \]

\[ ^4 \text{A function } f : \mathbb{R}^m \to \mathbb{R} \text{ is Lipschitz if there exists a constant } L > 0 \text{ such that for all } x, y \in \mathbb{R}^m, |f(x) - f(y)| \leq L \|x - y\| \text{.} \]
where $\eta : \mathbb{R}^\Lambda \to \mathbb{R}$ is the sliding-window denoiser and $Z \in \mathbb{R}^\Gamma$ has i.i.d. standard normal entries, independent of $\beta$, which implies that $Z_{\Lambda_i}$ is independent of $\beta_{\Lambda_i}$ and $\beta_i$ for all $i \in \Gamma$. Let $\beta^t \in E^\Lambda \sim \mu_\Lambda$ and define $Z^t \in \mathbb{R}^\Lambda$ with entries that are i.i.d. $\mathcal{N}(0,1)$. We notice that for all $i \in \Gamma^{\mid d}$, we have $\beta_{\Lambda_i} \overset{d}{=} \beta'$ and $Z_{\Lambda_i} \overset{d}{=} Z'$. Therefore, for all $i \in \Gamma^{\mid d}$, the expectations in (11) satisfy

\[
E \left[ (\eta_{t-1}(\beta + \tau_{t-1}Z) - \beta_i)^2 \right] = E \left[ (\eta_{t-1}(\beta' + \tau_{t-1}Z') - \beta'_c)^2 \right],
\]

where $\beta'_c$ is the center coordinate of $\beta'$. For $i \in \Gamma^{\text{edge}}$ with $\Gamma^{\text{edge}}$ defined in (6), it is not necessarily true that $\beta_{\Lambda_i} \overset{d}{=} \beta'$ since, by definition (7), some entries of $\beta_{\Lambda_i}$ are defined as the average of other entries.

The explicit expression for the definition of $\sigma^2_t$ in (11) is different when considering $\Gamma \subset \mathbb{Z}^p$ for different $p$ values, as the size of the set $\Gamma^{\text{edge}}$ depends on the dimension. In the following, we provide explicit expressions for $\sigma^2_t$ for the cases $p = 1, 2$, but in the proof we will use the general expression given in (11) for brevity. We emphasize that the definition of the state evolution sequence in (11) only uses the marginal distribution $\mu_{\Lambda}$ (or $\beta' \in E^\Lambda$) instead of the joint distribution $\mu$ (or $\beta \in E^\Gamma$), as demonstrated in the two examples below in (12) and (13).

**Examples for explicit expressions for $\sigma^2_t$:** Let $\beta'_{\Lambda}$ be the center coordinate of $\beta' \in E^\Lambda$ and $\Lambda_c$ the window $\Lambda \subset \mathbb{Z}^p$ translated with center $c \in \mathbb{Z}^p$. Recall that $\Lambda$ is the $p$-dimensional cube with length $(2k + 1)$ in each of the $p$ dimensions. Then we have $\beta' = \beta'_{\Lambda_c}$ and when we consider shifts $\beta'_{\Lambda_{c+\ell}}$ for $\ell \in \{-k, -k + 1, \ldots, k - 1, k\}$ we, analogous to the definition in (7), define “missing” entries to be replaced by the average of the existing entries. (Note that since $\beta'$ is exactly of size $\Lambda$, thus for any $\ell \neq 0$, there will be “missing” entries.) For example, when $p = 1$,

$\beta'_{\Lambda_c} = (\beta'_1, \beta'_2, \ldots, \beta'_{2k+1})$, while $\beta'_{\Lambda_{c-2}} = (\text{avg}, \text{avg}, \beta'_1, \beta'_2, \ldots, \beta'_{2k-1})$,

where $\text{avg} = \frac{1}{2k+1} \sum_{i=1}^{2k+1} \beta'_i$. Generalizing, we have $\beta'_{\Lambda_{c+\ell}} \in \mathbb{R}^{2k+1}$ with

\[
\beta'_{\Lambda_{c+\ell}} = \begin{cases} 
\frac{1}{2k+1} \sum_{i=1}^{2k+1+\ell} \beta'_i, \ldots, \frac{1}{2k+1+\ell} \sum_{i=1}^{2k+1+\ell} \beta'_i, \beta'_1, \beta'_2, \ldots, \beta'_{2k+1+\ell} & \text{if } \ell < 0, \\
(\beta'_1, \beta'_2, \ldots, \beta'_{2k+1}) & \text{if } \ell = 0, \\
(\beta'_{1+\ell}, \beta'_{2+\ell}, \ldots, \beta'_{2k+1}, \frac{1}{2k+1-\ell} \sum_{i=1+\ell}^{2k+1} \beta'_i, \ldots, \frac{1}{2k+1-\ell} \sum_{i=1+\ell}^{2k+1} \beta'_i) & \text{if } \ell > 0.
\end{cases}
\]

The same idea can be extended when $p > 1$.

For the case $p = 1$, we note that $\Gamma^{\mid d} = \{k+1, k+2, \ldots, N-k-1\}$ and $\Gamma^{\text{edge}} = \{1, 2, \ldots, k\} \cup \{N-k, N-k+1, \ldots, N\}$, hence $|\Gamma^{\mid d}| = N - 2k$ and $|\Gamma^{\text{edge}}| = 2k$. Therefore, we have

\[
\sigma^2_t = \frac{(N-2k)}{\delta N} E \left[ (\eta_{t-1}(\beta + \tau_{t-1}Z) - \beta')^2 \right] + \frac{1}{\delta N} \sum_{\ell \in \{-k, \ldots, k\} \setminus \{0\}} E \left[ (\eta_{t-1}(\beta' + \tau_{t-1}Z)_{\Lambda_{c+\ell}} - \beta'_{c+\ell})^2 \right],
\]

where $\{-k, \ldots, k\} \setminus \{0\} = \{-k, \ldots, -1\} \cup \{1, \ldots, k\}$. In the above the first term corresponds to the $N - 2k$ middle indices, while the second term sums over $2k$ terms corresponding to all the possible edge cases.

For the case $p = 2$, we note that $\Gamma^{\mid d} = \{(i, j) | k+1 \leq i, j \leq N-k+1\}$, hence $|\Gamma^{\mid d}| = (N-2k)^2$. Here we note $\ell = (\ell_1, \ell_2) \in \{-k, -k+1, \ldots, k-1, k\} \times \{-k, -k+1, \ldots, k-1, k\}$. Therefore,
\[
\sigma_t^2 = \frac{(N - 2k)^2}{\delta N^2} \mathbb{E} \left[ (\eta_{t-1}(\beta' + \tau_{t-1} Z') - \beta'_c)^2 \right] + \frac{1}{\delta N^2} \sum_{\ell_1, \ell_2 \in \{-k, \ldots, k\} \setminus \{0\}} \mathbb{E} \left[ (\eta_{t-1}([\beta' + \tau_{t-1} Z']_{\Lambda_{c+\ell}}) - \beta'_{c+\ell})^2 \right] \\
+ \frac{(N - 2k)}{\delta N^2} \sum_{\ell_1 \in \{-k, \ldots, k\} \setminus \{0\}} \mathbb{E} \left[ (\eta_{t-1}([\beta' + \tau_{t-1} Z']_{\Lambda_{c+\ell}}) - \beta'_{c+\ell})^2 \right] \\
+ \frac{(N - 2k)}{\delta N^2} \sum_{\ell_2 \in \{-k, \ldots, k\} \setminus \{0\}} \mathbb{E} \left[ (\eta_{t-1}([\beta' + \tau_{t-1} Z']_{\Lambda_{c+\ell}}) - \beta'_{c+\ell})^2 \right],
\]

where we notice that there are \((2k)^2\) terms in the second summand, \(2k\) terms in the third and fourth summands, and \(\frac{(N-2k)^2}{N^2} + \frac{(2k)^2}{N^2} + \frac{2k(N-2k)}{N^2} + \frac{2k(N-2k)}{N^2} = 1\). Again, in the above the first term sums over all the middle indices. In this case, the second term corresponds to the corner edge cases, while the third and fourth terms correspond to the edge cases in one dimension only. We note that \(\sigma_t^2\) is a function of \(N\), but do not explicitly represent this relationship to simplify the notation. Moreover, for fixed \(k\), the terms \(\frac{(2k)^2}{N^2}\), \(\frac{2k(N-2k)}{N^2}\), and \(\frac{2k(N-2k)}{N^2}\) vanish as \(N\) goes to infinity. Therefore, we have \(\lim_{N \to \infty} \sigma_t^2(N) = \frac{1}{2} \mathbb{E} \left[ (\eta_{t-1}(\beta' + \tau_{t-1} Z') - \beta'_c)^2 \right]\).

Similar to \([14]\), our performance guarantee, Theorem 1, is a concentration inequality for PL(2) loss functions at any fixed iteration \(t < T^*\), where \(T^*\) is the first iteration when either \((\sigma_t^2)\) or \((\tau_t^2)\) defined in (36) is smaller than a predefined quantity \(\hat{\delta}\). The precise definition of \((\sigma_t^2)\) and \((\tau_t^2)\) is deferred to Section 3.2. For now, we can understand \((\sigma_t^2)\) (respectively, \((\tau_t^2)\)) as a number that quantifies (in a probabilistic sense) how close an estimate \(\beta^t\) (respectively, a residual \(z^t\)) is to the subspace spanned by the previous estimates \(\{\beta^s\}_{s \leq t}\) (respectively, the previous residuals \(\{z^s\}_{s \leq t}\)). In the special case where \(\{\eta_t\}_{t \geq 0}\) are Bayes-optimal conditional expectation denoisers, it can be shown that small \((\sigma_t^2)\) implies that the difference between \(\sigma_t^2\) and \(\sigma_{t-1}^2\) is small \([14]\).

**Theorem 1.** Under the assumptions stated in Section 2.1, and for fixed half window-size \(k > 0\), then for any (order-2) pseudo-Lipschitz function \(\phi : \mathbb{R}^2 \to \mathbb{R}, \epsilon \in (0, 1), 0 \leq t < T^*\),

\[
P\left( \frac{1}{|\Gamma|} \sum_{i \in \Gamma} \left( \phi(\beta_{i}^{t+1}, \beta_i) - \mathbb{E}[\phi(\eta_t([\beta + \tau Z]_{\Lambda_i}), \beta_i)] \right) \geq \epsilon \right) \leq K_{k, t} e^{-\kappa_{k, t} n \epsilon^2},
\]

where \(\beta \in \mathbb{E} \sim \mu, Z \in \mathbb{R}^\Gamma\) has i.i.d. standard normal entries and is independent of \(\beta\), and the deterministic quantity \(\tau_t\) is defined in (11). The constants \(K_{k, t}, \kappa_{k, t} > 0\) do not depend on \(n\) or \(\epsilon\), but do depend on \(k\) and \(t\). Their values are not explicitly specified.

**Proof.** See Section 3. \(\square\)

**Remarks:**

1. The probability in (14) is w.r.t. the product measure on the space of the matrix \(A\), signal \(\beta\), and noise \(w\).

2. By choosing the following PL(2) loss function, \(\phi(a, b) = (a - b)^2\), Theorem 1 gives the following concentration result for the mean squared error of the estimates. For all \(t \geq 0\),

\[
P\left( \frac{1}{|\Gamma|} \|\beta_{i}^{t+1} - \beta\|^2 - \delta \sigma_{i}^{t+1} \geq \epsilon \right) \leq K_{k, t} e^{-\kappa_{k, t} n \epsilon^2},
\]

with \(\sigma_{i}^{t+1}\) defined in (11).
2.3 Numerical Examples

Before moving to the proof of Theorem 1, we first demonstrate the effectiveness of the AMP algorithm with sliding-window denoisers when used to reconstruct an image $\beta_0$ from its linear measurements acquired according to (1). We verify that state evolution accurately tracks the normalized estimation error of AMP, as is guaranteed by Theorem 1. While we use squared error as the error metric in our examples, which corresponds to the case where the PL(2) loss function $\phi$ in Theorem 1 is defined as $\phi(a, b) := (a - b)^2$, we remark that Theorem 1 also supports other PL(2) loss functions. Moreover, we apply AMP with sliding-window denoisers to reconstruct texture images, which are known to be well-modeled by MRFs in many cases [12, 13].

2.3.1 Verification of state evolution

We consider a class of stationary MRFs on $\mathbb{Z}^2$ whose neighborhood is defined as the eight-nearest neighbors, meaning this is a 2nd-order MRF per the definition in Section 2.1. The joint distribution of such an MRF on any finite $M \times N$ rectangular lattice in $\mathbb{Z}^2$ has the following expression [23]:

$$
\mu(x) = P(\beta = x) = \frac{\prod_{m=1}^{M-1} \prod_{n=1}^{N-1} \begin{bmatrix} x_{m,n} & x_{m,n+1} \\ x_{m+1,n} & x_{m+1,n+1} \end{bmatrix} \prod_{m=2}^{M-1} \prod_{n=2}^{N-1} \begin{bmatrix} x_{m,n} \\ x_{m+1,n} \end{bmatrix}}{\prod_{m=2}^{M-1} \prod_{n=1}^{N-1} \begin{bmatrix} x_{m,n} \\ x_{m,n+1} \end{bmatrix} \prod_{m=1}^{M-1} \prod_{n=2}^{N-1} \begin{bmatrix} x_{m,n} \\ x_{m+1,n} \end{bmatrix}},
$$

(15)

where we follow the notation in [23] for the generic measure $\begin{bmatrix} x_{m,n} & x_{m,n+1} \\ x_{m+1,n} & x_{m+1,n+1} \end{bmatrix}$ defined as

$$
\begin{bmatrix} x_{m,n} & x_{m,n+1} \\ x_{m+1,n} & x_{m+1,n+1} \end{bmatrix} := P(\beta_{m,n} = x_{m,n}, \beta_{m,n+1} = x_{m,n+1}, \beta_{m+1,n} = x_{m+1,n}, \beta_{m+1,n+1} = x_{m+1,n+1}),
$$

and the conditional distribution of the element in the box given the element(s) not in the box:

$$
\begin{bmatrix} x_{m,n} & x_{m,n+1} \\ x_{m+1,n} & x_{m+1,n+1} \end{bmatrix} := P(\beta_{m+1,n+1} = x_{m+1,n+1}|\beta_{m,n} = x_{m,n}, \beta_{m+1,n} = x_{m+1,n}, \beta_{m,n+1} = x_{m,n+1}).
$$

The generic measure needs to satisfy some consistency conditions to ensure the Markovian property and stationarity of the MRF on a finite grid; details can be found in [23]. For convenience, in simulations we use a $\Pi_+$ Binary MRF as defined in [23, Definition 7], for which the generic measure is conveniently parameterized by four parameters, namely,

$$
\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = p, \quad \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} = q, \quad \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} = r, \quad \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} = s.
$$

(16)

In the simulations, we set $\{p = 0.4, q = 0.5, r = 0.01, s = 0.4\}$. Using (9) and (10), it can be checked that the distribution measure of this MRF satisfies the Dobrushin uniqueness condition.

As mentioned previously, an attractive property of AMP, which is formally stated in Theorem 1, is the following: for large $n$ and $|\Gamma|$ and for $i \in \Gamma$, the observation vector $[A^* z^T + \beta']_{\Lambda_i}$ used as an input to the estimation function in (5) is approximately distributed as $\beta' + \tau_i Z'$, where $\beta' \sim \mu_\Lambda$, $Z'$ has i.i.d. standard normal entries, independent of $\beta'$, and $\tau_i$ is defined in (11). With this property in mind, a natural choice of denoiser functions $\{\eta_t\}_{t \geq 0}$ are those that calculate the conditional
Figure 3  From left to right: ground-truth image generated by the MRF described in Section 2.3.1, image reconstructed by AMP with a separable Bayesian denoiser (computed from the incorrect assumption that the signal is generated from an i.i.d. Bernoulli distribution), and image reconstructed by AMP with a Bayesian sliding-window denoiser with $k = 1$, hence $\Lambda = [3] \times [3]$. ($\Gamma = [128] \times [128], \delta = 0.5, \text{SNR} = 17 \text{dB}$.)

Figure 4  Numerical verification that the empirical MSE achieved by AMP with sliding-window denoisers is tracked by state evolution. The empirical MSE is averaged over 50 realizations of the MRF (as described in Section 2.3.1), measurement matrix, and measurement noise. ($\Gamma = [128] \times [128], \delta = 0.5, \text{SNR} = 17 \text{dB}$.)

The expectation of the signal given the value of the input argument, which we refer to as Bayesian sliding-window denoisers. Let $V_t = \beta' + \tau_t Z'$ and $v \in \mathbb{R}^\Lambda$, then for each $t \geq 0$ we define

$$\eta_t(v) := \mathbb{E} \left[ \beta'_c \big| V_t = v \right] = P \left( \beta'_c = 1 \big| V_t = v \right) = \frac{P \left( V_t = v, \beta'_c = 1 \right)}{P(V_t = v)}$$

(17)

where $x_c$ denotes the center coordinate of $x \in \mathbb{R}^\Lambda$, $x_{\Lambda \setminus c}$ denotes all coordinates in $x$ except the center, $f_{V_t|\beta'}(v|x) = \prod_{i \in \Lambda} \frac{1}{\sqrt{2\pi\tau_t}} \exp \left( -\frac{(v_i - \beta'_i)^2}{2\tau_t} \right)$ since coordinates of $Z'$ are i.i.d. normal, and $\mu(x)$ is computed according to (15) with $M = N = 2k + 1$ by using (16) and the property of $\Pi_+$ Binary MRF given in [23, Definition 7]. Figure 4 shows that the MSE achieved by AMP with the non-separable sliding-window denoiser defined above is tracked by state evolution at every iteration.

Notice that when $k = 0$, the denoisers $\{\eta_t\}_{t \geq 0}$ are separable and since the empirical distribution of $\beta_0$ converges to the stationary probability distribution $\mu_1$ on $E \subset \mathbb{R}$, the state evolution analysis for AMP with separable denoisers ($k = 0$) was justified by Bayati and Montanari [4]. However, it can be seen in Figures 3 and 4 that the MSE achieved by the separable denoiser ($k = 0$) is significantly higher (worse) than that achieved by the non-separable denoisers ($k = 1$).
2.3.2 Texture Image Reconstruction

We now use the Bayesian sliding-window denoiser defined in (17) to reconstruct binary texture images shown in Figure 5. The MRF prior is the same type as described in Section 2.3.1, namely the $\Pi_+$ Binary MRF, but we set the parameters $\{p = 0.18, q = 0.16, r = 0.034, s = 0.01\}$. Note that while it is possible to learn an MRF model for each of the images using well-established MRF learning algorithms, we do not include this procedure in our simulations since the study of texture image modeling is beyond the scope of this paper. Moreover, the reconstruction results obtained using the simple MRF defined above are sufficiently satisfactory, despite the fact that the prior may be inaccurate. In Figure 5, we begin with natural images of a cloud, a leaf, and wood (1st column) and then use thresholding to generate binary test images (2nd column). In addition to presenting the reconstructed images obtained by the Bayesian sliding-window denoisers with $k = 1$ (4th column) and $k = 0$ (5th column), respectively, we also present those obtained by AMP with a total variation denoiser [24] as a baseline approach (3rd column).

3 Proof of Theorem 1

The proof of Theorem 1 follows the work of Rush and Venkataramanan [14], with modifications for the dependent structure of the unknown vector $\beta$ in (1). For this reason, we use much of the same notation. We prove Theorem 1 using a technical lemma, Lemma 4, which corresponds to [14, Lemma 6]. Before stating the lemma, we cover some preliminary results and establish notation to be used in its proof.
3.1 Proof Notation

As in the previous work by Bayati and Montanari [4], as well as the work by Rush and Venkataramanan [14], the technical lemma is proved for a more general recursion, with AMP being a specific example of the general recursion as shown below. The connection between AMP and the general recursion will be explained in (25) and (26).

Fix the half-window size $0 \leq k \leq (N - 1)/2$, an integer. Let $\{f_t\}_{t \geq 0} : \mathbb{R}^{\Lambda \times A} \to \mathbb{R}$ and $\{g_t\}_{t \geq 0} : \mathbb{R}^2 \to \mathbb{R}$ be sequences of Lipschitz functions. Specifically, the arguments of $f_t$ are two variables in $\mathbb{R}^A$, for example, for $x, y \in \mathbb{R}^A$, we write $f_t(x, y)$ and call $x$ the first argument of $f_t$. Given noise $w \in \mathbb{R}^n$ and unknown signal $\beta \in E^\Gamma$, define vectors $h^{t+1}, q^{t+1} \in \mathbb{R}^{|\Gamma|}$ and $b^t, m^t \in \mathbb{R}^n$, as well as arrays $\hat{h}^{t+1}, \hat{q}^{t+1} \in \mathbb{R}^\Gamma$ (for which $h^{t+1}$ and $q^{t+1}$ are the vectorized versions) for $t \geq 0$ recursively as follows. Starting with initial condition $q^0 \in \mathbb{R}^\Gamma$:

\[
\begin{align*}
    h^{t+1} &= A^*m^t - \xi_0q^t, \quad q^t := \mathcal{V}(\hat{q}^t), \\
    \hat{h}^{t+1} &= \mathcal{V}^{-1}(h^{t+1}), \quad \hat{q}^t := f_t([\hat{h}^t]_\Lambda, \beta_\Lambda), \quad \text{for all } i \in \Gamma, \\
    b^t &= Aq^t - \lambda_t m^{t-1}, \quad m^t := g_t(b^t, w), \quad \text{for all } i \in [n],
\end{align*}
\]

with the scalars $\xi_t, \lambda_t$ defined as

\[
\begin{align*}
    \xi_t &= \frac{1}{n} \sum_{i=1}^n g_i(b^t_i, w_i), \quad \lambda_t &= \frac{1}{n} \sum_{i \in \Gamma} f_i([\hat{h}^t]_\Lambda, \beta_\Lambda),
\end{align*}
\]

where the derivative of $g_t$ is w.r.t. the first argument, and the derivative of $f_t$ is w.r.t. the center coordinate of the first argument. In the context of AMP, as made explicit in (25), the terms $\hat{h}^{t+1}$ and $\hat{q}^t$ measure the error in the observation $\mathcal{V}^{-1}(A^*z^t) + \beta^t$ and the estimate $\beta^t$ at time $t$, respectively, (the error w.r.t. the true $\beta$). The term $m^t$ measures the residual at time $t$ and the term $b^t$ is the difference between the noise and residual at time $t$.

Recall that the unknown vector $\beta \in E^\Gamma$ is assumed to have a stationary MRF prior with joint distribution measure $\mu$. Let $\beta \in E^\Gamma \sim \mu$ and $\mathbf{0} \in \mathbb{R}^A$ be an all-zero array. Define

\[
\begin{align*}
    \sigma_{\beta}^2 &= \mathbb{E}[\beta_0^2], \\
    \sigma_0^2 &= \frac{1}{\delta|\Gamma|} \sum_{i \in \Gamma} \mathbb{E}[f_i(\mathbf{0}, \beta_\Lambda)] > 0.
\end{align*}
\]

Further, for all $i \in \Gamma$ let

\[
\begin{align*}
    \hat{q}_i^0 &= f_i(\mathbf{0}, \beta_\Lambda), \quad \text{and} \quad q^0 := \mathcal{V}(\hat{q}^0),
\end{align*}
\]

and assume that there exist constants $K, \kappa > 0$ such that

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

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

\[
\begin{align*}
    \tau_t^2 &= \mathbb{E}[(g_t(\sigma_tZ, W))^2], \quad \sigma_t^2 := \frac{1}{\delta|\Gamma|} \sum_{i \in \Gamma} \mathbb{E}[(f_i(\tau_{t-1}Z_\Lambda, \beta_\Lambda))^2],
\end{align*}
\]

where random variables $W \sim p_w$ and $Z \sim \mathcal{N}(0, 1)$ are independent and random arrays $\beta \in E^\Gamma \sim \mu$ and $Z \in \mathbb{R}^\Gamma$ with i.i.d. $\mathcal{N}(0, 1)$ entries are also independent. We assume that both $\sigma_0^2$ and $\tau_0^2$ are
strictly positive. The technical lemma will show that \( \hat{h}^{t+1} \) can be approximated as i.i.d. \( \mathcal{N}(0, \tau_t^2) \) in functions of interest for the problem, namely when used as an input to PL functions, and \( b^t \) can be approximated as i.i.d. \( \mathcal{N}(0, \sigma_t^2) \) in PL functions. Moreover, it will be shown that the probability of the deviations of the quantities \( \frac{1}{n} \| m^t \|^2 \) and \( \frac{1}{n} \| q^t \|^2 \) from \( \tau_t^2 \) and \( \sigma_t^2 \), respectively, decay exponentially in \( n \).

We note that the AMP algorithm introduced in (4) and (5) is a special case of the general recursion of (18) and (19). Indeed, define the following vectors recursively for \( t \geq 0 \), starting with \( \beta^0 = 0 \) and \( z^0 = y \),

\[
\hat{h}^{t+1} = \beta - (\mathcal{N}(A^*z^t) + \beta^t), \quad \dot{q}^t = \beta^t - \beta, \\
b^t = w - z^t, \quad m^t = -z^t. \tag{25}
\]

It can be verified that these vectors satisfy (18) and (19) using Lipschitz functions

\[
f_t(a, \beta_t, a) = \eta_{t-1}(\beta_t - a) - \beta, \quad \text{and} \quad g_t(b, w_t) = b - w, \tag{26}
\]

where \( a \in \mathbb{R}^A \) and \( b \in \mathbb{R} \). Using the choice of \( f_t, g_t \) given in (26) also yields the expressions for \( \sigma_t^2, \tau_t^2 \) given in (11). In the remaining analysis, the general recursion given in (18) and (19) is used. Note that in AMP, \( q^0 = -\beta \) and \( \sigma_0^2 = \sigma_0^2 / \delta \), hence, assumption (23) for AMP requires

\[
P\left( \frac{1}{\| \beta \|^2} - \sigma_\beta^2 \right) \geq \delta \epsilon \leq K e^{-\kappa n \delta^2 \epsilon^2}. \tag{27}
\]

Under our assumptions for \( \beta \) as stated in Section 2.1, we see that (27) is satisfied using Lemma C.2 (Appendix C), since the function \( f(x) = x^2 \) is pseudo-Lipschitz. Finally, note that if we assume \( \sigma_\beta^2 > 0 \) and \( \delta < \infty \), then the condition of strict positivity of \( \sigma_\beta^2 \) and \( \tau_\beta^2 \) defined in (24) is satisfied.

Let \([c_1 \mid c_2 \mid \ldots \mid c_k]\) denote a matrix with columns \( c_1, \ldots, c_k \). For \( t \geq 1 \), define matrices

\[
M_t := [m^0 \mid \ldots \mid m^{t-1}], \quad Q_t := [q^0 \mid \ldots \mid q^{t-1}], \quad B_t := [b^0 \mid \ldots \mid b^{t-1}], \quad H_t := [h^0 \mid \ldots \mid h^t]. \tag{28}
\]

Moreover, \( M_0, Q_0, B_0, H_0 \) are defined to be the all-zero vector.

The values \( m^t_{\parallel} \) and \( q^t_{\parallel} \) are projections of \( m^t \) and \( q^t \) onto the column space of \( M_t \) and \( Q_t \), with \( m^t_{\parallel} := m^t - m^t_{\perp} \), and \( q^t_{\parallel} := q^t - q^t_{\perp} \) being the projections onto the orthogonal complements of \( M_t \) and \( Q_t \). Finally, define the vectors

\[
\alpha^t := (\alpha^t_0, \ldots, \alpha^t_{t-1})^*, \quad \gamma^t := (\gamma^t_0, \ldots, \gamma^t_{t-1})^*, \tag{29}
\]

to be the coefficient vectors of the parallel projections, i.e.,

\[
m^t_{\parallel} := \sum_{i=0}^{t-1} \alpha^t_i m^i, \quad q^t_{\parallel} := \sum_{i=0}^{t-1} \gamma^t_i q^i. \tag{30}
\]

The technical lemma, Lemma 4, shows that for large \( n \), the entries of the vectors \( \alpha^t \) and \( \gamma^t \) concentrate to constant values, which are defined in the following section.

### 3.2 Concentrating Constants

Recall that \( \beta \in E^{\Gamma} \) is the unknown vector to be recovered and \( w \in \mathbb{R}^m \) is the measurement noise. In this section we introduce the concentrating values for inner products of pairs of the vectors \( \{h^t, m^t, q^t, b^t\} \) that are used in Lemma 4.
Let \( \{ \tilde{Z}_t \}_{t \geq 0} \) be a sequence of zero-mean jointly Gaussian random variables taking values in \( \mathbb{R} \), and let \( \{ \tilde{Z}_t \}_{t \geq 0} \) be a sequence of zero-mean jointly Gaussian random arrays taking values in \( \mathbb{R}^\Gamma \). The covariance of the two random sequences is defined recursively as follows. For \( r, t \geq 0, i, j \in \Gamma \),

\[
\mathbb{E}[\tilde{Z}_r \tilde{Z}_t] = \frac{\tilde{E}_{r,t}}{\sigma_r \sigma_t}, \quad \mathbb{E}[\tilde{Z}_r[i] \tilde{Z}_t[j]] = \begin{cases} \frac{\tilde{E}_{r,t}}{\tau_r \tau_t}, & \text{if } i = j \\ 0, & \text{if } i \neq j, \end{cases}
\]

(31)

where

\[
\tilde{E}_{r,t} := \mathbb{E}\left[g_r(\sigma_r, \tilde{Z}_r, W) g_t(\sigma_t, \tilde{Z}_t, W)\right], \quad \tilde{E}_{r,t} := \frac{1}{\delta|\Gamma|} \sum_{i \in \Gamma} \mathbb{E}\left[ f_r(\tau_{r-1}[\tilde{Z}_{r-1}]_\Lambda_i, \beta_{\Lambda_i}) f_t(\tau_{t-1}[\tilde{Z}_{t-1}]_\Lambda_i, \beta_{\Lambda_i}) \right].
\]

(32)

Note that both terms of the above (32) are scalar values and we take \( f_0(\cdot, \beta_{\Lambda_i}) := f_0(0, \beta_{\Lambda_i}) \), the initial condition. Moreover, \( \tilde{E}_{r,t} = \sigma_r^2 \) and \( \tilde{E}_{t,r} = \tau_t^2 \), as can be seen from (24), thus for all \( i \in \Gamma \), we have \( \mathbb{E}[\tilde{Z}_i^2] = \mathbb{E}[\tilde{Z}_i^2] = 1 \). Therefore, \( \tilde{Z}_t \) has i.i.d. \( \mathcal{N}(0,1) \) entries.

Next, we define matrices \( \tilde{C}^t, \tilde{\sigma}^t \in \mathbb{R}^{\times t} \) and vectors \( \tilde{E}_t, \tilde{E}_t \in \mathbb{R}^t \) whose entries are \( \{ \tilde{E}_{r,t} \}_{r,t \geq 0} \) and \( \{ \tilde{E}_{r,t} \}_{r,t \geq 0} \) defined in (32): for \( 0 \leq i, j \leq t - 1 \),

\[
\tilde{C}^t_{i+1,j+1} := \tilde{E}_{i,j}, \quad \tilde{C}^t_{i+1,j+1} := \tilde{E}_{i,j},
\]

(33)

\[
\tilde{E}_t := (\tilde{E}_{0,t}, \ldots, \tilde{E}_{t-1,t})^*, \quad \tilde{E}_t := (\tilde{E}_{0,t}, \ldots, \tilde{E}_{t-1,t})^*.
\]

(34)

Lemma 1 below shows that \( \tilde{C}^t \) and \( \tilde{\sigma}^t \) are invertible. Therefore, we can define the concentrating values for \( \gamma^t \) and \( \alpha^t \) defined in (29) as

\[
\gamma^t := (\tilde{C}^t)^{-1} \tilde{E}_t \quad \text{and} \quad \alpha^t := (\tilde{\sigma}^t)^{-1} \tilde{E}_t,
\]

(35)

as well as the values of \( (\sigma^t)^2 \) and \( (\tau^t)^2 \) for \( t > 0 \):

\[
(\sigma^t)^2 := \sigma_t^2 - (\gamma^t)^* \tilde{E}_t = \tilde{E}_{t,t} - \tilde{E}_t^* (\tilde{C}^t)^{-1} \tilde{E}_t,
\]

(36)

\[
(\tau^t)^2 := \tau_t^2 - (\alpha^t)^* \tilde{E}_t = \tilde{E}_{t,t} - \tilde{E}_t^* (\tilde{\sigma}^t)^{-1} \tilde{E}_t.
\]

For \( t = 0 \), we let \( (\sigma^0)^2 := \sigma_0^2 \) and \( (\tau^0)^2 := \tau_0^2 \). Finally, define the concentrating values for \( \lambda_{t+1} \) and \( \xi_t \) defined in (19) as

\[
\xi_t := \mathbb{E}[g_0(\sigma_t, \tilde{Z}_t, W)], \quad \lambda_{t+1} := \frac{1}{\delta|\Gamma|} \sum_{i \in \Gamma} \mathbb{E}[f'_i(\tau_i [\tilde{Z}_t]_\Lambda_i, \beta_{\Lambda_i})].
\]

(37)

Lemma 1. If \( (\sigma^k)^2 \) and \( (\tau^k)^2 \) are bounded below by some positive constants for \( k \leq t \), then the matrices \( \tilde{C}^{k+1} \) and \( \tilde{\sigma}^{k+1} \) defined in (33) are invertible for \( k \leq t \).

Proof. The proof follows directly as that of [14, Lemma 1] and therefore is not restated here. To see that this is the case, note that the proof of [14, Lemma 1] relies only on the relationship between \( (\sigma^k)^2 \) (resp. \( (\tau^k)^2 \)) and \( \tilde{C}^k \) (resp. \( \tilde{\sigma}^k \)) as defined in [14, (4.19)], which is the same as (36), and not the actual values taken by these objects. Therefore, the proof for [14, Lemma 1] applies here. \( \square \)
3.3 Conditional Distribution Lemma

As mentioned previously, the proof of Theorem 1 relies on a technical lemma, Lemma 4, stated in Section 3.4 and proved in Section 4. Lemma 4 uses the conditional distribution of the vectors \( h^{t+1} \) and \( b^t \) given the matrices in (28) as well as \( \beta, w \). Two forms of the conditional distribution of \( h^{t+1} \) will be provided in Lemmas 2 and 3, which correspond to [14, Lemma 4] and [14, Lemma 5], respectively. Lemma 3 explicitly shows that the conditional distribution of \( h^{t+1} \) can be represented as the sum of a standard Gaussian vector and a deviation term, where the explicit expression of the deviation term is provided in Lemma 2. Then Lemma 4 shows that the deviation term is small, meaning that its normalized Euclidean norm concentrates on zero, and also provides concentration results for various inner products involving the other terms in recursion (18), namely \( \{h^{t+1}, q^t, b^t, m^t\} \).

The following notation is used. Considering two random variables \( X, Y \) and a sigma-algebra \( \mathcal{S} \), we denote the relationship that the conditional distribution of \( X \) with respect to \( \mathcal{S} \) equals the distribution of \( Y \) by \( X|\mathcal{S} \overset{d}= Y \). We represent a \( t \times t \) identity matrix as \( I_t \), dropping the \( t \) subscript when it is clear from the context. For a matrix \( A \) with full column rank, \( P_A^\parallel := A(A^*A)^{-1}A^* \) is the orthogonal projection matrix onto the column space of \( A \), and \( P_A^\perp := I - P_A^\parallel \). Define \( \mathcal{S}_{t_1,t_2} \) to be the sigma-algebra generated by the terms

\[
\{b^0, ..., b^{t_1-1}, m^0, ..., m^{t_1-1}, h^1, ..., h^{t_2}, q^0, ..., q^{t_2}, \beta, w\}.
\]

Lemma 2. For the vector \( h^{t+1} \) and \( b^t \) defined in (18), the following conditional distribution holds for \( t \geq 1 \):

\[
h^1|\mathcal{S}_{1,0} \overset{d}= \tau_0 Z_0 + \Delta_{1,0}, \quad \text{and} \quad h^{t+1}|\mathcal{S}_{t+1,t} \overset{d}= \sum_{r=0}^{t-1} \hat{\alpha}_r^t h^{r+1} + \tau_t^\perp Z_t + \Delta_{t+1,t},
\]

\[
b^0|\mathcal{S}_{0,0} \overset{d}= \sigma_0 Z_0^t + \Delta_{0,0}, \quad \text{and} \quad b^t|\mathcal{S}_{t,t} \overset{d}= \sum_{r=0}^{t-1} \hat{\gamma}_r^t b^r + \sigma_t^\perp Z_t^t + \Delta_{t,t}.
\]

where \( Z_0, Z_t \in \mathbb{R}^{|\Gamma|} \) and \( Z_0^t, Z_t^t \in \mathbb{R}^n \) are i.i.d. standard Gaussian random vectors that are independent of the corresponding conditioning sigma algebras. The term \( \hat{\gamma}_i^t \) and \( \hat{\alpha}_i^t \) for \( i = 0, ..., t - 1 \) is defined in (35) and the term \( (\tau_t^\perp)^2 \) and \( (\sigma_t^\perp)^2 \) in (36). The deviation terms are

\[
\Delta_{0,0} = \left( \frac{\|q^0\|}{\sqrt{n}} - \sigma_0 \right) Z_0^t,
\]

\[
\Delta_{1,0} = \left[ \left( \frac{\|m^0\|}{\sqrt{n}} - \tau_0 \right) 1_N - \frac{\|m^0\|}{\sqrt{n}} P_\parallel^\| \right] Z_0 + q^0 \left( \frac{\|q^0\|^2}{n} \right)^{-1} \left( \frac{(b^0)^* m_0}{n} - \xi_0 \frac{\|q^0\|^2}{n} \right),
\]

where \( I \) is the identity matrix and for any matrix \( A \), \( P_A^\parallel \) is the orthogonal projection matrix onto
the column space of $A$. For $t > 0$,

$$
\Delta_{t,t} = \sum_{r=0}^{t-1} (q_r^t - \hat{q}_r^t) b^r + \left[ \left( \frac{q_r^t}{\sqrt{n}} - \sigma^t_r \right) I_n - \frac{q_r^t}{\sqrt{n}} P_{M_t} \right] Z_t^r
+ M_t \left( \frac{M_t^*}{n} \right)^{-1} \left( \frac{H_t q_r^t}{n} - M_t^* \left[ \lambda_t m_{t-1} - \sum_{r=1}^{t-1} \lambda_r \gamma_r^{t-1} m_{r-1} \right] \right),
$$

(42)

$$
\Delta_{t+1,t} = \sum_{r=0}^{t-1} (\alpha_r^t - \hat{\alpha}_r^t) h^{r+1} + \left[ \left( \frac{m_r^t}{\sqrt{n}} - \tau^t_r \right) I_N - \frac{m_r^t}{\sqrt{n}} P_{Q_{t+1}} \right] Z_t^r
+ Q_{t+1} \left( \frac{Q_{t+1}^*}{n} \right)^{-1} \left( B_{t+1}^* m_{1}^t - \frac{Q_{t+1}^*}{n} \left( \xi_t q^t - \sum_{i=0}^{t-1} \xi_t \alpha_t^i q^i \right) \right).
$$

(43)

**Proof.** As in [14], the key theoretical insight in the proof is to study the distribution of $A$ conditioned on the sigma algebra $\mathcal{F}_{t+1,t}$ where $t_1$ is either $t+1$ or $t$, meaning one treats $A$ as random, and considers the output of the AMP algorithm up until the current iteration as fixed and given in the sigma-algebra. This is done by observing that conditioning on $\mathcal{F}_{t+1,t}$ is equivalent to conditioning on the linear constraints

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

(due to the relationships $b^s = Aq^s - \lambda_s m_{s-1}$ and $h^{r+1} = A^* m^r - \xi_r q^r$ for $0 \leq s \leq t$ and $0 \leq r \leq t-1$ given in (18)). Then it is straightforward to characterize the conditional distribution of a Gaussian matrix given linear constraints.

Since the sequences $\{\lambda_s; \xi_s; b^s; q^s; h^s; m^s\}_{0 \leq s \leq t}$ are all in the conditioning $\mathcal{F}_{t+1,t}$, this conditional distribution depends only on the relationship between the matrix $A$ and these fixed, given terms. This relationship, namely that specified via $b^s = Aq^s - \lambda_s m_{s-1}$ and $h^{r+1} = A^* m^r - \xi_r q^r$, is the same here as in [14, Lemma 3], and so the proofs are identical. We therefore do not repeat the details. Although $\{\lambda_s\}_{0 \leq s \leq t}$ and $\{q^s\}_{0 \leq s \leq t}$ are obtained from $\{f_s\}_{0 \leq s \leq t}$, which is separable in [14], but non-separable in our case, $\{\lambda_s\}_{0 \leq s \leq t}$ and $\{q^s\}_{0 \leq s \leq t}$ are simply treated as fixed elements in the conditioning sigma-algebra $\mathcal{F}_{t+1,t}$ and the fact that they are calculated via non-separable functions here does not change the proof.

Then one is able to specify the conditional distributions of $b^t$ and $h^{t+1}$ given $\mathcal{F}_{t,t}$ and $\mathcal{F}_{t+1,t}$, respectively, using the conditional distribution of $A$. Again since the relationship between $b^t$ and $h^{t+1}$ and $A$ is the same here as in [14], the details are identical to that provided in the proof of [14, Lemma 4] and are not repeated here.

Note that Lemma 2 holds only when $Q_{t+1}^* Q_{t+1}$ is invertible. The following lemma provides an alternative representation of the conditional distribution of $h^{t+1}_{t+1,t}$ for $t \geq 0$, and it explicitly shows that $h^{t+1}_{t+1,t}$ is distributed as an i.i.d. Gaussian random vector with $N(0, \tau_t^2)$ entries plus a deviation term.

**Lemma 3.** For $t \geq 0$, let $Z_t \in \mathbb{R}^{[1]}$ be i.i.d. standard normal random vectors. Let $h^t_{\text{pure}} := \tau_0 Z_0$. For $t \geq 1$, recursively define

$$
h^t_{\text{pure}} = \sum_{r=0}^{t-1} \hat{\alpha}_r^t h^{r+1} + \tau^t_t Z_t
$$

(44)

and a set of scalars $\{d^t_i\}_{0 \leq i \leq t}$ with $d^0_0 = 1$,

$$
d^t_i = \sum_{r=i}^{t-1} d^i_r \hat{\alpha}_r^t \quad \text{for} \quad 0 \leq i \leq (t-1), \quad \text{and} \quad d^t_t = 1.
$$

(45)
Let $\hat{h}_{\text{pure}}^{t+1} = V^{-1}(h_{\text{pure}}^{t+1}) \in \mathbb{R}^\Gamma$. Then for all $t \geq 0$ we have

$$
(\hat{h}_{\text{pure}}^1, \ldots, \hat{h}_{\text{pure}}^{t+1}) \overset{d}{=} (\tau_0 \tilde{Z}_0, \ldots, \tau_t \tilde{Z}_t),
$$

where $\{\tilde{Z}_t\}_{t \geq 0}$ are jointly Gaussian with correlation structure defined in (31). Moreover,

$$
|\hat{h}_{\text{pure}}^{t+1}|_{\mathcal{F}_{t+1,t}} \overset{d}{=} h_{\text{pure}}^{t+1} + \sum_{r=0}^{t} d_r \Delta r_{r+1},
$$

(47)

Proof. First, we prove (46) by induction. For $t = 1$, $\hat{h}_{\text{pure}}^1 = \tau_0 V^{-1}(Z_0) \overset{d}{=} \tau_0 \tilde{Z}_0$. As the inductive hypothesis, assume $(\hat{h}_{\text{pure}}^1, \ldots, \hat{h}_{\text{pure}}^{t}) \overset{d}{=} (\tau_0 \tilde{Z}_0, \ldots, \tau_{t-1} \tilde{Z}_{t-1})$. By (44), term $\hat{h}_{\text{pure}}^{t+1}$ is equal in distribution to $\sum_{r=0}^{t} \alpha_r \tau_r \tilde{Z}_r + \tau_t^1 Z$, where $Z \in \mathbb{R}^\Gamma$ is independent of $\tilde{Z}_r$ for all $r = 0, \ldots, t - 1$. In what follows, we show

$$(\tau_0 \tilde{Z}_0, \ldots, \tau_{t-1} \tilde{Z}_{t-1}, \sum_{r=0}^{t-1} \alpha_r \tau_r \tilde{Z}_r + \tau_t^1 Z) \overset{d}{=} (\tau_0 \tilde{Z}_0, \ldots, \tau_{t-1} \tilde{Z}_{t-1}, \tau_t \tilde{Z}_t).$$

Note that $\tilde{Z}_0, \ldots, \tilde{Z}_{t-1}, Z$ are all zero-mean Gaussian, and therefore so is the sum. We now study the variance and covariance of $\sum_{r=0}^{t-1} \alpha_r \tau_r \tilde{Z}_r + \tau_t^1 Z$ by demonstrating the following two results:

(i) For all $i, j \in \Gamma$,

$$
E \left[ \left( \sum_{r=0}^{t-1} \alpha_r \tau_r \tilde{Z}_r \right)_{i} \left( \sum_{r=0}^{t-1} \alpha_r \tau_r \tilde{Z}_r \right)_{j} \right] = \tau_t^2 E \left[ [\tilde{Z}_t], [\tilde{Z}_t] \right] = \begin{cases} \tau_t^2 & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases}
$$

(ii) For $0 \leq s \leq (t - 1)$ and all $i, j \in \Gamma$,

$$
E \left[ \tau_s [\tilde{Z}_s]_{i} \left( \sum_{r=0}^{t-1} \alpha_r \tau_r \tilde{Z}_r \right)_{j} \right] = \tau_s \tau_t E \left[ [\tilde{Z}_s], [\tilde{Z}_t] \right] = \begin{cases} \hat{E}_{s,t} & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases}
$$

First consider (i). We note,

$$
E \left[ \left( \sum_{r=0}^{t-1} \alpha_r \tau_r \tilde{Z}_r \right)_{i} \left( \sum_{r=0}^{t-1} \alpha_r \tau_r \tilde{Z}_r \right)_{j} \right] \overset{(a)}{=} \sum_{r=0}^{t-1} \sum_{s=0}^{t-1} \alpha_r \alpha_s \tau_r \tau_s E \left[ [\tilde{Z}_r], [\tilde{Z}_s] \right] + (\tau_t^1)^2 E \left[ Z, Z \right] \overset{(b)}{=} \sum_{r=0}^{t-1} \sum_{s=0}^{t-1} \alpha_r \alpha_s \tau_r \tau_s \tau_t^2, \quad \text{if } i = j,
$$

0, otherwise.

In the above, step (a) follows from the fact that $Z$ is independent of $\tilde{Z}_0, \ldots, \tilde{Z}_{t-1}$, step (b) from the covariance definition (31) and the i.i.d. standard normal nature of elements of $\tilde{Z}$, and step (c) from

$$
\sum_{r=0}^{t-1} \sum_{l=0}^{t-1} \alpha_r \alpha_l \hat{E}_{r,l} = (\hat{\alpha}^t)^* \hat{C}^t \hat{\alpha}^t = [\hat{C}^* (\hat{C}^t)^{-1}] (\hat{C}^t)^{-1} [\hat{C}^* (\hat{C}^t)^{-1} \hat{E}_t] = \hat{E}_t^* (\hat{C}^t)^{-1} \hat{E}_t = \hat{E}_{t,t} - (\tau_t^1)^2.
$$

Next, consider (ii). We see that

$$
E \left[ \tau_s [\tilde{Z}_s]_{i} \left( \sum_{r=0}^{t-1} \alpha_r \tau_r \tilde{Z}_r \right)_{j} \right] \overset{(a)}{=} \sum_{r=0}^{t-1} \alpha_r \tau_s \tau_r E \left[ [\tilde{Z}_s], [\tilde{Z}_r] \right] \overset{(b)}{=} \begin{cases} \sum_{r=0}^{t-1} \hat{E}_{s,r} \alpha_r \alpha_r \tau_r & \text{if } i = j, \\ 0, & \text{otherwise.} \end{cases}
$$


In the above, step (a) follows since \( Z \) is independent of \( \tilde{Z}_s \) and step (b) from (31). Finally, notice that \( \sum_{r=0}^{t-1} E_{s,r} \hat{\alpha}_r = [\hat{C}^{t} \hat{\alpha}^T]_{s+1} = \hat{E}_{s,t} \), where the first equality holds since the sum equals the inner product of the \((s+1)^{th}\) row of \( \hat{C}^t \) with \( \hat{\alpha}^T \) and the second equality by definition of \( \hat{\alpha}^T \) in (35).

Next, we prove (47), also by induction. For \( t = 0 \), by (38) we have \( h_{t+1}^{r+1}_{|\mathcal{A}_{t+1}} \overset{d}{=} \tau_0 Z_t + \Delta_{t,0} \overset{d}{=} h_{\text{pure}}^{t+1} + \Delta_{t,0} \). Assume that \( h_{r+1}^{t+1}_{|\mathcal{A}_{t+1}} \overset{d}{=} h_{r+1}^{t+1}_{\text{pure}} + \sum_{i=0}^{r} d_{i}^{t} \Delta_{i+1,i} \) holds for \( r = 0, \ldots, t - 1 \) as the inductive hypothesis. Then,

\[
h_{t+1}^{r+1}_{|\mathcal{A}_{t+1}} \overset{d}{=} \sum_{r=0}^{t-1} \hat{\alpha}_r^{t+1} Z_t + \Delta_{t+1,t}^{r+1} = \sum_{r=0}^{t-1} \hat{\alpha}_r^{t+1} \left( h_{r+1}^{t+1}_{\text{pure}} + \sum_{i=0}^{r} d_{i}^{t} \Delta_{i+1,i} \right) + \Delta_{t+1,t}^{r+1} Z_t + \Delta_{t+1,t}
\]

In the above, the first equality uses (38) and the second the inductive hypothesis. The last equality follows since \( \Delta_{t+1,t}^{r+1} = 0, \) by (38) and the inductive hypothesis. The last equality follows by noticing that \( \sum_{r=0}^{t-1} \sum_{i=0}^{r} d_{i}^{t} \Delta_{i+1,i} \) for \( (v_i, r)_{0 \leq i, r \leq t - 1} \) and using (45). \( \square \)

### 3.4 Main Concentration Lemma

**Lemma 4.** We use the shorthand \( X_n \doteq c \) to denote the concentration inequality \( P(|X_n - c| \geq \epsilon) \leq K_{k,t} e^{-\kappa_{k,t} n \epsilon^2} \), where \( K_{k,t}, \kappa_{k,t} \) denote constants depending on the iteration index \( t \) and the fixed half-window size \( k \), but not on \( n \) or \( \epsilon \). The following statements hold for \( 0 \leq t < T^* \) and \( \epsilon \in (0,1) \).

(a) For \( \Delta_{t+1,t} \) defined in (41) and (43),

\[
P \left( \frac{1}{|\Gamma|} \| \Delta_{t+1,t} \|^2 \geq \epsilon \right) \leq K_{k,t} e^{-\kappa_{k,t} n \epsilon^2}.
\]

(b) For (order-2) pseudo-Lipschitz functions \( \phi_k : \mathbb{R}^{(t+2)|\Lambda|} \rightarrow \mathbb{R} \),

\[
\frac{1}{|\Gamma|} \sum_{i \in \Gamma} \phi_k \left( [\hat{h}^t]_{\Lambda_i}, \ldots, [\hat{h}^{t+1}]_{\Lambda_i}, \beta_{\Lambda_i} \right) \overset{d}{=} \frac{1}{|\Gamma|} \sum_{i \in \Gamma} \mathbb{E} \left[ \phi_k \left( \tau_0 [\tilde{Z}_0]_{\Lambda_i}, \ldots, \tau_t [\tilde{Z}_t]_{\Lambda_i}, \beta_{\Lambda_i} \right) \right]. \tag{49}
\]

The random vectors \( \tilde{Z}_0, \ldots, \tilde{Z}_t \in \mathbb{R}^{\Gamma} \) are jointly Gaussian with zero mean entries, which are independent of the other entries in the same vector with covariance across iterations given by (31), and are independent of \( \beta \sim \mu \).

(c) Recall that the operator \( \mathcal{V} \) rearranges the elements of an array into a vector,

\[
\frac{(h^{t+1}_{|\Lambda}) q^0}{n} \overset{d}{=} 0, \quad \frac{(h^{t+1}_{|\Lambda}) \mathcal{V}(\beta)}{n} \overset{d}{=} 0, \quad \frac{(b^t)^* w}{n} \overset{d}{=} 0. \tag{50}
\]

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

\[
\frac{(h^{r+1})^{t+1}}{|\Gamma|} \overset{d}{=} \tilde{E}_{r,t}, \quad \frac{(b^r)^* b^t}{n} \overset{d}{=} \tilde{E}_{r,t}. \tag{52}
\]

19
(e) For all \(0 \leq r \leq t\),
\[
\frac{(q_r^0)^*q_r^{t+1}}{n} \equaln \tilde{E}_{0,t+1}, \quad \frac{(q_r^{t+1})^*q_r^{t+1}}{n} \equaln \tilde{E}_{r+1,t+1},
\]
(54)
\[
\frac{(m^r)^*m^t}{n} \equaln \tilde{E}_{r,t}.
\]
(55)

(f) For all \(0 \leq r \leq t\),
\[
\lambda_t \equaln \hat{\lambda}_t, \quad \frac{(h_t^{t+1})^*q_r^{t+1}}{n} \equaln \hat{\lambda}_{r+1}\tilde{E}_{r,t}, \quad \frac{(h_t^{r+1})^*q_r^{t+1}}{n} \equaln \hat{\lambda}_{t+1}\tilde{E}_{r,t},
\]
(56)
\[
\xi_t \equaln \hat{\xi}_t, \quad \frac{(b^r)^*m^t}{n} \equaln \hat{\xi}_t\tilde{E}_{r,t}, \quad \frac{(b^t)^*m^t}{n} \equaln \hat{\xi}_t\tilde{E}_{r,t}.
\]
(57)

(g) For \(Q_{t+1} = \frac{1}{n}Q_{t+1}Q_{t+1}\) and \(M_t = \frac{1}{n}\sigma_t M_t^\top M_t\), when the inverses exist, for all \(0 \leq i, j \leq t\) and \(0 \leq t', t'' \leq t - 1\):
\[
[Q_{t+1}]_{i+1,j+1} \equaln [(\tilde{C}_{t+1})^{-1}]_{i+1,j+1}, \quad \gamma_t^{i+1} \equaln \hat{\gamma}_t^{i+1},
\]
(58)
\[
[M_t]_{i',j'} \equaln [(\tilde{C}_{t})^{-1}]_{i'+1,j'+1}, \quad \alpha_t^{i'} \equaln \hat{\alpha}_t^{i'}, \quad t \geq 1,
\]
where \(\hat{\gamma}_t^{i+1}\) and \(\hat{\alpha}_t^{i'}\) are defined in (35),

(h) With \(\sigma_t^{1+1}, \tau_t^{1+1}\) defined in (36),
\[
\frac{1}{n} \norm{q_t^{t+1}}^2 \equaln (\sigma_t^{1+1})^2,
\]
(60)
\[
\frac{1}{n} \norm{m_t^t}^2 \equaln (\tau_t^{1+1})^2.
\]
(61)

3.5 Proof of Theorem 1

Proof. Applying part (b) of Lemma 4 to a PL(2) function \(\phi_h : \mathbb{R}N \to \mathbb{R}\),
\[
P\left(\frac{1}{\norm{\Gamma}} \sum_{i \in \Gamma} \left(\phi_h(h_{\Lambda_i}^{t+1}, \beta_{\Lambda_i}) - E[\phi_h(\tau Z_{\Lambda_i}, \beta_{\Lambda_i})]\right) \geq \epsilon\right) \leq K_{k,te^{-\kappa_k n^2/4}},
\]
where the random field \(\beta \in E\Gamma \sim \mu\) is independent of \(Z \in \mathbb{R}\Gamma\) having i.i.d. standard normal entries. Now for \(i \in \Gamma\) let
\[
\phi_h(h_{\Lambda_i}^{t+1}, \beta_{\Lambda_i}) := \phi(\eta_t(\beta_{\Lambda_i} - h_{\Lambda_i}^{t+1}), \beta_i),
\]
(62)
where \(\phi : \mathbb{R} \to \mathbb{R}\) is the PL(2) function in the statement of the theorem. The function \(\phi_h(h_{\Lambda_i}^{t+1}, \beta_{\Lambda_i})\) in (62) is PL(2) since \(\phi\) is PL(2) and \(\eta_t\) is Lipschitz. We therefore obtain
\[
P\left(\frac{1}{\norm{\Gamma}} \sum_{i \in \Gamma} \left(\phi(\eta_t(\beta_{\Lambda_i} - h_{\Lambda_i}^{t+1}), \beta_i) - E[\phi(\eta_t(\beta_{\Lambda_i} - \tau Z_{\Lambda_i}), \beta_i)]\right) \geq \epsilon\right) \leq K_{k,te^{-\kappa_k n^2/4}}.
\]
The proof is completed by noting from (5) and (25) that \(\beta_t^{i+1} = \eta_t(\gamma^{-1}(A^sz^t) + \beta_t^t)_{\Lambda_i} = \eta_t(\beta_{\Lambda_i} - h_{\Lambda_i}^{t+1})\).
4 Proof of Lemma 4

We make use of concentration results listed in Appendices A, B, and C, where Appendix C contains concentration results for dependent variables that were needed to provide the new results in this paper. Note that the lemmas that are stated in Appendices are labeled by capital letters with numbers (e.g., Lemma A.1), whereas the lemmas that are stated in the body are labeled by numbers (e.g., Lemma 1).

The proof of Lemma 4 proceeds by induction on \( t \). We label as \( \mathcal{H}^{t+1} \) the results (48), (49), (50), (52), (54), (56), (58), (60) and similarly as \( B^t \) the results (51), (53), (55), (57), (59), (61). The proof consists of four steps: (1) proving that \( B_0 \) holds, (2) proving that \( \mathcal{H}_1 \) holds, (3) assuming that \( B_r \) and \( \mathcal{H}_s \) hold for all \( r < t \) and \( s \leq t \), then proving that \( B_t \) holds, and (4) assuming that \( B_r \) and \( \mathcal{H}_s \) hold for all \( r \leq t \) and \( s \leq t \), then proving that \( \mathcal{H}_{t+1} \) holds.

The proof of steps (1) and (3) – the \( B \) steps – follow as in [14]. To see that this is the case, notice that in the proof for [14, Lemma 6], the results given by \( B_0(b) - B_t(h) \) involve the sequence of functions \( \{g_r\}_{0 \leq r \leq t} \) and the sequence of vectors \( \{b^r\}_{0 \leq r \leq t} \). In our case, the definition of the separable functions \( g_t(\cdot) \) in (18) is the same that in [14, (4.1)] and the conditional distribution of \( \{b^r\}_{0 \leq r \leq t} \) given in Lemma 2 has the same expression as that in [14, Lemma 6]. Therefore, the proof of \( B_0(b) - B_t(h) \) in [14] is directly applicable here.

Now consider \( B_t(a) \). When \( t = 0 \), it only involves \( \|q^0\| \), which has the same assumption in our case and in [14]. When \( t > 0 \), the proof uses induction hypothesis \( B_0(d) - B_{t-1}(d), B_{t-1}(e) \), and \( \mathcal{H}_t(g), \mathcal{H}_t(h) \). The statement of those hypotheses have the same form as in [14]. Although the concentration constants in those hypotheses have different definitions in this paper due to using non-separable denoisers, it does not change the proof for \( B_t(a) \), since the actual values of the concentration constants are not involved in the proof. Therefore, we do not repeat the proof of steps (1) and (3) here. In what follows, we only show steps (2) and (4).

For each step, in parts (a)–(h) of the proof, we use \( K \) and \( \kappa \) to label universal constants, meaning that they do not depend on \( n \) or \( \epsilon \), but may depend on \( t \) and \( k \), in the concentration upper bounds.

4.1 Step 2: Showing that \( \mathcal{H}_1 \) holds

Throughout the proof we will make use of a function \( S : \mathbb{R}^\Lambda \rightarrow \mathbb{R} \) that selects the center coordinate of its argument. For example, for \( v \in \mathbb{R}^{|\Lambda|} \),

\[
S(v_{\Lambda}) = v_i.
\]

We will only use \( S \) in cases where such a “center point” is well-defined. Notice that \( S \) is Lipschitz, since \( |S(x) - S(x')| = |x_c - x'_c| \leq \|x - x'\| \), for all \( x, x' \in \mathbb{R}^\Lambda \), where \( x_c \) (respectively, \( x'_c \)) is the center coordinate of \( x \) (respectively, \( x' \)). Moreover, if a function \( f : \mathbb{R}^\Lambda \rightarrow \mathbb{R} \) is defined as \( f(x, y) := S(x) \) with arbitrary but fixed \( \Lambda \), then \( f \) is Lipschitz, because \( |f(x, y) - f(x', y')| = |S(x) - S(x')| \leq \|x - x'\| \leq \|(x, y) - (x', y')\| \). We are now ready to prove \( \mathcal{H}_1(a) - \mathcal{H}_1(h) \).

(A) The definition of \( \Delta_{1,0} \) is given in (41). First notice that by Lemma B.1, we have \( \mathcal{B}_{q^0} Z_0 \overset{d}{=} \frac{q^0}{\|q^0\|^2} Z_0 \), where \( Z_0 \) is a standard normal random variable on \( \mathbb{R} \). Using this fact and (41), we have

\[
\Delta_{1,0} \overset{d}{=} \left( \frac{\|m^0\|}{\sqrt{n}} - \tau_0 \right) Z_0 - \frac{\|m^0\|}{\sqrt{n}} \left( \frac{q^0}{\|q^0\|^2} \right) Z_0 + q^0 \left( \frac{n}{\|q^0\|^2} \right) \left( \frac{(b^0)^* m^0}{n} - \frac{\xi_0 q^0}{n} \right).
\]
By applying the triangle inequality to the norm of the RHS of (64) and then applying Lemma A.1,
\[
P\left(\frac{\|\Delta_{1,0}\|^2}{|1|} \geq \epsilon\right) = P\left(\frac{\|\Delta_{1,0}\|}{\sqrt{|1|}} \geq \sqrt{\epsilon}\right)
\leq P\left(\frac{\|m_0\|}{\sqrt{n}} - \tau_0\left(\frac{\|Z_0\|}{\sqrt{|1|}} - 1\right) + 1 \geq \frac{\sqrt{\epsilon}}{3}\right) + P\left(\frac{\|Z_0\|}{\sqrt{|1|}} - 1 \geq \frac{\sqrt{\epsilon}}{3}\right)
\]
(65)
Label the three terms on the right-hand side (RHS) of (65) as \(T_1 - T_3\). We will show that each term is bounded by \(Ke^{-\kappa n\epsilon}\).

First consider \(T_1\).
\[
T_1 \leq P\left(\frac{\|m_0\|}{\sqrt{n}} - \tau_0\left(\frac{\|Z_0\|}{\sqrt{|1|}} - 1\right) + 1 \geq \frac{\sqrt{\epsilon}}{3}\right)
\leq P\left(\|m_0\| - \tau_0 \geq \frac{\sqrt{\epsilon}}{6}\right) + P\left(\|Z_0\| - 1 \geq \frac{\sqrt{\epsilon}}{3}\right)
\leq Ke^{-\kappa n\epsilon} + Ke^{-\kappa n\epsilon},
\]
(66)
where step (b) follows by Lemma A.3, Lemma A.7, and \(\mathcal{B}_0(\epsilon)\). To see that step (a) in (66) holds, we notice that
\[
\left\{\left|\frac{\|m_0\|}{\sqrt{n}} - \tau_0\right| < \frac{\sqrt{\epsilon}}{6}\right\} \cap \left\{\left|\frac{\|Z_0\|}{\sqrt{|1|}} - 1\right| < \frac{\sqrt{\epsilon}}{3}\right\} \subset \left\{\left|\frac{\|m_0\|}{\sqrt{n}} - \tau_0\right|\left(\left|\frac{\|Z_0\|}{\sqrt{|1|}} - 1\right| + 1\right) < \frac{\sqrt{\epsilon}}{3}\right\}
\]
(67)
since if the two events on the left-hand side (LHS) of (67) hold, then using that \(\epsilon < 1\),
\[
\left|\frac{\|m_0\|}{\sqrt{n}} - \tau_0\right|\left(\left|\frac{\|Z_0\|}{\sqrt{|1|}} - 1\right| + 1\right) < \frac{\sqrt{\epsilon}}{6}\left(\frac{\sqrt{\epsilon}}{3} + 1\right) < \frac{\sqrt{\epsilon}}{3}.
\]
Taking the complement on both sides of (67),
\[
\left\{\left|\frac{\|m_0\|}{\sqrt{n}} - \tau_0\right|\left(\left|\frac{\|Z_0\|}{\sqrt{|1|}} - 1\right| + 1\right) \geq \frac{\sqrt{\epsilon}}{3}\right\} \subset \left\{\left|\frac{\|m_0\|}{\sqrt{n}} - \tau_0\right| > \frac{\sqrt{\epsilon}}{6}\right\} \cup \left\{\left|\frac{\|Z_0\|}{\sqrt{|1|}} - 1\right| \geq \frac{\sqrt{\epsilon}}{3}\right\}.
\]
Then step (a) in (66) follows by the union bound.

Next consider \(T_2\).
\[
T_2 \leq P\left(\left|\frac{\|m_0\|}{\sqrt{n}} - \tau_0\right| + \tau_0\left|\frac{\|Z_0\|}{\sqrt{|1|}} - 1\right| \geq \frac{\sqrt{\epsilon}}{3}\right)
\leq P\left(\|m_0\| - \tau_0 \geq \frac{\sqrt{\epsilon}}{3}\right) + P\left(\|\hat{Z}_0\| - 1 \geq \frac{\sqrt{\epsilon}}{3}\right)
\leq Ke^{-\kappa n\epsilon} + Ke^{-\kappa n\epsilon},
\]
where step (a) follows by similar justification as that for step (a) in (66) and step (b) follows by Lemma A.3, Lemma A.6, and \(\mathcal{B}_0(\epsilon)\).

Finally consider \(T_3\).
\[
T_3 \leq P\left(\left|\frac{(b^0)^*m_0}{n}\frac{\sqrt{n}}{\|q_0\|} - \hat{\xi}_0\sigma_0\right| \geq \frac{\sqrt{\epsilon}}{6\sqrt{\delta}}\right) + P\left(\xi_0\frac{\|q_0\|}{\sqrt{n}} - \hat{\xi}_0\sigma_0 \geq \frac{\sqrt{\epsilon}}{6\sqrt{\delta}}\right)
\leq Ke^{-\kappa n\epsilon} + Ke^{-\kappa n\epsilon},
\]
where step (a) follows by Lemma A.1 and step (b) follows by Lemma A.2, \(\mathcal{B}_0(f)\), and the assumption on \(\|q_0\|\) given in (23).
(b) Let \( \hat{Z}_0 := \mathcal{V}^{-1}(Z_0) \in \mathbb{R}^\Gamma \) and \( \hat{\Delta}_{1,0} := \mathcal{V}^{-1}(\Delta_{1,0}) \in \mathbb{R}^\Gamma \) be the array versions of the vectors \( Z_0 \) and \( \Delta_{1,0} \), respectively. For \( t = 0 \), the left-hand side (LHS) of (49) can be bounded as

\[
P\left( \left| \frac{1}{|\Gamma|} \sum_{i \in \Gamma} \left( \phi_h(\tilde{h}_1^i, \beta_{\Lambda_i}) - \mathbb{E}[\phi_h(\tau_0|\tilde{Z}_0\Lambda_i, \beta_{\Lambda_i})] \right) \right| \geq \epsilon \right)
\]

\[
\leq P\left( \left| \frac{1}{|\Gamma|} \sum_{i \in \Gamma} \left( \phi_h(\tau_0\hat{Z}_0 + \hat{\Delta}_{1,0}\Lambda_i, \beta_{\Lambda_i}) - \mathbb{E}[\phi_h(\tau_0|\hat{Z}_0\Lambda_i, \beta_{\Lambda_i})] \right) \right| \geq \frac{\epsilon}{3} \right) + P\left( \left| \frac{1}{|\Gamma|} \sum_{i \in \Gamma} \left( \phi_h(\tau_0\hat{Z}_0 + \hat{\Delta}_{1,0}\Lambda_i, \beta_{\Lambda_i}) - \mathbb{E}[\phi_h(\tau_0|\hat{Z}_0\Lambda_i, \beta_{\Lambda_i})] \right) \right| \geq \frac{\epsilon}{3} \right).
\]

Step (a) follows from the conditional distribution of \( h^1 \) given in Lemma 2 (38) and since \( \hat{h}^1 = \mathcal{V}^{-1}(h^1) \) and \( \tau_0\hat{Z}_0 + \hat{\Delta}_{1,0} = \mathcal{V}^{-1}(\tau_0Z_0 + \Delta_{1,0}) \). Step (b) follows from Lemma A.1. Label the terms on the RHS of (68) as \( T_1 - T_3 \). We show that each of these terms is bounded above by \( K e^{-\kappa n^{2\epsilon}} \).

First, consider \( T_1 \). Recall the definition of the functions \( \mathcal{T}_i \) for \( i \in \Gamma \) in (8), which extends an array in \( \mathbb{R}^{\Lambda_i, \Gamma} \) to an array in \( \mathbb{R}^{\Lambda} \) by defining the extended entries to be the average of the entries in the original array. For arbitrary but fixed \( s \in \mathbb{R}^{\Lambda} \), the function \( \tilde{\phi}_{\hat{h},i}(v, s) := \phi_h(\mathcal{T}_i(v), s) \) is PL(2) by Lemma B.5. Then it follows from Lemma B.4 that the function \( \phi_{1,i} : \mathbb{R}^{\Lambda} \to \mathbb{R} \) defined as \( \phi_{1,i}(s) := \mathbb{E}_{\hat{Z}_0}[\tilde{\phi}_{\hat{h},i}(\tau_0|\hat{Z}_0\Lambda_i, \Gamma, s)] \) is PL(2), since \( [\hat{Z}_0\Lambda_i, \Gamma] \) is an array of i.i.d. standard norm random variables for all \( i \in \Gamma \). Notice that \( \mathbb{E}_{\hat{Z}_0}[\tilde{\phi}_{\hat{h},i}(\tau_0|\hat{Z}_0\Lambda_i, \Gamma, s)] = \mathbb{E}_{\hat{Z}_0}[\phi_h(\tau_0|\hat{Z}_0\Lambda_i, s)] \) by the definition of \( \tilde{\phi}_{\hat{h},i} \) and \( \mathcal{T}_i \). Therefore,

\[
T_1 \overset{(a)}{=} P\left( \left| \frac{1}{|\Gamma|} \sum_{i \in \Gamma} \left( \phi_{1,i}(\mathcal{T}_i(\beta_{\Lambda_i}, \Gamma)) - \mathbb{E}[\phi_{1,i}(\mathcal{T}_i(\beta_{\Lambda_i}, \Gamma))] \right) \right| \geq \epsilon \right) \overset{(b)}{\leq} Ke^{-\kappa n^{2\epsilon}},
\]

where in step (a) we use the definition of \( \mathcal{T}_i \) in (8) and step (b) follows from Lemma C.2 by noticing from Lemma B.5 that the function \( \phi_{1,i}(\mathcal{T}_i(\cdot)) \) is PL(2) for all \( i \in \Gamma \).

Next, consider \( T_2 \). We use iterated expectation to condition on the value of \( \beta \). Then \( T_2 \) an be expressed as an expectation as follows,

\[
T_2 = \mathbb{E}_{\beta}\left[ P\left( \left| \frac{1}{|\Gamma|} \sum_{i \in \Gamma} \left( \phi_h(\tau_0|\hat{Z}_0\Lambda_i, \beta_{\Lambda_i}) - \mathbb{E}_{\hat{Z}_0}[\phi_h(\tau_0|\hat{Z}_0\Lambda_i, \beta_{\Lambda_i})] \right) \right| \geq \frac{\epsilon}{3} \mid \beta \right) \right].
\]

Define the function \( f : \mathbb{R}^{\Gamma} \to [0, 1] \) as

\[
f(a) := P\left( \left| \frac{1}{|\Gamma|} \sum_{i \in \Gamma} \left( \phi_h(\tau_0|\hat{Z}_0\Lambda_i, \beta_{\Lambda_i}) - \mathbb{E}_{\hat{Z}_0}[\phi_h(\tau_0|\hat{Z}_0\Lambda_i, \beta_{\Lambda_i})] \right) \right| \geq \frac{\epsilon}{3} \mid \beta = a \right) = P\left( \left| \frac{1}{|\Gamma|} \sum_{i \in \Gamma} \left( \phi_h(\tau_0|\hat{Z}_0\Lambda_i, a_{\Lambda_i}) - \mathbb{E}_{\hat{Z}_0}[\phi_h(\tau_0|\hat{Z}_0\Lambda_i, a_{\Lambda_i})] \right) \right| \geq \frac{\epsilon}{3} \right).
\]

For any fixed \( a \in \mathbb{R}^{\Gamma} \), define a function \( \phi_{2,i} : \mathbb{R} \to \mathbb{R} \) as \( \phi_{2,i}(s) := \phi_h(s, a_{\Lambda_i}) \) for each \( i \in \Gamma \) and note that it is PL(2) with PL constant upper-bounded by \( L(1 + 2\sqrt{|\Lambda|M}) \), where \( L \) is the PL
constant for \( \phi_h \) and \( M \) is such that \( |x| \leq M \) for all \( x \in E \), since by the pseudo-Lipschitz property of \( \phi_h \) and the triangle inequality,

\[
|\phi_{2,i}(x) - \phi_{2,i}(y)| = |\phi_h(x, a_{\Lambda_i}) - \phi_h(y, a_{\Lambda_i})| \leq L \left( 1 + 2||a_{\Lambda_i}|| + ||x|| + ||y||\right)||x - y||
\[
\leq L(1 + 2\sqrt{|\Lambda|}M)(1 + ||x|| + ||y||)||x - y||.
\]

Using \( L(1 + 2\sqrt{|\Lambda|}M) \) as the PL constant for \( \phi_{2,i} \) for all \( i \in \Gamma \), then

\[
f(a) = P\left(\frac{1}{|\Gamma|} \sum_{i \in \Gamma} \left| \phi_h(\tau_0 T_i([\hat{Z}_0]_{\Lambda_i \cap \Gamma}), a_{\Lambda_i}) - \mathbb{E}_{\hat{Z}_0} [\phi_h(\tau_0 T_i([\hat{Z}_0]_{\Lambda_i \cap \Gamma}), a_{\Lambda_i})] \right| \geq \frac{\epsilon}{3}\right) \leq K e^{-\kappa |\Gamma| e^2},
\]

where the last inequality follows from Lemma C.4 by noticing that \( \mathbb{E}_{\hat{Z}_0} [\phi_h(\tau_0 T_i([\hat{Z}_0]_{\Lambda_i \cap \Gamma}), a_{\Lambda_i})] = \mathbb{E}_{\hat{Z}_0} [\phi_h(\tau_0 T_i([\hat{Z}_0]_{\Lambda_i \cap \Gamma}), a_{\Lambda_i})] \). Therefore, \( T_2 = \mathbb{E}[f(\beta)] \leq K e^{-\kappa |\Gamma| e^2} \), since \( K \) and \( \kappa \) don't depend on \( \beta \) (as it doesn't show up in the pseudo-Lipschitz constant \( L(1 + 2\sqrt{|\Lambda|}M) \)).

Finally, consider \( T_3 \), the third term on the RHS of (68).

\[
T_3 \leq P\left(\frac{1}{|\Gamma|} \sum_{i \in \Gamma} L(1 + ||\tau_0 \hat{Z}_0 + \hat{\Delta}_{1,0}||, ||\tau_0 \hat{Z}_0||, ||\hat{\Delta}_{1,0}|| \geq \frac{\epsilon}{3}\right)
\[
\leq P\left(\frac{1}{\sqrt{|\Gamma|}} ||\hat{\Delta}_{1,0}|| \left(1 + \frac{2d}{|\Gamma|} ||\hat{\Delta}_{1,0}|| + 2\tau_0 \frac{2d}{|\Gamma|} ||\hat{Z}_0|| \right) \geq \frac{\epsilon}{3|\Gamma|} \right).
\]

Step (a) follows from the fact that \( \phi_h \) is PL(2). Step (b) uses \( ||\tau_0 \hat{Z}_0 + \hat{\Delta}_{1,0}|| \leq \tau_0 ||\hat{Z}_0||, \sum_{i \in \Gamma} ||a_{\Lambda_i}||^2 \leq 2d ||a||^2 \), where \( d = |\Lambda| = (2k + 1)^p \), and the following application of Lemma B.6:

\[
\sum_{i \in \Gamma} (1 + ||\hat{\Delta}_{1,0}|| + 2 ||\tau_0 \hat{Z}_0||)^2 \leq 3(|\Gamma| + 2d ||\hat{\Delta}_{1,0}||^2 + 4\tau_0^2 d ||\hat{Z}_0||^2).
\]

From (69), we have

\[
T_3 \leq P\left(\frac{1}{\sqrt{|\Gamma|}} ||\hat{Z}_0|| \geq 2\right) + P\left(\frac{1}{\sqrt{|\Gamma|}} ||\hat{\Delta}_{1,0}|| \geq \frac{\epsilon}{3|\Gamma|} \right) = P\left(\frac{1}{\sqrt{|\Gamma|}} \sum_{i \in \Gamma} (\phi_1(\hat{h}_{i,1}, \beta_{\Lambda_i}) - \mathbb{E}[\phi_1(\tau_0 [\hat{Z}_0]_{\Lambda_i}, \beta_{\Lambda_i})]) \geq \delta\epsilon\right).
\]

Finally, note that \( \phi_1 \) is PL(2) since \( S \) is Lipschitz by Lemma B.3, hence, we can apply \( H_1(b) \) to give the desired upper bound.

Next, we show concentration for \( \frac{1}{n}(h_1)^* q^0 = \frac{1}{n} \sum_{i \in \Gamma} \hat{h}_{i,1} q_{i,1}^0 \). Recall, \( q_{i,1}^0 = f_0(0, \beta_{\Lambda_i}) \) for all \( i \in \Gamma \). The function \( \phi_2 : \mathbb{R}^{2|\Lambda|} \to \mathbb{R} \) defined as \( \phi_2(x, y) := S(x) f_0(0, y) \) is PL(2) by Lemma B.3 since \( S \) and \( f_0 \) are both Lipschitz. Notice that \( \phi_2(\hat{h}_{i,1}, \beta_{\Lambda_i}) = \hat{h}_{i,1} q_{i,1}^0 \) and \( \mathbb{E}[\phi_2(\tau_0 [\hat{Z}_0]_{\Lambda_i}, \beta_{\Lambda_i})] = \)

24
\[ \mathbb{E}[\tau_0|\mathbf{Z}_0|] \mathbb{E}[f_0(0, \beta_{\lambda})] = 0 \] for all \( i \in \Gamma \) since \( |\mathbf{Z}_0| \) has zero-valued mean and is independent of \( \beta \). Therefore, using \( \mathcal{H}_1(b) \),

\[
P\left( \left| \frac{1}{n} h_1^1 q^0 \right| \geq \epsilon \right) = P\left( \frac{1}{|\Gamma|} \sum_{i \in \Gamma} \left( \phi_2(h_1^1, \beta_{\lambda}) - \mathbb{E}[\phi_2(\tau_0|\mathbf{Z}_0|, \beta_{\lambda})] \right) \right) \geq \epsilon \geq K e^{-\kappa e^2}.
\]

(d) The function \( \phi_3 : \mathbb{R}^{2|\Lambda|} \rightarrow \mathbb{R} \) defined as \( \phi_3(x, y) := (S(x))^2 \) is PL(2) by Lemma B.3 since the operator \( \mathcal{S} \) defined in (63) is Lipschitz. Notice that

\[
P\left( \left| \frac{1}{|\Gamma|} h_1^1 q^0 \right| \geq \epsilon \right) = P\left( \frac{1}{|\Gamma|} \sum_{i \in \Gamma} \left( \phi_3(h_1^1, \beta_{\lambda}) - \mathbb{E}[\phi_3(\tau_0|\mathbf{Z}_0|, \beta_{\lambda})] \right) \right) \geq \epsilon \geq K e^{-\kappa e^2}.
\]

(e) We prove concentration for \( \frac{1}{n} (q_1^0)^* q^1 \), and the result for \( \frac{1}{n} (q_1^0)^* q^1 \) follows similarly. The function \( \phi_4 : \mathbb{R}^{2|\Lambda|} \rightarrow \mathbb{R} \) defined as \( \phi_4(x, y) := f_0(0, y) f_1(x, y) \) is PL(2) by Lemma B.3, since \( f_0 \) and \( f_1 \) are Lipschitz. Notice that

\[
P\left( \frac{1}{|\Gamma|} \sum_{i \in \Gamma} \mathbb{E}[f_0(0, \beta_{\lambda}) f_1(\tau_0|\mathbf{Z}_0|, \beta_{\lambda})] = \delta \tilde{E}_{0,1} \right) \geq \epsilon \geq K e^{-\kappa e^2}.
\]

(f) The concentration of \( \lambda_0 \) to \( \tilde{\lambda}_0 \) follows from \( \mathcal{H}_1(b) \) applied to the function \( \phi_5([h_1^1], \beta_{\lambda}) := f_0'(([h_1^1], \beta_{\lambda})), \) since \( f_0' \) is assumed to be Lipschitz, hence PL(2).

The only other result to prove is concentration for \( \frac{1}{n} (h_1^1)^* q^1 = \frac{1}{|\Gamma|} \sum_{i \in \Gamma} h_1^1 q^1_i \). The function \( \phi_5 : \mathbb{R}^{2|\Lambda|} \rightarrow \mathbb{R} \) defined as \( \phi_5(x, y) = \mathcal{S}(x) f_1(x, y) \) is PL(2) by Lemma B.3. Notice that \( \phi_5(h_1^1, \beta_{\lambda}) = h_1^1 q^1_i \). Moreover, let the function \( \tilde{f}_i : \mathbb{R} \rightarrow \mathbb{R} \) be defined as \( \tilde{f}_i(x) := \mathbb{E}[\mathbf{Z}_0, \lambda \backslash \{i\}] \mathbb{E}[\tilde{f}_i(\mathcal{R}(x, [\tau_0 \mathbf{Z}_0|], \beta_{\lambda}))], \) where the function \( \mathcal{R} : \mathbb{R} \times \Lambda \rightarrow \mathbb{R} \) replaces the center coordinate of the second argument, which is in \( \mathbb{R}^\Lambda \), with the first argument, which is in \( \mathbb{R} \), a scalar. For example, \( \tilde{f}_i([\tau_0 \mathbf{Z}_0|]) = \mathbb{E}[\mathbf{Z}_0, \lambda \backslash \{i\}], \mathbb{E}[\tilde{f}_i(\mathcal{R}(x, [\tau_0 \mathbf{Z}_0|], \beta_{\lambda})), \) Then we have

\[
\sum_{i \in \Gamma} \mathbb{E}[\phi_5(\tau_0|\mathbf{Z}_0|, \beta_{\lambda})] = \sum_{i \in \Gamma} \mathbb{E}[\tau_0|\mathbf{Z}_0|, f_1(\tau_0|\mathbf{Z}_0|, \beta_{\lambda})] = \sum_{i \in \Gamma} \mathbb{E}[\tilde{f}_i(\mathcal{R}(x, [\tau_0 \mathbf{Z}_0|], \beta_{\lambda}))] = \delta |\tilde{\lambda}_1| \tilde{E}_{0,1}.
\]

In the above, step (a) follows from Stein’s Method, Lemma B.2, step (b) follows from the definition of \( \mathbf{Z}_0 \) in (31) and the definition of \( f_1' \), which is the partial derivative w.r.t. the center coordinate of the first arguments, and step (c) follows from the definition of \( \tilde{\lambda}_1 \) in (37) and the definition of \( \tilde{E}_{0,1} \) in (32). Therefore, using \( \mathcal{H}_1(b) \), we have the desired upper bound, since

\[
P\left( \frac{1}{n} (h_1^1)^* q^1 \geq \tilde{\lambda}_1 \tilde{E}_{0,1} \right) \geq \epsilon \geq K e^{-\kappa e^2}.
\]

(g) Note that \( Q_1 = \frac{1}{n} \| q^0 \|^2 \) and \( \tilde{C}_1 = \tilde{E}_{0,1} = \sigma_0^2 > 0 \). By Lemma A.5 and (23),

\[
P\left( \left| \frac{1}{n} \| q^0 \|^2 - \sigma_0^{-2} \right| \geq \epsilon \right) \leq 2 e^{-\kappa e^2} \min(1, \sigma_0^2).
\]
By the definitions in Section 3.1, \( \gamma_0^1 = \frac{1}{n} Q_1^{-1}(q^0)^* q^1 \) and \( \gamma_0^1 = (\tilde{C}^1)^{-1} \tilde{E}_1 = \tilde{E}_{0,1} \sigma_0^{-2} \). Therefore,

\[
P\left( |\gamma_0^1 &- \gamma_0^1| \geq \epsilon \right) = P\left( \left| \frac{1}{n} Q_1^{-1}(q^0)^* q^1 - \tilde{E}_{0,1} \sigma_0^{-2} \right| \geq \epsilon \right)
\]

(a) \[ \leq P\left( |Q_1^{-1} - \sigma_0^{-2}| \geq \epsilon \right) + P\left( \left| \frac{1}{n}(q^0)^* q^1 - \tilde{E}_{0,1} \right| \geq \epsilon \right) \]

(b) \[ \leq Ke^{-\kappa n \epsilon^2} + Ke^{-\kappa n \epsilon^2}. \]

where (a) follows from Lemma A.2 with \( \tilde{\epsilon} := \min \left\{ \sqrt{\frac{\epsilon}{3 \epsilon_{0,1}}}, \frac{\epsilon_0^2}{3} \right\} \) and (b) from (70) and \( H_1(e) \).

(h) From the definitions in Section 3.1, we have \( \|q_1^1\|^2 = \|q_1^1\|^2 - \|q_0^1\|^2 = \|q_0^1\|^2 - (\gamma_0^1)^2 \|q_0^1\|^2 \), and \( (\sigma_1^1)^2 = \tilde{E}_{1,1} - \tilde{E}_{1}^* (\tilde{C}^1)^{-1} \tilde{E}_1 = \sigma_1^2 - (\tilde{E}_{0,1})^2 \tilde{E}_{0,0} = \sigma_1^2 - (\gamma_0^1)^2 \sigma_0^2 \). We therefore have

\[
P\left( \left| \frac{1}{n} \|q_1^1\|^2 - (\sigma_1^1)^2 \right| \geq \epsilon \right) \leq P\left( \left| \frac{1}{n} \|q_1^1\|^2 - \sigma_1^2 \right| \geq \frac{\epsilon}{2} \right) + P\left( (\gamma_0^1)^2 \frac{1}{n} \|q_0^1\|^2 - (\gamma_0^1)^2 \sigma_0^2 \right| \geq \frac{\epsilon}{2} \right) \]

\[ \leq Ke^{-\kappa n \epsilon^2} + Ke^{-\kappa n \epsilon^2}. \]

where the last inequality is obtained using \( H_1(e) \) for bounding the first term and by applying Lemma A.2 to the second term along with the concentration of \( \|q_0^1\|^2 \) in (23), \( H_1(g) \), and Lemma A.4 (for concentration of the square).

4.2 Step 4: Showing that \( \mathcal{H}_{t+1} \) holds

The probability statements in the lemma and the other parts of \( \mathcal{H}_{t+1} \) are conditioned on the event that the matrices \( Q_1, \ldots, Q_{t+1} \) are invertible, but for the sake of brevity, we do not explicitly state the conditioning in the probabilities. The following lemma will be used to prove \( \mathcal{H}_{t+1} \).

Lemma 5. Let \( v := \frac{1}{n} B_{t+1}^* m_{t+1}^\perp - \frac{1}{n} Q_{t+1}^* (\xi_t q^t - \sum_{i=0}^{t-1} \alpha_i^t \xi_t q^i) \) and \( Q_{t+1} := \frac{1}{n} Q_{t+1}^* Q_{t+1} \). Then for \( j \in [t+1] \),

\[ P\left( \|Q_{t+1}^{-1} v\| \geq \epsilon \right) \leq e^{-\kappa n \epsilon^2}. \]

Proof. We can infer from the proof of [14, Lemma 6] that the proof of Lemma 5 involves induction hypotheses \( H_1(g), H_1(h), H_h, H_e, \) and \( B_i(c), B_i(f), B_i(g) \). Notice that the statements of these hypotheses have the same form as the corresponding statements in [14]. While the definition of the concentration constants in these hypotheses may be different for separable and non-separable denoisers, the proof uses the result that the quantities concentrate with desired rate rather than what the concentration constants are. Therefore, the proof for [14, Lemma 12], which is similar to the proof of [14, Lemma 6], is directly applicable here. \( \Box \)

We are ready to prove \( H_{t+1}(a) - H_{t+1}(h) \).

(a) Recall the definition of \( \Delta_{t+1,t} \) from Lemma 2 (43). Using Lemma B.1, \( \frac{1}{\sqrt{n}} \left\| m_{t+1}^i \right\| P_{Q_{t+1}}^i Z_t \overset{d}{=} \frac{1}{\sqrt{n}} \left\| m_{t+1}^i \right\| \frac{1}{\sqrt{|\Gamma|}} \hat{Q}_{t+1}^* \tilde{Z}_{t+1} \), where columns of the matrix \( \hat{Q}_{t+1} \in \mathbb{R}^{|\Gamma| \times (t+1)} \) form an orthogonal basis for the column space of \( Q_{t+1} \), which are normalized such that \( \hat{Q}_{t+1}^* \hat{Q}_{t+1} = |\Gamma| I_{t+1} \), and \( Z_{t+1} \in \mathbb{R}^{t+1} \) is an independent random vector with i.i.d. \( \mathcal{N}(0, 1) \) entries. We can then write

\[
\Delta_{t+1,t} \overset{d}{=} \sum_{r=0}^{t-1} (\alpha_r^t - \hat{\alpha}_r^t) h_{r+1} + Z_t \left( \left\| m_{t+1}^i \right\| \frac{1}{\sqrt{n}} - \hat{\alpha}_t^i \right) - \frac{\left\| m_{t+1}^i \right\|}{\sqrt{n}} \hat{Q}_{t+1}^* \tilde{Z}_{t+1} + Q_{t+1}^{-1},
\]
where $Q_{t+1} \in \mathbb{R}^{(t+1) \times (t+1)}$ and $v \in \mathbb{R}^{t+1}$ are defined in Lemma 5. By Lemma B.6,

$$\|\Delta_{t+1,v}\|^2 \leq (2t + 3) \left[ \sum_{r=0}^{t-1} (\alpha^*_r - \hat{\alpha}_r^*)^2 \|h^{r+1}\|^2 + \|Z_t\|^2 \left( \frac{\|m^*_t\|}{\sqrt{n}} - \tau^+_t \right)^2 \right]$$

$$+ \frac{\|m^*_t\|^2}{n} \frac{\|\dot{Q}_{t+1} Z_{t+1}\|^2}{|\Gamma|} + \sum_{j=0}^{t} \|q^j\|^2 \left[\frac{Q_{t+1}^{-1} v}{j+1}\right],$$

where we have used $Q_{t+1} Q_{t+1}^{-1} v = \sum_{j=0}^{t} q^j [Q_{t+1}^{-1} v]_{j+1}$. Applying Lemma, with $\epsilon_t = \frac{\epsilon}{(2t+3)^{\frac{3}{2}}}$, A.1,

$$P\left( \left\| \frac{\Delta_{t+1,v}}{|\Gamma|} \right\|^2 \geq \epsilon \right) \leq \sum_{r=0}^{t-1} P\left( |\alpha^*_r - \hat{\alpha}_r^*| \frac{\|h^{r+1}\|}{|\Gamma|} \geq \sqrt{\epsilon_t} \right) + P\left( \left| \frac{\|m^*_t\|}{\sqrt{n}} - \tau^+_t \right| \geq \sqrt{\epsilon_t} \right)$$

$$+ P\left( \left| \frac{\|\dot{Q}_{t+1} Z_{t+1}\|}{|\Gamma|} \right| \geq \sqrt{\epsilon_t} \right) + \sum_{j=0}^{t} P\left( \left| \frac{Q_{t+1}^{-1} v}{j+1} \right| \geq \sqrt{\epsilon_t} \right).$$

(71)

We now show each of the terms in (71) has the desired upper bound. For $0 \leq r \leq t$,

$$P\left( |\alpha^*_r - \hat{\alpha}_r^*| \frac{\|h^{r+1}\|}{|\Gamma|} \geq \sqrt{\epsilon_t} \right) \leq P\left( |\alpha^*_r - \hat{\alpha}_r^*| \left( \frac{\|h^{r+1}\|}{|\Gamma|} \right) \geq \sqrt{\epsilon_t} \right)$$

$$\leq P\left( |\alpha^*_r - \hat{\alpha}_r^*| \geq \sqrt{\epsilon_t} \right) + P\left( \left| \frac{\|h^{r+1}\|}{|\Gamma|} \right| \geq \sqrt{\epsilon_t} \right) \leq Ke^{-\kappa n} + Ke^{-\kappa n},$$

where step (a) follows from similar justification as that for step (a) in (66) and step (b) follows from induction hypotheses $B_t(g), H_1(d) - H_t(d)$, and Lemma A.3. Next, the second term in (71) is bounded as

$$P\left( \left| \frac{\|m^*_t\|}{\sqrt{n}} - \tau^+_t \right| \geq \sqrt{\epsilon_t} \right) \leq P\left( \frac{\|m^*_t\|}{\sqrt{n}} \geq \epsilon_t \right) + P\left( \frac{\|Z_t\|}{|\Gamma|} \geq \epsilon_t \right) \leq Ke^{-\kappa n} - \kappa n \epsilon + Ke^{-\kappa n} \epsilon,$$

where step (b) is obtained using induction hypothesis $B_t(h)$, Lemma A.7, and Lemma A.3. Since $\frac{1}{\sqrt{n}} \|m^*_t\|$ concentrates on $\tau^+_t$ by $B_t(h)$, the third term in (71) can be bounded as

$$P\left( \left| \frac{\|\dot{Q}_{t+1} Z_{t+1}\|}{|\Gamma|} \right| \geq \sqrt{\epsilon_t} \right) \leq P\left( \left| \frac{\|\dot{Q}_{t+1} Z_{t+1}\|}{|\Gamma|} \right| \geq \frac{\epsilon_t}{\sqrt{2}} \right) \leq Pe^{-\kappa n} \epsilon + Pe^{-\kappa n} \epsilon,$$

(72)

For the second term in (72), first bound the norm of $\dot{Q}_{t+1} Z_{t+1}$ as follows. Letting $\hat{q}_i$ denote the $i^{th}$ column of $\dot{Q}_{t+1}$, we have

$$\left\| \dot{Q}_{t+1} Z_{t+1} \right\|^2 = \left\| \sum_{i=0}^{t} \hat{q}_i \dot{Z}_{t+1,i} \right\|^2 \leq (t+1) \sum_{i=0}^{t} \left\| \hat{q}_i \right\|^2 \left( \dot{Z}_{t+1,i} \right)^2 = \left| \Gamma \right| (t+1) \sum_{i=0}^{t} \left( \dot{Z}_{t+1,i} \right)^2,$$

where step (c) follows from Lemma B.6 and step (d) uses $\left\| \hat{q}_i \right\|^2 = |\Gamma|$ for all $0 \leq i \leq t$. Therefore,

$$P\left( \left| \frac{\|\dot{Q}_{t+1} Z_{t+1}\|}{|\Gamma|} \right| \geq \epsilon \right) \leq \sum_{i=0}^{t} P\left( \dot{Z}_{t+1,i} \geq \sqrt{\frac{|\Gamma| \epsilon}{(t+1)^2}} \right) \leq 2e^{-\frac{1}{2(t+1)^2}|\Gamma| \epsilon}.$$ 

(73)
Step (e) is obtained from Lemma A.1 and step (f) from Lemma A.6. Using (73), the RHS of (72) is bounded by $Ke^{-\kappa n^2}$. Finally, for $0 \leq j \leq t$, the last term in (71) can be bounded by

$$P\left(\left|\frac{\|q_j\|}{\sqrt{n}} - \sigma_j\right| \geq \sqrt{\epsilon t}\right) \leq P\left(\left|\frac{\|q_j\|}{\sqrt{n}} - \sigma_j\right| + \sigma_j \geq \sqrt{\epsilon t}\right) \leq P\left(\left|\frac{\|q_j\|}{\sqrt{n}} - \sigma_j\right| \geq \sqrt{\epsilon t}\right) + P\left(\left|\frac{\|q_j\|}{\sqrt{n}} - \sigma_j\right| \leq \sqrt{\epsilon t} - \frac{\epsilon}{2}\right) \leq 2e^{-n\epsilon^2} + 2e^{-n\epsilon^2},$$

where step (g) follows from Lemma 5, the induction hypothesis $H_t(e)$, and Lemma A.3. Thus we have bounded each term of (71) as desired.

(b) For brevity, we use the notation $E_{\phi_h} := \frac{1}{\Gamma} \sum_{i \in \Gamma} E\left[\phi_h(\tau_0[\mathbf{Z}_0]_{\Lambda_i}, ..., \tau_t[\mathbf{Z}_t]_{\Lambda_i})\right]$ and

$$\hat{a}_i := ([\hat{t}_{\text{pure}}]^i, \ldots, [\hat{t}_r^{t+1}]^i, \sum_{r=0}^t d_r^i \hat{\Delta}_{r+1,i}^j, \beta_i),$$

$$\hat{c}_i := ([\hat{t}_{\text{pure}}]^i, \ldots, [\hat{t}_r^{t+1}]^i, \beta_i),$$

for $i \in \Gamma$. Hence $\hat{a}$ and $\hat{c}$ are arrays in $\mathbb{R}^\Gamma$ with entries $\hat{a}_i, \hat{c}_i \in \mathbb{R}^{(t+2)}$. We note that by $\hat{a}_{\Lambda_i}$ we mean for the $p$-dimensional cube $\Lambda_i$ to be applied to each of the $(t + 2)$ elements of $\hat{a}$ and we define $\|\hat{a}_{\Lambda_i}\|^2 := \sum_{j \in \Lambda_i} \|\hat{a}_j\|^2$. Moreover, define $\hat{\Delta}_{r+1,i}^j = \mathcal{Y}^{-1}(\Delta_{r+1,i}^j)$, hence $\hat{\Delta}_{r+1,i}^j \in \mathbb{R}^\Gamma$, for all $r = 0, \ldots, t$. Then, using the conditional distribution of $h^{t+1}$ from Lemma 3 and Lemma A.1,

$$P\left(\left|\frac{1}{\Gamma} \sum_{i \in \Gamma} \phi_h(\hat{a}_{\Lambda_i}, \ldots, \hat{t}_r^{t+1}_{\Lambda_i}, \beta_{\Lambda_i}) - E_{\phi_h}\right| \geq \epsilon\right) \leq P\left(\left|\frac{1}{\Gamma} \sum_{i \in \Gamma} \phi_h(\hat{c}_{\Lambda_i}) - E_{\phi_h}\right| \geq \frac{\epsilon}{2}\right).$$

Label the terms of (75) as $T_1$ and $T_2$. We next show that both terms are bounded by $Ke^{-\kappa n^2}$.

First consider term $T_1$. Let $d = |\Lambda|$. Notice that

$$\|\hat{a} - \hat{c}\|^2 = \sum_{i \in \Gamma} \left[\|\hat{a}_{\Lambda_i}\|^2 + \sum_{r=0}^t (d_r^i)^2 \|\hat{\Delta}_{r+1,i}^j\|^2 + \sum_{r'=0}^t \|\hat{\Delta}_{r+1,r'}^j\|^2\right] \leq \|\hat{a}_{\Lambda_i}\|^2 + \sum_{r=0}^t (d_r^i)^2 \sum_{r'=0}^t \|\hat{\Delta}_{r+1,r'}^j\|^2.$$  

In the above, (a) follows from Cauchy-Schwartz and (b) by collecting the terms in the sums. Hence,

$$\frac{\|\hat{a} - \hat{c}\|^2}{|\Gamma|} \leq \left(\frac{\|\hat{a}_{\Lambda_i}\|}{\sqrt{|\Gamma|}} \sum_{k=0}^t \sum_{r=0}^k (d_r^i)^2 \frac{\|\hat{\Delta}_{k+1,i}\|}{\sqrt{|\Gamma|}} \sum_{k'=0}^t \sum_{r'=0}^k (d_r^{i'})^2 + \ldots + \frac{\|\hat{\Delta}_{t+1,i}\|}{\sqrt{|\Gamma|}} \sum_{k'=0}^t \sum_{r'=0}^k (d_r^{i'})^2\right)^2.$$  

Denote the RHS of (76) by $\Delta_{\text{total}}^2$, then using Lemma A.1 and $H_1(a) - H_{t+1}(a)$, we have

$$P\left(\left|\frac{\|\hat{a} - \hat{c}\|}{\sqrt{|\Gamma|}} \geq \epsilon\right| \leq P(\Delta_{\text{total}} \geq \epsilon) \leq Ke^{-\kappa |\Gamma|^2}.$$
Now, using the pseudo-Lipschitz property of $\phi_h$, we have

$$T_1 \leq P\left( \frac{1}{|\Gamma|} \sum_{i \in \Gamma} L(1 + \|\hat{\alpha}_i\| + \|\hat{\epsilon}_i\|) \|\hat{\alpha} - \hat{\epsilon}\| |\hat{\alpha}_i\| \geq \frac{\epsilon}{2} \right)$$

$$\leq P\left( \frac{1}{|\Gamma|} \sum_{i \in \Gamma} (1 + \|\hat{\alpha}_i\| + \|\hat{\epsilon}_i\|)^2 \right)^{1/2} \left( \frac{1}{|\Gamma|} \sum_{i \in \Gamma} \|\hat{\alpha} - \hat{\epsilon}\| |\hat{\alpha}_i\| \|\hat{\epsilon}_i\| \right)^{1/2} \geq \frac{\epsilon}{2L} \right) \right)

$$\leq P\left( 1 + \sqrt{2d} \frac{\|\hat{\alpha}\|}{|\Gamma|} + \sqrt{2d} \frac{\|\hat{\epsilon}\|}{|\Gamma|} \bigg| \Delta_{\text{total}} \geq \frac{\epsilon}{2\sqrt{6dL}} \right)$$

$$(78)$$

Above, step (a) follows from Cauchy-Schwartz and step (b) from an application of Lemma B.6:

$$\sum_{i \in \Gamma} (1 + \|\hat{\alpha}_i\| + \|\hat{\epsilon}_i\|)^2 \leq 3 \sum_{i \in \Gamma} (1 + \|\hat{\alpha}_i\|)^2 + 3(\sqrt{|\Gamma|} + \sqrt{2d\|\hat{\alpha}\| + \sqrt{2d\|\hat{\epsilon}\|}})^2,$$

and $\sum_{i \in \Gamma} \|\hat{\alpha} - \hat{\epsilon}\|_{\text{total}} \leq 2d\|\hat{\alpha} - \hat{\epsilon}\|^2$ along with (76). Step (c) follows from $\|\hat{\alpha}\| \leq \|\hat{\alpha} - \hat{\epsilon}\| + \|\hat{\epsilon}\| \leq \sqrt{|\Gamma|} \Delta_{\text{total}} + \|\hat{\epsilon}\|$. Notice that

$$\|\hat{\epsilon}\|^2 = \sum_{r=0}^t ||\hat{Z}_{r+1}\|^2 + \|\beta\|^2 = \sum_{r=0}^t \tau_r^2 \|\tilde{Z}_r\|^2 + \|\beta\|^2,$$

where the last step follows from Lemma 3. Define $E_c := \sum_{r=0}^t \tau_r^2 + \sigma_r^2$. Then

$$P\left( \left| \frac{\|\hat{\epsilon}\|^2}{|\Gamma|} - E_c \right| \geq \epsilon \right) \leq \sum_{r=0}^t P\left( \left| \frac{\|\tilde{Z}_r\|^2}{|\Gamma|} - 1 \right| \geq \frac{\epsilon}{(t+2)^2} \right) + P\left( \left| \frac{\|\beta\|^2}{|\Gamma|} - \sigma_r^2 \right| \geq \frac{\epsilon}{t+2} \right) \leq K e^{-\gamma |\Gamma| \epsilon^2},$$

where the last step follows from Lemma A.7 and (27). Therefore, using the bound in (78),

$$T_1 \leq P\left( 1 + 2\sqrt{2d} \left( \frac{\|\hat{\epsilon}\|}{|\Gamma|} - E_c^{1/2} \right) + 2\sqrt{2d} E_c^{1/2} + 2d E_c^{1/2} \bigg| \Delta_{\text{total}} \geq \frac{\epsilon}{2\sqrt{6dL}} \right)$$

$$\leq P\left( \left| \frac{\|\hat{\epsilon}\|}{|\Gamma|} - E_c^{1/2} \right| \geq \frac{\epsilon}{2d} \right) + P\left( \Delta_{\text{total}} \geq \frac{\epsilon}{4} \right) \leq K e^{-\gamma |\Gamma| \epsilon^2},$$

where the last step follows from (77) and (79).

Next, consider term $T_2$ of (75).

$$P\left( \left| \frac{1}{|\Gamma|} \sum_{i \in \Gamma} \left[ \phi_h(\hat{\alpha}_i) - \phi_h(\tau_0[\tilde{Z}_0]|_\Lambda, ..., \tau_\Gamma[\tilde{Z}_\Gamma]|_\Lambda, \beta_\Lambda) \right] \right| \geq \frac{\epsilon}{2} \right)$$

$$\leq P\left( \left| \frac{1}{|\Gamma|} \sum_{i \in \Gamma} \left[ \phi_h(\hat{\alpha}_i) - \phi_h(\tau_0[\tilde{Z}_0]|_\Lambda, ..., \tau_\Gamma[\tilde{Z}_\Gamma]|_\Lambda, \beta_\Lambda) \right] \right| \geq \frac{\epsilon}{4} \right)$$

$$+ P\left( \left| \frac{1}{|\Gamma|} \sum_{i \in \Gamma} \left[ \phi_h(\tau_0[\tilde{Z}_0]|_\Lambda, ..., \tau_\Gamma[\tilde{Z}_\Gamma]|_\Lambda, \beta_\Lambda) - \phi_h(\tau_0[\tilde{Z}_0]|_\Lambda, ..., \tau_\Gamma[\tilde{Z}_\Gamma]|_\Lambda, \beta_\Lambda) \right] \right| \geq \frac{\epsilon}{4} \right).$$

Label the two terms on the RHS as $T_{2a}$ and $T_{2b}$. $T_{2a}$ can be bounded in a similar way as $T_2$ in (68) and $T_{2b}$ has the desired bound by Lemma C.2, since the function $\tilde{\phi}_h : \mathbb{R}^\Lambda \rightarrow \mathbb{R}$ defined as

$$\tilde{\phi}(s) := \mathbb{E}_{\tilde{Z}_1, ..., \tilde{Z}_\Gamma} \left[ \phi_h(\tau_0[\tilde{Z}_0]|_\Lambda, ..., \tau_\Gamma[\tilde{Z}_\Gamma]|_\Lambda, s) \right] = \mathbb{E}_{\tilde{Z}_1, ..., \tilde{Z}_\Gamma} \left[ \phi_h(\tau_0\tilde{T}_i([\tilde{Z}_0]|_\Lambda|_\Gamma), ..., \tau_\Gamma\tilde{T}_i([\tilde{Z}_\Gamma]|_\Lambda|_\Gamma), s) \right]$$

30
is PL(2) by Lemmas B.4 and B.5.

(c) We first show the concentration of \( \frac{1}{n}(h_i^{t+1})^*\mathcal{V}(\beta) = \frac{1}{n}\sum_{i\in\mathcal{I}} h_i^{t+1}\beta_i \). Using the PL(2) function \( \phi_1 \) defined in \( \mathcal{H}_1(c) \), we have that \( \phi_1(h_i^{t+1}, \beta_i) = h_i^{t+1}\beta_i \) and \( \mathbb{E}[\phi_1(\tau_i[Z_i]_{\Lambda_i}, \beta_i)] = \mathbb{E}[[\tau_iZ_i]\mathbb{I}[\beta_i] = 0 \) for all \( i \in \Gamma \), since \([\tau_iZ_i] \), has zero-valued mean and is independent of \( \beta_i \). Therefore, \( \mathcal{H}_{t+1}(b) \) gives the desired upper bound, since

\[
P\left( \left| \frac{(h_i^{t+1})^*\mathcal{V}(\beta)}{n} \right| \geq \epsilon \right) = P\left( \left| \frac{1}{|\mathcal{I}|} \sum_{i\in\mathcal{I}} \phi_1(h_i^{t+1}, \beta_i) - \mathbb{E}[\phi_1(\tau_i[Z_i]_{\Lambda_i}, \beta_i)] \right| \geq \delta \epsilon \right).
\]

We now show the concentration of \( \frac{1}{n}(h_i^{t+1})^*\mathcal{V}(\beta) = \frac{1}{n}\sum_{i\in\mathcal{I}} h_i^{t+1}\mathcal{V}_i^0 \). Using the PL(2) function \( \phi_2 \) defined in \( \mathcal{H}_1(c) \), we have that \( \phi_2(h_i^{t+1}, \beta_i) = h_i^{t+1}\mathcal{V}_i^0 \) and \( \mathbb{E}[\phi_2(\tau_i[Z_i]_{\Lambda_i}, \beta_i)] = \mathbb{E}[[\tau_iZ_i]\mathbb{I}[\beta_i] = [\tau_iZ_i]\mathbb{I}[f_0(0, \beta_i)] = 0 \), since \([Z_i] \), has zero-valued mean and is independent of \( \beta_i \) for all \( i \in \Gamma \). Therefore, using \( \mathcal{H}_{t+1}(b) \), we have the desired upper bound, since

\[
P\left( \left| \frac{(h_i^{t+1})^*\mathcal{V}(\beta)}{n} \right| \geq \epsilon \right) = P\left( \left| \frac{1}{|\mathcal{I}|} \sum_{i\in\mathcal{I}} \phi_2(h_i^{t+1}, \beta_i) - \mathbb{E}[\phi_2(\tau_i[Z_i]_{\Lambda_i}, \beta_i)] \right| \geq \delta \epsilon \right).
\]

(d) Let a function \( \tilde{\phi}_3 : \mathbb{R}^{3|\Lambda|} \to \mathbb{R} \) be defined as \( \tilde{\phi}_3(x, y, z) = S(x)S(y) \). Since the operator \( S \) defined in (63) is Lipschitz, \( \tilde{\phi}_3 \) is PL(2) by Lemma B.3. Note that \( \tilde{\phi}_3([h_i^{t+1}]_{\Lambda_i}, h_i^{t+1}, \beta_i) = h_i^{t+1}\tilde{h}_i^{t+1} \) and \( \mathbb{E}[\tilde{\phi}_3(\tau_r[Z_r]_{\Lambda_r}, \tau_r[Z_t]_{\Lambda_t}, \beta_i)] = \tau_r\mathbb{E}[Z_r\mathbb{I}][Z_t\mathbb{I} = \bar{E}_{r,t}, \) where the last equality follows from the definition in (31). Therefore, the result follows from \( \mathcal{H}_{t+1}(b) \), since

\[
P\left( \left| \frac{1}{|\mathcal{I}|}(h_i^{t+1})^*h_i^{t+1} - \bar{E}_{r,t} \right| \geq \epsilon \right) = P\left( \left| \frac{1}{|\mathcal{I}|} \sum_{i\in\mathcal{I}} \tilde{\phi}_3([h_i^{t+1}]_{\Lambda_i}, h_i^{t+1}, \beta_i) - \mathbb{E}[\tilde{\phi}_3(\tau_r[Z_r]_{\Lambda_r}, \tau_r[Z_t]_{\Lambda_t}, \beta_i)] \right| \geq \delta \epsilon \right).
\]

(e) We will show the concentration of \( \frac{1}{n}(q_i^0)^*q_i^{t+1} = \frac{1}{n}\sum_{i\in\mathcal{I}} q_i^0\tilde{q}_i^{t+1} \); the concentration of \( \frac{1}{n}(q_i^{t+1})^*q_i^{t+1} \) follows similarly. The function \( \tilde{\phi}_4(x, y) : \mathbb{R}^{2|\Lambda|} \to \mathbb{R} \) defined as \( \tilde{\phi}_4(x, y) := f_0(0, y) f_{t+1}(x, y) \) is PL(2) by Lemma B.3 and \( \tilde{\phi}_4([h_i^{t+1}]_{\Lambda_i}, \beta_i) = q_i^0\tilde{q}_i^{t+1} \). Moreover,

\[
\sum_{i\in\mathcal{I}} \mathbb{E}[\tilde{\phi}_4(\tau_i[Z_i]_{\Lambda_i}, \beta_i)] = \sum_{i\in\mathcal{I}} \mathbb{E}[f_0(0, \beta_i) f_{t+1}(\tau_i[Z_i]_{\Lambda_i}, \beta_i)] = \delta |\mathcal{I}|\tilde{E}_{0,t+1},
\]

by definition (32). Therefore, using \( \mathcal{H}_{t+1}(b) \), we have the desired result, since

\[
P\left( \left| \frac{(q_i^0)^*q_i^{t+1}}{n} - \tilde{E}_{0,t+1} \right| \geq \epsilon \right) = P\left( \left| \frac{1}{|\mathcal{I}|} \sum_{i\in\mathcal{I}} \tilde{\phi}_4([h_i^{t+1}]_{\Lambda_i}, \beta_i) - \mathbb{E}[\tilde{\phi}_4(\tau_i[Z_i]_{\Lambda_i}, \beta_i)] \right| \geq \delta \epsilon \right).
\]

(f) The concentration of \( \lambda_t \) around \( \hat{\lambda}_t \) follows \( \mathcal{H}_{t+1}(b) \) applied to the function \( \phi_5(h_i^{t+1}, \beta_i) := f_{t+1}(h_i^{t+1}, \beta_i) \), since \( f_{t+1} \) is assumed to be Lipschitz, hence PL(2). Next, we show concentration for \( \frac{1}{n}(h_i^{t+1})^*h_i^{t+1} = \frac{1}{n}\sum_{i\in\mathcal{I}} h_i^{t+1}\mathcal{V}_i^0 \). Let \( \tilde{\phi}_5 : \mathbb{R}^{3|\Lambda|} \to \mathbb{R} \) be defined as \( \tilde{\phi}_5(x, y, z) := S(y) f_{t+1}(x, z) \), which is PL(2) by Lemma B.3. Note, \( \tilde{\phi}_5([h_i^{t+1}]_{\Lambda_i}, h_i^{t+1}, \beta_i) = h_i^{t+1}\tilde{h}_i^{t+1} \) and

\[
\sum_{i\in\mathcal{I}} \mathbb{E}[\tilde{\phi}_5(\tau_r[Z_r]_{\Lambda_r}, \tau_r[Z_t]_{\Lambda_t}, \beta_i)] = \sum_{i\in\mathcal{I}} \mathbb{E}[[\tau_iZ_i]\mathbb{I}[f_{t+1}(\tau_i[Z_r]_{\Lambda_r}, \beta_i)] = |\mathcal{I}|\hat{\lambda}_{r+1}\tilde{E}_{r,t}, \quad \text{(80)}
\]
where the last equality follows using Stein’s Method, Lemma B.2, as in \( \mathcal{H}_1(f) \). Therefore, \( \mathcal{H}_{t+1}(b) \) gives the desired result, since

\[
P\left( \frac{(h^{t+1})^* q^{t+1}}{n} - \tilde{\lambda}_{r+1} \tilde{E}_{r,t} \geq \epsilon \right) = P\left( \sum_{i \in \Gamma} \left( \tilde{\phi}_5(h_{\Lambda_i}^{t+1}, \tilde{h}_{\Lambda_i}^{t+1}, \beta_{\Lambda_i}) - \mathbb{E}[\tilde{\phi}_5(\tau_r Z_i^{\Lambda_i}, \tau_t Z_i^{\Lambda_i}, \beta_{\Lambda_i})] \right) \geq \delta \epsilon \right).
\]

(g) We can represent \( Q_{t+1} \) as follows.

\[
Q_{t+1} = \frac{1}{n} \left( \begin{array}{cc}
Q_t & Q_t^* q^t \\
(Q_t^* q^t)^* & \| q^t \|^2
\end{array} \right),
\]

Then, using \( \frac{1}{n} Q_t^{-1} Q_t^* q^t = \gamma^t \) and \( (Q_t^* q^t)^* \gamma^t = (q^t)^* q^t \), it follows by the block inversion formula that

\[
Q_{t+1}^{-1} = \left( \begin{array}{cc}
Q_t^{-1} + n \| q^t \|^2 \gamma^t (\gamma^t)^* & -n \| q^t \|^2 \gamma^t \\
-n \| q^t \|^2 \gamma^t (\gamma^t)^* & n \| q^t \|^2 \gamma^t
\end{array} \right).
\]  

(81)

Using definitions (35) and (36), block inversion can be similarly used to invert \( \tilde{C}^{t+1} \):

\[
(\tilde{C}^{t+1})^{-1} = \left( \begin{array}{cc}
\tilde{C} + \tilde{E}_t & \tilde{E}_t \\
\tilde{E}_t & \sigma_t^2
\end{array} \right)^{-1} = \left( \begin{array}{cc}
(\tilde{C}^t)^{-1} + (\sigma_t^{-1})^{-2} \gamma^t (\gamma^t)^* & -(\sigma_t^{-1})^{-2} \gamma^t \\
-(\sigma_t^{-1})^{-2} \gamma^t (\gamma^t)^* & (\sigma_t^{-1})^{-2}
\end{array} \right).
\]  

(82)

In what follows, we show concentration for each of the elements in (81) to the corresponding elements in (82).

First, \( n \| q^t_i \|^2 \gamma^t (\gamma^t)^* \) concentrates to \( (\sigma_t^{-1})^{-2} \) at rate \( Ke^{-\kappa n^2} \) by \( \mathcal{H}_t(h) \) and Lemma A.5. Next, consider the \( i^{th} \) element of \( -n \| q^t_i \|^2 \gamma^t \). For \( i \in [t] \), using Lemma A.2 and \( \mathcal{H}_t(g) \), as discussed in the previous paragraph,

\[
P\left( \left| n \| q^t_i \|^2 \gamma^t_{i-1} - (\sigma_t^{-1})^{-2} \gamma^t_{i-1} \right| \geq \epsilon \right) \leq Ke^{-\kappa n^2}. \]  

(83)

Consider element \( (i,j) \) of \( Q_t^{-1} + n \| q^t_i \|^2 \gamma^t (\gamma^t)^* \) for \( i, j \in [t] \).

\[
P\left( \left| \left[ Q_t^{-1} \right]_{i,j} + n \| q^t_i \|^2 \gamma^t_{i-1} \gamma^t_{j-1} - [(\tilde{C}^t)^{-1}]_{i,j} - (\sigma_t^{-1})^{-2} \gamma^t_{i-1} \gamma^t_{j-1} \right| \geq \epsilon \right) \leq P\left( \left| \left[ Q_t^{-1} \right]_{i,j} - [(\tilde{C}^t)^{-1}]_{i,j} \right| \geq \frac{\epsilon}{2} \right) + P\left( \left| \gamma^t_{j-1} - \tilde{\gamma}^t_{j-1} \right| \geq \frac{\epsilon'}{2} \right) + P\left( \left| \gamma^t_{j-1} - \tilde{\gamma}^t_{j-1} \right| \geq \frac{\epsilon'}{2} \right)
\]

\[
\leq Ke^{-\kappa n^2} + Ke^{-\kappa n^2} + Ke^{-\kappa n^2}.
\]

Step (a) follows from Lemma A.1 and Lemma A.2 with \( \epsilon' = \min\left\{ \sqrt{\frac{2}{3}}, \frac{\epsilon}{3\gamma^t_{i-1}}, \frac{\epsilon}{3\gamma^t_{j-1}} \right\} \). Step (b) follows from the inductive hypothesis \( \mathcal{H}_t(g) \), together with (83).

We now prove \( \gamma^{t+1} = \tilde{\gamma}^{t+1} \). Recall, \( \gamma^{t+1} = \frac{1}{n} Q_{t+1}^{-1} Q_{t+1}^* q^{t+1} \), where \( Q_{t+1} := \frac{1}{n} Q_{t+1}^* Q_{t+1} \). Thus,
\[ \gamma_{r}^{t+1} = \frac{1}{n} \sum_{i=0}^{t} \left[ Q_{t+1}^{-1} \right]_{r+1,i+1} (q^{i})^{*} q^{t+1}, \] for 0 \leq r \leq t. \] Then by the definition of \( \hat{\gamma}^{t+1} \), for 0 \leq r \leq t,

\[
P \left( | \hat{\gamma}_{r}^{t+1} - \hat{\gamma}_{r}^{t+1} | > \epsilon \right) = P \left( \left| \sum_{i=0}^{t} \left( \frac{1}{n} \left[ Q_{t+1}^{-1} \right]_{r+1,i+1} (q^{i})^{*} q^{t+1} - \left( \hat{C}^{t+1} \right)^{-1} \right]_{r+1,i+1} \bar{E}_{i,t+1} \right| \geq \frac{\epsilon}{t+1} \right)
\leq \sum_{i=0}^{t} P \left( \left| \frac{1}{n} \left[ Q_{t+1}^{-1} \right]_{r+1,i+1} (q^{i})^{*} q^{t+1} - \left( \hat{C}^{t+1} \right)^{-1} \right]_{r+1,i+1} \bar{E}_{i,t+1} \right| \geq \frac{\epsilon}{t+1} \right)
\leq \sum_{i=0}^{t} P \left( \left| \frac{1}{n} (q^{i})^{*} q^{t+1} - \bar{E}_{i,t+1} \right| \geq \hat{\epsilon}_{i} \right) + P \left( \left| \left[ Q_{t+1}^{-1} \right]_{r+1,i+1} - \left( \hat{C}^{t+1} \right)^{-1} \right]_{r+1,i+1} \right| \geq \hat{\epsilon}_{i} \right)
\leq K e^{-\kappa n \epsilon^{2}} + K e^{-\kappa n \epsilon^{2}}.

Step (a) follows from Lemma A.2, with \( \hat{\epsilon}_{i} := \min \left\{ \sqrt{\frac{t}{3(t+1)}}, \sqrt{\frac{t}{3(t+1)}} E_{i,t+1}, \frac{t}{3(t+1)}(\hat{C}^{t+1})^{-1} \right\} \), and

\( \gamma^{t+1} = \frac{1}{n} \sum_{i=0}^{t} \left[ Q_{t+1}^{-1} \right]_{r+1,i+1} \left( Q_{t+1}^{*} \right) q^{t+1}, \] for 0 \leq \epsilon \leq t+1.

\[ ||Q_{t+1}^{*}||^{2} = ||Q_{t+1}^{*}||^{2} - ||Q_{t+1}^{*}||^{2} = ||Q_{t+1}^{*}||^{2} - ||Q_{t+1}^{*}||^{2} = \sum_{i=0}^{t} \gamma^{t+1}_{i} (q^{i})^{*} q^{t+1}. \]

Hence

\[
P \left( \left| \frac{1}{n} \left[ Q_{t+1}^{*} \right] \right| - \frac{1}{n} \bar{E}_{t+1} \right| \geq \frac{1}{2} \epsilon \right) = P \left( \left| \sum_{i=0}^{t} \left( \frac{1}{n} \gamma^{t+1}_{i} (q^{i})^{*} q^{t+1} - \gamma^{t+1}_{i} \bar{E}_{i,t+1} \right) \right| \geq \frac{1}{2} \epsilon \right)
\leq \sum_{i=0}^{t} P \left( \left| \frac{1}{n} \gamma^{t+1}_{i} (q^{i})^{*} q^{t+1} - \gamma^{t+1}_{i} \bar{E}_{i,t+1} \right| \geq \frac{1}{2} \epsilon \right)
\leq \sum_{i=0}^{t} P \left( \left| \gamma^{t+1}_{i} - \gamma^{t+1}_{i} \right| \geq \hat{\epsilon}_{i} \right) + \sum_{i=0}^{t} P \left( \left| \frac{1}{n} (q^{i})^{*} q^{t+1} - \bar{E}_{i,t+1} \right| \geq \hat{\epsilon}_{i} \right)
\leq K e^{-\kappa n \epsilon^{2}} + K e^{-\kappa n \epsilon^{2}}.

Step (a) follows from the concentration of products, Lemma A.2, using \( \hat{\epsilon}_{i} := \min \left\{ \sqrt{\frac{t}{6(t+1)}}, \sqrt{\frac{t}{6(t+1)}} E_{i,t+1}, \frac{t}{6(t+1)} \gamma^{t+1}_{i} \right\} \), and

\( \gamma^{t+1} = \frac{1}{n} \sum_{i=0}^{t} \left[ Q_{t+1}^{-1} \right]_{r+1,i+1} \left( Q_{t+1}^{*} \right) q^{t+1}, \) for 0 \leq \epsilon \leq t+1.

A Concentration Lemmas

In the following, \( \epsilon > 0 \) is assumed to be a generic constant, with additional conditions specified whenever needed. The proof of the Lemmas in this section can be found in [25].
Lemma A.1. (Concentration of Sums.) If random variables $X_1, \ldots, X_M$ satisfy $P(|X_i| \geq \epsilon) \leq e^{-n\kappa_i\epsilon^2}$ for $1 \leq i \leq M$, then

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

Lemma A.2 (Concentration of Products). For random variables $X, Y$ and non-zero constants $c_X, c_Y$, if

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

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

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

Lemma A.3. (Concentration of Square Roots.) Let $c \neq 0$.

If $P\left(|X_n^2 - c^2| \geq \epsilon \right) \leq e^{-\kappa n\epsilon^2}$, then $P\left(\left|\frac{1}{n} \sum_{i=1}^{n} X_i \right| \geq \epsilon \right) \leq e^{-\kappa n|c|^2\epsilon^2}$.  

Lemma A.4 (Concentration of Powers). Assume $c \neq 0$ and $\epsilon \in (0, 1]$. Then for any integer $k \geq 2$,

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

Lemma A.5 (Concentration of Scalar Inverses). Assume $c \neq 0$ and $\epsilon \in (0, 1)$.

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

Lemma A.6. For a standard Gaussian random variable $Z$ and $\epsilon > 0$, $P(|Z| \geq \epsilon) \leq 2e^{-\frac{\epsilon^2}{2}}$.  

Lemma A.7. ($\chi^2$-concentration.) For $Z_i, i \in [n]$ that are i.i.d. $\sim \mathcal{N}(0, 1)$, and $0 \leq \epsilon \leq 1$,

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

Lemma A.8. [22] Let $X$ be a centered sub-Gaussian random variable with variance factor $\nu$, i.e.,

$$\ln \mathbb{E}[e^{tX}] \leq \frac{t^2}{2}, \quad \forall t \in \mathbb{R}. \quad \text{Then X satisfies:}$$

1. For all $x > 0$, $P(X > x) \vee P(X < -x) \leq e^{-\frac{x^2}{2\nu}}$, for all $x > 0$.

2. For every integer $k \geq 1$, $\mathbb{E}[X^{2k}] \leq 2(k!)(2\nu)^k \leq (k!)(4\nu)^k$.  

33
B Other Useful Lemmas

In this section, when the results are standard, they are presented without proof.

Lemma B.1. \cite{25, Fact 1} Let \( u \in \mathbb{R}^N \) be a deterministic vector and let \( \tilde{A} \in \mathbb{R}^{n \times N} \) be a matrix with independent \( \mathcal{N}(0, 1/n) \) entries. Moreover, let \( \mathcal{W} \) be a \( d \)-dimensional subspace of \( \mathbb{R}^n \) for \( d \leq n \). Let \( (w_1, \ldots, w_d) \) be an orthogonal basis of \( \mathcal{W} \) with \( \|w_\ell\| = n \) for \( \ell \in [d] \), and let \( P_\mathcal{W} \) denote the orthogonal projection operator onto \( \mathcal{W} \). Then for \( D = [w_1 | \ldots | w_d] \), we have \( P_\mathcal{W} \tilde{A}u = \frac{\|u\|}{\sqrt{n}} P_\mathcal{W}Z_u = \frac{\|u\|}{\sqrt{n}} Dx \) where \( x \in \mathbb{R}^d \) is a random vector with i.i.d. \( \mathcal{N}(0, 1/n) \) entries.

Lemma B.2. (Stein’s lemma.) For zero-mean jointly Gaussian random variables \( Z_1, Z_2, \ldots \), and any function \( f : \mathbb{R} \to \mathbb{R} \) for which \( \mathbb{E}[Z_1f(Z_2)] \) and \( \mathbb{E}[f'(Z_2)] \) both exist, we have \( \mathbb{E}[Z_1f(Z_2)] = \mathbb{E}[Z_1Z_2]E[f'(Z_2)] \).

Lemma B.3. (Products of Lipschitz Functions are PL(2).) Let \( f, g : \mathbb{R}^p \to \mathbb{R} \) be Lipschitz continuous. Then the product function \( h : \mathbb{R}^p \to \mathbb{R} \) defined as \( h(x) := f(x)g(x) \) is PL(2).

Lemma B.4. Let \( \Lambda \) be defined in (3). For each \( r = 1, \ldots, t \), let \( \tau_r > 0 \) be a constant and let \( Z_r \in \mathbb{R}^\Lambda \) have i.i.d. standard normal entries. Suppose \( f : \mathbb{R}^{|\Lambda|(t+1)} \to \mathbb{R} \) is PL(2) with PL constant \( L \), then the function \( \tilde{f} : \mathbb{R}^\Lambda \to \mathbb{R} \) defined as \( \tilde{f}(s) := \mathbb{E}[Z_1,\ldots,Z_t [f(\tau_1Z_1, \ldots, \tau_tZ_t, s)] \) is PL(2).

Proof. Take arbitrary \( x, y \in \mathbb{R}^\Lambda \),

\[
\|\tilde{f}(x) - \tilde{f}(y)\| = \mathbb{E}\left[ |f(\tau_1Z_1, \ldots, \tau_tZ_t, x) - f(\tau_1Z_1, \ldots, \tau_tZ_t, y)| \right] \\
\leq \mathbb{E}\left[ \left| f(\tau_1Z_1, \ldots, \tau_tZ_t, x) - f(\tau_1Z_1, \ldots, \tau_tZ_t, y) \right| \right] \leq \mathbb{E}[L(1 + 2 \sum_{r=1}^t \tau_r \|Z_r\| + \|x\| + \|y\|)|x - y|] \\
\leq L(1 + 2|\Lambda| \sqrt{\frac{2}{\pi} \sum_{r=1}^t \tau_r + \|x\| + \|y\|})\|x - y\| \leq L\left( 1 + 2|\Lambda| \sqrt{\frac{2}{\pi} \sum_{r=1}^t \tau_r} \right)(1 + \|x\| + \|y\|)\|x - y\|.
\]

In the above, step (a) follows from Jensen’s inequality, step (b) holds since \( f \) is PL(2) and using the triangle inequality, and step (c) follows from \( \mathbb{E}[\|Z_r\| \leq \sum_{i \in \Lambda} \mathbb{E}[|Z_i|] = |\Lambda|\sqrt{\frac{2}{\pi}}. \)

Lemma B.5. Let \( \Gamma \) and \( \Lambda \) be as defined in (2) and (3), and let \( \Lambda_i \) be \( \Lambda \) translated to be centered at \( i \) for each \( i \in \Gamma \). Let \( f : \mathbb{R}^\Lambda \to \mathbb{R} \) be a PL(2) function with constant \( L \) and define \( \tilde{f}_i : \mathbb{R}^{\Lambda_i \cap \Gamma} \to \mathbb{R} \) as \( \tilde{f}_i(v) := f(T_i(v)) \), where \( T_i : \mathbb{R}^{\Lambda_i \cap \Gamma} \to \mathbb{R}^\Lambda \) is defined in (8). Then \( f_i \) is PL(2) for all \( i \in \Gamma \).

Proof. Let \( i \) be an arbitrary but fixed index in \( \Gamma \). Let \( d = |\Lambda_i|, a_i := |\Lambda_i \cap \Gamma|, \) and \( b_i = d - a_i \), so that \( b_i \) counts the number of \"missing\" entries in \( \Lambda_i \). For any \( x, y \in \mathbb{R}^{\Lambda_i \cap \Gamma} \), we have that

\[
|\tilde{f}(x) - \tilde{f}(y)| = |f(T_i(x)) - f(T_i(y))| \leq L(1 + \|T_i(x)\| + \|T_i(y)\|)\|T_i(x) - T_i(y)\| \\
\leq L \left( 1 + \sqrt{b_i \left( \sum_{j \in \Lambda_i \cap \Gamma} x_j^2 \right)} + \|x\| + \sqrt{b_i \left( \sum_{j \in \Lambda_i \cap \Gamma} y_j^2 \right)} + \|y\| \right) \\
\times \sqrt{b_i \left( \sum_{j \in \Lambda_i \cap \Gamma} x_j - y_j \right)^2 + \|x - y\|^2} \\
\leq L \left( 1 + \sqrt{\frac{b_i}{a_i}} \|x\|^2 + \|x\|^2 + \sqrt{\frac{b_i}{a_i}} \|y\|^2 + \|y\|^2 \right) \sqrt{\frac{b_i}{a_i} \|x - y\|^2 + \|x - y\|^2} \\
= \frac{dL}{a_i} \left( \frac{a_i}{d} + \|x\| + \|y\| \right) \|x - y\| \leq \frac{dL}{a_i} (1 + \|x\| + \|y\|)\|x - y\|,
\]

\( \square \)
where step (a) follows from the pseudo-Lipschitz property of \( f \), step (b) from our definition of \( \mathcal{T}_i \) in (8), and step (c) from Lemma B.6.

\[ \square \]

**Lemma B.6.** For any scalars \( a_1, \ldots, a_t \) and positive integer \( m \), we have \( (|a_1| + \ldots + |a_t|)^m \leq t^{m-1} \sum_{i=1}^t |a_i|^m \). Consequently, for any vectors \( u_1, \ldots, u_t \in \mathbb{R}^N \), \( \| \sum_{k=1}^t u_k \|^2 \leq t \sum_{k=1}^t \| u_k \|^2 \).

**C Concentration with Dependencies**

We first state a concentration result, existing in the literature, for functions acting on random fields that satisfy the Dobrushin uniqueness condition in Lemma C.1. Then we use Lemma C.1 to obtain Lemma C.2, which is needed to prove \( \mathcal{H}_t(b) \).

**Lemma C.1.** [26, Theorem 1] Suppose that the random field \( X = (X_i)_{i \in \Gamma} \) taking values in \( E^\Gamma \) is distributed according to a Gibbs measure \( \mu \) that obeys the Dobrushin uniqueness condition with Dobrushin constant \( c \), and the transposed Dobrushin uniqueness condition with constant \( c^* \). Suppose that \( F \) is a real function on \( E^\Gamma \) with \( \mathbb{E}[\exp(tF(X))] < \infty \) for all real \( t \). Then we have

\[
P\left( F(X) - \mathbb{E}[F(X)] > r \right) \leq \exp \left( -\frac{r^2 (1-c)(1-c^*)}{\| \delta(F) \|^2_{\ell^2}} \right), \quad \forall r \geq 0. \tag{85} \]

Here \( \delta(F) := (\delta_i(F))_{i \in \Gamma} \) is the variation vector of \( F \), where \( \delta_i(F) := \sup_{\xi, \xi' \in \mathbb{R}^\Gamma} |F(\xi) - F(\xi')| \) denotes the variation of \( F \) at the site \( i \). Its \( \ell^2 \)-norm is defined as \( \| \delta(F) \|^2_{\ell^2} := \sum_{i \in \Gamma} (\delta_i(F))^2 \). If this norm is infinite, then the statement is empty (and thus correct).

**Lemma C.2.** Let \( \Gamma \) and \( \Lambda \) be defined in (2) and (3), respectively, and let \( X = (X_i)_{i \in \Gamma} \) be a stationary Markov random field with a unique Gibbs distribution measure \( \mu \) on \( E^\Gamma \subset \mathbb{R}^\Gamma \). Assume that \( \mu \) satisfies the Dobrushin uniqueness condition and the transposed Dobrushin uniqueness condition with constants \( c \) and \( c^* \), respectively. Suppose that the state space \( E \) is bounded, meaning that there exists an \( M \) such that \( |x| \leq M \), for all \( x \in E \). Let \( f_i : \mathbb{R}^{\Lambda_i \cap \Gamma} \rightarrow \mathbb{R} \), where \( \Lambda_i \) is \( \Lambda \) being translated to be centered at location \( i \in \Gamma \), be a PL(2) function with pseudo-Lipschitz constant \( L_i \), for all \( i \in \Gamma \). Then for all \( \epsilon \in (0,1) \) there exist \( K, \kappa > 0 \), such that

\[
P\left( \left| \frac{1}{|\Gamma|} \sum_{i \in \Gamma} \left( f_i(X_{\Lambda_i \cap \Gamma}) - \mathbb{E}[f_i(X_{\Lambda_i \cap \Gamma})] \right) \right| \geq \epsilon \right) \leq K e^{-\kappa |\Gamma| \epsilon^2}. \tag{86} \]

**Proof.** Let the function \( F \) in Lemma C.1 be defined as \( F(X) := \sum_{i \in \Gamma} f_i(X_{\Lambda_i \cap \Gamma}) \). In order to apply Lemma C.1, we need to calculate \( \| \delta(F) \|^2_{\ell^2} \). Let \( d := |\Lambda| \) and \( L := \max_{i \in \Gamma} L_i \), then we have that

\[
\delta_i(F) = \sup_{\xi, \xi' \in \mathbb{R}^\Gamma} \left| F(\xi) - F(\xi') \right| = \sup_{\xi, \xi' \in \mathbb{R}^\Gamma} \left| \sum_{j \in \Lambda_i \cap \Gamma} f_j(\xi_{\Lambda_j \cap \Gamma}) - f_j(\xi'_{\Lambda_j \cap \Gamma}) \right| \leq \sup_{\xi, \xi' \in \mathbb{R}^\Gamma} \left\| J \sum_{j \in \Lambda_i \cap \Gamma} \left( 1 + \| \xi_{\Lambda_j \cap \Gamma} \| + \| \xi'_{\Lambda_j \cap \Gamma} \| \right) \| \xi_{\Lambda_j \cap \Gamma} - \xi'_{\Lambda_j \cap \Gamma} \| \right\| \leq dL(1 + 2\sqrt{dM})2\sqrt{dM}. \tag{86} \]

In the above, step (a) uses the triangle inequality and the pseudo-Lipschitz property of \( f \). Step (b) follows from the fact that \( |x| \leq M \), for all \( x \in E \) and that \( L_j \leq L \) for all \( j \in \Gamma \). Therefore,

\[
\| \delta(F) \|^2_{\ell^2} \leq |\Gamma|(dL(1 + 2\sqrt{dM})2\sqrt{dM})^2.
\]
Now applying Lemma C.1, we have
\[
P\left(\left| \frac{1}{|\Gamma|} \sum_{i \in \Gamma} \left( f_i(X_{\Lambda_i \cap \Gamma}) - \mathbb{E}[f_i(X_{\Lambda_i \cap \Gamma})] \right) \right| \geq \epsilon \right) \leq 2 \exp\left( - \frac{|\Gamma| \epsilon^2 (1 - c)(1 - c^*)}{(dL(1 + 2\sqrt{dM})2\sqrt{dM})^2} \right). \tag{87}
\]

Lemma C.3 provides a technical result about pseudo-Lipschitz functions with sub-Gaussian inputs, which will be used to prove Lemma C.4.

**Lemma C.3.** [1, Lemma D.2] Let \( X \in \mathbb{R}^d \) be a random vector whose entries have a sub-Gaussian marginal distribution with variance factor \( \nu \) as in Lemma A.8. Let \( \tilde{X} \) be an independent copy of \( X \). If \( f : \mathbb{R}^d \to \mathbb{R} \) is a PL(2) function with pseudo-Lipschitz constant \( L \), then the expectation \( \mathbb{E}[\exp(rf(X))] \) satisfies the following for \( 0 < r < [5L(2d\nu + 24d^2\nu^2)^{1/2}]^{-1} \),
\[
\mathbb{E}[e^{rf(X)}] \leq \mathbb{E} [e^{r(f(X)-f(\tilde{X}))}] \leq \left[ 1 - 25r^2L^2(d\nu + 12d^2\nu^2) \right]^{-1} \leq e^{50r^2L^2(d\nu + 12d^2\nu^2)}. \tag{88}
\]

Lemma C.4 provides a concentration inequality for sums of pseudo-Lipschitz functions acting on overlapping subsets of jointly Gaussian random variables.

**Lemma C.4.** Let \( \Gamma \) and \( \Lambda \) be defined as in (2) and (3). For each \( r = 1, \ldots, t \), let \( (Z^i_\Gamma)_{i \in \Gamma} \) have i.i.d. \( \mathcal{N}(0,1) \) entries, and for all \( r, s = 1, \ldots, t \) and \( i \neq j \), \( Z^r_\Gamma \) is independent of \( Z^s_\Gamma \). Moreover, for each \( i \in \Gamma \), let \( (Z^i_1, \ldots, Z^i_t) \) be jointly Gaussian with covariance matrix \( K \in \mathbb{R}^{t \times t} \). For each \( i \in \Gamma \), define \( Y_i := (Z^i_{\Lambda_i \cap \Gamma}, \ldots, Z^i_{\Lambda_i \cap \Gamma}) \), where \( \Lambda_i \) is \( \Lambda \) translated to be centered at location \( i \in \Gamma \). Let \( f_i : \mathbb{R}^{|\Lambda_i \cap \Gamma|} \to \mathbb{R} \) be a PL(2) function for all \( i \in \Gamma \). Then for all \( \epsilon \in (0,1) \), there exist \( K, \kappa > 0 \) such that
\[
P\left(\left| \frac{1}{|\Gamma|} \sum_{i \in \Gamma} \left( f_i(Y_i) - \mathbb{E}[f_i(Y_i)] \right) \right| \geq \epsilon \right) \leq K e^{-\kappa |\Gamma|\epsilon^2}. \tag{89}
\]

**Proof.** In the following, we prove the case for \( p = 2 \) and the proof for other dimensions follows similarly. Without loss of generality, let \( i = (i_1, i_2) \) and \( \Gamma := \{(i_1, i_2)\}_{1 \leq i_1, i_2 \leq n} \), hence \( |\Gamma| = n^2 \). Further, assume without loss of generality that \( \mathbb{E}[f_i(Y_i)] = 0 \), for all \( i \in \Gamma \). In what follows, we demonstrate the upper-tail bound:
\[
P\left(\left| \frac{1}{|\Gamma|} \sum_{i \in \Gamma} f_i(Y_i) \geq \epsilon \right| \right) \leq K e^{-\kappa |\Gamma|\epsilon^2}, \tag{90}
\]
and the lower-tail bound follows similarly. Together they provide the desired result.

Using the Crámer-Chernoff method:
\[
P\left(\left| \frac{1}{|\Gamma|} \sum_{i \in \Gamma} f_i(Y_i) \geq \epsilon \right| \right) = P\left( e^{r\sum_{i \in \Gamma} f_i(Y_i)} \geq e^{r|\Gamma|\epsilon} \right) \leq e^{-r|\Gamma|\epsilon} \mathbb{E}[e^{r\sum_{i \in \Gamma} f_i(Y_i)}], \quad \forall r > 0. \tag{91}
\]

Let \( d^2 = |\Lambda| \) and \( L_i \) be the pseudo-Lipschitz parameters associated with functions \( f_i \) for \( i \in \Gamma \) and define \( L := \max_{i \in \Gamma} L_i \). In the following, we will show that
\[
\mathbb{E}[e^{r\sum_{i \in \Gamma} f_i(Y_i)}] \leq \exp(\kappa' |\Gamma| r^2), \quad \text{for } 0 < r < (10Ld^2\sqrt{2d^2 + 24d^4})^{-1}, \tag{92}
\]
where \( \kappa' \) is any constant that satisfies \( \kappa' \geq 450L^2d^2(d^2 + 12d^4) \). Then plugging (92) into (91), we can obtain the desired result in (90):
\[
P\left(\left| \frac{1}{|\Gamma|} \sum_{i \in \Gamma} f_i(Y_i) \geq \epsilon \right| \right) \leq \exp(-|\Gamma|(r\epsilon - \kappa' r^2)).
\]

36
Set \( r = \epsilon/(2\kappa') \), which is the choice that maximizes the term in the exponent in the above, i.e. it maximizes \( (re - \kappa' r^2) \) over \( r \). We can ensure that \( \forall \epsilon \in (0,1) \), \( r \) falls within the region required in (92) by choosing \( \kappa' \) large enough.

We now show (92). Define index sets

\[
I_{j_1,j_2} := \left\{ (j_1 + k_1 d, j_2 + k_2 d) \mid k_1 = 0, \ldots, \left\lfloor \frac{n - j_1}{d} \right\rfloor, k_2 = 0, \ldots, \left\lfloor \frac{n - j_2}{d} \right\rfloor \right\}
\]

(93)

for \( j_1, j_2 = 1, \ldots, d \), and let \( C_{j_1,j_2} \) denote the cardinality of \( I_{j_1,j_2} \). We notice that for any fixed \( (j_1, j_2) \), the \( Y_{i_1,i_2} \)'s are i.i.d. for all \( (i_1, i_2) \in I_{j_1,j_2} \). Also, we have \( \Gamma = \bigcup_{j_1,j_2=1}^d I_{j_1,j_2} \) and \( I_{j_1,j_2} \cap I_{s_1,s_2} = \emptyset \), for \( (j_1, j_2) \neq (s_1, s_2) \), making the collection \( I_{1,1}, I_{1,2}, \ldots, I_{d,d} \) a partition of \( \Gamma \). Therefore,

\[
\sum_{i \in \Gamma} f_i(Y_i) = \sum_{j_1,j_2=1}^d \sum_{(i_1,i_2) \in I_{j_1,j_2}} f_{i_1,i_2}(Y_{i_1,i_2}) = \sum_{j_1,j_2=1}^d p_{j_1,j_2} \frac{1}{p_{j_1,j_2}} \sum_{(i_1,i_2) \in I_{j_1,j_2}} f_{i_1,i_2}(Y_{i_1,i_2}),
\]

where \( 0 \neq p_{j_1,j_2} < 1 \) are probabilities satisfying \( \sum_{j_1,j_2=1}^d p_{j_1,j_2} = 1 \). Using the above,

\[
\mathbb{E}[\exp\left(\sum_{i \in \Gamma} r f_i(Y_i)\right)] = \mathbb{E}\left[\exp\left(\sum_{j_1,j_2=1}^d p_{j_1,j_2} \frac{r}{p_{j_1,j_2}} \sum_{(i_1,i_2) \in I_{j_1,j_2}} f_{i_1,i_2}(Y_{i_1,i_2})\right)\right]
\]

\[
\leq \sum_{j_1,j_2=1}^d p_{j_1,j_2} \mathbb{E}\left[\exp\left(\frac{r}{p_{j_1,j_2}} \sum_{(i_1,i_2) \in I_{j_1,j_2}} f_{i_1,i_2}(Y_{i_1,i_2})\right)\right]
\]

\[
= \prod_{j_1,j_2=1}^d \mathbb{E}\left[\exp\left(\frac{r}{p_{j_1,j_2}} f_{i_1,i_2}(Y_{i_1,i_2})\right)\right]
\]

\[
\leq \sum_{j_1,j_2=1}^d p_{j_1,j_2} \exp\left(\frac{50C_{j_1,j_2}^2 L^2 r^2 (td^2 + 12t^2 d^4)}{p_{j_1,j_2}^2}\right),
\]

(94)

where step (a) follows from Jensen’s inequality, step (b) from the independence of \( Y_{i_1,i_2} \)'s for \( (i_1, i_2) \in I_{j_1,j_2} \), and step (c) from Lemma C.3 with variance factor \( \nu = 1 \) and restriction

\[
0 < r < \left(5L \sqrt{2t \bar{d}^2 + 24t^2 d^4}\right)^{-1} \min_{(j_1,j_2)} p_{j_1,j_2}.
\]

(95)

Let \( p_{j_1,j_2} = \sqrt{C_{j_1,j_2}}/C \), where \( C = \sum_{j_1,j_2=1}^d \sqrt{C_{j_1,j_2}} \), ensuring \( \sum_{j_1,j_2=1}^d p_{j_1,j_2} = 1 \). Then,

\[
\sum_{j_1,j_2=1}^d p_{j_1,j_2} \exp\left(\frac{50C_{j_1,j_2}^2 L^2 r^2 (td^2 + 12t^2 d^4)}{p_{j_1,j_2}^2}\right) \leq e^{50C^2 L^2 r^2 (td^2 + 12t^2 d^4)} \leq e^{450d^2 L^2 (td^2 + 12t^2 d^4)r^2} \leq e^{\kappa' |\Gamma| r^2},
\]

whenever \( \kappa' \geq 450d^2 L^2 (td^2 + 12t^2 d^4) \). In the above, step (a) follows from:

\[
C^2 = \left(\sum_{j_1,j_2=1}^d \sqrt{C_{j_1,j_2}}\right)^2 = \sum_{j_1,j_2=1}^d C_{j_1,j_2} + \sum_{j_1,j_2=1}^d \sum_{(k_1,k_2) \neq (j_1,j_2)} \sqrt{C_{j_1,j_2} C_{k_1,k_2}} \leq n^2 + d^2 (d^2 - 1) C_{1,1} \leq d^2 n^2 + 4(d^2 - 1)(d^2 + nd) < 9d^2 n^2,
\]

(b)
where step (b) holds because $C_{1,1} = \max_{j_1, j_2} C_{j_1, j_2}$ and step (c) holds because $C_{1,1} = (\lfloor \frac{n-1}{d} \rfloor + 1)^2 \leq \frac{n}{d} + 2)$. Finally, we consider the effective region for $r$ as required in (95). Notice that

$$\min_{j_1, j_2} p_{j_1, j_2} = \frac{\sqrt{C_{d,d}}}{C} = \frac{\sqrt{C_{d,d}}}{\sum_{j_1, j_2=1}^{d} \sqrt{C_{j_1, j_2}}} \geq \frac{\sqrt{C_{d,d}}}{d^2 \sqrt{C_{1,1}}} = \frac{1}{d^2} \frac{\lfloor \frac{n}{d} \rfloor}{\lfloor \frac{n}{d} \rfloor + 1} \geq \frac{1}{d^2} \frac{\lfloor \frac{n}{d} \rfloor}{\lfloor \frac{n}{d} \rfloor} \geq \frac{1}{2d^2}.$$ 

Hence, if we require $0 < r < (10Ld^2\sqrt{2td^2} + 24t^2d^4)^{-1}$, then (95) is satisfied.

Acknowledgments

This work was supported by the National Science Foundation (NSF) under grants CCF-1217749 and ECCS-1611112.

References

[1] Y. Ma, C. Rush, and D. Baron, “Analysis of approximate message passing with a class of non-separable denoisers,” Proc. IEEE Int. Symp. Inf. Theory, June 2017. Full version: https://arxiv.org/abs/1705.03126.

[2] D. L. Donoho, A. Maleki, and A. Montanari, “Message passing algorithms for compressed sensing,” Proc. Nat. Academy Sci., vol. 106, pp. 18914–18919, Nov. 2009.

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

[4] M. Bayati and A. Montanari, “The dynamics of message passing on dense graphs, with applications to compressed sensing,” IEEE Trans. Inf. Theory, vol. 57, pp. 764–785, Feb. 2011.

[5] F. Krzakala, M. M´ezard, F. Sausset, Y. Sun, and L. Zdeborov´a, “Probabilistic reconstruction in compressed sensing: Algorithms, phase diagrams, and threshold achieving matrices,” J. Stat. Mech. – Theory E., vol. 2012, p. P08009, Aug. 2012.

[6] S. Rangan, “Generalized approximate message passing for estimation with random linear mixing,” in Proc. IEEE Int. Symp. Inf. Theory (ISIT), (St. Petersburg, Russia), pp. 2168–2172, July 2011.

[7] J. Tan, Y. Ma, and D. Baron, “Compressive imaging via approximate message passing with image denoising,” IEEE Trans. Signal Processing, vol. 63, pp. 2085–2092, April 2015.

[8] C. Metzler, A. Maleki, and R. G. Baraniuk, “From denoising to compressed sensing,” IEEE Trans. Inf. Theory, vol. 62, pp. 5117–5114, Apr. 2016.

[9] Y. Ma, J. Zhu, and D. Baron, “Approximate message passing algorithm with universal denoising and gaussian mixture learning,” IEEE Trans. Signal Processing, vol. 64, pp. 5611–5622, Nov. 2016.

[10] K. Sivaramakrishnan and T. Weissman, “Universal denoising of continuous amplitude signals with applications to images,” in Proc. Int. Conf. Image Process., pp. 2609–2612, Oct. 2006.

[11] K. Sivaramakrishnan and T. Weissman, “Universal denoising of discrete-time continuous-amplitude signals,” IEEE Trans. Inf. Theory, vol. 54, pp. 5632–5660, Dec. 2008.

[12] R. C. Dubes and A. K. Jain, “Random field models in image analysis,” J. Appl. Statist., vol. 16, no. 2, pp. 131–164, 1989.

[13] G. R. Cross and A. K. Jain, “Markov random field texture models,” IEEE Transactions on Pattern Analysis and Machine Intelligence, no. 1, pp. 25–39, 1983.
[14] C. Rush and R. Venkataramanan, “Finite sample analysis of approximate message passing algorithms,” IEEE Trans. Inf. Theory, vol. 64, no. 11, pp. 7264–7286, 2018.

[15] A. Javanmard and A. Montanari, “State evolution for general approximate message passing algorithms, with applications to spatial coupling,” Inf. and Inference, vol. 2, pp. 115 – 144, Dec. 2013.

[16] C. Rush, A. Greig, and R. Venkataramanan, “Capacity-achieving sparse regression codes via approximate message passing decoding,” Proc. IEEE Int. Symp. Inf. Theory, June 2015. Full version: http://arxiv.org/abs/1501.05892.

[17] J. Tan, Y. Ma, H. Rueda, D. Baron, and G. Arce, “Compressive hyperspectral imaging via approximate message passing,” IEEE J. Sel. Topics Signal Process., vol. 10, pp. 389–401, Mar. 2016.

[18] R. Berthier, A. Montanari, and P.-M. Nguyen, “State evolution for approximate message passing with non-separable functions,” arXiv:1708.03950, 2017.

[19] H. O. Georgii, Gibbs Measures and Phase Transitions. Berlin: de Gruyter, 1998.

[20] J. Hammersley and P. Clifford, “Markov fields on finite graphs and lattices.” Unpublished.

[21] S. Z. Li, Markov random field modeling in image analysis. Springer Science & Business Media, 2009.

[22] S. Boucheron, G. Lugosi, and P. Massart, Concentration inequalities: A nonasymptotic theory of independence. OUP Oxford, 2013.

[23] F. Champagnat, J. Idier, and Y. Goussard, “Stationary Markov random fields on a finite rectangular lattice,” IEEE Trans. Inf. Theory, vol. 44, no. 7, pp. 2901–2916, 1998.

[24] A. Beck and M. Teboulle, “A fast iterative shrinkage-thresholding algorithm for linear inverse problems,” SIAM J. Imag. Sci., vol. 2, pp. 183–202, Mar. 2009.

[25] C. Rush and R. Venkataramanan, “Finite sample analysis of approximate message passing,” Proc. IEEE Int. Symp. Inf. Theory, June 2016. Full version: https://arxiv.org/abs/1606.01800.

[26] C. K"ulske, “Concentration inequalities for functions of Gibbs fields with application to diffraction and random Gibbs measures,” Commun. Math. Phys., vol. 239, no. 1-2, pp. 29–51, 2003.