A Additional Experiments

Based on the feedback from the reviewers, we perform the following additional experiments which explore the robustness of the choice of buffer size in SGD−RER, choice of step sizes for GLMtron and the behavior of the said algorithms with heavy tailed noise with a similar setup as in Section 7.

We first perform an experimental study about the robustness of SGD−RER to the choice of buffer size in Figure 3a. Notice that the performance remains the same for a large range of buffer sizes (100 from to 2000). However the performance degrades when the buffer size is too large (∼10000). We believe this is the case since the number of buffers decreases as the buffer size increases and the output is averaged over too few number of iterates (In the case of B = 10000, the final output is just an average of 10 iterates).

Next, we consider the range of step sizes which allow GLMtron to converge in Figure 3b in order to supplement the discussion in Sections 7 and 3 regarding GLMtron requiring smaller step-sizes. In smooth convex optimization, it is typically the case that the iterates diverge to infinity if the step size is chosen to be too large. Theoretically, this largest step-size is $\frac{2}{\lambda}$ where $\lambda$ is the largest eigenvalue of the Hessian. In the case of GLMtron, it was experimentally observed that if the step size was chosen to be about 1.5 times the step size reported in Section 7, the iterates diverged. Quasi Newton method essentially normalizes the gradient with the inverse of the Hessian (or rather an approximation of the Hessian) in order to let it converge faster with large step sizes.

In Figure 4, we consider the same system as in Section 7 but with heavy tailed noise given by the student t distribution (scale $\nu = 4.1$) so that the 4-th moment exists but higher moments do not. The typical behavior of Forward SGD, SGD-ER, SGD-RER and Quasi Newton methods seems to be similar to that observed in the Sub-Gaussian noise case. However, GLMtron requires much smaller step sizes to ensure convergence and hence it takes much longer. We believe that the reason for this is related to the explanation given above for GLMtron step sizes. The largest eigenvalue of the Hessian depends on the quantity $X_i^T X_i$, which can be much larger than in the sub-gaussian case and hence we need to pick much smaller step sizes. However, further research is needed to confirm this.
phenomenon. We also note that we did not provide theoretical guarantees for SGD-RER in the heavy
tailed noise case. But it is still seen to typically perform very well.

B Analysis of the Quasi Newton Method

In this Section, we give the proofs of Theorems 1 and 2. Let \( e_1, \ldots, e_d \) be the standard basis vectors
for \( \mathbb{R}^d \). We will analyze the Quasi Newton method row by row.

**Definition 1.** Given a matrix \( A = [a_1, a_2, \ldots, a_d]^T \), let \( \mathcal{R}(A) = \{a_1, \ldots, a_d\} \) denote the set of vectors that are (transposes of) rows of the matrix \( A \). We use \( a^\top \) to represent a generic row of \( A \).

Follow Definition 1, we will consider the estimation of the \( i \)-th row \( a^\top_i \). Consider the gradient
\[ \nabla L_{\text{prox}}^{(i)} : \mathbb{R}^d \to \mathbb{R}^d \]
given by:
\[ \nabla L_{\text{prox}}^{(i)}(a) := \frac{1}{T} \sum_{t=0}^{T-1} (\phi(\langle a, X_t \rangle) - (e_i, X_{t+1})) X_t. \]

We can write
\[ \nabla L_{\text{prox}}^{(i)}(a) = \frac{1}{T} \sum_{t=0}^{T-1} (\phi(\langle a, X_t \rangle) - \phi(\langle a^\top_i, X_t \rangle)) X_t - \langle \eta_t, e_i \rangle X_t \]
\[ = \frac{1}{T} \sum_{t=0}^{T-1} \phi'(\beta_t)(a - a^\top_i, X_t)X_t - \langle \eta_t, e_i \rangle X_t \]
\[ = \tilde{K}_{a,i}(a - a^\top_i) - \tilde{N}_i \]
\[ \tag{8} \]

Where \( \beta_t \) exist because of the mean value theorem. We can make sense of \( \beta_t \) even when \( \phi \) is only weakly differentiable and check that the proof below still follows. Here, \( \tilde{K}_{a,i} := \frac{1}{T} \sum_{t=0}^{T-1} \phi'(\beta_t)X_tX^\top_t \) and \( \tilde{N}_i := \frac{1}{T} \sum_{t=0}^{T-1} \langle \eta_t, e_i \rangle X_t \). In the first step we have used the dynamics in Equation 1 to write down \( X_{t+1} \) in terms of \( X_t \) and \( \eta_t \).

We now define \( \tilde{G} \in \mathbb{R}^{d \times d} \) by \( \tilde{G} := \frac{1}{T} \sum_{t=0}^{T-1} X_tX^\top_t \). From the fact that \( \zeta \leq \phi'(\beta_t) \leq 1 \), we note that for every \( a \in \mathbb{R}^d \):
\[ \tilde{G} \succeq \tilde{K}_{a,i} \succeq \zeta \tilde{G} \]
\[ \tag{9} \]

Now consider the Quasi Newton Step given in Algorithm 1:
\[ a_i(l + 1) = a_i(l) - 2\gamma \tilde{G}^{-1}\nabla L_{\text{prox}}^{(i)}(a_i(l)) \]

Denoting \( \tilde{K}_{a, (l), i} \) by \( \tilde{K}_{i, l} \) we use Equation 8 to conclude:
\[ a_i(l + 1) - a_i^\top \leq a_i(l) - a_i^\top - 2\gamma \tilde{G}^{-1}\tilde{K}_{a,i}(a_i(l + 1) - a_i^\top) + 2\gamma \tilde{G}^{-1}\tilde{N}_i \]
\[ \implies \sqrt{\tilde{G}}(a_i(l + 1) - a_i^\top) = (1 - 2\gamma \tilde{G}^{-1/2}\tilde{K}_{i, l}\tilde{G}^{-1/2})\sqrt{\tilde{G}}(a_i(l) - a_i^\top) + 2\gamma \tilde{G}^{-1/2}\tilde{N}_i \]
\[ \tag{10} \]

Picking \( \gamma < \frac{1}{2} \), we conclude from Equation 9 that:
\[ (1 - 2\gamma)I \preceq I - 2\gamma \tilde{G}^{-1/2}\tilde{K}_{i, l}\tilde{G}^{-1/2} \preceq (1 - 2\gamma\zeta)I \]

We use the equation above in Equation 10 along with triangle inequality to conclude:
\[ \| \sqrt{\tilde{G}}(a_i(l + 1) - a_i^\top) \| \leq (1 - 2\gamma\zeta)\| \sqrt{\tilde{G}}(a_i(l) - a_i^\top) \| + 2\gamma\| \tilde{G}^{-1/2}\tilde{N}_i \| \]

Unrolling the recursion above, we obtain that:
\[ \| \sqrt{\tilde{G}}(a_i(m) - a_i^\top) \| \leq (1 - 2\gamma\zeta)^m\| \sqrt{\tilde{G}}(a_i(0) - a_i^\top) \| + \sum_{i=0}^{m-1} 2\gamma(1 - 2\gamma\zeta)^i\| \tilde{G}^{-1/2}\tilde{N}_i \|
\]
\[ \leq (1 - 2\gamma\zeta)^m\| \sqrt{\tilde{G}}(a_i(0) - a_i^\top) \| + \frac{1}{\zeta}\| \tilde{G}^{-1/2}\tilde{N}_i \| \]
Letting $A_m$ be the matrix with rows $a_i(m)$, we conclude:

$$
\|A_m - A^*\|_F^2 \leq 2 \lambda_{\max}(\hat{G}) (1 - 2\gamma\zeta)^{2m}\|A_0 - A^*\|_F^2 + \frac{2}{\zeta^2 \lambda_{\min}(\hat{G})} \sum_{i=1}^d \|\hat{G}^{-1/2} \hat{N}_i\|^2
$$

(11)

Proof of Theorems 1 and 2 follow once we provide high probability bounds for various terms in Equation (11). We will first define some notation. Let $S(\rho, T) := \sum_{t=0}^T \rho^{T-t}$. For $R, \kappa > 0$, we define the following events

1. $D_T(R) := \{ \sup_{0 \leq t \leq T} \|X_t\|^2 \leq R \}$
2. $E_T(\kappa) := \{ \hat{G} \geq \frac{2}{R^2} \}$
3. $D_T(R, \kappa) := D_T(R) \cap E_T(\kappa)$

**Lemma 1.** Under the Assumptions of Theorem 1, suppose $\delta \in (0, 1/2)$ and take $R = C_2 C_\eta S(\rho, T)^2 \operatorname{dlog}(\frac{T^2}{\delta})$, $\kappa = 2$, and $T \geq C_3 (d \log(\frac{R}{\sigma^2}) + \log(\frac{1}{\delta}))$

$$
\mathbb{P} \left( \sum_{i=1}^d \|\hat{G}^{-1/2} \hat{N}_i\|^2 \leq \frac{C_2}{T} [d^2 \log(1 + \frac{R}{\sigma^2}) + d \log(\frac{1}{\delta})] \cap D_T(R, \kappa) \right) \geq 1 - 2d\delta
$$

Where $C, C_\eta$ are constants depending only on $C_\eta$.

We refer to Section F for the proof.

**Lemma 2.** Suppose the Assumptions of Theorem 2 hold. There exist universal constants $C, C_1, c_0 > 0$ such that whenever $\delta \in (0, 1/2)$, $R := \frac{4T \delta C_2 \sigma^2}{(1-\rho)^2 \delta}$ and $T \geq C \log(\frac{1}{\delta}) \log(\frac{R}{\sigma^2}) \max \left( \frac{4C_2^2 M_4}{(1-\rho)^2 (1-\rho^2) \sigma^2}, \frac{\log \left( \frac{R C \eta}{\sigma^2} \right)}{\log \left( \frac{2}{\rho} \right)} \right)$, we have with probability at-least $1 - \delta$:

1. $\hat{G} \geq c_0 G$

2. $\lambda_{\max}(\hat{G}) \leq R$

Where $G := \mathbb{E}X_t X_t^\top$

We give the proof of this lemma in Section H.

**B.1 Proof of Theorem 1**

**Proof.** Note that $R$ in Lemma 1 is the same as $R^*$ in the statement of Theorem 1. We combine the result of Lemma 1 with the Equation (11). Under the event $D_T(R^*, \kappa)$, we have $\frac{\lambda_{\max}(\hat{G})}{\lambda_{\min}(\hat{G})} \leq \frac{2R}{\sigma^2}$ almost surely. Hence the result follows.

**B.2 Proof of Theorem 2**

**Proof.** We again begin with Equation (11).

$$
\|A_m - A^*\|_F^2 \leq 2 \frac{\lambda_{\max}(\hat{G})}{\lambda_{\min}(\hat{G})} (1 - 2\gamma\zeta)^{2m}\|A_0 - A^*\|_F^2 + \frac{2}{\zeta^2 \lambda_{\min}(\hat{G})} \sum_{i=1}^d \|\hat{G}^{-1/2} \hat{N}_i\|^2
$$
Let the event described in Lemma 2 be \( \mathcal{W} \). Under this event, \( \lambda_{\min} \left( \hat{G} \right) \geq c_0 \lambda_{\min} \left( G \right), \lambda_{\max} \left( G \right) \leq R \) and \( \hat{G}^{-1} \preceq \frac{1}{c_0} G^{-1} \). Therefore, we conclude:

\[
\|A_m - A^\ast\|_F^2 \leq 2 \frac{R}{c_0 \lambda_{\min} \left( G \right)} (1 - 2 \gamma \zeta)^{2m} \|A_0 - A^\ast\|_F^2 + \frac{2}{\zeta^2 c_0 \lambda_{\min} \left( G \right)} \sum_{i=1}^{d} \hat{N}_i^\top \hat{G}^{-1} \hat{N}_i
\]

Now, consider \( \mathbb{E}[\|A_m - A^\ast\|_F^2] \mathbb{I}(\mathcal{W}) \). To conclude the result of the Theorem, we will show that \( \mathbb{E}[\hat{N}_i^\top \hat{G}^{-1} \hat{N}_i] \mathbb{I}(\mathcal{W}) \leq \frac{d \sigma^2}{T} \). Indeed:

\[
\mathbb{E}[\hat{N}_i^\top \hat{G}^{-1} \hat{N}_i] \mathbb{I}(\mathcal{W}) \leq \mathbb{E}[\hat{N}_i^\top \hat{G}^{-1} \hat{N}_i] = \frac{1}{T} \sum_{t=s}^{T-1} \mathbb{E}[\langle \eta_t, e_i \rangle \langle \eta_s, e_i \rangle X_i^\top \hat{G}^{-1} X_s]
\]

Observe that whenever \( t > s \), \( \eta_t \) has mean zero and is independent of \( X_t, X_s \) and \( \eta_s \). Therefore, \( \mathbb{E}[\langle \eta_t, e_i \rangle \langle \eta_s, e_i \rangle X_i^\top \hat{G}^{-1} X_s = 0 \). When \( t = s \), we conclude using the fact that \( \mathbb{E}X_t X_i^\top = G \) that \( \mathbb{E}[\langle \eta_t, e_i \rangle^2 X_i^\top \hat{G}^{-1} X_i = d \sigma^2 \). Using the calculations above and Equation (12), we conclude that:

\[
\mathbb{E}[\|A_m - A^\ast\|_F^2] \mathbb{I}(\mathcal{W}) \leq \frac{2d^2 \sigma^2}{c_0 \zeta^2 T \lambda_{\min} \left( G \right)} + \frac{2R}{c_0 \lambda_{\min} \left( G \right)} (1 - 2 \gamma \zeta)^{2m} \|A_0 - A^\ast\|_F^2
\]

### C  Analysis of SGD-RER

In this section we consider the following \((X_0, X_1, \ldots, X_T)\) to be a stationary sequence from NLDS \( (A^\ast, \mu, \phi) \). We make Assumptions 1, 2, 3 and 5. We aim to analyze Algorithm 2 and then prove Theorem 3.

The data is divided into buffers of size \( B \) and the buffers have a gap of size \( u \) in between them. Let \( S = B + u \). The algorithm runs SGD with respect to the proxy loss \( L_{\text{prox}} \) in the order described in Section 6. Formally, let \( X_j^t \equiv X_{it+S+j} \) denote the \( j \)-th sample in buffer \( t \). We denote, for \( 0 \leq i \leq B-1 \), \( X_i^t \equiv X_i^{(s-1)-i} \), i.e., the \( i \)-th processed sample in buffer \( t \). We use similar notation for noise samples i.e., \( \eta^t_j \equiv \eta_{s+j} \) and \( \eta^t_{-j} \equiv \eta_{(s-1)-j} \).

The algorithm iterates are denoted by the sequence \((A^t_i : 0 \leq t \leq N-1, 0 \leq i \leq B-1)\) where \( A^t_i \) denotes the iterate obtained after processing \( i \)-th (reversed) sample in buffer \( t \) and \( N = T/S \) is the total number of buffers. Note that we enumerate buffers from 0, 1, \ldots, \( N \). Formally:

\[
A_t^{i+1} = A_t^{i-1} - 2 \gamma \left( \phi(A_t^{i-1} X_{-i}^{t-1}) - X_{-i}^{t-1}(i-1) \right) X_{-i}^{t-1,\top}
\]

for \( 1 \leq t \leq N, 0 \leq i \leq B-1 \) and we set \( A_0^i = A_B^t \) with \( A_0^0 = A_0 \).

The algorithm outputs the tail-averaged iterate at the end of each buffer \( t \): \( \hat{A}_{i_0,t} = \frac{1}{t-t_0} \sum_{t=t_0+1}^{t} A_{i_0,t}^{-1} \) where \( 1 \leq t \leq N \) and \( 0 \leq t_0 \leq t-1 \).

### C.1 Proof Strategy

The proof of Theorem 3 involves many intricate steps. Therefore, we give a detailed overview about the proof below.

1. In Section C.2, we first construct a fictitious coupled process \( \hat{X}_t \) such that for every data point within a buffer \( t \), \( \|\hat{X}_t - X_t\| \lesssim \frac{\alpha}{t^{1/2}} \) for some fixed \( \alpha > 0 \) chosen arbitrarily beforehand. We then show that the iterates \( A_t^i \) which are generated with SGD - RER is run with the coupled process \( \hat{X}_t \) is very close to the actual iterate \( A_t^i \). The coupled process has the advantage that the data in the successive buffers are independent. We then only deal with the coupled iterates \( \hat{A}_t^i \) and appeal to Lemma 4 to obtain bounds for \( A_t^i \).
2. In Section C.5 we give the bias variance decomposition as is standard in the linear regression literature. We extend it to the non-linear case using the mean value theorem and treat the buffers as independent data samples. Here, the matrices $H_{0,t-1}^t$ defined on the data in buffer $s$ ‘contracts’ the norm of $A^*_{t} - A^*$ giving the ‘bias term’ whereas the noise $σ^t_{i}$ presents the ‘variance’ term which is due to the inherent uncertainly in the estimation problem.

3. We refer to Section [1] where we develop the contraction properties of the matrices $\prod_{s=0}^{t} H_{0,t-1}^t$ where we show that $\| \prod_{s=0}^{t} H_{0,t-1}^t \| \lesssim (1 - \zeta \gamma B \lambda_{\min}(G))^t$ in Theorem 10 after developing some probabilistic results regarding NLDS($A^*, \mu, \phi$). This allows us show exponential decay of the bias.

4. We then turn to the squared variance term in Section C.6. We decompose it into ‘diagonal terms’ with non-zero expectation and ‘cross terms’ with a vanishing expectation. Bounding the diagonal term is straightforward using standard recursive arguments and we give the bound in Claim 1.

5. The ‘cross terms’ which vanish in expectation in the linear case, do not because of the coupling introduced by the non-linearities through the iterates (see Section 6 for a short description). However, we establish ‘algorithmic stability’ in Section C.7 where we show that the iterates depend only weakly on each of the noise vectors and hence the cross terms have expectation very close to zero. More specifically, we use the novel idea of re-sampling the whole trajectory ($\tilde{X}_t$) by re-sampling one noise vector only and show that the iterates of SGD – RER are not affected much.

6. We use the ‘algorithmic stability’ bounds to bound the cross terms in Sections C.8. We then combine the bounds to obtain the bound on the ‘variance term’.

7. Finally, we analyze the tail averaged output in Sections C.10, C.11 and C.12 and then combine these ingredients to prove Theorem 3.

C.2 Basic Notations and Coupled Process

Definition 2 (Coupled process). Given the co-variates $\{X_\tau : \tau = 0, 1, \cdots, T\}$ and noise $\{\eta_\tau : \tau = 1, 2, \cdots, T\}$, we define $\{\tilde{X}_\tau : \tau = 0, 1, \cdots, T\}$ as follows:

1. For each buffer $t$ generate, independently of everything else, $\tilde{X}_0^t \sim \pi$, the stationary distribution of the NLDS($A^*, \mu, \phi$) model.

2. Then, each buffer has the same recursion as eq (11):

$$X_{i+1}^t = \phi(A^*X_i^t) + \eta_i^t, i = 0, 1, \cdots, S - 1,$$

where the noise vectors as same as in the actual process $\{X_\tau\}$.

Lemma 3 (Coupling Lemma). Under Assumption 4 for any buffer $t$, we have $\|X_i^t - \tilde{X}_i^t\| \leq C_{\rho^t}, \|X_0^t - \tilde{X}_0^t\|$, a.s. Hence

$$\|X_i^t X_i^{T,T} - \tilde{X}_i^t \tilde{X}_i^{T,T}\| \leq 2(\sup_{\tau \leq T} \|X_\tau\|) \|X_i^t - \tilde{X}_i^t\| \leq 4 \sup_{\tau \leq T} \|X_\tau\|^2 C_{\rho^t} (15)$$

With the above notation, we can write (13) in terms of a generic row (say row $r$) $a_{i+1}^{r-1,T}$ of $A_{i+1}^{r-1}$ as follows. Let $\varepsilon_{i-1}^{r-1}$ denote the element of $\eta_{i-1}^{r-1}$ in row $r$. Similarly let $a^{r-1,T} \equiv (a^r)^T$ denote the row $r$ of $A^*$. Then

$$a_{i+1}^{r-1,T} = a_{i}^{r-1,T} - 2\gamma \left( \phi(X_{i}^{r-1,T} a_{i}^{r-1}) - \phi(X_{i}^{r-1,T} a^r) \right) X_{i-1}^{r-1,T} + 2\gamma \varepsilon_{i-1}^{r-1} X_{i-1}^{r-1,T} (16)$$

Now, by the mean value theorem we can write

$$\phi(X_{i-1}^{r-1,T} a_{i}^{r-1}) - \phi(X_{i-1}^{r-1,T} a^r) = \phi'(\xi_{i-1}^{r-1})(a_{i}^{r-1} - a^r)^T X_{i-1}^{r-1}$$

where $\xi_{i-1}^{r-1}$ lies between $X_{i-1}^{r-1,T} a_{i}^{r-1}$ and $X_{i-1}^{r-1,T} a^r$. Hence we obtain

$$(a_{i+1}^{r-1} - a^r)^T = (a_{i}^{r-1} - a^r)^T (I - 2\gamma \phi'(\xi_{i-1}^{r-1})X_{i-1}^{r-1} X_{i-1}^{r-1,T}) + 2\gamma \varepsilon_{i-1}^{r-1} X_{i-1}^{r-1,T} (18)$$
We make Assumptions 1, 2, 3 and 5 throughout. We set the parameters for SGD – RER iterates run with the coupled process \( \tilde{X}_i^t \) as follows:

\[
\tilde{A}_0^0 = A_0^0
\]

\[
\tilde{A}_{i+1}^t = \tilde{A}_i^{t-1} - 2\gamma \left( \phi(\tilde{A}_i^{t-1} \tilde{X}_i^{t-1}) - \tilde{X}_i^{t-1} \right) \tilde{X}_i^{t-1,\top}
\]  \hspace{1cm} (19)

Using Lemma 3, we can show that \( \tilde{A}_i^t \approx A_i^t \). Note that successive buffers for the iterates \( \tilde{A}_i^t \) are actually independent. We state the following lemma which shows that we can indeed just analyze \( \tilde{A}_i^t \) and then from this obtain error bounds for \( A_i^t \). We refer to Section H for the proof.

**Lemma 4.** Suppose \( \gamma < \frac{1}{2R_{\text{max}}} \). we have for every \( t \in [N] \) and \( i \in [B] \).

\[
\|\tilde{a}_i^t - a_i^t\| \leq (16\gamma^2 R_{\text{max}}^2 + 8\gamma R_{\text{max}}^2) \rho^t
\]

We note that we can just analyze the iterates \( \tilde{A}_i^t \) and then use Lemma 4 to infer error bounds for \( A_i^t \). Henceforth, we will only consider \( \tilde{A}_i^t \).

Before proceeding, we will set up some notation.

### C.3 Notations and Events

We define the following notations. Let \( R > 0 \) to be decided later.

\[
X_i^t = X_{(S-1)-i}^t, \quad 0 \leq i \leq S - 1, \quad \phi'(\tilde{\varepsilon}_i^t) = \frac{\phi(\tilde{a}_i^t, \tilde{X}_i^t) - \phi(a^*_i, \tilde{X}_i^t)}{(\tilde{a}_i^t - a^*_i)}
\]

\[
\bar{P}_{i,j}^t = \left( I - 2\gamma \phi'(\tilde{\varepsilon}_i^t) \tilde{X}_i^t \tilde{X}_i^{\top} \right), \quad \bar{H}_{i,j}^t = \left\{ \prod_{s=1}^t \bar{P}_{i,j}^s \ i \leq j \ i > j \right\},
\]

\[
\hat{\gamma} = 4\gamma (1 - \gamma R), \quad C_{i,j}^t = \left\{ \| X_i^t \|^2 \leq R \right\}, \quad \bar{C}_{i,j}^t = \left\{ \| \tilde{X}_i^t \|^2 \leq R \right\},
\]

\[
\mathcal{D}_{i,j}^t = \left\{ \| X_i^t \|^2 \leq R : j \leq i \leq B - 1 \right\} = \bigcap_{i=j}^{B-1} C_{i,j}^t,
\]

\[
\mathcal{D}_s^{*,t} = \bigcap_{s' = s}^{t} \mathcal{D}_{s'}^{*,0} \quad s \leq t, \quad \bar{D}_{s,j}^t = \left\{ \| \tilde{X}_i^t \|^2 \leq R : j \leq i \leq B - 1 \right\} = \bigcap_{i=j}^{B-1} \bar{C}_{i,j}^t,
\]

\[
\bar{D}_s^{*,t} = \bigcap_{s' = s}^{t} \bar{D}_{s'}^{*,0} \quad s \leq t, \quad \bar{D}_{s,j}^t = \mathcal{D}_{s,j}^t \cap \bar{D}_{s,j}^t, \quad \bar{D}_s^{*,t} = \mathcal{D}_s^{*,t} \cap \bar{D}_s^{*,t}.
\]

To execute algorithmic stability arguments, we will need to independently resample individual noise co-ordinates. To that end, define \( (\tilde{\eta}_i)_\tau \) drawn i.i.d from the noise distribution \( \mu \) and independent of everything else defined so far. We denote their generic rows by \( \tilde{\varepsilon} \). We use the following events which correspond to a generic row

\[
\mathcal{E}_{i,j}^t = \left\{ \| \varepsilon_{-k}^t \|^2 \leq \beta, \| \bar{\varepsilon}_{-k}^t \|^2 \leq \beta : i \leq k \leq j \right\}
\]

### C.4 Setting the Parameter Values:

We make Assumptions 1, 2, 3 and 5 throughout. We set the parameters for SGD – RER as follows for the rest of the analysis. We note that some of these parameter values were set in Section 2.

1. \( \alpha \geq 10 \)
2. \( \beta = 4C_\eta \sigma^2 (\alpha + 2) \log 2T \)
3. \( R \geq \frac{16(\alpha + 2) dC_\eta \sigma^2 \log T}{1 - \rho} \)
4. $\delta = 1/(2T^{\alpha+1})$
5. $u \geq \frac{2a\log T}{\log(\frac{1}{\beta})} = O(\tau_{\text{mix}} \log T)$
6. $B \geq \max \left(C_1 \frac{d}{(1-\rho)(1-\rho')}, 10u \right)$ where $C_1$ depends only on $C_p$ (see Theorem 10)
7. $\gamma \leq \min \left(\frac{C}{16R(1+\kappa)}, 1/2R \right)$ (see Theorem 10)

From Assumption 3 and Theorem 8, we conclude that for this choice of $R$ and $\beta$, we must have:

$$P \left[ \left( \tilde{D}_{0,N-1} \cap \bigcap_{r=0}^{N-1} E_{r[N,B-1]} \right)^C \right] \leq \frac{1}{2T^\alpha} \tag{20}$$

### C.5 Bias-variance decomposition

Using the above notation we can unroll the recursion in (18) as follows. We will only focus on the algorithmic iterated at the end of each buffer, i.e., we set $i = B - 1$ in (18).

$$(\tilde{a}_{B}^{t-1} - a^*)^\top = (a_0 - a^*) \prod_{s=0}^{t-1} \tilde{H}_{0,B-1}^s + 2\gamma \sum_{r=1}^{t} \sum_{j=0}^{B-1} \xi_{j-r} \tilde{X}_{j-r}^\top \tilde{H}^{t-r}_{j+1,B-1} \prod_{s=r-1}^{t-1} \tilde{H}_{0,B-1}^{t-s} \tag{21}$$

We call the above the bias-variance decomposition where

$$(\tilde{a}_{B}^{t-1,b} - a^*)^\top = (a_0 - a^*) \prod_{s=0}^{t-1} \tilde{H}_{0,B-1}^s \tag{22}$$

is the bias, and

$$(\tilde{a}_{B}^{t-1,v})^\top = 2\gamma \sum_{r=1}^{t} \sum_{j=0}^{B-1} \xi_{j-r} \tilde{X}_{j-r}^\top \tilde{H}^{t-r}_{j+1,B-1} \prod_{s=r-1}^{t-1} \tilde{H}_{0,B-1}^{t-s} \tag{23}$$

is the variance. We have the following simple lemma on bias-variance decomposition.

**Lemma 5.**

$$\|\tilde{a}_{B}^{t-1} - a^*\|^2 \leq 2 \left(\|\tilde{a}_{B}^{t-1,b} - a^*\|^2 + \|\tilde{a}_{B}^{t-1,v}\|^2\right) \tag{24}$$

### C.6 Variance of last iterate - Diagonal Terms

In this section our goal is to decompose $\|\tilde{a}_{B}^{t-1,v}\|^2$ into diagonal terms and cross terms. We will then proceed to bound the diagonal terms. First, we have a preliminary lemma, which can be shown via a simple recursion.

**Lemma 6.** For $k \leq t$ define $S_k^t$ as

$$S_k^t = \sum_{r=k}^{t-1} \sum_{j=0}^{B-1} \phi_i(\xi_{j-r}) \left( \prod_{s=1}^{r-1} \tilde{H}_{0,B-1}^{t-s} \right) \tilde{H}^{t-r}_{j+1,B-1} \tilde{X}_{j-r}^\top \tilde{X}_{j-r} \tilde{H}^{t-r}_{j+1,B-1} \prod_{s=r-1}^{t-1} \tilde{H}_{0,B-1}^{t-s} \tag{25}$$

Then, on the event $\tilde{D}_{0,t-1}$, we have

$$S_k^t \leq \frac{1}{\gamma} \left( I - \left( \prod_{s=1}^{t} \tilde{H}_{0,B-1}^{t-s} \right) \left( \prod_{s=t}^{1} \tilde{H}_{0,B-1}^{t-s} \right) \right) \tag{26}$$

where $\dot{\gamma} = 4\gamma(1 - \gamma R)$
Proof. The proof is similar to that of [19, Claim 1]. □

Next, we write \( \| \tilde{a}^{t-1,v} \| ^2 \) as

\[
\| \tilde{a}^{t-1,v} \| ^2 = \sum_{r=1}^{B-1} \sum_{j=0}^{B-1} Dg(t, r, j) + \sum_{r_1, r_2, j_1, j_2} Cr(t, r_1, r_2, j_1, j_2)
\]

where the second sum is over \((r_1, j_1) \neq (r_2, j_2)\) and

\[
Dg(t, r, j) = 4\gamma^2 \| \epsilon_{t-j} \| ^2 \cdot \tilde{X}_{t-r} \cdot \tilde{H}_{t-r}^T \tilde{H}_{j+1, B-1}^{-1} \left( \prod_{s=r-1}^{r-1} \tilde{H}_{s, B-1}^{-1} \right) \left( \prod_{s=1}^{r} \tilde{H}_{0, B-1}^{-1} \right) \tilde{H}_{j+1, B-1} \cdot \tilde{X}_{t-r}
\]

and

\[
Cr(t, r_1, r_2, j_1, j_2) = 4\gamma^2 \| \epsilon_{t-r_2} \| \cdot \tilde{X}_{t-r_2} \cdot \tilde{H}_{j+1, B-1}^{-1} \left( \prod_{s=r_2-1}^{r_2-1} \tilde{H}_{s, B-1}^{-1} \right) \left( \prod_{s=1}^{r_2} \tilde{H}_{0, B-1}^{-1} \right) \tilde{H}_{j+1, B-1} \cdot \tilde{X}_{t-r_2}
\]

Finally, we bound the diagonal term:

Claim 1.

\[
E \left[ \sum_{r=1}^{t} \sum_{j=0}^{B-1} Dg(t, r, j) \right] \leq \frac{\gamma d}{\zeta (1 - \gamma R)} \beta + 16C_\eta \sigma^2 \gamma^2 \cdot \frac{1}{T^{\alpha/2}}
\]

Proof. Notice that we can write

\[
Dg(t, r, j) \leq 4\gamma^2 \left( \beta + \| \epsilon_{t-r} \| \cdot \| \epsilon_{t-r} \| > \beta \right) \cdot \tilde{X}_{t-r} \cdot \tilde{H}_{t-r}^T \tilde{H}_{j+1, B-1}^{-1} \left( \prod_{s=r-1}^{r-1} \tilde{H}_{s, B-1}^{-1} \right) \left( \prod_{s=1}^{r} \tilde{H}_{0, B-1}^{-1} \right) \tilde{H}_{j+1, B-1} \cdot \tilde{X}_{t-r}
\]

Further

\[
\tilde{X}_{t-r} \cdot \tilde{H}_{t-r}^T \tilde{H}_{j+1, B-1}^{-1} \left( \prod_{s=r-1}^{r-1} \tilde{H}_{s, B-1}^{-1} \right) \left( \prod_{s=1}^{r} \tilde{H}_{0, B-1}^{-1} \right) \tilde{H}_{j+1, B-1} \cdot \tilde{X}_{t-r} \leq R
\]

Combining the above two we obtain

\[
\sum_{r=1}^{B-1} \sum_{j=0}^{B-1} Dg(t, r, j) \left[ \tilde{B}^{0, t-1} \right] \leq 4\gamma^2 \beta \frac{\text{Tr} S_1}{\zeta} + R \sum_{r=1}^{t} \sum_{j=0}^{B-1} \| \epsilon_{t-j} \| \cdot \| \epsilon_{t-j} \| > \beta
\]

where \( S_1 \) is defined in (25). Now taking expectation, and using lemma 6 and Cauchy-Schwarz inequality for the first and second terms, respectively, in (33) we obtain the claim. Here we use the fact that \( E \left[ \| \epsilon_{t-j} \| ^4 \right] \leq 16C_\eta \sigma^4 \) from [48, Theorem 2.1]

C.7 Algorithmic stability

In order to bound the cross terms in the variance, we need the notion of algorithmic stability. Here the idea is that if \( \phi \) was identity, then \( E \left[ Cr(t, r_1, r_2, j_1, j_2) \right] \) would vanish. But in the non-linear setting, this does not happen due to dependencies between \( \epsilon_{t-j} \) and \( \tilde{H}_{j+1, B-1}^{-1} \left( \prod_{s=r}^{r} \tilde{H}_{s, B-1}^{-1} \right) \) through the algorithmic iterates. We can still show that \( E \left[ Cr(t, r_1, r_2, j_1, j_2) \right] \approx 0 \) by showing that the iterates depend very weakly on each of the noise co-ordinates \( \epsilon_{t-j} \). So our idea is to use algorithmic stability: we re-sample the whole trajectory of \( X \) by re-sampling a single noise co-ordinate independently. We
then show that the iterates are not affected much by such a re-sampling, which shows that the iterates are only weakly coupled to each individual noise vector.

To that end, we need some additional notation. We have the data \((X_\tau)_r\) and the coupled process \((\tilde{X}_\tau)_r\). Let the corresponding (coupled) algorithmic iterates be \((\tilde{a}^*_i : 0 \leq s \leq N - 1, 0 \leq i \leq B - 1)\). Now \(\tilde{a}^*_i\) are functions of \(X_0\) and noise vectors \(\{\eta^r_{-j} : 0 \leq s \leq N - 1, 0 \leq i \leq S - 1\}\). Suppose we re-sample the noise \(\eta^r_{-j}\) independently of everything else to get \(\tilde{\eta}^r_{-j}\). So the new noise samples are:

\[
\left(\tilde{\eta}^0_0, \tilde{\eta}^0_1, \ldots, \tilde{\eta}^r_{n-1}, \tilde{\eta}^{r}_{n-1-(j+1)}, \tilde{\eta}^{r}_{n-1-(j+1)}, \tilde{\eta}^{r}_{n-1-(j+1)}, \ldots\right).
\]

We then run the dynamics in Equation 1 with the new noise samples to obtain \((\tilde{X}_\tau)_r\) and the new coupled process \((\tilde{X}_\tau)_r\) obtained through the new noise sequence (but same stationary renewal given in Definition 2), and they satisfy the following:

\[
\tilde{X}^*_r = \begin{cases} 
X^*_r, & s < r, 0 \leq i \leq S - 1 \\
X^r_r, & s = r, j \leq i \leq S - 1
\end{cases}
\]

\[
\tilde{X}^r_r = \begin{cases} 
\tilde{X}^*_r, & s \in \{1, \ldots, r - 1, r + 1, \ldots, N - 1\}, 0 \leq i \leq S - 1 \\
\tilde{X}^r_r, & s = r, j \leq i \leq S - 1
\end{cases}
\]

We obtain the iterates \(\tilde{\eta}^*_i\) by running the update Equation (13) with the data \(\tilde{X}_\tau\) instead of \(X_\tau\). Accordingly, the algorithmic iterates change to \((\tilde{a}^*_i : 0 \leq s \leq N - 1, 0 \leq i \leq B - 1)\) that satisfy

\[
\tilde{a}^*_i = \tilde{a}^*_i \quad \text{for } s < r, 0 \leq i \leq B - 1
\]

This is because, resampling \(\eta_\tau\) does not change the value of data \(\tilde{X}_\tau\) for \(\tau' \leq \tau\). Under the setting we have the following lemma:

**Lemma 7.** Let \(A^{-1}\) be the following event

\[
A^{t-1} = \cap_{r=0}^{t-1} \cap_{j=0}^{B-1} \left\{ \|\tilde{a}^*_j - a^*\| \leq \|a_0 - a^*\| + \tilde{C} \sqrt{R B} \right\}
\]

(34)

For some constant \(\tilde{C}\) depending only on \(\alpha\), we have for any \(1 \leq t \leq N\)

\[
\mathbb{P} \left[ \hat{D}_0^{N-1} \cap A^{N-1} \cap \cap_{r=0}^{N-1} \cap_{j=0}^{B-1} \right] \geq 1 - \frac{1}{T^\alpha}
\]

(35)

Further more, on the event \(\mathcal{E}_{0,j}^{r} \cap \hat{D}_r^{N-1} \cap A^r\) we have:

\[
\tilde{a}^*_i = \tilde{a}^*_i, 0 \leq s < r, 0 \leq i \leq B - 1
\]

(36)

\[
\tilde{a}^*_0 = \tilde{a}^*_0
\]

(37)

and for \(s \geq r\) we have

\[
\|\tilde{a}^*_i - \tilde{a}^*_i\| \leq \tilde{C}_2 \gamma R B \sqrt{R B} \frac{\sqrt{R B}}{\tilde{\alpha}_{\min}} + 8 \gamma R B \|a_0 - a^*\| \leq \tilde{C}_3 \gamma R B \sqrt{R B} \frac{\sqrt{R B}}{\tilde{\alpha}_{\min}}
\]

(38)

We give the proof in Section H

**Remark 3.** In expression (38), we have suppressed the dependence of \(\|a_0 - a^*\|\) for the ease of exposition with the rationale being that since \(a_0 = 0\), it should be lower order compared to \(\sqrt{R B}\).

Hence we see from the above lemma that changing a particular noise sample in a particular buffer perturbs the algorithmic iterates by \(O(\gamma \text{poly}(RB))\).

Let \(\mathcal{R}^r_{-j}\) denote the re-sampling operator corresponding to re-sampling \(\eta^r_{-j}\). That is, for any function \(f((a_r), (X_r), (\tilde{X}_r))\) we have

\[
\mathcal{R}^r_{-j} \left( f((a_r), (\tilde{a}_r), (X_r), (\tilde{X}_r)) \right) = f((\tilde{a}_r), (\tilde{a}_r), (X_r), (\tilde{X}_r))
\]

(39)

We will drop the subscripts and superscripts on \(\mathcal{R}\) when there is no ambiguity on which noise is re-sampled. First we will prove a lemma that bounds the effect of re-sampling.
Lemma 8. On the event $\mathcal{E}_{0,j}^t \cap \tilde{D}^{0,t-1} \cap \mathcal{A}^{t-1}$, for some constant $C$ depending only on $C_r$: 
\[
\left\| \tilde{H}_{j+1,B-1}^{t-r} \left( \prod_{s=r-1}^{1} \tilde{H}_{0,B-1}^{t-s} \right) - \mathcal{R}_{j}^{t-r} \tilde{H}_{j+1,B-1}^{t-r} \left( \prod_{s=r-1}^{1} \mathcal{R}_{j}^{t-s} \tilde{H}_{0,B-1}^{t-s} \right) \right\| 
\leq \tilde{C} B t \| \phi'' \| \gamma^2 R^3 B \sqrt{B} \quad (40)
\]

Proof. First, note that since we are re-sampling $\eta_{j-r}$, the only difference between $\mathcal{R}_{j}^{t-r} \tilde{H}_{j+1,B-1}^{t-r} \left( \prod_{s=r-1}^{1} \mathcal{R}_{j}^{t-s} \tilde{H}_{0,B-1}^{t-s} \right)$ and $\tilde{H}_{j+1,B-1}^{t-r} \left( \prod_{s=r-1}^{1} \tilde{H}_{0,B-1}^{t-s} \right)$ is that the algorithmic iterates $\tilde{a}_j^s$ that appear in the latter (through $\phi' (\cdot)$) are replaced by $\tilde{a}_j^s$ in the former, but the covariates remain the same in both.

Now, the matrix $\tilde{H}_{j+1,B-1}^{t-r} \left( \prod_{s=r-1}^{1} \tilde{H}_{0,B-1}^{t-s} \right)$ is of the form $\prod_{i=1}^{k} A_i$ where $\| A_i \| \leq 1$ under the conditioned events and is of the form $I - 2\gamma \phi' (\tilde{\xi}_{-i}^{t-s}) \tilde{X}_{-i}^{t-s} \tilde{X}_{-i}^{t-s,\top}$. Similarly, we write: $\mathcal{R}_{j}^{t-r} \tilde{H}_{j+1,B-1}^{t-r} \left( \prod_{s=r-1}^{1} \mathcal{R}_{j}^{t-s} \tilde{H}_{0,B-1}^{t-s} \right) = \prod_{i=1}^{k} A_i$ where $A_i = \mathcal{R}_{j-i}^{t-r} A_i$. Now consider the simple inequality under the condition that $\| A_i \|, \| \tilde{A}_i \| \leq 1$
\[
\left\| \prod_{i=1}^{k} A_i - \prod_{i=1}^{k} \tilde{A}_i \right\| \leq \sum_{i=1}^{k} \| A_i - \tilde{A}_i \| \quad (41)
\]

Therefore, we will just bound each of the component differences $\| A_i - \tilde{A}_i \|$. To this end, consider a typical term $I - 2\gamma \phi' (\tilde{\xi}_{-i}^{t-s}) \tilde{X}_{-i}^{t-s} \tilde{X}_{-i}^{t-s,\top}$. We have 
\[
\left( I - 2\gamma \phi' (\tilde{\xi}_{-i}^{t-s}) \tilde{X}_{-i}^{t-s} \tilde{X}_{-i}^{t-s,\top} \right) - \mathcal{R}_{j-i}^{t-r} \left( I - 2\gamma \phi' (\tilde{\xi}_{-i}^{t-s}) \tilde{X}_{-i}^{t-s} \tilde{X}_{-i}^{t-s,\top} \right) = 2\gamma (\phi' (\tilde{\xi}_{-i}^{t-s}) - \mathcal{R}_{j-i}^{t-r} \phi' (\tilde{\xi}_{-i}^{t-s})) \tilde{X}_{-i}^{t-s} \tilde{X}_{-i}^{t-s,\top} \quad (42)
\]

Now
\[
\phi' (\tilde{\xi}_{-i}^{t-s}) - \mathcal{R}_{j-i}^{t-r} \phi' (\tilde{\xi}_{-i}^{t-s}) = \frac{\phi'(a_{i}^{t-s,\top} \tilde{X}_{-i}^{t-s}) - \phi(a^{t-s,\top} \tilde{X}_{-i}^{t-s})}{(a_{i}^{t-s} - a^{t-s}) \top \tilde{X}_{-i}^{t-s}} - \frac{\phi(\tilde{a}_{i}^{t-s,\top} \tilde{X}_{-i}^{t-s}) - \phi(a^{t-s,\top} \tilde{X}_{-i}^{t-s})}{(\tilde{a}_{i}^{t-s} - a^{t-s}) \top \tilde{X}_{-i}^{t-s}} \quad (43)
\]

Now we can use the following simple result from calculus. Suppose $f$ is a real valued twice continuously differentiable function with bounded second derivative (denoted by $\| f'' \|$). Fix $x_0 \in \mathbb{R}$. Let $g(x) = \frac{f(x) - f(x_0)}{x - x_0}$. By the mean value theorem, there exists $\xi$ such that:
\[
g'(x) = \frac{f(x_0) - (f(x) + (x - x_0)f'(x))}{(x - x_0)^2} = \frac{1}{2} f''(\xi)
\]

Now for any $x, y$, we have 
\[
|g(x) - g(y)| = |g'(\xi_1)(x - y)| \leq \frac{1}{2} \| f'' \| | x - y |
\]

for some $\xi_1$ between $x$ and $y$. Here again we use the mean value theorem in the equality above. Now we will apply this result to $\phi$ with $x = \tilde{a}_{i}^{t-s,\top} \tilde{X}_{-i}^{t-s}, y = \tilde{a}_{i}^{t-s,\top} \tilde{X}_{-i}^{t-s}$ and $x_0 = a^{t-s,\top} \tilde{X}_{-i}^{t-s}$ to get 
\[
\left| \phi' (\tilde{\xi}_{-i}^{t-s}) - \mathcal{R}_{j-i}^{t-r} \phi' (\tilde{\xi}_{-i}^{t-s}) \right| \leq \frac{1}{2} \| \phi'' \| \| \tilde{a}_{i}^{t-s} - \tilde{a}_{i}^{t-s} \| \| \tilde{X}_{-i}^{t-s} \|
\]

(44)
Now we appeal to lemma \[7\] In particular, using equation \[38\] we see that, on the event $E_{0,j} \cap \mathcal{D}^{0,t-1} \cap \mathcal{A}^{t-1}$,
\[
\phi'(\tilde{x}_{t-s}^-) - \phi'(\tilde{x}_{t-s}^+)| \leq \frac{1}{2} \|\phi''\|_1 \gamma R \frac{1}{\lambda_{\min}} \sqrt{\rho B} \leq 64C \|\phi''\| \gamma R \frac{1}{\lambda_{\min}} B \sqrt{R} \leq 128C \|\phi''\| \gamma^2 R^3 B \frac{1}{\lambda_{\min}} (45)
\]

We now use Equation \[41\] with $k \leq Bt$ along with Equation \[46\] to conclude the statement of the lemma.

**C.8 Bound $C_r(t, r_1, r_2, j_1, j_2)$**

Next we will bound $\sum_r \sum_{j_1 \neq j_2} C_r(t, r, j_1, j_2)$

**Claim 2.**
\[
\mathbb{E} \left[ \sum_{r=1}^{t} \sum_{j_1 \neq j_2} C_r(t, r, j_1, j_2) \right] \leq \tilde{C} \left[ \frac{\sigma^2 \gamma^2 R B}{T^2} + \frac{\|\phi''\| \gamma T^2 R^3 B^2 \sigma^2}{\lambda_{\min}} \right] (47)
\]

Where $\tilde{C}$ is a constant depending only on $C_r$

**Proof.** Let $j_1 < j_2$. We will suppress the arguments of $C_r$ for brevity. First, we re-sample the noise which is ahead in the time, i.e., $\tilde{\eta}_{t-r}^{j_1}$ (and hence the entry $\tilde{\xi}_{t-r}^{j_1}$, in the row under consideration).

Let $Cr'$ denote the resampled version of $Cr$ as defined below
\[
Cr'(t, r, j_1, j_2) := 4 \gamma^2 \tilde{\eta}_{t-r}^{j_1} - \tilde{\eta}_{t-r}^{j_2} \mathcal{R}_{t-j_1}^{j_2} \tilde{X}_{t-r}^{j_1} \tilde{H}_{t-j_2}^{j_2} \left( \prod_{s=r-1}^{t-1} \tilde{H}_{t-s}^{j_1, B-1} \right) \cdot \left( \prod_{s=1}^{r-1} \tilde{H}_{t-s}^{j_1, B-1} \right) \tilde{H}_{t-j_1, B-1} \tilde{X}_{t-r}^{j_1} \mathcal{R}_{t-j_2}^{j_1} \left( \prod_{s=r-1}^{t-1} \tilde{H}_{t-s}^{j_2, B-1} \right), (48)
\]

where we have used the fact that $\mathcal{R}_{t-j_1}^{j_2}$ has no effect on the items from the process $(\tilde{X}_r)_r$ that appear in the expression above. Note the this is not $\mathcal{R}_{t-j_1}^{j_2} C_r$, since in $\mathcal{R}_{t-j_1}^{j_2} C_r$ we would have $\tilde{\eta}_{t-r}^{j_1}$ instead.

Now, since the new algorithmic iterates $(\tilde{A}_t)$ depend on $\tilde{\eta}_{t-r}^{j_1}$ but not on $\eta_{t-r}^{j_1}$, it is immediate that
\[
\mathbb{E} \left[ C_{r'}(t, r, j_1, j_2) \right] = 0
\]

For convenience, we introduce some notation which is only used in this proof. $Cr'(t, r, j_1, j_2)$ can be written in the form $4 \gamma^2 \tilde{\xi}_{t-r}^{j_1} - \tilde{\eta}_{t-r}^{j_2} K_1$, for some random variable $K_1$ independent of $\tilde{\xi}_{t-r}^{j_1}$. Under the event $\mathcal{D}^{0,t-1}$, we can easily show that $|K_1| \leq R$ almost surely. Let $F_K = \sigma(K_1, \tilde{\xi}_{t-r}^{j_2})$. Let $\mathcal{M} := \{|K_1| \leq R\}$. Clearly, $\mathcal{D}^{0,t-1} \subseteq \mathcal{M}$ and $\tilde{\xi}_{t-r}^{j_2} \perp \mathcal{F}_K$. We conclude:
\[
\left| \mathbb{E} \left[ C_{r'} 1 \left[ \mathcal{D}^{0,t-1} \right] \right] \right| = \left| \mathbb{E} \left[ C_{r'} 1 \left[ \mathcal{D}^{0,t-1} \right] \mathcal{M} \right] \right| = 4 \gamma^2 \mathbb{E} \left[ \mathbb{E} \left[ \mathcal{D}^{0,t-1} | F_K \right] \mathcal{M} \right] = 4 \gamma^2 \mathbb{E} \left[ \mathbb{E} \left[ \mathcal{D}^{0,t-1} | F_K \right] \mathcal{M} \right] \leq 4 \gamma^2 \mathbb{E} \left[ \mathbb{E} \left[ \mathcal{D}^{0,t-1} | F_K \right] \mathcal{M} \right] \leq 4 \gamma^2 \mathbb{E} \left[ \mathcal{D}^{0,t-1} | F_K \right] \mathcal{M} \right] (49)
\]
We refer to Section H for the proof.

We note that:

Therefore, on the event $E$ from Lemma 9, we have that on the event $r > r_1$ and arbitrary $j_1$ and $j_2$, we conclude:

Using similar technique as Lemma 8, we have that:

Next we want to bound $\mathbb{E}[\varepsilon_{r_1,j_1}^r \mid \bar{D}^{0,t-1} \cap \mathcal{A}^{t-1}]$, for $r > r_1$ and arbitrary $j_1$ and $j_2$.

Recall the definition of $\hat{a}_{B}^{t-1,v}$ from (23). Via simple rearrangement of summation, we can express $\sum_{r_2 > r_1} \sum_{j_1,j_2} \mathbb{E}[\varepsilon_{r_1,j_1}^r \mid \bar{D}^{0,t-1} \cap \mathcal{A}^{t-1}]$ in terms of $\hat{a}_{B}^{t-1,v}$ as follows.

Claim 3.

The proof of the claim essentially proceeds similar to that of Claim 2, but with additional complications. We refer to Section H for the proof.

Combining everything in this section we have the following proposition.
Proposition 1. Let

\[
\hat{V}_{t-1} = \mathbb{E} \left[ \left\| \hat{a}_{B}^{t-1,v} \right\|^2 1 \left[ \bar{D}^{0,t-1} \right] \right]
\]  

(56)

Then for some constant \( \hat{C} \) which depends only on \( C_{\eta} \):

\[
\sup_{s \leq N-1} \hat{V}_{s} \leq \frac{2\gamma d}{\zeta (1 - \gamma R)} + \hat{C} \| \phi'' \| \gamma^4 T^2 R^4 B^2 \frac{\sigma^2 \sqrt{\beta}}{\zeta \lambda_{\min}} + \hat{C} \sigma^2 R^2 T \frac{1}{T^{\alpha/2}} + \hat{C} R \sigma^2 \left( \| \phi'' \| \gamma^4 T^2 R^4 B^2 \frac{\sqrt{\beta}}{\zeta \lambda_{\min}} + \gamma^4 T^2 R^2 B^2 \right)
\]  

(57)

Proof. In the whole proof, we will denote any large enough constant depending on \( C_{\eta} \) by \( \hat{C} \). From claims 1, 2 and 3 along with equation (27) we have

\[
\mathbb{E} \left[ \left\| \hat{a}_{B}^{t-1,v} \right\|^2 1 \left[ \bar{D}^{0,t-1} \right] \right] \leq \frac{\gamma d}{\zeta (1 - \gamma R)} + \hat{C} \| \phi'' \| \gamma^4 T^2 R^4 B^2 \frac{\sigma^2 \sqrt{\beta}}{\zeta \lambda_{\min}} + \hat{C} \sigma^2 R^2 T \frac{1}{T^{\alpha/2}} (1 + B + T) + \hat{C} \left( \| \phi'' \| \gamma^4 T^2 R^4 B^2 \frac{\sqrt{\beta}}{\zeta \lambda_{\min}} + \gamma^4 T^2 R^2 B^2 \right) \sqrt{R \sigma^2} \sqrt{\sup_{s \leq N-1} \hat{V}_{s}} 
\]  

(58)

Thus

\[
\sup_{s \leq N-1} \hat{V}_{s} \leq \frac{\gamma d}{\zeta (1 - \gamma R)} + \hat{C} \| \phi'' \| \gamma^4 T^2 R^4 B^2 \frac{\sigma^2 \sqrt{\beta}}{\zeta \lambda_{\min}} + \hat{C} \sigma^2 R^2 T \frac{1}{T^{\alpha/2}} + \hat{C} \left( \| \phi'' \| \gamma^4 T^2 R^4 B^2 \frac{\sqrt{\beta}}{\zeta \lambda_{\min}} + \gamma^4 T^2 R^2 B^2 \right) \sqrt{R \sigma^2} \sqrt{\sup_{s \leq N-1} \hat{V}_{s}} 
\]  

(59)

Finally, we need to solve the above recursive relation. We note a simple fact: Let \( c_1, c_2 > 0 \) be constants and let \( x > 0 \) satisfy

\[
x^2 \leq c_1 + c_2 x
\]  

(60)

then

\[
x^2 \leq \frac{1}{4} \left( c_2 + \sqrt{c_2^2 + 4c_1} \right)^2 \leq c_2^2 + 2c_1
\]  

(61)

where in the last inequality above we used the fact that \( (a + b)^2 \leq 2(a^2 + b^2) \).

Thus,

\[
\sup_{s \leq N-1} \hat{V}_{s} \leq \frac{2\gamma d}{\zeta (1 - \gamma R)} + \hat{C} \| \phi'' \| \gamma^4 T^2 R^4 B^2 \frac{\sigma^2 \sqrt{\beta}}{\zeta \lambda_{\min}} + \hat{C} \sigma^2 R^2 T \frac{1}{T^{\alpha/2}} + \hat{C} R \sigma^2 \left( \| \phi'' \| \gamma^4 T^2 R^4 B^2 \frac{\sqrt{\beta}}{\zeta \lambda_{\min}} + \gamma^4 T^2 R^2 B^2 \right)
\]  

(62)

\[ \square \]

C.9 Bias of last iterate

In this part, we will bound the expectation of the bias term \( \left\| \hat{a}_{B}^{t-1,b} - a^* \right\|^2 \).

Theorem 5. For some universal constant \( c_0 \):

\[
\mathbb{E} \left[ \left\| \hat{a}_{B}^{t-1,b} - a^* \right\|^2 1 \left[ \bar{D}^{0,t-1} \right] \right] \leq \left\| a_0 - a^* \right\|^2 (1 - c_0 \zeta B \lambda_{\min})^t
\]  

(63)

where \( \hat{a}_{B}^{t-1,b} - a^* \) is defined in (22)
Proof. Define \( x_v = \prod_{s=0}^{v-1} H_{0,B-1}^s (a_0 - a^*) \). Here, we consider the event \( G_v \) considered in Claim 6 in the proof of Theorem 10, and show that for some universal constant \( q_0 > 0 \),

\[
\mathbb{P}(\|H_{0,B-1}^{v,T} x_v\|^2 \geq (1 - \zeta \lambda_{\min}(G)) \|x_v\|^2 \|\tilde{D}^{0,t-1}, x_v\| \geq q_0)
\]

(64)

From Theorem 10 we also note that conditioned on \( \tilde{D}^{0,t-1} \), almost surely:

\[
\|H_{0,B-1}^{v,T} x_v\|^2 \leq 1
\]

We let \( G_v \) be the event lower bounded in Equation (64).

\[
\mathbb{E} \left[ \|x_{v+1}\|^2 | \tilde{D}^{0,t-1} \right] = \mathbb{E} \left[ \|H_{0,B-1}^{v,T} x_v\|^2 | \tilde{D}^{0,t-1} \right]
\]

\[
= \mathbb{E} \left[ \|H_{0,B-1}^{v,T} x_v\|^2 \mathbb{1}[G_v] + \|H_{0,B-1}^{v,T} x_v\|^2 \mathbb{1}[\neg G_v] | \tilde{D}^{0,t-1} \right]
\]

\[
\leq \mathbb{E} \left[ (1 - \zeta \lambda_{\min}(G)) \|x_v\|^2 \mathbb{1}[G_v] + \|x_v\|^2 \mathbb{1}[\neg G_v] | \tilde{D}^{0,t-1} \right]
\]

\[
= \mathbb{E} \left[ \|x_v\|^2 \mathbb{1}[1 - \zeta \lambda_{\min}(G)] \mathbb{P}(G_v | \tilde{D}^{0,t-1}, x_v) | \tilde{D}^{0,t-1} \right]
\]

\[
\leq \mathbb{E} \left[ \|x_v\|^2 \mathbb{1}[1 - \zeta \lambda_{\min}(G)] q_0 | \tilde{D}^{0,t-1} \right]
\]

(65)

Unrolling the recursion given by Equation (65), and noting that \( \hat{a}^{t-1,b} - a^* = x_t \), we conclude

\[
\mathbb{E} \left[ \|\hat{a}^{t-1,b} - a^*\|^2 \mathbb{1}[\tilde{D}^{0,t-1}] \right] \leq (1 - c_0 \gamma B \lambda_{\min} \zeta)^t
\]

Hence we have the theorem.

\[\square\]

C.10 Average iterate: bias-variance decomposition

In the part, we will consider the tail-averaged iterate where a generic row is given by

\[
\hat{a}_{t_0,N} = \frac{1}{N - t_0} \sum_{t=t_0+1}^{N} \hat{a}_{t_B}^{t-1}
\]

(66)

where \( t_0 \in \{0, 1, \cdots, N - 1\} \).

Thus we can write \( \hat{a}_{t_0,N} - a^* \) as

\[
\hat{a}_{t_0,N} - a^* = (\hat{a}_v^{t_0}) + (\hat{a}_b^{t_0} - a^*)
\]

(67)

where

\[
\hat{a}_v^{t_0} = \frac{1}{N - t_0} \sum_{t=t_0+1}^{N} (\hat{a}_B^{t-1,v})
\]

(68)

\[
\hat{a}_b^{t_0} - a^* = \frac{1}{N - t_0} \sum_{t=t_0+1}^{N} (\hat{a}_B^{t-1,b} - a^*)
\]

(69)

C.11 Variance of average iterate

Remark 4. From now on we will use the following notation:

\[
\sum_t \equiv \sum_{t=t_0+1}^{N}
\]

28
\[
\sum_{t_1, t_2} = \sum_{t_1, t_2 = t_0 + 1}^{N}
\]
\[
\sum_{t_1 \neq t_2} = \sum_{t_1, t_2 = t_0 + 1}^{N}
t_1 \neq t_2
\]
\[
\sum_{t_2 > t_1} = \sum_{t_1 = t_0 + 1}^{N} \sum_{t_2 = t_1 + 1}^{N}
\]

Next we expand \( \| \tilde{\alpha}^v_{t_0, N} \|^2 \):
\[
\| \tilde{\alpha}^v_{t_0, N} \|^2 = \frac{1}{(N - t_0)^2} \sum_t \| \tilde{\alpha}^{t-1,v}_{B} \|^2 + \frac{1}{(N - t_0)^2} \sum_{t_1 \neq t_2} (\tilde{\alpha}^{t_2-1,v}_{B})^\top (\tilde{\alpha}^{t_1-1,v}_{B})
\]

Claim 4. For \( t_2 > t_1 \)
\[
E \left[ \left( (\tilde{\alpha}^{t_2-1,v}_{B})^\top - \left( \prod_{s = t_2 - t_1} \tilde{H}_{s}^{t_2-s} \right) (\tilde{\alpha}^{t_1-1,v}_{B}) \right) \right] 1\left[ \tilde{D}^{0,N-1} \right] \leq C_1 \text{Poly}(R, B, \beta, 1/\zeta, 1/\lambda_{\text{min}}, \| \phi' \|) \left( \gamma^{7/2}T^2 + \gamma^3T^3 + \gamma^6T^4 \right)
\]

Proof. From (23) we can write
\[
(\tilde{\alpha}^{t_2-1,v}_{B})^\top = (\tilde{\alpha}^{t_1-1,v}_{B})^\top \left( \prod_{s = t_2 - t_1} \tilde{H}_{s}^{t_2-s} \right) + 2\gamma \sum_{r=1}^{t_2-t_1} \sum_{j=0}^{B-1} \varepsilon_{j-r}^{t_2-t_1} \tilde{X}_{j-r}^{t_2-t_1} \tilde{H}_{j+1}^{t_2-r} \left( \prod_{s = r-1} \tilde{H}_{s}^{t_2-s} \right)
\]

Hence
\[
(\tilde{\alpha}^{t_2-1,v}_{B})^\top (\tilde{\alpha}^{t_1-1,v}_{B}) = (\tilde{\alpha}^{t_1-1,v}_{B})^\top \left( \prod_{s = t_2 - t_1} \tilde{H}_{s}^{t_2-s} \right) (\tilde{\alpha}^{t_1-1,v}_{B}) + 2\gamma \sum_{r=1}^{t_2-t_1} \sum_{j=0}^{B-1} \varepsilon_{j-r}^{t_2-t_1} \tilde{X}_{j-r}^{t_2-t_1} \tilde{H}_{j+1}^{t_2-r} \left( \prod_{s = r-1} \tilde{H}_{s}^{t_2-s} \right) (\tilde{\alpha}^{t_1-1,v}_{B})
\]

Now recall the noise re-sampling operator \( \mathcal{R}^{t_2-r}_{j} \) from (39). It is easy to see that
\[
E \left[ 2\gamma \sum_{r=1}^{t_2-t_1} \sum_{j=0}^{B-1} \varepsilon_{j-r}^{t_2-r} \mathcal{R}^{t_2-r}_{j} \left[ \tilde{X}_{j-r}^{t_2-r} \tilde{H}_{j+1}^{t_2-r} \left( \prod_{s = r-1} \tilde{H}_{s}^{t_2-s} \right) (\tilde{\alpha}^{t_1-1,v}_{B}) \right] \right] = 0
\]
(Note that \( \mathcal{R}^{t_2-r}_{j} \tilde{X}_{j-r}^{t_2-r} \tilde{H}_{j+1}^{t_2-r} = \tilde{X}_{j-r}^{t_2-r} \tilde{H}_{j+1}^{t_2-r} \))

Thus, using the decomposition
\[
\tilde{D}^{0,N-1} = \tilde{D}^{0,t_2-r-1} \cap \tilde{D}^{t_2-r+1,N-1} \cap \tilde{D}^{t_2-r} \cap \bigcap_{i=0}^{t_2-r-1} \tilde{C}_{i}^{t_2-r}
\]
we get
\[ E \left[ 2\gamma \sum_{r=1}^{t_2-t_1} \sum_{j=0}^{B-1} \epsilon_{x_j}^{t_2-r} \tilde{X}_{x_j}^{t_2-r,\top} \right] \]
\[ \leq 4\gamma^2 R(Bt_1)(B(t_2 - t_1))C_\eta \sigma^2 \frac{1}{T^{a/2}} \]
\[ \leq 4\gamma^2 R C_\eta \sigma^2 T^2 \frac{1}{T^{a/2}} \tag{75} \]

Next we need to bound the effect due to noise re-sampling. On the event \( \mathcal{D}^{0,N-1} \cap \mathcal{A}^{N-1} \cap \cap_{r=0}^{N-1} \mathcal{C}_r \), we have
\[ \left\| \tilde{H}^{t_2-r}_{j+1,B-1} \left( \prod_{s=r-1}^{1} \tilde{H}^{t_2-s}_{0,B-1} \right) - R^{t_2-r}_{x_j} \left[ \tilde{H}^{t_2-r}_{j+1,B-1} \left( \prod_{s=r-1}^{1} \tilde{H}^{t_2-s}_{0,B-1} \right) \right] (\tilde{a}^{t_{1-1},v}) \right\| \]
\[ \leq C \| \phi'' \| \gamma^2 T^2 B \frac{\sqrt{\beta}}{\zeta_{\min}} \]
\[ \tag{76} \]

Thus
\[ 2\gamma E \left[ \sum_{r=1}^{t_2-t_1} \sum_{j=0}^{B-1} \epsilon_{x_j}^{t_2-r} \tilde{X}_{x_j}^{t_2-r,\top} \frac{1}{\mathcal{D}^{0,N-1}} \right] \]
\[ \leq \tilde{C} \gamma \left( \| \phi'' \| \gamma^2 T^2 B \frac{\sqrt{\beta}}{\zeta_{\min}} \right) B(t_2 - t_1) \sqrt{R} \sigma^2 \mathbb{E} \left[ \left\| (\tilde{a}^{t_{1-1},v}) \right\| \frac{1}{\mathcal{D}^{0,t_1-1}} \right] \]
\[ + \tilde{C} \gamma^2 (2R)\sigma^2 (Bt_1)(B(t_2 - t_1)) \frac{1}{T^{a/2}} \tag{77} \]

Now from proposition[1] there is a constant \( C_1 \) such that
\[ \left( E \left[ \| \tilde{a}^{t_{1-1},v} \| \frac{1}{\mathcal{D}^{0,t_1-1}} \right] \right)^2 \leq \]
\[ C_1 \left( \frac{\gamma d}{\zeta(1 - \gamma R)} \beta + \| \phi'' \| \gamma^3 T^2 R^4 B^2 \frac{\sigma^2 \sqrt{\beta}}{\zeta_{\min}} + \sigma^2 \gamma^2 R^2 T^2 \frac{1}{T^{a/2}} + \right. \]
\[ \left. R \sigma^2 \left( \| \phi'' \| \gamma^6 T^4 R^6 B^2 \frac{\beta}{\zeta^2 \lambda_{\min}^2} + \gamma^4 T^2 R^2 B^2 \right) \right) \]
\[ \tag{78} \]

So
\[ 2\gamma E \left[ \sum_{r=1}^{t_2-t_1} \sum_{j=0}^{B-1} \epsilon_{x_j}^{t_2-r} \tilde{X}_{x_j}^{t_2-r,\top} \frac{1}{\mathcal{D}^{0,N-1}} \right] \]
\[ \leq \text{Poly}(R, B, \beta, 1/\zeta, 1/\lambda_{\min}, \| \phi'' \| ) \left( \gamma^{7/2} T^2 + \gamma^{5} T^3 + \gamma^{6} T^4 \right) \]
\[ \tag{79} \]

where we absorbed terms involving \( \frac{1}{T} \) since \( \alpha \) is taken to be large.
Claim 5.

$$\left| \mathbb{E} \left[ (\hat{a}_{t_B}^{t_1-1,v})^\top \sum_{t_2 > t_1} \left( \prod_{s = t_2 - t_1}^{t_2 - s} \tilde{H}_{0,B-1}^{t_2-s} \right) (\hat{a}_{B}^{t_1-1,v}) 1 \left[ \tilde{D}^{0,N-1}_{B} \right] \right] \right| \leq \mathcal{V}_{t_1-1} \frac{C}{\zeta \gamma B \lambda_{\min}} + 16(N - t_1) \gamma^2 RC_\eta \sigma^2 T^2 \frac{1}{T^{\alpha/2}}$$

(80)

where $\mathcal{V}_{t_1-1}$ is defined in (56).

Proof. Note that

$$\prod_{s = t_2 - t_1}^{t_2 - s} \tilde{H}_{0,B-1}^{t_2-s} = \prod_{s = t_1}^{t_2 - 1} \tilde{H}_{0,B-1}^s$$

From theorem [11] there is a universal constant $C$ such that with $\delta = \frac{1}{2T^\alpha}$ we have

$$\mathbb{P} \left[ \left\| \sum_{t_2 > t_1} \prod_{s = t_1}^{t_2 - 1} \tilde{H}_{0,B-1}^s \right\| > C \left( d + \log \frac{N}{\delta} + \frac{1}{\zeta \gamma B \lambda_{\min}} \right) \tilde{D}^{t_1,N-1} \right] \leq \frac{1}{2T^\alpha}$$

(81)

Since $\mathbb{P} \left[ \tilde{D}^{t_1,N-1} \right] \leq \frac{1}{2T^\alpha}$ we obtain

$$\mathbb{P} \left[ \left\| \sum_{t_2 > t_1} \prod_{s = t_1}^{t_2 - 1} \tilde{H}_{0,B-1}^s \right\| > C \left( d + \log \frac{N}{\delta} + \frac{1}{\zeta \gamma B \lambda_{\min}} \right) \right] \leq \frac{1}{T^\alpha}$$

(82)

Choosing $\gamma$ such that

$$d + \log \frac{N}{\delta} \leq \frac{1}{\zeta \gamma B \lambda_{\min}}$$

we get

$$\mathbb{P} \left[ \left\| \sum_{t_2 > t_1} \prod_{s = t_1}^{t_2 - 1} \tilde{H}_{0,B-1}^s \right\| > \frac{C}{\zeta \gamma B \lambda_{\min}} \right] \leq \frac{1}{T^\alpha}$$

(83)

Thus conditioning on the event

$$\left\| \sum_{t_2 > t_1} \prod_{s = t_1}^{t_2 - 1} \tilde{H}_{0,B-1}^s \right\| \leq \frac{C}{\zeta \gamma B \lambda_{\min}}$$

we obtain

$$\mathbb{E} \left[ (\hat{a}_{B}^{t_1-1,v})^\top \sum_{t_2 > t_1} \left( \prod_{s = t_2 - t_1}^{t_2 - s} \tilde{H}_{0,B-1}^{t_2-s} \right) (\hat{a}_{B}^{t_1-1,v}) 1 \left[ \tilde{D}^{0,N-1}_{B} \right] \right] \leq \mathcal{V}_{t_1-1} \frac{C}{\zeta \gamma B \lambda_{\min}} + 16(N - t_1) \gamma^2 RC_\eta \sigma^2 T^2 \frac{1}{T^{\alpha/2}}$$

(84)

Thus combining everything we have the following theorem

**Theorem 6.** Suppose $\gamma \geq \frac{1}{T}$. Then

$$\mathbb{E} \left[ \left\| \hat{a}_{t_0,N}^{v} \right\|^2 1 \left[ \tilde{D}^{0,N-1}_{B} \right] \right] \leq C_1 \frac{d\beta}{\zeta^2 \lambda_{\min} B (N - t_0)} + \bar{P} \cdot \left( \gamma^{7/2} T^2 + \gamma^6 T^4 \right)$$

(85)

$\bar{P} = \text{Poly}(R, B, \beta, 1/\zeta, 1/\lambda_{\min}, \|\phi''\|, C_\eta)$ and $C_1 > 0$ is some constant.
C.12 Bias of average iterate

**Theorem 7.** There are constants $C, c_1, c_2$ such that:

$$
E \left[ \| \hat{a}_{t_0,N} - a^* \|^2 \right] \leq C \| a_0 - a \|^2 \left[ e^{-c_2 \zeta \gamma B \lambda_{\min} t_0} \min \left\{ 1, \frac{1}{(N - t_0) \zeta \gamma B \lambda_{\min}} \right\} \right]
$$

(86)

The proof follows from an application of Theorem 5.

C.13 Proof of Theorem 3

**Proof.** Let $\tilde{P}$ denote the polynomial in Theorem 6. Theorems 7 and 6 imply for every row of the coupled iterate:

$$
E \left[ \| \hat{a}_{t_0,N} - a^* \|^2 \left[ \tilde{D}^{0,N-1} \right] \right] \leq 2E \left[ \| \hat{a}_{t_0,N} \|^2 \left[ \tilde{D}^{0,N-1} \right] \right] + 2E \left[ \| \hat{a}_{t_0,N} - a^* \|^2 \left[ \tilde{D}^{0,N-1} \right] \right]
$$

$$
\leq C_2 \zeta^2 \lambda_{\min} (G) B(N - t_0) + \tilde{P} \left( \gamma^{7/2} T^2 + \gamma^6 T^4 \right)
$$

Thus for the actual process we can use the following decomposition

$$
E \left[ \| \hat{a}_{t_0,N} - a^* \|^2 \left[ \tilde{D}^{0,N-1} \right] \right] \leq E \left[ \| \hat{a}_{t_0,N} - a^* \|^2 \left[ \tilde{D}^{0,N-1} \right] \right] + E \left[ \| \hat{a}_{t_0,N} - a^* \|^2 \left[ \tilde{D}^{0,N-1} \right] \right]
$$

$$
\leq E \left[ \| \hat{a}_{t_0,N} - a^* \|^2 \left[ \tilde{D}^{0,N-1} \right] \right] + C_2 \gamma^2 T^2 R C_2 \sigma^2 \frac{1}{T^{\alpha/2}}
$$

(87)

where we used the fact on the event $\tilde{D}^{0,N-1}$

$$
\| \hat{a}_{t_0,N} - a^* \|^2 \leq \frac{1}{N - t_0} \sum_{t = t_0 + 1}^{N} \left( 2 \| a_0 - a^* \|^2 + 2(Bt)(4\gamma^2 R \sum_{r = 1}^{t} \sum_{j = 0}^{B-1} | \varepsilon_t^j |^2) \right)
$$

(88)

and then used Cauchy-Schwarz inequality for the expectation over $\tilde{D}^{0,N-1,C}$. Now using lemma 3 we get

$$
E \left[ \| \hat{a}_{t_0,N} - a^* \|^2 \left[ \tilde{D}^{0,N-1} \right] \right] \leq E \left[ \| \hat{a}_{t_0,N} - a^* \|^2 \left[ \tilde{D}^{0,N-1} \right] \right] + C_2 \gamma^2 T^2 \frac{1}{T^{\alpha}}
$$

(89)

since we are choosing $u$ such that $\rho_u \leq \frac{1}{T^\alpha}$. Using $\tilde{D}^{0,N-1} \subset D_0, N - 1$ we get

$$
E \left[ \| \hat{a}_{t_0,N} - a^* \|^2 \left[ \tilde{D}^{0,N-1} \right] \right] \leq E \left[ \| \hat{a}_{t_0,N} - a^* \|^2 \left[ \tilde{D}^{0,N-1} \right] \right] + C_2 \gamma^2 T \frac{1}{T^{\alpha/2-2}}
$$

(90)

where we absorbed all terms of order $\frac{1}{T^\alpha}$ (including those depending on $\| a_0 - a^* \|$) to the last term in the above display. Thus

$$
E \left[ \| \hat{a}_{t_0,N} - a^* \|^2 \left[ \tilde{D}^{0,N-1} \right] \right] \leq C_1 \zeta^2 \lambda_{\min} (G) B(N - t_0) + \tilde{P} \left( \gamma^{7/2} T^2 + \gamma^6 T^4 \right)
$$

$$
+ C_2 \| a_0 - a \|^2 \left[ e^{-c_2 \zeta \gamma B \lambda_{\min} t_0} \min \left\{ 1, \frac{1}{(N - t_0) \zeta \gamma B \lambda_{\min}} \right\} \right]
$$

$$
+ C_3 \gamma^2 R \sigma^2 \frac{1}{T^{\alpha/2-2}}
$$

(91)
Summing over all the rows we get a bound on the Frobenius norm. Lastly if the event $D_{0,N-1}$ does not occur, the $A_{t_0,N}$ is the zero matrix and hence

$$
\mathbb{E}\left[\|\hat{A}_{t_0,N} - A^*\|^2 \right] \leq \|A^*\|^2 \frac{1}{T^\alpha}
$$

(92)

Therefore:

$$
\mathbb{E}\left[\|\hat{A}_{t_0,N} - A^*\|^2 \right] \leq C \frac{d^2 \beta}{\xi^2 \lambda_{\min}(G) B(N - t_0)} + \bar{P} \cdot d \left( \gamma^{7/2}T^2 + \gamma^6T^4 \right)
$$

$$+ C \|A_0 - A^*\|^2 F \left[ e^{-c_2 \gamma B \lambda_{\min} t_0} \right]
$$

$$+ C\gamma^2 R \sigma^2 d \frac{1}{T^\alpha/2-2}
$$

(93)

\[\square\]

### D Concentration Under Stationary Measure

In this section, we will consider the process NLDS($A^*, \mu, \phi$) and the concentration of measure under its stationary distribution. In what follows, we will use the fact that $\phi$ is 1-Lipschitz as in the definition of NLDS, even when we don’t explicitly use Assumption 1.

**Proposition 2.** Under Assumption 2 with $\rho < 1$, the process is exponentially Ergodic and has a stationary distribution $\pi$. Suppose $X \sim \pi$ then $\mathbb{E}\|X\|^4 < \infty$.

The result follows from a technique similar to the one used for Proposition 49 by considering the process in the space of measures endowed with the Wasserstein metric.

#### D.1 Sub-Gaussian Case: Stable Systems

We will first consider the process with $X_0 = 0$ and prove concentration for $X_t$ for arbitrary $t$, and then use distributional convergence results to prove the concentration results at stationarity. First, we prove some preparatory lemmas.

**Lemma 10.** Suppose $Y$ is a $\nu^2$ sub-Gaussian random variable with zero mean. Then, for any $\lambda \leq \frac{1}{\pi^2}$, we have:

$$
\mathbb{E}\exp(\lambda Y^2) \leq 1 + 8\nu^2.
$$

We refer Section \[4\] for the proof. Now, consider the random variable $Z_{t+1} = \|X_{t+1}\|^2 - \sum_{s=0}^{t} \rho^{t-s} \|\eta_s\|^2$. By assumption, we have $X_0 = 0$. Therefore we must have $Z_0 = 0$. We have the following lemma:

**Lemma 11.** Suppose that Assumptions 3 and 5 hold and $\rho$ be as given in Assumption 5. For any $\lambda$ such that $0 \leq \lambda \leq \frac{1 - \rho}{\pi^2 c \sigma^2}$, we have:

$$
\mathbb{E}\exp(\lambda Z_{t+1}) \leq 1.
$$

**Proof.** First by mean value theorem, we must have: $\phi(A^* X_t) = \phi(A^* X_t) - \phi(0) = DA^* X_t$ for some diagonal matrix $D$ with entries lying in $[0, 1]$. Therefore, $\|\phi(A^* X_t)\| \leq \|D\|\|A^*\|\|X_t\| \leq \rho\|X_t\|$. Using this in Equation \[1\], we conclude:

$$
\|X_{t+1}\|^2 - \|\eta_t\|^2 = \|\phi(A^* X_t)\|^2 + 2\langle \eta_t, \phi(A^* X_t) \rangle
$$

$$\leq \rho^2\|X_t\|^2 + 2\langle \eta_t, \phi(A^* X_t) \rangle
$$

(94)

Let $F_s = \sigma(X_0, \eta_0, \ldots, \eta_s)$. It is clear that $X_s \in F_{s-1}$. Using Equation \[94\], we conclude:
\[
\mathbb{E} \left[ \exp(\lambda Z_{t+1}) | \mathcal{F}_{t-1} \right] = \mathbb{E} \left[ \exp \left( \lambda \|X_{t+1}\|^2 - \lambda \|\eta_t\|^2 \right) | \mathcal{F}_{t-1} \right] \exp \left( -\lambda \sum_{s=0}^{t-1} \rho^{t-s} \|\eta_s\|^2 \right) \\
\leq \mathbb{E} \left[ \exp \left( \lambda \rho^2 \|X_t\|^2 + 2\lambda \langle \eta_t, \phi(A^* X_t) \rangle \right) | \mathcal{F}_{t-1} \right] \exp \left( -\lambda \sum_{s=0}^{t-1} \rho^{t-s} \|\eta_s\|^2 \right) \\
\leq \exp \left( \lambda \rho^2 \|X_t\|^2 + 2\lambda^2 C_d \sigma^2 \|\phi(A^* X_t)\|^2 \right) \exp \left( -\lambda \sum_{s=0}^{t-1} \rho^{t-s} \|\eta_s\|^2 \right) \\
\leq \exp \left( \lambda \rho \|X_t\|^2 \right) \exp \left( -\lambda \sum_{s=0}^{t-1} \rho^{t-s} \|\eta_s\|^2 \right) \\
= \exp(\lambda \rho Z_t) \\
\text{(95)}
\]

In the fourth step, we have used the fact that \(\|\phi(A^* X_t)\| \leq \rho \|X_t\|\). In the fifth step we have used the assumption that \(\rho \leq \frac{1}{2d C_d \sigma^2}\) to show \(\lambda \rho^2 + 2\lambda^2 \rho^2 C_d \sigma^2 \leq \rho \lambda\). In the last step, we have used the definition of \(Z_t\). We iterate over Equation (95) and use the fact that \(Z_0 = 0\) almost surely to conclude that whenever \(\lambda \leq \frac{1}{2d C_d \sigma^2}\), we must have:

\[
\mathbb{E} \exp(\lambda Z_{t+1}) \leq \mathbb{E} \exp(\lambda Z_0) = 1.
\]

Now, let \(Y_{t+1} = \sum_{s=0}^{t} \rho^{t-s} \|\eta_s\|^2\). We will now use Lemma \[10\] to bound \(\mathbb{E} \exp(\lambda Y_{t+1})\) for \(\lambda > 0\) small enough.

**Lemma 12.** Suppose that Assumptions [3] and [5] hold and \(\rho\) be as given in Assumption [5]. For any \(\lambda\) such that \(0 \leq \lambda \leq \frac{1}{4d C_d \sigma^2}\), we have:

\[
\mathbb{E} \exp(\lambda Y_{t+1}) \leq \exp(8\lambda d C_d \sigma^2 \frac{\rho^2}{1-\rho})
\]

**Proof.** Let \(N(\beta) := \mathbb{E} \exp(\beta \|\eta_s\|^2)\). By independence of the noise sequence, we have:

\[
\mathbb{E} \exp(\lambda Y_{t+1}) = \prod_{s=0}^{t} N(\rho^{t-s}\lambda)
\]

For \(\beta \leq \frac{1}{4d C_d \sigma^2}\),

\[
N(\beta) = \mathbb{E} \exp(\beta \|\eta_s\|^2) = \mathbb{E} \exp(\beta \sum_{i=1}^{d} \langle e_i, \eta_s \rangle^2) \\
\leq \frac{1}{d} \sum_{i=1}^{d} \mathbb{E} \exp(\beta d \langle e_i, \eta_s \rangle^2) \leq 1 + 8\beta d C_d \sigma^2
\]

In the last step, we have used Jensen’s inequality for the function \(x \rightarrow \exp(x)\) and then invoked the Lemma \[10\]. Plugging this into Equation (96), we conclude:

\[
\mathbb{E} \exp(\lambda Y_{t+1}) \leq \prod_{s=0}^{t} \left( 1 + 8\lambda d \rho^{t-s} C_d \sigma^2 \right) \leq \exp(\sum_{s=0}^{t} 8\lambda d \rho^{t-s} C_d \sigma^2) \\
\leq \exp(8\lambda d C_d \sigma^2 \frac{\rho^2}{1-\rho})
\]

\(\blacksquare\)
Based on Lemmas 12 and 11, we will now state the following concentration inequality:

**Theorem 8.** Suppose Assumptions 3 and 5 hold and ρ be as given in Assumption 5. Let X be distributed according π, the stationary distribution of NLDS(A∗, µ, φ). Then, for any 0 < λ ≤ λ∗ := min(\(\frac{1}{8dCnσ^2}, \frac{1-ρ}{4ρCnσ^2}\)), we have:

\[
E \exp(λ\|X\|^2) \leq \exp(\frac{8λdCnσ^2}{1-ρ}) .
\]

We conclude:

1. Applying Chernoff bound with \(λ = λ^∗\), we conclude:

\[
\mathbb{P}\left(\|X\|^2 > \frac{8dCnσ^2}{1-ρ} + β\right) \leq \exp(-λ^∗β) .
\]

2. \(E\|X\|^2 \leq \frac{8dCnσ^2}{1-ρ}\)

The conclusions still hold when X is replaced by \(X_t\) for any \(t \in \mathbb{N}\) for the process started at 0.

**Proof.** We first note that \(\|X_{t+1}\|^2 = Z_{t+1} + Y_{t+1}\). Therefore, by Cauchy-Schwarz inequality, we must have:

\[
E \exp(λ\|X_{t+1}\|^2) = E \exp(λ(Z_{t+1} + Y_{t+1})) \\
\leq \sqrt{E}\exp(2λZ_{t+1})\sqrt{E}\exp(2λY_{t+1}) \\
\leq \exp(\frac{8λdCnσ^2}{1-ρ}) . (99)
\]

Here we have used Lemmas 12 and 11 and the appropriate bounds on λ. Recall that we started the chain \((X_t)\) with \(X_0 = 0\). Denote the law of \(X_t\) by \(π_t\). By proposition 2, we show that \(π_t\) converges weakly to the stationary distribution \(π\). We invoke Skhorokhod representation theorem to show that there exist random variables \(\bar{X}_t \sim π_t\) and \(X \sim π\) for \(t \in \mathbb{N}\), defined on a common probability space such that \(\bar{X}_t \to X\) almost surely. Now, we have shown that:

\[
E \exp(λ\|\bar{X}_{t+1}\|^2) \leq \exp(\frac{8λdCnσ^2}{1-ρ}) .
\]

Now, applying Fatou’s Lemma to the equation above as \(t \to \infty\), we conclude:

\[
E \exp(λ\|X\|^2) \leq \exp(\frac{8λdCnσ^2}{1-ρ}) . (100)
\]

The concentration inequality follows from an application of Chernoff bound and the second moment bound follows from Jensen’s inequality to Equation (100) (i.e, \(E\exp(Y) ≥ \exp(EY)\)).

**D.2 Sub-Gaussian Case: Possibly Unstable Systems**

We consider the case with \((Cρ, ρ)\) regularity, but we allow \(ρ > 1\).

**Lemma 13.** Under Assumption 4 we have:

\[
\|X_t\| \leq Cρ \sum_{s=0}^{t-1} ρ^{t-s-1}\|η_s\| . (101)
\]

No suppose Assumption 3 also holds. Let \(δ ∈ (0, 1/2)\). Then with probability atleast \(1 − δ\), we must have:

\[
\sup_{0 ≤ t ≤ T} \|X_t\| ≤ CCρ√CnS(ρ, T)σ√d\log(\frac{T}{δ}) .
\]

Where \(S(ρ, T) := \sum_{t=0}^{T-1} ρ^{T-t-1}\) and C is some universal constant.
Theorem 9. Suppose Assumption 4 holds with $\rho < 1$. Suppose that $X$ is distributed as the stationary distribution $\pi$ of the system NLDS($A^*, \mu, \phi$). Then, we have:

1. $\mathbb{E}\|X\|^4 \leq \frac{C_4^2 M_4}{(1 - \rho)^2}.$

Where we recall $M_4 = \mathbb{E}\|\eta_t\|^4$.

2. $\mathbb{E}\|X\|^2 \leq \frac{C_2^2 \sigma^2}{(1 - \rho)^2}$.

Proof. We use Equation (101) to conclude the desired bound for $X_T$ when the process is started with $X_0 = 0$. We then use Fatou’s lemma along with Skhorokhod Representation theorem like in Theorem 8 to conclude the bound at stationarity.

D.3 Heavy Tailed Case: Stable Systems

Theorem 9. Suppose Assumption [4] holds with $\rho < 1$. Suppose that $X$ is distributed as the stationary distribution $\pi$ of the system NLDS($A^*, \mu, \phi$). Then, we have:

1. $\mathbb{E}\|X\|^4 \leq \frac{C_4^2 M_4}{(1 - \rho)^2}.$

Where we recall $M_4 = \mathbb{E}\|\eta_t\|^4$.

2. $\mathbb{E}\|X\|^2 \leq \frac{C_2^2 \sigma^2}{(1 - \rho)^2}$.

Proof. We use Equation (101) to conclude the desired bound for $X_T$ when the process is started with $X_0 = 0$. We then use Fatou’s lemma along with Skhorokhod Representation theorem like in Theorem 8 to conclude the bound at stationarity.

E Well Conditioned Second Moment Matrices

In this section we will consider a stationary sequence $X_0, \ldots, X_T$ derived from the process NLDS($A^*, \mu, \phi$), with the corresponding noise sequence $\eta_0, \ldots, \eta_T$. We want to show that the matrix $\frac{1}{T} \sum_{t=0}^{T-1} X_t X_t^\top$ behaves similar to $G := \mathbb{E} X_t X_t^\top$. To do this, we will first to control the quantity: $\mathbb{E} (X_t, x)^2 (X_s, x)^2$ for arbitrary fixed vector $x \in \mathbb{R}^d$. Clearly, $\mathbb{E} (X_t, x)^2 = x^\top G x$.

Lemma 14. Without loss of generality, we suppose that $t > s$. Suppose $X_0, \ldots, X_T$ be a stationary sequence from NLDS($A^*, \mu, \phi$).

1. Suppose Assumptions [3] and [5] hold. Then we have:

$$\mathbb{E} (X_t, x)^2 (X_s, x)^2 \leq 2 (x^\top G x)^2 + \tilde{C}_1 \rho^{2(t-s)} \frac{\sigma^2}{1 - \rho} x^\top G x \log \left( \frac{d}{1 - \rho} \right)$$

Where $\tilde{C}_1$ depends only on $C_\eta$.

2. Suppose Assumption [6] holds with $\rho < 1$. Recall that $M_4 = \mathbb{E}\|\eta_t\|^2$. We have:

$$\mathbb{E} (X_t, x)^2 (X_s, x)^2 \leq 2 (x^\top G x)^2 + 8 \|x\|^2 C_\rho \rho^{2(t-s)} \frac{M_4}{(1 - \rho)^4}$$
We will give the proof of this lemma in Section H. Now, consider the random matrix \( \hat{G}_B := \frac{1}{p} \sum_{t=0}^{B-1} X_t X_t^\top \). Clearly, \( \hat{G}_B \succeq 0 \) and because of stationarity, \( EG_B = G \). We write down the following lemma:

**Lemma 15.** Suppose \( X_0, \ldots, X_T \) be a stationary sequence from \( \text{NLDS}(A^*, \mu, \phi) \).

1. Suppose Assumptions 3 and 5 hold. Let \( \bar{C}_1 \) be as in Lemma 14. Suppose \( B \geq \bar{C}_1 \frac{d}{(1-\rho)(1-\rho^2)\log \left( \frac{d}{1-\rho} \right)} \). Then, for any fixed vector \( x \in \mathbb{R}^d \),

\[
\mathbb{P} \left( x^\top \hat{G}_B x \geq \frac{1}{2} x^\top G x \right) \geq p_0 > 0.
\]

Where \( p_0 \) is a universal constant which can be taken to be \( \frac{1}{10} \). Furthermore, for any event \( A \) such that \( \mathbb{P}(A) > 1 - p_0 \), we must have:

\[
\mathbb{P} \left( x^\top \hat{G}_B x \geq \frac{1}{2} x^\top G x \bigg| A \right) \geq q_0 := \frac{p_0 - \mathbb{P}(A^c)}{\mathbb{P}(A)} > 0.
\]

2. Suppose Assumption 4 holds with \( \rho < 1 \). Whenever \( B \geq \frac{8C_6 M_4}{(1-\rho)^2(1-\rho^2)\log \left( \frac{d}{1-\rho} \right)} \),

\[
\mathbb{P} \left( x^\top \hat{G}_B x \geq \frac{1}{2} x^\top G x \right) \geq p_0 > 0.
\]

**Proof.** Without loss of generality, take \( \|x\| = 1 \). We start with the Paley-Zygmund inequality. Let \( Z \) be any random variable such that \( Z \geq 0 \) almost surely and \( \mathbb{E}Z^2 < \infty \). For any \( \theta \in [0, 1] \) we must have:

\[
\mathbb{P}(Z \geq \theta \mathbb{E}Z) \geq (1 - \theta)^2 \frac{(\mathbb{E}Z)^2}{\mathbb{E}Z^2}.
\]

Now consider \( Z = x^\top \hat{G}_B x \) and \( \theta = \frac{1}{2} \).

1. The simple calculation shows that:

\[
\mathbb{P} \left( x^\top \hat{G}_B x \geq \frac{1}{2} x^\top G x \right) \geq \frac{1}{4} \frac{B^2(x^\top G x)^2}{\sum_{s,t=0}^{B-1} \mathbb{E}(X_s, x)^2(X_t, x)^2} \\
\geq \frac{1}{4} \frac{B^2(x^\top G x)^2}{2B^2(x^\top G x)^2 + \sum_{s,t=0}^{B-1} \bar{C}_1 \rho^{2(|t-s|)} \frac{ds^2}{1-\rho} x^\top G x \log \left( \frac{d}{1-\rho} \right)} \\
\geq \frac{1}{4} \frac{B^2(x^\top G x)^2}{2B^2(x^\top G x)^2 + 2 \sum_{t=0}^{B-1} \bar{C}_1 \frac{ds^2}{(1-\rho)(1-\rho^2)} x^\top G x \log \left( \frac{d}{1-\rho} \right)} \\
= \frac{1}{4} \frac{B^2(x^\top G x)^2}{2B^2(x^\top G x)^2 + 2B\bar{C}_1 \frac{ds^2}{(1-\rho)(1-\rho^2)} x^\top G x \log \left( \frac{d}{1-\rho} \right)} \\
= \frac{1}{8} \frac{1}{1 + \tau_B}.
\]

(103)

Here, \( \tau_B := \frac{\bar{C}_1}{x^\top G x} \frac{ds^2}{B(1-\rho)(1-\rho^2)} \log \left( \frac{d}{1-\rho} \right) \). In the second step we have used item 1 of Lemma 14. In the third step, we have summed the infinite series \( \sum_{s \geq 1} \rho^{2(t-s)} \). Using the hypothesis that \( B \geq \bar{C}_1 \frac{d}{(1-\rho)(1-\rho^2)\log \left( \frac{d}{1-\rho} \right)} \) and \( G \succeq \sigma^2 I \), we conclude the result.

2. We proceed similarly as above, but use item 2 in Lemma 14 instead.
We will now follow the method used to prove [19, Lemma 31]. We now consider the matrix $\hat{H}_{0,B-1}^s$ under the event $\hat{D}_{s,0}^a$ in order to prove Theorem [10], where the terms are as defined in Section C.3. For the sake of clarity, we will drop the superscripts $s$.

**Remark 5.** We prove the results below for $\hat{H}_{0,B-1}^s$ but they hold unchanged when the matrices are all replaced with $\hat{H}_{0,B-1}^{s,T}$ given that we reverse the order of taking products whenever they are encountered.

**Lemma 16.** Suppose Assumption [I] holds. Suppose that $\gamma RB < \frac{1}{4}$. Then, for any buffer $s$, under the event $\hat{D}_{s,0}^a$, we have:

$$I - 4\gamma \left(1 + \frac{2\gamma BR}{1 - 4\gamma BR}\right) \sum_{i=0}^{B-1} \tilde{X}_{i}^{s} \tilde{X}_{i}^{s,T} \leq \hat{H}_{0,B-1}^{s} \hat{H}_{0,B-1}^{s,T} \leq I - 4\gamma \left(\zeta - \frac{2\gamma BR}{1 - 4\gamma BR}\right) \sum_{i=0}^{B-1} \tilde{X}_{i}^{s} \tilde{X}_{i}^{s,T}$$

In particular, whenever we have $\gamma BR \leq \frac{\zeta}{4(1+\gamma)}$, we must have:

$$I - 4\gamma \left(1 + \frac{\zeta}{2}\right) \sum_{i=0}^{B-1} \tilde{X}_{i}^{s} \tilde{X}_{i}^{s,T} \leq \hat{H}_{0,B-1}^{s} \hat{H}_{0,B-1}^{s,T} \leq I - 2\gamma \zeta \sum_{i=0}^{B-1} \tilde{X}_{i}^{s} \tilde{X}_{i}^{s,T}$$

**Proof.** The proof follows from the proof of [19, Lemma 28] with minor modifications to account for the fact that $\phi'(\beta) \in [\zeta, 1]$. □

Combining Lemma 16 with Lemma 14, we will show that $\hat{H}_{0,B-1}^s$ contracts any given vector with probability at-least $p_0 > 0$.

**Theorem 10.** Suppose Assumptions [I], [II] and [III] hold. Assume that $B$ and $\gamma$ are such that: $B \geq \bar{C}_1 \frac{d}{(1-\rho)(1-\rho^2)} \log \left(\frac{d}{1-\rho}\right)$ and $\gamma BR \leq \frac{\zeta}{4(1+\gamma)}$, where $\bar{C}_1$ is as given in Lemma 15. We also assume that $\mathbb{P}(\hat{D}_{b,a}^b) > \max\left(\frac{3}{2}, 1 - \frac{p_0}{2}\right)$, where $p_0$ is as given in Lemma 15. Let $\alpha \geq b$. Let $\lambda_{\min}(G)$ denote the smallest eigenvalue of $G$. Conditioned on the event $\hat{D}_{b,a}^b$,

(1) $||\prod_{s=0}^{b} \hat{H}_{0,B-1}^{s,T}|| \leq 1$ almost surely.

(2) Whenever $b - a + 1$ is larger than some universal constant $C_0$,

$$\mathbb{P}\left(||\prod_{s=a}^{b} \hat{H}_{0,B-1}^{s,T}|| \geq 2(1 - \zeta \gamma B \lambda_{\min}(G))^{c_4(a-b+1)} \bigg| \hat{D}_{b,a}^b\right) \leq \exp(-c_3(a-b+1) + c_5d)$$

Where $c_3, c_4$ and $c_5$ are universal constants.

**Proof.** The proof of (1) above follows from an application of Lemma 16. So we will just prove (2). We will prove this with an $\epsilon$ net argument over the unit $\ell^2$ sphere in $\mathbb{R}^d$.

Suppose we have arbitrary $x \in \mathbb{R}^d$ such that $||x|| = 1$. Let $K_v := \prod_{s=0}^{b} \hat{H}_{0,B-1}^{s,T}$. When $v \leq b$, we take this product to be identity. Now, define $\hat{G}_{b}^v := \frac{1}{B} \sum_{j=0}^{B-1} X_v^j X_j^{v,T}$. Consider the class of events indexed by $v$: $\mathcal{G}_v := \{||\hat{H}_{0,B-1}^{v,T} K_v x||^2 \leq ||K_v x||^2 (1 - \gamma \zeta \lambda_{\min}(G))\}$. From Lemma 15, we will prove the following claim:

**Claim 6.** Whenever $v \in [b, a] \cap \mathbb{Z}$:

$$\mathbb{P}(\mathcal{G}_v \cap \hat{D}_{b,a}^b, \hat{H}_{0,B-1}^{s,T} : s < v) \leq 1 - q_0$$

(104)

Where $q_0 > 0$ is as given in Lemma 15 and can be taken to be a universal constant under the present hypotheses.
Proof. We will denote $K_{v-1}x$ by $x_v$, for the sake of convenience. We note that when conditioned on \( \hat{H}^{s,T}_{0,B-1} \) for \( s < v \), \( x_v \) is fixed. Using Lemma 16, we note that:

\[
\mathbb{P}(G^v_B|\hat{D}^{b,a},\hat{H}^{s,T}_{0,B-1} : s < v) \leq \mathbb{P}(x_v \geq \hat{G}^v_Bx_v < \frac{1}{2}x_vGx_v|\hat{D}^{b,a},\hat{H}^{s,T}_{0,B-1} : s < v)
\]

We note that $\hat{G}^v_B$ is independent of $\hat{H}^{s,T}_{0,B-1}$ for \( s \leq v \) (eventhough $\hat{H}^{v}_{0,B-1}$ is not necessarily). Now we also note that $\hat{G}^v_B$ is independent of $\hat{D}^s$ for \( s \neq v \). Therefore, we can apply Lemma 15 to conclude the claim.

Let $D \subseteq \{b, \ldots, a\}$ such that $|D| = r$. It is also clear from item 1 and the definitions above that whenever the event $\cap_{v \in D}G_v$ holds, we have:

\[
\| \prod_{s=a}^{b} \hat{H}^{s,T}_{0,B-1}x \| \leq (1 - \gamma B\lambda_{\min}(G))^{\frac{39}{2}} .
\]

(105)

Therefore, whenever Equation (105) is violated, we must have a set $D^c \subseteq \{b, \ldots, a\}$ such that $|D^c| \geq b - a - r$ and the event $\cap_{v \in D^c}G_v$ holds. We will union bound all such events indexed by $D^c$ to obtain an upper bound on the probability that Equation (105) is violated. Therefore, using Equation (104) along with the union bound, we have:

\[
\mathbb{P}(\| \prod_{s=a}^{b} \hat{H}^{s,T}_{0,B-1}x \| \geq (1 - \gamma B\lambda_{\min}(G))^{\frac{39}{2}D^{b,a}} | \hat{D}^{b,a}) \leq \exp(-c_3(b - a + 1))
\]

Whenever $a - b + 1$ is larger than some universal constant, we can pick $r = c_2(b - a + 1)$ for some constant $c_2 > 0$ small enough such that:

\[
\mathbb{P}(\| \prod_{s=a}^{b} \hat{H}^{s,T}_{0,B-1}x \| \geq (1 - \gamma B\lambda_{\min}(G))^{\frac{39}{2}D^{b,a}} | \hat{D}^{b,a}) \leq \exp(-c_3(a - b + 1))
\]

(106)

Now, let $\mathcal{N}$ be a $1/2$-net of the sphere $S^{d-1}$. Using Corollary 4.2.13 in [50], we can choose $|\mathcal{N}| \leq 6^d$. By Lemma 4.4.1 in [50] we show that:

\[
\| \prod_{s=a}^{b} \hat{H}^{s,T}_{0,B-1} \| \leq 2\sup_{x \in \mathcal{N}} \| \prod_{s=a}^{b} \hat{H}^{s,T}_{0,B-1}x \|
\]

(107)

By union bounding Equation (106) for every $x \in \mathcal{N}$, we conclude that:

\[
\mathbb{P}(\| \prod_{s=a}^{b} \hat{H}^{s,T}_{0,B-1} \| \geq 2(1 - \zeta B\lambda_{\min}(G))^{c_3(b-a+1)} | \hat{D}^{b,a}) \leq |\mathcal{N}| \exp(-c_3(a - b + 1))
\]

Equation (108)

We will now state the equivalent of [19] Lemma 32. The proof proceeds similarly, but using Theorem 10 instead. Consider the following operator:

\[
 Fa,N := \sum_{t=a}^{N-1} \prod_{s=t}^{a+1} \hat{H}^{s,T}_{0,B-1}
\]

(109)

Here we choose the convention that whenever $s > t$, then in any product involving $\hat{H}^{s,T}_{0,B-1}$ and $\hat{H}^{t}_{0,B-1}$, $s$ appears to the right of $t$. Hence, we use the take $\prod_{s=t}^{a+1} \hat{H}^{s,T}_{0,B-1} = I$

Theorem 11. Suppose all the conditions in Theorem 10 hold. Then, for any $\delta \in (0, 1)$, we have:

\[
\mathbb{P}(\| Fa,N \| \geq C \left(d + \log \frac{N}{\delta} + \frac{1}{\zeta B\lambda_{\min}(G)}\right) | \hat{D}^{a,N}) \leq \delta
\]

Where $C$ is a universal constant.
F  Self Normalized Noise Concentration

We recall the events defined in Section 3:

1. $\mathcal{D}_T(R) := \{\sup_{0 \leq t \leq T} \|X_t\|^2 \leq R\}$
2. $\mathcal{E}_T(\kappa) := \{\hat{G} \geq \frac{\sigma^2}{T}\}$
3. $\mathcal{D}_T(R, \kappa) := \mathcal{D}_T(R) \cap \mathcal{E}_T(\kappa)$

From Lemma 13, we conclude that taking $R \geq C^2_3 \sigma^2 (S(\rho, T))^2 \sigma^2 \log(\frac{2T}{\delta})$ ensures that $\mathbb{P}(\mathcal{D}_T(R)) \geq 1 - \frac{\delta}{4}$. Only in this section, we define the following:

1. $\hat{X}_t := \phi(A^* X_t)$
2. $\hat{K}_X := \frac{1}{T} \sum_{t=0}^{T-2} X_t \eta_t\eta_t^\top + \eta_t \hat{X}_t^\top$
3. $\hat{G} := \frac{1}{T} \sum_{t=0}^{T-2} \hat{X}_t \hat{X}_t^\top$
4. $\hat{K}_\eta := \frac{1}{T} \sum_{t=0}^{T-2} \eta_t \eta_t^\top$

Lemma 17. Let $\delta \in (0, \frac{1}{2})$ Take $R = C^2_3 \sigma^2 (S(\rho, T))^2 \sigma^2 \log(\frac{2T}{\delta})$, $\kappa = 2$ and suppose $T \geq C_3 (d \log \left( \frac{N}{\delta^2} \right) + \log \frac{1}{\delta})$ for some constant $C_3$ depending only on $C_\eta$. Then, we have:

$$\mathbb{P}(\mathcal{D}_T(R, \kappa)) \geq 1 - \delta$$

Proof. Consider $\hat{G} = \frac{1}{T} \sum_{t=0}^{T-1} X_t X_t^\top = \frac{1}{T} \sum_{t=0}^{T-2} X_t \eta_t\eta_t^\top + \eta_t \hat{X}_t^\top$. For this proof only, we will define. To show the result, we will prove that $\hat{K}_X$ is not too negative with high probability and that $\hat{K}_\eta$ dominates identity with high probability. Let $x \in S^{d-1}$ and $\lambda \in \mathbb{R}$ Note that due to the sub-Gaussianity of $\eta_t$ and the definition of the process,

$$M_s := \exp \left( \sum_{s=0}^{t} \lambda \langle x, \eta_s \rangle \langle x, \hat{X}_s \rangle - \frac{C_\eta \sigma^2 \lambda^2}{2} \langle \hat{X}_s, x \rangle^2 \right)$$

is a super martingale with respect to the filtration $\mathcal{F}_t := \sigma(X_0, \eta_0, \ldots, \eta_t)$. We conclude that $\mathbb{E} M_{T-1} \leq 1$. An application of Chernoff bound shows that for every $\lambda, \beta > 0$, we must have:

$$\mathbb{P} \left( \sup_{t \in [0,T]} \langle x, \hat{K}_X x \rangle \geq 2C_\eta \sigma^2 \lambda x^\top \hat{G} x + \frac{\beta}{T} \right) \leq \frac{2}{1 - \delta} \exp(-\lambda \beta) \quad (109)$$

We will now invoke Theorem 5.39 in [51] to conclude that for some constant $\tilde{C}_2$ which depends only on $C_\eta$:

$$\mathbb{P} \left( \hat{K}_\eta \preceq \left( 1 - \tilde{C}_2 \left( \sqrt{\frac{d}{T}} + \sqrt{\frac{\log \frac{1}{\delta}}{T}} \right) \sigma^2 \right) \mathcal{D}_T(R) \right) \leq \frac{\delta}{4} \quad (110)$$

Consider any $\epsilon$ net $\mathcal{N}_\epsilon$ over $S^{d-1}$. By Corollary 4.2.13 in [51], we can take $|\mathcal{N}_\epsilon| \leq (1 + \frac{\epsilon}{\delta})^d$. From Equations (109) and (110), we conclude that conditioned on $\mathcal{D}_T(R)$, with probability at-least $1 - \frac{\delta}{4} - |\mathcal{N}_\epsilon| \exp(-\lambda \beta)$ we have:

$$\inf_{x \in S^{d-1}} x^\top \hat{G} x \geq \inf_{y \in \mathcal{N}_\epsilon} y^\top \hat{G} y - 2\|\hat{G}\| \epsilon$$

$$\geq \inf_{y \in \mathcal{N}_\epsilon} y^\top \hat{G} y - \|y^\top \hat{K}_X y\| + y^\top \hat{K}_\eta y - 2\|\hat{G}\| \epsilon$$

$$\geq \inf_{y \in \mathcal{N}_\epsilon} y^\top \hat{G} y - \|y^\top \hat{K}_X y\| + y^\top \hat{K}_\eta y - 2R\epsilon$$

$$\geq \inf_{y \in \mathcal{N}_\epsilon} y^\top \hat{G} y (1 - 2\lambda \sigma^2 C_\eta) - \frac{\beta}{T} + \sigma^2 \left( 1 - \tilde{C}_2 \left( \sqrt{\frac{d}{T}} + \sqrt{\frac{\log \frac{1}{\delta}}{T}} \right) \right) - 2R\epsilon$$

$$\quad (111)$$
In the third step, we have used the fact that under the event $\mathcal{D}_T(R)$, $\|\hat{G}\| \leq R$. Take $\lambda = \frac{1}{2\sigma^2 C_{\eta}}$ and $\epsilon = \frac{1}{\sqrt{d} \sigma}$ and $\beta = 2\sigma^2 d C_{\eta} \log (16\frac{R}{\sigma^2} + 1) + 2\sigma^2 C_{\eta} \log \frac{8}{\delta}$. We conclude that whenever $T \geq C_3 (d \log (\frac{R}{\sigma^2}) + \log \frac{1}{\delta})$ for some constant $C_3$ depending only on $C_{\eta}$, with probability at-least $1 - \frac{\delta}{2}$ conditioned on $\mathcal{D}_T(R)$, we have: $\hat{G} \succeq \frac{\sigma^2}{2} I$. In the definition of $\mathcal{E}_T(\kappa)$, we take $\kappa = 2$. Therefore, we must have:

$$\mathbb{P}(\mathcal{E}_T(\kappa) \cap \mathcal{D}_T(R)) = \mathbb{P}(\mathcal{E}_T(\kappa) | \mathcal{D}_T(R)) \mathbb{P}(\mathcal{D}_T(R)) \geq (1 - \frac{\delta}{2})^2 \geq 1 - \delta.$$  

We conclude the result from the equation above.

We now give the proof of Lemma 1.

### F.1 Proof of Lemma 1

**Proof.** We invoke Theorem 1 in [20] with $S_t = T \tilde{N}_t$, $V = T \sigma^2 I$, $\tilde{V}_t = V + T \hat{G}$, $\langle \eta, e_i \rangle$ is $C_{\eta} \sigma^2$ sub-Gaussian. So, we take ‘$R$’ in the reference to be $C_{\eta} \sigma^2$. Therefore, we conclude that with probability at least $1 - \delta$:

$$\tilde{N}_t^\top \tilde{V}_t^{-1} \tilde{N}_i \leq \frac{2 C_{\eta} \sigma^2}{T^2} \log \left( \frac{\det(\tilde{V}_t)^{1/2} \det(V)^{-1/2}}{\delta} \right). \quad (112)$$

Under the event $\mathcal{D}_T(R, \kappa)$, we must have: $\tilde{V}_t \leq \sigma^2 TI + TRI$. This implies:

$$\det(\tilde{V}_t)^{1/2} \det(V)^{-1/2} \leq (1 + \frac{R}{\sigma^2}) \frac{\delta}{2} \quad (113)$$

Now, observe that under the event $\mathcal{D}_T(R, \kappa)$, $\hat{G} \succeq \frac{\sigma^2}{2} I$. Therefore, $\tilde{V}_t \leq 3T \hat{G}$. This implies:

$$\frac{1}{3T} \tilde{N}_t^\top \hat{G}^{-1} \tilde{N}_i \leq \tilde{N}_t^\top \tilde{V}_t^{-1} \tilde{N}_i \quad (114)$$

Combining Equations (112), (113) and (114) and using Lemma 17, we conclude that with probability at-least $1 - 2\delta$, we have:

$$\tilde{N}_t^\top \hat{G}^{-1} \tilde{N}_i \leq \frac{6C_{\eta} \sigma^2}{T} \left[ d \log(1 + \frac{R}{\sigma^2}) + \log \left( \frac{1}{\delta} \right) \right]$$

Using union bound, we conclude the result.

### G Parameter Recovery Lower Bounds for ReLU-AR Model

We will show that in the case of non-expansive activation functions, parameter recovery can be hard information theoretically. More specifically, even when $\phi = \text{ReLU}$ and $\eta_i \sim \mathcal{N}(0, I)$, and $\| A^* \|_2 = \rho < 1$, we will need exponentially many samples with respect to $d$ in order estimate $A^*$ up to any vanishing accuracy. We do this via the two point method. Henceforth, we will assume that $d \geq 2$. Given $\epsilon > 0$, define with matrix $A(\epsilon) \in \mathbb{R}^{d \times d}$ as:

$$A_{ij}(\epsilon) = \begin{cases} 1 & \text{if } i = j; i \leq d - 1 \\ \frac{1}{4} & \text{if } i \neq j; i \leq d - 1 \\ \frac{1}{\sqrt{2}} & \text{if } i \neq j; i = d \\ 0 & \text{if } i = j = d \end{cases} \quad (115)$$

We will consider $\text{NLDS}(A(\epsilon), \mathcal{N}(0, I), \text{ReLU})$ and $\text{NLDS}(A(0), \mathcal{N}(0, I), \text{ReLU})$ as the two points in the two point method. As usual, we will consider $X_0 = 0$ almost surely for the sake of convenience. But it will be shown that this process is rapidly mixing since we intend to pick $\epsilon$ such that $\| A(\epsilon) \| < \frac{1}{2}$, so all the results should easily extend to stationary sequences. We collect some useful results in the following lemma:

**Lemma 18.**  

1. $\| A(\epsilon) \| = \sqrt{\frac{1}{16} + \epsilon^2}$
2. Suppose \( \epsilon \) is small enough such that \( \|A(\epsilon)\| < 1 \). Let \( X_0, X_1, \ldots, X_T \sim \text{NLDS}(A(\epsilon), \mathcal{N}(0, I), \text{ReLU}) \) with \( X_0 = 0 \). For some universal constant \( p_1 > 0 \), we have for every \( i \leq d, t \geq 0 \):

\[
\mathbb{P}\left( \langle X_{t+1}, e_i \rangle \geq \frac{1}{2} \left| X_t \right| \right) \geq p_1
\]

**Proof.**

1. Proof follows from elementary calculations.

2. \( X_{t+1} = \text{ReLU}(A(\epsilon)X_t) + \eta_t \). Therefore, for every \( i \leq d - 1 \), must have: \( \langle X_{t+1}, e_i \rangle \geq \langle \eta_t, e_i \rangle \). Therefore, we conclude that

\[
\mathbb{P}\left( \langle X_{t+1}, e_i \rangle \geq \frac{1}{2} \left| X_t \right| \right) \geq \mathbb{P}\left( \langle \eta_t, e_i \rangle \geq \frac{1}{2} \left| X_t \right| \right) \geq p_1 > 0 .
\]

We will now show that the last co-ordinate \( \langle X_{t+1}, e_d \rangle \) is just noise with a large probability and hence, we cannot estimate the last row of \( A(\epsilon) \) even with a large number of samples. Note that the event \( \langle X_{t+1}, e_d \rangle \neq \langle \eta_t, e_d \rangle \) is the same as the event \( \langle a_d(\epsilon), X_t \rangle > 0 \). Let \( a_d(\epsilon) \) denote the last row of \( A(\epsilon) \), in the form of a column vector.

**Lemma 19.** Suppose \( t \geq 2 \). Then, for some universal constants \( C_0, C_1 > 0 \),

\[
\mathbb{P}\left( \langle a_d(\epsilon), X_t \rangle > 0 \right) \leq C_0 \exp(-C_1 d) .
\]

**Proof.** \( \langle a_d, X_t \rangle > 0 \) if and only if \( \langle a_d, \text{ReLU}(A(\epsilon)X_{t-1}) \rangle + \langle a_d, \eta_{t-1} \rangle > 0 \). We first note that the first \( d - 1 \) rows of \( X_{t-1} \) are i.i.d. by the definition of \( A(\epsilon) \). Therefore, we conclude using Lemma 18 that:

\[
\mathbb{P}\left( \langle a_d, \text{ReLU}(A(\epsilon)X_{t-1}) \rangle \geq -c_0 \epsilon \sqrt{d} \right) \leq \exp(-c_1 d) .
\]

for some universal constants \( c_0, c_1 \). Now, \( \langle a_d, \eta_{t-1} \rangle \) is distributed as \( \mathcal{N}(0, \epsilon^2) \) and is independent of \( X_{t-1} \). Therefore, with a large probability, \( \langle a_d, \text{ReLU}(A(\epsilon)X_{t-1}) \rangle \) takes a large negative value and by Gaussian concentration, \( \langle a_d, \eta_{t-1} \rangle \) concentrates near 0. Therefore, using elementary calculations we conclude the result.

In what follows, by \( X_T^e := (X_0^e, \ldots, X_T^e) \) we denote the process such that \( X_0^e = 0 \) and \( (X_0^e, \ldots, X_T^e) \sim \text{NLDS}(A(\epsilon), \mathcal{N}(0, I), \text{ReLU}) \).

**Lemma 20.** When \( \epsilon < \frac{1}{T} \), for some universal constants \( c_0, c_1 > 0 \), we have:

\[
\text{TV}(X_T^e, X_T^0) \leq c_0 \left( \epsilon + \epsilon \sqrt{T} \exp(-c_1 d) \right) .
\]

**Proof.** Following the proof of Lemma 36 in [19], we conclude that:

\[
\text{KL}(X_T^e || X_T^0) = \frac{1}{2} \sum_{t=0}^{T-1} \mathbb{E}\|\text{ReLU}(A(\epsilon)X_t^e) - \text{ReLU}(A(0)X_t^e)\|^2
\]

(116)

Note here that the expectation is with respect to the randomness in the trajectory \( X^e \). Applying the definition of \( A(\epsilon) \) and \( \text{ReLU} \) to Equation (116), we further simplify to show that:

\[
\text{KL}(X_T^e || X_T^0) = \frac{1}{2} \sum_{t=0}^{T-1} \mathbb{E}|\text{ReLU}(\langle a_d(\epsilon), X_t^e \rangle)|^2 = \frac{1}{2} \sum_{t=1}^{T-1} \mathbb{E}|\text{ReLU}(\langle a_d(\epsilon), X_t^e \rangle)|^2
\]

(117)

In the equation above, we have used the fact that \( X_0^e = 0 \) almost surely. Now, when \( t = 1 \), we have \( X_1^e = \eta_1 \sim \mathcal{N}(0, I) \) almost surely. Therefore, \( \mathbb{E}|\text{ReLU}(\langle a_d(\epsilon), X_1^e \rangle)|^2 = \frac{\epsilon^2}{2} \). For \( t \geq 2 \), we will use Lemma 19 and Theorem 8. In this proof only, define \( \mathcal{P}_1 \) to be the event \( \{ \langle a_d, X_1^e \rangle > 0 \} \). Therefore, we conclude that:

42
\[
E[\text{ReLU}(a_d(\epsilon), X_c^\epsilon)]^2 = E[I(P_t) | (a_d(\epsilon), X_c^\epsilon)]^2 \\
\leq E[I(P_t) \| a_d(\epsilon) \|^2 | X_c^\epsilon]^2 \\
= E[I(P_t) \epsilon^2 \| X_c^\epsilon]^2 \\
\leq \epsilon^2 \sqrt{P(P_t)} \sqrt{E[\|X_c\|^4]} (118)
\]

From Lemma 19 we conclude that \(P(P_t) \leq C_0 \exp(-C_1 t)\) and from Theorem 8 it is clear that the concentration inequalities hold we show that \(\sqrt{E[\|X_t\|^4]} \leq C_2 t\) for some universal constant \(C_2\). Combining the results above with Equation (117), we obtain that for universal constants \(C_0, C_1\):

\[
\text{KL}(X_t^0 | X_t^0) \leq C_0 \left( \epsilon^2 d \exp(-C_1 t) + \epsilon^2 \right) (119)
\]

We then use Pinsker’s inequality to conclude the result.

**G.1 Proof of Theorem 4**

**Proof.** Let the notations below be as defined in the statement of the theorem. Since the minimax loss upper bounds any Bayesian loss, we will consider the uniform prior over \{NLDS(A(\epsilon), N(0, I), \text{ReLU}), NLDS(A(0), N(0, I), \text{ReLU})\}. For any estimator \(A\) with input \(X_0, \ldots, X_T\), we denote \(A(X)\) as its output. We use the notation \(X^\epsilon\) and \(X^0\) as defined in Lemma 20.

Therefore,

\[
\mathcal{L}(\Theta(\frac{1}{2}), T) \geq \inf_{A} \frac{1}{2} \mathcal{L}(A, T, A(0)) + \frac{1}{2} \mathcal{L}(A, T, A(\epsilon)) \\
= \frac{1}{2} \inf_{A} E[\|A(X^0) - A(0)\|^2] + E[\|A(X^\epsilon) - A(\epsilon)\|^2] (120)
\]

Where the expectation is over the respective trajectories \(X^\epsilon\) and \(X^0\). Now, from Lemma 20 and the coupling calculation of TV distance, we can define the trajectories \(X^\epsilon\) and \(X^0\) on a common probability space such that \(P(X^0 \neq X^\epsilon) \leq TV(X^0, X^\epsilon)\). Picking \(\epsilon\) to be small enough such that \(TV(X^0, X^\epsilon) \leq \frac{1}{2}\) (which is true for a choice of \(\epsilon^2 = c_0 \min\left(\frac{\exp(\epsilon^2 t)}{T}, 1\right)\) for universal constants \(c_0, c_1\)). Define the event \(S_T := \{X^0 = X^\epsilon\}\), we conclude from Equation (120) that:

\[
\mathcal{L}(\Theta(\frac{1}{2}), T) \geq \frac{1}{2} \inf_{A} \mathbb{E} I(S_T) \|A(X^0) - A(0)\|^2 + \mathbb{E} I(S_T) \|A(X^\epsilon) - A(\epsilon)\|^2 (121)
\]

Note that over the event \(S_T\), we must have \(A(X^\epsilon) = A(X^0)\) in the case of a deterministic estimator. In case of a random estimator, the proof is same once we observe that their distribution is same and hence can be coupled almost surely. By convexity, we must have: \(\|a - b\|^2 + \|b - c\|^2 \geq \frac{1}{2} \|a - c\|^2\). Combining these considerations into the equation above, we have:

\[
\mathcal{L}(\Theta(\frac{1}{2}), T) \geq \frac{P(S_T)}{4} \|A(\epsilon) - A(0)\|^2 \\
\geq \epsilon^2 \frac{8}{3}. (122)
\]

From the choice of \(\epsilon^2\) above, we conclude the result.
H   Proofs of Technical Lemmas

H.1 Proof of Lemma 10

Proof. The proof follows from integrating the tails. Let $Z := \exp(\lambda Y^2)$. For any $\gamma \in \mathbb{R}^+$, we have from the definition of sub-Gaussianity.

$$\mathbb{P}(Z \geq \gamma) = \begin{cases} 1 & \text{if } \gamma \leq 1 \\ \mathbb{P}(|Y| \geq \sqrt{\frac{\log(\gamma)}{\lambda}}) & \text{if } \gamma > 1 \end{cases}$$ (123)

Now,

$$E[Z] = \int_0^\infty \mathbb{P}(Z \geq \gamma) d\gamma$$

$$= \int_0^1 d\gamma + \int_1^\infty \mathbb{P}(|Y| \geq \sqrt{\frac{\log(\gamma)}{\lambda}}) d\gamma$$

$$\leq 1 + \int_1^\infty 2\exp\left(-\frac{\log(\gamma)}{2\nu^2}\right) d\gamma$$

$$= 1 + 2\int_1^\infty \gamma^{-\frac{1}{2\nu^2}} d\gamma$$

$$= 1 + \frac{4\nu^2\lambda}{1 - 2\nu^2\lambda}$$

$$\leq 1 + 8\lambda\nu^2$$ (124)

\[\square\]

H.2 Proof of Lemma 14

Proof. We draw $\tilde{X}_s \sim \pi$, independent of $X_s$. We obtain $\tilde{X}_{s+k}$ by running the markov chain with the same noise sequence. i.e., $\tilde{X}_{s+k+1} = \phi(A^*\tilde{X}_{s+k}) + \eta_{s+k}$. We then obtain $\tilde{X}_t$. Then, it is clear that:

$$(X_t, x)^2(X_s, x)^2 = (X_t - \tilde{X}_t + \tilde{X}_t, x)^2(X_s, x)^2$$

$$\leq 2(\tilde{X}_t, x)^2(X_s, x)^2 + 2(X_t - \tilde{X}_t, x)^2(X_s, x)^2$$

Taking expectation on both sides and noting that $\tilde{X}_t$ is independent of $X_s$, we conclude:

$$E(X_t, x)^2(X_s, x)^2 \leq 2(x^\top G x)^2 + 2E(X_t - \tilde{X}_t, x)^2(X_s, x)^2$$ (125)

By Assumption 4, we have: $\|X_t - \tilde{X}_t\|^2 \leq C\rho^2\rho^{2(t-s)}\|X_s - \tilde{X}_s\|^2$. Plugging this into Equation (125), we conclude:

$$E(X_t, x)^2(X_s, x)^2 \leq 2(x^\top G x)^2 + 2E\|x\|^2C\rho^2\rho^{2(t-s)}\|X_s - \tilde{X}_s\|^2(X_s, x)^2$$

$$\leq 2(x^\top G x)^2 + 4\|x\|^2C\rho^2\rho^{2(t-s)}E\left(\|X_s\|^2 + \|\tilde{X}_s\|^2\right)(X_s, x)^2$$

$$= 2(x^\top G x)^2 + 4\|x\|^2C\rho^2\rho^{2(t-s)}\left[E\|X_s\|^2(X_s, x)^2 + x^\top G x E\|\tilde{X}_s\|^2\right]$$ (126)

We can evaluate $E\|\tilde{X}_s\|^2$ from Theorems 8 and 9.

1. First we consider the Sub-Gaussian setting with $\|A^*\| = \rho < 1$ and $C\rho = 1$. Fix $R > 0$. We will use the notation from Theorem 8 below. We can then write,
will divide the times into buffers of size $B$. Recall that:

Consider the event and Theorem 9 before proceeding further, we will give a high probability bound on $M \geq B$ definition of the coupling, we have that:

$\hat{G}_t$ is a constant which depends only on $C_n$. Using Equation (128) in Equation (126), we conclude that:

$$\mathbb{E}(X_t, x) = 2(x^\top G x) \leq 2(x^\top G x)^2 + 2\|x\|^2 C^2 \rho^2 \delta (1-\rho)^2 \mathbb{E}\|X_s\|^4$$

We then use the bound on $\mathbb{E}\|X_s\|^4$ given in Theorem 8 to conclude the result.

H.3 Proof of Lemma 2

**Proof.** Consider $\hat{G} = \frac{1}{T} \sum_{t=0}^{T-1} X_t X_t^\top$ and consider the coupled process as in Definition 2. We will divide the times into buffers of size $B$, with gaps of size $u$ and also consider related notation as given in Section C. Now, $\hat{G} \geq \frac{1}{SN} \sum_{t=1}^N G^{(t)}$, where $G^{(t)}$ is the empirical second moment matrix of the buffer $t$ given by $\hat{G}^{(t)} := \frac{1}{B} \sum_{i=0}^{B-1} X_i^t X_i^t$. Now, consider the coupled second moment matrix defined on buffer $t$ given by $\tilde{G}^{(t)} := \frac{1}{B} \sum_{i=0}^{B-1} \tilde{X}_i^t \tilde{X}_i^t$. By Lemma 3, we know that $\|\tilde{G}^{(t)} - \hat{G}^{(t)}\| \leq 4 M^* C \rho u$. Where $M^* = \max_{t \leq T} \|X_t\|^2, \|X_t\|^2$. Now, observe that by definition of the coupling, we have that $\tilde{G}_t$ are i.i.d. Combining the considerations above, and letting $B \geq u$ to conclude $\frac{B}{u} \geq 1$

$$\hat{G} \geq \frac{B}{SN} \sum_{t=1}^N \hat{G}^{(t)} \geq \frac{1}{2N} \sum_{t=1}^N \hat{G}^{(t)} - 4 M^* C \rho u I$$

Before proceeding further, we will give a high probability bound on $M^*$. By Markov’s inequality, and Theorem 9

$$\mathbb{P}(M^* > R) \leq \frac{2T \mathbb{E}\|X_s\|^2}{R} \leq 2TC^2 \mathbb{E}\|X_s\|^2 \delta (1-\rho)^2 \alpha.$$ 

Consider the event $\hat{D}_T(R) = \{M^* \leq R\}$. Letting $R = \frac{4TC^2 \delta^2}{(1-\rho)^2 \alpha}$ only in this proof, we conclude that:

$$\mathbb{P}(\hat{D}_T(R)) \geq 1 - \delta$$

Recall that $G = \mathbb{E}X_t X_t^\top$. Now, as shown by item 2 in Lemma 15, we have whenever $B \geq \frac{4C^2 \mathbb{E}X_t X_t^\top}{(1-\rho)^2 (1-\rho^2)^2 \sigma^2}$

$$\mathbb{P}(x^\top \tilde{G}(t) x \geq \frac{1}{2} x^\top G x) \geq p_0 > 0.$$ (131)
Now, consider any arbitrary, fixed vector $x \in \mathbb{R}^d$. Using independence of $\hat{G}(t)$ and Equation \eqref{eq:inequality}, we conclude that for some universal constants $c_0, c_1$, we must have:

$$
P(\frac{1}{N} \sum_{t=1}^{N} x^T \hat{G}(t)x \leq c_0 x^T G x) \leq \exp(-c_1 N). \quad (132)$$

Rewriting the equation above by taking $x = G^{-1/2}y$, we have:

$$
P(\frac{1}{N} \sum_{t=1}^{N} y^T G^{-1/2} \hat{G}(t)G^{-1/2} y \leq c_0 \|y\|^2) \leq \exp(-c_1 N). \quad (133)$$

For simplicity of exposition, we will take $J := \frac{1}{N} \sum_{t=1}^{N} G^{-1/2} \hat{G}(t)G^{-1/2}$ in the calculations below. Before proceeding with a bound on the operator norm, we will give a bound on $\|J\|$. Since $G \succeq \sigma^2 I$, we must have: $\|J\| \leq \frac{M}{\sigma^2}$. Now, we will apply an epsilon net argument. Let $\mathcal{N}_\epsilon$ be an $\epsilon$-net over $S^{d-1}$. We can take $|\mathcal{N}_\epsilon| \leq (1 + \frac{2}{\epsilon})^d$.

$$
\inf_{y \in S^{d-1}} y^T J y \geq \inf_{y \in \mathcal{N}_\epsilon} y^T J y - 2\|J\|\epsilon.
$$

We let $\epsilon = \frac{c_2}{4}$ for some constant $c > 0$ small enough and $R$ as defined earlier in this proof. By union bound over $\mathcal{D}_\mathcal{N}^\epsilon$ and the event given in Equation \eqref{eq:inequality}, we conclude that for some universal constant $c_2 > 0$ small enough:

$$
P(\inf_{y \in S^{d-1}} y^T J y > c_2, M^* \leq R) \geq 1 - \exp(Cd \log \left( \frac{R}{\sigma^2} \right) - c_1 N) - \frac{\delta}{2} \quad (134)$$

Now, using Equation \eqref{eq:inequality}, we conclude that:

$$G^{-1/2} \hat{G} G^{-1/2} \succeq J - 4M^* C_C \rho^\mu G^{-1} - 4M^* C_C \rho^\mu G^{-1}$$

Using the fact that $G^{-1} \succeq \frac{I}{\sigma^2}$ and using Equation \eqref{eq:inequality}, we conclude that whenever $T \geq Cd \log \left( \frac{1}{K} \right) B \log (R/\sigma^2)$, $B \succeq \frac{4C_C M_4}{(1-\rho)^2(1-\rho^2)\sigma^2}$ and $u \geq \frac{\log \left( \frac{R}{\sigma^2} \right)}{\log \left( \frac{1}{\delta} \right)}$, we conclude that for some constant $c_0 > 0$ small enough, with probability at least $1 - \delta$, we have:

$$\hat{G} \succeq c_0 G$$

\[\square\]

**H.4 Proof of Lemma 7**

First, we will obtain a crude upper bound on $\|\hat{a}^{t-1} - a^*\|$ using Theorem 10. That is, we want to show that $\|\hat{a}^{t-1} - a^*\|$ does not grow too large with high probability.

**Proposition 3.** Let $\lambda_{\min} \equiv \lambda_{\min}(G)$. Conditional on $\hat{D}^{0,t-1} \cap \bigcap_{r=0}^{t-1} \mathcal{C}_0^{t-1}$, with probability at least $1 - N\delta$, for all $1 \leq t \leq N$, all $1 \leq j \leq B$ we have

$$
\|\hat{a}^{t-1} - a^*\| \leq \|a_0 - a\| + 2\gamma B \sqrt{R\beta C \left( d + \log \frac{N}{\delta} + \frac{1}{\zeta \gamma B \lambda_{\min}} \right)} \quad (135)
$$

where $C$ is constant depending only on $C_\eta$. 46
Proof. Let us start with the expression for $\hat{a}^{-1}_j - a^*$
\[ (\hat{a}^{-1}_j - a^*)^T = (a_0 - a^*)^T \left( \prod_{s=0}^{t-2} \tilde{H}_{0,s+1} + 2\gamma \sum_{i=0}^{j-1} \phi(\tilde{c}_i) \tilde{X}_{i+1}^T \tilde{H}_{i+1,j-1} \right) \tilde{H}^{t-1}_{0,j-1} + 2\gamma \sum_{r=2}^t \sum_{i=0}^{j-1} \phi(\tilde{c}_i)^r \tilde{X}_{i+1}^T \tilde{H}_{i+1,r-1}^r \left( \prod_{s=r-1}^{t} \tilde{H}_{s+1}^s \right) \]
(136)

We will work on the event $D^{t-1}_0 \cap \cap_{r=0}^{t-1} E_r$. It is clear from Equation (136) that:
\[ \|\hat{a}^{-1}_j - a^*\| \leq \|a_0 - a^*\| + 2\gamma B\sqrt{R\beta} \leq \|a_0 - a^*\| + 2\gamma B\sqrt{R\beta} \]
Thus taking union bound we get that conditional on $D^{t-1}_0 \cap \cap_{r=0}^{t-1} E_r$, for fixed $t$, with probability at least $1 - \delta$, for all $1 \leq j \leq B$
\[ \|\hat{a}^{-1}_j - a^*\| \leq \|a_0 - a^*\| + 2\gamma B\sqrt{R\beta} \]

Proof of Lemma 3 On the event $E_{0,j} \cap \tilde{D}^r, N^{-1}$, we note the following inequalities
\[ \tilde{a}^s_i = \tilde{a}^s_i, 0 \leq s < r, 0 \leq i \leq B - 1 \]
\[ \tilde{a}^r_0 = \tilde{a}^r_0 \]
\[ \|\tilde{a}^r_i - a^s_i\| \leq \begin{cases} 4\gamma \sqrt{R\beta} + \sum_{k=0}^{j-1} 4\gamma R \|\hat{a}^r_k - a^s_k\|, & s = r, 1 \leq i \leq j \\ 4(\gamma + 1) \sqrt{R\beta} + \sum_{k=0}^{j-1} 4\gamma R \|\hat{a}^r_k - a^s_k\|, & s = r, \ j + 1 \leq i \leq B - 1 \\ 4(\gamma + 1) \sqrt{R\beta} + \sum_{k=0}^{j-1} 4\gamma R \|\hat{a}^r_k - a^s_k\|, & r < s, \ 0 \leq i \leq B - 1 \end{cases} \]
(139)

The result then follows from an application of Proposition 3 with $\delta$ chosen as in C.4

H.5 Proof of Lemma 4

We first state and prove the following result:

**Lemma 21.** Let $R_{\text{max}} := \sup_{t \leq T}(\|X_t\|^2, \|X_t\|)$ and suppose $\gamma \leq \frac{1}{2R_{\text{max}}}$. For every $t \in [N]$ and $i \in [B]$ we have:
\[ \|a^*_i\| \leq 2\gamma R_{\text{max}} T \]

**Proof.** Let the row under consideration be the $k$-th row and $e_k$ be the standard basis vector. Consider the SGD – RER iteration:
\[ a^*_{t+1} = a^*_t - 2\gamma \left( \phi(a^*_t, X^*_t) - X^*_{(i-1)} \right) X^*_i \]
\[ = (I - 2\gamma \zeta_{i,t} X^*_i X^*_i^T) a^*_t + 2\gamma (X^*_t (i-1), e_k) X^*_i \]
(140)
Where \( \zeta_{t,i} := \frac{\phi(a^i_t, X^t_i)}{(a^i_t, X^t_i)} \in [\zeta, 1] \) exists in a weak sense due to our assumptions on \( \phi \). Observe that for our choice of \( \gamma \), we have \( \| \langle I - 2\gamma \zeta_{t,i} X^t_i, X^t_{i-1} \rangle \| \leq 1 \) and \( \| (X^t_{i-1}, e_k) X^t_{i-1} \| \leq R_{\max} \). Therefore, triangle inequality implies:

\[
\| a^t_{i+1} \| \leq \| a^t_i \| + 2\gamma R_{\max}
\]

We conclude the bound in the Lemma.

\[\square\]

**Proof of Lemma** [4] Let the row under consideration be the \( k \)-th row and \( e_k \) be the standard basis vector.

\[
a^t_{i+1} = a^t_i - 2\gamma(\langle a^t_i, X^t_i \rangle - \langle e_k, X^t_{i-(i-1)} \rangle)X^t_i
\]

\[
= a^t_i - 2\gamma(\langle a^t_i, \tilde{X}^t_i \rangle - \langle e_k, \tilde{X}^t_{i-(i-1)} \rangle)\tilde{X}^t_i + \Delta_{t,i}
\]

(141)

Where

\[
\Delta_{t,i} := 2\gamma \left( \phi(\langle a^t_i, \tilde{X}^t_i \rangle) \tilde{X}^t_i - \phi(\langle a^t_i, X^t_i \rangle) X^t_i \right) + 2\gamma \left( \langle X^t_{i-(i-1)}, e_k \rangle X^t_{i-1} - \langle \tilde{X}^t_{i-(i-1)}, e_k \rangle \tilde{X}^t_i \right).
\]

Using Lemmas 2 and 3 we conclude that:

\[
\| \Delta_{t,i} \| \leq (16\gamma^2 R_{\max}^2 T + 8\gamma R_{\max}) \rho^u
\]

(142)

In the last step we have used the fact that under the conditions on \( \gamma \), we must have \( \| (I - 2\gamma \zeta_{t,i} \tilde{X}^t_i \tilde{X}^t_{i-1}^T) \| \leq 1 \). We conclude the statement of the lemma from Equation (142).

\[\square\]

**H.6 Proof of Claim** [3]

**Proof.** Let \( r_2 > r_1 \). As in proof of Claim 2 let \( C^t_r \) denote the resampled version of \( C^t_r \) obtained by re-sampling \( \eta_{t-r_1} \) i.e.,

\[
C^t_r(t, r_1, r_2, j_1, j_2) := 4\gamma^2 \epsilon_{t-r_1} \epsilon_{t-r_2} \left( \prod_{s=r_2-1}^{r_1-1} \hat{H}^{t-s}_{0,B-1} \right) \left( \prod_{s=r_2-1}^{r_1-1} \hat{H}^{t-r_1,T}_{j_1+1,B-1} \right) \hat{X}^{t-r_2}_{j_2} \hat{H}^{t-r_2}_{j_2+1,B-1} \left( \prod_{s=r_2-1}^{r_1-1} \hat{H}^{t-s}_{0,B-1} \right).
\]

(143)

Here we have used the fact that \( R^{t-r_1} \) does not affect the buffers up to \( t - r_1 - 1 \) and only \( \tilde{X} \)'s that are affected are in the term \( \hat{H}^{t-r_1}_{0,j_1-1} \). Like in Claim 2 notice that

\[
\mathbb{E} \left[ \sum_{r_2 > r_1} \sum_{j_1, j_2} C^t_r(t, r_1, r_2, j_1, j_2) \right] = 0
\]

48
Applying Lemma 9, we conclude that:

\[
\sum_{r_2 > r_1} \sum_{j_1, j_2} C r_i'(t, r_1, r_2, j_1, j_2) = 2\gamma \sum_{r_1=1}^{t-1} \sum_{j_1=0}^{B-1} (\tilde{a}_{B}^{r_1-1, r_1})^\top \mathcal{R}_{-j_1}^{t-r_1} \left( \tilde{H}_{0, B-1}^t \right) \mathcal{R}_{-j_1}^{t-r_1} \left( \prod_{s=r_1-1}^{1} \tilde{H}_{0, B-1}^s \right) \mathcal{R}_{-j_1}^{t-r_1} \left( \tilde{H}_{j_1+1, B-1}^t \right) \tilde{X}_{j_1}^{t-r_1, r_1, j_1}.
\]

We cannot continue our analysis like in Claim 2 because due to resampling of \(\tilde{\varepsilon}_{j_1}^t\), \(\tilde{H}_{0, B-1}^t\) changes not just because of the iterates \(\tilde{a}_{j_1}^{t-r_1}\) but also due to \(\tilde{X} \rightarrow \tilde{\tilde{X}}\).

Further

\[
\mathbb{E} \left[ \sum_{r_2 > r_1} \sum_{j_1, j_2} C r_i'(t, r_1, r_2, j_1, j_2) \right] = 0 \quad (145)
\]

Next we have simple lemma

**Lemma 22.** Consider for each \((r_1, j_1)\), the re-sampling operator \(\mathcal{R}_{-j_1}^{t-r_1}\)

\[
\mathbb{E} \left[ \sum_{r_2 > r_1} \sum_{j_1, j_2} C r_i(t, r_1, r_2, j_1, j_2) \right] \leq 4\gamma^2 R \frac{(Bt)^2}{2} C \sigma^2 \frac{1}{T^{\alpha/2}} +
\]

\[
\mathbb{E} \left[ \sum_{r_2 > r_1} \sum_{j_1, j_2} C r_i(t, r_1, r_2, j_1, j_2) \right] \mathcal{R}_{-j_1}^{t-r_1} \left( \tilde{D}_{j_1}^{0, t-1} \right) \mathcal{R}_{-j_1}^{t-r_1} \left( \tilde{D}_{j_1}^{t-r_1, C} \right) \leq 4\gamma^2 R \frac{(Bt)^2}{2} C \sigma^2 \frac{1}{T^{\alpha/2}} \quad (146)
\]

**Proof.** We have

\[
1 \left[ \tilde{D}_{j_1}^{0, t-1} \right] = 1 \left[ \tilde{D}_{j_1}^{0, t-1} \right] \mathcal{R}_{-j_1}^{t-r_1} \left[ \tilde{D}_{j_1}^{t-r_1} \right] + 1 \left[ \tilde{D}_{j_1}^{0, t-1} \right] \mathcal{R}_{-j_1}^{t-r_1} \left[ \tilde{D}_{j_1}^{t-r_1, C} \right] \quad (147)
\]

Hence

\[
\mathbb{E} \left[ \sum_{r_2 > r_1} \sum_{j_1, j_2} C r_i(t, r_1, r_2, j_1, j_2) \right] \leq 4\gamma^2 R \frac{(Bt)^2}{2} C \sigma^2 \frac{1}{T^{\alpha/2}} \quad (148)
\]

where we used \(\mathcal{R}_{-j_1}^{t-r_1} \left[ \tilde{D}_{j_1}^{t-r_1, C} \right]\) is identically distributed as \(1 \left[ \tilde{D}_{j_1}^{t-r_1, C} \right]\) and hence

\[
\mathbb{E} \left[ \mathcal{R}_{-j_1}^{t-r_1} \left[ \tilde{D}_{j_1}^{t-r_1, C} \right] \right] \leq \frac{1}{\tau^n}
\]

\[\Box\]

So, based on the above lemma, we focus on bounding

\[
\mathbb{E} \left[ \sum_{r_2 > r_1} \sum_{j_1, j_2} C r_i(t, r_1, r_2, j_1, j_2) \right] \mathcal{R}_{-j_1}^{t-r_1} \left( \tilde{D}_{j_1}^{0, t-1} \right) \mathcal{R}_{-j_1}^{t-r_1} \left( \tilde{D}_{j_1}^{t-r_1} \right) \]

49
Now notice that
\[
\mathbb{E} \left[ \sum_{r_2 > r_1} \sum_{j_1, j_2} Cr'(t, r_1, r_2, j_1, j_2) \right] 1 [\mathcal{D}_{t-r_1}^0, t-r_1-1] 1 [\mathcal{D}_{t-j_1}^{-r_1}, 1] 1 [\mathcal{D}_{t-0}^{-r_1}] = 0
\]

(149)

Hence
\[
\mathbb{E} \left[ \sum_{r_2 > r_1} \sum_{j_1, j_2} Cr'(t, r_1, r_2, j_1, j_2) 1 [\mathcal{D}_{t-r_1}^0, t-r_1-1] 1 [\mathcal{D}_{t-j_1}^{-r_1}, 1] \right] = 0 - \mathbb{E} \left[ \sum_{r_2 > r_1} \sum_{j_1, j_2} Cr'(t, r_1, r_2, j_1, j_2) 1 [\mathcal{D}_{t-r_1}^0, t-r_1-1] 1 [\mathcal{D}_{t-j_1}^{-r_1}, 1] \right]
\]

(150)

Thus
\[
\mathbb{E} \left[ \sum_{r_2 > r_1} \sum_{j_1, j_2} Cr'(t, r_1, r_2, j_1, j_2) 1 [\mathcal{D}_{t-r_1}^0, t-r_1-1] 1 [\mathcal{D}_{t-j_1}^{-r_1}, 1] \right] \leq 2\gamma^2 R \frac{(Bt)^2}{2} C \sigma^2 \frac{1}{T^{\alpha/2}}
\]

(151)

Now, similar to lemma 8 on the event \( \mathcal{E}_{0:r_1} \cap \mathcal{D}_{t-r_1}^0 \cap \mathcal{A}^t \) we have:
\[
\left\| \prod_{s=r_1-1}^{t-1} \tilde{H}_{0,B-1}^{t-s} \right\| - \mathcal{R}^{t-r_1}_{t-j_1} \left( \prod_{s=r_1-1}^{t-1} \tilde{H}_{0,B-1}^{t-s} \right) = C B t \| \phi' \|^2 R^3 B \frac{\sqrt{\beta}}{\zeta \lambda_{\min}}
\]

(152)

Next, similar to lemma 8 for \( R \gamma \leq \frac{1}{2} \), on the event \( \mathcal{D}_{t-r_1}^0 \cap \mathcal{A}_{t-1} \) we have
\[
\left\| \tilde{H}_{0,B-1}^{t-r_1, 1} \mathcal{R}^{t-r_1}_{t-j_1} \left( \tilde{H}_{0,B-1}^{t-r_1, 1} \right) \right\| \leq 4\gamma R B
\]

(153)

Finally we can bound the norm of the expected difference of sums of \( Cr \) and \( Cr' \) using lemma 9 and (144) as
\[
\mathbb{E} \left[ \sum_{r_2 > r_1} \sum_{j_1, j_2} (Cr - Cr') 1 [\mathcal{D}_{t-r_1}^0, t-r_1-1] 1 [\mathcal{D}_{t-j_1}^{-r_1}, 1] 1 [\mathcal{D}_{t-0}^{-r_1}] \right] \\
\leq 2\gamma \mathbb{E} \left[ \sum_{r_2 > r_1} \sum_{j_1, j_2} \sqrt{R} \| e_{r_1-j_1}^{t-r_1} \| 1 [\mathcal{D}_{t-r_1}^0, t-r_1-1] \right] \\
\leq \left( C \| \phi' \|^2 \gamma^2 T R^3 B \frac{\sqrt{\beta}}{\zeta \lambda_{\min}} + C \gamma RB \right)
\]

(154)
Thus

\[
\begin{align*}
E \left[ \sum_{r_2 > r_1} \sum_{j_1, j_2} (C_r - C_r') \left[ \tilde{D}^{0,t-1} \right] R_{r_2 - j_1} \left[ \tilde{D}^{0,r_1} \right] \right] \\
\leq C \left( \| \phi'' \| \gamma^3 T^2 R^3 B \frac{\sqrt{B}}{\zeta \lambda_{\min}} + \gamma^2 T R B \right) \sqrt{C_n \sigma^2} \sqrt{\sup_{s \leq N-1} E \left[ \| \tilde{a}_{B, s} \|_{1} \| \tilde{D}^{0,s} \| \right]} \\
+ C(Bt)^2 \left[ \gamma^2 R C_n \sigma^2 \left( \sqrt{\mathbb{P} \left[ \bigcup_{s=0}^{N-1} \bigcup_{i=0}^{B-1} \xi_{s,i} \right]} + \sqrt{\mathbb{P} \left[ A^{t-1,C} \right]} \right) \right] \\
\leq C \left( \| \phi'' \| \gamma^3 T^2 R^3 B \frac{\sqrt{B}}{\zeta \lambda_{\min}} + \gamma^2 T R B \right) \sqrt{C_n \sigma^2} \sqrt{\sup_{s \leq N-1} E \left[ \| \tilde{a}_{B, s} \|_{1} \| \tilde{D}^{0,s} \| \right]} \\
+ C(Bt)^2 \gamma^2 R C_n \sigma^2 \frac{1}{T^{\alpha/2}}
\end{align*}
\]  

(155)

Combining everything we conclude the claim.

\( \square \)