Approximate Message Passing with A Class of Non-Separable Denoisers

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

January 23, 2017

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 β_0 from noisy, linear measurements $y = A\beta_0 + w$. AMP has the attractive feature that its performance (for example, the mean squared error of its estimates) can be accurately tracked by a simple, scalar iteration referred to as state evolution when the unknown signal has independent and identically distributed (i.i.d.) entries. However, in many real-world applications, like image or audio signal reconstruction, the unknown signal contains dependencies between entries and so a coordinate-wise independence structure is not a good approximation for the prior of the unknown signal. In this paper we study the case where the unknown signal has dependent entries using a class of non-separable sliding-window denoisers and prove that a new form of state evolution still accurately predicts AMP performance in this scenario. This is an early step in understanding the role of non-separable denoisers within AMP, and will lead to a characterization of more general denoisers in problems including compressive image reconstruction.

1 Introduction

1.1 Motivation

In this work, we study the high-dimensional regression model, where one wishes to recover an unknown vector $\beta_0 \in \mathbb{R}^N$ from noisy observations as in the following model:

$$y = A\beta_0 + w, \tag{1}$$

where $y \in \mathbb{R}^n$ is the output, $A \in \mathbb{R}^{n \times N}$ is a known measurement matrix, and $w \in \mathbb{R}^n$ is zero-mean noise with finite variance σ^2. We assume that the ratio of the dimensions of the measurement is a constant value, $\delta := n/N$, where $\delta \in (0, \infty)$.

Approximate message passing (AMP) [1–5] is a class of low-complexity, iterative algorithms studied to solve the high-dimensional regression task of (1). The performance of AMP mainly depends on a sequence of functions $\{\eta_t\}_{t\geq 0}$ used to generate a sequence of estimates $\{\beta_t\}_{t\geq 0}$ from auxiliary observation vectors computed in every iteration of the algorithm. A nice property of AMP is that under some technical conditions these observation vectors can be approximated as the input signal β_0 plus independent and identically distributed (i.i.d.) Gaussian noise. This fact allows one to choose functions $\{\eta_t\}_{t\geq 0}$ based on statistical knowledge of β_0, for example, a common choice is for η_t to be the Bayes optimal estimate of β_0 conditional on the value of the observation vector. For this reason, the functions $\{\eta_t\}_{t\geq 0}$ are referred to as 'denoisers.'

Previous analysis of the performance of AMP only considers the case when the unknown signal has a prior with i.i.d. entries. In this case, there is no loss of performance by limiting choices of the denoiser functions $\{\eta_t\}_{t\geq 0}$ to the class of functions that acts coordinate-wise when applied to a vector; such functions are referred to as *separable*. However, in many real-world applications, the unknown signal β_0 contains dependencies between entries and so such a coordinate-wise independence structure is not a good approximation for the prior of β_0. For example, when the signals are images [6,7] or sound clips [8], *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 chain prior. In this paper, we extend the previous performance guarantees for AMP to a class of non-separable sliding-window denoisers introduced by Ma *et al.* [8] when the unknown signal is produced by a Markov chain in its stationary distribution.

When the measurement matrix A has i.i.d. Gaussian entries and the unknown signal β_0 is i.i.d., Bayati and Montanari [3] proved that the performance of AMP at each iteration can be accurately predicted by a simple, scalar iteration referred to as *state evolution* in the large system limit ($n, N \to \infty$ such that $\frac{n}{N} = \delta$ is a constant). For example, if β^t is the estimate produced by AMP at time t, their result implies that the ℓ^2-error, $\frac{1}{N}\left\|\beta^t - \beta_0\right\|^2$, and other performance measures converge to known values predicted by state evolution using the prior distribution of β_0. Recently, Rush and Venkataramanan [9] provided a concentration version of the asymptotic result when the prior distribution of β_0 is i.i.d. sub-Gaussian. The result implies that the probability of ϵ-deviation between various performance measures and their limiting constant values fall exponentially in n. Extensions of AMP performance guarantees beyond separable denoisers have been considered in special cases [10,11] for certain classes of block-separable denoisers that allow dependencies within blocks of the signal β_0 with independence across blocks. However these settings are more restricted than the types of dependencies we consider.

1.2 AMP Algorithm for Sliding-Window Denoiser

The AMP algorithm, in the case of a dependent signal, generates successive estimates of the unknown vector denoted by $\beta^t \in \mathbb{R}^N$ for $t = 1, 2, \ldots$. These values are calculated as follows: given the observed vector y, set $\beta^0 = 0$, the all-zeros vector. For $t = 0, 1, \ldots$, fix $k \geq 0$ an integer, and AMP computes

$$z^t = y - A\beta^t + \frac{z^{t-1}}{n} \sum_{i=k+1}^{N-k} \eta'_{t-1}([A^* z^{t-1} + \beta^{t-1}]_{i-k}^{i+k}), \tag{2}$$

$$\beta^{t+1} = \begin{cases} \eta_t([A^* z^t + \beta^t]_{i-k}^{i+k}) & \text{if } k+1 \leq i \leq N-k, \\ 0 \quad \text{otherwise,} \end{cases} \tag{3}$$

for an appropriately-chosen sequence of non-separable denoiser functions $\{\eta_t\}_{t\geq 0} : \mathbb{R}^{2k+1} \to \mathbb{R}$, where the notation

$$[x]_{i-k}^{i+k} = (x_{i-k}, \ldots, x_{i+k}) \in \mathbb{R}^{2k+1} \quad \text{for} \quad x \in \mathbb{R}^N,$$

and A^* denotes the transpose of A. We let η'_t denote the (weak) partial derivative of η_t with respect to (w.r.t.) the $(k+1)^{th}$ coordinate, or the center element. Quantities with a negative index in (2) and (3) are set to zero.

1.3 Contributions and Outline

Our main result proves concentration for order-2 pseudo-Lipschitz (PL) loss functions[1] for the AMP estimate of (3) at any iteration t of the algorithm to constant values predicted by the state evolution equations. We envision that our work in understanding the role of sliding-window denoisers within AMP is an early step in characterizing the role of non-separable denoisers within AMP. This work will lead to a characterization of more general denoisers in problems including compressive image reconstruction [6, 7].

To characterize AMP performance in this way for sliding-window denoisers when the input signal is a Markov chain, we need concentration inequalities for PL functions of Markov chains and sequences of Gaussian vectors that are constructed in a certain way. Specifically, in the constructed sequences, successive elements are successive $(2k+1)$-length overlapping blocks of some original sequences (another Markov chain or Gaussian sequence, respectively), as suggested by the structure of the denoiser η_t in (3). These concentration results are stated in Lemmas D.6 and D.7 in Section 3. Because we also need to characterize the bounded derivative of the denoiser function, concentration results for bounded functions of the sequences constructed as described above are also needed; these results are stated in Lemmas D.2 and D.8 in Section 3.

The rest of the paper is organized as follows. Section 2 provides model assumptions, state evolution for sliding-window denoisers, and the main performance guarantee (Theorem 1), which is a concentration result for PL loss functions acting on the AMP outputs from (2)-(3) to the state evolution predictions. Section 3 proves Theorem 1. The proof is based on two technical lemmas: Lemma 2 and Lemma 3. The proof of Lemma 2 is the same as [9, Lemma 4] and we prove Lemma 3 in Section 4.

2 Main Results

2.1 Definitions and Assumptions

Before proceeding, let us clarify the assumptions under which our result is proved. The assumptions on the measurement matrix A and noise w are the same as in Rush and Venkataramanan [9] and are not repeated here. We state our assumptions on the non-i.i.d. signal β_0 and the non-separable denoiser η_t below. First we include a definition on properties of Markov chains, which will be useful throughout the work.

Definition 2.1. *Consider a Markov chain taking values in a state space S with probability transition measure $r(x, dy)$ and stationary probability measure γ. The chain is said to be* ***geometrically ergodic on*** *$L^2(\gamma)$ where $L^2(\gamma) := \{\nu \ll \gamma : \int_S \left|\frac{d\nu}{d\gamma}\right|^2 d\gamma < \infty\}$ if there exists $0 < \rho < 1$ such that for each probability measure $\nu \in L^2(\gamma)$,[2] there is $C_\nu < \infty$ such that*

$$\sup_{A \in \mathcal{B}(S)} \left| \int_S r^n(x, A)\nu(dx) - \gamma(A) \right| < C_\nu \rho^n, \qquad n \in \mathbb{N},$$

where $\mathcal{B}(S)$ is the Borel sigma-algebra on S and $r^n(x, dy)$ denotes the n-step transition probability measure. The chain is said to be ***reversible*** *if $r(x, dy)\gamma(dx) = r(y, dx)\gamma(dy)$.*

[1] A function $f : \mathbb{R}^m \to \mathbb{R}$ is order-2 *pseudo-Lipschitz* if there exists a constant $L > 0$ such that for all $x, y \in \mathbb{R}^m$, $|\phi(x) - \phi(y)| \le L(1 + \|x\| + \|y\|)\,\|x - y\|$, where $\|\cdot\|$ denotes the Euclidean norm.

[2] For two measures ν and γ, $\nu \ll \gamma$ denotes that ν is absolutely continuous w.r.t. γ, and $\frac{d\nu}{d\gamma}$ denotes the Radon-Nikodym derivative.

Assumptions:

Signal: Let $S \subset \mathbb{R}$ be a bounded state space (countable or uncountable). We assume that the signal $\beta_0 \in S^N$ is produced by a time-homogeneous, reversible, geometrically ergodic Markov chain in its (unique) stationary distribution. Note that this means the 'sequence' $\beta_{0_1}, \beta_{0_2}, \ldots, \beta_{0_N}$ forms a Markov chain. We refer to the stationary distribution as γ_β and assume that γ_β has finite fourth moment. Moreover, we assume that for any $\epsilon \in (0,1)$,

$$P\left(\left|\frac{1}{N}\sum_{i=1}^{N}\beta_{0_i} - \mathbb{E}[\beta^2]\right| \geq \epsilon\right) \leq Ke^{-\kappa N\epsilon^2},$$

for some constants $K, \kappa > 0$, where $\beta \sim \gamma_\beta$ and β_{0_i} is element i of $\beta_0 \in \mathbb{R}^N$ for $i = 1, 2, \ldots, N$.

Denoiser functions: The denoiser functions $\eta_t : \mathbb{R}^{2k+1} \to \mathbb{R}$ in (3) are assumed to be Lipschitz[3] for each $t > 0$ and, therefore, are also weakly differentiable with bounded derivative. The weak partial derivative w.r.t. the $(k+1)^{th}$ (middle) coordinate is denoted by η_t'. Further, the derivative η_t' is assumed to be differentiable with bounded derivative, meaning that all $2k+1$ partial derivatives exist and are bounded. (It is possible to weaken this condition to allow η_t' to have a finite number of discontinuities, if needed, as in [9].)

2.2 Performance Guarantee

As noted in Section 1, the behavior of the AMP algorithm is predicted by a simple, scalar iteration referred to as state evolution introduced here. Let the stationary distribution γ_β and the transition probability measure $r(x, dy)$ define the prior distribution for the unknown vector β_0 in (1). Let scalar $\beta \in \mathbb{R}$ be distributed as γ_β and vector $\underline{\beta} \in \mathbb{R}^{2k+1}$ be distributed as π, where

$$\pi(dx) = \pi((dx_1, ..., dx_{2k+1})) = \prod_{i=2}^{2k+1} r(x_{i-1}, dx_i)\gamma_\beta(dx_1), \tag{4}$$

is the probability of seeing such a length-$(2k+1)$ sequence in the β_0 Markov chain. Note that you could equivalently think of β as being the k^{th} entry of $\underline{\beta}$ since $\underline{\beta}$ is a Markov chain starting in its stationary distribution. Define $\sigma_\beta^2 = \mathbb{E}[\beta^2] > 0$, and $\sigma_0^2 = \sigma_\beta^2/\delta$. Iteratively define the quantities $\{\tau_t^2\}_{t\geq 0}$ and $\{\sigma_t^2\}_{t\geq 1}$ as follows.

$$\begin{aligned} \tau_t^2 &= \sigma^2 + \sigma_t^2, \\ \sigma_t^2 &= \frac{1}{\delta}\left((1-w_k)\mathbb{E}\left\{\left(\eta_{t-1}(\underline{\beta} + \tau_{t-1}\underline{Z}) - \beta\right)^2\right\} + w_k\sigma_\beta^2\right), \end{aligned} \tag{5}$$

where $\underline{Z} \in \mathbb{R}^{2k+1}$ is a vector of i.i.d. standard normal random variables, $w_k = 2k/N$, and $\delta = n/N$.

Theorem 1 provides our main performance guarantee which is a concentration inequality for PL loss functions

Theorem 1. *With the assumptions of Section 1, for any (order-2) pseudo-Lipschitz function* $\phi : \mathbb{R}^2 \to \mathbb{R}$, $\epsilon \in (0,1)$, *and* $t \geq 0$,

$$P\left(\left|\sum_{i=k+1}^{N-k}\frac{\phi(\beta_i^{t+1}, \beta_{0_i})}{N-2k} - \mathbb{E}[\phi(\eta_t(\underline{\beta} + \tau_t\underline{Z}), \beta)]\right| \geq \epsilon\right) \leq K_t e^{-\kappa_t n\epsilon^2}. \tag{6}$$

[3] A function $f : \mathbb{R}^m \to \mathbb{R}$ is *Lipschitz* if there exists a constant $L > 0$ such that for all $x, y \in \mathbb{R}^m$, $|\phi(x) - \phi(y)| \leq L\|x - y\|$, where $\|\cdot\|$ denotes the Euclidean norm.

In the expectation in (6), $\beta \in \mathbb{R} \sim \gamma_\beta$, $\underline{\beta} \in \mathbb{R}^{2k+1} \sim \pi$, *and* $\underline{Z} \in \mathbb{R}^{2k+1}$ *is a vector of i.i.d. standard normal random variables that is independent of* β *and* $\underline{\beta}$. *Constant* τ_t *is defined in* (5). *Constants* $K_t, \kappa_t > 0$ *do not depend on* n *or* ϵ, *but their values are not exactly specified.*

Remarks:

(1) The probability in (6) is w.r.t. the product measure on the space of the matrix A, signal β_0, and noise w.

(2) Theorem 1 shows concentration for the loss when considering only the inner $N-2k$ elements of the signal. This is due to the nature of the sliding-window denoiser, which updates each element of the estimate β^t using the k elements on either side of that location. In practice, as in Ma *et al.* [8], one could run a slightly different algorithm than that given in (2)-(3): instead of setting the end elements, meaning the first k and last k elements, of the estimate β^t equal to 0, update these elements using the sliding-window denoiser but with missing input values replaced by the median of the other inputs. Such a strategy shows good empirical performance – even at the end elements – and suggests that the concentration result of Theorem 1 could be extended to show concentration for the loss of the full signal. Proving this requires a delicate handling of the end elements and is left for future research.

(3) The state evolution constants $\{\tau_t^2\}_{t\geq 0}$ defined in (5) are the sum of σ^2 and two weighted terms, where the weight depends on k, the length of the window in the sliding-window denoiser. As k increases, the state evolution values depend more on σ_β^2, the second moment of the stationary distribution γ_β. The increasing dependency on σ_β^2 as k grows is because the nature of the sliding-window denoiser makes estimating the end elements of β_0 difficult. For our theoretical results, we only estimate the middle $N-2k$ elements of the signal, and so as k increases the state evolution constants $\{\tau_t^2\}_{t\geq 0}$ depend more on σ_β^2, the second moment of the elements of the original signal, corresponding to the estimation error in the un-estimated part of the signal.

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

$$P\left(\left|\sum_{i=k+1}^{N-k} \frac{(\beta_i^{t+1} - \beta_{0_i})^2}{N-2k} - \frac{n(\tau_{t+1}^2 - \sigma^2) - 2k\sigma_\beta^2}{N-2k}\right| \geq \epsilon\right) \leq K_t e^{-\kappa_t n \epsilon^2},$$

with τ_{t+1}^2 defined in (5). Numerical verification for the mean squared error loss when the denoisers $\{\eta_t\}_{t\geq 0}$ are Bayes-optimal denoisers can be found in [8].

3 Proof of Theorem 1

The proof of Theorem 1 follows the work of Rush and Venkataramanan [9] closely, with modifications for the dependent structure of the unknown vector β_0 in (1). For this reason, we use much of the same notation. The main ingredients in the proof of Theorem 1 are two technical lemmas corresponding to [9, Lemmas 4 and 5]. We first cover some preliminary results and establish notation used in the proof. We then discuss the lemmas used to prove Theorem 1.

3.1 Proof Notation

As mentioned above, in order to streamline the proof of our technical lemmas we use notation similar to [9] and consequently to [3]. As in the previous work, the technical lemmas are proved

for a more general recursion which we define in the following, with AMP being a specific example of the general recursion as shown below.

Given noise $w \in \mathbb{R}^n$ and unknown signal $\beta_0 \in \mathbb{R}^N$, fix the half-window-size $k > 0$ an integer, define column vectors $h^{t+1}, q^{t+1} \in \mathbb{R}^N$ and $b^t, m^t \in \mathbb{R}^n$ for $t \geq 0$ recursively as follows, starting with initial condition $q^0 \in \mathbb{R}^N$:

$$\begin{aligned} h^{t+1} &:= A^* m^t - \xi_t q^t, \qquad q^t := \begin{cases} f_t([h^t]_{i-k}^{i+k}, [\beta_0]_{i-k}^{i+k}), & \text{if } i = k+1, ..., N-k, \\ -\beta_{0_i}, & \text{otherwise,} \end{cases} \\ b^t &:= Aq^t - \lambda_t m^{t-1}, \qquad m^t := g_t(b^t, w), \end{aligned} \tag{7}$$

with scalar values ξ_t and λ_t defined as

$$\xi_t := \frac{1}{n} \sum_{i=1}^{n} g_t'(b_i^t, w_i), \quad \lambda_t := \frac{1 - w_k}{\delta(N - 2k)} \sum_{i=k+1}^{N-k} f_t'([h^t]_{i-k}^{i+k}, [\beta_0]_{i-k}^{i+k}), \tag{8}$$

where $w_k = 2k/N$. For the derivatives in (8), the derivative of $g_t : \mathbb{R}^2 \to \mathbb{R}$ is with respect to the first argument and the derivative of $f_t : \mathbb{R}^{2(2k+1)} \to \mathbb{R}$ is with respect to the $(k+1)^{th}$ or center coordinate of the first argument. The functions $\{f_t\}_{t \geq 0}$, and $\{g_t\}_{t \geq 0}$ are assumed to be Lipschitz continuous[4], meaning the the weak derivatives g_t' and f_t' exist and are bounded. Further, g_t' and f_t' are each assumed to be differentiable in the first argument with bounded derivative. For f_t' this means that we assume the first $2k+1$ partial derivatives exist and are bounded.

Recall that the unknown vector $\beta_0 \in \mathbb{R}^N$ is assumed to have a Markov chain prior with transition probability measure $r(x, dy)$ and stationary probability measure γ_β. Let $\beta \in \mathbb{R} \sim \gamma_\beta$ and $\underline{\beta} \in \mathbb{R}^{2k+1} \sim \pi$ where π is defined in (4). Note that π is the $(2k+1)$-dimension marginal distribution of β_0 and γ_β is the one-dimensional marginal distribution.

Let $\underline{0} \in \mathbb{R}^{2k+1}$ be a vector of zeros. Define

$$\sigma_\beta^2 := \mathbb{E}[\beta^2], \tag{9}$$

$$\sigma_0^2 := \frac{1}{\delta}((1 - w_k)\mathbb{E}[f_0^2(\underline{0}, \underline{\beta})] + w_k \sigma_\beta^2) > 0. \tag{10}$$

Further, let

$$q^0 := \begin{cases} f_0(\underline{0}, [\beta_0]_{i-k}^{i+k}), & \text{if } i = k+1, ..., N-k \\ -\beta_{0_i}, & \text{otherwise} \end{cases} \tag{11}$$

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

$$P\left(\left|\frac{1}{n}\left\|q^0\right\|^2 - \sigma_0^2\right| \geq \epsilon\right) \leq Ke^{-\kappa n \epsilon^2}. \tag{12}$$

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

$$\tau_t^2 := \mathbb{E}[(g_t(\sigma_t Z, W))^2], \qquad \sigma_t^2 := \frac{1}{\delta}((1 - w_k)\mathbb{E}[(f_t(\tau_{t-1}\underline{Z}, \underline{\beta}))^2] + w_k \sigma_\beta^2), \tag{13}$$

where random variables $W \sim p_w$ and $Z \sim \mathcal{N}(0, 1)$ and random vectors $\underline{\beta} \in \mathbb{R}^{2k+1} \sim \pi$ and $\underline{Z} \in \mathbb{R}^{2k+1} \sim \mathcal{N}(0, \mathrm{I})$ are independent. We assume that both σ_0^2 and τ_0^2 are strictly positive.

[4]A function $f : \mathbb{R}^m \to \mathbb{R}$ is *Lipschitz* if there exists a constant $L > 0$ such that for all $x, y \in \mathbb{R}^m$, $|\phi(x) - \phi(y)| \leq L \|x - y\|$, where $\|\cdot\|$ denotes the Euclidean norm.

We note that the AMP algorithm introduced in (2) and (3) is a special case of the general recursion introduced (7) and (8). Indeed, define the following vectors recursively for $t \geq 0$, starting with $\beta^0 = 0$ and $z^0 = y$.

$$\begin{aligned} h^{t+1} &= \beta_0 - (A^* z^t + \beta^t), \qquad & q^t &= \beta^t - \beta_0, \\ b^t &= w - z^t, & m^t &= -z^t. \end{aligned} \tag{14}$$

It can be verified that these vectors satisfy (7) and (8) using Lipschitz functions

$$f_t(\underline{a}, [\beta_0]_{i-k}^{i+k}) = \eta_{t-1}([\beta_0]_{i-k}^{i+k} - \underline{a}) - \beta_{0_i}, \quad \text{and} \quad g_t(b, w) = b - w, \tag{15}$$

where $\underline{a} \in \mathbb{R}^{2k+1}$ and $b \in \mathbb{R}$. Using this choice of f_t, g_t in (15) also yields the expressions for σ_t^2, τ_t^2 given in (5).

In what follows, the notation matches that of [9] but is repeated here for completeness. In the remaining analysis, the general recursion given in (7) and (8) is used. We can write vector equations to represent the recursion as follows: for all $t \geq 0$,

$$b^t + \lambda_t m^{t-1} = Aq^t, \quad \text{and} \quad h^{t+1} + \xi_t q^t = A^* m^t. \tag{16}$$

This yields matrix equations $X_t = A^* M_t$ and $Y_t = AQ_t$, where we define the individual matrices as

$$\begin{aligned} X_t &:= [h^1 + q^0 \mid h^2 + q^1 \mid \ldots \mid h^t + q^{t-1}], & Q_t &:= [q^0 \mid \ldots \mid q^{t-1}], \\ Y_t &:= [b^0 \mid b^1 + \lambda_1 m^0 \mid \ldots \mid b^{t-1} + \lambda_{t-1} m^{t-2}], & M_t &:= [m^0 \mid \ldots \mid m^{t-1}] \\ \Xi_t &:= \operatorname{diag}(\xi_0, \ldots, \xi_{t-1}) & H_t &:= [h^1 | \ldots | h^t], \\ \Lambda_t &:= \operatorname{diag}(\lambda_0, \ldots, \lambda_{t-1}) & B_t &:= [b^0 | \ldots | b^{t-1}]. \end{aligned} \tag{17}$$

In the above, $[c_1 \mid c_2 \mid \ldots \mid c_k]$ denotes a matrix with columns $c_1, \ldots, c_k$ and M_0, Q_0, B_0, H_0, and Λ_0 are defined to be the all-zero vector. From the above matrix definitions we have the following matrix equations $Y_t = B_t + \Lambda_t[0|M_{t-1}]$ and $X_t = H_t + \Xi_t Q_t$.

The values $m_\parallel^t$ and $q_\parallel^t$ are projections of m^t and q^t onto the column space of M_t and Q_t, with $m_\perp^t := m^t - m_\parallel^t$, and $q_\perp^t := q^t - q_\parallel^t$ being the projections onto the orthogonal complements of M_t and Q_t. Finally, define the vectors

$$\alpha^t := (\alpha_0^t, \ldots, \alpha_{t-1}^t)^*, \qquad \gamma^t := (\gamma_0^t, \ldots, \gamma_{t-1}^t)^* \tag{18}$$

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

$$m_\parallel^t := \sum_{i=0}^{t-1} \alpha_i^t m^i, \qquad q_\parallel^t := \sum_{i=0}^{t-1} \gamma_i^t q^i. \tag{19}$$

The technical lemma, Lemma 3, shows that for large n, the entries of the vectors α^t and γ^t concentrate to constant values which are defined in the following section.

3.2 Concentrating Constants

Recall that $\beta_0 \in \mathbb{R}^N$ is the unknown vector to be recovered and $w \in \mathbb{R}^n$ is the measurement noise. Using the definitions in (14), note that the vector h^{t+1} is the noise in the observation $A^* z^t + \beta^t$ (from the true β_0), while q^t is the error in the estimate β^t. The technical lemma will show that h^t and m^t are approximately i.i.d. $\mathcal{N}(0, \tau_t^2)$, while q^t and b^t are approximately i.i.d. $\mathcal{N}(0, \sigma_t^2)$. In this

section we introduce the concentrating values for various inner products of the values $\{h^t, m^t, q^t, b^t\}$ that are used in Lemma 3.

First define the concentrating values for λ_{t+1} and ξ_t defined in (8) as

$$\hat{\lambda}_{t+1} := \frac{1-w_k}{\delta}\mathbb{E}[f_t'(\tau_t \tilde{\underline{Z}}_t, \underline{\beta})], \quad \text{and } \hat{\xi}_t = \mathbb{E}[g_t'(\sigma_t \breve{Z}_t, W)]. \tag{20}$$

Next, let $\{\breve{Z}_t\}_{t\geq 0}$ be a sequence of zero-mean jointly Gaussian random variables on $\mathbb{R}$, and let $\{\tilde{\underline{Z}}_t\}_{t\geq 0}$ be a sequence of zero-mean jointly Gaussian random vectors on $\mathbb{R}^{2k+1}$, where $\tilde{\underline{Z}}_t$ has i.i.d. coordinates $\forall t \geq 0$, and $\tilde{\underline{Z}}_{t_i}$ and $\tilde{\underline{Z}}_{r_j}$ are independent when $i \neq j$. The covariance of the two random sequences is defined recursively as follows. For $r, t \geq 0$,

$$\mathbb{E}[\breve{Z}_r \breve{Z}_t] = \frac{\tilde{E}_{r,t}}{\sigma_r \sigma_t}, \qquad \mathbb{E}[\tilde{\underline{Z}}_{r_i} \tilde{\underline{Z}}_{t_i}] = \frac{\breve{E}_{r,t}}{\tau_r \tau_t}, \quad \forall i = 1, ..., 2k+1 \tag{21}$$

where

$$\tilde{E}_{r,t} := \frac{(1-w_k)\mathbb{E}[f_r(\tau_{r-1}\tilde{\underline{Z}}_{r-1}, \underline{\beta}) f_t(\tau_{t-1}\tilde{\underline{Z}}_{t-1}, \underline{\beta})] + w_k \sigma_\beta^2}{\delta}, \qquad \breve{E}_{r,t} := \mathbb{E}[g_r(\sigma_r \breve{Z}_r, W) g_t(\sigma_t \breve{Z}_t, W)], \tag{22}$$

where $w_k = 2k/N$ and σ_β^2 was defined in (9). Note that both terms of the above are scalar values and we take $f_0(\cdot, \underline{\beta}) := f_0(\underline{0}, \underline{\beta})$, the initial condition. Moreover, $\tilde{E}_{t,t} = \sigma_t^2$ and $\breve{E}_{t,t} = \tau_t^2$, as can be seen from (13), thus $\mathbb{E}[\tilde{Z}_{t_i}^2] = \mathbb{E}[\breve{Z}_t^2] = 1$.

Define matrices $\tilde{C}^t, \breve{C}^t \in \mathbb{R}^{t\times t}$ for $t \geq 1$ taking values from (22) as

$$\tilde{C}^t_{i+1,j+1} = \tilde{E}_{i,j}, \quad \text{and} \quad \breve{C}^t_{i+1,j+1} = \breve{E}_{i,j}, \quad 0 \leq i, j \leq t-1. \tag{23}$$

Then, concentrating values for γ^t and α^t defined in (18) (as long as $\tilde{C}^t$ and $\breve{C}^t$ are invertible) are

$$\hat{\gamma}^t := (\tilde{C}^t)^{-1}\tilde{E}_t, \quad \text{and} \quad \hat{\alpha}^t := (\breve{C}^t)^{-1}\breve{E}_t, \tag{24}$$

where

$$\tilde{E}_t := (\tilde{E}_{0,t} \ldots, \tilde{E}_{t-1,t})^*, \quad \text{and} \quad \breve{E}_t := (\breve{E}_{0,t} \ldots, \breve{E}_{t-1,t})^*. \tag{25}$$

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

$$\begin{aligned}(\sigma_t^\perp)^2 &:= \sigma_t^2 - (\hat{\gamma}^t)^* \tilde{E}_t = \tilde{E}_{t,t} - \tilde{E}_t^* (\tilde{C}^t)^{-1}\tilde{E}_t, \\ (\tau_t^\perp)^2 &:= \tau_t^2 - (\hat{\alpha}^t)^* \breve{E}_t = \breve{E}_{t,t} - \breve{E}_t^* (\breve{C}^t)^{-1}\breve{E}_t.\end{aligned} \tag{26}$$

Lemma 1. *If $(\sigma_k^\perp)^2$ and $(\tau_k^\perp)^2$ are bounded below by some positive constants for $k \leq t$, then the matrices $\tilde{C}^{k+1}$ and $\breve{C}^{k+1}$ defined in (23) are invertible for $k \leq t$.*

Proof. The proof can be found in [9].

□

3.3 Conditional Distribution Lemma

As mentioned, the proof of Theorem 1 relies on two technical lemmas. The first lemma, presented in this section, provides the conditional distribution of the vectors h^{t+1} and b^t given the matrices in (17) as well as β_0, w. Lemma 2 shows that these conditional distributions can be represented as the

sum of a standard Gaussian vector an a deviation term. Then the second technical lemma, Lemma 3, shows that the deviation terms are small, meaning that their standardized norms concentrate on zero, and also provides concentration results for various inner products involving the other terms in recursion (7), namely $\{h^{t+1}, q^t, b^t, m^t\}$.

The following notation is used for the concentration lemmas. Considering two random vectors X, Y and a sigma-algebra $\mathscr{S}$, we denote the fact that that conditional distribution of X given $\mathscr{S}$ equals the distribution of Y as $X|_{\mathscr{S}} \stackrel{d}{=} Y$. We represent a $t \times t$ identity matrix as I_t, dropping the t subscript when it's obvious. For a matrix A with full column rank, $\mathsf{P}_A^{\|} := A(A^*A)^{-1}A^*$ is the orthogonal projection matrix onto the column space of A, and $\mathsf{P}_A^{\perp} := \mathsf{I} - \mathsf{P}_A^{\|}$.

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

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

Lemma 2. *[9, Lemma 4] For vectors h^{t+1} and b^t defined in (7), the following conditional distributions hold for $t \geq 1$:*

$$h^1|_{\mathscr{S}_{1,0}} \stackrel{d}{=} \tau_0 Z_0 + \Delta_{1,0}, \quad \text{and} \quad h^{t+1}|_{\mathscr{S}_{t+1,t}} \stackrel{d}{=} \sum_{r=0}^{t-1} \hat{\alpha}_r^t h^{r+1} + \tau_t^{\perp} Z_t + \Delta_{t+1,t}, \tag{27}$$

$$b^0|_{\mathscr{S}_{0,0}} \stackrel{d}{=} \sigma_0 Z_0' + \Delta_{0,0}, \quad \text{and} \quad b^t|_{\mathscr{S}_{t,t}} \stackrel{d}{=} \sum_{r=0}^{t-1} \hat{\gamma}_r^t b^r + \sigma_t^{\perp} Z_t' + \Delta_{t,t}. \tag{28}$$

where $Z_0, Z_t \in \mathbb{R}^N$ and $Z_0', Z_t' \in \mathbb{R}^n$ are i.i.d. standard Gaussian random vectors that are independent of the corresponding conditioning sigma algebras. The terms $\hat{\gamma}_i^t$ and $\hat{\alpha}_i^t$ for $i = 0, ..., t-1$ are defined in (24) and the terms $(\tau_t^{\perp})^2$ and $(\sigma_t^{\perp})^2$ in (26). The deviation terms are

$$\Delta_{0,0} = \left(\frac{\|q^0\|}{\sqrt{n}} - \sigma_0 \right) Z_0', \tag{29}$$

$$\Delta_{1,0} = \left[\left(\frac{\|m^0\|}{\sqrt{n}} - \tau_0 \right) \mathsf{I}_N - \frac{\|m^0\|}{\sqrt{n}} \mathsf{P}_{q^0}^{\|} \right] Z_0 + q^0 \left(\frac{\|q^0\|^2}{n} \right)^{-1} \left(\frac{(b^0)^* m_0}{n} - \xi_0 \frac{\|q^0\|^2}{n} \right), \tag{30}$$

and for $t > 0$,

$$\begin{aligned} \Delta_{t,t} = \sum_{r=0}^{t-1} (\gamma_r^t - \hat{\gamma}_r^t) b^r + \left[\left(\frac{\|q_{\perp}^t\|}{\sqrt{n}} - \sigma_t^{\perp} \right) \mathsf{I}_n - \frac{\|q_{\perp}^t\|}{\sqrt{n}} \mathsf{P}_{M_t}^{\|} \right] Z_t' \\ + M_t \left(\frac{M_t^* M_t}{n} \right)^{-1} \left(\frac{H_t^* q_{\perp}^t}{n} - \frac{M_t}{n}^* \left[\lambda_t m^{t-1} - \sum_{r=1}^{t-1} \lambda_r \gamma_r^t m^{r-1} \right] \right), \end{aligned} \tag{31}$$

$$\begin{aligned} \Delta_{t+1,t} = \sum_{r=0}^{t-1} (\alpha_r^t - \hat{\alpha}_r^t) h^{r+1} + \left[\left(\frac{\|m_{\perp}^t\|}{\sqrt{n}} - \tau_t^{\perp} \right) \mathsf{I}_N - \frac{\|m_{\perp}^t\|}{\sqrt{n}} \mathsf{P}_{Q_{t+1}}^{\|} \right] Z_t \\ + Q_{t+1} \left(\frac{Q_{t+1}^* Q_{t+1}}{n} \right)^{-1} \left(\frac{B_{t+1}^* m_{\perp}^t}{n} - \frac{Q_{t+1}^*}{n} \left[\xi_t q^t - \sum_{i=0}^{t-1} \xi_i \alpha_i^t q^i \right] \right). \end{aligned} \tag{32}$$

Proof. The proof can be found in [9]. □

Lemma 2 holds only when $M_t^* M_t$ and $Q_{t_1}^* Q_{t_1}$ are invertible.

3.4 Main Concentration Lemma

We use the shorthand $X_n \doteq c$ to denote the concentration inequality $P(|X_n - c| \geq \epsilon) \leq K_t e^{-\kappa_t n \epsilon^2}$. As specified in the theorem statement, the lemma holds for all $\epsilon \in (0, 1)$, with K_t, κ_t denoting generic constants depending on t, but not on n or ϵ.

Lemma 3. *With the $\doteq$ notation defined above, the following statements hold for $t \geq 0$.*

(a)

$$P\left(\frac{1}{N}\left\|\Delta_{t+1,t}\right\|^2 \geq \epsilon\right) \leq K_t e^{-\kappa_t n \epsilon}, \tag{33}$$

$$P\left(\frac{1}{n}\left\|\Delta_{t,t}\right\|^2 \geq \epsilon\right) \leq K_t e^{-\kappa_t n \epsilon}. \tag{34}$$

(b) i) *For pseudo-Lipschitz functions $\phi_h : \mathbb{R}^{(t+2)(2k+1)} \to \mathbb{R}$*

$$\frac{1}{N-2k}\sum_{i=k+1}^{N-k} \phi_h\left([h^1]_{i-k}^{i+k}, \ldots, [h^{t+1}]_{i-k}^{i+k}, [\beta_0]_{i-k}^{i+k}\right) \doteq \mathbb{E}\left[\phi_h\left(\tau_0 \underline{\tilde{Z}}_0, \ldots, \tau_t \underline{\tilde{Z}}_t, \underline{\beta}\right)\right]. \tag{35}$$

The random vectors $\underline{\tilde{Z}}_0, \ldots, \underline{\tilde{Z}}_t \in \mathbb{R}^{2k+1}$ are jointly Gaussian with zero mean entries which are independent of the other entries in the same vector with covariance across iterations given by (21), and are independent of $\underline{\beta} \sim \pi$.

ii) *Let $\psi_h : \mathbb{R}^{2(2k+1)} \to \mathbb{R}$ be a bounded function that is differentiable in the first argument with bounded derivative, meaning that the first $2k+1$ partial derivatives exist and are bounded. Then,*

$$\frac{1}{N-2k}\sum_{i=k+1}^{N-k} \psi_h([h^{t+1}]_{i-k}^{i+k}, [\beta_0]_{i-k}^{i+k}) \doteq \mathbb{E}\left[\psi_h(\tau_t \underline{\tilde{Z}}_t, \underline{\beta})\right]. \tag{36}$$

As above, $\underline{\tilde{Z}}_t \in \mathbb{R}^{2k+1}$ has entries that are i.i.d. $\sim \mathcal{N}(0,1)$ and $\underline{\beta} \in \mathbb{R}^{2k+1} \sim \pi$ are independent.

iii) *For pseudo-Lipschitz functions $\phi_b : \mathbb{R}^{t+2} \to \mathbb{R}$*

$$\frac{1}{n}\sum_{i=1}^{n} \phi_b\left(b_i^0, \ldots, b_i^t, w_i\right) \doteq \mathbb{E}\left[\phi_b\left(\sigma_0 \breve{Z}_0, \ldots, \sigma_t \breve{Z}_t, W\right)\right]. \tag{37}$$

The random variables $\breve{Z}_0, \ldots, \breve{Z}_t$ are jointly Gaussian with zero mean and covariance given by (21), and are independent of $W \sim p_w$.

iv) *Let $\psi_b : \mathbb{R} \to \mathbb{R}$ be a bounded function that is differentiable in the first argument with bounded derivative. Then,*

$$\frac{1}{n}\sum_{i=1}^{n} \psi_b(b_i^t, w_i) \doteq \mathbb{E}\left[\psi_b(\sigma_t \breve{Z}_t, W)\right]. \tag{38}$$

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

(c)

$$\frac{(h^{t+1})^* q^0}{n} \doteq 0, \quad \frac{(h^{t+1})^* \beta_0}{n} \doteq 0, \tag{39}$$

$$\frac{(b^t)^* w}{n} \doteq 0. \tag{40}$$

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

$$\frac{(h^{r+1})^* h^{t+1}}{N} \doteq \breve{E}_{r,t}, \tag{41}$$

$$\frac{(b^r)^* b^t}{n} \doteq \tilde{E}_{r,t}. \tag{42}$$

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

$$\frac{(q^0)^* q^{t+1}}{n} \doteq \tilde{E}_{0,t+1}, \quad \frac{(q^{r+1})^* q^{t+1}}{n} \doteq \tilde{E}_{r+1,t+1}, \tag{43}$$

$$\frac{(m^r)^* m^t}{n} \doteq \breve{E}_{r,t} \tag{44}$$

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

$$\lambda_t \doteq \hat{\lambda}_t, \quad \frac{(h^{t+1})^* q^{r+1}}{n} \doteq \hat{\lambda}_{r+1} \breve{E}_{r,t}, \quad \frac{(h^{r+1})^* q^{t+1}}{n} \doteq \hat{\lambda}_{t+1} \breve{E}_{r,t}, \tag{45}$$

$$\xi_t \doteq \hat{\xi}_t, \quad \frac{(b^r)^* m^t}{n} \doteq \hat{\xi}_t \tilde{E}_{r,t}, \quad \frac{(b^t)^* m^r}{n} \doteq \hat{\xi}_r \tilde{E}_{r,t}. \tag{46}$$

(g) For $\boldsymbol{Q}_{t+1} = \frac{1}{n} Q_{t+1}^* Q_{t+1}$ *and* $\boldsymbol{M}_t = \frac{1}{n} M_t^* M_t$, *when the inverses exist, for all* $0 \le i, j \le t$ *and* $0 \le i', j' \le t-1$*:*

$$\left[\mathbf{Q}_{t+1}^{-1}\right]_{i+1,j+1} \doteq [(\tilde{C}^{t+1})^{-1}]_{i+1,j+1}, \quad \gamma_i^{t+1} \doteq \hat{\gamma}_i^{t+1}, \tag{47}$$

$$\left[\mathbf{M}_t^{-1}\right]_{i'+1,j'+1} \doteq [(\breve{C}^t)^{-1}]_{i'+1,j'+1}, \quad \alpha_{k'}^t \doteq \hat{\alpha}_{i'}^t, \quad t \ge 1, \tag{48}$$

where $\hat{\gamma}_k^{t+1}$ *and* $\hat{\alpha}_{k'}^t$ *are defined in* (24),

(h) With $\sigma_{t+1}^{\perp}, \tau_t^{\perp}$ *defined in* (26),

$$\frac{1}{n} \left\| q_{\perp}^{t+1} \right\|^2 \doteq (\sigma_{t+1}^{\perp})^2, \tag{49}$$

$$\frac{1}{n} \left\| m_{\perp}^{t} \right\|^2 \doteq (\tau_t^{\perp})^2. \tag{50}$$

3.5 Proof of Theorem 1

Proof. Applying Part (b)(i) of Lemma 3 to a pseudo-Lipschitz (PL) function $\phi_h : \mathbb{R}^{2(2k+1)} \to \mathbb{R}$,

$$P\left(\left| \frac{1}{N-2k} \sum_{i=k+1}^{N-k} \phi_h([h^{t+1}]_{i-k}^{i+k}, [\beta_0]_{i-k}^{i+k}) - \mathbb{E}\left[\phi_h(\tau_t \underline{Z}, \underline{\beta})\right] \right| \ge \epsilon \right) \le K_t e^{-\kappa_t n \epsilon^2}$$

where the random vectors $\underline{\beta} \in \mathbb{R}^{2k+1} \sim \pi$ and $\underline{Z} \in \mathbb{R}^{2k+1}$, whose entries are i.i.d. standard normal random variables, are independent. Now for $i = k+1, ..., N-k$ let

$$\phi_h([h^{t+1}]_{i-k}^{i+k}, [\beta_0]_{i-k}^{i+k}) := \phi(\eta_t([\beta_0 - h^{t+1}]_{i-k}^{i+k}), \beta_{0_i}), \tag{51}$$

where $\phi : \mathbb{R}^2 \to \mathbb{R}$ is the PL function in the statement of the theorem. The function $\phi_h([h^{t+1}]_{i-k}^{i+k}, [\beta_0]_{i-k}^{i+k})$ in (51) is PL since ϕ is PL and η_t is Lipschitz. We therefore obtain

$$P\left(\left|\frac{1}{N-2k}\sum_{i=k+1}^{N-k}\phi(\eta_t([\beta_0 - h^{t+1}]_{i-k}^{i+k}), \beta_{0_i}) - \mathbb{E}\left[\phi(\eta_t(\underline{\beta} - \tau_t\underline{Z}), \beta)\right]\right| \geq \epsilon\right) \leq K_t e^{-\kappa_t n\epsilon^2},$$

where $\beta \in \mathbb{R} \sim \gamma_\beta$. The proof is completed by noting from (3) and (14) that $\beta_i^{t+1} = \eta_t([A^*z^t + \beta^t]_{i-k}^{i+k}) = \eta_t([\beta_0 - h^{t+1}]_{i-k}^{i+k})$. □

4 Proof of Lemma 3

4.1 Mathematical Preliminaries

Fact 1. *[9, Fact 1] Let $u \in \mathbb{R}^N$ and $v \in \mathbb{R}^n$ be deterministic vectors, and let $\tilde{A} \in \mathbb{R}^{n\times N}$ be a matrix with independent $\mathcal{N}(0, 1/n)$ entries. Then:*

(a)

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

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

(b) Let $\mathcal{W}$ be a d-dimensional subspace of $\mathbb{R}^n$ for $d \leq n$. Let $(w_1, ..., w_d)$ be an orthogonal basis of $\mathcal{W}$ with $\|w_\ell\|^2 = n$ for $\ell \in [d]$, and let $\mathsf{P}_{\mathcal{W}}^{\|}$ denote the orthogonal projection operator onto $\mathcal{W}$. Then for $D = [w_1 \mid \ldots \mid w_d]$, we have $\mathsf{P}_{\mathcal{W}}^{\|}\tilde{A}u \stackrel{d}{=} \frac{\|u\|}{\sqrt{n}}\mathsf{P}_{\mathcal{W}}^{\|}Z_u \stackrel{d}{=} \frac{\|u\|}{\sqrt{n}}Dx$ where $x \in \mathbb{R}^d$ is a random vector with i.i.d. $\mathcal{N}(0, 1/n)$ entries.

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

Fact 3. *Let $g : \mathbb{R}^{2k+1} \to \mathbb{R}$ be differentiable function with bounded partial derivatives. For all $s, \Delta \in \mathbb{R}^{2k+1}$, there exists a constant $\boldsymbol{B} > 0$ such that $|g(s + \Delta) - g(s)| \leq \boldsymbol{B}\sqrt{2k+1}\,\|\Delta\|$.*

Proof of Fact 3. From Taylor's Theorem, for any $s, \Delta \in \mathbb{R}^{2k+1}$,

$$g(s + \Delta) = g(s) + \Delta^*\nabla g(s + \xi\Delta), \tag{52}$$

for some $\xi \in [0, 1]$. Moreover, for any $\tilde{s}$ on the line segment joining s and $s + \Delta$, we have that the partial derivatives of $g(\tilde{s})$ are upper bounded by some constant $\mathbf{B} > 0$. From (52) it follows that

$$|g(s + \Delta) - g(s)| = |\Delta^*\nabla g(s + \xi\Delta)| \stackrel{(a)}{\leq} \|\Delta\|\,\|\nabla g(s + \xi\Delta)\| \leq \|\Delta\|\,\mathbf{B}\sqrt{2k+1},$$

where step (a) follows from Cauchy-Schwarz. The above gives the desired result. □

We also make use of concentration results that are listed in Appendices A, B, and C. Many of these results and their proofs can be found in Rush and Venkataramanan [9]. Appendix D holds concentration results for dependent variables that were needed to provide the new results in this paper, many of which we believe to be interesting results on their own, such as concentration for psuedo-Lipschitz functions acting on Markovian input.

The proof of Lemma 3. proceeds by induction on t. We label as $\mathcal{H}^{t+1}$ the results (33), (35), (36), (39), (41), (43), (45), (47), (49) and similarly as $\mathcal{B}^t$ the results (34), (37), (38), (40), (42), (44), (46), (48), (50). The proof consists of four steps:

1. $\mathcal{B}_0$ holds.

2. $\mathcal{H}_1$ holds.

3. If $\mathcal{B}_r, \mathcal{H}_s$ holds for all $r < t$ and $s \leq t$, then $\mathcal{B}_t$ holds.

4. if $\mathcal{B}_r, \mathcal{H}_s$ holds for all $r \leq t$ and $s \leq t$, then $\mathcal{H}_{t+1}$ holds.

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

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

We wish to show results (a)–(h) in (34), (37), (38), (40), (42), (44), (46), (48), (50) for the case $t = 0$. The proof of these results is the same as in the step $\mathcal{B}_0$ of the proof in [9] and therefore is not repeated here.

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

We wish to show results (a)–(h) in (33), (35), (36), (39), (41), (43), (45), (47), (49) for the case $t = 0$.

(a) The proof of $\mathcal{H}_1(a)$ follows as the corresponding proof in [9].

(b)(i) For $t = 0$, the LHS of (35) can be bounded as

$$
\begin{aligned}
&P\left(\left|\frac{1}{N-2k}\sum_{i=k+1}^{N-k}\phi_h([h^1]_{i-k}^{i+k},[\beta_0]_{i-k}^{i+k}) - \mathbb{E}[\phi_h(\tau_0\tilde{\underline{Z}}_0,\underline{\beta})]\right| \geq \epsilon\right)\\
&\overset{(a)}{=} P\left(\left|\frac{1}{N-2k}\sum_{i=k+1}^{N-k}\phi_h([\tau_0 Z_0+\Delta_{1,0}]_{i-k}^{i+k},[\beta_0]_{i-k}^{i+k}) - \mathbb{E}[\phi_h(\tau_0\tilde{\underline{Z}}_0,\underline{\beta})]\right| \geq \epsilon\right)\\
&\overset{(b)}{\leq} P\left(\left|\frac{1}{N-2k}\sum_{i=k+1}^{N-k}\mathbb{E}_{\tilde{\underline{Z}}_0}[\phi_h(\tau_0\tilde{\underline{Z}}_0,[\beta_0]_{i-k}^{i+k})] - \mathbb{E}_{\tilde{\underline{Z}}_0,\underline{\beta}}[\phi_h(\tau_0\tilde{\underline{Z}}_0,\underline{\beta})]\right| \geq \frac{\epsilon}{3}\right) \qquad (53)\\
&\quad + P\left(\left|\frac{1}{N-2k}\sum_{i=k+1}^{N-k}\left[\phi_h(\tau_0[Z_0]_{i-k}^{i+k},[\beta_0]_{i-k}^{i+k}) - \mathbb{E}_{\tilde{\underline{Z}}_0}[\phi_h(\tau_0\tilde{\underline{Z}}_0,[\beta_0]_{i-k}^{i+k})]\right]\right| \geq \frac{\epsilon}{3}\right)\\
&\quad + P\left(\left|\frac{1}{N-2k}\sum_{i=k+1}^{N-k}\left[\phi_h([\tau_0 Z_0+\Delta_{1,0}]_{i-k}^{i+k},[\beta_0]_{i-k}^{i+k}) - \phi_h(\tau_0[Z_0]_{i-k}^{i+k},[\beta_0]_{i-k}^{i+k})\right]\right| \geq \frac{\epsilon}{3}\right).
\end{aligned}
$$

Step (a) follows from the conditional distribution of h^1 given in Lemma 2 (27) and step (b) Lemma A.1. Label the terms on the RHS of (53) as $T_1 - T_3$. We show that each of these terms is bounded above by $Ke^{-\kappa N\epsilon^2}$. Term T_1 is upper bounded by $Ke^{-\kappa(N-2k)\epsilon^2}$ using Lemma D.7 since the function $\tilde{\phi}_h : \mathbb{R}^{2k+1} \to \mathbb{R}$ defined as $\tilde{\phi}_h(\underline{s}) := \mathbb{E}_{\underline{Z}}[\phi_h(\tau_0\underline{Z},\underline{s})]$ is PL(2) by Lemma C.2. Term T_2 is upper bounded by $Ke^{-\kappa(N-2k)\epsilon^2}$ using Lemma D.6 since the function $\hat{\phi}_{h,i} : \mathbb{R}^{2k+1} \to \mathbb{R}$ defined as

$$
\hat{\phi}_{h,i}(\underline{s}) := \phi_h(\underline{s},[\beta]_{i-k}^{i+k}) \in PL(2), \quad \text{for } k+1 \leq i \leq N-k, \qquad (54)
$$

by Lemma C.2, where we have used the fact that $\mathbb{E}_{[Z_0]_{i-k}^{i+k}}[\phi_h(\tau_0[Z_0]_{i-k}^{i+k}, [\beta_0]_{i-k}^{i+k})] = \mathbb{E}_{\underline{\tilde{Z}}_0}[\phi_h(\tau_0 \underline{\tilde{Z}}_0, [\beta_0]_{i-k}^{i+k})]$ for each $k+1 \le i \le N-k$. Finally consider T_3, the third term on the RHS of (53).

$$
\begin{aligned}
T_2 &\overset{(a)}{\le} P\left(\frac{1}{N-2k}\sum_{i=k+1}^{N-k} L\left(1 + \left\|[\tau_0 Z_0 + \Delta_{1,0}]_{i-k}^{i+k}\right\| + \left\|\tau_0[Z_0]_{i-k}^{i+k}\right\|\right)\left\|[\Delta_{1,0}]_{i-k}^{i+k}\right\| \ge \frac{\epsilon}{3}\right) \\
&\overset{(b)}{\le} P\left(\frac{\|\Delta_{1,0}\|}{\sqrt{N-2k}} \cdot \left(1 + \sqrt{2k+1}\frac{\|\Delta_{1,0}\|}{\sqrt{N-2k}} + 2\tau_0\sqrt{2k+1}\frac{\|Z_0\|}{\sqrt{N-2k}}\right) \ge \frac{\epsilon}{3L\sqrt{3(2k+1)}}\right). \quad (55)
\end{aligned}
$$

Step (a) follows by the triangle inequality and (54) with the PL constant denoted by L and step (b) by $\left\|[\tau_0 Z_0 + \Delta_{1,0}]_{i-k}^{i+k}\right\| = \left\|\tau_0[Z_0]_{i-k}^{i+k}\right\| + \left\|[\Delta_{1,0}]_{i-k}^{i+k}\right\|$, the Cauchy-Schwarz inequality, the fact that for $a \in \mathbb{R}^N$, $\sum_{i=k+1}^{N-k}\left\|[a]_{i-k}^{i+k}\right\|^2 \le (2k+1)\|a\|^2$, and the following application of Lemma C.3:

$$
\sum_{i=k+1}^{N-k}\left(1 + \left\|[\Delta_{1,0}]_{i-k}^{i+k}\right\| + 2\left\|\tau_0[Z_0]_{i-k}^{i+k}\right\|\right)^2 \le 3\left((N-2k) + (2k+1)\|\Delta_{1,0}\|^2 + 4\tau_0^2(2k+1)\|Z_0\|^2\right).
$$

From (55), we have

$$
T_2 \le P\left(\frac{\|Z_0\|}{\sqrt{N-2k}} \ge 2\right) + P\left(\frac{\|\Delta_{1,0}\|}{\sqrt{N-2k}} \ge \frac{\frac{\epsilon}{\sqrt{2k+1}}\min\left\{1, \frac{1}{6L\sqrt{3}}\right\}}{2 + 4\tau_0\sqrt{2k+1}}\right) \overset{(a)}{\le} e^{-(N-2k)} + Ke^{-\kappa(N-2k)\epsilon^2},
$$

where to obtain (a),we use Lemma B.2 and $\mathcal{H}_1(a)$.

(b)(ii) Using the conditional distribution of h^1 stated in Lemma 2 (27), we have

$$
\begin{aligned}
&P\left(\left|\frac{1}{N-2k}\sum_{i=k+1}^{N-k}\psi_h([h^1]_{i-k}^{i+k}, [\beta_0]_{i-k}^{i+k}) - \mathbb{E}[\psi_h(\tau_0\underline{\tilde{Z}}_0, \underline{\beta})]\right| \ge \epsilon\right) \\
&= P\left(\left|\frac{1}{N-2k}\sum_{i=k+1}^{N-k}\psi_h([\tau_0 Z_0 + \Delta_{1,0}]_{i-k}^{i+k}, [\beta_0]_{i-k}^{i+k}) - \mathbb{E}[\psi_h(\tau_0\underline{\tilde{Z}}_0, \underline{\beta})]\right| \ge \epsilon\right) \\
&\overset{(a)}{\le} P\left(\left|\frac{1}{N-2k}\sum_{i=k+1}^{N-k}\left[\psi_h([\tau_0 Z_0 + \Delta_{1,0}]_{i-k}^{i+k}, [\beta_0]_{i-k}^{i+k}) - \psi_h(\tau_0[Z_0]_{i-k}^{i+k}, [\beta_0]_{i-k}^{i+k})\right]\right| \ge \frac{\epsilon}{3}\right) \quad (56) \\
&\quad + P\left(\left|\frac{1}{N-2k}\sum_{i=k+1}^{N-k}\left[\psi_h(\tau_0[Z_0]_{i-k}^{i+k}, [\beta_0]_{i-k}^{i+k}) - \mathbb{E}_{\underline{\tilde{Z}}_0}[\psi_h(\tau_0\underline{\tilde{Z}}_0, [\beta_0]_{i-k}^{i+k})]\right]\right| \ge \frac{\epsilon}{3}\right) \\
&\quad + P\left(\left|\frac{1}{N-2k}\sum_{i=k+1}^{N-k}\mathbb{E}_{\underline{\tilde{Z}}_0}[\psi_h(\tau_0\underline{\tilde{Z}}_0, [\beta_0]_{i-k}^{i+k})] - \mathbb{E}[\psi_h(\tau_0\underline{\tilde{Z}}_0, \underline{\beta})]\right| \ge \frac{\epsilon}{3}\right).
\end{aligned}
$$

Step (a) follows from Lemma A.1. Label the terms on the RHS as $T_1 - T_3$. To complete the proof we show that each term is upper bounded by $Ke^{-\kappa n\epsilon^2}$. The function $\hat{\psi}_{i,h} : \mathbb{R}^{2k+1} \to \mathbb{R}$ defined as $\hat{\psi}_{i,h}(\underline{s}) := \psi_h(\underline{s}, [\beta_0]_{i-k}^{i+k})$ is bounded, hence using Lemma D.2 term T_2 has the desired upper bound using the fact that $\mathbb{E}_{[Z_0]_{i-k}^{i+k}}[\psi_h(\tau_0[Z_0]_{i-k}^{i+k}, [\beta_0]_{i-k}^{i+k})] = \mathbb{E}_{\underline{\tilde{Z}}_0}[\psi_h(\tau_0\underline{\tilde{Z}}_0, [\beta_0]_{i-k}^{i+k})]$ for all $k+1 \le i \le N-k$. The function $\tilde{\psi}_h : \mathbb{R}^{2k+1} \to \mathbb{R}$ defined $\tilde{\psi}_h(\underline{s}) := \mathbb{E}_{\underline{Z}}[\psi_h(\tau_0\underline{Z}, \underline{s})]$ is also bounded, hence using Lemma D.8 term T_3 has the desired bound.

Finally consider T_1, the first term in (56). Recall the definition of $\Delta_{1,0}$ in Lemma 2, and let $u \in \mathbb{R}^N$ be a vector such that $u_i = q_i^0 \left(\frac{(b^0)^* m^0}{\|q^0\|^2} - \xi_0 \right)$, for all $k+1 \le i \le N-k$. Then we have

$$[\tau_0 Z_0 + \Delta_{1,0}]_{i-k}^{i+k} = \frac{\left\|m^0\right\|}{\sqrt{n}} \left[\left(\mathsf{I} - \mathsf{P}_{q^0}^{\|} \right) Z_0 \right]_{i-k}^{i+k} + [u]_{i-k}^{i+k}.$$

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

$$\mathcal{F} := \left\{ \left| \frac{1}{\sqrt{n}} \left\|m^0\right\| - \tau_0 \right| \ge \epsilon_0 \right\} \cup \left\{ \frac{1}{\sqrt{n}} \|u\| \ge \epsilon_0 \right\}. \tag{57}$$

Denoting the event we are considering in T_1 by Π_1, we have

$$T_1 = P(\Pi_1) \le P(\mathcal{F}) + P(\Pi_1 \mid \mathcal{F}^c) \le K e^{-\kappa n \epsilon_0^2} + P(\Pi_1 \mid \mathcal{F}^c). \tag{58}$$

where the last inequality is by $\mathcal{B}_0$ (e) and

$$\begin{aligned} P\left(\frac{\|u\|}{\sqrt{n}} \ge \epsilon_0 \right) &= P\left(\left| \frac{(b^0)^* m^0}{\sqrt{n}\|q^0\|} - \xi_0 \frac{\|q^0\|}{\sqrt{n}} \right| \ge \epsilon_0 \right) \\ &\overset{(a)}{\le} P\left(\left| \frac{(b^0)^* m^0}{n} \cdot \frac{\sqrt{n}}{\|q^0\|} - \hat{\xi}_0 \sigma_0^2 \cdot \frac{1}{\sigma_0} \right| \ge \frac{\epsilon_0}{2} \right) + P\left(\left| \xi_0 \frac{\|q^0\|}{\sqrt{n}} - \hat{\xi}_0 \sigma_0 \right| \ge \frac{\epsilon_0}{2} \right) \\ &\overset{(b)}{\le} 2K \exp\left\{ \frac{-\kappa n \epsilon_0^2/4}{9 \max(1, \hat{\xi}_0^2 \sigma_0^4, \sigma_0^{-2})} \right\} + 2K \exp\left\{ \frac{-\kappa n \epsilon_0^2/4}{9 \max(1, \hat{\xi}_0^2, \sigma_0^2)} \right\}. \end{aligned}$$

Step (a) follows from Lemma A.1, and step (b) follows from Lemma A.2, $\mathcal{B}_0(f)$, the assumption on $\left\|q^0\right\|$ given in (12), and Lemma A.5.

Writing $P(\Pi_1|\mathcal{F}^c) = \mathbb{E}_{\cdot|\mathcal{F}^c}[P(\Pi_1|\mathcal{F}^c, \mathscr{S}_{1,0})]$, we now bound $P(\Pi_1|\mathcal{F}^c, \mathscr{S}_{1,0})$ which equals

$$P\left(\left| \frac{1}{N-2k} \sum_{i=k+1}^{N-k} \left(\psi_h\left(\frac{\left\|m^0\right\|}{\sqrt{n}} [(\mathsf{I} - \mathsf{P}_{q^0}^{\|}) Z_0]_{i-k}^{i+k} + [u]_{i-k}^{i+k}, [\beta_0]_{i-k}^{i+k} \right) - \psi_h(\tau_0 [Z_0]_{i-k}^{i+k}, [\beta_0]_{i-k}^{i+k}) \right) \right| \ge \frac{\epsilon}{3} \right),$$

conditional on the event $\mathcal{F}^c$ and sigma-field $\mathscr{S}_{1,0}$. In the above only Z_0 is random as the other terms are all in $\mathscr{S}_{1,0}$. In what follows, we drop the explicit conditioning on $\mathcal{F}^c, \mathscr{S}_{1,0}$ for brevity. Using Lemma A.1, we can upper bound $P(\Pi_1|\mathcal{F}^c, \mathscr{S}_{1,0})$ with

$$\begin{aligned} &P\left(\left| \frac{1}{N-2k} \sum_{i=k+1}^{N-k} \psi_h\left(\frac{\left\|m^0\right\|}{\sqrt{n}} [(\mathsf{I} - \mathsf{P}_{q^0}^{\|}) Z_0]_{i-k}^{i+k} + [u]_{i-k}^{i+k}, [\beta_0]_{i-k}^{i+k} \right) - \psi_h\left(\frac{\left\|m^0\right\|}{\sqrt{n}} [Z_0]_{i-k}^{i+k} + [u]_{i-k}^{i+k}, [\beta_0]_{i-k}^{i+k} \right) \right| \ge \frac{\epsilon}{6} \right) \\ &\quad + P\left(\left| \frac{1}{N-2k} \sum_{i=k+1}^{N-k} \psi_h\left(\frac{\left\|m^0\right\|}{\sqrt{n}} [Z_0]_{i-k}^{i+k} + [u]_{i-k}^{i+k}, [\beta_0]_{i-k}^{i+k} \right) - \psi_h(\tau_0 [Z_0]_{i-k}^{i+k}, [\beta_0]_{i-k}^{i+k}) \right| \ge \frac{\epsilon}{6} \right). \end{aligned} \tag{59}$$

Label the two terms in (59) as $T_{1,a}$ and $T_{1,b}$. To complete the proof we show that both are upper bounded by $K e^{-\kappa n \epsilon^2}$.

First consider $T_{1,a}$.

$$T_{1,a} \overset{(a)}{\leq} P\left(\frac{C\sqrt{2k+1}}{N-2k}\sum_{i=k+1}^{N-k}\frac{\left\|m^0\right\|}{\sqrt{n}}\left\|[\mathsf{P}^{\parallel}_{q^0}Z_0]_{i-k}^{i+k}\right\| \geq \frac{\epsilon}{6}\right)$$

$$\overset{(b)}{\leq} P\left(\frac{\sqrt{2k+1}}{N-2k}\sum_{i=k+1}^{N-k}|\tau_0+\epsilon_0|\left\|[\mathsf{P}^{\parallel}_{q^0}Z_0]_{i-k}^{i+k}\right\| \geq \frac{\epsilon}{6C}\right) \overset{(c)}{\leq} P\left(\frac{\sqrt{2k+1}}{N-2k}\sum_{i=k+1}^{N-k}\frac{\left\|[q^0]_{i-k}^{i+k}\right\|}{\|q^0\|}|Z| \geq \frac{\epsilon}{6C\,|\tau_0+\epsilon_0|}\right)$$

$$\overset{(d)}{\leq} P\left(\frac{|Z|}{\sqrt{N-2k}} \geq \frac{\epsilon}{6C(2k+1)\,|\tau_0+\epsilon_0|}\right) \overset{(e)}{\leq} e^{-\frac{\kappa(N-2k)\epsilon^2}{(2k+1)^2}}.$$

Step (a) holds by Fact 3 for a suitable constant $C>0$. Step (b) follows because we are conditioning on $\mathcal{F}^c$ defined in (57). Step (c) is obtained by writing out the expression for the vector $\mathsf{P}^{\parallel}_{q^0}Z_0$:

$$\mathsf{P}^{\parallel}_{q^0}Z_0 = \frac{q^0}{\|q^0\|}\left[\sum_{j=1}^{N}\frac{q_j^0}{\|q^0\|}Z_{0_j}\right] \overset{d}{=} \frac{q^0}{\|q^0\|}Z.$$

where from Fact 1, $Z\in\mathbb{R}$ is an independent standard Gaussian random variable. Step (d) follows from Cauchy-Schwarz as follows:

$$\left(\sum_{i=k+1}^{N-k}\left\|[q^0]_{i-k}^{i+k}\right\|\right)^2 \leq (N-2k)\sum_{i=k+1}^{N-k}\left\|[q^0]_{i-k}^{i+k}\right\|^2 \leq (N-2k)(2k+1)\left\|q^0\right\|^2,$$

and step (e) by Lemma B.1.

For $T_{1,b}$ the second term of (59), noting that all quantities except Z_0 are in $\mathscr{S}_{1,0}$, define the shorthand $\mathsf{diff}_i([Z_0]_{i-k}^{i+k}) := \psi_h(\frac{1}{\sqrt{n}}\left\|m^0\right\|[Z_0]_{i-k}^{i+k}+[u]_{i-k}^{i+k},[\beta_0]_{i-k}^{i+k}) - \psi_h(\tau_0[Z_0]_{i-k}^{i+k},[\beta_0]_{i-k}^{i+k})$. Then the upper tail of $T_{1,b}$ can be written as

$$P\left(\frac{1}{N-2k}\sum_{i=k+1}^{N-k}\mathsf{diff}_i([Z_0]_{i-k}^{i+k}) - \mathbb{E}[\mathsf{diff}_i([Z_0]_{i-k}^{i+k})] \geq \frac{\epsilon}{6} - \frac{1}{N-2k}\sum_{i=k+1}^{N-k}\mathbb{E}[\mathsf{diff}_i([Z_0]_{i-k}^{i+k})]\,\Big|\,\mathcal{F}^c,\mathscr{S}_{1,0}\right). \tag{60}$$

We now show that $\frac{1}{N-2k}\sum_{i=k+1}^{N-k}\left|\mathbb{E}[\mathsf{diff}_i([Z_0]_{i-k}^{i+k})]\right| \leq \frac{\epsilon}{12}$ for all $k+1\leq i\leq N-k$. From here on, we suppress the conditioning on $\mathcal{F}^c,\mathscr{S}_{0,0}$ for brevity. Denoting the standard normal density by ϕ, we have

$$\frac{1}{N-2k}\sum_{i=k+1}^{N-k}\left|\mathbb{E}[\mathsf{diff}_i([Z_0]_{i-k}^{i+k})]\right| \leq \frac{1}{N-2k}\sum_{i=k+1}^{N-k}\int_{\mathbb{R}^{2k+1}}\phi(\underline{z})\;|\mathsf{diff}_i(\underline{z})|\,d\underline{z}$$

$$\overset{(a)}{\leq} \frac{C\sqrt{2k+1}}{N-2k}\sum_{i=k+1}^{N-k}\int_{\mathbb{R}^{2k+1}}\phi(\underline{z})\left(\|\underline{z}\|\left|\frac{\left\|m^0\right\|}{\sqrt{n}}-\tau_0\right| + \|[u]_{i-k}^{i+k}\|\right)d\underline{z} \overset{(b)}{\leq} C(2k+1)\epsilon_0\left(1+\sqrt{\frac{n}{N-2k}}\right).$$

The above is upper bounded by $\epsilon/12$ if we choose

$$\epsilon_0 \leq \epsilon\left[12C(2k+1)\left(1+\sqrt{\frac{n}{N-2k}}\right)\right]^{-1}. \tag{61}$$

In the chain above, (a) follows from Fact 3 for a suitable constant $C>0$ as ψ_h is differentiable with bounded derivatives. The inequality (b) follows since $\mathcal{F}^c$ holds, from the fact $\mathbb{E}\,\|\underline{Z}\| \leq \sqrt{2k+1}$ when

$\underline{Z} \in \mathbb{R}^{2k+1}$ is a vector of i.i.d. standard Gaussian random variables, and since $\sum_{i=k+1}^{N-k} \|[u]_{i-k}^{i+k}\| \leq \sqrt{2k+1}\|u\|\sqrt{N-2k}$ by Cauchy-Schwarz.

The probability in (60) can then be upper bounded using Lemma D.2:

$$P\left(\frac{1}{N-2k}\sum_{i=k+1}^{N-k} \mathsf{diff}_i([Z_0]_{i-k}^{i+k}) - \mathbb{E}[\mathsf{diff}_i([Z_0]_{i-k}^{i+k})] \geq \frac{\epsilon}{6} \,\middle|\, \mathcal{F}^c, \mathscr{S}_{1,0}\right) \leq e^{-\kappa(N-2k)\epsilon^2}.$$

Using a similar bound for the lower tail and plugging into (59), we have shown that $P(\Pi_1 \mid \mathcal{F}^c) \leq 2e^{-\kappa(N-2k)\epsilon^2}$. Using this in (58) with ϵ_0 as in (61) proves that the first term in (56) is upper bounded by $Ke^{-\kappa(N-2k)\epsilon^2}$.

(c) We first show concentration for $(h^1)^*\beta_0/n$. Note that this result follows directly from $\mathcal{H}_1(b)(i)$. For example we can write $\left|(h^1)^*\beta_0\right| = \left|\sum_{i=1}^N h_i^1\beta_{0,i}\right| \leq \left|\sum_{i=1}^{N/2} h_i^1\beta_{0,i}\right| + \left|\sum_{j=N/2+1}^N h_j^1\beta_{0,j}\right|$ and therefore it follows by Lemma A.1,

$$P\left(\left|\frac{(h^1)^*\beta_0}{n}\right| \geq \epsilon\right) \leq P\left(\left|\sum_{i=1}^{N/2}\frac{h_i^1\beta_{0,i}}{N/2}\right| \geq \epsilon\delta\right) + P\left(\left|\sum_{j=N/2+1}^{N}\frac{h_j^1\beta_{0,j}}{N/2}\right| \geq \epsilon\delta\right) \overset{(a)}{\leq} Ke^{-\kappa N\epsilon^2\delta^2} + Ke^{-\kappa N\epsilon^2\delta^2}.$$

In the above, step (a) follows by applying $\mathcal{H}_1(b)(i)$ using PL(2) functions $\phi_{1,h}, \phi_{2,h}$ both defined from $\mathbb{R}^{2(2k+1)} \to \mathbb{R}$ equal to $\phi_{1,h}(\underline{x}, \underline{y}) = \underline{x}_1\underline{y}_1$, and $\phi_{2,h}(\underline{x}, \underline{y}) = \underline{x}_{2k+1}\underline{y}_{2k+1}$. Note that $\mathbb{E}[\tau_0\tilde{\underline{Z}}_{0,1}\underline{\beta}_1] = 0$.

Next we show concentration for $(h^1)^*q^0/n$. Note that

$$(h^1)^*q^0 = \sum_{i=1}^N h_i^1 q_i^0 = \sum_{i=k+1}^{N-k} h_i^1 f_0(\underline{0}, [\beta_0]_{i-k}^{i+k}) - \sum_{i=1}^{k} h_i^1\beta_{0,i} - \sum_{i=N-k+1}^{N} h_i^1\beta_{0,i}$$

where the last equality follows by definition of q^0 provided in (11). It follows by Lemma A.1,

$$\begin{aligned}
&P\left(\left|\frac{(h^1)^*q^0}{n}\right| \geq \epsilon\right) \\
&\leq P\left(\left|\sum_{i=k+1}^{N-k}\frac{h_i^1 f_0(\underline{0}, [\beta_0]_{i-k}^{i+k})}{N-2k}\right| \geq \frac{\epsilon n}{3(N-2k)}\right) + P\left(\left|\sum_{i=1}^{k}\frac{h_i^1\beta_{0,i}}{k}\right| \geq \frac{\epsilon n}{3k}\right) + P\left(\left|\sum_{i=N-k+1}^{N}\frac{h_i^1\beta_{0,i}}{k}\right| \geq \frac{\epsilon n}{3k}\right) \\
&\overset{(a)}{\leq} K\exp\left\{-\kappa n\epsilon^2\right\} + K\exp\left\{-\kappa n\epsilon^2\right\} + K\exp\left\{-\kappa n\epsilon^2\right\}.
\end{aligned}$$

In the above, step (a) follows from $\mathcal{H}_1(b)(i)$ using PL(2) functions $\phi_{1,h}, \phi_{2,h}, \phi_{3,h} : \mathbb{R}^{2(2k+1)} \to \mathbb{R}$ equal to $\phi_{1,h}(\underline{x}, \underline{y}) = \underline{x}_i f_0(\underline{0}, \underline{y})$, $\phi_{2,h}(\underline{x}, \underline{y}) = \underline{x}_1\underline{y}_1$, $\phi_{3,h}(\underline{x}, \underline{y}) = \underline{x}_{2k+1}\underline{y}_{2k+1}$ which are all PL(2) since products of Lipschitz functions are PL(2) by Lemma C.1. Note that $\mathbb{E}[\tau_0\tilde{\underline{Z}}_{0,k+1}f(\underline{0}, \underline{\beta})] = 0$.

(d) The result follows as in $\mathcal{H}_1(c)$. We can write $\left\|h^1\right\|^2 = |\sum_{i=1}^N (h_i^1)^2| \leq |\sum_{i=1}^{N/2}(h_i^1)^2| + |\sum_{j=N/2+1}^N (h_j^1)^2|$ and therefore it follows by Lemma A.1,

$$\begin{aligned}
P\left(\left|\frac{\left\|h^1\right\|^2}{N} - \tau_0^2\right| \geq \epsilon\right) &\leq P\left(\left|\sum_{i=1}^{N/2}\frac{(h_i^1)^2}{N/2} - \tau_0^2\right| \geq \epsilon\right) + P\left(\left|\sum_{j=N/2+1}^{N}\frac{(h_j^1)^2}{N/2} - \tau_0^2\right| \geq \epsilon\right) \\
&\overset{(a)}{\leq} Ke^{-\frac{\kappa N\epsilon^2}{4}} + Ke^{-\frac{\kappa N\epsilon^2}{4}}.
\end{aligned}$$

In the above, step (a) follows by applying $\mathcal{H}_1(b)(i)$ using PL(2) functions $\phi_{1,h}, \phi_{2,h}$ both defined from $\mathbb{R}^{2(2+1)} \to \mathbb{R}$ equal to $\phi_{1,h}(\underline{x}, \underline{y}) = (\underline{x}_1)^2$, and $\phi_{2,h}(\underline{x}, \underline{y}) = (\underline{x}_{2k+1})^2$.

(e) We prove concentration for $(q^0)^* q^1$ first. Notice that

$$(q^0)^* q^1 = \sum_{i=1}^{N} q_i^0 q_i^1 = \sum_{i=k+1}^{N-k} f_0(\underline{0}, [\beta_0]_{i-k}^{i+k}) f_1([h^1]_{i-k}^{i+k}, [\beta_0]_{i-k}^{i+k}) + \sum_{i=1}^{k} \beta_{0,i}^2 + \sum_{i=N-k+1}^{N} \beta_{0,i}^2.$$

Therefore it follows by Lemma A.1,

$$\begin{aligned}
&P\left(\left|\frac{(q^0)^* q^1}{n} - \tilde{E}_{0,1}\right| \geq \epsilon\right) \\
&\leq P\left(\left|\sum_{i=k+1}^{N-k} \frac{f_0(\underline{0}, [\beta_0]_{i-k}^{i+k}) f_1([h^1]_{i-k}^{i+k}, [\beta_0]_{i-k}^{i+k})}{N-2k} - \mathbb{E}[f_0(\underline{0}, \underline{\beta}) f_1(\tau_0 \tilde{\underline{Z}}_0, \underline{\beta})]\right| \geq \frac{\epsilon n}{3(N-2k)}\right) \\
&\quad + P\left(\left|\sum_{i=1}^{k} \frac{\beta_{0,i}^2}{k} - \sigma_\beta^2\right| \geq \frac{\epsilon n}{3k}\right) + P\left(\left|\sum_{i=N-k+1}^{N} \frac{\beta_{0,i}^2}{k} - \sigma_\beta^2\right| \geq \frac{\epsilon n}{3k}\right) \\
&\overset{(a)}{\leq} K \exp\left\{-\kappa n \epsilon^2\right\} + K \exp\left\{-\kappa n \epsilon^2\right\} + K \exp\left\{-\kappa n \epsilon^2\right\}.
\end{aligned}$$

In the above, step (a) follows from $\mathcal{H}_1(b)(i)$ using PL(2) functions $\phi_{1,h}, \phi_{2,h}, \phi_{3,h} : \mathbb{R}^{2(2k+1)} \to \mathbb{R}$ equal to $\phi_{1,h}(\underline{x}, \underline{y}) = f_0(\underline{0}, \underline{y}) f_1(\underline{x}, \underline{y})$, $\phi_{2,h}(\underline{x}, \underline{y}) = \underline{y}_1^2$, $\phi_{3,h}(\underline{x}, \underline{y}) = \underline{y}_{2k+1}^2$ which are all PL(2) since products of Lipschitz functions are PL(2) by Lemma C.1. The result follows by noting $\mathbb{E}[\beta^2] = \sigma_\beta^2$.

Concentration for $\left\|q^1\right\|^2$ follows similarly, by representing

$$\left\|q^1\right\|^2 = \sum_{i=1}^{N} (q_i^1)^2 = \sum_{i=k+1}^{N-k} (f_1([h^1]_{i-k}^{i+k}, [\beta_0]_{i-k}^{i+k}))^2 + \sum_{i=1}^{k} \beta_{0,i}^2 + \sum_{i=N-k+1}^{N} \beta_{0,i}^2,$$

then applying $\mathcal{H}_1(b)(i)$.

(f) The concentration of λ_0 around $\hat{\lambda}_0$ follows from $\mathcal{H}_1(b)(ii)$ applied to the function $\psi_h([h^1]_{i-k}^{i+k}, [\beta_0]_{i-k}^{i+k}) := f_0'([h^1]_{i-k}^{i+k}, [\beta_0]_{i-k}^{i+k})$. The only other result to prove is concentration for $(h^1)^* q^1$. Notice that

$$(h^1)^* q^1 = \sum_{i=1}^{N} h_i^1 q_i^1 = \sum_{i=k+1}^{N-k} h_i^1 f_1([h^1]_{i-k}^{i+k}, [\beta_0]_{i-k}^{i+k}) + \sum_{i=1}^{k} h_i^1 \beta_{0,i} + \sum_{i=N-k+1}^{N} h_i^1 \beta_{0,i}.$$

Therefore it follows by Lemma A.1,

$$\begin{aligned}
P\left(\left|\frac{(h^1)^* q^1}{n} - \hat{\lambda}_1 \breve{E}_{0,0}\right| \geq \epsilon\right) &\leq P\left(\left|\sum_{i=k+1}^{N-k} \frac{h_i^1 f_1([h^1]_{i-k}^{i+k}, [\beta_0]_{i-k}^{i+k})}{N-2k} - \frac{n \hat{\lambda}_1 \breve{E}_{0,0}}{N-2k}\right| \geq \frac{\epsilon n}{3(N-2k)}\right) \\
&\quad + P\left(\left|\sum_{i=1}^{k} \frac{h_i^1 \beta_{0,i}}{k}\right| \geq \frac{\epsilon n}{3k}\right) + P\left(\left|\sum_{i=k+1}^{N-k} \frac{h_i^1 \beta_{0,i}}{k}\right| \geq \frac{\epsilon n}{3k}\right) \\
&\overset{(a)}{\leq} K \exp\left\{-\kappa n \epsilon^2\right\} + K \exp\left\{-\kappa n \epsilon^2\right\} + K \exp\left\{-\kappa n \epsilon^2\right\},
\end{aligned}$$

In the above, step (a) follows from $\mathcal{H}_1(b)(i)$ using PL(2) functions $\phi_{1,h}, \phi_{2,h}, \phi_{3,h} : \mathbb{R}^{2(2k+1)} \to \mathbb{R}$ equal to $\phi_{1,h}(\underline{x}, \underline{y}) = \underline{x}_{k+1} f_1(\underline{x}, \underline{y})$, $\phi_{2,h}(\underline{x}, \underline{y}) = \underline{x}_1 \underline{y}_1$, $\phi_{3,h}(\underline{x}, \underline{y}) = \underline{x}_{2k+1} \underline{y}_{2k+1}$ which are all PL(2) since products of Lipschitz functions are PL(2) by Lemma C.1. The result follows by noting that $\mathbb{E}[\tau_0 \tilde{\underline{Z}}_{k+1} \underline{\beta}_{k+1}] = 0$ and $\mathbb{E}[\tau_0 \tilde{\underline{Z}}_{0,k+1} f_1(\tau_0 \tilde{\underline{Z}}_0, \underline{\beta})] = \left(\frac{n}{N-2k}\right) \hat{\lambda}_1 \breve{E}_{0,0}$, which follows by Stein's Lemma

given in Fact 2. We demonstrate this in the following. Think of a function $\tilde{f}:\mathbb{R}\to\mathbb{R}$ defined as $\tilde{f}(x)=f_1(\tau_0\underline{\tilde{Z}}_{0,1},\ldots,\tau_0\underline{\tilde{Z}}_{0,k},x,\tau_0\underline{\tilde{Z}}_{0,k+1},\ldots,\tau_0\underline{\tilde{Z}}_{0,2k+1},\underline{\beta})$. Then,

$$\mathbb{E}[\tau_0\underline{\tilde{Z}}_{0,k+1}f_1(\tau_0\underline{\tilde{Z}}_0,\underline{\beta})]=\mathbb{E}[\tau_0\underline{\tilde{Z}}_{0,k+1}\tilde{f}(\tau_0\underline{\tilde{Z}}_{0,k+1})]\overset{(b)}{=}\tau_0^2\mathbb{E}[f_1'(\tau_0\underline{\tilde{Z}}_0,\underline{\beta})]=\breve{E}_{0,0}\left(\frac{n}{N-2k}\right)\hat{\lambda}_1.$$

In the above, step (b) follows by Fact 2

(g), (h) The proof of $\mathcal{H}_1(g),(h)$ follow as the corresponding proofs in [9].

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

We wish to show results (a) – (h) in (34), (37), (38), (40), (42), (44), (46), (48), (50) assuming that $\mathcal{B}_r$ and $\mathcal{H}_{r+1}$ hold for all $0\le r\le t-1$ due to the inductive hypothesis. The proof of these results is the same as in the step $\mathcal{B}_t$ of the proof in [9] and therefore is not repeated here.

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

We wish to show results (a) – (h) in (33), (35), (36), (39), (41), (43), (45), (47), (49) assuming $\mathcal{B}_r$ holds for all $0\le r\le t$ and $\mathcal{H}_{s+1}$ holds for all $0\le s\le t-1$.

The following two lemmas, whose proof is the same as in [9], will be used to prove $\mathcal{H}_{t+1}$.

Lemma 4. *[9, Lemma 6] The symmetric matrix $\mathbf{Q}_{t+1}:=\frac{1}{n}Q_{t+1}^*Q_{t+1}$ is invertible with high probability. In particular,*

$$P\left(\mathbf{Q}_{t+1}\ \textit{singular}\right)\le Ke^{-\kappa n\epsilon^2} \tag{62}$$

Lemma 5. *[9, Lemma 8] Let $v:=\frac{1}{n}B_{t+1}^*m_t^\perp-\frac{1}{n}Q_{t+1}^*(\xi_tq^t-\sum_{i=0}^{t-1}\alpha_i^t\xi_iq^i)$ and $\mathbf{Q}_{t+1}:=\frac{1}{n}Q_{t+1}^*Q_{t+1}$. Then for $j\in[t+1]$,*

$$P\left(\left|\left[\mathbf{Q}_{t+1}^{-1}v\right]_j\right|\ge\epsilon\right)\le e^{-\kappa n\epsilon^2}.$$

(a) Recall the definition of $\Delta_{t+1,t}$ from Lemma 2 (32). Using Fact 1, we have

$$\frac{\left\|m_\perp^t\right\|}{\sqrt{n}}\mathsf{P}^{\parallel}_{Q_{t+1}}Z_t\overset{d}{=}\frac{\left\|m_\perp^t\right\|}{\sqrt{n}}\frac{1}{\sqrt{N}}\tilde{Q}_{t+1}\bar{Z}_{t+1},$$

where matrix $\tilde{Q}_{t+1}\in\mathbb{R}^{N\times(t+1)}$ forms an orthogonal basis for the column space of Q_{t+1} such that $\tilde{Q}_{t+1}^*\tilde{Q}_{t+1}=N\mathsf{I}_{t+1}$ and $\bar{Z}_{t+1}\in\mathbb{R}^{t+1}$ is an independent random vector with i.i.d. $\mathcal{N}(0,1)$ entries. We can then write

$$\Delta_{t+1,t}\overset{d}{=}\sum_{r=0}^{t-1}(\alpha_r^t-\hat{\alpha}_r^t)h^{r+1}+Z_t\left(\frac{\left\|m_\perp^t\right\|}{\sqrt{n}}-\tau_t^\perp\right)-\frac{\left\|m_\perp^t\right\|}{\sqrt{n}}\frac{\tilde{Q}_{t+1}\bar{Z}_{t+1}}{\sqrt{N}}+Q_{t+1}\mathbf{Q}_{t+1}^{-1}v,$$

where $\mathbf{Q}_{t+1}\in\mathbb{R}^{(t+1)\times(t+1)}$ and $v\in\mathbb{R}^{t+1}$ are defined in Lemma 5. By Lemma C.3,

$$\begin{aligned}\left\|\Delta_{t+1,t}\right\|^2\le(2t+3)\Bigg[&\sum_{r=0}^{t-1}(\alpha_r^t-\hat{\alpha}_r^t)^2\left\|h^{r+1}\right\|^2+\left\|Z_t\right\|^2\left(\frac{\left\|m_\perp^t\right\|}{\sqrt{n}}-\tau_t^\perp\right)^2\\&+\frac{\left\|m_\perp^t\right\|^2}{n}\frac{\left\|\tilde{Q}_{t+1}\bar{Z}_{t+1}\right\|^2}{N}+\sum_{j=0}^{t}\left\|q^j\right\|^2\left[\mathbf{Q}_{t+1}^{-1}v\right]_{j+1}^2\Bigg],\end{aligned}$$

where we have used $Q_{t+1}\mathbf{Q}_{t+1}^{-1}v = \sum_{j=0}^{t} q^j \left[\mathbf{Q}_{t+1}^{-1}v\right]_{j+1}$. Applying Lemma A.1,

$$P\left(\frac{\|\Delta_{t+1,t}\|^2}{N} \geq \epsilon\right) \leq \sum_{r=0}^{t-1} P\left(\left|\alpha_r^t - \hat{\alpha}_r^t\right| \frac{\left\|h^{r+1}\right\|}{\sqrt{N}} \geq \sqrt{\epsilon_t}\right) + P\left(\left|\frac{\left\|m_\perp^t\right\|}{\sqrt{n}} - \tau_t^\perp\right| \frac{\|Z_t\|}{\sqrt{N}} \geq \sqrt{\epsilon_t}\right)$$
$$+ P\left(\frac{\left\|m_\perp^t\right\|}{\sqrt{n}} \frac{\left\|\tilde{Q}_{t+1}\bar{Z}_{t+1}\right\|}{N} \geq \sqrt{\epsilon_t}\right) + \sum_{j=0}^{t} P\left(\left|\left[\mathbf{Q}_{t+1}^{-1}v\right]_{j+1}\right| \frac{\left\|q^j\right\|}{\sqrt{n}} \geq \sqrt{\epsilon_t}\right), \tag{63}$$

where $\epsilon_t = \frac{\epsilon}{(2t+3)^2}$. We now show each of the terms in (63) has the desired upper bound. For $0 \leq r \leq t$,

$$P\left(\left|\alpha_r^t - \hat{\alpha}_r^t\right| \frac{\left\|h^{r+1}\right\|}{\sqrt{N}} \geq \sqrt{\epsilon_t}\right) \leq P\left(\left|\alpha_r^t - \hat{\alpha}_r^t\right| \left(\left|\frac{\left\|h^{r+1}\right\|}{\sqrt{N}} - \tau_r\right| + \tau_r\right) \geq \sqrt{\epsilon_t}\right)$$
$$\leq P\left(\left|\alpha_r^t - \hat{\alpha}_r^t\right| \geq \frac{\sqrt{\epsilon_t}}{2}\min\{1, \tau_r^{-1}\}\right) + P\left(\left|\frac{\left\|h^{r+1}\right\|}{\sqrt{N}} - \tau_r\right| \geq \sqrt{\epsilon}\right) \overset{(a)}{\leq} Ke^{-\kappa N\epsilon} + Ke^{-\kappa N\epsilon},$$

where step (a) follows from induction hypotheses $\mathcal{B}_t(g)$, $\mathcal{H}_1(d) - \mathcal{H}_t(d)$, and Lemma A.3. Next, the second term in (63) is bounded as

$$P\left(\left|\frac{\left\|m_\perp^t\right\|}{\sqrt{n}} - \tau_t^\perp\right| \frac{\|Z_t\|}{\sqrt{N}} \geq \sqrt{\tilde{\epsilon}_t}\right) \leq P\left(\left|\frac{\left\|m_\perp^t\right\|}{\sqrt{n}} - \tau_t^\perp\right| \geq \epsilon_t^{\frac{1}{4}}\right) + P\left(\frac{\|Z_t\|}{\sqrt{N}} \geq \epsilon_t^{\frac{1}{4}}\right) \overset{(b)}{\leq} Ke^{-\kappa n\sqrt{\epsilon}} + Ke^{-\kappa N\sqrt{\epsilon}},$$

where step (b) is obtained using induction hypothesis $\mathcal{B}_t(h)$, Lemma B.2, and Lemma A.3. Since $\frac{\left\|m_\perp^t\right\|}{\sqrt{n}}$ concentrates on $\tau_t^\perp$ by $\mathcal{B}_t(h)$, the third term in (63) can be bounded as

$$\begin{aligned} P\left(\frac{\left\|m_\perp^t\right\|}{\sqrt{n}} \cdot \frac{\left\|\tilde{Q}_{t+1}\bar{Z}_{t+1}\right\|}{N} \geq \sqrt{\epsilon_t}\right) &= P\left(\left(\left|\frac{\left\|m_\perp^t\right\|}{\sqrt{n}} - \tau_t^\perp\right| + \tau_t^\perp\right) \cdot \frac{\|\tilde{Q}_{t+1}\bar{Z}_{t+1}\|}{N} \geq \sqrt{\epsilon_t}\right) \\ &\leq P\left(\left|\frac{\left\|m_\perp^t\right\|}{\sqrt{n}} - \tau_t^\perp\right| \geq \sqrt{\epsilon}\right) + P\left(\frac{1}{N}\left\|\tilde{Q}_{t+1}\bar{Z}_{t+1}\right\| \geq \frac{\sqrt{\epsilon_t}}{2}\min\{1, (\tau_t^\perp)^{-1}\}\right). \end{aligned} \tag{64}$$

For the second term in (64), first bound the norm of $\tilde{Q}_{t+1}\bar{Z}_{t+1}$ as follows.

$$\left\|\tilde{Q}_{t+1}\bar{Z}_{t+1}\right\|^2 = \left\|\sum_{i=0}^{t} \tilde{q}_i \bar{Z}_{t+1,i}\right\|^2 \overset{(c)}{\leq} (t+1)\sum_{i=0}^{t} \|\tilde{q}_i\|^2 (\bar{Z}_{t+1,i})^2 \overset{(d)}{=} N(t+1)\sum_{i=0}^{t} (\bar{Z}_{t+1,i})^2,$$

where step (c) follows from Lemma C.3 and step (d) uses the fact that $\|\tilde{q}_i\|^2 = N$ for all $0 \leq i \leq t$. Therefore,

$$P\left(\frac{1}{N^2}\left\|\tilde{Q}_{t+1}\bar{Z}_{t+1}\right\|^2 \geq \epsilon'\right) \overset{(e)}{\leq} \sum_{i=0}^{t} P\left(|\bar{Z}_{t+1,i}| \geq \sqrt{\frac{N\epsilon'}{(t+1)^2}}\right) \overset{(f)}{\leq} 2e^{-\frac{1}{2(t+1)^2}N\epsilon'}. \tag{65}$$

Step (e) is obtained from Lemma A.1 and step (f) from Lemma B.1. Using (65), the RHS of (64) is bounded by $K\exp\{-\kappa n\epsilon\}$. Finally, for $0 \leq j \leq t$, the last term in (63) can be bounded by

$$P\left(\left|\left[\mathbf{Q}_{t+1}^{-1}v\right]_{j+1}\right| \frac{\left\|q^j\right\|}{\sqrt{n}} \geq \sqrt{\epsilon_t}\right) = P\left(\left|\left[\mathbf{Q}_{t+1}^{-1}v\right]_{j+1}\right| \left(\left|\frac{\left\|q^j\right\|}{\sqrt{n}} - \sigma_j\right| + \sigma_j\right) \geq \sqrt{\epsilon_t}\right)$$
$$\leq P\left(\left|\frac{\left\|q^j\right\|}{\sqrt{n}} - \sigma_j\right| \geq \sqrt{\epsilon}\right) + P\left(\left|\left[\mathbf{Q}_{t+1}^{-1}v\right]_{j+1}\right| \geq \frac{\sqrt{\epsilon_t}}{2}\min\{1, \sigma_j^{-1}\}\right) \overset{(g)}{\leq} Ke^{-\kappa n\epsilon^2} + Ke^{-\kappa n\epsilon^2},$$

where step (g) follows from Lemma 5, the induction hypothesis $\mathcal{H}_t(e)$, and Lemma A.3. Thus we have bounded each term of (63) as desired.

(b)(i) For brevity we define the shorthand notation $\mathbb{E}_{\phi_h} := \mathbb{E}\left[\phi_h(\tau_0\underline{\tilde{Z}}_0, ..., \tau_t\underline{\tilde{Z}}_t, \underline{\beta})\right]$, and

$$a_i = \left(h_i^1, ..., h_i^t, \sum_{r=0}^{t-1}\hat{\alpha}_r^t h_i^{r+1} + \tau_t^{\perp} Z_{t_i} + [\Delta_{t+1,t}]_i, \beta_{0_i}\right), \quad c_i = \left(h_i^1, ..., h_i^t, \sum_{r=0}^{t-1}\hat{\alpha}_r^t h_i^{r+1} + \tau_t^{\perp} Z_{t_i}, \beta_{0_i}\right), \tag{66}$$

for $i = 1, ..., N$. Hence a, c are length-N vectors with entries a_i, $c_i \in \mathbb{R}^{(t+2)}$.

Then, using the conditional distribution of h^{t+1} from Lemma 2 and Lemma A.1, we have

$$\begin{aligned}
&P\left(\left|\frac{1}{N-2k}\sum_{i=k+1}^{N-k}\phi_h([h^1]_{i-k}^{i+k}, ..., [h^{t+1}]_{i-k}^{i+k}, [\beta_0]_{i-k}^{i+k}) - \mathbb{E}_{\phi_h}\right| \geq \epsilon\right) \\
&\leq P\left(\left|\frac{1}{N-2k}\sum_{k+1}^{N-k}\left(\phi_h([a]_{i-k}^{i+k}) - \phi_h([c]_{i-k}^{i+k})\right)\right| \geq \frac{\epsilon}{2}\right) + P\left(\left|\frac{1}{N-2k}\sum_{i=k+1}^{N-k}\phi_h([c]_{i-k}^{i+k}) - \mathbb{E}_{\phi_h}\right| \geq \frac{\epsilon}{2}\right).
\end{aligned} \tag{67}$$

Label the two terms of (67) as T_1 and T_2. To complete the proof we show both are bounded by $Ke^{-\kappa n\epsilon^2}$. First consider term T_1. Using the pseudo-Lipschitz property of ϕ_h, we have

$$\begin{aligned}
T_1 &\leq P\left(\frac{1}{N-2k}\sum_{i=k+1}^{N-k} L(1 + \|[a]_{i-k}^{i+k}\| + \|[c]_{i-k}^{i+k}\|)\|[a-c]_{i-k}^{i+k}\| \geq \frac{\epsilon}{2}\right) \\
&\overset{(a)}{\leq} P\left(\frac{1}{N-2k}\left(\sum_{i=k+1}^{N-k}\left(1 + \|[a]_{i-k}^{i+k}\| + \|[c]_{i-k}^{i+k}\|\right)^2\right)^{1/2}\left(\sum_{i=k+1}^{N-k}\|[a-c]_{i-k}^{i+k}\|^2\right)^{1/2} \geq \frac{\epsilon}{2L}\right) \\
&\overset{(b)}{\leq} P\left(\frac{1}{N-2k}\left(\sum_{i=k+1}^{N-k}\left(1 + \|[\Delta_{t+1,t}]_{i-k}^{i+k}\|^2 + 4\|[c]_{i-k}^{i+k}\|^2\right)\right)^{1/2}\left(\sum_{i=k+1}^{N-k}\|[\Delta_{t+1,t}]_{i-k}^{i+k}\|^2\right)^{1/2} \geq \frac{\epsilon}{2\sqrt{3}L}\right) \\
&\leq P\left(\left(1 + \sqrt{2k+1}\frac{\|\Delta_{t+1,t}\|}{\sqrt{N-2k}} + 2\sqrt{2k+1}\frac{\|c\|}{\sqrt{N-2k}}\right)\left(\sqrt{2k+1}\frac{\|\Delta_{t+1,t}\|}{\sqrt{N-2k}}\right) \geq \frac{\epsilon}{2\sqrt{3}L}\right).
\end{aligned} \tag{68}$$

Step (a) follows by Cauchy-Schwarz and step (b) uses $\|[a]_{i-k}^{i+k}\| \leq \|[c]_{i-k}^{i+k}\| + \|[\Delta_{t+1,t}]_{i-k}^{i+k}\|$, $\|[a-c]_{i-k}^{i+k}\|^2 = \|[\Delta_{t+1,t}]_{i-k}^{i+k}\|^2$, and Lemma C.3.

From (66) and using Lemma C.3, we have

$$\|c\|^2 \leq \sum_{r=0}^{t-1}\|h^{r+1}\|^2 + 2\sum_{r=0}^{t-1}\sum_{l=0}^{t-1}\hat{\alpha}_r\hat{\alpha}_l(h^{r+1})^* h^{l+1} + 2(\tau_t^{\perp})^2\|Z_t\|^2 + \|\beta_0\|^2 \tag{69}$$

Denote the RHS of above by $\tilde{c}^2$. From the induction hypothesis, $\frac{1}{N}(h^{r+1})^* h^{l+1}$ concentrates on $\breve{E}_{r,l}$ for $0 \leq r, l \leq (t-1)$. Using this in (69), we will argue that $\frac{1}{N}\tilde{c}^2$ concentrates on

$$\mathbb{E}_{\tilde{c}} := \sum_{l=0}^{t-1}\breve{E}_{l,l} + 2\sum_{r=0}^{t-1}\sum_{l=0}^{t-1}\hat{\alpha}_r^t\hat{\alpha}_l^t\breve{E}_{l,r} + 2(\tau_t^{\perp})^2 + \sigma_\beta^2 = \sum_{l=0}^{t-1}\tau_l^2 + 2\tau_t^2 + \sigma_\beta^2, \tag{70}$$

where the last equality is obtained using $\breve{E}_{l,l} = \tau_l^2$, and by rewriting the double sum as follows:

$$\sum_{r=0}^{t-1}\sum_{l=0}^{t-1}\hat{\alpha}_r^t\hat{\alpha}_l^t\breve{E}_{r,l} = (\hat{\alpha}^t)^*\breve{C}^t\hat{\alpha}^t = [\breve{E}_t^*(\breve{C}^t)^{-1}](\breve{C}^t)^{-1}[(\breve{C}^t)^{-1}\breve{E}_t] = \breve{E}_t^*(\breve{C}^t)^{-1}\breve{E}_t = \breve{E}_{t,t} - (\tau_t^{\perp})^2. \quad (71)$$

Using Lemma A.1, let $\epsilon_t = \epsilon/(t+t^2+2)$,

$$P\left(\left|\frac{\tilde{c}^2}{N} - \mathbb{E}_{\tilde{c}}\right| \geq \epsilon\right) \leq \sum_{l=0}^{t-1} P\left(\left|\frac{\|h^{l+1}\|^2}{N} - \tau_l^2\right| \leq \epsilon_t\right) + P\left(\left|\frac{\|\beta_0\|^2}{N} - \sigma_\beta^2\right| \geq \epsilon_t\right)$$
$$+ \sum_{r=0}^{t-1}\sum_{l=0}^{t-1} P\left(\left|\frac{(h^{r+1})^* h^{l+1}}{N} - \breve{E}_{r,l}\right| \geq \frac{\epsilon_t}{2\hat{\alpha}_r^t\hat{\alpha}_l^t}\right) + P\left(\left|\frac{\|Z_t\|^2}{N} - 1\right| \geq \frac{\epsilon_t}{2(\tau_t^{\perp})^2}\right) \overset{(a)}{\leq} Ke^{-\kappa N\epsilon^2}. \quad (72)$$

In step (a), we used induction hypothesis $\mathcal{H}_1(d) - \mathcal{H}_t(d)$, our assumptions on β_0, and Lemma B.2.

Therefore, using (68), term T_1 of (67) can be bounded as

$$\begin{aligned} T_1 &\leq P\left(\left(1 + \frac{\|\Delta_{t+1,t}\|}{\sqrt{N}} + 2\frac{\tilde{c}}{\sqrt{N}}\right)\cdot\frac{\|\Delta_{t+1,t}\|}{\sqrt{N}} \geq \frac{\epsilon(1-2k/N)}{2\sqrt{3}(2k+1)L}\right) \\ &= P\left(\left(1 + \frac{\|\Delta_{t+1,t}\|}{\sqrt{N}} + 2\left(\frac{\tilde{c}}{\sqrt{N}} - \mathbb{E}_{\tilde{c}}^{1/2}\right) + 2\mathbb{E}_{\tilde{c}}^{1/2}\right)\cdot\frac{\|\Delta_{t+1,t}\|}{\sqrt{N}} \geq \frac{\epsilon(1-2k/N)}{2\sqrt{3}(2k+1)L}\right) \\ &\leq P\left(\left|\frac{\tilde{c}}{\sqrt{N}} - \mathbb{E}_{\tilde{c}}^{1/2}\right| \geq \epsilon\right) + P\left(\frac{\|\Delta_{t+1,t}\|}{\sqrt{N}} \geq \frac{\epsilon(1-2k/N)}{2\sqrt{3}(2k+1)L(4+2\mathbb{E}_{\tilde{c}}^{1/2})}\right) \overset{(a)}{\leq} Ke^{-\kappa N\epsilon^2}. \end{aligned}$$

In step (a), we used (72), $\mathcal{H}_{t+1}(a)$, and Lemma A.3.

Next consider term T_2 of (67). Define function $\tilde{\phi}_{h_i} : \mathbb{R}^{2k+1} \to \mathbb{R}$ as

$$\tilde{\phi}_{h_i}(\underline{z}) := \phi_h([h^1]_{i-k}^{i+k}, ..., [h^t]_{i-k}^{i+k}, \sum_{r=0}^{t-1}\hat{\alpha}_r^t[h^{r+1}]_{i-k}^{i+k} + \tau_t^{\perp}\underline{z}, [\beta_0]_{i-k}^{i+k}) \in PL(2), \quad (73)$$

for each $i = k+1, ..., N-k$, where we treat all arguments except $\underline{z}$ as fixed. Let $\underline{Z} \in \mathbb{R}^{2k+1}$ be a random vector of i.i.d. $\mathcal{N}(0,1)$ entries, and assume that $\underline{Z}$ is independent of $\underline{\tilde{Z}}_0, ..., \underline{\tilde{Z}}_{t-1}$, then

$$\begin{aligned} T_2 &= P\left(\left|\frac{1}{N-2k}\sum_{i=k+1}^{N-k}\tilde{\phi}_{h_i}([Z_t]_{i-k}^{i+k}) - \mathbb{E}_{\phi_h}\right| \geq \frac{\epsilon}{2}\right) \\ &\leq P\left(\left|\frac{1}{N-2k}\sum_{i=k+1}^{N-k}\left(\tilde{\phi}_{h_i}([Z_t]_{i-k}^{i+k}) - \mathbb{E}_{\underline{Z}}\left[\tilde{\phi}_{h_i}(\underline{Z})\right]\right)\right| \geq \frac{\epsilon}{4}\right) + P\left(\left|\frac{1}{N-2k}\sum_{i=k+1}^{N-k}\mathbb{E}_{\underline{Z}}\left[\tilde{\phi}_{h_i}(\underline{Z})\right] - \mathbb{E}_{\phi_h}\right| \geq \frac{\epsilon}{4}\right). \end{aligned}$$

The first term on the RHS of the above has the desired bound using Lemma D.6. We now bound the second term.

$$\begin{aligned} &P\left(\left|\frac{1}{N-2k}\sum_{i=k+1}^{N-k}\mathbb{E}_{\underline{Z}}\left[\tilde{\phi}_{h_i}(\underline{Z})\right] - \mathbb{E}_{\phi_h}\right| \geq \frac{\epsilon}{4}\right) \\ &= P\left(\left|\frac{1}{N-2k}\sum_{i=k+1}^{N-k}\mathbb{E}_{\underline{Z}}\left[\phi_h\left([h^1]_{i-k}^{i+k}, ..., [h^t]_{i-k}^{i+k}, \sum_{r=0}^{t-1}\hat{\alpha}_r^t[h^{r+1}]_{i-k}^{i+k} + \tau_t^{\perp}\underline{Z}, [\beta_0]_{i-k}^{i+k}\right)\right] - \mathbb{E}_{\phi_h}\right| \geq \frac{\epsilon}{4}\right) \\ &\overset{(a)}{=} P\left(\left|\frac{1}{N-2k}\sum_{i=k+1}^{N-k}\phi_h'\left([h^1]_{i-k}^{i+k}, ..., [h^t]_{i-k}^{i+k}, [\beta_0]_{i-k}^{i+k}\right) - \mathbb{E}_{\phi_h}\right| \geq \frac{\epsilon}{4}\right). \end{aligned} \quad (74)$$

Step (a) uses the function $\phi'_h : \mathbb{R}^{(2k+1)(t+1)} \to \mathbb{R}$ defined as

$$\phi'_h\left([h^1]_{i-k}^{i+k}, ..., [h^t]_{i-k}^{i+k}, [\beta_0]_{i-k}^{i+k}\right) := \mathbb{E}_{\underline{Z}}\left[\phi_h\left([h^1]_{i-k}^{i+k}, ..., [h^t]_{i-k}^{i+k}, \sum_{r=0}^{t-1}\hat{\alpha}_r^t[h^{r+1}]_{i-k}^{i+k} + \tau_t^{\perp}\underline{Z}, [\beta_0]_{i-k}^{i+k}\right)\right],$$

which is $PL(2)$ by Lemma C.2. We will now show that

$$\mathbb{E}\left[\phi'_h\left(\tau_0\tilde{\underline{Z}}_0, ..., \tau_{t-1}\tilde{\underline{Z}}_{t-1}, \underline{\beta}\right)\right] = \mathbb{E}\left[\phi_h(\tau_0\tilde{\underline{Z}}_0, ..., \tau_t\tilde{\underline{Z}}_t, \underline{\beta})\right] = \mathbb{E}_{\phi_h}, \tag{75}$$

and then the probability in (74) can be upper bounded by $Ke^{-\kappa n\epsilon^2}$ using the inductive hypothesis $\mathcal{H}_t(b)(i)$. We have

$$\mathbb{E}\left[\phi'_h\left(\tau_0\tilde{\underline{Z}}_0, ..., \tau_{t-1}\tilde{\underline{Z}}_{t-1}, \underline{\beta}\right)\right] = \mathbb{E}\left[\phi_h\left(\tau_0\tilde{\underline{Z}}_0, ..., \tau_{t-1}\tilde{\underline{Z}}_{t-1}, \sum_{r=0}^{t-1}\hat{\alpha}_r^t\tau_r\tilde{\underline{Z}}_r + \tau_t^{\perp}\underline{Z}, \underline{\beta}\right)\right],$$

where we recall that $\underline{Z}$ is independent of $\tilde{\underline{Z}}_0, ..., \tilde{\underline{Z}}_{t-1}$. To prove (75), we need to show that

$$\left(\tau_0\tilde{\underline{Z}}_0, ..., \tau_{t-1}\tilde{\underline{Z}}_{t-1}, \tau_t\tilde{\underline{Z}}_t, \underline{\beta}\right) \overset{d}{=} \left(\tau_0\tilde{\underline{Z}}_0, ..., \tau_{t-1}\tilde{\underline{Z}}_{t-1}, \sum_{r=0}^{t-1}\hat{\alpha}_r^t\tau_r\tilde{\underline{Z}}_r + \tau_t^{\perp}\underline{Z}, \underline{\beta}\right).$$

We do this by demonstrating that: (i) the covariance matrix of $\sum_{r=0}^{t-1}\hat{\alpha}_r^t\tau_r\tilde{\underline{Z}}_r + \tau_t^{\perp}\underline{Z}$ is $\tau_t^2\mathsf{I}$; and (ii) the covariance $\text{Cov}\left(\tau_l\tilde{\underline{Z}}_l, \sum_{r=0}^{t-1}\hat{\alpha}_r^t\tau_r\tilde{\underline{Z}}_r + \tau_t^{\perp}\underline{Z}\right) = \text{Cov}\left(\tau_l\tilde{\underline{Z}}_l, \tau_t\tilde{\underline{Z}}_t\right) = \breve{E}_{l,t}\mathsf{I}$, for $0 \le l \le (t-1)$. First consider (i). The $(i,j)^{th}$ entry of the covariance matrix is

$$\mathbb{E}\left[\left(\sum_{r=0}^{t-1}\hat{\alpha}_r^t\tau_r\tilde{\underline{Z}}_{r,i} + \tau_t^{\perp}\underline{Z}_i\right)\left(\sum_{r=0}^{t-1}\hat{\alpha}_r^t\tau_r\tilde{\underline{Z}}_{r,j} + \tau_t^{\perp}\underline{Z}_j\right)\right] = \sum_{r=0}^{t-1}\sum_{l=0}^{t-1}\hat{\alpha}_r^t\hat{\alpha}_l^t\tau_r\tau_l\mathbb{E}\left[\tilde{\underline{Z}}_{r,i}\tilde{\underline{Z}}_{l,j}\right] + (\tau_t^{\perp})^2\mathbb{E}\left[\underline{Z}_i\underline{Z}_j\right]$$
$$\overset{(a)}{=} \begin{cases}\sum_{r=0}^{t-1}\sum_{l=0}^{t-1}\hat{\alpha}_r^t\hat{\alpha}_l^t\breve{E}_{r,l} + (\tau_t^{\perp})^2 \overset{(b)}{=} \tau_t^2, & \text{if } i = j\\ 0, \quad \text{otherwise}\end{cases},$$

where step (a) follows from (21) and step (b) follows from (71). Therefore, we have showed that the covariance matrix is $\tau_t^2\mathsf{I}$. Next consider (ii), for any $0 \le l \le (t-1)$, the $(i,j)^{th}$ entry of the covariance matrix is

$$\mathbb{E}\left[\tau_l\tilde{\underline{Z}}_{l,i}\sum_{r=0}^{t-1}\hat{\alpha}_r^t\tau_r\tilde{\underline{Z}}_{r,j} + \tau_t^{\perp}\underline{Z}_j\right] = \sum_{r=0}^{t-1}\hat{\alpha}_r^t\tau_l\tau_r\mathbb{E}\left[\tilde{\underline{Z}}_{l,i}\tilde{\underline{Z}}_{r,j}\right] \overset{(a)}{=} \begin{cases}\sum_{r=0}^{t-1}\breve{E}_{l,r}\hat{\alpha}_r^t, & \text{if } i = j\\ 0, \quad \text{otherwise}\end{cases},$$

where step (a) follows from (21). Moreover, notice that $\sum_{r=0}^{t-1}\breve{E}_{l,r}\hat{\alpha}_r^t = [\breve{C}^t\hat{\alpha}^t]_{l+1} = \breve{E}_{l,t}$, where the first equality holds because the required sum is the inner product of the $(l+1)^{th}$ row of $\breve{C}^t$ and $\hat{\alpha}^t$, and the second inequality follows the definition of $\hat{\alpha}^t$ in (24).

(b)(ii) Let $\mathsf{h}_r := \sum_{l=0}^{r-1}\hat{\alpha}_l^r h^{l+1}$ for $1 \le r \le t$, and $\underline{Z} \in \mathbb{R}^{2k+1}$ be a vector of i.i.d. $\mathcal{N}(0,1)$ random variables that is indepedent of $\tilde{\underline{Z}}_r, \forall r = 0, ..., t-1$. Then using the conditional distribution of h^{t+1}

in (27) of Lemma 2, we have

$$
\begin{aligned}
&P\left(\left|\frac{1}{N-2k}\sum_{i=k+1}^{N-k}\psi_h([h^{t+1}]_{i-k}^{i+k},[\beta_0]_{i-k}^{i+k})-\mathbb{E}[\psi_h(\tau_t\tilde{\underline{Z}}_t,\underline{\beta})]\right|\geq\epsilon\right)\\
&=P\left(\left|\frac{1}{N-2k}\sum_{i=k+1}^{N-k}\psi_h([\mathsf{h}_t+\tau_t^{\perp}Z_t+\Delta_{t+1,t}]_{i-k}^{i+k},[\beta_0]_{i-k}^{i+k})-\mathbb{E}[\psi_h(\tau_t\tilde{\underline{Z}}_t,\underline{\beta})]\right|\geq\epsilon\right)\\
&\overset{(a)}{\leq}P\left(\left|\frac{1}{N-2k}\sum_{i=k+1}^{N-k}\left(\psi_h([\mathsf{h}_t+\tau_t^{\perp}Z_t+\Delta_{t+1,t}]_{i-k}^{i+k},[\beta_0]_{i-k}^{i+k})-\psi_h([\mathsf{h}_t+\tau_t^{\perp}Z_t]_{i-k}^{i+k},[\beta_0]_{i-k}^{i+k})\right)\right|\geq\frac{\epsilon}{3}\right)\\
&+P\left(\left|\frac{1}{N-2k}\sum_{i=k+1}^{N-k}\left(\psi_h([\mathsf{h}_t+\tau_t^{\perp}Z_t]_{i-k}^{i+k},[\beta_0]_{i-k}^{i+k})-\mathbb{E}_{\underline{Z}}[\psi_h([\mathsf{h}_t]_{i-k}^{i+k}+\tau_t^{\perp}\underline{Z},[\beta_0]_{i-k}^{i+k})]\right)\right|\geq\frac{\epsilon}{3}\right) \quad (76)\\
&+P\left(\left|\frac{1}{N-2k}\sum_{i=k+1}^{N-k}\mathbb{E}_{\underline{Z}}[\psi_h([\mathsf{h}_t]_{i-k}^{i+k}+\tau_t^{\perp}\underline{Z},[\beta_0]_{i-k}^{i+k})]-\mathbb{E}[\psi_h(\tau_t\tilde{\underline{Z}}_t,\underline{\beta})]\right|\geq\frac{\epsilon}{3}\right).
\end{aligned}
$$

Step (a) follows from Lemma A.1. Label the terms on the RHS as T_1-T_3.

First consider T_2. Treat $\{h^1,...,h^t,\beta_0\}$ as constants, hence $\mathsf{h}_t=\sum_{r=0}^{t-1}\hat{\alpha}_r^t h^{r+1}$ is a constant. Since ψ_h is bounded, we apply Lemma D.2 to obtain $T_2\leq Ke^{-\kappa n\epsilon^2}$.

To bound T_3, we first note that the function $\tilde{\psi}_h(\underline{x},\underline{y}):\mathbb{R}^{2(2k+1)}\to\mathbb{R}$ defined as $\tilde{\psi}_h(\underline{x},\underline{y}):=\mathbb{E}_{\underline{Z}}\left[\psi_h(\underline{x}+\underline{Z},\underline{y})\right]$, $\underline{Z}\sim\mathcal{N}(0,\mathsf{I}_{2k+1})$, is bounded and differentiable in the first argument with bounded derivative, since the Gaussian density function is smooth. Hence, using the induction hypothesis $\mathcal{H}_1(b)(ii)-\mathcal{H}_t(b)(ii)$, we sequentially obtain the following concentration inequalities:

$$
\begin{aligned}
&\frac{1}{N-2k}\sum_{i=k+1}^{N-k}\mathbb{E}[\psi_h([\mathsf{h}_{t-1}]_{i-k}^{i+k}+\tau_t^{\perp}\tilde{\underline{Z}}+\hat{\alpha}_{t-1}^t[h^t]_{i-k}^{i+k},[\beta_0]_{i-k}^{i+k})]\\
&\qquad\doteq\frac{1}{N-2k}\sum_{i=k+1}^{N-k}\mathbb{E}[\psi_h([\mathsf{h}_{t-1}]_{i-k}^{i+k}+\tau_t^{\perp}\tilde{\underline{Z}}+\hat{\alpha}_{t-1}^t\tau_{t-1}\tilde{\underline{Z}}_{t-1},\underline{\beta})]\\
&\frac{1}{N-2k}\sum_{i=k+1}^{N-k}\mathbb{E}[\psi_h([\mathsf{h}_{t-2}]_{i-k}^{i+k}+\tau_t^{\perp}\tilde{\underline{Z}}+\hat{\alpha}_{t-1}^t\tau_{t-1}\tilde{\underline{Z}}_{t-1}+\hat{\alpha}_{t-2}^t[h^{t-1}]_{i-k}^{i+k},\underline{\beta})]\\
&\qquad\doteq\frac{1}{N-2k}\sum_{i=k+1}^{N-k}\mathbb{E}[\psi_h([\mathsf{h}_{t-2}]_{i-k}^{i+k}+\tau_t^{\perp}\tilde{\underline{Z}}+\hat{\alpha}_{t-1}^t\tau_{t-1}\tilde{\underline{Z}}_{t-1}+\hat{\alpha}_{t-2}^t\tau_{t-2}\tilde{\underline{Z}}_{t-2},\underline{\beta})]\\
&\vdots\\
&\frac{1}{N-2k}\sum_{i=k+1}^{N-k}\mathbb{E}[\psi_h([\hat{\alpha}_0^t h^1]_{i-k}^{i+k}+\tau_t^{\perp}\tilde{\underline{Z}}+\sum_{r=1}^{t-1}\hat{\alpha}_r^t\tau_r\tilde{\underline{Z}}_r,\underline{\beta})]\doteq\mathbb{E}[\psi_h(\sum_{r=0}^{t-1}\hat{\alpha}_r^t\tau_r\tilde{\underline{Z}}_r+\tau_t^{\perp}\tilde{\underline{Z}},\underline{\beta})]
\end{aligned}
$$

where the expectation in each term is over the random variables denoted in upper case. Recall from the proof of $(b)(i)$ above that $\sum_{r=0}^{t-1}\hat{\alpha}_r^t\tau_r\tilde{\underline{Z}}_r+\tau_t^{\perp}\underline{Z}\overset{d}{=}\tau_t\tilde{\underline{Z}}_t$. Therefore, we have showed that $T_3\leq Ke^{-\kappa n\epsilon^2}$.

Next consider T_1. Let $u\in\mathbb{R}^N$ be such that $u_i=\sum_{r=0}^{t-1}(\alpha_r^t-\hat{\alpha}_r^t)h_i^{r+1}+\sum_{r=0}^{t}q_i^r[\mathbf{Q}_{t+1}^{-1}v]_r$, for all $i\in[N]$, where $\mathbf{Q}_{t+1}$ and v defined as in Lemma 5. Then from the definition of $\Delta_{t+1,t}$ in Lemma 2,

$$
\mathsf{h}_t+\tau_t^{\perp}Z_t+\Delta_{t+1,t}=\mathsf{h}_t+\frac{\|m_{\perp}^t\|}{\sqrt{n}}\left(\mathsf{I}_N-\mathsf{P}_{Q_{t+1}}^{||}\right)Z_t+u.
$$

Using Lemma C.3, we have $\|u\|^2 \leq (2t+1)\left(\sum_{r=0}^{t-1}(\alpha_r^t - \hat{\alpha}_r^t)^2\|h^{r+1}\|^2 + \sum_{r=0}^{t}\|q^r\|^2([\mathbf{Q}_{t+1}^{-1}v]_r)^2\right)$. Therefore, by Lemma A.1,

$$P\left(\frac{\|u\|^2}{n} \geq \epsilon\right) \leq \sum_{r=0}^{t-1} P\left(\left|\alpha_r^t - \hat{\alpha}_r^t\right| \frac{\|h^{r+1}\|}{\sqrt{n}} \geq \sqrt{\epsilon_t}\right) + \sum_{r=0}^{t} P\left(\frac{\|q^r\|}{\sqrt{n}}\left|[\mathbf{Q}_{t+1}^{-1}v]_r\right| \geq \sqrt{\epsilon_t}\right), \tag{77}$$

where $\epsilon_t = \frac{\epsilon}{(2t+1)^2}$. We show that each term on the RHS of the above is upper bounded by $Ke^{-\kappa n\epsilon^2}$.

$$\begin{aligned} P\left(\left|\alpha_r^t - \hat{\alpha}_r^t\right| \frac{\|h^{r+1}\|}{\sqrt{n}} \geq \sqrt{\epsilon_t}\right) &\leq P\left(\left|\alpha_r^t - \hat{\alpha}_r^t\right|\left(\left|\frac{\|h^{r+1}\|}{\sqrt{n}} - \tau_r\right| + \tau_r\right) \geq \sqrt{\epsilon_t}\right) \\ &\leq P\left(\left|\alpha_r^t - \hat{\alpha}_r^t\right| \geq \frac{\sqrt{\epsilon_t}}{2}\min\{1, \tau_r^{-1}\}\right) + P\left(\left|\frac{\|h^{r+1}\|}{\sqrt{n}} - \tau_r\right| \geq \sqrt{\epsilon}\right) \overset{(a)}{\leq} Ke^{-\kappa n\epsilon^2}, \end{aligned} \tag{78}$$

where step (a) follows from induction hypothesis $\mathcal{B}_t(g)$, $\mathcal{H}_1(d) - \mathcal{H}_t(d)$, and Lemma A.3. Next consider the summands in the second summation on the RHS of (77).

$$\begin{aligned} P\left(\frac{\|q^r\|}{\sqrt{n}}\left|[\mathbf{Q}_{t+1}^{-1}v]_r\right| \geq \sqrt{\epsilon_t}\right) &\leq P\left(\left(\left|\frac{\|q^r\|}{\sqrt{n}} - \sigma_r\right| + \sigma_r\right)\left|[\mathbf{Q}_{t+1}^{-1}v]_r\right| \geq \sqrt{\epsilon_t}\right) \\ &\leq P\left(\left|\frac{\|q^r\|}{\sqrt{n}} - \sigma_r\right| \geq \sqrt{\epsilon}\right) + P\left(\left|[\mathbf{Q}_{t+1}^{-1}v]_r\right| \geq \frac{\sqrt{\epsilon_t}}{2}\min\{1, \sigma_r^{-1}\}\right) \overset{(a)}{\leq} Ke^{-\kappa n\epsilon^2}, \end{aligned} \tag{79}$$

where step (a) follows from Lemma 5, induction hypothesis $\mathcal{H}_1(e) - \mathcal{H}_t(e)$, our assumption on q^0, and Lemma A.3. Therefore, combining (77), (78), (79), and using Lemma A.3, we have

$$P\left(\frac{1}{\sqrt{n}}\|u\| \geq \epsilon\right) \leq Ke^{-\kappa n\epsilon^2}. \tag{80}$$

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

$$\mathcal{F} := \left\{\left|\frac{1}{\sqrt{n}}\left\|m_\perp^t\right\| - \tau_t^\perp\right| \geq \epsilon_0\right\} \cup \left\{\frac{1}{\sqrt{n}}\|u\| \geq \epsilon_0\right\} \cup_{r=0}^{t-1}\left\{\left|\frac{1}{\sqrt{n}}\left\|h^{r+1}\right\| - \tau_r\right| \geq \epsilon_0\right\}. \tag{81}$$

Denoting the event we are considering in T_1 by Π_{t+1} and following steps analogous to (58) in $\mathcal{H}_1(b)(ii)$, we obtain

$$T_1 = P(\Pi_{t+1}) \leq P(\mathcal{F}) + \mathbb{E}_{\cdot|\mathcal{F}^c}\left[P(\Pi_{t+1}|\mathcal{F}^c, \mathscr{S}_{t+1,t})\right] \overset{(a)}{\leq} Ke^{-\kappa n\epsilon_0^2} + \mathbb{E}_{\cdot|\mathcal{F}^c}\left[P(\Pi_{t+1}|\mathcal{F}^c, \mathscr{S}_{t+1,t})\right],$$

where step (a) follows from (80), Lemma A.3, and induction hypothesis $\mathcal{B}_t(h)$, $\mathcal{H}_0(d) - \mathcal{H}_t(d)$. We now bound $P(\Pi_{t+1}|\mathcal{F}^c, \mathscr{S}_{t+1,t})$. Using the definition of u from above, we have

$$\begin{aligned} T_1 &= P\left(\left|\sum_{i=k+1}^{N-k}\frac{\psi_h([\mathsf{h}_t + \frac{\|m_\perp^t\|}{\sqrt{n}}(\mathsf{I} - \mathsf{P}_{Q_{t+1}}^{||})Z_t + u]_{i-k}^{i+k}, [\beta_0]_{i-k}^{i+k}) - \psi_h([\mathsf{h}_t + \tau_t^\perp Z_t]_{i-k}^{i+k}, [\beta_0]_{i-k}^{i+k})}{N-2k}\right| \geq \frac{\epsilon}{3}\right) \\ &\leq P\left(\left|\sum_{i=k+1}^{N-k}\frac{\psi_h([\mathsf{h}_t + \frac{\|m_\perp^t\|}{\sqrt{n}}Z_t + u]_{i-k}^{i+k}, [\beta_0]_{i-k}^{i+k}) - \psi_h([\mathsf{h}_t + \tau_t^\perp Z_t]_{i-k}^{i+k}, [\beta_0]_{i-k}^{i+k})}{N-2k}\right| \geq \frac{\epsilon}{6}\right) \\ &+ P\left(\left|\sum_{i=k+1}^{N-k}\frac{\psi_h([\mathsf{h}_t + \frac{\|m_\perp^t\|}{\sqrt{n}}(\mathsf{I} - \mathsf{P}_{Q_{t+1}}^{||})Z_t + u]_{i-k}^{i+k}, [\beta_0]_{i-k}^{i+k}) - \psi_h([\mathsf{h}_t + \frac{\|m_\perp^t\|}{\sqrt{n}}Z_t + u]_{i-k}^{i+k}, [\beta_0]_{i-k}^{i+k})}{N-2k}\right| \geq \frac{\epsilon}{6}\right), \end{aligned}$$

where we have omitted the conditioning to shorten notation. Label the terms on the RHS of the inequality above as $T_{1,a}$ and $T_{1,b}$. To complete the proof, we show both are upper bounded by $Ke^{-\kappa n\epsilon^2}$. First consider $T_{1,b}$.

$$\begin{aligned}
T_{1,b} &\overset{(a)}{\leq} P\left(\frac{C\sqrt{2k+1}}{N-2k}\sum_{i=k+1}^{N-k}\frac{\|m_\perp^t\|}{\sqrt{n}}\|[\mathsf{P}^{||}_{Q_{t+1}}Z_t]_{i-k}^{i+k}\| \geq \frac{\epsilon}{6}\right)\\
&\overset{(b)}{\leq} P\left(\frac{C\sqrt{2k+1}}{N-2k}\sum_{i=k+1}^{N-k}\left|\tau_t^\perp+\epsilon_0\right|\frac{\|[\tilde{Q}_{t+1}\underline{Z}]_{i-k}^{i+k}\|}{\sqrt{N}} \geq \frac{\epsilon}{6}\right)\\
&\overset{(c)}{\leq} P\left(\frac{\|\tilde{Q}_{t+1}\bar{Z}_{t+1}\|}{N} \geq \frac{\epsilon\sqrt{1-2k/N}}{6C(2k+1)\left|\tau_t^\perp+\epsilon_0\right|}\right)\overset{(d)}{\leq} Ke^{-\kappa N\epsilon^2}.
\end{aligned}$$

In the above, step (a) follows Fact 3 for some constant $C>0$. Step (b) holds since we are conditioning on event $\mathcal{F}^c$ defined in (81) and since using Fact 1, it follows $\mathsf{P}^{||}_{Q_{t+1}}Z_t \overset{d}{=} \frac{1}{\sqrt{N}}\tilde{Q}_{t+1}\underline{Z}$, where $\underline{Z}\in\mathbb{R}^{t+1}$ has i.i.d. $\mathcal{N}(0,1)$ entries. Step (c) holds since $\sum_{i=k+1}^{N-k}\|[\tilde{Q}_{t+1}\underline{Z}]_{i-k}^{i+k}\| \leq \sqrt{2k+1}\|\tilde{Q}_{t+1}\underline{Z}\|\sqrt{N-2k}$ by Cauchy-Schwarz, and step (d) follows from (65).

Next we bound $T_{1,a}$. Noting that all quantities except Z_t are in $\mathscr{S}_{t+1,t}$, define the shorthand

$$\mathsf{diff}_i([Z_t]_{i-k}^{i+k}) := \psi_h([\mathsf{h}_t+\frac{\|m_\perp^t\|}{\sqrt{n}}Z_t+u]_{i-k}^{i+k},[\beta_0]_{i-k}^{i+k}) - \psi_h([\mathsf{h}_t+\tau_t^\perp Z_t]_{i-k}^{i+k},[\beta_0]_{i-k}^{i+k}). \tag{82}$$

Then the upper tail of $T_{1,a}$ can be written as

$$P\left(\frac{1}{N-2k}\sum_{i=k+1}^{N-k}(\mathsf{diff}_i([Z_t]_{i-k}^{i+k})-\mathbb{E}[\mathsf{diff}_i([Z_t]_{i-k}^{i+k})]) \geq \frac{\epsilon}{6}-\frac{1}{N-2k}\sum_{i=k+1}^{N-k}\mathbb{E}[\mathsf{diff}_i([Z_t]_{i-k}^{i+k})]\,\middle|\,\mathcal{F}^c,\mathscr{S}_{t+1,t}\right). \tag{83}$$

Using the conditioning on $\mathcal{F}^c$ and steps similar to those in $\mathcal{H}_1(b)(ii)$, we can show that

$$\frac{1}{N-2k}\sum_{i=k+1}^{N-k}\left|\mathbb{E}[\mathsf{diff}_i([Z_t]_{i-k}^{i+k})]\right| \leq \epsilon/12, \quad \text{for } \epsilon_0 \leq \epsilon\left[12C(2k+1)\left(1+\sqrt{\frac{n}{N-2k}}\right)\right]^{-1},$$

where $C>0$ is a constant. For such ϵ_0, we can bound (83) by $Ke^{-\kappa n\epsilon^2}$ using Lemma D.2 since the $\mathsf{diff}_i(\cdot)$ function defined in (82) is bounded (following from the fact that ψ_h is bounded). A bound for the lower tail corresponding to (83) follows similarly, providing the desired bound for $T_{1,a}$.

(c) We first show the concentration of $(h^{t+1})^*\beta_0/n$. Note, $\left|\sum_{i=1}^N h_i^{t+1}\beta_{0,i}\right| \leq \left|\sum_{i=1}^{N/2} h_i^{t+1}\beta_{0,i}\right| + \left|\sum_{i=N/2+1}^N h_i^{t+1}\beta_{0,i}\right|$. Then we have

$$P\left(\left|\frac{(h^{t+1})^*\beta_0}{n}\right| \geq \epsilon\right) \overset{(a)}{\leq} P\left(\left|\sum_{i=1}^{N/2}\frac{h_i^{t+1}\beta_{0,i}}{N/2}\right| \geq \delta\epsilon\right) + P\left(\left|\sum_{i=N/2+1}^{N}\frac{h_i^{t+1}\beta_{0,i}}{N/2}\right| \geq \delta\epsilon\right) \overset{(b)}{\leq} 2Ke^{-\kappa N\delta^2\epsilon^2},$$

where step (a) follows Lemma A.1 and step (b) follows $\mathcal{H}_{t+1}b(i)$ by considering $PL(2)$ functions $\phi_{1,h},\phi_{2,h}:\mathbb{R}^{2(2k+1)}\to\mathbb{R}$ defined as $\phi_{1,h}(\underline{x},\underline{y}) := \underline{x}_1\underline{y}_1$ and $\phi_{2,h}(\underline{x},\underline{y}) := \underline{x}_{2k+1}\underline{y}_{2k+1}$. Note that $\mathbb{E}[\tau_t\underline{\tilde{Z}}_{t,1}\underline{\beta}_1]=0$.

We now show the concentration of $(h^{t+1})^* q^0/n$. Rewrite $(h^{t+1})^* q^0$ as

$$(h^{t+1})^* q^0 = \sum_{i=k+1}^{N-k} h_i^{t+1} f_0(\underline{0}, [\beta_0]_{i-k}^{i+k}) + \sum_{i=1}^{k} h_i^{t+1}\beta_{0,i} + \sum_{i=N-k+1}^{N} h_i^{t+1}\beta_{0,i}.$$

Then we have

$$\begin{aligned} P\left(\left|\frac{(h^{t+1})^* q^0}{n}\right| \geq \epsilon\right) &\overset{(a)}{\leq} P\left(\left|\sum_{i=k+1}^{N-k} \frac{h_i^{t+1} f_0(\underline{0}, [\beta_0]_{i-k}^{i+k})}{N-2k}\right| \geq \frac{n\epsilon}{3(N-2k)}\right) + P\left(\left|\sum_{i=1}^{k} \frac{h_i^{t+1}\beta_{0,i}}{k}\right| \geq \frac{n\epsilon}{3k}\right) \\ &\quad + P\left(\left|\sum_{i=N-k+1}^{N} \frac{h_i^{t+1}\beta_{0,i}}{k}\right| \geq \frac{n\epsilon}{3k}\right) \\ &\overset{(b)}{\leq} Ke^{-\kappa N\epsilon^2} + Ke^{-\kappa n\epsilon^2} + Ke^{-\kappa n\epsilon^2}, \end{aligned}$$

where step (a) follows Lemma A.1 and step (b) follows $\mathcal{H}_{t+1}(b)(i)$ by considering $PL(2)$ functions $\phi_{1,h}, \phi_{2,h}, \phi_{3,h} : \mathbb{R}^{2(2k+1)} \to \mathbb{R}$ defined as $\phi_{1,h}(\underline{x}, \underline{y}) := \underline{x}_{k+1} f_0(\underline{0}, \underline{y})$, $\phi_{2,h}(\underline{x}, \underline{y}) := \underline{x}_1 \underline{y}_1$, and $\phi_{3,h}(\underline{x}, \underline{y}) := \underline{x}_{2k+1}\underline{y}_{2k+1}$. Note that $\mathbb{E}[\tau_t \tilde{\underline{Z}}_{t,k+1} f(\underline{0}, \underline{\beta})] = 0$.

(d) Similar to $\mathcal{H}_{t+1}(c)$, we split the inner product $(h^{r+1})^* h^{t+1}$ and then from Lemma A.1,

$$\begin{aligned} P\left(\left|\frac{(h^{r+1})^* h^{t+1}}{N} - \breve{E}_{r,t}\right| \geq \epsilon\right) &\leq P\left(\left|\sum_{i=1}^{N/2} \frac{h_i^{r+1} h_i^{t+1}}{N/2} - \breve{E}_{r,t}\right| \geq \epsilon\right) + P\left(\left|\sum_{i=N/2+1}^{N} \frac{h_i^{r+1} h_i^{t+1}}{N/2} - \breve{E}_{r,t}\right| \geq \epsilon\right) \\ &\overset{(a)}{\leq} Ke^{-\kappa N\epsilon^2} + Ke^{-\kappa N\epsilon^2}, \end{aligned}$$

where step (a) follows $\mathcal{H}_{t+1}(b)(i)$ by considering $PL(2)$ functions $\phi_{1,h}, \phi_{2,h} : \mathbb{R}^{2(2k+1)} \to \mathbb{R}$ defined as $\phi_{1,h}(\underline{x}, \underline{y}) := \underline{x}_1 \underline{y}_1$ and $\phi_{1,h}(\underline{x}, \underline{y}) := \underline{x}_{2k+1}\underline{y}_{2k+1}$.

(e) We first show the concentration of $(q^0)^* q^{t+1}/n$. Recall that for $0 \leq r, s \leq t+1$

$$\delta \tilde{E}_{r,s} = \frac{N-2k}{N}\mathbb{E}[f_r(\tau_{r-1}\tilde{\underline{Z}}_{r-1}, \underline{\beta}) f_s(\tau_{s-1}\tilde{\underline{Z}}_{s-1}, \underline{\beta})] + \frac{2k}{N}\sigma_\beta^2. \tag{84}$$

Then splitting $(q^0)^* q^{t+1}$ as in $\mathcal{H}_1(e)$, we have

$$\begin{aligned} &P\left(\left|\frac{(q^0)^* q^{t+1}}{n} - \tilde{E}_{0,t+1}\right| \geq \epsilon\right) \\ &\overset{(a)}{\leq} P\left(\left|\frac{1}{N-2k}\sum_{i=k+1}^{N-k} f_0(\underline{0}, [\beta_0]_{i-k}^{i+k}) f_{t+1}([h^{t+1}]_{i-k}^{i+k}, [\beta_0]_{i+k}^{i-k}) - \mathbb{E}[f_0(\underline{0}, \underline{\beta}) f_{t+1}(\tau_t \tilde{\underline{Z}}_t, \underline{\beta})]\right| \geq \frac{n\epsilon}{3(N-2k)}\right) \\ &\quad + P\left(\left|\sum_{i=1}^{k} \frac{(\beta_{0,i})^2}{k} - \sigma_\beta^2\right| \geq \frac{n\epsilon}{3k}\right) + P\left(\left|\sum_{i=N-k+1}^{N} \frac{(\beta_{0,i})^2}{k} - \sigma_\beta^2\right| \geq \frac{n\epsilon}{3k}\right) \\ &\overset{(b)}{\leq} Ke^{-\kappa n\epsilon^2} + Ke^{-\kappa n\epsilon^2}, \end{aligned}$$

where step (a) follows Lemma A.1 and step (b) follows $\mathcal{H}_{t+1}(b)(i)$ by considering the $PL(2)$ functions $\phi_{1,h}, \phi_{2,h}, \phi_{3,h} : \mathbb{R}^{2(2k+1)} \to \mathbb{R}$ defined as $\phi_{1,h}(\underline{x}, \underline{y}) := f_0(\underline{0}, \underline{y}) f_{t+1}(\underline{x}, \underline{y})$, $\phi_{2,h}(\underline{x}, \underline{y}) := \underline{y}_1^2$, and $\phi_{3,h}(\underline{x}, \underline{y}) := \underline{y}_{2k+1}^2$.

Concentration of $(q^{r+1})^* q^{t+1}/n$ can be obtained similarly by representing

$$(q^{r+1})^* q^{t+1} = \sum_{i=k+1}^{N-k} f_{r+1}([h^{r+1}]_{i-k}^{i+k}, [\beta_0]_{i-k}^{i+k}) f_{t+1}([h^{t+1}]_{i-k}^{i+k}, [\beta_0]_{i-k}^{i+k}) + \sum_{i=1}^{k} \beta_{0,i}^2 + \sum_{i=N-k+1}^{N} \beta_{0,i}^2 \quad (85)$$

(f) The concentration of λ_t around $\hat{\lambda}_t$ follows $\mathcal{H}_{t+1}(b)(ii)$ applied to the function $\psi_h([h^{t+1}]_{i-k}^{i+k}, [\beta_0]_{i-k}^{i+k}) := f'_{t+1}([h^{t+1}]_{i-k}^{i+k}, [\beta_0]_{i-k}^{i+k})$. Next, for $r \leq t$, splitting $(h^{t+1})^* q^{r+1}$ as in $\mathcal{H}_1(f)$,

$$\begin{aligned}
&P\left(\left|\frac{(h^{t+1})^* q^{r+1}}{n} - \hat{\lambda}_{r+1} \breve{E}_{r,t}\right| \geq \epsilon\right) \\
&\overset{(a)}{\leq} P\left(\left|\frac{1}{N-2k}\left[\sum_{i=k+1}^{N-k} h_i^{t+1} f_{r+1}([h^{r+1}]_{i-k}^{i+k}, [\beta_0]_{i-k}^{i+k}) - n\hat{\lambda}_{r+1}\breve{E}_{r,t}\right]\right| \geq \frac{n\epsilon}{3(N-2k)}\right) \\
&\quad + P\left(\left|\sum_{i=1}^{k} \frac{h_i^{t+1}\beta_{0,i}}{k}\right| \geq \frac{n\epsilon}{3k}\right) + P\left(\left|\frac{1}{k}\sum_{i=N-k+1}^{N} \frac{h_i^{t+1}\beta_{0,i}}{k}\right| \geq \frac{n\epsilon}{3k}\right) \\
&\overset{(b)}{\leq} Ke^{-\kappa N\epsilon^2} + Ke^{-\kappa n\epsilon^2} + Ke^{-\kappa n\epsilon^2},
\end{aligned}$$

where step (a) follows from Lemma A.1 and step (b) from $\mathcal{H}_{t+1}(b)(i)$ by considering $PL(2)$ functions $\phi_{1,h}, \phi_{2,h}, \phi_{3,h} : \mathbb{R}^{2(2k+1)} \to \mathbb{R}$ defined as $\phi_{1,h}(\underline{x}, \underline{y}) := \underline{x}_{k+1} f_{r+1}(\underline{x}, \underline{y})$, $\phi_{2,h}(\underline{x}, \underline{y}) := \underline{x}_1 \underline{y}_1$, $\phi_{3,h}(\underline{x}, \underline{y}) := \underline{x}_{2k+1}\underline{y}_{2k+1}$. The result follows by noticing $\mathbb{E}[\tau_t \tilde{\underline{Z}}_{t,i}\underline{\beta}_i] = 0$, for all $i \in [N]$, and

$$\mathbb{E}[\tau_t \tilde{\underline{Z}}_{t,k+1} f_{r+1}(\tau_r \tilde{\underline{Z}}_r, \underline{\beta})] = \frac{n}{N-2k}\hat{\lambda}_{r+1}\breve{E}_{r,t},$$

which follows by Stein's Lemma given in Fact 2. We demonstrate this in the following. Think of a function $\tilde{f} : \mathbb{R} \to \mathbb{R}$ defined as $\tilde{f}(x) := f_{r+1}(\tau_r \tilde{\underline{Z}}_{r,1}, ..., \tilde{\underline{Z}}_{r,k}, x, \tilde{\underline{Z}}_{r,k+2}, ..., \tilde{\underline{Z}}_{r,2k+1}, \underline{\beta})$. Then,

$$\mathbb{E}[\tau_t \tilde{\underline{Z}}_{t,k+1} f_{r+1}(\tau_r \tilde{\underline{Z}}_r, \underline{\beta})] = \mathbb{E}[\tau_t \tilde{\underline{Z}}_{t,k+1} \tilde{f}(\tau_r \tilde{\underline{Z}}_{r,k+1})] \overset{(a)}{=} \tau_t\tau_r \mathbb{E}[\tilde{\underline{Z}}_{t,k+1}\tilde{\underline{Z}}_{r,k+1}]\mathbb{E}[\tilde{f}'(\tau_r \tilde{\underline{Z}}_r, \underline{\beta})] \overset{(b)}{=} \frac{n\hat{\lambda}_{r+1}\breve{E}_{r,t}}{N-2k}.$$

Step (a) applies Stein's Lemma, Fact 2. Step (b) uses the facts that $\tau_t\tau_r\mathbb{E}[\tilde{\underline{Z}}_{t,k+1}\tilde{\underline{Z}}_{r,k+1}] = \breve{E}_{r,t}$ from (21) and that the derivative of $\tilde{f}$ is the derivative of f_t with respect to the middle coordinate of the first argument, along with the definition of $\hat{\lambda}_{r+1}$ in (20). Therefore, we have obtained the desired result.

(g) We can represent $\mathbf{Q}_{t+1}$ as follows.

$$\mathbf{Q}_{t+1} = \frac{1}{n}\begin{pmatrix} n\mathbf{Q}_t & Q_t^* q^t \\ (Q_t^* q^t)^* & \left\|q^t\right\|^2 \end{pmatrix},$$

Then by block inversion formula

$$\mathbf{Q}_{t+1}^{-1} = \begin{pmatrix} \mathbf{Q}_t^{-1} + n\left\|q_\perp^t\right\|^{-2}\gamma^t(\gamma^t)^* & -n\left\|q_\perp^t\right\|^{-2}\gamma^t \\ -n\left\|q_\perp^t\right\|^{-2}(\gamma^t)^* & n\left\|q_\perp^t\right\|^{-2} \end{pmatrix}, \quad (86)$$

where we have used facts $\frac{1}{n}\mathbf{Q}_t^{-1}Q_t^* q^t = \gamma^t$ and $(Q_t^* q^t)^*\gamma^t = (q^t)^* q_\parallel^t$. Block inversion can be similarly used to decompose the elements of $\tilde{C}^{t+1}$:

$$(\tilde{C}^{t+1})^{-1} = \begin{pmatrix} \tilde{C}^t & \tilde{E}_t \\ \tilde{E}_t^* & \sigma_t^2 \end{pmatrix}^{-1} = \begin{pmatrix} (\tilde{C}^t)^{-1} + (\sigma_t^\perp)^{-2}\hat{\gamma}^t(\hat{\gamma}^t)^* & -(\sigma_t^\perp)^{-2}\hat{\gamma}^t \\ -(\sigma_t^\perp)^{-2}(\hat{\gamma}^t)^* & (\sigma_t^\perp)^{-2} \end{pmatrix}, \quad (87)$$

where we have used (24) and (26). In what follows we prove concentrates for each of the elements in (86) to the corresponding elements in (87).

First, by $\mathcal{H}_t(h)$ and Lemma A.5, $n\left\|q_\perp^t\right\|^{-2}$ concentrates to $(\sigma_t^\perp)^{-2}$ at rate $K\exp\left\{-\kappa n\epsilon^2\right\}$. Next, consider the i^{th} element of $-n\left\|q_\perp^t\right\|^{-2}\gamma^t$. For $i\in[t]$,

$$P\left(\left|n\left\|q_\perp^t\right\|^{-2}\gamma_{i-1}^t-(\sigma_t^\perp)^{-2}\hat{\gamma}_{i-1}^t\right|\geq\epsilon\right)\overset{(a)}{\leq}Ke^{-\kappa n\epsilon^2}. \tag{88}$$

Step (a) follows from Lemma A.2, and $\mathcal{H}_t(g),(h)$ as discussed in the previous paragraph. Consider element (i,j) of $\mathbf{Q}_t^{-1}+n\left\|q_\perp^t\right\|^{-2}\gamma^t(\gamma^t)^*$ for $i,j\in[t]$.

$$\begin{aligned}
&P\left(\left|[\mathbf{Q}_t^{-1}]_{i,j}+n\left\|q_\perp^t\right\|^{-2}\gamma_{i-1}^t\gamma_{j-1}^t-[(\tilde{C}^t)^{-1}]_{i,j}-(\sigma_t^\perp)^{-2}\hat{\gamma}_{i-1}^t\hat{\gamma}_{j-1}^t\right|\geq\epsilon\right)\\
&\overset{(a)}{\leq}P\left(\left|[\mathbf{Q}_t^{-1}]_{i,j}-[(\tilde{C}^t)^{-1}]_{i,j}\right|\geq\frac{\epsilon}{2}\right)+P\left(\left|\gamma_{j-1}^t-\hat{\gamma}_{j-1}^t\right|\geq\frac{\epsilon'}{2}\right)\\
&\quad+P\left(\left|n\left\|q_\perp^t\right\|^{-2}\gamma_{i-1}^t-(\sigma_t^\perp)^{-2}\hat{\gamma}_{i-1}^t\right|\geq\frac{\epsilon'}{2}\right)\\
&\overset{(b)}{\leq}Ke^{-\kappa n\epsilon^2}+Ke^{-\kappa n\epsilon^2}+Ke^{-\kappa n\epsilon^2}.
\end{aligned}$$

Step (a) follows from Lemma A.1 and Lemma A.2 with

$$\epsilon'=\min\left(\sqrt{\frac{\epsilon}{3}},\frac{\epsilon(\sigma_t^\perp)^2}{3\hat{\gamma}_{i-1}^t},\frac{\epsilon}{3\hat{\gamma}_{j-1}^t}\right).$$

Step (b) follows from the inductive hypothesis $\mathcal{H}_t(g)$, together with (88).

We now prove $\gamma^{t+1}\doteq\hat{\gamma}^{t+1}$. Recall that $\gamma^{t+1}=\mathbf{Q}_{t+1}^{-1}\frac{Q_{t+1}^*q^{t+1}}{n}$ where $\mathbf{Q}_{t+1}:=\frac{Q_{t+1}^*Q_{t+1}}{n}$. Thus for $0\leq r\leq t$, $\gamma_r^{t+1}=\sum_{i=0}^t[\mathbf{Q}_{t+1}^{-1}]_{r+1,i+1}\frac{(q^i)^*q^{t+1}}{n}$. Then form the definition of $\hat{\gamma}^{t+1}$, we have for $0\leq r\leq t$,

$$\begin{aligned}
&P\left(\left|\gamma_r^{t+1}-\hat{\gamma}_r^{t+1}\right|\geq\epsilon\right)=P\left(\left|\sum_{i=0}^t\left([\mathbf{Q}_{t+1}^{-1}]_{r+1,i+1}\frac{(q^i)^*q^{t+1}}{n}-[(\tilde{C}^{t+1})^{-1}]_{r+1,i+1}\tilde{E}_{i,t+1}\right)\right|\geq\frac{\epsilon}{t+1}\right)\\
&\leq\sum_{i=0}^tP\left(\left|[\mathbf{Q}_{t+1}^{-1}]_{r+1,i+1}\frac{(q^i)^*q^{t+1}}{n}-[(\tilde{C}^{t+1})^{-1}]_{r+1,i+1}\tilde{E}_{i,t+1}\right|\geq\frac{\epsilon}{t+1}\right)\\
&\overset{(a)}{\leq}\sum_{i=0}^tP\left(\left|\frac{(q^i)^*q^{t+1}}{n}-\tilde{E}_{i,t+1}\right|\geq\tilde{\epsilon}_i\right)+P\left(\left|[\mathbf{Q}_{t+1}^{-1}]_{r+1,i+1}-[(\tilde{C}^{t+1})^{-1}]_{r+1,i+1}\right|\geq\tilde{\epsilon}_i\right)\\
&\overset{(b)}{\leq}Ke^{-\kappa N\epsilon^2}+Ke^{-\kappa N\epsilon^2}.
\end{aligned}$$

Step (a) follows from Lemma A.2, with $\tilde{\epsilon}_i:=\min\{\sqrt{\frac{\epsilon}{3(t+1)}},\frac{\epsilon}{3(t+1)\tilde{E}_{i,t+1}},\frac{\epsilon}{3(t+1)[(\tilde{C}^{t+1})^{-1}]_{r+1,i+1}}\}$, and step (b) uses $\mathcal{H}_{t+1}(e)$ and what we have just demonstrated in the previous paragraphs.

(h) First, note that $\left\|q_\perp^{t+1}\right\|^2=\left\|q^{t+1}\right\|^2-\left\|q_\parallel^{t+1}\right\|^2=\left\|q^{t+1}\right\|^2-\left\|Q_{t+1}\gamma^{t+1}\right\|^2$. Using the definition of $\sigma_{t+1}^\perp$ in (26), we then have

$$\begin{aligned}
&P\left(\left|\frac{\left\|q_\perp^{t+1}\right\|^2}{n}-(\sigma_{t+1}^\perp)^2\right|\geq\epsilon\right)=P\left(\left|\frac{\left\|q^{t+1}\right\|^2}{n}-\frac{\left\|Q_{t+1}\gamma^{t+1}\right\|^2}{n}-\sigma_{t+1}^2+(\hat{\gamma}^{t+1})^*\tilde{E}_{t+1}\right|\geq\epsilon\right)\\
&\leq P\left(\left|\frac{\left\|q^{t+1}\right\|^2}{n}-\sigma_{t+1}^2\right|\geq\frac{\epsilon}{2}\right)+P\left(\left|\frac{\left\|Q_{t+1}\gamma^{t+1}\right\|^2}{n}-(\hat{\gamma}^{t+1})^*\tilde{E}_{t+1}\right|\geq\frac{\epsilon}{2}\right).
\end{aligned} \tag{89}$$

By $\mathcal{H}_{t+1}(e)$, the first term in (89) is bounded by $Ke^{-\kappa n\epsilon^2}$. For the second term, using $\gamma^{t+1} = \mathbf{Q}_{t+1}^{-1}\frac{Q_{t+1}^* q^{t+1}}{n}$,

$$\left\|Q_{t+1}\gamma^{t+1}\right\|^2 = n(\gamma^{t+1})^*\mathbf{Q}_{t+1}\gamma^{t+1} = (\gamma^{t+1})^*\mathbf{Q}_{t+1}\mathbf{Q}_{t+1}^{-1}Q_{t+1}^* q^{t+1} = (\gamma^{t+1})^* Q_{t+1}^* q^{t+1} = \sum_{i=0}^{t}\gamma_i^{t+1}(q^i)^* q^{t+1}.$$

Hence

$$\begin{aligned}
&P\left(\left|\frac{1}{n}\left\|Q_{t+1}\gamma^{t+1}\right\|^2 - (\hat{\gamma}^{t+1})^*\tilde{E}_{t+1}\right| \geq \frac{1}{2}\epsilon\right) = P\left(\left|\sum_{i=0}^{t}\left(\frac{1}{n}\gamma_i^{t+1}(q^i)^* q^{t+1} - \hat{\gamma}_i^{t+1}\tilde{E}_{i,t+1}\right)\right| \geq \frac{1}{2}\epsilon\right)\\
&\leq \sum_{i=0}^{t} P\left(\left|\frac{1}{n}\gamma_i^{t+1}(q^i)^* q^{t+1} - \hat{\gamma}_i^{t+1}\tilde{E}_{i,t+1}\right| \geq \frac{1}{2}\epsilon(t+1)^{-1}\right)\\
&\overset{(a)}{\leq} \sum_{i=0}^{t} P\left(\left|\gamma_i^{t+1} - \hat{\gamma}_i^{t+1}\right| \geq \tilde{\epsilon}_i\right) + \sum_{i=0}^{t} P\left(\left|\frac{1}{n}(q^i)^* q^{t+1} - \tilde{E}_{i,t+1}\right| \geq \tilde{\epsilon}_i\right)\\
&\overset{(b)}{\leq} Ke^{-\kappa n\epsilon^2} + Ke^{-\kappa n\epsilon^2}.
\end{aligned}$$

Step (a) follows from the concentration of products, Lemma A.2, using $\tilde{\epsilon}_i := \min\left\{\sqrt{\frac{\epsilon}{6(t+1)}}, \frac{\epsilon}{6(t+1)\tilde{E}_{i,t+1}}, \frac{\epsilon}{6(t+1)\hat{\gamma}_i^{t+1}}\right\}$, and step (b) using $\mathcal{H}_{t+1}(e)$ and $\mathcal{H}_{t+1}(g)$.

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 [9].

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 $P\left(|XY - c_X c_Y| \geq \epsilon\right)$ is bounded by

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

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

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

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

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

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

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

B Gaussian and Sub-Gaussian Concentration

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

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

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

Lemma B.3. *[12] Let X be a centered sub-Gaussian random variable with variance factor ν, i.e., $\ln \mathbb{E}[e^{tX}] \leq \frac{t^2\nu}{2}$, $\forall t \in \mathbb{R}$. Then X satisfies:*

1. *For all $x > 0$, $P(X > x) \vee P(X < -x) \leq e^{-\frac{x^2}{2\nu}}$, for all $x > 0$.*
2. *For every integer $k \geq 1$,*
$$\mathbb{E}[X^{2k}] \leq 2(k!)(2\nu)^k \leq (k!)(4\nu)^k. \tag{90}$$

C Other Useful Lemmas

Lemma C.1. *(Products of Lipschitz Functions are PL2) Let $f : \mathbb{R}^p \to \mathbb{R}$ and $g : \mathbb{R}^p \to \mathbb{R}$ be Lipschitz continuous. Then the product function $h : \mathbb{R}^p \to \mathbb{R}$ defined as $h(x) := f(x)g(x)$ is pseudo-Lipschitz of order 2.*

Lemma C.2. *Let $\phi : \mathbb{R}^{t+2} \to \mathbb{R}$ be $PL(2)$. Let $(c_1, \ldots, c_{t+1})$ be constants. The function $\tilde{\phi} : \mathbb{R}^{t+1} \to \mathbb{R}$ defined as*

$$\tilde{\phi}\left(v_1, \ldots, v_t, w\right) = \mathbb{E}_Z\left[\phi\left(v_1, \ldots, v_t, \sum_{r=1}^{t} c_r v_r + c_{t+1} Z, w\right)\right] \tag{91}$$

where $Z \sim \mathcal{N}(0,1)$, is then also PL(2).

Lemma C.3. *For any scalars $a_1, ..., a_t$ 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 $\underline{u}_1, \ldots, \underline{u}_t \in \mathbb{R}^N$, $\left\|\sum_{k=1}^{t}\underline{u}_k\right\|^2 \leq t\sum_{k=1}^{t}\|\underline{u}_k\|^2$.*

D Concentration with Dependencies

We first list some notation that will be used frequently in this section. Let $S \subset \mathbb{R}^d$ for some $d \in \mathbb{N}$ be a state space and π a probability measure on S. Let $f : S \to \mathbb{R}$ be a measurable function. We use the following notation:

- The sup-norm: $\|f\|_\infty := \sup_{x \in S} |f(x)|$;
- The $L^2(\pi)$-norm: $\|f\|_{2,\pi}^2 := \int_S |f(x)|^2 \pi(dx)$;
- The expected value: $\mathbb{E}_\pi f := \int_S f(x)\pi(dx)$;
- The function space: $L_0^2(\pi) := \{f : S \to \mathbb{R} : \mathbb{E}_\pi f = 0, \|f\|_{2,\pi} < \infty\}$;

The following two lemmas exist in the literature and are reproduced here, without proofs, for completeness. The proofs can be found in the citations.

Lemma D.1. *[13, Theorem 2.1] Consider a Markov chain taking values in a state space S with probability transition measure $r(x, dx')$ and stationary probability measure γ. If the linear operator R associated with $r(x, dx')$ is defined as $Rh(x) = \int_S h(x')r(x, dx')$ and the Markov chain is geometrically ergodic as defined in Definition 2.1, then R has an $L^2(\gamma)$ spectral gap. That is, for each signed measure $\nu \in L^2(\gamma)$ with $\nu(S) = 0$, there is a $0 < \rho < 1$ such that*

$$\int_S \left| \frac{\int_S r(x', dx)\nu(dx')}{\gamma(dx)} \right|^2 \gamma(dx) \leq \rho \int_S \left| \frac{\nu(dx)}{\gamma(dx)} \right|^2 \gamma(dx).$$

Lemma D.2. *[14, Theorem 2.1] Let $d, n > 0$ be integers and*

$$X = \frac{1}{n - 2d} \sum_{i=1+d}^{n-d} f_i([\xi]_{i-d}^{i+d}),$$

for some functions f_i and independent random variables $\xi_1, \xi_2, \ldots, \xi_n$. Suppose the functions f_i are bounded, meaning $a_i \leq f_i([\xi]_{i-d}^{i+d}) \leq b_i$ for every $1 + d \leq i \leq n - d$ and some real numbers a_i, b_i. Then for $\epsilon > 0$,

$$P\left(|X - \mathbb{E}X| \geq \epsilon\right) \leq 2 \exp\left\{ -\frac{2(n - 2d)^2 \epsilon^2}{\left(\lfloor \frac{n}{d} \rfloor + 1\right) \sum_{i=1+d}^{n-d} (b_i - a_i)^2} \right\}.$$

The following three lemmas will be used to prove the latter three lemmas, which are our new contributions.

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

$$\mathbb{E}[e^{rf(X)}] \leq \mathbb{E}[e^{r(f(X) - f(\tilde{X})}] \leq [1 - 25r^2L^2(d\nu + 12d^2\nu^2)]^{-1} \leq e^{50r^2L^2(d\nu + 12d^2\nu^2)}. \tag{92}$$

Proof. Assume, without loss of generality $\mathbb{E}[f(X)] = 0$. By Jensen's inequality, $\mathbb{E}[\exp(-rf(\tilde{X}))] \leq \exp(-r\mathbb{E}[f(\tilde{X})]) = 1$. Therefore,

$$\mathbb{E}\left[\exp\left(rf(X)\right)\right] \leq \mathbb{E}\left[\exp\left(rf(X)\right)\right]\mathbb{E}[\exp(rf(\tilde{X}))] = \mathbb{E}[\exp(r(f(X) - f(\tilde{X})))],$$

which provides the first upper bound in (92). Next,

$$\begin{aligned}\mathbb{E}[e^{r(f(X)-f(\tilde{X}))}] &\overset{(a)}{\leq} \mathbb{E}[e^{rL(1+\|X\|+\|\tilde{X}\|)\|X-\tilde{X}\|}] = \sum_{q=0}^{\infty}\frac{(rL)^q}{q!}\mathbb{E}[((1+\|X\|+\|\tilde{X}\|)\|X-\tilde{X}\|)^q] \\ &\overset{(b)}{=} \sum_{k=0}^{\infty}\frac{(rL)^{2k}}{(2k)!}\mathbb{E}[((1+\|X\|+\|\tilde{X}\|)(\|X\|+\|\tilde{X}\|))^{2k}],\end{aligned} \tag{93}$$

where step (a) follows pseudo-Lipschitz property and step (b) holds because the odd order terms are zero along with triangle inequality. Now consider the expectation in the last term in the string given in (93).

$$\begin{aligned}\mathbb{E}[((1+\|X\|+\|\tilde{X}\|)(\|X\|+\|\tilde{X}\|))^{2k}] &= \mathbb{E}[(\|X\|+\|\tilde{X}\|+\|X\|^2+\|\tilde{X}\|^2+2\|X\|\|\tilde{X}\|)^{2k}] \\ &\overset{(c)}{\leq} 5^{2k-1}(2\mathbb{E}\|X\|^{2k} + 2\mathbb{E}\,\|X\|^{4k} + 2^{2k}\mathbb{E}[\|X\|^{2k}\|\tilde{X}\|^{2k}]) \\ &\overset{(d)}{\leq} 5^{2k-1}(4(k!)(2d\nu)^k + 4(2k)!(2d\nu)^{2k} + 4(k!)^2(4d\nu)^{2k})\end{aligned}$$

In the above step (c) follows from Lemma C.3 and step (d) from another application of Lemma C.3 and Lemma B.3. Now plugging the above back into (93), we find

$$\begin{aligned}&\mathbb{E}[e^{r(f(X)-f(\tilde{X}))}] \leq \sum_{k=0}^{\infty}\frac{(5rL)^{2k}}{5(2k)!}(4(k!)(2d\nu)^k + 4(2k)!(2d\nu)^{2k} + 4(k!)^2(4d\nu)^{2k}) \\ &\overset{(e)}{\leq} 1 + \frac{4}{5}\sum_{k=1}^{\infty}(5rL)^{2k}((d\nu)^k + (2d\nu)^{2k} + 2^k(2d\nu)^{2k}) \leq \sum_{k=0}^{\infty}(25r^2L^2)^k(d\nu + 12d^2v^2)^k \\ &\overset{(f)}{=} \frac{1}{1-25r^2L^2(d\nu+12d^2\nu^2)} \overset{(g)}{\leq} \exp(50r^2L^2(d\nu+12d^2\nu^2)r^2), \text{ for } 0 < r < \frac{1}{5L\sqrt{2d\nu+24d^2\nu^2}},\end{aligned}$$

where step (e) follows from the fact that $2^k(k!)^2 \leq (2k)!$, which can be seen by noting

$$\frac{(2k)!}{k!} = \prod_{j=1}^{k}(k+j) = k!\prod_{j=1}^{k}\left(\frac{k}{j}+1\right) \geq (k!)2^k,$$

step (f) follows for $0 < r < [25L^2(d\nu+12d^2\nu^2)]^{-1/2}$ providing the second bound in (92), and step (g) uses the inequality $(1-x)^{-1} \leq e^{2x}$ for $x \in [0, 1/2]$ for the final bound in (92). □

Lemma D.4. *Let $\{X_i\}_{i\in\mathbb{N}}$ be a time-homogeneous Markov chain on a state space S with transition probability measure $r(x, dx')$ and stationary probability measure γ. Assume that $\{X_i\}_{i\in\mathbb{N}}$ is reversible, geometrically ergodic on $L^2(\gamma)$ as defined in Definition 2.1. Let $\{\tilde{X}_i\}_{i\in\mathbb{N}}$ be an independent copy of $\{X_i\}_{i\in\mathbb{N}}$. Then the new sequence defined as $\{(X_i, \tilde{X}_i)\}_{i\in\mathbb{N}}$ is a Markov chain on $S \times S$ that is reversible and geometrically ergodic on $L^2(\gamma \times \gamma)$.*

Proof. Since $\{X_i\}_{i\in\mathbb{N}}$ is independent of $\{\tilde{X}_i\}_{i\in\mathbb{N}}$, we have that the transition probability measure of $\{(X_i,\tilde{X}_i)\}_{i\in\mathbb{N}}$ is $\tilde{r}((x,\tilde{x}),(dx',d\tilde{x}')) = r(x,dx')r(\tilde{x},d\tilde{x}')$, and the stationary probability measure of $\{(X_i,\tilde{X}_i)\}_{i\in\mathbb{N}}$ is $\tilde{\gamma}(dx,d\tilde{x}) = \gamma(dx)\gamma(d\tilde{x})$. In what follows, we demonstrate that $\tilde{r}((x,\tilde{x}),(dx',d\tilde{x}'))$ and $\tilde{\gamma}(dx,d\tilde{x})$ satisfy the reversibility and geometric ergodicity as defined in Definition 2.1.

The reversibility of the coupled chain follows directly from the reversibility of the individual chains:

$$\begin{aligned}\tilde{r}((x,\tilde{x}),(dx',d\tilde{x}'))\tilde{\gamma}(dx,d\tilde{x}) = r(x,dx')\gamma(dx)r(\tilde{x},d\tilde{x}')\gamma(d\tilde{x}) &= r(x',dx)\gamma(dx')r(\tilde{x}',d\tilde{x})\gamma(d\tilde{x}')\\ &= \tilde{r}((x',\tilde{x}'),(dx,d\tilde{x}))\tilde{\gamma}(dx',d\tilde{x}').\end{aligned}$$

To prove geometric ergodicity, we want to show that there is $\rho<1$ such that for each *probability measure* $\tilde{\nu}=\nu\times\nu\in L^2(\tilde{\gamma}) = \{\tilde{\nu}\ll\tilde{\gamma} : \int_{S\times S}\left|\frac{\tilde{\nu}(dz)}{\tilde{\gamma}(dz)}\right|\tilde{\gamma}(dz)<\infty\}$, there is $\tilde{C}_\nu<\infty$ such that

$$\sup_{(A,\tilde{A})\in\mathcal{B}(S\times S)}\left|\int_{S\times S}\tilde{r}^n(z,(A,\tilde{A}))\tilde{\nu}(dz)-\tilde{\gamma}(A,\tilde{A})\right|\le\tilde{C}_\nu\rho,$$

where $\mathcal{B}(S\times S)$ is the Borel sigma-algebra on $S\times S$. Notice that

$$\begin{aligned}&\left|\int_{S\times S}\tilde{r}^n(z,(A,\tilde{A}))\tilde{\nu}(dz)-\tilde{\gamma}(A,\tilde{A})\right| = \left|\int_S r^n(x,A)\nu(dx)\int_S r^n(\tilde{x},\tilde{A})\nu(d\tilde{x})-\gamma(A)\gamma(\tilde{A})\right|\\
&=\left|\left(\int_S r^n(x,A)\nu(dx)-\gamma(A)\right)\left(\int_S r^n(\tilde{x},\tilde{A})\nu(d\tilde{x})-\gamma(\tilde{A})\right)\right.\\
&\qquad\left.+\gamma(\tilde{A})\left(\int_S r^n(x,A)\nu(dx)-\gamma(A)\right)+\gamma(A)\left(\int_S r^n(\tilde{x},\tilde{A})\nu(d\tilde{x})-\gamma(\tilde{A})\right)\right|\\
&\overset{(a)}{\le}\left|\int_S r^n(x,A)\nu(dx)-\gamma(A)\right|\left|\int_S r^n(\tilde{x},\tilde{A})\nu(d\tilde{x})-\gamma(\tilde{A})\right|\\
&\qquad+\left|\int_S r^n(x,A)\nu(dx)-\gamma(A)\right|+\left|\int_S r^n(\tilde{x},\tilde{A})\nu(d\tilde{x})-\gamma(\tilde{A})\right|,\end{aligned}$$

where step (a) used triangle inequality and $0\le\gamma(A)\le 1$ for all $A\in\mathcal{B}(S)$. Taking the supremum of both sides of the above,

$$\begin{aligned}&\sup_{(A,\tilde{A})\in\mathcal{B}(S\times S)}\left|\int_{S\times S}\tilde{r}^n(z,(A,\tilde{A}))\tilde{\nu}(dz)-\tilde{\gamma}(A,\tilde{A})\right|\\
&\qquad\le\sup_{A\in\mathcal{B}(S)}\left|\int_S r^n(x,A)\nu(dx)-\gamma(A)\right|\sup_{\tilde{A}\in\mathcal{B}(S)}\left|\int_S r^n(\tilde{x},\tilde{A})\nu(d\tilde{x})-\gamma(\tilde{A})\right|\\
&\qquad\quad+\sup_{A\in\mathcal{B}(S)}\left|\int_S r^n(x,A)\nu(dx)-\gamma(A)\right|+\sup_{\tilde{A}\in\mathcal{B}(S)}\left|\int_S r^n(\tilde{x},\tilde{A})\nu(d\tilde{x})-\gamma(\tilde{A})\right|\\
&\qquad\overset{(a)}{\le}C_\nu^2\rho^{2n}+2C_\nu\rho^n\overset{(b)}{<}(C_\nu^2+2C_\nu)\rho^n,\end{aligned}$$

where we have $\tilde{C}_\nu := C_\nu^2+2C_\nu<\infty$. Step (a) follows from the fact that $\{X_i\}_{i\in\mathbb{N}}$ is geometrically ergodic and the definition of such in Definition 2.1 and step (b) since $0<\rho<1$. □

Lemma D.5. *Let $\{X_i\}_{i\in\mathbb{N}}$ be a time-homogeneous Markov chain on a state space S with transition probability measure $r(x,dx')$ and stationary probability measure γ. Assume that $\{X_i\}_{i\in\mathbb{N}}$ is reversible, geometrically ergodic on $L^2(\gamma)$ as defined in Definition 2.1. Define $\{Y_i\}_{i\in\mathbb{N}}$ as $Y_i =$*

$(X_{di-d+1}, ..., X_{di}) \in S^d$, *where* d *is an integer. Then* $\{Y_i\}_{i\in\mathbb{N}}$ *is a stationary, time-homogeneous Markov chain with transition probability measure* $p(y, dy')$ *and stationary probability measure* π. *Moreover, the linear operator* P *defined as* $Ph(y) := \int_{S^d} h(y')p(y, dy')$ *satisfies*

$$\beta_P := \sup_{h\in L_0^2(\pi)} \frac{\|Ph\|_{2,\pi}}{\|h\|_{2,\pi}} < 1. \tag{94}$$

Proof. The Markov property and time-homogeneous property follow directly by the construction of $\{Y_i\}_{i\in\mathbb{N}}$. We now verify that π is a stationary distribution for $p(y, dy')$. That is, we need to show that $\int_{S^d} p(y, dy')\pi(dy) = \pi(dy')$. First we write $p(y, dy')$ and π in terms of $r(x, dx')$ and γ:

$$\begin{aligned}
\pi(dy) = \pi(dy_1, ..., dy_d) &= \prod_{i=2}^{d} r(y_{i-1}, dy_i)\gamma(dy_1) \\
p(y, dy') = P\left(Y_2 \in dy' | Y_1 = y\right) &= P\left(X_{d+1} \in dy'_1, ..., X_{2d} \in dy'_d | X_1 = y_1, ..., X_d = y_d\right) \\
&= P\left(X_{d+1} \in dy'_1, ..., X_{2d} \in dy'_d | X_d = y_d\right) = r(y_d, dy'_1) \prod_{i=2}^{d} r(y'_{i-1}, dy'_i).
\end{aligned} \tag{95}$$

Then we have

$$\begin{aligned}
&\int_{y\in S^d} p(y, dy')\pi(dy) \overset{(a)}{=} \int_{y\in S^d} r(y_d, dy'_1) \prod_{i=2}^{d} r(y'_{i-1}, dy'_i) \prod_{i=2}^{d} r(y_{i-1}, dy_i)\gamma(dy_1) \\
&= \prod_{i=2}^{d} r(y'_{i-1}, dy'_i) \int_{y\in S^d} r(y_d, dy'_1) \prod_{i=2}^{d} r(y_{i-1}, dy_i)\gamma(dy_1) \overset{(b)}{=} \prod_{i=2}^{d} r(y'_{i-1}, dy'_i)\gamma(dy'_1) = \pi(dy'),
\end{aligned}$$

where step (a) follows from (95), and step (b) since γ is the stationary probability measure for $r(x, dx')$. Hence, we have verified that π is a stationary probability measure for $p(y, dy')$.

We now prove (94). Note β_P is a property of the Markov chain $\{Y_i\}_{i\in\mathbb{N}}$. If $\{Y_i\}_{i\in\mathbb{N}}$ is reversible and geometrically ergodic, then we would be able show (94) using Lemma D.1 directly. However, $\{Y_i\}_{i\in\mathbb{N}}$ is non-reversible, hence, we instead relate β_P to a similar property for the original $\{X_i\}_{i\in\mathbb{N}}$ chain, which we assume is reversible and geometrically ergodic, then use Lemma D.1.

Take arbitrary $h \in L_0^2(\pi)$, we have

$$\frac{\|Ph\|_{2,\pi}^2}{\|h\|_{2,\pi}^2} = \frac{\int_{S^d} \left(\int_{S^d} h(y')p(y, dy')\right)^2 \pi(dy)}{\int_{S^d} h^2(y)\pi(dy)}. \tag{96}$$

First consider the numerator of (96). Plugging in the expressions for $p(y, dy')$ and $\pi(dy)$ defined in (95), we write the numerator as

$$\begin{aligned}
&\int_{S^d} \left(\int_{S^d} h(y')r(y_d, dy'_1) \prod_{i=2}^{d} r(y'_{i-1}, dy'_i)\right)^2 \prod_{i=2}^{d} r(y_{i-1}, dy_i)\gamma(dy_1) \\
&\overset{(a)}{=} \int_{S} \left(\int_{S^d} h(y')r(y_d, dy'_1) \prod_{i=2}^{d} r(y'_{i-1}, dy'_i)\right)^2 \gamma(dy_d) \overset{(b)}{=} \int_{S} \left(\int_{S} \tilde{h}(y'_1)r(y_d, dy'_1)\right)^2 \gamma(dy_d) \overset{(c)}{=} \|R\tilde{h}\|_{2,\gamma}^2.
\end{aligned} \tag{97}$$

Step (a) holds because γ is the stationary probability measure for $r(x, dx')$ and the integrand inside the square does not involve $(y_1, ..., y_{d-1})$. In step (b), the function $\tilde{h} : \mathbb{R} \to \mathbb{R}$ is defined as

$$\tilde{h}(y_1') := \int_{S^{d-1}} h((y_1', ..., y_d')) \prod_{i=2}^{d} r(y_{i-1}', dy_i'). \tag{98}$$

In step (c), the operator R is defined as $R\tilde{h}(x) := \int_S \tilde{h}(x')r(x, dx')$.

We next show that $\tilde{h} \in L_0^2(\gamma)$ for $\tilde{h}$ defined in (98). Notice that

$$\int_S \tilde{h}(y_1')\gamma(dy_1') \overset{(a)}{=} \int_{S^d} h((y_1', ..., y_d'))\pi(dy_1', ..., dy_d') = \int_{S^d} h(y')\pi(dy') \overset{(b)}{=} 0.$$

Step (a) follows by plugging in the definition of $\tilde{h}$ given in (98) and the expression for π from (95). Step (b) holds because $h \in L_0^2(\pi)$. The fact that $\|\tilde{h}\|_{2,\gamma} < \infty$ follows by an application of Jensen's Inequality and the original assumption $\|h\|_{2,\pi} < \infty$. Hence, $\tilde{h} \in L_0^2(\gamma)$.

Next we consider the denominator of (96).

$$\begin{aligned}\int_{S^d} h^2(y)\pi(dy) &= \int_{S^d} h^2((y_1, ..., y_d)) \prod_{i=2}^{d} r(y_{i-1}, dy_i)\gamma(dx_1) \\ &\overset{(a)}{\geq} \int_S \left(\int_{S^{d-1}} h((y_1, ..., y_d)) \prod_{i=2}^{d} r(y_{i-1}, dy_i) \right)^2 \gamma(dy_1) \overset{(b)}{=} \int_S \tilde{h}^2(y_1)\gamma(dy_1) = \|\tilde{h}\|_{2,\gamma}^2, \end{aligned} \tag{99}$$

where step (a) follows from Jensen's inequality and step (b) uses the definition of $\tilde{h}$ given in (98). Combining (97) and (99), we have $\forall h \in L_0^2(\pi)$,

$$\frac{\|Ph\|_{2,\pi}}{\|h\|_{2,\pi}} \leq \frac{\|R\tilde{h}\|_{2,\gamma}}{\|\tilde{h}\|_{2,\gamma}},$$

where $\tilde{h}$ is defined in (98) and we have $\tilde{h} \in L_0^2(\gamma)$ as demonstrated above. Let $\tilde{\mathsf{H}} \subset L_0^2(\gamma)$ be the collection of functions defined in (98) for all $h \in L_0^2(\pi)$. Then we have

$$\beta_P = \sup_{h \in L_0^2(\pi)} \frac{\|Ph\|_{2,\pi}}{\|h\|_{2,\pi}} \leq \sup_{\tilde{h} \in \tilde{\mathsf{H}}} \frac{\|R\tilde{h}\|_{2,\gamma}}{\|\tilde{h}\|_{2,\gamma}} \overset{(a)}{\leq} \sup_{\tilde{h} \in L_0^2(\gamma)} \frac{\|R\tilde{h}\|_{2,\gamma}}{\|\tilde{h}\|_{2,\gamma}} = \beta_R, \tag{100}$$

where step (a) holds because $\tilde{\mathsf{H}} \subset L_0^2(\gamma)$.

Finally, let us show $\beta_R < 1$. By Lemma D.1, we have that for each *signed measure* $\nu \in L^2(\gamma)$ with $\nu(S) = 0$, we have

$$\int_S \left| \frac{\int_S r(x', dx)\nu(dx')}{\gamma(dx)} \right|^2 \gamma(dx) \leq \rho \int_S \left| \frac{\nu(dx)}{\gamma(dx)} \right|^2 \gamma(dx). \tag{101}$$

Define $h(x) := \nu(dx)/\gamma(dx)$, which is well-defined since $\nu \ll \gamma$. By the reversibility, we have

$$\frac{\int_S r(x', dx)\nu(dx')}{\gamma(dx)} = \int_S \frac{r(x, dx')\nu(dx')}{\gamma(dx')} = \int_S h(x')r(x, dx'),$$

Therefore, (101) can be written as

$$\int_S \left(\int_S h(x')r(x, dx') \right)^2 \gamma(dx) \leq \rho \int_S (h(x))^2 \gamma(dx),$$

for all ν such that $0 = \nu(S) = \int_S (\nu(dx)/\gamma(dx))\gamma(dx) = \int_S h(x)\gamma(dx)$. Therefore,

$$\beta_R = \sup_{h \in L_0^2(\gamma)} \frac{\|Rh\|_{2,\gamma}}{\|h\|_{2,\gamma}} \leq \rho < 1.$$

We have shown the result of (94) by showing that that $\beta_P \leq \beta_R < 1$. □

The following three lemmas are the key lemmas for proving Lemma 3 and so our main result, Theorem 1.

Lemma D.6. *Let $Z_1, Z_2...$ be i.i.d. standard normal random variables. Define $Y_i = (Z_i, ..., Z_{i+d-1})$, for $i = 1, ..., n$ and let $f_i : \mathbb{R}^d \to \mathbb{R}$ be pseudo-Lipschitz functions. Then, $\forall \epsilon \in (0, 1)$, there exists constants $K, \kappa > 0$, independent of n, ϵ, such that*

$$P\left(\left|\frac{1}{N}\sum_{i=1}^{N}\left(f_i(Y_i) - \mathbb{E}\left[f_i(Y_i)\right]\right)\right| \geq \epsilon\right) \leq Ke^{-\kappa n\epsilon^2}.$$

Proof. Without loss of generality, assume $\mathbb{E}\left[f_i(Y_i)\right] = 0$, for all $i \in [n]$. In what follows we demonstrate the upper-tail bound:

$$P\left(\frac{1}{n}\sum_{i=1}^{n} f_i(Y_i) \geq \epsilon\right) \leq Ke^{-\kappa n\epsilon^2}, \tag{102}$$

and the lower-tail bound follows similarly. Together they provide the desired result.

Using the Cramér-Chernoff method:

$$P\left(\frac{1}{n}\sum_{i=1}^{n} f_i(Y_i) \geq \epsilon\right) = P\left(e^{r\sum_{i=1}^{n} f_i(Y_i)} \geq e^{nr\epsilon}\right) \leq e^{-nr\epsilon}\mathbb{E}\left[e^{r\sum_{i=1}^{n} f_i(Y_i)}\right] \quad \forall r > 0. \tag{103}$$

Let L_i be the pseudo-Lipschitz parameters associated with functions f_i for $i = 1, ..., n$ and define $L := max_{i\in[n]} L_i$. In the following, we will show that

$$\mathbb{E}\left[e^{r\sum_{i=1}^{n} f_i(Y_i)}\right] \leq \exp\left(\kappa' n r^2\right), \text{ for } 0 < r < [5Ld\sqrt{2d + 24d^2}]^{-1}, \tag{104}$$

where κ' is any constant that satisfies $\kappa' \geq 150L^2d(d + 12d^2)$. Then plugging (104) into (103), we can obtain the desired result in (102):

$$P\left(\frac{1}{n}\sum_{i=1}^{n} f_i(Y_i) \geq \epsilon\right) \leq \exp\{-n(r\epsilon - \kappa' r^2)\}.$$

Set $r = \epsilon/(2\kappa')$, the choice that maximizes the term $(r\epsilon - \kappa' r^2)$ over r in the exponent in the above. We can ensure that $\forall \epsilon \in (0, 1)$, r falls within the region required in (104) by choosing κ' large enough.

Now we show (104). Define index sets $I_j := \{j + kd \,|\, k = 0, ..., \lfloor \frac{n-j}{d} \rfloor\}$ for $j = 1, ..., d$, let C_j denote the cardinality of I_j. We notice that for any fixed j, the Y_i's are i.i.d. for $i \in I_j$. For example, if $j = 1$ then the index set $I_1 = \{1, 1 + d, 1 + 2d, \ldots, 1 + \lfloor \frac{n-1}{d} \rfloor d\}$ and $Y_1 = (Z_1, \ldots, Z_d)$ is independent of $Y_{1+d} = (Z_{1+d}, \ldots, Z_{2d})$, which are both independent of $Y_{1+2d} = (Z_{2d+1}, \ldots, Z_{3d})$,

and so on. Also, we have $[n] = \cup_{j=1}^{d} I_j$, and $I_j \cap I_s = \emptyset$, for $j \neq s$, making the collection $I_1, I_2, \ldots, I_d$ a partition of $i \in [n]$. Therefore,

$$\sum_{i=1}^{n} f_i(Y_i) = \sum_{j=1}^{d} \sum_{i \in I_j} f_i(Y_i) = \sum_{j=1}^{d} p_j \cdot \frac{r}{p_j} \sum_{i \in I_j} f_i(Y_i),$$

where $0 < p_1, ..., p_d < 1$ are probabilities satisfying $\sum_{j=1}^{d} p_j = 1$. Using the above,

$$\mathbb{E}\left[\exp\left(r \sum_{i=1}^{n} f_i(Y_i)\right)\right] = \mathbb{E}\left[\exp\left(\sum_{j=1}^{d} p_j \cdot \frac{r}{p_j} \sum_{i \in I_j} f_i(Y_i)\right)\right] \overset{(a)}{\leq} \sum_{j=1}^{d} p_j \mathbb{E}\left[\exp\left(\frac{r}{p_j} \sum_{i \in I_j} f_i(Y_i)\right)\right]$$

$$\overset{(b)}{=} \sum_{j}^{d} p_{j=1}^{d} \prod_{i \in I_j} \mathbb{E}\left[\exp\left(\frac{r}{p_j} f_i(Y_i)\right)\right] \overset{(c)}{\leq} \sum_{j=1}^{d} p_j \exp\left(\frac{50 C_j L^2 r^2 (d + 12d^2)}{p_j^2}\right), \tag{105}$$

where step (a) follows from Jensen's inequality, step (b) from the fact that the Y_i's are independent for $i \in I_j$, and step (c) from Lemma D.3 with variance factor $\nu = 1$ and restriction

$$0 < r < [5L\sqrt{2d + 24d^2}]^{-1} \max_j p_j. \tag{106}$$

Let $p_j = \sqrt{C_j}/C$, where $C = \sum_{j=1}^{d} \sqrt{C_j}$ ensuring that $\sum_{j=1}^{d} p_j = 1$. Then, we have

$$\sum_{j=1}^{d} p_j \exp\left(\frac{50 C_j L^2 r^2 (d + 12d^2)}{p_j^2}\right) = e^{50 C^2 L^2 r^2 (d+12d^2)} \overset{(a)}{\leq} e^{150 d L^2 (d+12d^2) n r^2} \leq e^{\kappa' n r^2},$$

whenever $\kappa' \geq 150 d L^2 (d + 12d^2)$. In the above, step (a) follows from:

$$C^2 = \left(\sum_{j=1}^{d} \sqrt{C_j}\right)^2 = \sum_{j=1}^{d} C_j + \sum_{j=1}^{d} \sum_{k \neq j} \sqrt{C_j C_k} \overset{(b)}{\leq} n + d(d-1) C_1 \overset{(c)}{\leq} dn + 2d(d-1) < 3dn,$$

where step (b) holds because $C_1 = \max_{j \in [d]} C_j$ and step (c) holds because $C_1 = \lfloor \frac{n-1}{d} \rfloor + 1 \leq \frac{n}{d} + 2$.

Finally, we consider the effective region for r as required in (106). Notice that $\max_j p_j = \sqrt{C_1}/C > 1/d$. Hence, if we require $0 < r < [5Ld\sqrt{2d + 24d^2}]^{-1}$, then (106) is satisfied.

□

Lemma D.7. *Let $\{\beta_i\}_{i \in \mathbb{N}}$ be a time-homogeneous, stationary Markov chain on a bounded state space $S \subset \mathbb{R}$. Denote the transition probability measure of $\{\beta_i\}_{i \in \mathbb{N}}$ by $r(x, dy)$ and stationary probability measure by γ. Assume that the Markov chain is reversible and geometrically ergodic on $L^2(\gamma)$ as defined in Definition 2.1. Further assume that γ has finite second and fourth moment.*

Define $\{X_i\}_{i \in [n]}$ as $X_i = (\beta_i, ..., \beta_{i+d-1}) \in S^d$. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a measurable function that satisfies the pseudo-Lipschitz condition Then, for all $\epsilon \in (0, 1)$, there exists constants $K, \kappa > 0$ that are independent of n, ϵ, such that

$$P\left(\left|\frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}_\pi f\right| \geq \epsilon\right) \leq K e^{-\kappa n \epsilon^2},$$

where the probability measure π is defined as

$$\pi(dx) = \pi(dx_1, ..., dx_d) := \prod_{i=2}^{d} r(x_{i-1}, dx_i) \gamma(dx_1). \tag{107}$$

Proof. First, we split $\{X_i\}_{i\in[n]}$ into d subsequences, each containing every d^{th} term of $\{X_i\}_{i\in[n]}$, beginning from $1, 2, \ldots, d$. Label these $\{X_i^{(1)}\}_{i\in[n_1]}, ..., \{X_i^{(d)}\}_{i\in[n_d]}$ with $\{X_i^{(s)}\}_{i\in[n_s]} := \{X_{s+kd} : k = 1, ..., n_s\}$, where $n_s = \lfloor \frac{n-d-s+1}{d} \rfloor$, for $s = 1, ..., d$.

Notice that $\sum_{i=1}^n f(X_i) = \sum_{s=1}^d \sum_{i=1}^{n_s} f(X_i^{(s)})$. Using Lemma A.1, we have

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}_\pi f\right| \geq \epsilon\right) \leq \sum_{s=1}^{d} P\left(\left|\frac{1}{n_s}\sum_{i=1}^{n_s} f(X_i^{(s)}) - \mathbb{E}_\pi f\right| \geq \frac{n\epsilon}{dn_s}\right). \tag{108}$$

Without loss of generality, assume $\mathbb{E}_\pi f = 0$. In the following, we demonstrate the upper-tail bound for $\{X_i^{(1)}\}_{i\in[n_1]}$:

$$P\left(\frac{1}{n_1}\sum_{i=1}^{n_1} f(X_i^{(1)}) \geq \epsilon\right) \leq Ke^{-\kappa n_1 \epsilon^2}. \tag{109}$$

The lower-tail bound follows similarly, as do the corresponding results for $s = 2, 3, \ldots, d$. Together using (108) these provide the desired result. Using the Cramér-Chernoff method:

$$P\left(\frac{1}{n_1}\sum_{i=1}^{n_1} f(X_i^{(1)}) \geq \epsilon\right) = P\left(e^{r\sum_{i=1}^{n_1} f(X_i^{(1)})} \geq e^{rn_1\epsilon}\right) \leq e^{-rn_1\epsilon}\mathbb{E}\left[e^{r\sum_{i=1}^{n_1} f(X_i^{(1)})}\right], \quad \forall r > 0. \tag{110}$$

In what follows we will upper bound the expectation $\mathbb{E}\left[e^{r\sum_{i=1}^{n_1} f(X_i^{(1)})}\right]$ to show (109).

Let $\{\tilde{X}_i^{(1)}\}_{i\in[n_1]}$ be an independent copy of $\{X_i^{(1)}\}_{i\in[n_1]}$. By Jensen's inequality, we have

$$\mathbb{E}\left[e^{-r\sum_{i=1}^{n_1} f(\tilde{X}_i^{(1)})}\right] \geq e^{-r\mathbb{E}\left[\sum_{i=1}^{n_1} f(\tilde{X}_i^{(1)})\right]} = e^{-r\sum_{i=1}^{n_1}\mathbb{E}\left[f(\tilde{X}_i^{(1)})\right]} = 1.$$

Therefore,

$$\mathbb{E}\left[e^{r\sum_{i=1}^{n_1} f(X_i^{(1)})}\right] \leq \mathbb{E}\left[e^{r\sum_{i=1}^{n_1} f(X_i^{(1)})}\right]\mathbb{E}\left[e^{-r\sum_{i=1}^{n_1} f(\tilde{X}_i^{(1)})}\right] = \mathbb{E}\left[e^{r\sum_{i=1}^{n_1}\left(f(X_i^{(1)}) - f(\tilde{X}_i^{(1)})\right)}\right]. \tag{111}$$

Let $Z_i^{(1)} := (X_i^{(1)}, \tilde{X}_i^{(1)})$, and $g(Z_i^{(1)}) := f(X_i^{(1)}) - f(\tilde{X}_i^{(1)})$ for $i = 1, 2, \ldots, n_1$. We have shown $\mathbb{E}[\exp\{r\sum_{i=1}^{n_1} f(X_i^{(1)})\}] \leq \mathbb{E}[\exp\{r\sum_{i=1}^{n_1} g(Z_i^{(1)})\}]$ and therefore, in what follows we provide an upper bound for $\mathbb{E}[\exp\{r\sum_{i=1}^{n_1} g(Z_i^{(1)})\}]$ which can be used in (110).

We begin by demonstrating some properties of the sequence $\{Z_i^{(1)}\}_{i\in[n_1]}$, which will be used in the proof. By construction, $\{Z_i^{(1)}\}_{i\in[n_1]}$ is a time-homogeneous Markov chain on state space $D = S^d \times S^d$. Denote its marginal probability measure by μ and transition probability measure by $q(z, dz')$. In order to obtain more useful properties, it is helpful to relate $\{Z_i^{(1)}\}_{i\in[n_1]}$ to the original Markov chain $\{\beta_i\}_{i\in\mathbb{N}}$, which we have assumed to be reversible and geometrically ergodic.

The construction of $\{Z_i^{(1)}\}_{i\in[n_1]}$ can alternatively be thought of as follows. Let $\{\tilde{\beta}_i\}_{i\in\mathbb{N}}$ be an independent copy of $\{\beta_i\}_{i\in\mathbb{N}}$. Then by Lemma D.4, $\{(\beta_i, \tilde{\beta}_i)\}_{i\in\mathbb{N}}$ is reversible and geometrically ergodic. Also notice that the elements of $\{Z_i^{(1)}\}_{i\in[n_1]}$ consist of successive non-overlapping elements of $\{(\beta_i, \tilde{\beta}_i)\}_{i\in\mathbb{N}}$, same as the construction of $\{Y_i\}_{i\in\mathbb{N}}$ in Lemma D.5. Therefore, the results in Lemma D.5 imply that the marginal probability measure μ is a stationary measure of the transition probability measure $q(z, dz')$. Moreover, the linear operator Q defined as

$$Qh(z) := \int_D h(z')q(z, dz') \tag{112}$$

satisfies:

$$\beta_Q := \sup_{h \in L_0^2(\mu)} \frac{\|Qh\|_{2,\mu}}{\|h\|_{2,\mu}} < 1. \tag{113}$$

With the result $\beta_Q < 1$, we are now ready to bound $\mathbb{E}[\exp\{r\sum_{i=1}^{n_1} g(Z_i^{(1)})\}]$, where we will use a method similar to the one introduced in [15, Section 4].

Define $m(z) := \exp(rg(z)), \forall z \in D$, and so we can represent the expectation that we hope to upper bound in the following way:

$$\mathbb{E}[\exp\{r\sum_{i=1}^{n_1} g(Z_i^{(1)})\}] = \mathbb{E}[\prod_{i=1}^{n_1} m(Z_i^{(1)})]. \tag{114}$$

To provide an upper bound for (114), we first define a sequence $\{a_i\}_{i\in[n_1]}$ as $a_0 = 1$ and

$$a_i = \mathbb{E}[\exp\{r\sum_{j=1}^{i} g(Z_j^{(1)})\}] = \mathbb{E}[\prod_{j=1}^{i} m(Z_j^{(1)})], \quad 1 \le i \le n_1. \tag{115}$$

Note then that a_{n_1} equals the expectation in (114) and we have

$$\begin{aligned} a_{n_1} = \mathbb{E}[\prod_{i=1}^{n_1} m(Z_i^{(1)})] &\overset{(a)}{=} \int_{D^{n_1}} \mu(dz_1)m(z_1)\prod_{i=2}^{n_1} q(z_{i-1}, dz_i)m(z_i) \\ &= \int_{D^{n_1-1}} \mu(dz_1)m(z_1)\prod_{i=2}^{n_1-1} q(z_{i-1}, dz_i)m(z_i)\int_D q(z_{n_1-1}, dz_{n_1})m(z_{n_1}). \end{aligned} \tag{116}$$

In step (a) we use the fact that $\{Z_i^{(1)}\}_{i\in[n_1]}$ is a Markov Chain in its stationary distribution, μ, with probability transition measure $q(z, dz')$. Now, let $b_1 := \mathbb{E}_\mu m$, which is a constant value, and $m_1 := m - b_1$. Then $m(z_{n_1}) = b_1 + m_1(z_{n_1})$, and so it follows from (116),

$$\begin{aligned} a_{n_1} &= \int_{D^{n_1-1}} \mu(dz_1)m(z_1)\prod_{i=2}^{n_1-1} q(z_{i-1}, dz_i)m(z_i)\int_D q(z_{n_1-1}, dz_{n_1})\left(b_1 + m_1(z_{n_1})\right) \\ &= b_1\int_{D^{n_1-1}} \mu(dz_1)m(z_1)\prod_{i=2}^{n_1-1} q(z_{i-1}, dz_i)m(z_i) \\ &\quad + \int_{D^{n_1-1}} \mu(dz_1)m(z_1)\prod_{i=1}^{n_1-1} q(z_{i-1}, dz_i)m(z_i)\int_D q(z_{n_1-1}, dz_{n_1})m_1(z_{n_1}) \\ &\overset{(b)}{=} a_{n_1-1}b_1 + \int_{D^{n_1-1}} \mu(dz_1)m(z_1)\prod_{i=2}^{n_1-1} q(z_{i-1}, dz_i)m(z_i)Qm_1(z_{n_1-1}). \end{aligned} \tag{117}$$

Step (b) uses the definition of a_{n_1-1} given in (115) and the linear operator defined in (112). Now consider the integral in (117), which we split as in (116) in the following:

$$\begin{aligned} &\int_{D^{n_1-1}} \mu(dz_1)m(z_1)\prod_{i=2}^{n_1-1} q(z_{i-1}, dz_i)m(z_i)Qm_1(z_{n_1-1}) \\ &\quad = \int_{D^{n_1-1}} \mu(dz_1)m(z_1)\prod_{i=2}^{n-2} q(z_{i-1}, dz_i)m(z_i)\int_D q(z_{n_1-2}, dz_{n_1-1})m(z_{n_1-1})Qm_1(z_{n_1-1}). \end{aligned}$$

Then by defining $b_2 := \mathbb{E}_\mu[mQm_1]$, which is again a constant value, and $m_2 := mQm_1 - b_2$, we can represent a_{n_1} as the following sum using the above and step like those in (117).

$$a_{n_1} = a_{n_1-1}b_1 + a_{n_1-2}b_2 + \int_{D^{n_1-2}} \mu(dz_1)m(z_1)\prod_{i=2}^{n-2} q(z_{i-1}, dz_i)m(z_i)Qm_2(z_{n_1-2}). \tag{118}$$

Continuing in this way – defining constant values $b_i := \mathbb{E}_\mu[mQm_{i-1}]$ and $m_i := mQm_{i-1} - b_i$ for $i = 2, ..., n_1$, then splitting the integral as in (118) – we represent a_{n_1} recursively as $a_{n_1} = \sum_{i=1}^{n_1} b_i a_{n_1-i}$.

Again, our goal is to provide an upper bound for a_{n_1} which we can establish through the recursive relationship if we are able to upper bound $b_1, ..., b_{n_1}$. First consider b_1. Let $Z \sim \mu$.

$$b_1 = \mathbb{E}\left[e^{rg(Z)}\right] = \mathbb{E}\left[\lim_{n\to\infty}\sum_{k=0}^{n}\frac{r^k}{k!}(g(Z))^k\right].$$

Define the partial sum as $s_n := \sum_{k=0}^{n}\frac{r^k}{k!}(g(Z))^k$. Moreover, notice that

$$\sup_{z\in D}|g(z)| = \sup_{x\in S^d}\sup_{\tilde{x}\in S^d}|f(x) - f(\tilde{x})| \overset{(a)}{\leq} \sup_{x\in S^d}\sup_{\tilde{x}\in S^d} L(1 + \|x\| + \|\tilde{x}\|)\|x - \tilde{x}\| \overset{(b)}{\leq} L(1 + 2\sqrt{d}M)(2\sqrt{d}M),$$

where step (a) holds since $f(\cdot)$ is pseudo-Lipschitz with constant L and step (b) due to $\|x - \tilde{x}\| \leq \|x\| + \|\tilde{x}\|$ and the boundedness of S^d: $\|x\| \leq M\sqrt{d}$ for some constant $M > 0$ and all $x \in S^d$. Let $M_g = L(1 + 2\sqrt{d}M)(2\sqrt{d}M)$. Then for each n, $|s_n|$ is bounded by

$$\sup_{z\in D}\sum_{k=0}^{n}\frac{r^k}{k!}|g(z)|^k \leq \sum_{k=0}^{n}\frac{r^k}{k!}M_g^k \leq \sum_{k=0}^{\infty}\frac{r^k}{k!}M_g^k = e^{rM_g}.$$

Since the constant e^{rM_g} is integrable with respect to any proper probability measure, we have

$$\begin{aligned} b_1 &= \mathbb{E}\left[\lim_{n\to\infty}\sum_{k=0}^{n}\frac{r^k}{k!}(g(Z))^k\right] \overset{(a)}{=} \lim_{n\to\infty}\sum_{k=0}^{n}\frac{r^k}{k!}\mathbb{E}[(g(Z))^k] \overset{(b)}{\leq} 1 + \mathbb{E}[(g(Z))^2]\sum_{k=2}^{\infty}\frac{r^k M_g^{k-2}}{k!} \\ &= 1 + \frac{\mathbb{E}[(g(Z))^2]r^2}{2}\sum_{k=2}^{\infty}\frac{(rM_g)^{k-2}}{k!/2} \overset{(c)}{\leq} 1 + \frac{\mathbb{E}[(g(Z))^2]r^2}{2}\sum_{k=2}^{\infty}\frac{(rM_g)^{k-2}}{(k-2)!} = 1 + \frac{\mathbb{E}[(g(Z))^2]r^2}{2}e^{rM_g}, \end{aligned} \tag{119}$$

where step (a) follows the dominated convergence theorem, step (b) holds since $\mathbb{E}[g(Z)] = 0$ and $\mathbb{E}[(g(Z))^k] \leq M_g^{k-2}\mathbb{E}[(g(Z))^2]$, and step (c) holds since $(k-2)! = k!/(k(k-1)) \leq k!/2$ for $k \geq 2$ with the convention $0! = 1$.

Next we'll bound b_i for $i = 2, 3, \ldots$. To do this we first establish an upper bound on $\|m_i\|_{2,\mu}$ with the norm defined in (113).

$$\begin{aligned} \|m_i\|_{2,\mu} = \|mQm_{i-1} - b_i\|_{2,\mu} = \sqrt{\|mQm_{i-1}\|_{2,\mu}^2 - b_i^2} &\leq \|mQm_{i-1}\|_{2,\mu} \\ &\overset{(a)}{\leq} e^{rM_g}\|Qm_{i-1}\|_{2,\mu} \overset{(b)}{\leq} e^{rM_g}\beta_Q\|m_{i-1}\|_{2,\mu}. \end{aligned}$$

Step (a) holds since $\sup_{z\in D} m(z) = \sup_{z\in D} e^{rg(z)} \leq e^{rM_g}$. Step (b) holds since $E_\mu m_i = 0$, for all $i = 1, ..., n$ by construction, and so $\|Qm_i\|_{2,\mu} \leq \beta_Q\|m_i\|_{2,\mu}$ by (113). Hence, extending the above result recursively, we find

$$\|m_i\|_{2,\mu} \leq (e^{rM_g}\beta_Q)^{i-1}\|m_1\|_{2,\mu}. \tag{120}$$

Let $\langle f_1, f_2\rangle_\mu = \int f_1(z)f_2(z)\mu(dz)$. We use this to bound b_i in the following by noting that $b_i = \mathbb{E}_\mu[mQm_{i-1}] = \langle m, Qm_{i-1}\rangle_\mu$. Then,

$$b_i = \langle m, Qm_{i-1}\rangle_\mu \overset{(c)}{\leq} \|m\|_{2,\mu}\|Qm_{i-1}\|_{2,\mu} \overset{(d)}{\leq} \|m\|_{2,\mu}\beta_Q(\beta_Q e^{rM_g})^{i-2}\|m_1\|_{2,\mu} \overset{(e)}{\leq} \beta_Q(\beta_Q e^{rM_g})^{i-2}\|m_1\|_{2,\mu}^2, \tag{121}$$

where step (c) follows Cauchy-Schwarz inequality, step (d) follows from the fact that $\|Qm_{i-1}\|_{2,\mu} \leq \beta_Q\|m_{i-1}\|_{2,\mu}$ by (113) and (120), and step (e) from the fact that $\|m_1\|_{2,\mu}^2 = \|m\|_{2,\mu}^2 - b_1^2$ and therefore $\|m\|_{2,\mu}^2 \leq \|m_1\|_{2,\mu}^2$. Now let $Z \sim \mu$ and we bound $\|m_1\|_{2,\mu}^2$ as follows

$$\|m_1\|_{2,\mu}^2 = \mathbb{E}[e^{2rg(Z)}] - (\mathbb{E}[e^{rg(Z)}])^2 \overset{(f)}{\leq} 1 + \frac{4r^2\mathbb{E}[(g(Z))^2]}{2}e^{2rM_g} - e^{2r\mathbb{E}[g(Z)]} \overset{(g)}{=} \frac{4r^2\mathbb{E}[(g(Z))^2]}{2}e^{2rM_g}, \tag{122}$$

where step (f) uses similar approach to that used to bound b_1 in (119) and Jensen's inequality, and step (g) follows since $\mathbb{E}[g(Z)] = 0$.

Therefore, from (119), (121), and (122) we have

$$b_1 \leq 1 + \frac{r^2\mathbb{E}[(g(Z))^2]}{2}e^{rM_g} \quad \text{and} \quad b_i \leq \beta_Q(\beta_Q e^{rM_g})^{i-2}\frac{4r^2\mathbb{E}[(g(Z))^2]}{2}e^{2rM_g}. \tag{123}$$

Let $X, \tilde{X} \sim \pi$ independent. Notice that

$$\begin{aligned}
\mathbb{E}[(g(Z))^2] = \mathbb{E}[(f(X) - f(\tilde{X}))^2] &\overset{(a)}{\leq} L^2\mathbb{E}[((1 + \|X\| + \|\tilde{X}\|)\|X - \tilde{X}\|)^2] \\
&\overset{(b)}{\leq} 5L^2\left(2\mathbb{E}[\|X\|^2] + 2\mathbb{E}[\|X\|^4] + 4\mathbb{E}[\|X\|^2]\mathbb{E}[\|\tilde{X}\|^2]\right) \\
&\overset{(c)}{\leq} 10L^2\left(\sum_{i=1}^d \mathbb{E}[X_i^2] + d\sum_{i=1}^d \mathbb{E}[X_i^4] + 2\left(\sum_{i=1}^d \mathbb{E}[X_i^2]\right)\left(\sum_{i=1}^d \mathbb{E}[\tilde{X}_i^2]\right)\right) \\
&\overset{(d)}{=} 10L^2\left(d\mathsf{m}_2 + d^2\mathsf{m}_4 + 2d^2\mathsf{m}_2^2\right),
\end{aligned}$$

where step (a) holds since $f(\cdot)$ is pseudo-Lipschitz with constant $L > 0$, step (b) uses $\|X - \tilde{X}\| \leq \|X\| + \|\tilde{X}\|$, Lemma C.3, and the fact that X and $\tilde{X}$ are i.i.d., step (c) uses Lemma C.3, and step (d) uses the assumption of bounded second and fourth moments of the probability measure γ in the lemma statement, denoted by m_2 and m_4, respectively.

Let $\mathsf{b}^2 = 10L^2\left(d\mathsf{m}_2 + d^2\mathsf{m}_4 + 2d^2\mathsf{m}_2^2\right)$, $\mathsf{a} = \frac{1}{2}\mathsf{b}^2e^{rM_g}$, and $\alpha = \beta_Q e^{rM_g}$. Choose $r < (1-\beta_Q)/M_g$, then we have $0 < \alpha < 1$ since $1 - \beta_Q < -\ln\beta_Q$. Using these bounds and notation, (123) becomes

$$b_1 \leq 1 + \mathsf{a}r^2 \quad \text{and} \quad b_i \leq \alpha^{i-1}4\mathsf{a}r^2. \tag{124}$$

We now bound $a_1, ..., a_{n_1}$ by induction. We will show $a_i \leq [\phi(r)]^i$, where $\phi(r) = 1 + Cr^2$ for some $C \geq 4\mathsf{a}$ that is independent of i. For $i = 1$,

$$a_1 = b_1 \leq 1 + 4\mathsf{a}r^2.$$

Hence, the hypothesis $a_i \leq [\phi(r)]^i$ is true for $i = 1$. Suppose that the hypothesis is true for $i \leq n_1 - 1$, then

$$a_{n_1} = b_1a_{n_1-1} + \sum_{i=2}^{n_1} b_ia_{n_1-i} \leq (1 + 4\mathsf{a}r^2)[\phi(r)]^{n_1-1} + \sum_{i=2}^{n_1} 4\mathsf{a}r^2\alpha^{i-1}[\phi(r)]^{n_1-i}, \tag{125}$$

where the final inequality in the above follows by (124) and the inductive hypothesis. Consider only the second term on the right side of (125),

$$\sum_{i=2}^{n_1} 4\mathsf{a}r^2\alpha^{i-1}[\phi(r)]^{n_1-i} = 4\mathsf{a}r^2\alpha^{n_1-1}\sum_{i=2}^{n_1}[\alpha^{-1}\phi(r)]^{n_1-i}$$
$$= 4\mathsf{a}r^2\alpha^{n_1-1}\left(\frac{1-\left(\phi(r)\alpha^{-1}\right)^{n_1-1}}{1-\phi(r)\alpha^{-1}}\right) = 4\mathsf{a}r^2\left(\frac{\alpha[\phi(r)]^{n_1-1}-\alpha^{n_1}}{\phi(r)-\alpha}\right) \leq \frac{4\mathsf{a}r^2\alpha[\phi(r)]^{n_1-1}}{\phi(r)-\alpha},$$

where the final inequality follows since $\mathsf{a}, \alpha > 0$. Then plugging the above result into (125), we find

$$a_{n_1} \leq (1+4\mathsf{a}r^2)[\phi(r)]^{n_1-1} + \frac{4\mathsf{a}r^2\alpha[\phi(r)]^{n_1-1}}{\phi(r)-\alpha} \leq [\phi(r)]^{n_1-1}\left(1+\frac{4\mathsf{a}r^2\phi(r)}{\phi(r)-\alpha}\right) \leq [\phi(r)]^{n_1-1}\left(1+\frac{4\mathsf{a}r^2}{1-\alpha}\right),$$

where the final inequality follows since $\phi(r) \geq 1$. Therefore, let $C = 4\mathsf{a}(1-\alpha)^{-1} > 4\mathsf{a}$, since $0 < \alpha < 1$, and so $\phi(r) = 1 + 4\mathsf{a}r^2(1-\alpha)^{-1}$. It follows from the above then,

$$a_{n_1} \leq \left(1+\frac{4\mathsf{a}r^2}{1-\alpha}\right)^{n_1} = e^{n_1\ln\left(1+4\mathsf{a}r^2(1-\alpha)^{-1}\right)} \leq e^{n_1 4\mathsf{a}r^2(1-\alpha)^{-1}}, \tag{126}$$

where the final inequality uses the fact that $\ln(1+x) \leq x$ for $x \geq 0$.

Finally, from (110), (111), and the bound in (126),

$$P\left(\frac{1}{n_1}\sum_{i=1}^{n_1} f(X_i^{(1)}) \geq \epsilon\right) \leq \exp\left(-n_1\left(r\epsilon - 4\mathsf{a}r^2(1-\alpha)^{-1}\right)\right) \overset{(a)}{=} \exp\left(-n_1\left(r\epsilon - \frac{2\mathsf{b}^2r^2e^{rM_g}}{1-\beta_Qe^{rM_g}}\right)\right),$$

where step (a) follows from the fact that $\mathsf{a} = \mathsf{b}^2e^{rM_g}/2$ and $\alpha = \beta_Qe^{rM_g}$. Now let us consider the term in the exponent in the above for the cases where (i) $\mathsf{b}^2 \geq M_g$ and (ii) $\mathsf{b}^2 < M_g$ separately, and then combine the results in the two cases to obtain a desired bound for all $\epsilon \in (0,1)$.

First (i) $\mathsf{b}^2 \geq M_g$. Notice for every $0 < \epsilon < 4\mathsf{b}^2/M_g$, if we let $r = (1-\beta_Q)\epsilon/(4\mathsf{b}^2)$, then $r < (1-\beta_Q)/M_g$ as required before. We show whenever $0 < \epsilon \leq \mathsf{b}^2/M_g$, we can obtain a desired bound. Then the condition in the lemma statement, $\epsilon \in (0,1)$, falls within this effective region.

$$r\epsilon - \frac{2\mathsf{b}^2r^2e^{rM_g}}{1-\beta_Qe^{rM_g}} \overset{(a)}{=} \frac{(1-\beta_Q)\epsilon^2}{4\mathsf{b}^2} - \frac{(1-\beta_Q)^2\epsilon^2}{8\mathsf{b}^2}\cdot\frac{\exp\left(\frac{(1-\beta_Q)M_g\epsilon}{4\mathsf{b}^2}\right)}{1-\beta_Q\exp\left(\frac{(1-\beta_Q)M_g\epsilon}{4\mathsf{b}^2}\right)}$$
$$= \frac{(1-\beta_Q)\epsilon^2}{8\mathsf{b}^2}\left(1-\frac{\exp\left(\frac{(1-\beta_Q)M_g\epsilon}{4\mathsf{b}^2}\right)-1}{1-\beta_Q\exp\left(\frac{(1-\beta_Q)M_g\epsilon}{4\mathsf{b}^2}\right)}\right) \overset{(b)}{\geq} \frac{(1-\beta_Q)\epsilon^2}{8\mathsf{b}^2}\left(1-\frac{\frac{(1-\beta_Q)M_g\epsilon}{3\mathsf{b}^2}}{1-\beta_Q\left(1+\frac{(1-\beta_Q)M_g\epsilon}{3\mathsf{b}^2}\right)}\right)$$
$$= \frac{(1-\beta_Q)\epsilon^2}{8\mathsf{b}^2}\left(1-\frac{M_g\epsilon}{2\mathsf{b}^2+(\mathsf{b}^2-\beta_QM_g\epsilon)}\right) \overset{(c)}{\geq} \frac{(1-\beta_Q)\epsilon^2}{8\mathsf{b}^2}\left(1-\frac{\epsilon}{2}\right), \qquad \text{for } 0 < \epsilon \leq \mathsf{b}^2/M_g.$$

In the above, step (a) by plugging in $r = (1-\beta_Q)\epsilon/(4\mathsf{b}^2)$, step (b) holds since $e^x \leq 1 + 4x/3$ for $x \leq 1/2$, and step (c) holds since $\epsilon \leq \mathsf{b}^2/M_g$, so $(\mathsf{b}^2 - \beta_QM_g\epsilon) > 0$, and the fact $\mathsf{b}^2 \geq M_g$.

Next consider (ii) $\mathsf{b}^2 < M_g$. In this case, set $r = (1-\beta_Q)\epsilon/(4M_g)$. Hence, $r < (1-\beta_Q)/M_g$ for

$\epsilon \in (0,1)$, and then

$$r\epsilon - \frac{2\mathsf{b}^2 r^2 e^{rM_g}}{1-\beta_Q e^{rM_g}} \overset{(a)}{>} r\epsilon - \frac{2M_g r^2 e^{rM_g}}{1-\beta_Q e^{rM_g}} \overset{(b)}{=} \frac{(1-\beta_Q)\epsilon^2}{4M_g} - \frac{(1-\beta_Q)^2\epsilon^2}{8M_g} \cdot \frac{\exp\left(\frac{(1-\beta_Q)\epsilon}{4}\right)}{1-\beta_Q \exp\left(\frac{(1-\beta_Q)\epsilon}{4}\right)}$$

$$= \frac{(1-\beta_Q)\epsilon^2}{8M_g}\left(1 - \frac{\exp\left(\frac{(1-\beta_Q)\epsilon}{4}\right)-1}{1-\beta_Q \exp\left(\frac{(1-\beta_Q)\epsilon}{4}\right)}\right) \overset{(c)}{\geq} \frac{(1-\beta_Q)\epsilon^2}{8M_g}\left(1-\frac{\epsilon}{2}\right), \qquad \text{for } 0<\epsilon\leq 1.$$

In the above, step (a) holds since $\mathsf{b}^2 < M_g$, step (b) by plugging in $r = (1-\beta_Q)\epsilon/(4M_g)$, and step (c) follows similar calculation as in case (i).

Combining the results in the two cases, we conclude that for all $\epsilon \in (0,1)$, the following is satisfied:

$$r\epsilon - \frac{2\mathsf{b}^2 r^2 e^{rM_g}}{1-\beta_Q e^{rM_g}} \geq \frac{(1-\beta_Q)\epsilon^2}{8\max(M_g, \mathsf{b}^2)}\left(1-\frac{\epsilon}{2}\right).$$

Hence, for $\epsilon \in (0,1)$,

$$P\left(\frac{1}{n_1}\sum_{i=1}^{n_1} f(X_i^{(1)}) \geq \epsilon\right) \leq \exp\left(\frac{-(1-\beta_Q)n_1\epsilon^2}{8\max(M_g,\mathsf{b}^2)}\left(1-\frac{\epsilon}{2}\right)\right) \leq \exp\left(\frac{-(1-\beta_Q)n_1\epsilon^2}{16\max(M_g,\mathsf{b}^2)}\right). \tag{127}$$

Therefore, using (108) and the fact that we can show a similar result for each $s = 2, 3, \ldots, d$, we have for $\epsilon \in (0,1)$,

$$\begin{aligned} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}_\pi f\right| \geq \epsilon\right) &\leq \sum_{s=1}^{d} P\left(\left|\frac{1}{n_s}\sum_{i=1}^{n_s} f(X_i^{(s)}) - \mathbb{E}_\pi f\right| \geq \frac{n\epsilon}{dn_s}\right) \\ &\overset{(a)}{\leq} \sum_{s=1}^{d} \exp\left(\frac{-(1-\beta_Q)n^2\epsilon^2}{16 n_s \max(M_g,\mathsf{b}^2)}\right) \overset{(b)}{\leq} d\exp\left(\frac{-(1-\beta_Q)n\epsilon^2}{16d\max(M_g,\mathsf{b}^2)}\right), \end{aligned} \tag{128}$$

where step (a) follows (127) and step (b) holds since $n/n_s \geq n/n_1 = n/(\lfloor n/d \rfloor - 1) \geq d$, for all $s \in [d]$. To complete the proof, we recall that $\mathsf{b}^2 = 10L^2\left(d\mathsf{m}_2 + d^2\mathsf{m}_4 + 2d^2\mathsf{m}_2^2\right)$ and $M_g = L(1+2\sqrt{d}M)(2\sqrt{d}M)$.

□

Lemma D.8. *Consider the sequence $\{X_i\}_{i\in[n]}$ and the probability measure π defined in Lemma D.7. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a bounded function such that $\|f\|_\infty \leq M_f$ and $\|f\|_{2,\pi}^2 \leq b_f^2$. Then, for all $\epsilon \in (0,1)$, there exists constants $K, \kappa > 0$ that are independent of n, ϵ, such that*

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}_\pi f\right| \geq \epsilon\right) \leq Ke^{-\kappa n\epsilon^2}.$$

Proof. We split $\{X_i\}_{i\in[n]}$ into d subsequences $\{X_i^{(s)}\}_{i\in[n_s]}$, $s = 1, ..., d$, in the same way as in the proof of Lemma D.7, and therefore

$$P\left(\left|\frac{1}{n}\sum_{i=i}^{n} f(X_i) - \mathbb{E}_\pi f\right| \geq \epsilon\right) \leq \sum_{s=1}^{d} P\left(\left|\frac{1}{n_s}\sum_{i=i}^{n_s} f(X_i^{(s)}) - \mathbb{E}_\pi f\right| \geq \frac{n\epsilon}{dn_s}\right). \tag{129}$$

Without loss of generality, assume $\mathbb{E}_\pi f = 0$, and we demonstrate the upper-tail bound for $\{X_i^{(1)}\}_{i\in[n_1]}$:

$$P\left(\frac{1}{n_1}\sum_{i=1}^{n_1} f(X_i^{(1)}) \geq \epsilon\right) \leq Ke^{-\kappa n_1 \epsilon^2}. \tag{130}$$

The lower-tail bound, as well as the corresponding results for $s = 2, 3..., d$, follow similarly. Then plugging these bounds into (129), we can obtain the desired result.

Recall that the original Markov chain $\{\beta_i\}_{i\in\mathbb{N}}$ is assumed to be reversible and geometrically ergodic. Also notice that the elements in $\{X_i^{(1)}\}_{i\in[n_1]}$ consists of successive and non-overlapping elements of $\{\beta_i\}_{i\in\mathbb{N}}$. Therefore, applying Lemma D.5 to $\{X_i^{(1)}\}_{i\in[n_1]}$, the marginal probability measure of $\{X_i^{(1)}\}_{i\in[n_1]}$ defined in (107) is a stationary probability measure for the transition probability measure of $\{X_1^{(1)}\}_{i\in[n_1]}$, denoted by $p(x, dx')$. Moreover, the linear operator P defined as $Ph(x) := \int_{S^d} h(x')p(x, dx')$ satisfies $\beta_P := \sup_{h\in L_0^2(\pi)} \frac{\|Ph\|_{2,\pi}}{\|h\|_{2,\pi}} < 1$. Then, by similar demonstration as in (114) - (127), with the boundedness of the function in this case playing the role of the boundedness of the state space, we have

$$P\left(\frac{1}{n_1}\sum_{i=1}^{n_1} f(X_i^{(1)}) \geq \epsilon\right) \leq \exp\left(-\frac{(1-\beta_P)n_1\epsilon^2}{16\max(b_f^2, M_f)}\right), \quad \text{for } \epsilon \in (0, 1).$$

Then by similar demonstration as in (128), we have

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i)\right| \geq \epsilon\right) \leq d\exp\left(-\frac{(1-\beta_P)n\epsilon^2}{16d\max(b_f^2, M_f)}\right), \quad \text{for } \epsilon \in (0, 1). \tag{131}$$

□

References

[1] 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.

[2] 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.

[3] 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.

[4] F. Krzakala, M. Mézard, F. Sausset, Y. Sun, and L. Zdeborová, "Probabilistic reconstruction in compressed sensing: Algorithms, phase diagrams, and threshold achieving matrices," *J. Stat. Mech. – Theory E.*, vol. 2012, p. P08009, Aug. 2012.

[5] 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.

[6] 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.

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

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

[9] C. Rush and R. Venkataramanan, "Finite sample analysis of approximate message passing," *Proc. IEEE Int. Symp. Inf. Theory*, June 2015. Full version: `https://arxiv.org/abs/1606.01800`.

[10] A. Javanmard and A. Montanari, "State evolution for general approximate message passing algorithms, with applications to spatial coupling," *Arxiv preprint arXiv:1211.5164*, Dec. 2012.

[11] 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`.

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

[13] G. O. Roberts and J. S. Rosenthal, "Geometric ergodicity and hybrid markov chains," *Electronic Communications in Probability*, pp. 13–25, 1997.

[14] S. Janson, "Large deviations for sums of partly dependent random variables," *Random Structures Algorithms*, vol. 24, pp. 234–248, Sept. 2004.

[15] P. Lezaud, "Chernoff-type bound for finite markov chains," *The Annals of Applied Probability*, pp. 849–867, 1998.