Non-asymptotic Behavior in Massive Multiple Access and Streaming System Identification

by

Suhas Subramanya Kowshik

B.Tech & M.Tech. (Dual), Indian Institute of Technology Madras (2016)
S.M., Massachusetts Institute of Technology (2018)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2022

Author .....
Department of Electrical Engineering and Computer Science
May 13, 2022

Certified by .....
Yury Polyanskiy
Associate Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by .....
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Non-asymptotic Behavior in Massive Multiple Access and Streaming System Identification

by

Suhas Subramanya Kowshik

Submitted to the Department of Electrical Engineering and Computer Science
on May 13, 2022, in partial fulfillment of the
requirements for the degree of
Doctor of Science

Abstract

Non-asymptotic understanding of information theoretic and algorithmic limits of estimation in statistical problems is indispensable for practical applications in engineering. There are two broad approaches to this end. The first: tools and advances in modern probability theory aid in deriving fully non-asymptotic bounds for such problems. This approach, while useful, is sometimes insufficient as the bounds it yields usually have sub-optimal constants and unavoidable logarithmic factors. Consequently, in some situations where the primary goal is to obtain sharp characterizations, e.g. error exponents, it can be highly non-trivial to derive fully non-asymptotic results that serve this purpose, particularly in high dimensional problems of recent times. In aforesaid circumstances there is a second approach: recourse to asymptotics that can serve as a reasonable substitute for finite length behavior. In this thesis, we employ both these approaches. First we provide high dimensional asymptotic bounds for massive multiple access which is an important consideration in upcoming wireless networks. Then we turn towards streaming system identification where we develop a novel algorithm provide tight non-asymptotic bounds showing the optimality of our method.

Massive multiple access is an important problem in current and upcoming wireless networks. Also known as massive machine type communication (mMTC) in 5G, it envisions a scenario of a large number of transmitters (usually small sensors in IoT for instance) with small payloads communicating sporadically with a base station. Information theoretic understanding of such a problem is of paramount importance for evaluating existing multiple access schemes and developing new strategies that handle such drastic interference. To this end, many-user multiple access channel (MAC) is a crucial model that captures the new effects in massive multiple access. Previous works have focused on the additive white Gaussian noise (AWGN) many-user MAC. In this thesis, we aim to understand the fundamental limits of energy efficiency in the quasi-static Rayleigh fading many-user MAC. In particular, we provide tight achievability and converse bounds on the minimum energy-per-bit required to support a certain user density, fixed payload and target per-user error (in the limit as blocklength grows to infinity). Although asymptotic in nature, the results are expected to serve as a good proxy for true finite length behavior. We confirm the presence of the promising almost perfect multi-user interference cancellation, first observed in the AWGN setting, in the quasi-static case. Further we also provide a new achievability bound for the AWGN many-user MAC.

Next we turn towards problem of streaming or online system identification with the goal of designing optimal algorithms and providing non-asymptotic rates on the convergence. In particular, we consider a class of linear and generalized linear (nonlinear) parametric discrete time dynamical systems. Observing a single trajectory from such a system, the aim is recover the system parameters in a streaming fashion. Our work shows that one-pass forward stochastic gradient descent (SGD) algorithm where samples are read in order is sub-optimal compared to the offline ordinary least squares (OLS) estimator. More importantly, based on the observation that reading samples in reverse order mitigates the effect of temporal dependencies, we develop a novel algorithm called SGD with reverse experience replay (SGD-RER) and derive fully non-asymptotic bounds that show

it to be near minimax optimal for both stable linear and generalized linear models. Furthermore, we consider a Quasi-Newton style offline algorithm for the generalized linear setting and show that is near optimal even when the process is unstable.

Thesis Supervisor: Yury Polyanskiy
Title: Associate Professor of Electrical Engineering and Computer Science

Acknowledgments

This thesis marks the culmination of a wonderful journey that I began when I started my graduate school at MIT. It is said the life is nothing but a voyage of self discovery. This has never been more apparent to me than my journey as a graduate student. Perhaps this deeper understanding of myself is my biggest takeaway from my time at MIT. So I am thankful to each and everyone I was fortunate to meet in this journey, for being a mirror to see within myself. Nonetheless, I came to MIT to learn, become a better researcher, and get myself a doctorate. So this acknowledgement is a humble effort to thank some of those who made it happen.

First and foremost, I am eternally grateful to my parents for just being there, as they always have. None of anything I have pursued would have been possible without their consistent support. I am also grateful to my sister, Sahana, for being there in my life journey.

I am fortunate to have been advised by Prof. Yury Polyanskiy throughout my time here at MIT. I still fondly remember the first email I received from him after getting an admit to MIT. It is not possible to overstate how much I have learnt from interacting with him. I still remember the first couple of years, where I had a hard time progressing on my research. Yury stood by me patiently, providing valuable advice and feedback which resulted in publishing my first ever research paper. This also gave me a platform to stand on, as a researcher. People who know or have interacted with Yury are always amazed by his breadth, depth of knowledge, the ability to link many different areas, and his sheer energy and enthusiasm. I consider myself fortunate to have witnessed these for all these years in literally every meeting I have had with him. I am really grateful to him for taking me as his student and advising me all along.

I am also thankful to my committee members Prof. Gregory Wornell and Prof. Devavrat Shah. I met Prof. Greg for the first time when I came to MIT for the visit days. Later, I was fortunate to be a student in his course 6.437 on inference which I thoroughly enjoyed. Prof. Devavrat has been my academic advisor throughout my time here at MIT. I have enjoyed our academic advising sessions (accompanied with delicious pizza). I want to thank both Prof. Greg and Prof. Devavrat for being part of my thesis committee and providing valuable feedback on my work.

Next, I am grateful and fortunate to have worked with wonderful collaborators Prateek Jain, Praneeth Netrapalli and Dheeraj Nagaraj. Praneeth and Prateek mentored me as an intern at Microsoft Research (India), and I had an amazing and productive time there. I started working with them on online learning of dynamical systems, which was a new area for me. They have consistently guided me in my research on this front, and I have learnt a great deal from my interactions with them. I am very thankful for their mentorship. I am also grateful to Dheeraj Nagaraj, who was my flatmate at MIT and also my dear friend. Dheeraj joined this project on learning dynamical systems and helped me push it far ahead. I have always had such illuminating interactions with him, not just during our collaboration, but going back to my days in undergrad. I am very thankful to him

for being a wonderful collaborator.

I had an enriching experience in my extracurricular life here at MIT, thanks to Sangam (Indian graduate students association) and MIT Hindu Students Council (HSC). I was fortunate to serve as president and treasurer on the boards of Sangam and HSC, respectively. These gave me a chance to be in touch with my cultural and traditional roots. Further, they also helped me meet so many people that I would not have met otherwise.

Next I would like to thank my previous and current flatmates who are also my friends: Dheeraj, Prashanth, Kunal and Arpit. They have helped me make my dorm a home away from home.

I am very grateful to the amazing set of friends, wonderful peers and colleagues that I have had at MIT. As much as I wish to thank each one, it would take a couple of pages and I am sure that I would miss some of them.

I would like to express my gratitude to my extended family, grandparents and my amazing group of cousins, who have enriched my life all along.

Lastly, I am ever indebted to Devi: in Her infinite grace, among Her infinite creation, She has held me, a nobody, closely in her arms, always. May She ever reside in my heart.

Funding Acknowledgements

This work was supported in part by the following grants:

- NSF CCF-1253205
- NSF CCF-1717842
- MIT-Skoltech Initiative
- Skolkovo Inst of Sci & Tech

Part of the work on linear system identification was done when I was an intern at Microsoft Research Lab India Pvt. Ltd. during the summer of 2020.

सदाशिव समारम्भां शङ्कराचार्य मध्यमाम् |
अस्मदाचार्य पर्यन्तां वन्दे गुरु परम्पराम् | |

ॐ ग॒णानाँ॑ त्वा ग॒णप॑तिग्ं॒ हवामहे क॒विं क॑वी॒नामु॑प॒मश्र॑वस्तमम् |
ज्ये॒ष्ठरा॑ज॒ं ब्रह्म॑णां ब्रह्मणस्पत॒ आ नः॑ शृ॒ण्वन्नू॒तिभि॑स्सीद॒ सादं॑नम् | |

शुक्लां ब्रह्मविचारसारपरमामाद्यां जगद्व्यापिनीं
वीणापुस्तकधारिणीमभयदां जाड्यान्धकारापहाम् |
हस्ते स्फाटिकमालिकां च दधतीं पद्मासने संस्थितां
वन्दे तां परमेश्वरीं भगवतीं बुद्धिप्रदां शारदाम् | |

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Tractable statistical models play an important role in modeling real world engineering problems in communication, control, learning etc. Theoretical understanding of non-asymptotic information theoretic and algorithmic limits of relevant tasks on such models is of considerable importance in development and deployment of practical solutions. As examples , consider the task of parameter estimation based on observations from a model, or decoding messages using signal received over a communication channel. Deriving tight sample complexity bounds on estimation error in the former and finite blocklength bounds on probability of error in the latter is a challenging and fairly non-trivial exercise. Further complications arise due to the high dimensional nature of modern problems. Fully non-asymptotic bounds on algorithms usually depict the dependence on all the problem parameters but with sub-optimal constants (like in the exponential rate) and logarithmic factors, and such bounds are nonetheless useful in validating those algorithms. But in situations where one desires very precise bounds, like the error exponent, it is hard to obtain such non-asymptotic results especially in a high-dimensional setting. In such a scenario, asymptotics that can serve as an accurate proxy for finite length behavior is invaluable. In this thesis we undertake both the above paths of obtaining bounds that explain non-asymptotic behaviour in two different topics of massive multiple access and streaming system identification.

1.1 Massive multiple access : Many-user MAC

First we consider the problem of massive multiple access. The upcoming generation of wireless networks (such as 5G and beyond) face a unique challenge of supporting massive connectivity with diverse quality of service (QoS) requirements. One of them is known as massive machine type communication (mMTC): a large number of sporadically transmitting terminals are serviced by a single base-station (BS). The characteristics of such a system are high node density, small payloads, stringent energy constraints, sporadic transmission and constrained computational capabilities (at

transmission) [1]. A typical example is a large scale internet-of-things. For instance, a smart city scenario would envision a massive number of battery powered sensors connected to a BS. The increasing dense deployment of miniaturized radio-equipped sensors result in progressively worsening interference environment and stringent demands on communication energy efficiency. This suggests a bleak picture for the future networks, where a chaos of packet collisions and interference contamination prevents reliable connectivity.

Classical information theory of multiple access channels (MAC) [2] is not adequate to understand this new paradigm. This is due to the fact that short packets lead to finite blocklength effects and the number of *active* users users can be comparable to the blocklength. Furthermore, it is often difficult to compare network theoretic solutions (like ALOHA) to this problem with information theoretic results, mainly due to the models being different in each of these.

A concrete step towards addressing the above issues and providing a firm information theoretic footing to the question of massive multiple access was carried out in [3, 4]. In particular, [3] provided the notion of a random access code for a permutation-invariant MAC with a twist that decoding is done only up to permutation and the error metric is the fraction of decoded messages that are erroneous – called the per-user probability of error (PUPE). This model is now known as unsourced MAC in the literature [5] referring to the property that user identities are immaterial. Furthermore, [3] also gave finite blocklength (FBL) achievability bounds in the case of unsourced Gaussian MAC: minimum energy-per-bit (E_b/N_0) required when K_a active users want to transmit k bits each to a base-station over n degrees of freedom (or blocklength) such that PUPE is at most ϵ. [3, 4] also provided reasonable values for K_a, n and k that are relevant for mMTC: few tens to hundred bits of payload, $\sim 10^4$ real degrees of freedom n and $\sim 10^2$ active users. These numbers are now a standard in unsourced MAC literature. The proof of the main achievability in [3] uses random coding with an ML decoder which is computationally infeasible to implement. Evaluation of this bound revealed an interesting phenomenon: the required E_b/N_0 to achieve a target PUPE increases only negligibly for small values of K_a and this ballpark value of E_b/N_0 is dictated by the minimum energy required to communicate fixed payload for a single user channel [6] rather than the multi-user interference (MUI) effects of MAC. Comparing this result with ALOHA, TDMA/FDMA and TIN showed that the latter are sub-optimal and are susceptible to severe MUI even at small K_a. This has led to a series of works [4, 5, 7–12] aimed at designing practical low-complexity schemes that achieve the theoretical bound.

In order to rigorously explain the observed FBL behavior in unsourced MAC, [3] also considered an asymptotic linear scaling regime of the Gaussian (non-random access) MAC which we describe next. Consider a problem of K nodes communicating over a frame-synchronized multiple-access channel. When K is fixed and the blocklength n is taken to infinity we get the classical regime [2], in which the fundamental limits are given by well-known mutual information expressions. A new

regime, deemed *many-access*, was put forward by Chen, Chen and Guo [13]. In this regime the number of nodes K grows with blocklength n. It is clear that the most natural scaling is linear: $K = \mu n, n \to \infty$, corresponding to the fact that in time n there are linearly many users that will have updates/traffic to send [3]. That is, if each device wakes up once in every T seconds and transmits over a frame of length t, then in time (proportional to) t there are $K \approx t/T$ users where t is large enough for this approximation to hold but small that no device wakes up twice. Further, asymptotic results obtained from this linear scaling have been shown to approximately predict behavior of the fundamental limit at finite blocklength, e.g. at $n = 30000$ and $K <= 300$ [3, 14]. The analysis of [13] focused on the regime of infinitely large payloads (see also [15] for a related massive MIMO MAC analysis in this setting) along with the classical joint probability of error. In contrast [3] proposed to focus on a model where each of the $K = \mu n$ nodes has only finitely many bits to send in conjunction with the per-user probability of error (PUPE). In this regime, it turned out, one gets the relevant engineering trade-offs. Namely, the communication with finite energy-per-bit is possible as $n \to \infty$ and the optimal energy-per-bit depends on the user density μ.

These two modifications (the scaling $K = \mu n$ and the PUPE) were investigated in the case of the AWGN channel in [3, 14, 16]. We next describe the main discovery of that work. The channel model is[1]:

$$Y^n = \sum_{i=1}^{K} X_i + Z^n\,, \qquad Z^n \sim \mathcal{CN}(0, I_n)\,, \tag{1.1}$$

and $X_i = f_i(W_i) \in \mathbb{C}^n$ is the codeword of i-th user corresponding to $W_i \in [2^k]$ chosen uniformly at random. The system is said to have PUPE ϵ if there exist decoders $\hat{W}_i = \hat{W}_i(Y^n)$ such that

$$P_{e,u} = \frac{1}{K}\sum_{i=1}^{K} \mathbb{P}\left[W_i \neq \hat{W}_i\right] \leq \epsilon\,. \tag{1.2}$$

The energy-per-bit is defined as

$$\frac{E_b}{N_0} = \frac{1}{k} \sup_{i\in[K], w\in[2^k]} \|f_i(w)\|^2.$$

The goal in [3, 16] was to characterize the asymptotic limit

$$\mathcal{E}^*(\mu, k, \epsilon) \triangleq \limsup_{n\to\infty} \inf \frac{E_b}{N_0} \tag{1.3}$$

where infimum is taken over all possible encoders $\{f_i\}$ and decoders $\{\hat{W}_i\}$ achieving the PUPE ϵ for $K = \mu n$ users. (Note that this problem may be recast in the language of compressed sensing and sparse regression codes (SPARCs) – see Section 2.1.1 below.)

To predict how $\mathcal{E}^*(\mu, \epsilon)$ behaves, first consider a naive Shannon-theoretic calculation [17]: if K

[1] Although real AWGN channel is considered in [3], we state the results for the complex AWGN case.

users want to send k bits in n degrees of freedom, then their sum-power P_{tot} should satisfy

$$n \log(1 + P_{tot}) = kK \,.$$

In turn, the sum-power $P_{tot} = \frac{kK}{n} \frac{E_b}{N_0}$. Overall, we get

$$\mathcal{E}^* \approx \frac{2^{\mu k} - 1}{k\mu} \,.$$

This turns out to be a correct prediction, but only in the large-μ regime. The true behavior of the fundamental limit is roughly given by

$$\mathcal{E}^*(\mu, k, \epsilon) \approx \max\left(\frac{2^{\mu k} - 1}{k\mu}, \mathcal{E}_{\text{s.u.}}\right) , \tag{1.4}$$

where $\mathcal{E}_{\text{s.u.}} = \mathcal{E}_{\text{s.u.}}(k, \epsilon)$ does not depend on μ and corresponds to the single-user minimal energy-per-bit for sending k bits with error ϵ, for which a very tight characterization is given in [18]. In particular, with good precision for $k \geq 10$ we have

$$\mathcal{E}_{\text{s.u.}}(k, \epsilon) = \frac{1}{k} \left(\mathcal{Q}^{-1}\left(2^{-k}\right) - \mathcal{Q}^{-1}\left(1 - \epsilon\right)\right)^2 \tag{1.5}$$

where $\mathcal{Q}$ is the complementary CDF of the standard normal distribution: $\mathcal{Q}(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-\frac{u^2}{2}} du$.

In all, results of [3, 14, 16] suggest that the minimal energy-per-bit has a certain "inertia": as the user density μ starts to climb from zero up, initially the energy-per-bit should stay the same as in the single-user $\mu = 0$ limit. In other words, optimal multiple-access architectures should be able to *almost perfectly cancel all multi-user interference (MUI)*, achieving an essentially single-user performance for each user, *provided the user density is below a critical threshold*. Note that this is much better than orthogonalization, which achieves the same effect at the expense of shortening the available (to each user) blocklength by a factor of $\frac{1}{K}$. Quite surprisingly, standard approaches to multiple-access such as TDMA and TIN[2], while having an optimal performance at $\mu \to 0$ demonstrated a significant suboptimality for $\mu > 0$ regime. In particular, no "inertia" was observed and the energy-per-bit for those suboptimal architectures is always a monotonically increasing function of the user density μ. This opens the (so far open) quest for finding a future-proof MAC architecture that would achieve $\mathcal{E}_{\text{s.u.}}$ energy-per-bit for a strictly-positive $\mu > 0$. A thorough discussion of this curious behavior and its connections to replica-method predicted phase transitions is contained in Section 4.3.

[2]Note that pseudo-random CDMA systems without multi-user detection and large load factor provide an efficient implementation of TIN. So throughout our discussions, conclusions about TIN also pertain to CDMA systems of this kind.

Contributions A contribution of this thesis is in demonstrating the same almost perfect MUI cancellation effect in a much more practically relevant communication model, in which the ideal unit power-gains of (1.1) are replaced by random (but static) fading gain coefficients. We consider two cases of the channel state information: known at the receiver (CSIR) and no channel state information (noCSI).

Key technical ideas: For handling the noCSI case we employ the subspace projection decoder similar to the one proposed in [19], which can be seen as a version of the maximum-likelihood decoding (without prior on fading coefficients) – an idea often used in support recovery literature [20–22]. Another key idea is to decode only a subset of users corresponding to the strongest channel gains – a principle originating from Shamai-Bettesh [23]. While the randomness of channel gains increases the energy-per-bit requirements, in a related paper we find [24] an unexpected advantage: the inherent randomization helps the decoder disambiguate different users and improves performance of the belief propagation decoder. Our second achievability bound improves projection decoder regime by applying the Approximate Message Passing (AMP) algorithm [25] along with spatially coupled codebook design [26, 27]. The rigorous analysis of its performance is made possible by results in [22, 26–28]. On the converse side, we leverage the recent finite blocklength results for the noCSI channel from [6, 19].

We note that although the bounds are asymptotic, the high dimensional limit considered here can serve as a reasonable benchmark for the FBL behavior (for the parameter values relevant to unsourced MAC).

Organization Chapter 2 deals with quasi-static fading many-user MAC with section 2.1 containing the system model, section 2.2 on the classical regime of fixed K and $n \to \infty$, section 2.3 containing the main results and proofs of the many-user MAC. Chapter 3 deals with the AWGN many-user MAC with system model in 3.1 and main results in 3.2. The experimental results are provided in chapter 4.

1.2 System identification

Learning with Markovian data is important for problems in time series analysis, control and reinforcement learning (RL). Furthermore, methods to learn in a streaming fashion or on-the-go are necessary for many modern problems in these fields. In this thesis, we consider the problem of streaming system identification in linear and certain nonlinear time variant dynamical systems. In particular, we consider observing a single trajectory $(X_0, \cdots X_T)$ of vectors in $\mathbb{R}^d$ from the dynamical system:

$$X_{t+1} = \phi(A^* X_t) + \eta_t, \tag{1.6}$$

where $A^* \in \mathbb{R}^{d\times d}$ is the unknown system matrix, ϕ is a known component-wise (possibly nonlinear) function and η_t are i.i.d noise vectors. The goal is to estimate the system matrix A^*. If $\phi(x) = x$, the the above system is a linear time variant dynamical system with full state observation. Next we discuss each of these models in detail and place our work in the vast literature on this subject.

1.2.1 Linear system identification

First, we study the problem of learning linear-time invariant (LTI) systems, where the goal is to estimate the matrix $A^* \in \mathbb{R}^{d\times d}$ from the given samples $(X_0, \ldots, X_T)$ that obey:

$$X_{\tau+1} = A^* X_\tau + \eta_\tau, \quad X_\tau \in \mathbb{R}^d, \quad \eta_\tau \overset{i.i.d.}{\sim} \mu, \tag{1.7}$$

where μ is an unbiased noise distribution. The problem is central in control theory and reinforcement learning (RL) literature [29, 30]. It is also equivalent to estimating Vector Autoregressive (VAR) model popular in the time-series analysis literature [31], where it has been used in several applications like finding gene regulatory information network [32].

A natural estimator for the system parameters is the ordinary least squares (OLS) estimator

$$\hat{A}_{OLS} = \arg\min_A \sum_{t=0}^{T-1} \|AX_t - X_{t+1}\|^2 \tag{1.8}$$

This problem has decades of rich history in many disciplines like control theory and econometrics with the earliest works on the asymptotic consistency and limiting distribution of the OLS estimator going back to at least 1940's (see [33–38] and references therein). Furthermore, online estimation using recursive methods like stochastic approximation [39] or stochastic gradient descent (SGD) style methods have also been considered and analyzed for this model (see [40–44] and references therein). But all of these classical results are asymptotic in nature i.e., they prove strong consistency and asymptotic normality of the online estimators but do not talk about when the asymptotic effects kick in, in terms of the sample size.

Early non-asymptotic results focused on time series forecasting i.e., obtaining generalization bounds instead of parameter estimation, for general stationary mixing time series based on uniform convergence type arguments [45, 46]. Further extensions have been made in this direction in [47–49]. Since these bounds are developed in more generality, they are either not sharp in the problem parameters like mixing time and dimension or involve boundedness assumptions on the loss functions which does not apply to the case considered in this thesis.

Relegating more extensive literature discussion to a later section, we now present the known finite time results about the OLS estimator that helps place our bounds in context. These non-asymptotic near minimax optimal bounds for the OLS estimator of (5.1) i.e., on the operator norm

$\|A^* - \hat{A}_{OLS}\|_{\text{op}}$ are given in [50–53]. The main takeaways from these works are as follows. Let $\mathbb{E}\left[\eta_t \eta_t^\top\right] = \sigma^2 I$ and let $\Gamma = \sum_{k=0}^{\infty} (A^*)^k (A^{*,\top})^k$ be the controllability grammian.

- When the system is stable i.e., spectral radius of A^* is strictly smaller than 1, then the recovery guarantees on $\hat{A}_{OLS}$ are quantitatively similar to what one expects when the data are i.i.d i.e., in a standard linear regression. In particular,

$$\|\hat{A}_{OLS} - A^*\|_{op} \lesssim \sqrt{\frac{d + \log(1/\delta)}{T\lambda_{\min}(\Gamma)}}$$

  with probability at least $1 - \delta$.

- When all eigenvalues of A^* are in $(1 - 1/T, 1 + 1/T)$, faster rates are possible: $\|\hat{A}_{OLS} - A^*\| \leq O(d/T)$.

- When the matrix A^* is *regular* and all eigenvalues are larger than 1, the OLS estimator has exponential rates of convergence.

Notice that the above bounds are independent of the mixing time of the process. In fact, estimation becomes easier as the system becomes less and less stable. Furthermore, they are minimax optimal as shown in [51, 54, 55].

The OLS estimator is an offline method that requires access to all the data. Of course, one can implement OLS in an online fashion using Sherman-Morrison formula. But such a solution is limited and does not apply to practically important settings like *generalized non-linear dynamical* system or when A^* is high-dimensional and has special structure like low-rank or sparsity [56, 57]. Hence with the goal of potential applicability to certain non-linear dynamical systems, the question remains as to how do stochastic approximation or stochastic gradient descent (SGD) style *one pass streaming* algorithms perform for the online estimation of A^*. This is important for applications like RL, large-scale forecasting systems, recommendation systems [58, 59].

In [60], the authors consider a projected SGD style algorithm for estimating parameters of a more general partially observed linear state space model using multiple independent trajectories. The obtained bounds have sub-optimal dependence on dimension and mixing time. [61, 62] consider online prediction or output tracking of adversarial linear dynamical systems and prove $O(\sqrt{T})$ regret for a trajectory of length T. This is different from the focus of this thesis, which is specifically on system identification in a non-adversarial situation.

Contributions In this thesis, we study the above mentioned problem of learning LTI systems via first order gradient oracle with streaming data. The goal is to design an estimator that provides accurate estimation while ensuring nearly optimal time complexity and space complexity that is nearly *independent* of T. In particular, we focus on designing Stochastic Gradient Descent (SGD)

style methods that can work directly with first order gradient oracle, and hence is more widely applicable to the settings mentioned above. In fact the algorithm ($\mathsf{SGD} - \mathsf{RER}$) and the techniques introduced for the analysis here are to obtain near-optimal guarantees for learning certain classes of *non-linear dynamical systems* [63] described in the next section. Furthermore, this algorthims has been used to get near-optimal guarantees for Q-learning tabular MDPs in RL [64].

SGD is a popular method for general streaming settings, and has been shown to be *optimal* for problems like streaming linear regression [65]. However, when the data has temporal dependencies, as in the estimation of linear dynamical systems, such a naive implementation of SGD may not perform well as observed in [66, 67]. In fact, for linear system identification, our experiments suggest that SGD suffers from a non-zero bias (Section 7.1). In order to address temporal dependencies in data, practitioners use a heuristic called *experience replay*, which maintains a *buffer* of points, and samples points *randomly* from the buffer. However, for linear system identification, experience replay does not seem to provide an accurate unbiased estimator for reasonable buffer sizes (see Section 7.1).

In this work, we propose *reverse experience replay* for linear system identification. Our method maintains a small *buffer* of points, but instead of random ordering, we replay the points in a *reverse* order. We show that this algorithm exactly unravels the temporal correlations to obtain a consistent estimator for A^*. Similar to the standard linear regression problem with *i.i.d.* samples, we can break the error in two parts: a) bias: that depends on the initial error $\|A_0 - A^*\|$, b) variance: the steady state error due to noise η. We show that our proposed method, under fairly standard assumptions and with a small buffer size, is able to decrease the bias at fast rate, while the variance error is nearly optimal (see Theorem 5.3.1), matching the information theoretic lower bounds [51, Theorem 2.3]. To the best of our knowledge, we provide first non-trivial analysis for a purely streaming SGD-style algorithm with optimal computation complexity and nearly bounded space complexity that is dependent logarithmically on T. We note here that the idea of reverse experience replay was independently discovered in experimental reinforcement learning by [68] based on reverse replay observed in Hippocampal place cells [69] in Neurobiology. We also refer to [70] for more on this connection.

In addition to the transition matrix estimation error $\|A - A^*\|$, we also provide analysis of prediction error, i.e., $E[\|AX - A^*X\|^2]$ (see Theorem 5.3.2). Here again, we bound the *bias* and the *variance* part of the error separately. We further derive new lower bounds for prediction error (see Theorem 5.3.4) and show that our algorithm is minimax optimal, under standard assumptions on the model. As mentioned earlier, our method work with general first order oracles, hence applies to more general problems like *sparse LTI estimation* with known sparsity structure and unlike online OLS methods, $\mathsf{SGD} - \mathsf{RER}$ has nearly optimal time complexity. Finally, we also provide empirical validation of our method on simulated data, and demonstrate that the proposed method is indeed able to provide error rate similar to the OLS method while methods like SGD and standard

experience replay, lead to biased estimates.

Related Work. Due to applications in RL, recently LTI system identification has been widely studied. In particular, [71] studied the problem in offline setting under the "stability" condition, i.e., the spectral radius ($\rho(A^*)$) of A^* is a constant bounded away from 1. The sequence of papers [50–53] provide optimal analyses of the offline OLS estimator beyond assumptions of stability. That is, they show that OLS recovers A^* near optimally even the process defined by (1.7) is stable but does not mix within time T (when $\rho(A^*)$ is $1 - O(1/T)$) or is unstable (when $\rho(A^*)$ is larger than 1). Further [51, 54] provide information theoretic lower bounds for the LTI system identification problem. [63, 72, 73] consider the problem of identifying non-linear dynamical systems of the form $X_{t+1} = \phi(A^* X_t) + \eta_t$ where ϕ is a one dimensional link function which acts co-ordinate wise. In this setting, however, there is no closed for expressions for the estimator of A^*. [72, 73] give offline algorithms whose error guarantees are worse off by factors of mixing time whereas [63] obtains near optimal offline and streaming algorithms for this setting. In fact, [63] uses $\mathsf{SGD} - \mathsf{RER}$ which was first introduced in this work in order to obtain the streaming algorithm.

LTI identification problem has been studied in time series forecasting literature as well. For example, [35] obtains asymptotic consistency results for system identification problem and [45, 46] consider the problem of finite time recovery. Both consider a certain parameterized predictor for a linear system with empirical risk minimization for the parameter and analyzes the deviation from population risk. Similarly, [49] also studies generalization error guarantees. In contrast, our work is able to provide precise bias and variance (similar to generalization error) of the estimator in the streaming setting, and show that the asymptotic error is minimax optimal.

[74] studied SISO systems with observations $(x_\tau, y_\tau) \in \mathbb{R}^2$ and a hidden state h_τ which is high dimensional, thus their model and applications are significantly different than the LTI system we study. For the SISO system, [74] analyzes SGD to provide error bounds contain (a large) polynomial in the hidden state dimension. Here, the hidden state has an evolution similar to Equation 1.7 whereas $x_1, \ldots, x_T$ are drawn i.i.d from some distribution.

System identification has been studied in the context of partially observed LTI systems as well. Recent works [71, 75–79] focus on identifying a certain Hankel-like matrix of the system. These are not directly comparable to the fully observed setting in this work since the model parameters are identifiable only upto a similarity transformation in the partially observed setting.

Recently, there has been an exciting line of work in the related domain of online control (see [80–83] and references therein). The state equation studied in these papers also contain an additive term of Bu_τ for some unknown matrix B and a control signal u_τ and the noise η_τ is either stochastic (as in [80]) or adversarial (as in [81–83]). The goal is to output control signals u_τ after observing $X_1, \ldots, X_\tau$, such that the cost $\sum_\tau c_\tau(X_\tau, u_\tau)$ is minimized for some sequence of convex costs c_τ. We

focus on the LTI system identification(or estimation) problem while the goal of the above mentioned line of work is to design an online controller.

We also note here another line of works [61, 77, 84–88] focused on online prediction of both fully observed and partially observed LTI systems, and the similar problem of time series forecasting by regret minimization [49, 89]. In particular, the main goal there is to design online prediction algorithms minimizing regret against a certain class (for instance, against a Kalman filter with knowledge of the system parameters in the case of partially observed LTI systems). The situation considered in our work is different in atleast two aspects: 1) we focus significantly on parameter recovery or system identification and 2) our notion of prediction is *prediction at stationarity* which can be thought of as one-step regret (compared to T–step regret for instance in [84, 85]).

Next, [56] considers *offline* sparse linear regression with ℓ_1 penalty where the feature vector is derived from an auto regressive model. Similarly, [66] considers the problem of linear regression where the feature vectors come from a Markov chain. This line of work is different from ours in that we try to estimate the parameters of the Markov process itself.

Finally we note here some recent works on estimation or forecasting in VAR models with structured system matrices [57, 90–94] and robust estimation [95, 96].

Organization We provide the problem definition and introduce the notations in section 5.1. We then present our algorithm and the key intuition behind it in Section 5.2. We then present our main result in Section 5.3 and provide a proof sketch in Section 5.4. The proofs are in sections 5.5-5.11. Finally, we present simulation results in Section 7.1.

1.2.2 Generalized linear system

Now we consider the generalized linear dynamical system where the data points $(X_0, X_1, \ldots, X_T)$ evolve as:

$$X_{t+1} = \phi(A^* X_t) + \eta_t, \tag{1.9}$$

where $\eta_t \in \mathbb{R}^d$ are i.i.d. noise vectors, and $\phi : \mathbb{R} \to \mathbb{R}$ is a known increasing nonlinear function, called the 'link function', that acts component wise. The goal is to estimate A^* from observing a single trajectory. This model is a simplified version of recurrent neural networks based models used in nonlinear system identification [97–99], in dynamic behavioral modeling of RF power amplifiers [100] and for long term time series predictions [101].

The general nonlinear system identification problem is extensively studied in control theory [42, 45, 46, 98, 102, 103] as well as time-series analysis [104]. As with the case of linear system identification, early non-asymptotic results [45, 46] as well as modern extensions to general time series [47–49] are suboptimal in various problem parameters.

As in the case of linear systems, the problem is challenging due to temporal dependencies of the data, but also compounded by the presence of nonlinearity. If the mixing time τ_{mix} of the process is finite ($\tau_{\mathsf{mix}} < \infty$), then we can make the data approximately i.i.d. by considering only the points separated by $\tilde{O}(\tau_{\mathsf{mix}})$ time. While this allows using standard techniques for i.i.d. data, it reduces the effective number of samples to $O(\frac{T}{\tau_{\mathsf{mix}}})$, which typically gives an error of the order $O(\frac{\tau_{\mathsf{mix}}}{T})$. In fact, even the state-of-the-art results (prior to our work [105]) have error bounds which are sub-optimal by a factor of τ_{mix}.

Interestingly, as seen before for the special case of linear systems, i.e., when $\phi(x) = x$, the results are significantly stronger i.e., the matrix A^* can be estimated with an error $O(1/T)$ even when the mixing time $\tau_{\mathsf{mix}} > T$. But these results rely on the fact that for linear systems, the estimation problem reduces to an ordinary least squares (OLS) problem for which a closed form expression is available and can be analyzed effectively.

On the other hand, NLDS do not admit such closed form expression. In fact the existing techniques mostly rely on mixing time arguments to induce i.i.d. like behavior in a subset of the points which leads to sub-optimal rates by τ_{mix} factor. Similarly, a direct application of uniform convergence results [106] to show that the minimizer of the empirical risk is close to the population minimizer still gives sub-optimal rates as off-the-shelf concentration inequalities (cf. [107]) incur an additional factor of mixing time. Finally, existing results are mostly focused on offline setting, and do not apply to the case where the data points are streaming which is critical in several practical problems like reinforcement learning (RL) and control theory.

In this thesis, we provide algorithms and their corresponding error rates for the NLDS system identification problem in both offline and online setting, assuming the link function to be expansive (Assumption 4). The main highlight of our results is that the error rates are *independent* of the mixing time τ_{mix}, up to leading order, which to the best of our knowledge is first such result for any non-linear system identification in any setting. In fact, for offline setting, our analysis holds even for systems which do not mix within time T and even for marginally stable systems which do not mix at all. Furthermore, in the streaming setting, we adapt and analyze SGD-Reverse Experience Replay (SGD-RER) that we developed for the case of linear system to NLDS identification and show that error rate that is independent of τ_{mix} in the leading order while still ensuring small space and time complexity. We argue that expansivity is necessary for learning with polynomial (in d) samples due to a lower bound on ReLU (a non-expansive function) from [105].

Instead of mixing time arguments, our proofs for offline learning of NLDS use a natural exponential martingale of the kind considered in the analysis of self normalized process ([108, 109]). For streaming setting, while we do use mixing time arguments (proof of Theorem 6.3.1), we combine them with a delicate stability analysis of the specific algorithm and the machinery developed for the linear case [110] to obtain strong error bounds. See Section 6.5 for a description of these techniques.

Our Contributions. Key contributions of the paper are summarized below:

1. Assuming expansive and monotonic link function ϕ and sub-Gaussian noise, we show that the offline Quasi Newton Method (Algorithm 2) estimates the parameter A^* with near optimal errors of the order $O(1/T)$, even when the dynamics does not mix within time T.

2. We give a one-pass, streaming algorithm inspired by $\mathsf{SGD} - \mathsf{RER}$ method by [110], and show that it achieves near-optimal error rates under the assumption of sub-Gaussian noise, NLDS stability (see section 6.1.1 for the definition), uniform expansivity and second differentiability of the link function.

Related Works. NLDS has been studied in a variety of domains like time-series and recurrent neural networks (RNN). [104] studies specific NLDS models from a time series perspective and establishes non-asymptotic convergence bounds for natural estimators; their error rates suffer from mixing time factor τ_{mix}. [98] considers asymptotic learning of NLDS via neural networks trained using SGD, whereas [111] shows that overparametrized LSTMs trained with SGD learn to memorize the given data. [112–114] consider learning dynamical systems of the form $h_{t+1} = \phi(A^*h_t + B^*u_t)$ for states h_t and inputs u_t; this setting is different from standard NLDS model we study. [115] considers the non linear dynamical systems of the form $x_{t+1} = A\phi(x_t, u_t) + \eta_t$ which ϕ is a known non-linearity and matrix A is to be estimated. [116, 117] consider essentially linear dynamics but allow for certain non-linearities that can be modeled as process noise. All these again differ from the model we consider.

Standard NLDS identification (1.9) has received a lot of attention recently, with results by [72, 73] being the most relevant. [72] uses uniform convergence results via. mixing time arguments to obtain parameter estimation error for offline SGD. [73] obtains similar bounds via. the analysis of the GLMtron algorithm [118]. However, both these works suffer from sub-optimal dependence on the mixing time. We refer to Table 1.1 for a comparison of the results.

When ϕ is not uniformly expansive, [73] obtains within sample prediction error, along with parameter recovery bounds when ϕ is the ReLU function and the driving noise is Gaussian. However, the parameter estimation bounds for ReLU suffer from an exponential dependence on the dimension d and mixing time τ_{mix}. In Theorem 6.4.1 we establish that indeed we cannot improve the exponential dependence in the dimension d for the case of parameter estimation. We note that the exponential dependence arises due to the dynamics present in the system since ReLU regression with isotropic i.i.d. data in well specified case has only a polynomial dependence in d [119].

In linear system identification considered in the previous section as well as in literature [51, 120–122], parameter recovery is made by optimizing the (convex) empirical square loss which also has a closed form solution. However, the square loss in the non-linear case is non-convex. Under the assumption that the link function is increasing, we consider a convex proxy loss which is widely used

Paper	Guarantee	Link Function	System	Noise	Algorithm
[72] THEOREM 6.2	$\frac{d^2\tau_{\mathsf{mix}}}{T}$	INCREASING,LIPSCHITZ EXPANSIVE	MIXING	SUB-GAUSSIAN	OFFLINE
[73] THEOREM 2	$\frac{d^2\tau_{\mathsf{mix}}}{T}$	INCREASING,LIPSCHITZ EXPANSIVE	MIXING	SUB-GAUSSIAN	OFFLINE
This thesis THEOREM 6.2.1	$\frac{d^2\sigma^2}{T\lambda_{\min}(\hat{G})}$	INCREASING,LIPSCHITZ EXPANSIVE	NON-MIXING	SUB-GAUSSIAN	OFFLINE
This thesis THEOREM 6.3.1	$\frac{d^2\sigma^2}{T\lambda_{\min}(G)}$	INCREASING,LIPSCHITZ EXPANSIVE BOUNDED SECOND DERIVATIVE	MIXING	SUB-GAUSSIAN	STREAMING

Table 1.1: Comparison of our results with existing results in terms of mixing time τ_{mix}, stability and number of samples T. Here, we take $\tau_{\mathsf{mix}} = \tilde{\Omega}(\frac{1}{1-\|A^*\|_{op}})$ as a proxy for the mixing time. Note that $\lambda_{\min}(G) \geq \sigma^2$ in the worst case, and hence our bounds are better by a factor of τ_{mix}.

in generalized linear regression literature [118, 119, 123]. Similarly, GLMtron algorithm for learning NLDS in[73] (see Equation (6.2)) also minimizes a similar proxy loss. In [124], the authors consider a family of GLMtron-like algorithms call Reflectron under the i.i.d. data setting. But they compare the performance of these algorithms experimentally on an NLDS similar to one considered in this work under low rank assumption on the system matrix.

Our offline algorithm, Quasi Newton method, is a standard technique in optimization where the Hessian in the Newton Method is replaced with an approximation to the Hessian. We refer to [125–127] and references therein. Finally, streaming setting for linear system identification has been recently studied in different model settings [66, 110]. These methods observe that by exploiting techniques like experience replay ([128]) along with squared loss error, one can obtain strong error rates.

Organization We setup the problem in section 6.1. In section 6.2 we give the Quasi Newton Method and state results regarding offline estimation of NLDS. In section 6.3 we consider streaming estimation using $\mathsf{SGD} - \mathsf{RER}$ and state its estimation guarantees. In section 6.4 we present the lower bound on ReLU from [105]. Section 6.5 contains the sketch of the proofs and full proofs are in sections 6.6-6.9. Finally the experimental results are in 7.2.

1.3 Bibliographical notes

Chapters 2, 3 and 4 are primarily based on the journal paper [129] and the manuscript [130] (used here with permission from ©IEEE). The work in [130] will be published at the Proceedings of the IEEE International Symposium on Information Theory (ISIT) 2022.

Chapters 5, 6 and 7 are primarily based on the conference publications [122] and [105].

Chapter 2

Many-user quasi-static fading MAC

In this chapter, we will consider the quasi-static Rayleigh fading many-user MAC as discussed in the introduction, and provide tight bounds on the asymptotic minimum energy-per-bit required to achieve a target per-user error at a given user density and payload size.

2.1 System Model

Fix an integer $K \geq 1$ – the number of users. Let $\{P_{Y^n|X^n} = P_{Y^n|X_1^n,X_2^n,\ldots,X_K^n} : \prod_{i=1}^K \mathcal{X}_i^n \to \mathcal{Y}^n\}_{n=1}^\infty$ be a multiple access channel (MAC). In this work we consider only the quasi-static fading AWGN MAC: the channel law $P_{Y^n|X^n}$ is described by

$$Y^n = \sum_{i=1}^K H_i X_i^n + Z^n \tag{2.1}$$

where $Z^n \sim \mathcal{CN}(0, I_n)$, and $H_i \overset{iid}{\sim} \mathcal{CN}(0,1)$ are the fading coefficients which are independent of $\{X_i^n\}$ and Z^n. Naturally, we assume that there is a maximum power constraint:

$$\|X_i^n\|^2 \leq nP. \tag{2.2}$$

We consider two cases: 1) no channel state information (no-CSI): neither the transmitters nor the receiver knows the realizations of channel fading coefficients, but they both know the law; 2) channel state information only at the receiver (CSIR): only the receiver knows the realization of channel fading coefficients. The special case of (2.1) where $H_i = 1, \forall i$ is called the Gaussian MAC (GMAC).

In the rest of the thesis on many-user MAC, we drop the superscript n unless it is unclear.

Definition 1. *An $((M_1, M_2, ..., M_K), n, \epsilon)_U$ code for the MAC $P_{Y^n|X^n}$ is a set of (possibly randomized) maps $\{f_i : [M_i] \to \mathcal{X}_i^n\}_{i=1}^K$ (the encoding functions) and $g : \mathcal{Y}^n \to \prod_{i=1}^K [M_i]$ (the decoder) such that if for $j \in [K]$, $X_j = f_j(W_j)$ constitute the input to the channel and W_j is chosen uniformly (and independently of other W_i, $i \neq j$) from $[M_j]$ then the average (per-user) probability of error satisfies*

$$P_{e,u} = \frac{1}{K} \sum_{j=1}^{K} \mathbb{P}\left[W_j \neq (g(Y))_j\right] \leq \epsilon \tag{2.3}$$

where Y is the channel output.

We define an $((M_1, M_2, ..., M_K), n, \epsilon)_J$ code similarly, where $P_{e,u}$ is replaced by the usual joint error

$$P_{e,J} = \mathbb{P}\left[\bigcup_{j \in [K]} \left\{W_j \neq (g(Y))_j\right\}\right] \leq \epsilon \tag{2.4}$$

Further, if there are cost constraints, we naturally modify the above definitions such that the codewords satisfy the constraints.

Remark 1. *Note that in* (2.3)*, we only consider the* average *per-user probability. But in some situations, it might be relevant to consider* maximal *per-user error (of a codebook tuple) which is the maximum of the probability of error of each user. Formally, let $\mathcal{C}_{[K]} = \{\mathcal{C}_1, ..., \mathcal{C}_K\}$ denote the set of codebooks. Then*

$$\begin{aligned} P_{e,u}^{\max} &= P_{e,u}^{\max}(\mathcal{C}_{[K]}) \\ &= \max\left\{\mathbb{P}\left[W_1 \neq \hat{W}_1\right], ..., \mathbb{P}\left[W_K \neq \hat{W}_K\right]\right\} \end{aligned} \tag{2.5}$$

where the probabilities are with respect to the channel and possibly random encoding and decoding functions. In this paper we only consider the fundamental limits with respect to $P_{e,u}$ and PUPE always refers to this unless otherwise noted. But we note here that for both asymptotics and FBL the difference is not important. See appendix 2.5 for a discussion on this – there we show that by random coding $\mathbb{E}\left[P_{e,u}^{\max}\right]$ is asymptotically equal to $\mathbb{E}\left[P_{e,u}\right]$ (expectations are over random codebooks).

2.1.1 Connection to compressed sensing and sparse regression codes

The system model and coding problem considered in this work (see eqn. (2.1)) can be cast as a support recovery problem in compressed sensing. Suppose we have K users each with a codebook of size M and blocklength n. Let A_i be the $n \times M$ matrix consisting of the codewords of user i as columns. Then the codeword transmitted by the user can be represented as $X_i = A_i W_i$ where

$W_i \in \{0,1\}^M$ with a single nonzero entry. Since each codeword is multiplied by a scalar random gain H_i, we let $U_i = H_i W_i$ which is again a 1 sparse vector of length M. Finally the received vector Y can be represented as

$$Y = \sum_{i \in [K]} H_i X_i + Z = AU + Z \tag{2.6}$$

where $A = [A_1, \cdots, A_K]$ is $n \times KM$ matrix obtained by concatenating the codebooks, $U = [U_1^T, \cdots, U_K^T]^T$ is the $MK \times 1$ length vector denoting the codewords (and fading gains) of each user. In our problem, the vector U has a *block-sparse* structure, namely U has K sections, each of length M, and there is only a single non-zero entry in each section. (Majority of compressed sensing literature focuses on the *non-block-sparse* case, where U has just K non-zero entries, which can be spread arbitrarily inside KM positions.) Decoding of the codewords, then, is equivalent to the support recovery problem under the block-sparse structure, a problem considered in compressed sensing. In our setup, we keep M fixed and let $K, n \to \infty$ with constant $\mu = K/n$. Hence $1/M$ is the sparsity rate and $M\mu$ is the measurement rate.

This connection is not new and has been observed many times in the past [131, 132]. In [132] the authors consider a the exact support recovery problem in the case when the vector U is just sparse (with or without random gains). This corresponds to the random access version of our model where the users share a same codebook [133]. They analyze the fundamental limits in terms of the rate (i.e., ratio of logarithm of signal size to number of measurements) necessary and sufficient to ensure exact recovery in both cases when sparsity is fixed and growing with the signal size. For the fixed sparsity case and U having only $0, 1$ entries, this fundamental limit is exactly the symmetric capacity of an AWGN multiple access channel with same codebook (with non colliding messages). With fading gains, they recover the outage capacity of quasi-static MAC [134, 135] (but with same codebook).

In [131], the authors discuss necessary and sufficient conditions for the exact and approximate support recovery (in Hamming distortion), and L_2 signal recovery with various conditions on signal X and matrix A (deterministic versus random, discrete versus continuous support etc.). These results differ from ours in the sense that they are not for block sparse setting and more importantly, they do not consider approximate support recovery with Hamming distortion when the entries of the support of the signal are sampled from a continuous distribution, which is the case we analyze. Hence our results are not directly comparable.

Work [22] comes closest to our work in terms of the flavor of results of achievability. As pointed out in [22] itself, many other works like [131] focus on the necessary and sufficient scalings (between sparsity, measurements and signal dimension) for various forms of support recovery. But the emphasis in [22] and this work is on the precise constants associated with these scalings. In particular, the authors in [22] consider the approximate support recovery (in Hamming distortion)

problem when the entries in the support of the signal come from a variety of distributions. They analyze various algorithms, including matched filter and AMP, to find the minimum measurement rate required to attain desired support distortion error in terms of signal to noise ratio and other parameters. Furthermore, they compare these results to that of the optimal decoder predicted by the replica method [136].

The result on using replica method in [22] is not directly applicable since our signal has block sparse (as opposed to i.i.d.) coordinates. But the AMP analysis presented there can be extended to our setting. Because of the generality of the analysis in [28], it turns out to be possible to derive rigorous claims (and computable expressions) on the performance of the (scalar) AMP even in the block sparse setting. This is the content of Section 2.3.2 below. Unlike the achievability side, for the converse we cannot rely on bounds in [22] proven for the i.i.d. coordinates of X. Even ignoring the difference between the structural assumptions on X, we point out also that our converse bounds leverage finite-length results from [6], which makes them tighter than the genie-based bounds in [137].

The block-sparse assumption, however, comes very naturally in the area of SPARCs [138–140]. The section error rate (SER) of a SPARC is precisely our PUPE. The vector AMP algorithm has been analyzed for SPARC with i.i.d Gaussian design matrix in [140] and for the spatially-coupled matrix in [141] but for the AWGN channel (i.e., when non-zero entries of U in (2.6) are all 1). In[142], heuristic derivation of state evolution of the vector-AMP decoder for spatially-coupled SPARCS was presented for various signal classes (this includes our fading scenario). However, the the resulting fixed point equations may not be possible to solve for our block size as it amounts to computing 2^{100} dimensional integrals (and this also prevents evaluation of replica-method predictions from [142]).

2.2 Classical regime: K fixed, $n \to \infty$

In this section, we focus on the channel under classical asymptotics where K is fixed (and large) and $n \to \infty$. Further, we consider two distinct cases of joint error and per-user error. We show that subspace projection decoder (2.9) achieves a) ϵ–capacity region ($C_{\epsilon,J}$) for the joint error and b) the best known bound for ϵ–capacity region $C_{\epsilon,PU}$ under per-user error. This motivates using projection decoder in the many-user regime.

2.2.1 Joint error

A rate tuple $(R_1, ..., R_K)$ is said to be ϵ–achievable [135] for the MAC if there is a sequence of codes whose rates are asymptotically at least R_i such that joint error is asymptotically smaller than ϵ. Then the ϵ–capacity region $C_{\epsilon,J}$ is the closure of the set of ϵ–achievable rates. For our channel (2.1), the $C_{\epsilon,J}$ does not depend on whether or not the channel state information (CSI) is available at the receiver since the fading coefficients can be reliably estimated with negligible rate penalty as $n \to \infty$

[134][23]. Hence from this fact and using [135, Theorem 5] it is easy to see that, for $0 \leq \epsilon < 1$, the ϵ–capacity region is given by

$$C_{\epsilon,J} = \{R = (R_1, ..., R_K) : \forall i, R_i \geq 0 \text{ and } P_0(R) \leq \epsilon\} \tag{2.7}$$

where the outage probability $P_0(R)$ is given by

$$P_0(R) = \mathbb{P}\left[\bigcup_{S\subset[K],S\neq\emptyset}\left\{\log\left(1 + P\sum_{i\in S}|H_i|^2\right) \leq \sum_{i\in S} R_i\right\}\right] \tag{2.8}$$

Next, we define a subspace projection based decoder, inspired from [19]. The idea is the following. Suppose there were no additive noise. Then the received vector will lie in the subspace spanned by the sent codewords no matter what the fading coefficients are. To formally define the decoder, let C denote a set of vectors in $\mathbb{C}^n$. Denote P_C as the orthogonal projection operator onto the subspace spanned by C. Let $P_C^{\perp} = I - P_C$ denote the projection operator onto the orthogonal complement of $span(C)$ in $\mathbb{C}^n$.

Let $\mathcal{C}_1, ..., \mathcal{C}_K$ denote the codebooks of the K users respectively. Upon receiving Y from the channel the decoder outputs $g(Y)$ which is given by

$$\begin{aligned} g(Y) &= \left(f_1^{-1}(\hat{c}_1), ..., f_K^{-1}(\hat{c}_K)\right) \\ (\hat{c}_1, ...\hat{c}_K) &= \arg\max_{(c_i\in\mathcal{C}_i)_{i=1}^K} \left\|P_{\{c_i:i\in[K]\}}Y\right\|^2 \end{aligned} \tag{2.9}$$

where f_i are the encoding functions.

In this section, we show that using spherical codebook with projection decoding, $C_{\epsilon,J}$ of the K–MAC is achievable. We prove the following theorem

Theorem 2.2.1 (Projection decoding achieves $C_{\epsilon,J}$)**.** *Let $R \in C_{\epsilon,J}$ of* (2.1)*. Then R is ϵ–achievable through a sequence of codes with the decoder being the projection decoder* (2.9)*.*

Proof. We generate codewords iid uniformly on the power sphere and show that (2.9) yields a small $P_{e,J}$. See [129, Appendix A.A] for details. □

Remark 2. *Note that [132] also analyzed capacity region of the quasi-static MAC, but under the same codebook requirement, for the joint error probability (as opposed to PUPE), and with a different decoder.*

2.2.2 Per-user error

In this subsection, we consider the case of per-user error under the classical setting. Further, we assume availability of CSI at receiver (CSIR) which again can be estimated with little penalty.

The ϵ–capacity region for the channel under per-user error, $C_{\epsilon,PU}$ is defined similarly as $C_{\epsilon,J}$ but with per-user error instead of joint error. $C_{\epsilon,PU}$ is unknown, but the best lower bound is given by the Shamai-Bettesh capacity bound [23]: given a rate tuple $R = (R_1, ..., R_K)$, an upper bound on the per-user probability of error under the channel (2.1), as $n \to \infty$, is given by

$$\begin{aligned} P_{e,u} &\leq P_e^S(R) \\ &\equiv 1 - \frac{1}{K}\mathbb{E}\sup\left\{|D| : D \subset [K], \forall S \subset D, S \neq \emptyset, \right. \\ &\qquad \left. \sum_{i\in S} R_i < \log\left(1 + \frac{P\sum_{i\in S}|H_i|^2}{1 + P\sum_{i\in D^c}|H_i|^2}\right)\right\} \end{aligned} \tag{2.10}$$

where the maximizing set, among all those that achieve the maximum, is chosen to contain the users with largest fading coefficients. The corresponding achievability region is

$$C_{\epsilon,PU}^{S.B} = \left\{R : P_e^S(R) \leq \epsilon\right\} \tag{2.11}$$

and hence it is an inner bound on $C_{\epsilon,PU}$.

We note that, in [23], only the symmetric rate case i.e, $R_i = R_j \, \forall i, j$ is considered. So (2.10) is the extension of that result to the general non-symmetric case.

Here, we show that the projection decoding (suitably modified to use CSIR) achieves the same asymptotics as (2.10) for per-user probability of error i.e., achieves the Shamai-Bettesh capacity bound. Next we describe the modification to the projection decoder to use CSIR.

Let $\{\mathcal{C}_i\}_{i=1}^K$ denote the codebooks of the K users with $|\mathcal{C}_i| = M_i$. We have a maximum power constraint given by (2.2). Using the idea of joint decoder from [23], our decoder works in 2 stages. The first stage finds the following set

$$\begin{aligned} D \in \ \arg\max &\left\{|D| : D \subset [K], \forall S \subset D, S \neq \emptyset, \right. \\ &\left. \sum_{i\in S} R_i < \log\left(1 + \frac{P\sum_{i\in S}|H_i|^2}{1 + P\sum_{i\in D^c}|H_i|^2}\right)\right\} \end{aligned} \tag{2.12}$$

where D is chosen to contain users with largest fading coefficients. The second stage is similar to (2.9) but decodes only those users in D. Formally, let ? denote an error symbol. The decoder output $g_D(Y) \in \prod_{i=1}^K \mathcal{C}_i$ is given by

$$\begin{aligned} (g_D(Y))_i &= \begin{cases} f_i^{-1}(\hat{c}_i) & i \in D \\ ? & i \notin D \end{cases} \\ (\hat{c}_i)_{i\in D} &= \arg\max_{(c_i\in\mathcal{C}_i)_{i\in D}} \left\|P_{\{c_i : i\in D\}}Y\right\|^2 \end{aligned} \tag{2.13}$$

where f_i are the encoding functions. Our error metric is the average per-user probability of error (2.4).

The following theorem is the main result of this section.

Theorem 2.2.2. *For any $R \in C_{\epsilon,PU}^{S.B}$ there exists a sequence of codes with projection decoder* (2.12)(2.13) *with asymptotic rate R such that the per-user probability of error is asymptotically smaller than ϵ*

Proof. We generate iid (complex) Gaussian codebooks $\mathcal{CN}(0, P'I_n)$ with $P' < P$ and show that for $R \in C_{\epsilon,PU}^{S.B}$, (2.13) gives small $P_{e,u}$. See [129, Appendix A.B] for details. □

In the case of symmetric rate, an outer bound on $C_{\epsilon,PU}$ can be given as follows.

Proposition 1. *If the symmetric rate R is such that $P_{e,u} \leq \epsilon$, then*

$$R \leq \min\Bigg\{ \frac{1}{K(\theta-\epsilon)} \mathbb{E}\left[\log_2 \left(1 + P \min_{\substack{S\subset[K] \\ |S|=\theta K}} \sum_{i\in S} |H_i|^2 \right) \right], \\ \log_2\left(1 - P\ln(1-\epsilon)\right) \Bigg\}, \forall \theta \in (\epsilon, 1] \tag{2.14}$$

Proof. The first of the two terms in the min in (2.14) follows from Fano's inequality (see (2.159), with $\mu = K/n$, $M = 2^{nR}$ and taking $n \to \infty$). The second is a single-user based converse using a genie argument. See appendix 2.4.1 for details. □

Remark 3. *We note here that the second term inside the minimum in* (2.14) *is the same as the one we would obtain if we used strong converse for the MAC. To be precise, let $\{|H_{(1)}| > |H_{(2)}| > ... > |H_{(K)}|\}$ denote the order statistics of the fading coefficients. If $R > \log(1 + P|H_{(t)}|^2)$ then, using a Genie that reveals the codewords (and fading gains) of $t-1$ users corresponding to $t-1$ largest fading coefficients, it can be seen that $P_{e,u} \geq \frac{K-t+1}{K}$. Setting $t = \theta K$ and considering the limit as $K \to \infty$ (with $P = P_{tot}/K$) we obtain $S \leq -P_{tot}\log_2(1-\epsilon)$ which is same as that obtained from the second term in* (2.14) *under these limits.*

2.2.3 Numerical evaluation

First notice that $C_{\epsilon,J}$ (under joint error) tends to $\{0\}$ as $K \to \infty$ because, it can be seen, for the symmetric rate, by considering that order statistics of the fading coefficients that $P_0(R) \to 1$ for $R_i = O(1/K)$. $C_{\epsilon,PU}$, however, is more interesting. We evaluate trade-off between system spectral efficiency and the minimum energy-per-bit required for a target per-user error for the symmetric rate, in the limit $K \to \infty$ and power scaling as $O(1/K)$.

In the above figure we have also presented the performance of TDMA. That is, if we use orthogonalization then for any number of users K (not necessarily large), we have

$$\epsilon = \mathbb{P}\left[R > 1/K \log(1 + KP|H|^2)\right] \tag{2.15}$$

where ϵ is the PUPE. Thus the sum-rate vs E_b/N_0 formula for orthogonalization is

$$E_b/N_0 = \frac{2^S - 1}{S} \frac{1}{-\ln(1-\epsilon)} \tag{2.16}$$

where S is the sum-rate or the spectral efficiency.

We see that orthogonalization is suboptimal under the PUPE criterion. The reason is that it fails to exploit the multi-user diversity by allocating resources even to users in deep fades. Indeed, under orthogonalized setting the resources allocated to a user that happens to experience a deep fade become completely wasted, while non-orthogonal schemes essentially adapt to the fading realization: the users in deep fades create very little interference for the problem of decoding strong users. This is the effect stemming from the PUPE criterion for error rate.

2.3 Many user MAC: $K = \mu n$, $n \to \infty$

This is our main section. We consider the linear scaling regime where the number of users K scales with n, and $n \to \infty$. We are interested in the tradeoff of minimum E_b/N_0 required for the PUPE to be smaller than ϵ, with the user density μ ($\mu < 1$). So, we fix the message size k. Let $S = k\mu$ be the spectral efficiency.

We focus on the case of different codebooks, but under symmetric rate. So if M denotes the size of the codebooks, then $S = \frac{K \log M}{n} = \mu \log M$. Hence, given S and μ, M is fixed. Let $P_{tot} = KP$ denote the total power. Therefore denoting by $\mathcal{E}$ the energy-per-bit, $\mathcal{E} = E_b/N_0 = \frac{nP}{\log_2 M} = \frac{P_{tot}}{S}$. For finite E_b/N_0, we need finite P_{tot}, hence we consider the power P decaying as $O(1/n)$.

Let $\mathcal{C}_j = \{c_1^j, ..., c_M^j\}$ be the codebook of user j, of size M. The power constraint is given by $\left\|c_i^j\right\|^2 \leq nP = \mathcal{E} \log_2 M, \forall j \in [K], i \in [M]$. The collection of codebooks $\{\mathcal{C}_j\}$ is called an $(n, M, \epsilon, \mathcal{E}, K)$–code if it satisfies the power constraint described before, and the per-user probability of error is smaller than ϵ. Then, we can define the following fundamental limit for the channel

$$\mathcal{E}^*(M, \mu, \epsilon) = \lim_{n \to \infty} \inf \left\{\mathcal{E} : \exists (n, M, \epsilon, \mathcal{E}, K = \mu n) - code\right\}.$$

We make an important remark here that all the following results also hold for maximal per-user error (PUPE-max) (2.5) as discussed in appendix 2.5.

2.3.1 No-CSI: Projection decoding

In this subsection, we focus on the no-CSI case. The difficulty here is that, a priori, we do not know which subset of the users to decode. We have the following theorem.

Theorem 2.3.1. *Consider the channel* (2.1) *(no-CSI) with* $K = \mu n$ *where* $\mu < 1$. *Fix the spectral efficiency* S *and target probability of error (per-user)* ϵ. *Let* $M = 2^{S/\mu}$ *denote the size of the codebooks and* $P_{tot} = KP$ *be the total power. Fix* $\nu \in (1-\epsilon, 1]$. *Let* $\epsilon' = \epsilon - (1-\nu)$. *Then if* $\mathcal{E} > \mathcal{E}^*_{no-CSI} = \sup_{\frac{\epsilon'}{\nu} < \theta \leq 1} \sup_{\xi \in [0, \nu(1-\theta)]} \frac{P_{tot,\nu}(\theta,\xi)}{S}$, *there exists a sequence of* $(n, M, \epsilon_n, \mathcal{E}, K = \mu n)$ *codes such that* $\limsup_{n\to\infty} \epsilon_n \leq \epsilon$, *where, for* $\frac{\epsilon'}{\nu} < \theta \leq 1$ *and* $\xi \in [0, \nu(1-\theta)]$,

$$P_{tot,\nu}(\theta,\xi) = \frac{\hat{f}(\theta,\xi)}{1 - \hat{f}(\theta,\xi)\alpha\left(\xi + \nu\theta, \xi + 1 - \nu(1-\theta)\right)} \tag{2.17}$$

$$\hat{f}(\theta,\xi) = \frac{f(\theta)}{\alpha(\xi, \xi+\nu\theta)} \tag{2.18}$$

$$f(\theta) = \frac{\frac{1+\delta_1^*(1-V_\theta)}{V_\theta} - 1}{1-\delta_2^*} \tag{2.19}$$

$$V_\theta = e^{-\tilde{V}_\theta} \tag{2.20}$$

$$\tilde{V}_\theta = \delta^* + \frac{\theta\mu\nu \ln M}{1-\mu\nu} + \frac{1-\mu\nu(1-\theta)}{1-\mu\nu} h\left(\frac{\theta\mu\nu}{1-\mu\nu(1-\theta)}\right) + \frac{\mu(1-\nu(1-\theta))}{1-\mu\nu} h\left(\frac{\theta\nu}{1-\nu(1-\theta)}\right) \tag{2.21}$$

$$\delta^* = \frac{\mu h(1-\nu(1-\theta))}{1-\mu\nu} \tag{2.22}$$

$$c_\theta = \frac{2V_\theta}{1-V_\theta} \tag{2.23}$$

$$q_\theta = \frac{\mu h(1-\nu(1-\theta))}{1-\mu\nu(1-\theta)} \tag{2.24}$$

$$\delta_1^* = q_\theta(1+c_\theta) + \sqrt{q_\theta^2(c_\theta^2+2c_\theta) + 2q_\theta(1+c_\theta)} \tag{2.25}$$

$$\delta_2^* = \inf\left\{x : 0 < x < 1, -\ln(1-x) - x > \frac{\mu h(1-\nu(1-\theta))}{1-\mu\nu(1-\theta)}\right\} \tag{2.26}$$

$$\alpha(a,b) = a\ln(a) - b\ln(b) + b - a. \tag{2.27}$$

Hence $\mathcal{E}^* \leq \mathcal{E}^*_{no-CSI}$.

Proof Idea. Before we present the full proof, the main ideas are presented here. Also, over the course, we explain the quantities that are present in the statement of the theorem. We start with choosing independent random Gaussian codebooks for all users. That is, for each message of each user there is an independent complex Gaussian $\mathcal{CN}(0, P'I_n)$ codeword where $P' < P$. The choice $P' < P$ is to ensure we can control the maximum power constraint violation events.

For simplicity we will consider $\nu = 1$. Here ν represents the fraction of users that the decoder can choose to decode. Due to random coding, we can assume that a particular tuple of codewords

$(c_1, c_2, \cdots, c_K)$ were transmitted i.e., the received vector at the decoder is $Y = \sum_{i=1}^{K} H_i c_i + Z$. Then the decoder performs subspace projection decoding. The idea is that in the absence of noise, the received vector lies in the subspace spanned by the K codewords. Since we assume $\mu = K/n < 1$, and the K codewords are linearly independent, we can uniquely decode them by projecting the received vector onto various K dimensional subspaces formed by taking a codeword from each of the codebooks. Formally,

$$\{\hat{c}_i : i \in [K]\} = \arg \max_{(c_i \in \mathcal{C}_i : i \in [K])} \left\| P_{\{c_i : i \in S\}} Y \right\|^2 \tag{2.28}$$

Notice that the PUPE is given by

$$P_e = \frac{1}{K} \sum_{i \in [K]} \mathbb{P}\left[c_i \neq \hat{c}_i\right].$$

We will bound this error with the probability of events F_t– event that exactly t users were misdecoded. That is

$$P_e \leq \epsilon + \mathbb{P}\left[\bigcup_{t > \epsilon K} F_t \right] \tag{2.29}$$

Hence it is enough to find conditions under which the second term (call it p_1) in the above display goes to 0 in our scaling. To analyze F_t, we consider subsets $S \subset [K]$ with $|S| = t$ and a choice of *incorrect* codewords $(c_i' \in \mathcal{C}_i : i \in S)$ where $c_i' \neq c_i$, and bound F_t as union (over S and $(c_i' : i \in [S])$) of events $\left\{ \left\| P_{c'_{[S]}, c_{[[K]\setminus S]}} Y \right\|^2 > \left\| P_{c_{[[K]]}} Y \right\|^2 \right\}$. With abuse of notation, denote this set as $F(S,t)$.

Let $c_{[S]} = \{c_i : i \in S\}$, similarly we have $H_{[S]}$. We make a crucial observation that, conditioned on $c_{[K]}$, $H_{[K]}$ and Z, the random variable $\left\| P_{c'_{[S]}, c_{[[K]\setminus S]}} Y \right\|^2$ can be written as $\left\| P_{c_{[K\setminus S]}} Y \right\|^2 + \left\| P^{\perp}_{c_{[K\setminus S]}} Y \right\|^2 \text{Beta}(t, n-K)$ where $\text{Beta}(a,b)$ is a beta distributed random variable with parameters a and b.

Let $G_S = \frac{\left\| P^{\perp}_{c_{[[K]]}} Y \right\|^2}{\left\| P^{\perp}_{c_{[K\setminus S]}} Y \right\|^2}$. Then we show that

$$\begin{aligned} \mathbb{P}\left[F(S,t) | c_{[K]}, H_{[K]}, Z\right] &= \mathbb{P}\left[\text{Beta}(n-K, t) < G_S | c_{[K]}, H_{[K]}, Z\right] \\ &\leq \binom{n-K+t-1}{t-1} (G_S)^{n-K} \end{aligned} \tag{2.30}$$

Next, we use the idea of random coding union (RCU) bound [143] to get

$$\mathbb{P}\left[\bigcup_t F_t \right] \leq \mathbb{E}\left[\min\left\{ 1, \sum_{t,S} \mathbb{P}\left[F(S,t) | c_{[K]}, H_{[K]}, Z\right] \right\} \right] \tag{2.31}$$

Let $\theta = t/K$, which is the fraction of misdecoded users. Now, by thresholding the value of G_S (this threshold is parameterized by a $\delta > 0$) we get from (2.31) a sum of an exponentially decaying

term with combinatorial factors and the probability that G_S violates this threshold for some S and t (call this probability p_2). Choosing the right threshold (δ^* and corresponding threshold value V_θ in the theorem) the first term vanishes (in the limit) and we are left with p_2.

This is analyzed by conditioning on $c_{[K]}$ and $H_{[K]}$ along with using concentration of non-central chi-squared distributed variables (see claim 3). We follow similar procedure to above (using RCU and thresholding) multiple times to obtain thresholds parameterized by δ_1^* and δ_2^* to vanish combinatorial factors (like q_θ in the theorem which is the exponent of a binomial coefficient) and finally we are left with the *bottleneck* term:

$$\limsup_n P_e \leq \epsilon + \limsup_n \mathbb{P}\left[\bigcup_{t,S}\left\{P' \sum_{i\in[S]} |H_i|^2 < g\left(\delta_1^*, \delta_2^*, \delta_3^*, M, \mu, \theta\right)\right\}\right] \tag{2.32}$$

where g is some specific function. In essence, this bottleneck term is precisely the event that $> \epsilon$ fraction of users are outside the Gaussian capacity region!

Next step is to replace $\cup_S$ with $\min_S$ and use the convergence of order statistics of fading coefficients i.e., $|H_{(1)}| > \cdots > |H_{(K)}|$:

$$\limsup_n P_e \leq \epsilon + \limsup_n \mathbb{P}\left[\bigcup_{t}\left\{P' \sum_{i=K-t+1}^{K} |H_{(i)}|^2 < g\left(\delta_1^*, \delta_2^*, \delta_3^*, M, \mu, t/K\right)\right\}\right] \tag{2.33}$$

Then we show that, for $t = \theta K$, $\frac{1}{K}\sum_{i=K-t+1}^{K} |H_{(i)}|^2 \to \int_{1-\theta}^{1} F^{-1}_{|H|^2}(1-\gamma)\, d\gamma \equiv \alpha(1-\theta, 1)$ in probabilty as $n \to \infty$. Hence the bottleneck term becomes deterministic in the limit. The choice P_{tot} such that this terms vanishes is precisely the one given in the statement of the theorem. $\square$

Proof. The proof uses random coding. Let each user generate a Gaussian codebook of size M and power $P' < P$ independently such that $KP' = P'_{tot} < P_{tot}$. Let W_j denote the random (in $[M]$) message of user j. So, if $\mathcal{C}_j = \{c_i^j : i \in [M]\}$ is the codebook of user j, he/she transmits $X_j = c^j_{W_j} 1\left\{\left\|c^j_{W_j}\right\|^2 \leq nP\right\}$. For simplicity let $(c_1, c_2, ..., c_K)$ be the sent codewords. Hence the received vector is $Y = \sum_{i\in[K]} H_i c_i + Z$ where Z is the noise vector. Fix $\nu \in (1-\epsilon, 1]$. Let $K_1 = \nu K$ be the number of users that are decoded. Since there is no knowledge of CSIR, it is not possible to, a priori, decide what set to decode. Instead, the decoder searches of all K_1 sized subsets of $[M]$. Formally, let ? denote an error symbol. The decoder output $g_D(Y) \in \prod_{i=1}^{k} \mathcal{C}_i$ is given by

$$\left[\hat{S}, (\hat{c}_i)_{i\in\hat{S}}\right] = \arg \max_{\substack{S\subset[K]\\|S|=K_1}} \max_{(c_i\in\mathcal{C}_i)_{i\in S}} \left\|P_{\{c_i : i\in S\}} Y\right\|^2$$

$$(g_D(Y))_i = \begin{cases} f_i^{-1}(\hat{c}_i) & i \in \hat{S} \\ ? & i \notin \hat{S} \end{cases} \tag{2.34}$$

where f_i are the encoding functions. The probability of error (averaged over random codebooks) is given by

$$P_e = \frac{1}{K}\sum_{j=1}^{K}\mathbb{P}\left[W_j \neq \hat{W}_j\right] \tag{2.35}$$

where $\hat{W}_j = (g(Y))_j$ is the decoded message of user j.

We perform a change of measure to $X_j = c^j_{W_j}$. Since P_e is the expectation of a non-negative random variable bounded by 1, this measure change adds a total variation distance which can bounded by $p_0 = K\mathbb{P}\left[\frac{\chi_2(2n)}{2n} > \frac{P}{P'}\right] \to 0$ as $n \to \infty$, where $\chi_2(d)$ is the distribution of sum of squares of d iid standard normal random variables (the chi-square distribution). The reason is as follows. If we have two random vectors U_1 and U_2 on a the same probability space such that $U_1 = U_2 1[U_2 \in E]$, where E is a Borel set, then for any Borel set A, we have

$$\begin{aligned}|\mathbb{P}\left[U_1 \in A\right] - \mathbb{P}\left[U_2 \in A\right]| &= |1[0 \in A]\mathbb{P}\left[U_2 \in E^c\right] - \mathbb{P}\left[U_2 \in A \cap E^c\right]| \\ &\leq \mathbb{P}\left[U_2 \in E^c\right].\end{aligned} \tag{2.36}$$

Henceforth we only consider the new measure.

Let $\epsilon > 1 - \nu$ and $\epsilon' = \epsilon - (1 - \nu)$. Now we have

$$\begin{aligned}P_e &\leq \epsilon + \mathbb{P}\left[\frac{1}{K}\sum_{j=1}^{K} 1[W_j \neq \hat{W}_j] > \epsilon\right] \\ &= \epsilon + \mathbb{P}\left[\sum_{j=1}^{K} 1[W_j \neq \hat{W}_j] > K\epsilon' + K - K_1\right] \\ &= \epsilon + p_1.\end{aligned} \tag{2.37}$$

where

$$p_1 = \mathbb{P}\left[\bigcup_{t=\epsilon' K}^{\nu K}\left\{\sum_{j=1}^{K} 1[W_j \neq \hat{W}_j] = K - K_1 + t\right\}\right].$$

Let $F_t = \left\{\sum_{j=1}^{K} 1[W_j \neq \hat{W}_j] = K - K_1 + t\right\}$. Let $c_{[S]} \equiv \{c_i : i \in S\}$ and $H_{[S]} \equiv \{H_i : i \in S\}$, where $S \subset [K]$. Conditioning on $c_{[K]}, H_{[K]}$ and Z, we have

$$\begin{aligned}\mathbb{P}\left[F_t|c_{[K]},H_{[K]},Z\right] \leq \mathbb{P}\Bigg[&\exists S\subset[K]:|S|=K-K_1+t,\exists S_1\subset S:|S_1|=t,\\ &\exists\{c_i'\in\mathcal{C}_i:i\in S_1,c_i'\neq c_i\}:\left\|P_{c'_{[S_1]},c_{[[K]\setminus S]}}Y\right\|^2>\\ &\max_{\substack{S_2\subset S\\|S_2|=t}}\left\|P_{c_{[S_2]},c_{[[K]\setminus S]}}Y\right\|^2\Bigg|c_{[K]},H_{[K]},Z\Bigg]\\ \leq \mathbb{P}\Bigg[&\bigcup_{\substack{S\subset[K]\\|S|=K-K_1+t}}\bigcup_{\substack{S_1\subset S\\|S_1|=t}}\bigcup_{\substack{\{c_i'\in\mathcal{C}_i:\\ i\in S_1,c_i'\neq c_i\}}}F(S,S_2^*,S_1,t)\Bigg|c_{[K]},H_{[K]},Z\Bigg]\end{aligned}\tag{2.38}$$

where

$$F(S,S_2^*,S_1,t)=\left\{\left\|P_{c'_{[S_1]},c_{[[K]\setminus S]}}Y\right\|^2>\left\|P_{c_{[S_2^*]},c_{[[K]\setminus S]}}Y\right\|^2\right\}$$

and $S_2^*\subset S$ is a possibly random (depending only on $H_{[K]}$) subset of size t, to be chosen later. Next we will bound $\mathbb{P}\left[F(S,S_2^*,S_1,t)|c_{[K]},H_{[K]},Z\right]$.

For the sake of brevity, let $A_0=c_{[S_2^*]}\cup c_{[[K]\setminus S]}$, $A_1=c_{[[K]\setminus S]}$ and $B_1=c'_{[S_1]}$. We have the following claim.

Claim 1. *For any $S_1\subset S$ with $|S_1|=t$, conditioned on $c_{[K]}$, $H_{[K]}$ and Z, the law of $\left\|P_{c'_{[S_1]},c_{[[K]\setminus S]}}Y\right\|^2$ is same as the law of $\|P_{A_1}Y\|^2+\|(I-P_{A_1})Y\|^2\,\mathrm{Beta}(t,n-K_1)$ where $\mathrm{Beta}(a,b)$ is a beta distributed random variable with parameters a and b.*

Proof. Let us write $V=span\{A_1,B_1\}=A\oplus B$ where $A\perp B$ are subspaces of dimension K_1-t and t respectively, with $A=span(A_1)$ and B is the orthogonal complement of A_1 in V. Hence $\|P_VY\|^2=\|P_AY\|^2+\|P_BY\|^2$ (by definition, $P_A=P_{A_1}$). Now we analyze $\|P_BY\|^2$. We can further write $P_BY=P_BP_A^\perp Y$. Observe that the subspace B is the span of $P_A^\perp B_1$, and, conditionally, $P_A^\perp B_1\sim\mathcal{CN}^{\otimes|S|}(0,P'P_A^\perp)$ which is the product measure of $|S|$ complex normal vectors in a subspace of dimension $n-K_1+t$. Hence, the conditional law of $\left\|P_BP_A^\perp Y\right\|^2$ is the law of squared length of projection of a fixed $n-K_1+t$ dimensional vector of length $\|(I-P_A)Y\|^2$ onto a (uniformly) random t dimensional subspace.

Further, the law of the squared length of the orthogonal projection of a fixed unit vector in $\mathbb{C}^d$ onto a random t–dimensional subspace is same as the law of the squared length of the orthogonal projection of a random unit vector in $\mathbb{C}^d$ onto a fixed t–dimensional subspace, which is $\mathrm{Beta}(t,d-t)$ (see for e.g. [144, Eq. 79]): that is, if u is a unit random vector in $\mathbb{C}^d$ and L is a fixed t dimensional subspace, then $\mathbb{P}\left[\|P_Lu\|^2\leq x\right]=\mathbb{P}\left[\frac{\sum_{i=1}^t|Z_i|^2}{\sum_{i=1}^d|Z_i|^2}\leq x\right]=F_\beta(x;t,n-K_1)$ where $Z_i\overset{iid}{\sim}\mathcal{CN}(0,1)$ and

$F_\beta(x;a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^x w^{a-1}(1-w)^{b-1}dw$ denotes the CDF of the beta distribution with parameters a and b. Hence the conditional law of $\left\|P_B P_A^{\perp} Y\right\|^2$ is $\|(I-P_A)Y\|^2 \operatorname{Beta}(t, n-K_1)$.

□

Therefore we have,

$$\begin{aligned}\mathbb{P}\left[F(S,S_2^*,S_1,t)|c_{[K]},H_{[K]},Z\right] &= \mathbb{P}\left[Beta(n-K_1,t) < G_S|c_{[K]},H_{[K]},Z\right] \\ &= F_\beta\left(G_S;n-K_1,t\right) \end{aligned} \tag{2.39}$$

where

$$G_S = \frac{\|Y\|^2 - \|P_{A_0}Y\|^2}{\|Y\|^2 - \|P_{A_1}Y\|^2}. \tag{2.40}$$

Since $t \geq 1$, we have

$$F_\beta\left(G_S;n-K_1,t\right) \leq \binom{n-K_1+t-1}{t-1} G_S^{n-K_1}. \tag{2.41}$$

Let us denote $\bigcup_{t=\epsilon' K}^{\nu K}$ as $\bigcup_t$, $\bigcup_{\substack{S\subset[K]\\|S|=K-K_1+t}}$ as $\bigcup_{S,K_1}$, and $\bigcup_t \bigcup_{\substack{S\subset[K]\\|S|=K-K_1+t}}$ as $\bigcup_{t,S,K_1}$; similarly for $\sum$ and $\bigcap$ for the ease of notation. Using the above claim, we get,

$$\mathbb{P}\left[F_t|c_{[K]},H_{[K]},Z\right] \leq \sum_{S,K_1} \binom{K-K_1+t}{t} M^t \binom{n-K_1+t-1}{t-1} G_S^{n-K_1}. \tag{2.42}$$

Therefore p_1 can be bounded as

$$\begin{aligned} p_1 &= \mathbb{P}\left[\bigcup_t F_t\right] \\ &\leq \mathbb{E}\left[\min\left\{1, \sum_{t,S,K_1} \binom{K-K_1+t}{t} M^t \binom{n-K_1+t-1}{t-1} G_S^{n-K_1}\right\}\right] \\ &= \mathbb{E}\left[\min\left\{1, \sum_{t,S,K_1} e^{(n-K_1)s_t} M^t G_S^{n-K_1}\right\}\right] \end{aligned} \tag{2.43}$$

where $s_t = \frac{\ln\left(\binom{K-K_1+t}{t}\binom{n-K_1+t-1}{t-1}\right)}{n-K_1}$.

Now we can bound the binomial coefficient [145, Ex. 5.8] as

$$\begin{aligned}\binom{n-K_1+t-1}{t-1} &\leq \sqrt{\frac{n-K_1+t-1}{2\pi(t-1)(n-K_1)}} e^{(n-K_1+t-1)h(\frac{t-1}{n-K_1+t-1})} \\ &= O\left(\frac{1}{\sqrt{n}}\right) e^{n(1-\mu\nu(1-\theta))h(\frac{\theta\mu\nu}{1-\mu\nu(1-\theta)})}. \end{aligned} \tag{2.44}$$

Similarly,

$$\binom{K-K_1+t}{t} \leq O\left(\frac{1}{\sqrt{n}}\right) e^{n\mu(1-\nu(1-\theta))h\left(\frac{\theta\nu}{1-\nu(1-\theta)}\right)} \tag{2.45}$$

Let $r_t = s_t + \frac{t \ln M}{n-K_1}$. For $\delta > 0$, define $\tilde{V}_{n,t} = r_t + \delta$ and $V_{n,t} = e^{-\tilde{V}_{n,t}}$. Let E_1 be the event

$$E_1 = \bigcap_{t,S,K_1} \{-\ln G_S - r_t > \delta\} = \bigcap_{t,S,K_1} \{G_S < V_{n,t}\}. \tag{2.46}$$

Let $p_2 = \mathbb{P}\left[\bigcup_{t,S,K_1} \{G_S > V_{n,t}\}\right]$. Then

$$\begin{aligned} p_1 &\leq \mathbb{E}\left[\min\left\{1, \sum_{t,S,K_1} e^{(n-K_1)r_t} G_S^{n-K_1}\right\} (1[E_1] + 1[E_1^c])\right] \\ &\leq \mathbb{E}\left[\sum_{t,S,K_1} e^{-(n-K_1)\delta}\right] + p_2 \\ &= \sum_t \binom{K}{K-K_1+t} e^{-(n-K_1)\delta} + p_2. \end{aligned} \tag{2.47}$$

Observe that, for $t = \theta K_1 = \theta\nu K$,

$$\begin{aligned} s_t = &\frac{1-\mu\nu(1-\theta)}{1-\mu\nu} h\left(\frac{\theta\mu\nu}{1-\mu\nu(1-\theta)}\right) + \\ &\frac{\mu(1-\nu(1-\theta))}{1-\mu\nu} h\left(\frac{\theta\nu}{1-\nu(1-\theta)}\right) - O\left(\frac{\ln(n)}{n}\right) \end{aligned}$$

and $r_t = s_t + \frac{\theta\mu\nu}{1-\mu\nu} \ln M$. Therefore $n \to \infty$ with θ fixed, we have

$$\lim_{n\to\infty} \tilde{V}_{n,\theta\nu\mu} = \tilde{V}_\theta \tag{2.48}$$

where $\tilde{V}_\theta$ is given in (2.21).

Now, note that, for $1 < t < K_1$,

$$\binom{K}{K-K_1+t} \leq \sqrt{\frac{K}{2\pi(K-K_1+t)(K_1-t)}} e^{Kh(\frac{K-K_1+t}{K})}. \tag{2.49}$$

Hence choosing $\delta > \frac{Kh(\frac{K-K_1+t}{K})}{n-K_1}$ will ensure that the first term in (2.47) goes to 0 as $n \to \infty$. So for $t = \theta K_1 = \theta\nu K$, we need to have

$$\delta > \delta^*. \tag{2.50}$$

where δ^* is given in (2.22).

Let us bound p_2. Let $\hat{Z} = Z + \sum_{i\in S\setminus S_2^*} H_i c_i$. We have

Claim 2.

$$p_2 = \mathbb{P}\Big[\bigcup_{t,S,K_1}\{G_S > V_{n,t}\}\Big]$$
$$\leq \mathbb{P}\Big[\bigcup_{t,S,K_1}\Big\{\Big\|(1-V_{n,t})P_{A_1}^{\perp}\hat{Z} - V_{n,t}P_{A_1}^{\perp}\sum_{i\in S_2^*}H_i c_i\Big\|^2 \geq V_{n,t}\Big\|P_{A_1}^{\perp}\sum_{i\in S_2^*}H_i c_i\Big\|^2\Big\}\Big] \tag{2.51}$$

Proof. See appendix 2.4.2.

□

Let $\chi_2'(\lambda, d)$ denote the non-central chi-squared distributed random variable with non-centrality λ and degrees of freedom d. That is, if $W_i \sim \mathcal{N}(\mu_i, 1), i \in [d]$ and $\lambda = \sum_{i\in[d]}\mu_i^2$, then $\chi_2'(\lambda, d)$ has the same distribution as that of $\sum_{i\in[d]} W_i^2$. We have the following claim.

Claim 3. *Conditional on $H_{[K]}$ and A_0,*

$$\Big\|P_{A_1}^{\perp}\Big(\hat{Z} - \frac{V_{n,t}}{1-V_{n,t}}\sum_{i\in S_2^*}H_i c_i\Big)\Big\|^2 \sim \Big(1 + P'\sum_{i\in S\setminus S_2^*}|H_i|^2\Big)\frac{1}{2}\chi_2'\left(2F, 2n'\right) \tag{2.52}$$

where

$$F = \frac{\Big\|\frac{V_{n,t}}{1-V_{n,t}}P_{A_1}^{\perp}\sum_{i\in S_2^*}H_i c_i\Big\|^2}{\Big(1 + P'\sum_{i\in S\setminus S_2^*}|H_i|^2\Big)} \tag{2.53}$$
$$n' = n - K_1 + t. \tag{2.54}$$

Hence its conditional expectation is

$$\mu = n' + F. \tag{2.55}$$

Proof. See appendix 2.4.2. □

Now let

$$T = \frac{1}{2}\chi_2'(2F, 2n') - \mu \tag{2.56}$$
$$U = \frac{V_{n,t}}{(1-V_{n,t})}\frac{\Big\|P_{A_1}^{\perp}\sum_{i\in S_2^*}H_i c_i\Big\|^2}{\Big(1 + P'\sum_{i\in S\setminus S_2^*}|H_i|^2\Big)} - n' \tag{2.57}$$
$$U^1 = \frac{1}{1-V_{n,t}}\left(V_{n,t}W_S - 1\right) \tag{2.58}$$

where

$$W_S = \left(1 + \frac{\left\| P_{A_1}^{\perp} \sum_{i \in S_2^*} H_i c_i \right\|^2}{n' \left(1 + P' \sum_{i \in S \setminus S_2^*} |H_i|^2\right)}\right).$$

Notice that $U = n'U^1$ and $F = \frac{V_{n,t}}{1-V_{n,t}} n'(1+U^1)$.

Then we have

$$\begin{aligned} \text{RHS of (2.51)} &= \mathbb{P}\Big[\bigcup_{t,S,K_1} \Big\{ \Big\| P_{A_1}^{\perp} \hat{Z} - \frac{V_{n,t}}{(1-V_{n,t})} P_{A_1}^{\perp} \sum_{i \in S_2^*} H_i c_i \Big\|^2 - \mu \geq U \Big\} \Big] \\ &= \mathbb{P}\Big[\bigcup_{t,S,K_1} \{T \geq U\} \Big]. \end{aligned} \tag{2.59}$$

Now, let $\delta_1 > 0$, and $E_2 = \cap_{t,S,K_1} \left\{U^1 > \delta_1\right\}$. Taking expectations over E_1 and its complement, we have

$$\begin{aligned} \mathbb{P}\Big[\bigcup_{t,S,K_1} \{T \geq U\} \Big] &\leq \sum_{t,S,K_1} \mathbb{P}\left[T > U, U^1 > \delta_1\right] + \mathbb{P}\left[E_2^c\right] \\ &= \sum_{t,S,K_1} \mathbb{E}\left[\mathbb{P}\left[T > U \middle| H_{[K]}, A_0\right] 1[U^1 > \delta_1]\right] + \mathbb{P}\left[E_2^c\right] \end{aligned} \tag{2.60}$$

which follows from the fact that $\{U^1 > \delta_1\} \in \sigma(H_{[K]}, A_0)$. To bound this term, we use the following concentration result from [146, Lemma 8.1].

Lemma 2.3.2 ([146]). *Let $\chi = \chi_2'(\lambda, d)$ be a non-central chi-squared distributed variable with d degrees of freedom and non-centrality parameter λ. Then $\forall x > 0$*

$$\begin{aligned} \mathbb{P}\left[\chi - (d+\lambda) \geq 2\sqrt{(d+2\lambda)x} + 2x\right] &\leq e^{-x} \\ \mathbb{P}\left[\chi - (d+\lambda) \leq -2\sqrt{(d+2\lambda)x}\right] &\leq e^{-x} \end{aligned} \tag{2.61}$$

Hence, for $x > 0$, we have

$$\mathbb{P}\left[\chi - (d+\lambda) \geq x\right] \leq e^{-\frac{1}{2}\left(x + d + 2\lambda - \sqrt{d+2\lambda}\sqrt{2x+d+2\lambda}\right)}. \tag{2.62}$$

and for $x < (d+\lambda)$, we have

$$\mathbb{P}\left[\chi \leq x\right] \leq e^{-\frac{1}{4}\frac{(d+\lambda-x)^2}{d+2\lambda}}. \tag{2.63}$$

Observe that, in (2.62), the exponent is always negative for $x > 0$ and finite λ due to AM-GM inequality. When $\lambda = 0$, we can get a better bound for the lower tail in (2.63) by using [22, Lemma 25].

Lemma 2.3.3 ([22]). *Let $\chi = \chi_2(d)$ be a chi-squared distributed variable with d degrees of freedom.*

Then $\forall x > 1$

$$\mathbb{P}\left[\chi \leq \frac{d}{x}\right] \leq e^{-\frac{d}{2}\left(\ln x + \frac{1}{x} - 1\right)} \tag{2.64}$$

Therefore, from (2.51), (2.59), (2.60) and (2.62), we have

$$p_2 \leq \sum_{t,S,K_1} \mathbb{E}\left[e^{-n' f_n(U^1)} 1[U^1 > \delta_1]\right] + \mathbb{P}\Big[\bigcup_{t,S,K_1}\Big\{U^1 \leq \delta_1\Big\}\Big] \tag{2.65}$$

where f_n is given by

$$\begin{aligned} f_n(x) &= x + 1 + \frac{2V_{n,t}}{1 - V_{n,t}}(1 + x) \\ &- \sqrt{1 + \frac{2V_{n,t}}{1 - V_{n,t}}(1 + x)}\sqrt{2x + 1 + \frac{2V_{n,t}}{1 - V_{n,t}}(1 + x)}. \end{aligned} \tag{2.66}$$

Next, we have the following claim.

Claim 4. *For* $0 < V_{n,t} < 1$ *and* $x > 0$, $f_n(x)$ *is a monotonically increasing function of* x.

Proof. See appendix 2.4.2. □

From this claim, we get

$$p_2 \leq \sum_{t,S,K_1} e^{-n' f_n(\delta_1)} + p_3 \tag{2.67}$$

where $p_3 = \mathbb{P}\left[E_2^c\right]$.

Now, if, for each t, δ_1 is chosen such that $f_n(\delta_1) > \frac{Kh(\frac{K-K_1+t}{K})}{n-K_1+t}$, then the first term in (2.144) goes to 0 as $n \to \infty$. Therefore, for $t = \theta K_1$, setting c_θ and q_θ as in (2.23) and (2.24) respectively, and choosing δ_1 such that

$$\delta_1 > \delta_1^* \tag{2.68}$$

with δ_1^* given by (2.25), will ensure that the first term in (2.67) goes to 0 as $n \to \infty$.

Note that

$$p_3 = \mathbb{P}\left[E_2^c\right] = \mathbb{P}\Big[\bigcup_{t,S,K_1}\Big\{V_{n,t}W_S - 1 \leq \delta_1(1 - V_{n,t})\Big\}\Big]. \tag{2.69}$$

Conditional on $H_{[K]}$,

$$\left\|P_{A_1}^{\perp} \sum_{i \in S_2^*} H_i c_i\right\|^2 \sim \frac{1}{2} P' \sum_{i \in S_2^*} |H_i|^2 \chi_2^{S_2^*}(2n')$$

where $\chi_2(2n')$ is a chi-squared distributed random variable with $2n'$ degrees of freedom (here the superscript S_2^* denotes the fact that this random variable depends on the codewords corresponding to S_2^*). For $1 > \delta_2 > 0$, consider the event $E_4 = \bigcap_{t,S,K_1} \left\{ \frac{\chi_2^{S_2^*}(2n')}{2n'} > 1 - \delta_2 \right\}$. Using (2.64) , we can bound p_3 as

$$p_3 \leq \sum_t \binom{K}{K - K_1 + t} e^{-n'(-\ln(1-\delta_2)-\delta_2)} + p_4 \tag{2.70}$$

where

$$\begin{aligned} p_4 &= \mathbb{P}\left[E_4^c\right] \\ &= \mathbb{P}\Bigg[\bigcup_{t,S,K_1} \left\{ V_{n,t} \left(1 + \frac{P' \sum_{i \in S_2^*} |H_i|^2 (1-\delta_2)}{\left(1 + P' \sum_{i \in S \setminus S_2^*} |H_i|^2\right)} \right) \leq 1 + \delta_1 (1 - V_{n,t}) \right\} \Bigg]. \end{aligned} \tag{2.71}$$

Again, it is enough to choose δ_2 such that

$$\delta_2 > \delta_2^* \tag{2.72}$$

with δ_2^* given by (2.26), to make sure that the first term in (2.70) goes to 0 as $n \to \infty$.

Note that the union bound over S is the minimum over S, and this minimizing S should be contiguous amongst the indices arranged according the decreasing order of fading powers. Further, S_2^* is chosen to be corresponding to the top t fading powers in S. Hence, we get

$$p_4 = \mathbb{P}\Bigg[\bigcup_t \Bigg\{ \min_{0 \leq j \leq K_1 - t} \left(\frac{P' \sum_{i=j+1}^{j+t} |H_{(i)}|^2 (1-\delta_2)}{1 + P' \sum_{i=j+t+1}^{j+t+K-K_1} |H_{(i)}|^2} \right) \leq \frac{1 + \delta_1 (1 - V_{n,t})}{V_{n,t}} - 1 \Bigg\} \Bigg]. \tag{2.73}$$

We make the following claim

Claim 5.

$$\begin{aligned} &\limsup_{n \to \infty} p_4 \leq \\ &1 \Bigg[\bigcup_{\theta \in (\frac{\epsilon'}{\nu}, 1] \cap \mathbb{Q}} \Bigg\{ \inf_{\xi \in [0, \nu(1-\theta)]} \left(\frac{(1-\delta_2) P'_{tot} \alpha(\xi, \xi + \nu\theta)}{1 + P'_{tot} \alpha(\xi + \nu\theta, \xi + 1 - \nu(1-\theta))} \right) \leq \frac{1 + \delta_1 (1 - V_\theta)}{V_\theta} - 1 \Bigg\} \Bigg] \end{aligned} \tag{2.74}$$

where $\alpha(a, b)$ is given by (2.27).

Proof. We have $|H_1|^2, ..., |H_K|^2$ with CDF $F(x) = (1 - e^{-x}) 1[x >= 0]$. Let $\tilde{F}_K(x) = \frac{1}{K} \sum_{i=1}^K 1[|H_i|^2 \leq$

$x]$ be the empirical CDF (ECDF). Then standard Chernoff bound gives, for $0 < r < 1$,

$$\mathbb{P}\left[|\tilde{F}_K(x) - F(x)| > rF(x)\right] \leq 2e^{-KcF(x)r^2} \tag{2.75}$$

where c is some constant.

From [147], we have the following representation. Let $0 < \gamma < 1$. Then

$$|H_{(\lceil n\gamma \rceil)}|^2 = F^{-1}(1-\gamma) - \frac{\tilde{F}_K(F^{-1}(1-\gamma)) - (1-\gamma)}{f\left(F^{-1}(1-\gamma)\right)} + R_K \tag{2.76}$$

where f is the pdf corresponding to F, and with probability 1, we have $R_K = O(n^{-3/4}\log(n))$ as $n \to \infty$.

Let $\tau > 0$. Then using (2.75) and (2.76), we have

$$\left||H_{(\lceil n\gamma \rceil)}|^2 - F^{-1}(1-\gamma)\right| \leq O\left(\frac{1}{n^{\frac{1-\tau}{2}}}\right) \tag{2.77}$$

with probability atleast $1 - e^{-O(n^\tau)}$.

Hence, for $0 < \xi < \zeta < 1$, we have, with probability $1 - e^{-O(n^\tau)}$,

$$\frac{1}{K}\sum_{i=\lceil \alpha K \rceil}^{\lceil \beta K \rceil} |H_{(i)}|^2 = \left[\frac{1}{K}\sum_{i=1}^{K} |H_i|^2 1\left[b \leq |H_i|^2 \leq a\right]\right] + o(1) \tag{2.78}$$

where $a = F^{-1}(1-\xi)$ and $b = F^{-1}(1-\zeta)$. Now, by law of large numbers (and Bernstein's inequality [148]), with overwhelming probability (exponentially close to 1), we have

$$\frac{1}{K}\sum_{i=1}^{K} |H_i|^2 1\left[b \leq |H_i|^2 \leq a\right] = \int_b^a x dF(x) + o(1) \tag{2.79}$$

and $\int_b^a x dF(x) = \int_\xi^\zeta F^{-1}(1-\gamma)d\gamma = \alpha(\xi, \zeta)$.

Define the events

$$J_{n,\theta,\xi} = \left\{ \left(\frac{P' \sum_{i=\lceil \xi K \rceil + 1}^{\lceil (\xi+\nu\theta)K \rceil} |H_{(i)}|^2 (1-\delta_2)}{1 + P' \sum_{i=\lceil (\xi+\nu\theta)K \rceil + 1}^{\lceil (\xi+1-\nu(1-\theta))K \rceil} |H_{(i)}|^2} \right) \leq \frac{1 + \delta_1 (1 - V_{n,\lceil \theta\nu K \rceil})}{V_{n,\lceil \theta\nu K \rceil}} - 1 \right\} \tag{2.80}$$

$$I_{n,\theta,\xi} = \left\{ \left(\frac{(1-\delta_2) P'_{tot} \alpha(\xi, \xi+\nu\theta)}{1 + P'_{tot} \alpha(\xi+\nu\theta, \xi+1-\nu(1-\theta))} \right) \leq \frac{1 + \delta_1 (1 - V_{n,\lceil \theta\nu K \rceil})}{V_{n,\lceil \theta\nu K \rceil}} - 1 \right\} \tag{2.81}$$

$$I_{\theta,\xi} = \left\{ \left(\frac{(1-\delta_2) P'_{tot} \alpha(\xi, \xi+\nu\theta)}{1 + P'_{tot} \alpha(\xi+\nu\theta, \xi+1-\nu(1-\theta))} \right) \leq \frac{1 + \delta_1 (1 - V_{\theta})}{V_{\theta}} - 1 \right\} \tag{2.82}$$

$$\begin{aligned} E_{n,\theta,\xi} = & \left\{ \left| \frac{1}{K} \sum_{i=\lceil \xi K \rceil + 1}^{\lceil (\xi+\nu\theta)K \rceil} |H_{(i)}|^2 - \alpha\left(\xi, \xi+\nu\theta\right) \leq o(1) \right| \right\} \\ & \bigcap \left\{ \left| \frac{1}{K} \sum_{i=\lceil (\xi+\nu\theta)K \rceil + 1}^{\lceil (\xi+1-\nu(1-\theta))K \rceil} |H_{(i)}|^2 - \alpha\left(\xi+\nu\theta, \xi+1-\nu(1-\theta)\right) \leq o(1) \right| \right\} \end{aligned} \tag{2.83}$$

$$E_n = \left(\bigcap_{\theta \in A_n} \bigcap_{\xi \in B_{K,\theta}} E_{n,\theta,\xi} \right) \tag{2.84}$$

where $A_n = \left(\frac{\epsilon'}{\nu}, 1 \right] \cap \left\{ \frac{i}{K_1} : i \in [K_1] \right\}$ and $B_{K,\theta} = [0, \nu(1-\theta)] \cap \left\{ \frac{i}{K} : i \in [K] \right\}$. Note that, from (2.78) and (2.79), $\mathbb{P}\left[E^c_{n,\theta,\xi} \right]$ is exponentially small in n.

Then we have

$$\begin{aligned} p_4 &= \mathbb{P}\left[\bigcup_{\theta \in A_n} \bigcup_{\xi \in B_{K,\theta}} J_{n,\theta,\xi} \right] \\ &\leq \mathbb{P}\left[\bigcup_{\theta \in A_n} \bigcup_{\xi \in B_{K,\theta}} J_{n,\theta,\xi} \cap E_{n,\theta,\xi} \right] + \sum_{\theta \in A_n} \sum_{\xi \in B_{K,\theta}} \mathbb{P}\left[E^c_{n,\theta,\xi} \right] \\ &\leq 1\left[\bigcup_{\theta \in A_n} \bigcup_{\xi \in B_{K,\theta}} I_{n,\theta,xi} \right] + o(1) \\ &\leq 1\left[\bigcup_{\theta \in (\frac{\epsilon'}{\nu}, 1]} \bigcup_{\xi \in [0, \nu(1-\theta)]} I_{n,\theta,\xi} \right] + o(1). \end{aligned} \tag{2.85}$$

Therefore

$$\limsup_{n \to \infty} p_4 \leq 1\left[\bigcup_{\theta \in (\frac{\epsilon'}{\nu}, 1]} \bigcup_{\xi \in [0, \nu(1-\theta)]} I_{\theta,\xi} \right] \tag{2.86}$$

This concludes the proof of claim 5.

□

The statement of the theorem follows by choosing P'_{tot} to make sure that $\limsup_{n \to \infty} p_4 = 0$.

Remark 4. *In retrospect, our analysis is rather similar to the one in [22]. We remind that the problem considered there can be seen (as argued in [3]) as a version of the many-MAC problem with random-access, cf. Section 2.1.1 for more.*

□

2.3.2 No-CSI: Scalar AMP with i.i.d Gaussian codebook

In this section, we will given an achievability bound on E_b/N_0 for the no-CSI case by the asymptotic analysis of the scalar AMP algorithm [22, 25, 28, 142]. Here, we recall the compressed sensing view of our model (2.6) where U is block sparse. As discussed in section 2.1.1, a better algorithm to use in this case would be the vector or block version of AMP, whose analysis is also well studied, e.g. [142]. However, as we discussed in Section 2.1.1 evaluation of performance of this block-AMP requires computing $M = 2^{100}$ dimensional integrals, and thus does not result in computable bounds. Instead, here we take a different approach by analyzing the scalar AMP algorithm, whose asymptotic analysis in [28] in fact only requires that the empirical distribution of entries of U be convergent – a fact emphasized in [22]. Let us restate the signal model we have:

$$Y = AU + Z, \qquad A_{i,j} \overset{iid}{\sim} \mathcal{CN}(0, E/n), i \in [n], j \in [KM]\,, \tag{2.87}$$

where $E = \frac{P_{tot}}{\mu}$ is the total energy of each codeword, $U \in \mathbb{C}^{KM}$ is block sparse with $K = \mu n$ blocks each of length M, with a single non-zero entry U_j in each block with $U_j \sim \mathcal{CN}(0,1)$ (Rayleigh fading), and $Z \sim \mathcal{CN}(0, I_n)$. The support of U, denoted by $S \in \{0,1\}^{KM}$, is sampled uniformly from all such block sparse supports (there are M^K of them). The goal is to get an estimate $\hat{S} = \hat{S}(Y, A)$ of S where our figure of merit is the following:

$$\mathrm{PUPE}(\hat{S}) = \frac{1}{K}\sum_{k=1}^{K} \mathbb{P}\left[S_{1+(k-1)M}^{kM} \neq \hat{S}_{1+(k-1)M}^{kM}\right]\,, \tag{2.88}$$

which is also known as section error rate (SER) in the SPARC literature [139].

The AMP-based algorithm operates as follows. First we estimate U iteratively, then after estimating U we threshold its values to obtain an estimator for S.

To describe scalar AMP we first introduce the following scalar problem. For each $\sigma > 0$ define $\mu^{(\sigma)} = P_{X,V}$ to be the joint distribution of variables X and V:

$$V = X + \sigma W, \qquad X \perp W \sim \mathcal{CN}(0,1) \tag{2.89}$$

and

$$X \sim \mathrm{BG}(1, 1/M) = \begin{cases} \mathcal{CN}(0,1) & \text{w.p. } \frac{1}{M} \\ 0 & \text{w.p. } 1 - \frac{1}{M} \end{cases} \tag{2.90}$$

We also define

$$\eta(z, \sigma^2) \triangleq \mathbb{E}[X|V = z]\,, \mathrm{mmse}(\sigma^2) \triangleq \mathbb{E}[(X - \mathbb{E}[X|V])^2]\,. \tag{2.91}$$

Next, start with $U^{(0)} = 0 \in \mathbb{C}^{KM}$, $R^{(0)} = Y$, $\hat{\sigma}_0^2 = \frac{1}{E} + \mu$. Then for $t = 1, 2, \cdots$ we have the following iterations

$$U^{(t)} = \eta\left(A^* R^{(t-1)} + U^{(t-1)}, \hat{\sigma}_{t-1}^2\right) \tag{2.92}$$

$$R^{(t)} = Y - AU^{(t)} + \mu M R^{(t-1)} \frac{1}{KM} \sum_{i=1}^{KM} \eta'\left(\left(A^* R^{(t-1)} + U^{(t-1)}\right)_i, \hat{\sigma}_{t-1}^2\right) \tag{2.93}$$

$$\hat{\sigma}_t^2 = \frac{1}{n} \left\| R^{(t)} \right\|^2 \tag{2.94}$$

where $\eta'(x + iy, \sigma^2)$ denotes $\frac{1}{2}\left(\frac{\partial \eta(x+iy,\sigma^2)}{\partial x} - i\frac{\partial \eta(x+iy,\sigma^2)}{\partial y}\right)$ and $i = \sqrt{-1}$ is the imaginary unit (see [149, 150] for a more general derivation of complex AMP). The estimate of U after t steps is given by (see [22] for more details)

$$\hat{U}^{(t)} = A^* R^{(t)} + U^{(t)} \tag{2.95}$$

To convert $\hat{U}^{(t)}$ into $\hat{S}^{(t)}$ we perform a simple thresholding:

$$\hat{S}^{(t)}(\theta) = \{i \in [KM] : |\hat{U}_i^{(t)}|^2 > \theta\}\,. \tag{2.96}$$

Theorem 2.3.4 (Scalar AMP achievability)**.** *Fix any $\mu > 0$, $P_{tot} > 0$ and $M \geq 1$. Then for every $\mathcal{E} > \frac{E}{\log_2 M} = \frac{P_{tot}}{\mu \log_2 M}$ there exist a sequence of $(n, M, \epsilon_n, \mathcal{E}, K = \mu n)$ codes (noCSI) such that AMP decoder* (2.96) *(with a carefully chosen $\theta = \theta(\mathcal{E}, M, \mu)$ and sufficiently large t) achieves*

$$\limsup_{n\to\infty} \epsilon_n \leq \pi^*(\sigma_\infty^2, M)\,,$$

where $\pi^*(\tau, M) = 1 - \frac{1}{1+\tau}\left((M-1)\left(\frac{1}{\tau} + 1\right)\right)^{-\tau}$ and σ_∞^2 is found from

$$\begin{aligned} \sigma_\infty^2 &\equiv \sigma_\infty^2(\mu, E, M) \\ &= \sup\left\{\tau \geq 0 : \tau = \frac{1}{E} + \mu M \,\mathrm{mmse}(\tau)\right\} \end{aligned} \tag{2.97}$$

Proof. Denote the Hamming distance

$$d_H(S,\hat{S}) = \frac{1}{KM}\sum_{i=1}^{KM} 1[S_i \neq \hat{S}_i] \tag{2.98}$$

Note that according to the definition (2.88) we have a bound

$$\mathrm{PUPE}(\hat{S}^{(t)}(\theta)) \leq M\mathbb{E}\left[d_H(S,\hat{S}^{(t)}(\theta)))\right] \tag{2.99}$$

Indeed, this is a simple consequence of upper bounding each probability in (2.88) by the union bound.

The key result of [28] shows the following. Let the empirical joint distribution of entries in $(U,\hat{U}^{(t)})$ be denoted by

$$\hat{\mu}_{U,\hat{U}^{(t)}} \triangleq \frac{1}{KM}\sum_{i=1}^{KM} \delta_{(U_i,\hat{U}_i^{(t)})},$$

where δ_x is the Dirac measure at x. Then as $n\to\infty$ this (random) distribution on $\mathbb{C}^2$ converges weakly to a deterministic limit $\mu_{X,V}$ almost surely. More precisely, from [28, Lemma 1(b)] and proof of [22, Theorem 5] for any bounded Lipschitz continuous function $f:\mathbb{C}^2\to\mathbb{R}$ we have

$$\lim_{n\to\infty}\int f\, d\hat{\mu}_{U,\hat{U}^{(t)}} = \int f\, d\mu^{(\sigma_t)} \text{ a.s.} \tag{2.100}$$

where $\mu^{(\sigma_t)}$ is the joint distribution of $(X, V = X+\sigma_t W)$ defined in (2.89), and σ_t can further be determined from the so called state evolution sequence: Set $\sigma_0^2 = \frac{1}{E}+\mu$ and then

$$\sigma_t^2 = \frac{1}{E} + \mu M\,\mathrm{mmse}(\sigma_{t-1}^2) \tag{2.101}$$

where mmse is defined in (2.91).

Note that the assumptions on U A and Z in [28, Lemma 1(b)] hold in our case. In particular, since the support of U is sampled uniformly from all block sparse supports of size K and the entries in the support are iid $\mathcal{CN}(0,1)$ random variables, we have that the empirical distribution of entries of U converge weakly almost surely to the distribution P_X of X defined in (2.90). Further the moment conditions in [28, Theorem 2] are also satisfied. We note here that although [22, 28] consider only real valued signals, the results there also hold for the complex case (see [151, Theorem III.15], [152, Chapter 7]).

We next consider the support recovery in the scalar model (2.89). Let $S_0 = 1[X\neq 0]$ denote the indicator of the event when X is non-zero. Let $\hat{S}_0 \equiv \hat{S}_0(\theta) = 1[|V|^2>\theta]$ denote an estimator of S_0 using the observation $V = X+\sigma W$ in (2.89). Let

$$\psi(\sigma^2,\theta,M) = \mathbb{P}\left[S_0\neq\hat{S}_0\right] \tag{2.102}$$

denote the probability of error in the scalar model (2.89) with σ dependence made explicit as an argument of ψ. The from the convergence of $\hat{\mu}_{U,\hat{U}^{(t)}}$ we conclude as in [22] that for any number t of steps of the AMP algorithm $\hat{S}^{(t)}(\theta)$ achieves

$$\lim_{n\to\infty} \mathrm{PUPE}(\hat{S}^{(t)}(\theta)) \le M\psi(\sigma_t^2, \theta, M)\,. \tag{2.103}$$

Since this holds for any t and any θ we can optimize both by taking $t \to \infty$ and $\inf_{\theta>0}$. From the proof of [22, Theorem 6] it follows that $\lim_{t\to\infty} \sigma_t^2 = \sigma_\infty^2$ exists and σ_∞ satisfies (2.97). The proof is completed by the application of the following Claim, which allows us to compute infimum over θ in closed form. □

Claim 6.

$$M \inf_\theta \psi(\tau, \theta, M) = 1 - \frac{1}{1+\tau} \frac{1}{\left((M-1)\left(\frac{1}{\tau}+1\right)\right)^\tau} \tag{2.104}$$

Proof. Let us define $\tau = \sigma^2$. We have

$$\begin{aligned}\psi(\tau, \theta, M) &= \mathbb{P}\left[S_0 \neq \hat{S}_0\right] \\ &= \frac{1}{M}\mathbb{P}\left[\hat{S}_0 = 0 | S_0 = 1\right] + \\ &\quad \left(1 - \frac{1}{M}\right)\mathbb{P}\left[\hat{S}_0 = 1 | S_0 = 0\right]\end{aligned}$$

Now conditioned on $S_0 = 1$, $|V|^2 \sim (1+\tau)\mathrm{Exp}(1)$ and conditioned on $S_0 = 0$, $|V|^2 \sim \tau\mathrm{Exp}(1)$ where $\mathrm{Exp}(1)$ is the Exponential distribution with density function $p(x) = e^{-x}1[x \ge 0]$. Hence

$$\psi(\tau, \theta, M) = \frac{1}{M}\left(1 - e^{-\frac{\theta}{1+\tau}}\right) + \left(1 - \frac{1}{M}\right)e^{-\frac{\theta}{\tau}} \tag{2.105}$$

The claim follows by optimizing (2.105) over θ. The optimum occurs at

$$\theta^* = \tau(1+\tau)\ln\left(\frac{1+\tau}{\tau}(M-1)\right)$$

Substituting θ^* in (2.105) proves the claim. □

2.3.3 No-CSI: Scalar AMP with spatially coupled codebook

Previously, we presented an achievability bound using the fixed point of the state evolution of the AMP algorithm that is designed for i.i.d Gaussian codebooks. This is known to be suboptimal [152], and this is also evident from the figure 4-1. Instead, we can leverage recent rigorous results on the optimality of AMP with spatially coupled codebook design to get an almost optimal achievability

bound. In particular, we build on the spatial coupling idea from [27] and the scalar AMP from section 2.3.2 to provide new achievability bound for the many-user quasi-static fading MAC.

Spatially coupled codebook

Now we describe the spatially coupled codebook design based on [27]. Let $R, C \in \mathbb{N}$ be such that R divides n and C divides $p = KM$. The codebook A is divided into blocks of size $\frac{n}{R} \times \frac{p}{C}$ and hence can be considered as a block matrix of size $R \times C$. Let $B \in \mathbb{R}^{R \times C}$ be the base matrix with nonnegative entries $B_{r,c}$ such that $\sum_{r=1}^{R} B_{r,c} = 1$ for all $c \in [C]$. Further, with abuse of notation, let $r : [n] \to [R]$ and $c : p \to [C]$ denote functions that map a particular row or column index to its corresponding block. Let $E = \frac{P_{tot}}{\mu}$ denote the total energy. Then the matrix A is constructed as $A_{i,j} \sim \mathcal{CN}\left(0, E\frac{R}{n}B_{r(i),c(j)}\right)$ with $\{A_{i,j}\}$ independent across tuples (i, j). In particular we use the (ω, Λ, ρ) base matrix from [27] as the choice of B. This is as follows. Let $\rho \in [0, 1)$, $\omega \geq 1$ and $\Lambda \geq 2\omega - 1$. Then we choose $R = \Lambda + \omega - 1$ and $C = \Lambda$. Finally we have

$$B_{r,c} = \begin{cases} \frac{1-\rho}{\omega}, & c \leq r \leq c + \omega - 1 \\ \frac{\rho}{\Lambda - 1}, & \text{o/w} \end{cases} \tag{2.106}$$

Let $\tilde{\mu} = \frac{R}{C}\mu$ be the effective user density. Since usually $\omega > 1$ we have that $\tilde{\mu} > \mu$ and thus the effective user density is higher in such spatially coupled systems [27].

Next we present the AMP algorithm adapted to the spatially coupled codebook (SC-AMP) from [27] (see also [26, 153]; refer to [154, Sec 4.4] for extension to the complex case). In particular the section size (denoted as B in [27]) is set to 1 in the main SC-AMP algorithm in [27]. But the main difference from [27] is that we ignore the block sparse structure of U and just use the fact that the empirical distribution of entries in U converge to $\mathrm{BG}(1, 1/M)$ defined in (2.90). This is sufficient for the state evolution to be valid [153], and similar to the results in previous section, the state evolution is defined on the scalar channel (2.89) (instead of the vector channel in [27] which would be infeasible for numerical evaluation when M is very large but finite). The scalar equivalent channel and related quantities are restated here for convenience.

$$V_{\sigma^2} = X + \sigma W \tag{2.107}$$

where X independent of W, and $X \sim \mathrm{BG}(1, 1/M)$, $W \sim \mathcal{CN}(0, 1)$. We denote the joint distribution of X and V_{σ^2} by $P_{X,V_{\sigma^2}}$. The scalar denoising function is

$$\eta(v, \sigma^2) = \mathbb{E}\left[X | V_{\sigma^2} = v\right] = \mathbb{E}\left[X | X + \sigma W = v\right]. \tag{2.108}$$

The minimum mean squared error of estimating X from V is given by

$$\text{mmse}(\sigma^2) = \mathbb{E}\left[(X - \eta(V_{\sigma^2}, \sigma^2))^2\right] \tag{2.109}$$

Lastly we define the equivalent of support recovery. Let $S_0 = 1[X \neq 0]$. Let $\hat{S}_0(\theta)$ be an estimate of S_0 based on observation V_{σ^2}. In particular we use the following estimator $\hat{S}_0(\theta) = 1[|V_{\sigma^2}|^2 > \theta]$. Then we denote the probability of error in support recovery by ψ:

$$\psi(\sigma^2, \theta, M) = \mathbb{P}\left[S_0 \neq \hat{S}_0(\theta)\right] \tag{2.110}$$

AMP algorithm: Start with $U^{(0)} = 0 \in \mathbb{C}^{KM}$, $R^{(0)} = Y$. Then for $t = 1, 2, \cdots$ we have the following iterations

$$U^{(t)} = \eta^{(t)}\left((\tilde{Q}^{(t-1)} \odot A)^* R^{(t-1)} + U^{(t-1)}\right) \tag{2.111}$$

$$R^{(t)} = Y - AU^{(t)} + \frac{R}{C}\mu M(\tilde{b}^{(t)} \odot R^{(t-1)}) \tag{2.112}$$

where $\odot$ denotes element wise product, and matrix $\tilde{Q}^{(t)}$, vector $\tilde{b}^{(t)}$ and denoiser $\eta^{(t)} : \mathbb{C}^{KM} \to \mathbb{C}^{KM}$ will be defined next via the state evolution.

Let $\psi_c^{(0)} = \infty$. Then for $t \geq 1$, for each $r \in [R]$ and $c \in [C]$ we define

$$\gamma_r^{(t)} = \sum_{c=1}^{C} B_{r,c}\psi_c^{(t)} \tag{2.113}$$

$$\phi_r^{(t)} = \frac{1}{E} + \tilde{\mu}M\gamma_r^{(t)} \tag{2.114}$$

$$\tau_c^{(t)} = \frac{1}{\sum_{r=1}^{R} B_{r,c}\left(\phi_r^{(t)}\right)^{-1}} \tag{2.115}$$

$$\psi_c^{(t+1)} = \text{mmse}(\tau_c^{(t)}) \tag{2.116}$$

where $\text{mmse}(\cdot)$ is defined in (2.109).

Now the matrices $\tilde{Q}^{(t)}$ and vectors $\tilde{b}^{(t)}$ are defined as follows. For each $i \in [n]$ and $j \in [KM]$

$$\tilde{b}_i^{(t)} = \tilde{\mu}M\frac{\gamma_{r(i)}^{(t)}}{\phi_{r(i)}^{(t-1)}}$$

$$\tilde{Q}_{i,j}^{(t)} = \frac{\tau_{c(j)}^{(t)}}{\phi_{r(i)}^{(t)}}$$

The denoiser at time t is given by $\eta^{(t)} = (\eta_1^{(t)}, \cdots, \eta_{KM}^{(t)})$ with

$$\eta_i^{(t)}(z) = \mathbb{E}\left[X | X + \sqrt{\tau_{c(i)}^{(t)}}W = z\right] = \eta(z, \tau_{c(i)}^{(t)}) \tag{2.117}$$

where $\eta(\cdot,\cdot)$ was defined in (2.108).

The estimate of U after t steps is given by (see [22, 27] for more details on hard decision estimate)

$$\hat{U}^{(t)} = (\tilde{Q}^t \odot A)^* R^{(t)} + U^{(t)} \tag{2.118}$$

To convert $\hat{U}^{(t)}$ into $\hat{S}^{(t)}$ we perform a simple thresholding for each $c \in [C]$ i.e., for each i

$$\hat{S}_i^{(t)}(\theta_{c(i)}) = 1[|\hat{U}_i^{(t)}|^2 > \theta_{c(i)}] \tag{2.119}$$

where $\{\theta_c : c \in [C]\}$ is a set of thresholds.

We have the following lemma that follows directly from [27, Theorem 2] (with $B = 1$ in their notation which in turn relies on [26, 153]). Define the replica potential

$$\mathcal{F}(\tau; \mu, E, M) = (\mu M) I(X; V_\tau) + \left(\ln \tau + \frac{1}{\tau E} - 1 \right) \tag{2.120}$$

where $(X, V_\tau) \sim P_{X,V_\tau}$. Further, let $\mathcal{M}$ denote the maximum of the global minimizers of $\mathcal{F}$:

$$\mathcal{M}(\mu, E, M) = \max(\arg\min_{\tau > \frac{1}{E}} \mathcal{F}(\tau; \mu, E, M)) \tag{2.121}$$

Lemma 2.3.5 ([27])**.** *For any (ω, Λ, ρ) base matrix B, for each $c \in [C]$, $\tau_c^{(t)}$ is non-increasing in t and converges to a fixed point τ_c^∞. Furthermore, for any $\delta > 0$, there exists $\omega_0 < \infty$, $\Lambda_0 < \infty$ and $\rho_0 > 0$ such that for all $\omega > \omega_0$, $\Lambda > \Lambda_0$ and $\rho < \rho_0$, the fixed points $\{\tau_c^\infty : c \in [C]\}$ satisfy*

$$\tau_c^\infty \leq \tau^\infty(\tilde{\mu}) + \tilde{\mu} M \delta \tag{2.122}$$

where $\tau^{(\infty)}(\tilde{\mu}) = \mathcal{M}(\tilde{\mu}, E, M)$.

Notice that $\tau_c^{(t)}$ tracks the noise variance (and hence also mmse) of estimation in the scalar channel (2.107). Thus the above lemma says that the fixed points of the spatially coupled system are at least as good as the uncoupled system (i.e., with A having i.i.d entries as in the previous section) but with user density increased from μ to $\tilde{\mu} = \left(1 + \frac{\omega - 1}{\Lambda}\right) \mu$. Notice that if take limits as $\Lambda \to \infty$ *and then* $\omega \to \infty$ we obtain that $\tau^\infty(\tilde{\mu}) \to \tau^\infty(\mu)$. This known in the literature as threshold saturation (see [27, Remark 3.3]).

Finally we have the main achievability bound based on spatial coupling

Theorem 2.3.6. *Fix any $\mu > 0$, $E > 0$ and $k = \log_2 M \geq 1$. Then for every $\mathcal{E} > \frac{E}{k}$ there exist a sequence of $(n, M, \epsilon_n, \mathcal{E}, K = \mu n)$ codes for the quasi-static fading MAC such that*

$$\limsup_{n \to \infty} \epsilon_n \leq \pi^*(\tau^{(\infty)}(\mu), M) \tag{2.123}$$

where $\pi^*(\tau, M) = 1 - \frac{1}{1+\tau}\left((M-1)\left(\frac{1}{\tau}+1\right)\right)^{-\tau}$ *and*

$$\tau^{(\infty)}(\mu) \equiv \tau^{(\infty)}(\mu; E, M) = \mathcal{M}(\mu, E, M) \tag{2.124}$$

Proof. The idea is to use random coding along with the spatially coupled codebook described in the beginning of this section. The proof is similar to that of theorem 2.3.4 but uses the result on convergence of the empirical joint distribution of entries in $(U, \hat{U}^{(t)})$ in the spatially coupled systems from [26, 153] (adapted to the complex number setting). Recall that if S is the support of the true signal U, and $\hat{S}^{(t)} \equiv (\hat{S}_i^{(t)}(\theta_{c(i)}))_{i=1}^{KM}$ (see (2.119)) is the estimate of the support, then from (2.99) we have that

$$\mathrm{PUPE}(\hat{S}^{(t)}) \leq M\mathbb{E}\left[d_H(S, \hat{S}^{(t)}))\right] \tag{2.125}$$

Notice that

$$d_H(S, \hat{S}^{(t)}) = \frac{1}{C}\sum_{c=1}^{C}\left[\frac{C}{KM}\sum_{i=(c-1)\frac{KM}{C}+1}^{c\frac{KM}{C}} 1[S_i \neq \hat{S}_i^{(t)}]\right]$$

Moreover, from [26, Theorem 1] (see proof of lemma 1 there) we have that for any Lipschitz function $f : \mathbb{C}^2 \to \mathbb{R}$ (or more generally any pseudo-Lipschitz function [28]) the following holds almost surely (with $K = \mu n$):

$$\lim_{n\to\infty}\frac{C}{KM}\sum_{i=(c-1)\frac{KM}{C}+1}^{c\frac{KM}{C}} f(U_i, \hat{U}_i^{(t)}) = \mathbb{E}\left[f(X, V_{\tau_c^{(t)}})\right] \tag{2.126}$$

where $X \sim \mathrm{BG}(1, 1/M)$ and $V_{\tau(t)_c}$ is the the output of the scalar channel (2.89) with dependence on $\sigma^2 = \tau_c^{(t)}$ made explicit for convenience:

$$V_{\tau_c^{(t)}} = X + \sqrt{\tau_c^{(t)}}W, \quad X \perp V_{\tau_c^{(t)}}$$

Remark 5. *We note here that [26] deal only with real valued system. But as noted in [151, Theorem III.15] and [154, Sec 4.4], the proofs in [26] go through for complex valued systems as well.*

By a standard approximation argument as in proof of [27, Theorem 1 (3)] we get that

$$\lim_{n\to\infty}\frac{C}{KM}\sum_{i=(c-1)\frac{KM}{C}+1}^{c\frac{KM}{C}} \mathbb{P}\left[S_i \neq \hat{S}_i^{(t)}\right] = \psi(\tau_c^{(t)}, \theta_c, M) \tag{2.127}$$

where the function $\psi()$ is defined in (2.110) and $\{\theta_c : c \in [C]\}$ are thresholds (2.119).

Thus for any $\{\theta_c > 0 : c \in [C]\}$

$$\lim_{n\to\infty} \text{PUPE}(\hat{S}^{(t)}) \leq \frac{1}{C}\sum_{c=1}^{C} M\psi(\tau_c^{(t)}, \theta_c, M) \tag{2.128}$$

Now we take $t \to \infty$ and use lemma 2.3.5 to obtain

$$\lim_{t\to\infty}\lim_{n\to\infty} \text{PUPE}(\hat{S}^{(t)}) \leq \frac{1}{C}\sum_{c=1}^{C} M\psi(\tau_c^{(\infty)}, \theta_c, M) \tag{2.129}$$

Since $\{\theta_c\}$ are arbitrary we can minimize over $\{\theta_c > 0 : c \in [C]\}$ and use claim 6 to obtain

$$\lim_{t\to\infty}\lim_{n\to\infty} \text{PUPE}(\hat{S}^{(t)}) \leq \frac{1}{C}\sum_{c=1}^{C} \pi^*(\tau_c^{(\infty)}, M) \tag{2.130}$$

where

$$\pi^*(\tau, M) = \left[1 - \frac{1}{1+\tau}\frac{1}{\left((M-1)\left(\frac{1}{\tau}+1\right)\right)^{\tau}}\right] \tag{2.131}$$

Notice that π^* is non-decreasing in τ. Thus for any fixed $\delta > 0$, from second item in lemma 2.3.5 we have that for all large enough ω and Λ, and all small enough ρ:

$$\lim_{t\to\infty}\lim_{n\to\infty} \text{PUPE}(\hat{S}^{(t)}) \leq \pi^*(\tau^{(\infty)}(\tilde{\mu}) + \tilde{\mu}M\delta, M) \tag{2.132}$$

Lastly, we take limit as $\Lambda \to \infty$ and then $\omega \to \infty$ to obtain that for every $\delta > 0$ there is a $\rho_0 > 0$ such that for all $0 < \rho < \rho_0$ we have

$$\lim_{\omega\to\infty}\lim_{\Lambda\to\infty}\lim_{t\to\infty}\lim_{n\to\infty} \text{PUPE}(\hat{S}^{(t)}) \leq \pi^*(\tau^{(\infty)}(\mu) + \mu M\delta, M) \tag{2.133}$$

The theorem is proved by noticing that $\delta > 0$ is arbitrary.

□

2.3.4 CSIR

In this subsection, we focus on the CSIR scenario. We could use projection decoding to decode a fraction of users where decoding set is a function of CSIR. But a better bound is obtained by directly using euclidean metric to decode, similar to [3]. Then have the following theorem.

Theorem 2.3.7. *Consider the channel* (2.1) *(with CSIR) with* $K = \mu n$ *where* $\mu < 1$. *Fix the spectral efficiency* S *and target probability of error (per-user)* ϵ. *Let* $M = 2^{S/\mu}$ *denote the size of the codebooks and* $P_{tot} = KP$ *be the total power. Fix* $\nu \in (1-\epsilon, 1]$. *Let* $\epsilon' = \epsilon - (1-\nu)$. *Then if* $\mathcal{E} > \mathcal{E}^*_{CSIR} = \sup_{\frac{\epsilon'}{\nu} < \theta \leq 1} \inf_{0\leq\rho\leq 1} \frac{P_{tot,\nu}(\theta,\rho)}{S}$, *there exists a sequence of* $(n, M, \epsilon_n, \mathcal{E}, K = \mu n)$ *codes such that* $\limsup_{n\to\infty} \epsilon_n \leq \epsilon$, *where, for* $\frac{\epsilon'}{\nu} < \theta \leq 1$,

$$P_{tot,\nu}(\theta,\rho)=\frac{(1+\rho)\left(e^{\mu\nu\left(\frac{h(\theta)}{\rho}+\theta\ln M\right)}-1\right)}{\alpha(\nu(1-\theta),\nu)-\left(e^{\mu\nu\left(\frac{h(\theta)}{\rho}+\theta\ln M\right)}-1\right)\alpha(\nu,1)(1+\rho)} \tag{2.134}$$

$$\alpha(a,b)=a\ln(a)-b\ln(b)+b-a. \tag{2.135}$$

Hence $\mathcal{E}^*\leq\mathcal{E}^*_{CSIR}$.

The proof idea is a combination of techniques similar to [3] and theorem 2.3.1

Proof. Let each user generate a Gaussian codebook of size M and power $P'<P$ independently such that $KP'=P'_{tot}<P_{tot}$. Let W_j denote the random (in $[M]$) message of user j. So, if $\mathcal{C}_j=\{c_i^j:i\in[M]\}$ is the codebook of user j, he/she transmits $X_j=c^j_{W_j}1\left\{\left\|c^j_{W_j}\right\|^2\leq nP\right\}$. For simplicity let $(c_1,c_2,...,c_K)$ be the sent codewords. Fix $\nu\in(1-\epsilon,1]$. Let $K_1=\nu K$ be the number of users that are decoded. Fix a decoding set $D\subset[K]$, possibly depending on $H_{[K]}$ such that $|D|=K_1,a.s.$ Since the receives knows $H_{[K]}$, we can use the euclidean distance used in [3] as the decoding metric. Formally, the decoder output $g_D(Y)\in\prod_{i=1}^K\mathcal{C}_i$ is given by

$$(g_D(Y))_i=\begin{cases}f_i^{-1}(\hat{c}_i) & i\in D\\ ? & i\notin D\end{cases}$$
$$(\hat{c}_i)_{i\in D}=\arg\min_{(c_i\in\mathcal{C}_i)_{i\in D}}\left\|Y-\sum_{i\in D}H_ic_i\right\|^2.$$

The probability of error is given by

$$P_e=\frac{1}{K}\sum_{j=1}^K\mathbb{P}\left[W_j\neq\hat{W}_j\right] \tag{2.136}$$

where $\hat{W}_j=(g(Y))_j$ is the decoded message of user j. Similar to the no-CSI case, we perform a change of measure to $X_j=c^j_{W_j}$ by adding a total variation distance bounded by $p_0=K\mathbb{P}\left[\frac{\chi_2(2n)}{2n}>\frac{P}{P'}\right]\to0$ as $n\to\infty$.

Let $\epsilon'=\epsilon-(1-\nu)$. Now we have

$$\begin{aligned}P_e&=\mathbb{E}\left[\frac{1}{K}\sum_{j=1}^K1\{W_j\neq\hat{W}_j\}\right]\\&=\frac{K-K_1}{K}+\mathbb{E}\left[\frac{1}{K}\sum_{j\in D}1\{W_j\neq\hat{W}_j\}\right]\end{aligned}$$

$$\leq (1-\nu) + \epsilon' + \nu \mathbb{P}\left[\frac{1}{K}\sum_{j\in D} 1\{W_j \neq \hat{W}_j\} \geq \epsilon'\right]$$
$$= \epsilon + \nu p_1 \tag{2.137}$$

where $p_1 = \mathbb{P}\left[\bigcup_{t=\epsilon' K}^{\nu K}\left\{\sum_{j\in D} 1\{W_j \neq \hat{W}_j\} = t\right\}\right]$.

From now on, we just write $\bigcup_t$ to denote $\bigcup_{t=\epsilon' K}^{\nu K}$, $\sum_t$ for $\sum_{t=\epsilon' K}^{\nu K}$, and $\sum_S$ for $\sum_{\substack{S\subset D\\|S|=t}}$. Let $c_{[S]} \equiv \{c_i : i \in [S]\}$ and $H_{[K]} = \{H_i : i \in [K]\}$. Let $F_t = \left\{\sum_{j\in D} 1\{W_j \neq \hat{W}_j\} = t\right\}$. Let $\rho \in [0,1]$. We bound $\mathbb{P}[F_t]$ using Gallager's rho trick similar to [3] as

$$\begin{aligned}
&\mathbb{P}\left[F_t | Z, c_{[K]}, H_{[K]}\right] \\
&\leq \mathbb{P}\Bigg[\exists S \subset D : |S| = t, \exists\{c_i' \in \mathcal{C}_i : i \in S, c_i' \neq c_i\} : \\
&\left\|Y - \sum_{i\in S} H_i c_i' - \sum_{i\in D\setminus S} H_i c_i\right\|^2 < \left\|Y - \sum_{i\in D} H_i c_i\right\|^2 \Bigg| Z, c_{[K]}, H_{[K]}\Bigg] \\
&\leq \sum_S \mathbb{P}\Bigg[\bigcup_{\substack{c_i'\in\mathcal{C}_i : i\in S\\ c_i'\neq c_i}} \left\{\left\|Z_D + \sum_{i\in S} H_i c_i - \sum_{i\in S} H_i c_i'\right\|^2 < \left\|Z_D\right\|^2\right\} \Bigg| Z, c_{[K]}, H_{[K]}\Bigg] \\
&\leq \sum_S M^{\rho t}\mathbb{P}\left[\left\|Z_D + \sum_{i\in S} H_i c_i - \sum_{i\in S} H_i c_i'\right\|^2 < \left\|Z_D\right\|^2 \Bigg| Z, c_{[K]}, H_{[K]}\right]^\rho
\end{aligned} \tag{2.138}$$

where $Z_D = Z + \sum_{i\in[K]\setminus D} H_i c_i$ and $c_{[S]}'$ in the last display denotes a generic set of unsent codewords corresponding to codebooks of users in set S.

We use the following simple lemma which is a trivial extension of a similar result used in [3] to compute the above probability.

Lemma 2.3.8. *Let $Z \sim \mathcal{CN}(0, I_n)$ and $u \in \mathbb{C}^n$. Let $D = diag(d_1, ..., d_n) \in \mathbb{C}^{n\times n}$ be a diagonal matrix. If $\gamma > \sup_{j\in[n]} -\frac{1}{|d_j|^2}$, then*

$$\mathbb{E}\left[e^{-\gamma\|DZ+u\|^2}\right] = \frac{1}{\prod_{j\in[n]}(1+\gamma|d_j|^2)} e^{-\gamma\sum_{j\in[n]}\frac{|u_j|^2}{1+\gamma|d_j|^2}}$$

Proof. Omitted

□

So, using the above lemma, we have, for $\lambda_1 > 0$,

$$\begin{aligned}
&\mathbb{E}_{\{c'_S\}}\left[\mathbb{P}\left[\left\|Z_D + \sum_{i\in S} H_i c_i - \sum_{i\in S} H_i c'_i\right\|^2 < \|Z_D\|^2 \middle| Z, c_{[K]}, H_{[K]}\right]^\rho\right] \\
&= \mathbb{E}_{\{c'_S\}}\left[\mathbb{P}\left[\exp\left(-\lambda_1\left\|Z_D + \sum_{i\in S} H_i c_i - \sum_{i\in S} H_i c'_i\right\|^2\right) > \right.\right. \\
&\quad \left.\left. \exp\left(-\lambda_1 \|Z_D\|^2\right) \middle| Z, c_{[K]}, H_{[K]}\right]^\rho\right] \\
&\leq \frac{e^{\rho\lambda_1\|Z_D\|^2}}{\left(1+\lambda_1 P' \sum_{i\in S}|H_i|^2\right)^{\rho n}} e^{\frac{-\rho\lambda_1\left\|Z_D+\sum_{i\in S} H_i c_i\right\|^2}{1+\lambda_1 P' \sum_{i\in S}|H_i|^2}}
\end{aligned} \tag{2.139}$$

where $\mathbb{E}_{c'_S}$ denotes taking expectation with respect to $\{c'_i : i \in S\}$ alone, and $1+\lambda_1 P' \sum_{i\in S}|H_i|^2 > 0$.

Let $\lambda_2 = \frac{\rho\lambda_1}{1+\lambda_1 P' \sum_{i\in S}|H_i|^2}$. Note that λ_2 is a function of H_S. Now using lemma 2.3.8 again to take expectation over c_S, we get

$$\begin{aligned}
&\mathbb{E}_{c_S}\left[\frac{e^{\rho\lambda_1\|Z_D\|^2}}{\left(1+\lambda_1 P' \sum_{i\in S}|H_i|^2\right)^{\rho n}} e^{\frac{-\rho\lambda_1\left\|Z_D+\sum_{i\in S} H_i c_i\right\|^2}{1+\lambda_1 P' \sum_{i\in S}|H_i|^2}}\right] \\
&\leq \frac{1}{\left(1+\lambda_1 P' \sum_{i\in S}|H_i|^2\right)^{\rho n}} \frac{1}{\left(1+\lambda_2 P' \sum_{i\in S}|H_i|^2\right)^{n}} e^{\left(\rho\lambda_1 - \frac{\lambda_2}{1+\lambda_2 P' \sum_{i\in S}|H_i|^2}\right)\|Z_D\|^2}
\end{aligned} \tag{2.140}$$

with $1 + \lambda_2 P' \sum_{i\in S}|H_i|^2 > 0$. Finally, taking expectation over Z, we get

$$\mathbb{P}\left[F_t | H_{[K]}\right] \leq \sum_S M^{\rho t} e^{-nE_0(\lambda_1;\rho,H_{[K]},S)} \tag{2.141}$$

where

$$\begin{aligned}
E_0(\lambda_1;\rho,H_{[K]},S) = \rho\ln\left(1+\lambda_1 P' \sum_{i\in S}|H_i|^2\right) + \ln\left(1+\lambda_2 P' \sum_{i\in S}|H_i|^2\right) + \\
\ln\left(1 - \left(1+P' \sum_{i\in D^c}|H_i|^2\right)\left(\rho\lambda_1 - \frac{\lambda_2}{1+\lambda_2 P' \sum_{i\in S}|H_i|^2}\right)\right)
\end{aligned} \tag{2.142}$$

with

$$1 > \left(1+P' \sum_{i\in D^c}|H_i|^2\right)\left(\rho\lambda_1 - \frac{\lambda_2}{1+\lambda_2 P' \sum_{i\in S}|H_i|^2}\right).$$

It is easy to see that the optimum value of λ_1 that maximizes E_0 is given by

$$\lambda_1^* = \frac{1}{\left(1+P' \sum_{i\in D^c}|H_i|^2\right)(1+\rho)} \tag{2.143}$$

and hence the maximum value of the exponent

$$E_0(\rho, H_{[K]}, S) = E_0(\lambda_1^*; \rho, H_{[K]}, S)$$

is given by

$$E_0(\rho, H_{[K]}, S) = \rho \ln \left(1 + \frac{P' \sum_{i \in S} |H_i|^2}{(1+\rho)\left(1 + P' \sum_{i \in D^c} |H_i|^2\right)} \right).$$

Therefore, we have

$$p_1 \leq \mathbb{E}\left[\sum_t \sum_S e^{\rho t \ln M} e^{-n E_0(\rho, H_{[K]}, S)} \right]. \tag{2.144}$$

Since we want an upper bound for (2.144), we would like to take minimum over $S \subset D : |S| = t$. For a given choice of D, this corresponds to minimizing $P' \sum_{i \in S} |H_i|^2$ which mean we take S to contain indices in D which correspond to t smallest fading coefficients (within D). Then, the best such bound is obtained by choosing D that maximizes $\frac{P' \sum_{i \in S} |H_i|^2}{\left(1 + P' \sum_{i \in D^c} |H_i|^2\right)}$. Clearly this corresponds to choosing D to contain indices corresponding to top K_1 fading coefficients.

Therefore, we get

$$p_1 \leq \mathbb{E}\left[\sum_t \binom{K_1}{t} e^{\rho t \ln M} e^{-n\rho \ln\left(1 + \frac{P' \sum_{i=K_1 - t + 1}^{K_1} |H_{(i)}|^2}{(1+\rho)\left(1 + P' \sum_{i=K_1+1}^{K} |H_{(i)}|^2\right)}\right)} \right].$$

Let $A_n = [\frac{\epsilon'}{\nu}, 1] \cap \left\{ \frac{i}{K_1} : i \in [K_1] \right\}$. For $\theta \in A_n$ and $t = \theta K_1$, using [145, Ex. 5.8] again, we have

$$\binom{K_1}{t} \leq \sqrt{\frac{K_1}{2\pi t (K_1 - t)}} e^{K_1 h(\frac{t}{K_1})} = O\left(\frac{1}{\sqrt{n}}\right) e^{n\mu\nu h(\theta)}. \tag{2.145}$$

The choice of ρ was arbitrary, and hence,

$$\begin{aligned}
& p_1 \\
& \leq \mathbb{E}\Bigg[\min\Bigg\{ 1, \sum_{\theta \in A_n} \exp\Bigg(- n \sup_{\rho \in [0,1]} \Bigg(\\
& \quad \rho \ln\left(1 + \frac{P' \sum_{i=\nu(1-\theta)K+1}^{\nu K} |H_{(i)}|^2}{(1+\rho)(1 + P' \sum_{i=\nu K+1}^{K} |H_{(i)}|^2)}\right) - \mu\nu h(\theta) - \rho\mu\nu\theta \ln M \Bigg)\Bigg)\Bigg\}\Bigg] \\
& \leq \mathbb{E}\Bigg[\min\Bigg\{ 1, |A_n| \exp\Bigg(- n \inf_{\theta \in A_n} \sup_{\rho \in [0,1]}
\end{aligned}$$

$$\left(\rho \ln\left(1+\frac{P'\sum_{i=\nu(1-\theta)K+1}^{\nu K}|H_{(i)}|^2}{(1+\rho)(1+P'\sum_{i=\nu K+1}^{K}|H_{(i)}|^2)}\right)-\mu\nu h(\theta)-\rho\mu\nu\theta\ln M\right)\right)\Bigg\}\Bigg] \tag{2.146}$$

where we have used min since $p_1 \leq 1$. Now, using similar arguments as in the proof of claim 5 and taking limits, we can see that

$$\begin{aligned}
&\inf_{\theta\in A_n}\sup_{\rho\in[0,1]}\left(\rho\ln\left(1+\frac{P'\sum_{i=\nu(1-\theta)K+1}^{\nu K}|H_{(i)}|^2}{(1+\rho)(1+P'\sum_{i=\nu K+1}^{K}|H_{(i)}|^2)}\right)-\mu\nu h(\theta)-\rho\mu\nu\theta\ln M\right)\\
&=\inf_{\theta\in A_n}\sup_{\rho\in[0,1]}\left(\rho\ln\left(1+\frac{P'_{tot}\alpha(\nu(1-\theta),\nu)}{(1+\rho)(1+P'_{tot}\alpha(\nu,1))}\right)-\mu\nu h(\theta)-\rho\mu\nu\theta\ln M\right)\\
&\quad+o(1)
\end{aligned} \tag{2.147}$$

with exponentially high probability. Hence,

$$\begin{aligned}
p_1 &\leq \mathbb{E}\Bigg[|A_n|\exp\Bigg(o(n)-n\inf_{\theta\in A_n}\sup_{\rho\in[0,1]}\\
&\quad\left(\rho\ln\left(1+\frac{P'_{tot}\alpha(\nu(1-\theta),\nu)}{(1+\rho)(1+P'_{tot}\alpha(\nu,1))}\right)-\mu\nu h(\theta)-\rho\mu\nu\theta\ln M\right)\Bigg)\Bigg]+o(1)\\
&\leq \mathbb{E}\Bigg[|A_n|\exp\Bigg(o(n)-n\inf_{\theta\in A}\sup_{\rho\in[0,1]}\\
&\quad\left(\rho\ln\left(1+\frac{P'_{tot}\alpha(\nu(1-\theta),\nu)}{(1+\rho)(1+P'_{tot}\alpha(\nu,1))}\right)-\mu\nu h(\theta)-\rho\mu\nu\theta\ln M\right)\Bigg)\Bigg]+o(1)]]
\end{aligned} \tag{2.148}$$

where $A=[\frac{\epsilon'}{\nu},1]$.

Therefore, choosing $P'_{tot} > \sup_{\theta\in A}\inf_{\rho\in[0,1]} P_{tot}(\theta,\rho)$ will ensure that $\limsup_{n\to\infty} p_1 = 0$.

□

Remark 6. *Note that the analysis of the CSIR case in this paper and the AWGN case in [3] are similar, in particular both analyze a (suboptimal for PUPE) maximum likelihood decoder. However, there are two new subtleties, compared to [3]. First, [3] applies Gallager's ρ-trick twice, where the second application (with parameter ρ_1 in the notation of [3]) is applied just before taking the expectation over Z in* (2.141)*. In the CSIR case, the summands of $\sum_S$ actually depend on the subset S through the fading gains, which makes the ρ-trick less appealing, and that is why we omitted it here. Secondly, because the summands depend on S, we upper bound each by taking the maximum over S, and this requires analysis of order statistics which is, of course, not present in the AWGN case.*

2.3.5 Converse

In this section we derive a converse for $\mathcal{E}^*$, based on the Fano inequality and the results from [6].

Theorem 2.3.9. *Let M be the codebook size. Given ϵ and μ, let $S = \mu \log M$. Then assuming that the distribution of $|H|^2$ has a density with $\mathbb{E}\left[|H|^2\right] = 1$ and $\mathbb{E}\left[|H|^4\right] < \infty$, $\mathcal{E}^*(M, \mu, \epsilon)$ satisfies the following two bounds*

1.

$$\mathcal{E}^*(M, \mu, \epsilon) \geq \inf \frac{P_{tot}}{S} \tag{2.149}$$

where infimum is taken over all $P_{tot} > 0$ that satisfies

$$\begin{aligned} &\theta S - \epsilon\mu \log\left(2^{S/\mu} - 1\right) - \mu h_2(\epsilon) \leq \\ &\log\left(1 + P_{tot}\alpha\left(1-\theta, 1\right)\right), \forall \theta \in [0,1] \end{aligned} \tag{2.150}$$

where $\alpha(a,b) = \int_a^b F^{-1}_{|H|^2}(1-\gamma)d\gamma$, and $F_{|H|^2}$ is the CDF of squared absolute value of the fading coefficients.

2.

$$\mathcal{E}^*(M, \mu, \epsilon) \geq \inf \frac{P_{tot}}{S} \tag{2.151}$$

where infimum is taken over all $P_{tot} > 0$ that satisfies

$$\epsilon \geq 1 - \mathbb{E}\left[Q\left(Q^{-1}\left(\frac{1}{M}\right) - \sqrt{\frac{2P_{tot}}{\mu}|H|^2}\right)\right] \tag{2.152}$$

where $\mathcal{Q}$ is the complementary CDF function of the standard normal distribution.

Proof. First, we use the Fano inequality.

Let $W = (W_1, ..., W_K)$, where $W_i \overset{iid}{\sim} Unif[M]$ denote the sent messages of K users. Let $X = (X_1, ..., X_K)$ where $X_i \in \mathbb{C}^n$ be the corresponding codewords, $Y \in \mathbb{C}^n$ be the received vector. Let $\hat{W} = \left(\hat{W}_1, ...\hat{W}_K\right)$ be the decoded messages. Then $W \to X \to Y \to \hat{W}$ forms a Markov chain. Then $\epsilon = P_e = \frac{1}{K}\sum_{i\in[K]} \mathbb{P}\left[W_i \neq \hat{W}_i\right]$.

Suppose a genie G reveal a set $S_1 \subset [K]$ for transmitted messages $W_{S_1} = \{W_i : i \in S_1\}$ and the corresponding fading coefficients H_{S_1} to the decoder. So, a converse bound in the Genie case is a converse bound for our problem (when there is no Genie). Further, the equivalent channel at the

receiver is

$$Y_G = \sum_{i \in S_2} H_i X_i + Z \tag{2.153}$$

where $S_2 = [K] \setminus S_1$, and the decoder outputs a $[K]$ sized tuple. So, PUPE with Genie is given by

$$P_e^G = \frac{1}{K} \sum_{i \in [K]} \mathbb{P}\left[W_i \neq \hat{W}_i^G\right]. \tag{2.154}$$

Now, it can be seen that the optimal decoder must have the codewords revealed by the Genie in the corresponding locations in the output tuple, i.e., if $\hat{W}^G$ denotes the output tuple (in the Genie case), for $i \in S_1$, we must have that $W_i = \hat{W}_i^G$. Otherwise, PUPE can be strictly decreased by including these Genie revealed codewords.

So, letting $E_i = 1[W_i \neq \hat{W}_i^G]$ and $\epsilon_i^G = \mathbb{E}[E_i]$, we have that $\epsilon_i^G = 0$ for $i \in S_1$. For $i \in S_2$, a Fano type argument gives

$$I(W_i; \hat{W}_i^G) \geq \log M - \epsilon_i^G \log(M-1) - h_2(\epsilon_i^G). \tag{2.155}$$

So, using the fact that

$$\begin{aligned}\sum_{i \in S_2} I(W_i; \hat{W}_i^G) &\leq I(W_{S_2}; \hat{W}_{S_2}^G) \\ &\leq n\mathbb{E}\left[\log(1 + P\sum_{i \in S_2} |H_i|^2)\right]\end{aligned}$$

we have

$$\begin{aligned}&|S_2| \log M - \sum_{i \in S_2} \epsilon_i^G \log(M-1) - \sum_{i \in S_2} h_2(\epsilon_i^G) \\ &\leq n\mathbb{E}\left[\log(1 + P\sum_{i \in S_2} |H_i|^2)\right].\end{aligned} \tag{2.156}$$

By concavity of h_2, we have

$$\frac{1}{K}\sum_{i \in S_2} h_2(\epsilon_i^G) = \frac{1}{K}\sum_{i \in [K]} h_2(\epsilon_i^G) \leq h_2(P_e^G). \tag{2.157}$$

Hence we get

$$\begin{aligned}&\frac{|S_2|}{K} \log M - P_e^G \log(M-1) - h_2(P_e^G) \\ &\leq \frac{n}{K}\mathbb{E}\left[\log(1 + P\sum_{i \in S_2} |H_i|^2)\right].\end{aligned} \tag{2.158}$$

Next, notice that $P_e^G \leq P_e \leq 1 - \frac{1}{M}$ and hence $P_e^G \log(M-1) + h_2(P_e^G) \leq P_e \log(M-1) + h_2(P_e)$. Further the inequality above hols for all $S_2 \subset [K]$ (which can depend of $H_{[K]}$ as well). Hence, letting $|S_2| = \theta K$

$$\begin{aligned} &\theta \log M - P_e \log(M-1) - h_2(P_e) \\ &\leq \frac{1}{\mu} \mathbb{E}\left[\log\left(1 + \inf_{S_2 : |S_2| = \theta K} \frac{P_{tot}}{K} \sum_{i \in S_2} |H_i|^2 \right) \right]. \end{aligned} \tag{2.159}$$

Now, taking limit as $K \to \infty$ and using results on strong laws of order statistics [155, Theorem 2.1], we get that

$$\begin{aligned} &\log\left(1 + \inf_{S_2 : |S_2| = \theta K} \frac{P_{tot}}{K} \sum_{i \in S_2} |H_i|^2 \right) \\ &\to \log\left(1 + P_{tot} \alpha(1-\theta, 1)\right). \end{aligned} \tag{2.160}$$

For any $a, b \in [0,1]$ with $a < b$, let $S_K \equiv S_K(a,b) = \frac{1}{K} \sum_{i=aK}^{bK} |H_{(i)}|^2$. Note that $S_K \to \alpha(a,b)$ as $K \to \infty$. Then

$$\begin{aligned} \mathbb{E}\left[S_K^2\right] &\leq \mathbb{E}\left[\left(\frac{1}{K} \sum_{i=1}^{K} |H_i|^2 \right)^2 \right] \\ &= 1 + \frac{\mathbb{E}\left[|H|^4\right] - 1}{K} \leq \mathbb{E}\left[|H|^4\right]. \end{aligned} \tag{2.161}$$

Hence the family of random variables $\{S_K : K \in \mathbb{N}\}$ is *uniformly integrable*. Further

$$0 \leq \log(1 + P_{tot} S_K) \leq P_{tot} S_K.$$

Hence the family $\{\log(1 + P_{tot} S_K) : K \geq 1\}$ is also uniformly integrable. Then from theorem [156, Theorem 9.1.6],

$$\mathbb{E}\left[\log(1 + P_{tot} S_K)\right] \to \log(1 + P_{tot} \alpha(a,b)).$$

Using this in (2.159) with $a = 1 - \theta$ and $b = 1$, we obtain (2.150).

Next we use the result from [6] to get another bound.

Using the fact that S/μ bits are needed to be transmitted under a per-user error of ϵ, we can get a converse on the minimum E_b/N_0 required by deriving the corresponding results for a single user quasi-static fading MAC. In [6], the authors gave the following non-asymptotic converse bound on the minimum energy required to send k bits for an AWGN channel. Consider the single user AWGN channel $Y = X + Z$, $Y, X \in \mathbb{R}^\infty$, $Z_i \overset{iid}{\sim} \mathcal{N}(0,1)$. Let $M^*(E, \epsilon)$ denote the largest M such that there exists a (E, M, ϵ) code for this channel: codewords $(c_1, ..., c_M)$ with $\|c_i\|^2 \leq E$ and a decoder such that probability of error is smaller than ϵ. The following is a converse bound from [6].

Lemma 2.3.10 ([6]). *Any (E, M, ϵ) code satisfies*

$$\frac{1}{M} \geq Q\left(\sqrt{2E} + Q^{-1}\left(1-\epsilon\right)\right) \tag{2.162}$$

Translating to our notations, for the channel $Y = HX + Z$, conditioned on H, if $\epsilon(H)$ denotes the probability of error for each realization of H, then we have

$$\frac{1}{M} \geq Q\left(\sqrt{\frac{2P_{tot}}{\mu}|H|^2} + Q^{-1}\left(1-\epsilon(H)\right)\right). \tag{2.163}$$

Further $\mathbb{E}\left[\epsilon(H)\right] = \epsilon$. Therefore we have

$$\epsilon \geq 1 - \mathbb{E}\left[Q\left(Q^{-1}\left(\frac{1}{M}\right) - \sqrt{\frac{2P_{tot}}{\mu}|H|^2}\right)\right]. \tag{2.164}$$

Hence we have the required converse bound.

Remark 7. *We also get the following converse from [19, theorem 7] by taking the appropriate limits $P = \frac{P_{tot}}{\mu n}$ and $n \to \infty$.*

$$\log M \leq -\log\left(\mathbb{E}\left[\mathcal{Q}\left(\frac{c + \frac{P_{tot}|H|^2}{\mu}}{\sqrt{\frac{2P_{tot}|H|^2}{\mu}}}\right)\right]\right) \tag{2.165}$$

where c satisfies

$$\mathbb{E}\left[\mathcal{Q}\left(\frac{c - \frac{P_{tot}|H|^2}{\mu}}{\sqrt{\frac{2P_{tot}|H|^2}{\mu}}}\right)\right] = 1 - \epsilon. \tag{2.166}$$

But this is strictly weaker than (2.164). *This is because, using lemma 2.3.10, we perform hypothesis testing (in the meta-converse) for each realization of H but in the bound used in [19], hypothesis testing is performed over the joint distribution (including the distribution of H). This is to say that if H is presumed to be constant (and known), then in* (2.165) *and* (2.166) *we can remove the expectation over H and this gives precisely the same bound as* (2.163).

□

Bounds tighter than (2.150) can be obtained if further assumptions are made on the codebook. For instance, if we assume that each codebook consists of iid entries of the form $\frac{C}{K}$ where C is sampled from a distribution with zero mean and finite variance, then using ideas similar to [137, Theorem 3] we have the following converse bound.

Theorem 2.3.11. *Let M be the codebook size, and let μn users ($\mu < 1$) generate their codebooks independently with each code symbol iid of the form $\frac{C}{K}$ where C is of zero mean and variance P_{tot}. Then in order for the iid codebook to achieve PUPE ϵ with high probability, the energy-per-bit $\mathcal{E}$ should satisfy*

$$\mathcal{E} \geq \inf \frac{P_{tot}}{\mu \log M} \tag{2.167}$$

where infimum is taken over all $P_{tot} > 0$ that satisfies

$$\begin{aligned} &\ln M - \epsilon \ln(M-1) - h(\epsilon) \\ &\leq \left(M\mathcal{V}\left(\frac{1}{\mu M}, P_{tot}\right) - \mathcal{V}\left(\frac{1}{\mu}, P_{tot}\right)\right) \end{aligned} \tag{2.168}$$

where $\mathcal{V}$ is given by [137]

$$\begin{aligned} \mathcal{V}(r,\gamma) &= r\ln\left(1+\gamma-\mathcal{F}(r,\gamma)\right) + \ln\left(1+r\gamma-\mathcal{F}(r,\gamma)\right) \\ &\quad - \frac{\mathcal{F}(r,\gamma)}{\gamma} \end{aligned} \tag{2.169}$$

$$\mathcal{F}(r,\gamma) = \frac{1}{4}\left(\sqrt{\gamma\left(\sqrt{r}+1\right)^2+1} - \sqrt{\gamma\left(\sqrt{r}-1\right)^2+1}\right)^2 \tag{2.170}$$

Proof sketch. The proof is almost the same as in [137, Theorem 3] (see [137, Remark 3] as well). We will highlight the major differences here. First, our communication system can be modeled as a support recovery problem as follows. Let A be the $n \times KM$ matrix consisting of $n \times M$ blocks of codewords of users. Let $\boldsymbol{H}$ be the $KM \times KM$ block diagonal matrix with block i being a diagonal $M \times M$ matrix with all diagonal entries being equal to H_i. Finally let $W \in \{0,1\}^{KM}$ with K blocks of size M each and within each M sized block, there is exactly one 1. So the position of 1 in block i of W denotes the message or codeword corresponding to the user i which is the corresponding column in block i of matrix A. Hence our channel can be represented as

$$Y = A\boldsymbol{H}W + Z \tag{2.171}$$

with the goal of recovering W.

Next the crucial step is bound $R^K(\epsilon, M)$ in (2.155) as

$$\begin{aligned} R^K(\epsilon, M) &\leq I(W;Y|A) \\ &= I(\boldsymbol{H}W;Y|A) - I(\boldsymbol{H}W;Y|A,W) \end{aligned} \tag{2.172}$$

where the equality in the above display follows from [137, equation (78)]. The first term in above display is bounded as

$$\begin{aligned} I(\boldsymbol{H}W; Y|A = A_1) &= I(\boldsymbol{H}W; A_1 HW + Z) \\ &\leq \sup_U I(U; A_1 U + Z) \end{aligned} \tag{2.173}$$

where A_1 is a realization of A and supremum is over random vectors $U \in \mathbb{C}^{KM}$ such that $\mathbb{E}[U] = 0$ and $\mathbb{E}[UU^*] = \mathbb{E}[(\boldsymbol{H}W)(\boldsymbol{H}W)^*] = \frac{\mathbb{E}\left[|H_1|^2\right]}{M} I_{KM\times KM}$. Now similar to [137], the supremum is achieved when

$$U \sim \mathcal{CN}\left(0, \frac{\mathbb{E}\left[|H_1|^2\right]}{M} I_{KM\times KM}\right).$$

Hence

$$I(\boldsymbol{H}W; Y|A = A_1) \leq \log\det\left(I_{n\times n} + \frac{1}{M} AA^*\right). \tag{2.174}$$

Next, for any realization A_1 and W_1 of A and W respectively, we have

$$\begin{aligned} I(\boldsymbol{H}W; Y|A = A_1, W = W_1) &= I(\boldsymbol{H}W_1; A_1 \boldsymbol{H}W_1 + Z) \\ &= I(\tilde{\boldsymbol{H}}; (A_1)_{W_1} \tilde{\boldsymbol{H}} + Z) \\ &\geq I(\tilde{\boldsymbol{H}}; \tilde{\boldsymbol{H}} + (A_1)^\dagger_{W_1} Z) \end{aligned} \tag{2.175}$$

where $\tilde{\boldsymbol{H}} = [H_1, ..., H_K]^T$ and $(A_1)_{W_1}$ is the $n\times K$ submatrix of A_1 formed by columns corresponding to the support of W_1 and $\dagger$ denotes the Moore-Penrose inverse (pseudoinverse). The last equality in the above follows from the data processing inequality. Now, by standard mutual information of Gaussians, we have

$$I(\tilde{\boldsymbol{H}}; \tilde{\boldsymbol{H}} + (A_1)^\dagger_{W_1} Z) = \log\det\left(I_{K\times K} + \left((A_1)_{W_1}\right)^* (A_1)_{W_1}\right). \tag{2.176}$$

Hence

$$I(\boldsymbol{H}W; Y|A, W) = \mathbb{E}\left[\log\det\left(I_{K\times K} + A^*_W A_W\right)\right]. \tag{2.177}$$

Hereafter, the we can proceed similarly to the proof of [137, Theorem 3] using results from random-matrix theory [157, 158] to finish the proof. □

We remark here that for a general fading distribution, the term $I(\tilde{\boldsymbol{H}}; \tilde{\boldsymbol{H}} + (A_1)^\dagger_{W_1} Z)$ can be lower

bounded similar to the proof of [137, Theorem 3] using EPI (and its generalization [159]) to get

$$I(\tilde{\boldsymbol{H}};\tilde{\boldsymbol{H}}+((A_1)_{W_1})^{\dagger}Z)\geq K\log\left(1+N_H\left(\det\left(((A_1)_{W_1})^{*}(A_1)_{W_1}\right)\right)^{\frac{1}{K}}\right) \tag{2.178}$$

where $N_H=\frac{1}{\pi e}\exp(h(H))$ is the entropy power of fading distribution. Hence

$$I(\boldsymbol{H}W;Y|A,W)\geq K\mathbb{E}\left[\log\left(1+N_H\left(\det\left(A_W^{*}A_W\right)\right)^{\frac{1}{K}}\right)\right]. \tag{2.179}$$

Again, we can use results from random-matrix theory [158] and proceed similarly to the proof of [137, Theorem 3] to get a converse bound with the second term in (2.168) replaced by $\mathcal{V}_{LB}\left(\frac{1}{\mu},P_{tot}\right)$ and

$$\mathcal{V}_{LB}(r,\gamma)=\ln\left(1+\gamma r\left(\frac{r}{r-1}\right)^{r-1}\frac{1}{e}\right) \tag{2.180}$$

We make a few observations regarding the preceding theorem. First and foremost, this hold only for the case of no-CSI because the term analogous to $I(\boldsymbol{H}W;Y|A,W)$ in the case of CSIR is $I(\boldsymbol{H}W;Y|A,\boldsymbol{H},W)$ which is zero. Next, it assumes that the codebooks have iid entries with variance scaling $\Theta(1/n)$. This point is crucial to lower bounding $I(\boldsymbol{H}W;Y|A,W)$, and this is where a significant improvement comes when compared to (2.150). Indeed, EPI and results from random matrix theory give $O(n)$ lower bound for $I(\boldsymbol{H}W;Y|A,W)$. This once again brings to focus the the difference between classical regime and the scaling regime, where in the former, this term is negligible. Further this leaves open the question of whether we could improve performance in the high-density of users case by using non-iid codebooks.

Now, as to what types of codebooks give a $\Theta(n)$ lower bound for $I(\boldsymbol{H}W;Y|A,W)$, a partial answer can be given by carefully analyzing the full proof of the theorem. In particular, if $\mathcal{S}=\text{supp}W$ i.e, the support of W, then as seen from [137, equation (85)], any non zero lower bound on $\det(A_{\mathcal{S}}^{*}A_{\mathcal{S}})^{1/K}$ in the limit is enough. So if the matrix $A_{\mathcal{S}}^{*}A_{\mathcal{S}}$ possesses strong diagonal dominance then it is possible to have such a non zero lower bound on $\det(A_{\mathcal{S}}^{*}A_{\mathcal{S}})^{1/K}$ for every S [160]. These could be ensured by having codewords that are overwhelmingly close to orthogonal.

2.4 Technical results

2.4.1 Proof of proposition 1

Proof. We prove the second upper bound in (2.14). This is based on a single-user converse using the genie argument. Formally, since we consider per-user error, it is enough to look at the event that a particular user is not decoded. Let $W_i\overset{iid}{\sim}\text{unif}[M]$ be the message of user i. The channel (2.1) can be written as $Y=H_1X_1+\hat{Z}+Z$ where $\hat{Z}=\sum_{i=2}^{K}H_iX_i$ denotes the interference. Let $L(Y)$ be the decoder output. Also, let $L(Y,\hat{Z})$ be the decoder output when it has knowledge of $\hat{Z}$. Hence a

converse bound $\mathbb{P}\left[W_1 \neq (L(Y))_1\right] \geq \epsilon$ is implied by $\mathbb{P}\left[W_1 \neq \left(L(Y,\hat{Z})\right)_1\right] \geq \epsilon$ for all $L(\cdot,\cdot)$. Since $Y - \hat{Z}$ is a sufficient statistic of $(Y,\hat{Z})$ for W_1, we have, equivalently, $\mathbb{P}\left[W_1 \neq \left(L(Y-\hat{Z})\right)_1\right] \geq \epsilon$ for all $L(\cdot)$. Letting $\hat{Y} = Y - \hat{Z}$, this is equivalent to a converse for the channel $\hat{Y} = H_1 X_1 + Z$: $\mathbb{P}\left[W_1 \neq \left(L(\hat{Y})\right)_1\right] \geq \epsilon$ for all $L(\cdot)$. This is just the usual single user converse, and hence the bound is given by $R \leq C_\epsilon = \sup\{\xi : \mathbb{P}\left[\log_2(1+P|H_1|^2) \leq \xi\right] \leq \epsilon\} = \log_2(1 - P\ln(1-\epsilon))$[19].

□

2.4.2 Proofs of certain claims

Proof of claim 2. We have $\|Y\|^2 - \|P_{A_0}Y\|^2 = \left\|P_{A_0}^\perp \hat{Z}\right\|^2 \leq \left\|\hat{Z}\right\|^2 - \left\|P_{A_1}\hat{Z}\right\|^2 = \left\|P_{A_1}^\perp \hat{Z}\right\|^2$.

Also, $\left\|P_{A_1}^\perp Y\right\|^2 = \left\|P_{A_1}^\perp \sum_{i\in S_2^*} H_i c_i + P_{A_1}^\perp \hat{Z}\right\|^2$. Hence we have

$$
\begin{aligned}
p_2 &= \mathbb{P}\left[\bigcup_{t,S,K_1}\left\{\frac{\|Y\|^2 - \|P_{A_0}Y\|^2}{\|Y\|^2 - \|P_{A_1}Y\|^2} \geq V_{n,t}\right\}\right] \\
&= \mathbb{P}\left[\bigcup_{t,S,K_1}\left\{\left\|\hat{Z}\right\|^2 - \left\|P_{A_0}\hat{Z}\right\|^2 \geq V_{n,t}\left\|P_{A_1}^\perp \sum_{i\in S_2^*} H_i c_i + P_{A_1}^\perp \hat{Z}\right\|^2\right\}\right] \\
&\leq \mathbb{P}\left[\bigcup_{t,S,K_1}\left\{\left\|P_{A_1}^\perp \hat{Z}\right\|^2 \geq V_{n,t}\left\|P_{A_1}^\perp \sum_{i\in S_2^*} H_i c_i + P_{A_1}^\perp \hat{Z}\right\|^2\right\}\right] \\
&= \mathbb{P}\left[\bigcup_{t,S,K_1}\left\{(1-V_{n,t})\left\|P_{A_1}^\perp \hat{Z}\right\|^2 - 2V_{n,t}Re\left\langle P_{A_1}^\perp \hat{Z}, P_{A_1}^\perp \sum_{i\in S_2^*} H_i c_i\right\rangle \geq V_{n,t}\left\|P_{A_1}^\perp \sum_{i\in S_2^*} H_i c_i\right\|^2\right\}\right] \\
&= \mathbb{P}\left[\bigcup_{t,S,K_1}\left\{(1-V_{n,t})^2\left\|P_{A_1}^\perp \hat{Z}\right\|^2 - 2V_{n,t}(1-V_{n,t})Re\left\langle P_{A_1}^\perp \hat{Z}, P_{A_1}^\perp \sum_{i\in S_2^*} H_i c_i\right\rangle \geq V_{n,t}(1-V_{n,t})\left\|P_{A_1}^\perp \sum_{i\in S_2^*} H_i c_i\right\|^2\right\}\right] \\
&= \mathbb{P}\left[\bigcup_{t,S,K_1}\left\{\left\|(1-V_{n,t})P_{A_1}^\perp \hat{Z} - V_{n,t}P_{A_1}^\perp \sum_{i\in S_2^*} H_i c_i\right\|^2 \geq V_{n,t}\left\|P_{A_1}^\perp \sum_{i\in S_2^*} H_i c_i\right\|^2\right\}\right] \qquad (2.181)
\end{aligned}
$$

□

Proof of claim 3. Conditional of $H_{[K]}$ and A_0, $\hat{Z} \sim \mathcal{CN}\left(0, \left(1 + P'\sum_{i\in S\setminus S_2^*}|H_i|^2\right)\right)$. Hence

$$
P_{A_1}^\perp\left(\hat{Z} - \frac{V_{n,t}}{1-V_{n,t}}\sum_{i\in S_2^*} H_i c_i\right) \sim \mathcal{CN}\left(-\frac{V_{n,t}}{1-V_{n,t}}P_{A_1}^\perp \sum_{i\in S_2^*} H_i c_i, \left(1 + P'\sum_{i\in S\setminus S_2^*}|H_i|^2\right)P_{A_1}^\perp\right). \qquad (2.182)
$$

Now, the rank of $P_{A_1}^\perp$ is $n - K_1 + t$ because the vectors in A_1 are linearly independent almost surely. Let $\mathcal{U}$ be a unitary change of basis matrix that rotates the range space of $P_{A_1}^\perp$ to the space corresponding to first $(n - K_1 + t)$ coordinates. Then

$$
\begin{aligned}
&\left\| \mathcal{CN}\left(-\frac{V_{n,t}}{1-V_{n,t}} P_{A_1}^{\perp} \sum_{i\in S_2^*} H_i c_i, \left(1 + P' \sum_{i\in S\setminus S_2^*} |H_i|^2\right) P_{A_1}^{\perp}\right)\right\|^2 \\
&= \left\| \mathcal{U}\left(\mathcal{CN}\left(-\frac{V_{n,t}}{1-V_{n,t}} P_{A_1}^{\perp} \sum_{i\in S_2^*} H_i c_i, \left(1 + P' \sum_{i\in S\setminus S_2^*} |H_i|^2\right) P_{A_1}^{\perp}\right)\right)\right\|^2 \\
&= \left\| \mathcal{CN}\left(-\frac{V_{n,t}}{1-V_{n,t}} \mathcal{U} P_{A_1}^{\perp} \sum_{i\in S_2^*} H_i c_i, \left(1 + P' \sum_{i\in S\setminus S_2^*} |H_i|^2\right) \mathcal{U} P_{A_1}^{\perp} \mathcal{U}^*\right)\right\|^2 . \qquad (2.183)
\end{aligned}
$$

Observe that $\mathcal{U} P_{A_1}^{\perp} \mathcal{U}^*$ is a diagonal matrix with first $(n - K_1 + t)$ diagonal entries being ones and rest all 0. Also, if $W = P + iQ \sim \mathcal{CN}(\mu, \Gamma)$ (with pseudo-covariance being 0) then

$$
\begin{bmatrix} P \\ Q \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} Re(\mu) \\ Im(\mu) \end{bmatrix}, \frac{1}{2} \begin{bmatrix} Re(\Gamma) & -Im(\Gamma) \\ Im(\Gamma) & Re(\Gamma) \end{bmatrix} \right). \qquad (2.184)
$$

Using this and the definition of non-central chi-squared distribution the claim follows.

□

Proof of Claim 4. We have

$$
\begin{aligned}
f_n(x) &= x + 1 + \frac{2V_{n,t}}{1-V_{n,t}}(1+x) - \sqrt{1 + \frac{2V_{n,t}}{1-V_{n,t}}(1+x)} \sqrt{2x + 1 + \frac{2V_{n,t}}{1-V_{n,t}}(1+x)} \\
&= \frac{1}{1-V_{n,t}} \left[(1+V_{n,t})(x+1)) - 2\sqrt{V_{n,t}} \sqrt{\left(x + \frac{(1+V_{n,t})^2}{4V_{n,t}}\right)^2 - \frac{(1-V_{n,t}^2)^2}{16V_{n,t}^2}} \right] \qquad (2.185)
\end{aligned}
$$

Hence

$$
\begin{aligned}
f'(x) &= \frac{1}{1-V_{n,t}} \left[1 + V_{n,t} - 2\sqrt{V_{n,t}} \frac{a}{\sqrt{a^2 - b^2}} \right] \\
&= \frac{1}{1-V_{n,t}} \left(\sqrt{V_{n,t}} - \sqrt{\frac{a+b}{a-b}} \right) \left(\sqrt{V_{n,t}} - \sqrt{\frac{a-b}{a+b}} \right) \qquad (2.186)
\end{aligned}
$$

where $a = \left(x + \frac{(1+V_{n,t})^2}{4V_{n,t}}\right)$ and $b = \frac{1-V_{n,t}^2}{4V_{n,t}}$. Also $a > 0$ and $b > 0$. Further $a + b > a - b$ and

$$
\begin{aligned}
&\sqrt{V_{n,t}} < \sqrt{\frac{a-b}{a+b}} = \sqrt{\frac{V_{n,t}(1 + V_{n,t} + 2x)}{1 + V_{n,t} + 2V_{n,t}x}} \\
&\iff 2V_{n,t}x + 1 + V_{n,t} < 2x + 1 + V_{n,t} \\
&\iff 0 < V_{n,t} < 1
\end{aligned}
$$

which is true. Hence both the factors in (2.186) are negative. Therefore $f'(x) > 0$. □

2.5 Maximal per-user error

In this chapter we briefly describe relations between maximal per-user error (PUPE-max) defined in (2.5) and PUPE. First, we represent our system as in (2.171)

$$Y = AHW + Z. \tag{2.187}$$

Let $P_{e,i}(A) = \mathbb{P}\left[W_i \neq \hat{W}_i\right]$. We are interested in bounding the variance of $P_{e,i}(A)$ so that

$$\mathbb{E}\left[P_{e,u}^{\max}(A)\right] = \mathbb{E}\left[\max\{P_{e,i}(A) : i \in [K]\}\right]$$

can be related to $\mathbb{E}\left[P_{e,i}(A)\right] = \mathbb{E}\left[P_{e,u}\right]$ due to symmetry on users by random codebook generation. Consider two coupled systems

$$Y = AHW + Z \tag{2.188}$$

$$Y' = A'HW + Z \tag{2.189}$$

where A and A' are fixed so that the channels are dependent on these.

Now we have

$$|P_{e,i}(A) - P_{e,i}(A')| \leq d_{TV}(P_{Y,H,W}, P_{Y',H,W}) \leq \sqrt{\frac{1}{2} D\left(P_{Y,H,W} || P_{Y',H,W}\right)} \tag{2.190}$$

where $d_{TV}(P,Q) = \sup\{|P(A) - Q(A)| : A \,\text{is measurable}\}$ is the total variation distance between measures P and Q, $D(P||Q) = \mathbb{E}_P\left[\ln \frac{dP}{dQ}\right]$ is the Kullback-Leibler divergence (in nats) and the last inequality is the Pinsker's inequality (see [161]). Now using properties of D (see [162, Theorem 2.2])

$$\begin{aligned} D\left(P_{Y,H,W} || P_{Y',H,W}\right) &= D(P_{Y|H,W} || P_{Y'|H,W} | P_{H,W}) \\ &= \int_{H,W} D(P_{Y|H=h,W=w} || P_{Y'|H=h,W=w}) dP_{H,W}(h,w) \end{aligned} \tag{2.191}$$

Now note that conditioned on $H = h, W = w$, we have $Y \sim \mathcal{CN}(Ahw, I_n)$ and $Y' \sim \mathcal{CN}(A'hw, I_n)$. Hence a simple computation shows that $D(P_{Y|H=h,W=w} || P_{Y'|H=h,W=w}) = \|Ahw - A'hw\|^2$. Therefore we have

$$D\left(P_{Y,H,W} || P_{Y',H,W}\right) = \mathbb{E}\left[\|(A - A')HW\|^2\right]. \tag{2.192}$$

Now let $B = A - A'$ and $X = HW$. Then

$$\mathbb{E}\left[\|BX\|^2\right] = \sum_{i\in[n]} \mathbb{E}\left[\sum_{j,k\in[KM]} B_{i,j}\bar{B}_{i,k}X_i\bar{X}_k\right] \tag{2.193}$$

Note that $\mathbb{E}\left[X_j\bar{X}_k\right]$ is zero if $j \neq k$ and it is $1/M$ otherwise. Hence

$$\mathbb{E}\left[\|BX\|^2\right] = \frac{1}{M}\sum_{i\in[n]}\sum_{j\in[KM]} B_{i,j}\bar{B}_{i,j} = \frac{1}{M}\|B\|_F^2 . \tag{2.194}$$

Therefore

$$D\left(P_{Y,H,W}||P_{Y',H,W}\right) = \frac{1}{M}\|A - A'\|_F^2 . \tag{2.195}$$

So combining this with (2.190), we obtain

$$|P_{e,i}(A) - P_{e,i}(A')| \leq \sqrt{\frac{1}{2M}}\|A - A'\|_F . \tag{2.196}$$

Now let each entry of A and A' to be distributed iid as $\mathcal{CN}(0, P)$ where $P = P_{tot}/K$. Further, let $\tilde{A} = \sqrt{\frac{K}{P_{tot}}}A$ and $\tilde{A}' = \sqrt{\frac{K}{P_{tot}}}A'$. So the entries of $\tilde{A}$ and $\tilde{A}'$ are iid $\mathcal{CN}(0, 1)$. Therefore, with slight abuse of notation, we can rewrite (2.196) as

$$|P_{e,i}(\tilde{A}) - P_{e,i}(\tilde{A}')| \leq \sqrt{\frac{P_{tot}}{2MK}}\left\|\tilde{A} - \tilde{A}'\right\|_F . \tag{2.197}$$

Hence the function $P_{e,i}$ is Lipschitz with Lipschitz constant $L = \sqrt{\frac{P_{tot}}{2MK}}$. By concentration of Lipschitz functions of Gaussian random vectors [148, Theorems 5.5, 5.6], we have that $P_{e,i}(\tilde{A})$ is sub-Gaussian with

$$\mathrm{Var}(P_{e,i}(A)) \leq 4L^2 = \frac{2P_{tot}}{KM}. \tag{2.198}$$

Hence, using bounds on expected maximum of sub-Gaussian random variables (see [148, Section 2.5]), we obtain

$$\begin{aligned}\mathbb{E}\left[\max_{i\in[K]} P_{e,i}(A)\right] &\leq \mathbb{E}\left[P_{e,u}\right] + \sqrt{\mathrm{Var}(P_{e,i}(A))\ln K} \\ &= \mathbb{E}\left[P_{e,u}\right] + \sqrt{\frac{2P_{tot}}{M}\frac{\ln K}{K}} \xrightarrow{K\to\infty} \mathbb{E}\left[P_{e,u}\right].\end{aligned} \tag{2.199}$$

Therefore, a random coding argument along with (2.199) shows that PUPE-max has same asymptotics as PUPE in the linear scaling regime. For FBL performance, if each user sends $k = 100$ bits then $M = 2^k$ and hence $\mathbb{E}\left[P_{e,u}^{\max}\right] \approx \mathbb{E}\left[P_{e,u}\right]$

Chapter 3

Many user AWGN MAC

In this chapter we present new achievability bounds for the many user AWGN MAC based on spatially coupled codes and scalar AMP decoding; this bound turns out to be tighter than [14].

3.1 System model

The AWGN MAC is defined as follows. The channel law $P_{Y^n|X^n}$ is described by

$$Y^n = \sum_{i=1}^{K} X_i^n + Z^n \tag{3.1}$$

where $Z^n \sim \mathcal{N}(0, I_n)$. We assume that there is a maximum power constraint:

$$\|X_i^n\|^2 \leq nP. \tag{3.2}$$

We drop superscript n for brevity. As mentioned in section 2.1.1, we can represent the system in the language of compressed sensing as

$$Y = AU + Z \tag{3.3}$$

where $A \in \mathbb{R}^{n \times KM}$ denotes the codebook and $U \in \{0,1\}^{KM}$ is block sparse: for each $i \in [K]$ it satisfies $\sum_{j=1}^{M} U_{(i-1)M+j} = 1$.

As before, we consider the linear scaling regime where the number of users K scales with n, and $n \to \infty$. We are interested in the tradeoff of minimum E_b/N_0 required for the PUPE to be smaller than ϵ, with the user density μ. So, we fix the message size k. Let $S = k\mu$ be the spectral efficiency.

We focus on the case of different codebooks, but under symmetric rate. So if M denotes the size of the codebooks, then $S = \frac{K \log M}{n} = \mu \log M$. Hence, given S and μ, M is fixed. Let $P_{tot} = KP$

denote the total power. Therefore denoting by $\mathcal{E}$ the energy-per-bit, $\mathcal{E} = E_b/N_0 = \frac{nP}{2\log_2 M}$. We consider the power P decaying as $O(1/n)$ (required for finite E_b/N_0).

Let $\mathcal{C}_j = \{c_1^j, ..., c_M^j\}$ be the codebook of user j, of size M. The power constraint is given by $\|c_i^j\|^2 \leq nP = 2\mathcal{E}\log_2 M, \forall j \in [K], i \in [M]$. The collection of codebooks $\{\mathcal{C}_j\}$ is called an $(n, M, \epsilon, \mathcal{E}, K)$–code if it satisfies the power constraint described before, and the per-user probability of error is smaller than ϵ. Then, we can define the following fundamental limit for the channel

$$\mathcal{E}^*(M, \mu, \epsilon) = \lim_{n\to\infty} \inf \{\mathcal{E} : \exists(n, M, \epsilon, \mathcal{E}, K = \mu n) - code\}.$$

3.2 Achievability bound: Spatially coupled AMP

Similar to section 2.3.3 we obtain an achievability bound on $\mathcal{E}$ based on the spatially coupled codebook design in conjunction with AMP decoder.

The spatially coupled codebook is almost the same as in 2.3.3. We state it again for convenience. Let $R, C \in \{1, 2, \cdots\}$ be such that R divides n and C divides KM. The codebook A is divided into blocks of size $\frac{n}{R} \times \frac{KM}{C}$ and hence can be considered as a block matrix of size $R \times C$. Let $B \in \mathbb{R}^{R\times C}$ be the base matrix with nonnegative entries $B_{r,c}$ such that $\sum_{r=1}^{R} B_{r,c} = 1$ for all $c \in \{1, 2, \cdots, C\}$. Further, with abuse of notation, let $r : [n] \to [R]$ and $c : [KM] \to [C]$ denote functions that map a particular row or column index to its corresponding block. Then the matrix A is constructed as $A_{i,j} \sim \mathcal{N}\left(0, E\frac{R}{n}B_{r(i),c(j)}\right)$ (where $E = \frac{P_{tot}}{\mu}$) with $\{A_{i,j}\}$ independent across tuples (i, j). As in 2.3.3 we use the (ω, Λ, ρ) base matrix from [27] as the choice of B (see (2.106)). Let $\tilde{\mu} = \frac{R}{C}\mu$ be the effective user density.

Next we present the AMP algorithm adapted to the spatially coupled codebook (SC-AMP) in the AWGN setting from [27]. Consider the scalar channel (with abuse of notation from (2.107))

$$V_{\sigma^2} = X + \sigma W \tag{3.4}$$

where now $X \sim \text{Ber}(1/M)$, $W \sim \mathcal{N}(0, 1)$ with $X \perp W$. Again, the main difference from [27] is that we ignore the block sparse structure of U and just use the fact that the empirical distribution of entries in U converge to $\text{Ber}(1/M)$. This is sufficient for the state evolution to be valid [153], and similar to the results in previous section, the state evolution is defined on the scalar channel (3.4) (instead of the vector channel in [27] which would be infeasible for numerical evaluation when M is very large but finite).

Start with $U^{(0)} = 0 \in \mathbb{C}^{KM}$, $R^{(0)} = Y$. Then for $t = 1, 2, \cdots$ we have the following iterations

$$U^{(t)} = \eta^{(t)}\left((\tilde{Q}^{(t-1)} \odot A)^T R^{(t-1)} + U^{(t-1)}\right) \tag{3.5}$$

$$R^{(t)} = Y - AU^{(t)} + \frac{R}{C}\mu M(\tilde{b}^{(t)} \odot R^{(t-1)}) \tag{3.6}$$

where $\odot$ denotes element wise product, and matrix $\tilde{Q}^{(t)}$, vector $\tilde{b}^{(t)}$ and denoiser $\eta^{(t)} : \mathbb{R}^{KM} \to \mathbb{R}^{KM}$ will be defined next via the state evolution.

Let $\psi_c^{(0)} = \infty$. Then for $t \geq 1$, for each $r \in [R]$ and $c \in [C]$ we define

$$\gamma_r^{(t)} = \sum_{c=1}^{C} B_{r,c} \psi_c^{(t)} \tag{3.7}$$

$$\phi_r^{(t)} = \frac{1}{E} + \tilde{\mu} M \gamma_r^{(t)} \tag{3.8}$$

$$\tau_c^{(t)} = \frac{1}{\sum_{r=1}^{R} B_{r,c} \left(\phi_r^{(t)}\right)^{-1}} \tag{3.9}$$

$$\psi_c^{(t+1)} = \text{mmse}(\tau_c^{(t)}) \tag{3.10}$$

where $\text{mmse}(\cdot)$ is defined as

$$\text{mmse}(\sigma^2) = \mathbb{E}\left[\left(X - \mathbb{E}\left[X|V_{\sigma^2}\right]\right)^2\right] \tag{3.11}$$

where V is the output of the scalar channel (3.4).

Now the matrices $\tilde{Q}^{(t)}$ and vectors $\tilde{b}^{(t)}$ are defined as follows. For each $i \in [n]$ and $j \in [KM]$

$$\tilde{b}_i^{(t)} = \tilde{\mu} M \frac{\gamma_{r(i)}^{(t)}}{\phi_{r(i)}^{(t-1)}}$$

$$\tilde{Q}_{i,j}^{(t)} = \frac{\tau_{c(j)}^{(t)}}{\phi_{r(i)}^{(t)}}$$

The denoiser at time t is given by $\eta^{(t)} = (\eta_1^{(t)}, \cdots, \eta_{KM}^{(t)})$ with

$$\eta_i^{(t)}(z) = \mathbb{E}\left[X | X + \sqrt{\tau_{c(i)}^{(t)}} W = z\right] = \eta(z, \tau_{c(i)}^{(t)}) \tag{3.12}$$

where $\eta(\cdot, \cdot)$ is scalar denoiser

$$\eta(z, \sigma^2) = \mathbb{E}\left[X | V_{\sigma^2} = z\right] = \mathbb{E}\left[X | X + \sigma W = z\right] \tag{3.13}$$

The estimate of U after t steps is given by (see [22, 27] for more details on hard decision estimate)

$$\hat{U}^{(t)} = (\tilde{Q}^t \odot A)^T R^{(t)} + U^{(t)} \tag{3.14}$$

Let $\hat{S}^{(t)}$ denote the estimate of U. To convert $\hat{U}^{(t)}$ into $\hat{S}^{(t)}$ we perform a simple thresholding for each $c \in [C]$ i.e., for each i

$$\hat{S}_i^{(t)}(\theta_{c(i)}) = 1[\hat{U}_i^{(t)} > \theta_{c(i)}] \tag{3.15}$$

where $\{\theta_c : c \in [C]\}$ is a set of thresholds.

To state the main achievability bound, define the replica potential

$$\mathcal{F}_{\mathsf{awgn}}(\tau;\mu,E,M) = (\mu M)I(X;X+\sqrt{\tau}W) + \frac{1}{2}\left(\ln\tau + \frac{1}{\tau E} - 1\right) \tag{3.16}$$

where $X \sim \mathrm{Ber}(1/M)$, $W \sim \mathcal{N}(0,1)$ and $X \perp W$.

Theorem 3.2.1. *Fix any $\mu > 0$, $E > 0$ and $k = \log_2 M \geq 1$. Then for every $\mathcal{E} > \frac{E}{2k}$ there exist a sequence of $(n, M, \epsilon_n, \mathcal{E}, K = \mu n)$ codes for the AWGN MAC such that*

$$\limsup_{n\to\infty} \epsilon_n \leq 2\epsilon^*(\tau^{(\infty)}(\mu), M)\,,$$

where $\epsilon^*(\tau, M)$ is the solution to

$$\frac{1}{\sqrt{\tau}} = \mathcal{Q}^{-1}(\epsilon^*) + \mathcal{Q}^{-1}\left(\frac{\epsilon^*}{M-1}\right) \tag{3.17}$$

and $\tau^{(\infty)}(\mu) = \mathcal{M}_{\mathsf{AWGN}}(\mu, E, M) \equiv \max(\arg\min_{\tau > \frac{1}{E}} \mathcal{F}_{\mathsf{awgn}}(\tau;\mu,E,M))$.

Proof. The proof is similar to that of theorem 2.3.6 and theorem 2.3.4. In particular, we have

$$\mathrm{PUPE}(\hat{S}^{(t)}) \leq M\mathbb{E}\left[d_H(U, \hat{S}^{(t)}))\right] \tag{3.18}$$

From state evolution, it can be shown that

$$\lim_{n\to\infty} \frac{C}{KM} \sum_{i=(c-1)\frac{KM}{C}+1}^{c\frac{KM}{C}} \mathbb{P}\left[U_i \neq \hat{S}_i^{(t)}\right] = \mathbb{P}\left[X \neq \hat{S}_0\right] \tag{3.19}$$

where $\hat{S}_0 = 1[X + \sqrt{\tau_c^{(t)}}W > \theta_c]$ is the estimator for X is the scalar channel (3.4). Notice that the Bayes' optimal estimator for X is of the form $\hat{S}_0$ for some carefully chosen θ_c.

As in the proof of theorem 2.3.6, we take limit as $t \to \infty$, apply (an equivalent of) lemma 2.3.5 and then optimize over θ_c we obtain

$$\lim_{t\to\infty}\lim_{n\to\infty} \mathrm{PUPE}(\hat{S}^{(t)}) \leq \frac{1}{C}\sum_{c=1}^{C} M\tilde{\epsilon}^*(\tau_c^{(\infty)}, M) \tag{3.20}$$

where $\tilde{\epsilon}^*(\tau, M)$ is the minimum probability of error for decoding X from the scalar channel (3.4): $V = X + \sqrt{\tau}W$. It can be shown that $\tilde{\epsilon}^*(\tau, M)$ satisfies

$$\frac{1}{\sqrt{\tau}} = \mathcal{Q}^{-1}\left(\frac{M\tilde{\epsilon}^*}{2}\right) + \mathcal{Q}^{-1}\left(\frac{M\tilde{\epsilon}^*}{2(M-1)}\right) \tag{3.21}$$

Now define $\epsilon^*(\tau, M) = \frac{M\tilde{\epsilon}^*(\tau,M)}{2}$. Using monotonicity of ϵ^* with respect to τ and the second

item in lemma 2.3.5 (along with threshold saturation) concludes the proof. □

Chapter 4

Numerical evaluation

4.1 Fading MAC: Numerical evaluation and discussion

In this section, we provide the results of numerical evaluation of the bounds in the chapter 2. We focus on the trade-off of user density μ with the minimum energy-per-bit $\mathcal{E}^*$ for a given message size k and target probability of error P_e.

For $k = 100$ bits, we evaluate the trade-off from the bounds in this paper for $P_e = 0.1$ and $P_e = 0.001$ in figures 4-2 and 4-1 respectively. For TDMA, we split the frame of length n equally among K users, and compute the smallest P_{tot} the ensures the existence of a single user quasi-static AWGN code of rate S, blocklength $\frac{1}{\mu}$ and probability of error ϵ using the bound from [19]. The simulations of the single user bound is performed using codes from [163]. TIN is computed using a method similar to theorem 2.3.7. In particular, the codeword of user i is decoded as $\hat{c}_i = \arg\min_{c' \in \mathcal{C}_i} \|Y - H_i c_i\|^2$ where we assume that the decoder has the knowledge of CSI. The analysis proceeds in a similar way as theorem 2.3.7.

Achievability bounds. It can be seen that for small μ the scalar-AMP bound of Theorem 2.3.4 is better than the projection decoder bounds of Theorems 2.3.7 and 2.3.1. The latter bounds have another artifact. For example, the no-CSI bound on $\mathcal{E}^*$ from Theorem 2.3.1 increases sharply as $\mu \downarrow 0$, in fact one can show that the said bound behaves as $\mathcal{E} = \Omega(\sqrt{-\ln \mu})$.

Engineering insights. From these figures, we clearly observe the *perfect MUI cancellation* effect mentioned in the introduction and previously observed for the non-fading model [3, 16]. Namely, as μ increases from 0, the $\mathcal{E}^*$ is almost a constant, $\mathcal{E}^*(\mu, \epsilon, k) \approx \mathcal{E}_{\text{s.u.}}(\epsilon, k)$ for $0 < \mu < \mu_{\text{s.u.}}$. As μ increases beyond $\mu_{\text{s.u.}}$ the tradeoff undergoes a "phase transition" and the energy-per-bit $\mathcal{E}^*$ exhibits a more familiar increase with μ. Further, standard schemes for multiple-access like TDMA and TIN do not have this behavior. Moreover, although these suboptimal schemes have an optimal trade-off at $\mu \to 0$ they show a significant suboptimality at higher μ. We note again that this perfect MUI cancellation

which was observed in standard AWGN MAC [3, 16] is also present in the more practically relevant quasi-static fading model. So, we suspect that this effect is a general characteristic of the many-user MAC.

Suboptimality of orthogonalization. The fact that orthogonalization is not optimal is one of the key practical implications of our work. It was observed before in the GMAC and here we again witness it in the more relevant QS-MAC. How to understand this suboptimality? First, in the fading case we have already seen this effect even in the classical regime (but under PUPE) – see (2.16). To give another intuition we consider a $K = \mu n$ user binary adder MAC

$$Y = \sum_{i=1}^{K} X_i \tag{4.1}$$

where $X_i \in \{0, 1\}$ and addition is over $\mathbb{Z}$. Now, using TDMA on this channel, each user can send at most $n/K = 1/\mu$ bits. Hence the message size is bounded by

$$\log M \leq \frac{1}{\mu}. \tag{4.2}$$

Next, let us consider TIN. Assume $X_i \sim \text{Ber}(1/2)$. For user 1, we can treat $\sum_{i=2}^{\mu n} X_i$ as noise. By central limit theorem, this noise can be approximated as $\sqrt{\frac{1}{4}\mu n}Z$ where $Z \sim N(0, 1)$. Thus we have a binary input AWGN (BIAWGN) channel

$$Y = X_1 + \sqrt{\frac{1}{4}\mu n}Z. \tag{4.3}$$

Therefore, the message size is bounded as

$$\log M \leq nC_{\text{BIAWGN}}\left(1 + \frac{4}{\mu n}\right) \leq \frac{n}{2}\log\left(1 + \frac{4}{\mu n}\right) \to \frac{2}{\mu \ln 2} \tag{4.4}$$

where C_{BIAWGN} is the capacity of the BIAWGN channel. Note that in both the above schemes the achievable message size is a constant as $n \to \infty$.

On the other hand, the true sum-capacity of the K-user adder MAC is given by

$$C_{sum} = \max_{X_1, \cdots, X_K} H(X_1 + \cdots + X_K)\,.$$

As shown in [164] this maximum as achieved at $X_i \overset{iid}{\sim} \text{Ber}(1/2)$. Since the the entropy of binomial distributions [165] can be computed easily, we obtain

$$C_{sum} = \frac{1}{2}\log K + o(\log K)\,.$$

In particular, for our many-user MAC setting we obtain from the Fano inequality (and assuming

PUPE is small)

$$\log M \lessapprox \frac{\log(\mu n)}{2\mu}\,.$$

Surprisingly, there exist explicit codes that achieve this limit and with a very low-complexity (each message bit is sent separately),– a construction rediscovered several times [166–168]. Hence the optimal achievable message size is

$$\log M \approx \frac{\log n}{2\mu} \to \infty \tag{4.5}$$

as $n \to \infty$. And again, we see that TDMA and TIN are severely suboptimal for the many-user adder MAC as well.

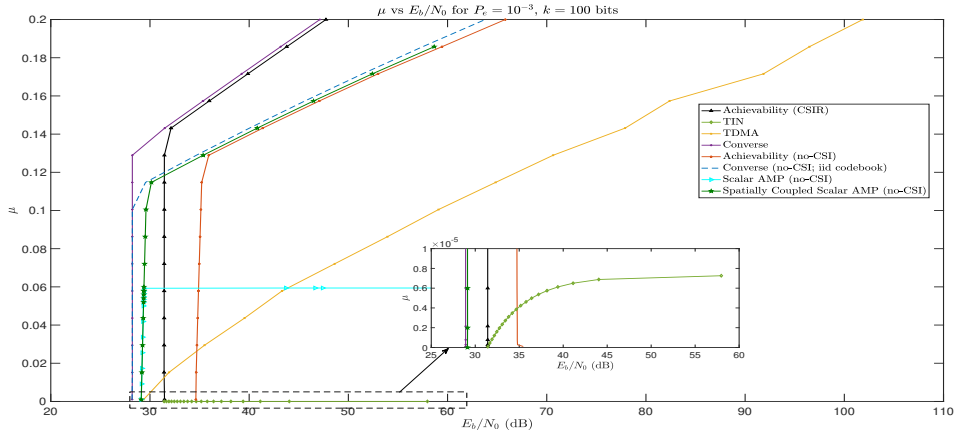

Figure 4-1: Fading Many MAC: μ vs E_b/N_0 for $\epsilon \leq 10^{-3}$, $k = 100$

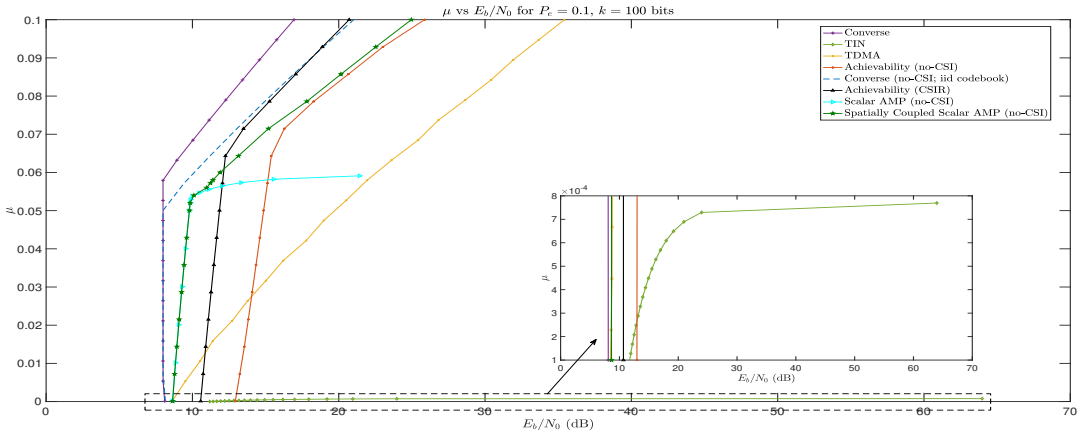

Figure 4-2: Fading Many MAC: μ vs E_b/N_0 for $\epsilon \leq 10^{-1}$, $k = 100$

4.2 Numerical evaluation of AWGN MAC

In this section we present the numerical evaluation of the bound from theorem 3.2.1 and compare it to the main achievability bound from [14].

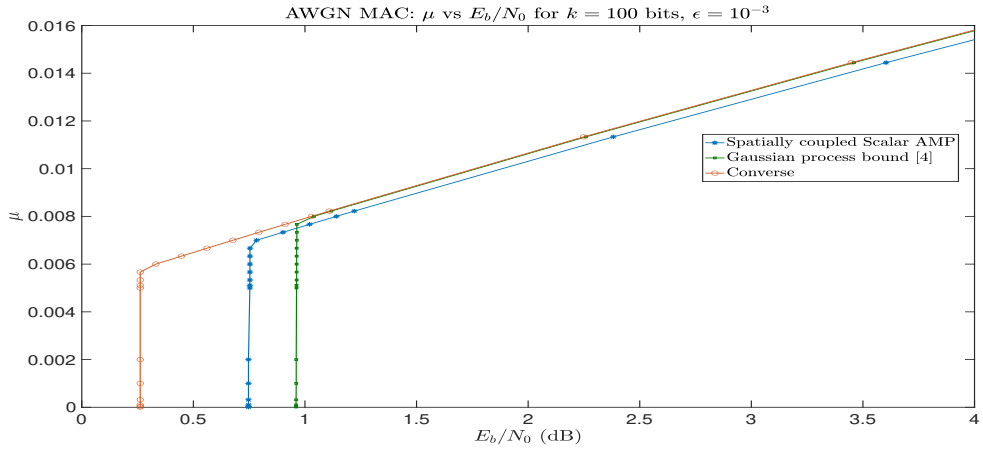

Figure 4-3: AWGN Many-MAC:μ vs E_b/N_0 for $\epsilon \leq 10^{-3}$, $k = 100$

4.3 The "curious behavior" in phase transition

As we emphasized in the Introduction, the most exciting conclusion of our work is the existence of the almost vertical part on the μ vs $\frac{E_b}{N_0}$ plots of Fig. 4-2 and 4-1. In this section we want to explain how this effect arises, why it can be called the "almost perfect MUI cancellation" and how it relates (but is not equivalent) to well-known phase transitions in compressed sensing.

To make things easier to evaluate, however, we depart from the model in the previous sections and do two relaxations. First, we consider a non-fading AWGN. Second, we endow all users with the same codebook. The second assumption simply means that the decoding from now on is only considered up to permutation of messages, see [3] for more on this. Technically, these two assumptions mean that we are considering a model (2.6) with U vector that is K-sparse (as opposed to block-sparse) and that all non-zero entries of U are equal to 1. Finally, we will consider the real-valued channel. In all, we get the following signal model [14, Section IV]:

$$Y^n = AU^p + Z^n\,, \qquad Z^n \sim \mathcal{N}(0, I_n)\,, \tag{4.6}$$

with $A_{i,j} \overset{iid}{\sim} \mathcal{N}(0, b^2/n)$, $(i,j) \in [n] \times [p]$, $U_i \overset{iid}{\sim} \mathrm{Ber}(K/p)$, so that $\mathbb{E}[\|U^p\|_0] = K$. We take the proportional scaling limit with $K = \mu n$ and $p = KM$. Interpretation of these parameters in the context of communication problem are:

- M as the number of messages that each user wantes to communicate
- μ is the user density per (real) degree of freedom
- $b^2 = \frac{P_{tot}}{\mu}$ where P_{tot} is the total received power from all K users at the receiver.

Consequently, we may define energy-per-bit as $\frac{E_b}{N_0} \triangleq \frac{b^2}{2\log M}$.

Given (Y^n, A), the decoder outputs an estimate $\hat{U}^p \in \{0,1\}^p$ with $\mathbb{E}[\left\|\hat{U}^p\right\|_0] = K$ and we are interested in the minimal achievable PUPE, or

$$P_e^*(\mu, M, b) \triangleq \limsup_{n\to\infty} \min_{\hat{U}^p:\mathbb{E}[\|\hat{U}\|_0]=K} \frac{1}{K}\sum_{i\in[p]} \mathbb{P}\left[U_i = 1, \hat{U}_i = 0\right] \tag{4.7}$$

To discuss performance of the optimal decoder, we need to return to the scalar channel (2.89) with the following modifications: $X \sim \mathrm{Ber}(1/M)$, $W \sim \mathcal{N}(0,1)$. Now, for every value of σ in (2.89) we may ask for the smallest possible error $\epsilon^*(\sigma, M) = \min \mathbb{P}\left[\hat{X} \neq X\right]$ where minimum is taken over all estimators $\hat{X} = \hat{X}(V)$ such that $\mathbb{P}\left[\hat{X} = 1\right] = \frac{1}{M}$. As discussed in [14, Section IV.B], this minimal $\epsilon^*(\sigma, M)$ satisfies [14] is found from solving:

$$\frac{1}{\sigma} = Q^{-1}\left(\frac{\epsilon^*}{M-1}\right) + Q^{-1}\left(\epsilon^*\right) \tag{4.8}$$

where $Q(\cdot)$ is the complementary CDF of the standard normal distribution.

Now the limit P_e^* in (4.7) can be computed via the replica method.[1] Namely, replica predictions tell us that

$$P_e^*(b, \mu) = \epsilon^*(\sigma, M)\,,$$

where $\sigma^2 = \frac{1}{\eta^* b^2}$ and the *multiuser efficiency* $\eta^* = \eta^*(M, b, \mu)$ is given by

$$\eta^* = \arg\min_{\eta\in[0,1]} \mu M I\left(\frac{1}{\eta b^2}\right) + \frac{1}{2}(\eta - 1 - \ln\eta) \tag{4.9}$$

where $I(\sigma^2) = I(X; X + \sigma W)$ is the mutual information between the signal and observation in the scalar channel (2.89).

In the figure 4-4 we have shown the plots of optimal PUPE P_e for the model (4.6) versus E_b/N_0 for various values of μ when $M = 2^{100}$, computed via replica predictions. What is traditionally referred to as the phase transition in compressed sensing is the step-function drop from $P_e \approx 1$ to $P_e \ll 1$. However, there is a *second effect* here as well. Namely that all the curves with different μ

[1] Note that in [169, 170] it was shown that the replica-method prediction is correct for estimating $I(U_i; Y^n, A)$ and $\mathrm{Var}[U_i|Y^n, A]$, but what we need for computing the P_e is asymptotic distribution of a random variable $\mathbb{P}[U_i = 1|Y^n, A]$. First, it is known that AMP initialized at the true value U converges to an asymptotically MMSE-optimal estimate. Second, distribution of the AMP estimates are known to belong to a $P_{X|V}$ in (2.89) (with σ identified from the replica method). Finally, any asymptotically MMSE-optimal estimator $\hat{U}$ should satisfy $\hat{U}_i \overset{(d)}{\to} \mathbb{E}[U_i|Y^n, A] = \mathbb{P}[U_i = 1|Y^n, A]$, and thus $\mathbb{P}[U_i = 1|Y^n, A]$ should match the replica-method predicted one.

seem to have a *common envelope*. The former has not only been observed in compressed sensing, e.g. [22, 137, Fig.1] and [171, Fig.4] among others, but also in a number of other inference problems: randomly-spread CDMA [136], LDPC codes [172] and random SAT [173]. However, the second effect appears to be a rather different phenomenon, and in fact it is exactly the one that corresponds to the existence of the vertical part of the curves on Fig. 4-1-4-2.

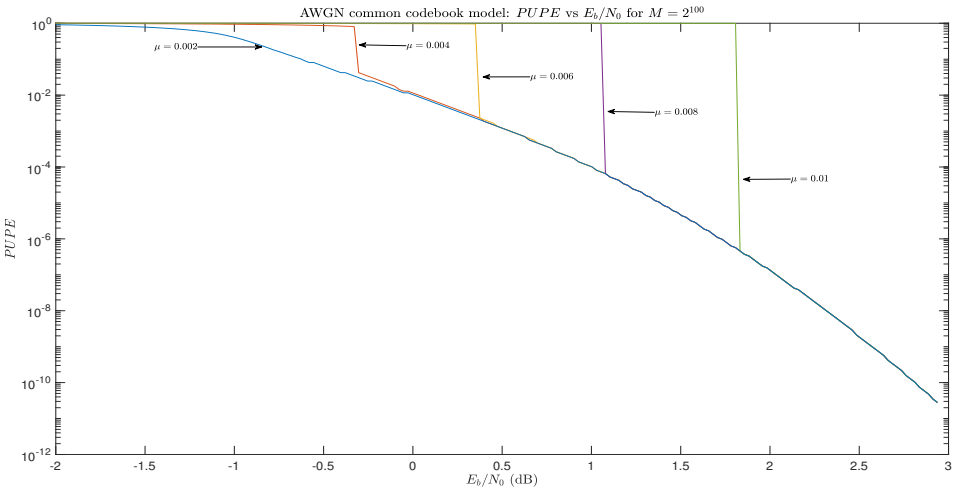

Figure 4-4: AWGN same codebook model: P_e vs E_b/N_0 for $M = 2^{100}$

Let us, for the moment, assume that the envelope is actually exactly the same for all μ. Fix a value of PUPE $P_e = 10^{-3}$ (say) and consider how the intercept of the horizontal line at $P_e = 10^{-3}$ on Fig. 4-4 changes with μ. It is easy to see that as long as the value of μ is small enough the intercept will not be moving (corresponding to constancy of the E_b/N_0 as a function of μ). However, once the value of μ exceeds a value (dependent on the fixed value of P_e) the intercept starts moving to the right together with the step-drop portion of the curves. From this we conclude that indeed, existence of the (almost) common envelope on Fig. 4-4 results in the (almost) vertical part on Fig. 4-1-4-2. (As a side note, we also note that since the slanted portions of the tradeoff curves on those figures correspond to the vertical drop on the Fig. 4-4 and hence the slanted portion is virtually independent of the fixed value of P_e – as predicted by (1.4).)

How can the curves have common envelope? Notice that in the expression for P_e^* only η^* is a function of μ. Thus, we conclude that for small μ we must have $\eta^*(\mu) \approx$ const. But as $\mu \to 0$ we should get a $\eta^* \to 1$. Thus we see that common envelopes are only possible if to the right of the step-drops on Fig. 4-4 we get $\eta^* \approx 1$. Is that indeed so? Fig. 4-5 provides an affirmative answer.

In reality, the "vertical part" is not truly vertical and the common envelope is not exactly common. In truth the right portions of the curves on Fig. 4-4 (following the drop) are all very slightly different, but this difference is imperciptible to the eye (and irrelevant to an engineer).

What makes them so close is the incredible degree of sparsity $\frac{1}{M} = 2^{-100}$. Indeed, as Fig. 4-5 demonstrates that as $M \to \infty$ the value of η^* to the right of the step transition approaches 1.

To summarize, we conclude that what determines our "curious behavior" is not a sudden change in the estimation performance (typically credited as "phase transition" in compressed sensing), but rather a more subtle effect arising in the super-low sparsity limit: the step-transition of the parameter η^* from a moderate value in the interior of $(0, 1)$ to a value close to 1. The fact that only the incredibly low sparsity values $\frac{1}{M}$ are relevant for the many-MAC problems makes this new effect practically interesting.

To close our discussion, we want to further argue that the transition $\eta^* \approx 1$ is not related to the so-called "all-or-nothing" property in the sublinear-sparsity regime, cf. [174]. More formally, that property corresponds to a transition of $\frac{\mathrm{Var}[U_i|Y^n,A]}{\mathrm{Var}[U_i]} \approx 0$ to $\frac{\mathrm{Var}[U_i|Y^n,A]}{\mathrm{Var}[U_i]} \approx 1$ in a certain limiting regime of sparsity $\to 0$. To understand relation to our previous discussion, let us again consider the scalar channel (2.89). We are interested (because of finite $\frac{E_b}{N_0}$) in the following regime:

$$\sigma^2 \ln M \to c\,, M \to \infty$$

Using the approximation

$$Q^{-1}(\delta) \approx \sqrt{2\ln\frac{1}{\delta} - \ln\left(4\pi \ln\frac{1}{\delta}\right)} \tag{4.10}$$

one can easily check that in this scaling

$$\lim_{M\to\infty} \epsilon^*\left(\sigma^2 = \frac{c}{\ln M}, M\right) \to \begin{cases} 0, & c > 1/2 \\ 1, & c < 1/2 \end{cases},$$

which incidentally corresponds to the two cases $\frac{E_b}{N_0} \lessgtr -1.59\ dB$. Since $\epsilon^* \to \{0, 1\}$ is tantamount to normalized MMSE converging to $\{0, 1\}$, we can see that from replica-prediction the all-or-nothing transitions corresponds to whether

$$\lim_{M\to\infty} \frac{\ln M}{\eta^* b^2} \lessgtr \frac{1}{2}$$

and has no immediate bearing on $\lim \eta^* = 1$. It would be interesting to formulate a conjecture that would be morally equivalent to $\eta^* \to 1$ but without going into the unnatural double limit that we discussed above (i.e. first taking $n \to \infty$ with $p = \Theta(n)$ and $K = \Theta(n)$ and then taking sparsity $\frac{1}{M}$ to zero).

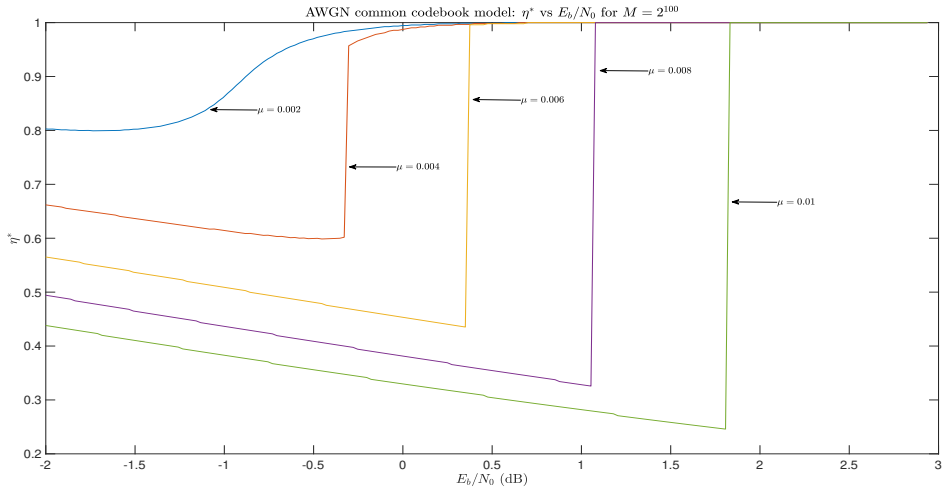

Figure 4-5: AWGN same codebook model: η^* vs E_b/N_0 for $M = 2^{100}$

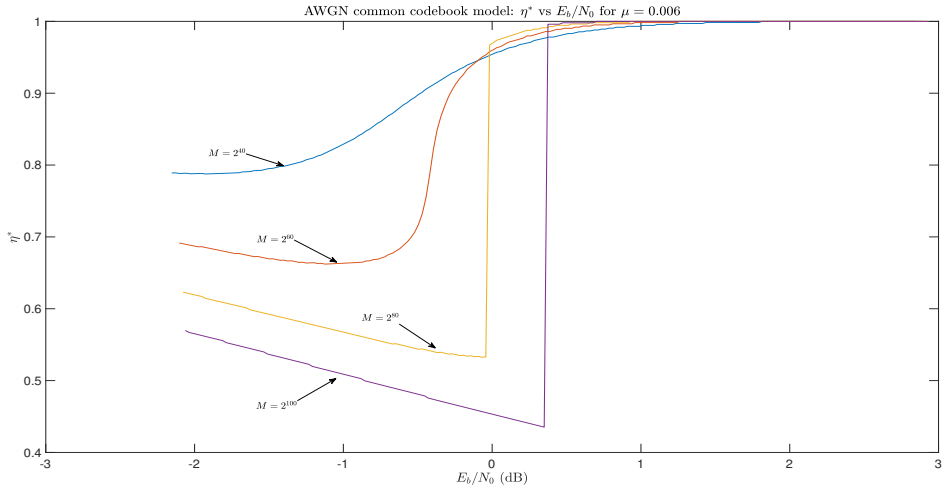

Figure 4-6: AWGN same codebook model: η^* vs E_b/N_0 for $\mu = 0.006$

Chapter 5

Linear system identification

5.1 Problem setting and notation

In this section, we first introduce the data generation model, the required assumptions and then provide the precision problem definition. Throughout the paper, we use $\|A\|$ to denote the operator norm of A unless otherwise specified. $\|A\|_F$ denotes the Frobenius norm of A. $\sigma_i(A)$ denotes the i-th largest singular value of A, i.e., $\sigma_{\max}(A) = \sigma_1(A)$. $\kappa(A) := \sigma_{\max}(A)/\sigma_{\min}(A)$ denotes the condition number of A. $\rho(A)$ denotes the spectral radius of A. For two symmetric matrices $A, B \in \mathbb{R}^{d\times d}$ we say $A \preceq B$ if $B - A$ is positive semidefinite (psd). For notational simplicity, we use C to denote a constant, and it's value can be different in different equations.

Linear dynamical system/VAR(1) model. Given an initial (possibly random) data point X_0 which is independent of the noise sequence, we generate the $(X_0, \ldots, X_T)$ from the VAR model as:

$$X_{\tau+1} = A^* X_\tau + \eta_\tau, \quad 0 \le \tau \le T-1, \tag{5.1}$$

where $A^* \in \mathbb{R}^{d\times d}$ be the transition matrix. Let $\eta_1, \ldots, \eta_T \in \mathbb{R}^d$ be an i.i.d noise sequence with 0 mean and finite second moment with probability measure μ. We will denote this model by $\mathsf{VAR}(A^*, \mu)$. We also make the following assumptions about A^*, μ, and X_0:

Assumption 1. ***External Stability.*** $\|A^*\| < 1$

Assumption 2. ***Sub-Gaussian Noise.*** *μ has co-variance Σ and for all $x \in \mathbb{R}^d$, $\langle x, \eta_\tau\rangle$ is $C_\mu\langle x, \Sigma\cdot x\rangle$ sub-Gaussian. Further, Σ is full rank. Also, let $\mu_4 := \mathbb{E}\left[\|\eta_\tau\|^4\right]$ be the fourth moment of the noise.*

Assumption 3. ***Stationarity.*** *$X_0 \sim \pi$, the stationary distribution corresponding to (A^*, μ). Let $M_4 := \mathbb{E}\left[\|X_0\|^4\right]$.*

Due to Assumption 1, we can show that the law of the iterate X_T from the VAR model defined above converges to a stationary distribution π as $T \to \infty$ for arbitrary choice of X_0 and has a mixing time of the order $\tau_{\mathsf{mix}} = O\left(\frac{1}{1-\|A^*\|}\right)$. For simplicity, we will absorb C_μ into other constants. Finally, we will use $(Z_0, \ldots, Z_T) \sim \mathsf{VAR}(A^*, \mu)$ to mean that $Z_0, \ldots, Z_T$ is a stationary sequence corresponding to $\mathsf{VAR}(A^*, \mu)$. We also note that the covariance matrix under stationarity, $G := \mathbb{E}_{X\sim\pi} XX^\top = \sum_{s=0}^{\infty} A^{*s}\Sigma(A^{*\top})^s \succeq \Sigma$.

Remark 8. *It is indeed possible to replace Assumption 1 with the weaker condition on the spectral radius of A^*: $\rho(A^*) < 1$. While our results still hold in this case, the bound might have additional condition number factors. See Section 5.3.1 for more details.*

Remark 9. *The full rank assumption on Σ is needed for polynomial sample complexity [55].*

Problem statement. Let $(X_0, X_1, \cdots, X_T)$ be sampled from $\mathsf{VAR}(A^*, \mu)$ model for a fixed horizon T. Then, the goal is to design and analyze an online algorithm that uses only first order gradient oracle to estimate the system matrix A^*. That is, at each time-step τ, we obtain gradient for the transition $(X_\tau, X_{\tau+1})$ and output estimate A_τ. The goal is to ensure that each A_τ has small estimation error wrt A^*; naturally, we would expect better estimation error with increasing τ. We quantify estimation error using the following two loss functions:

1. Parameter error: $\mathcal{L}_{\mathsf{op}}(A; A^*, \mu) = \|A - A^*\|$
2. Prediction error at stationarity: $\mathcal{L}_{\mathsf{pred}}(A; A^*, \mu) := \mathbb{E}_{X_\tau\sim\pi}\|X_{\tau+1} - AX_\tau\|^2$

Note that the problem is equivalent to d linear regression problems, but with *dependent* samples, making it significantly more challenging. Whenever Assumption 1 holds, stationary distribution π exists, so the prediction error $\mathcal{L}_{\mathsf{pred}}$ is meaningful. Furthermore: $\mathcal{L}_{\mathsf{pred}}(A) - \mathcal{L}_{\mathsf{pred}}(A^*) = \mathrm{Tr}\left[(A - A^*)^\top (A - A^*)G\right]$ where $G := \mathbb{E}_{X\sim\pi} XX^\top$.

5.2 Algorithm

As mentioned in related works, the standard OLS estimator that minimizes the empirical loss is known to be nearly optimal in the *offline setting* [51]:

$$\hat{A}_{OLS} = \arg\min_A \sum_{\tau=0}^{T-1} \|AX_\tau - X_{\tau+1}\|^2 . \tag{5.2}$$

Note that for least squares loss, one can indeed maintain covariance matrix and residual vector to compute the OLS solution *online*. But such a solution does not work if we have access to only gradients and breaks down even for generalized linear models, whereas as the techniques introduced in this work has been extended to non-linear systems [63].

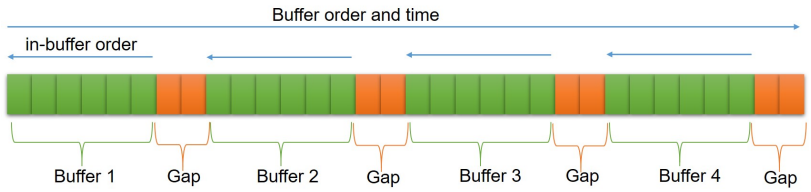

Figure 5-1: Data Processing Order in SGD – RER. A cell represents a data point. Time goes from left to right, buffers are also considered from left to right. Within each buffer, the data is processed in the reverse order. Gaps ensure that data in successive buffers are approximately independent.

On the other hand, using standard SGD we can obtain update to A efficiently by using gradient at the current point. That is, assuming $A_0 = 0$, we get the following SGD update (for all $\tau \geq 0$):

$$A_{\tau+1} = A_\tau - 2\gamma(A_\tau X_\tau - X_{\tau+1})X_\tau^\top, \tag{5.3}$$

where γ is the stepsize. While SGD is known to be an optimal estimator in certain streaming problems with i.i.d. data, for the $\mathsf{VAR}(A^*, \mu)$ problem the standard SGD does not apply, as samples $(X_\tau, X_{\tau+1})$ and $(X_{\tau+1}, X_{\tau+2})$ are highly correlated. To see why this is the case, let us unroll the recursion for two steps and using Equation (5.1):

$$A_2 - A^* = (A_0 - A^*)(I - 2\gamma X_0 X_0^\top)(I - 2\gamma X_1 X_1^\top) + 2\gamma\eta_1 X_1^\top + 2\gamma\eta_0 X_0^\top (I - 2\gamma X_1 X_1^\top).$$

Note that the last term does not have 0 mean because X_1 depends on η_0 by Equation (5.1). Even in the case when $A_0 = A^*$, this means that $\mathbb{E}A_2 \neq A^*$ in general. In fact, in Section 7.1, we show empirically that SGD with constant step-size converges to a significantly larger error than OLS, even when T is very large. This shows that we cannot naively treat this problem as a collection of d linear regressions. This is consistent with the results in [66, 67] which show a similar behavior for constant step-size SGD with dependent data. Now, one can use techniques like *data drop* that drops a large fraction of points (either explicitly or during the mathematical analysis) from the stream to obtain nearly independent samples [66, 175], but such methods waste a lot of samples and have significantly suboptimal error rate than OLS.

So, the goal is to design a streaming method for the problem of learning dynamical systems that at each time-step t provides an accurate estimate of A^*, while also ensuring small space+time complexity.We now present a novel algorithm that addresses the above mentioned problem.

5.2.1 SGD with Reverse Experience Replay

We now discuss a novel algorithm called SGD with Reverse Experience Replay (SGD – RER) that addresses the problem of learning stationary auto-regressive models (or linear dynamical systems) in the streaming setting. Our method is inspired by the experience replay technique [128], used

extensively in RL to break temporal correlations between dependent data. We make the following crucial observation. Suppose in Equation (5.3), instead of processing the samples in the order $(X_1, X_2) \to (X_2, X_3) \to \cdots \to (X_{T-1}, X_T)$, we process it in the reverse order. That is: $(X_{T-1}, X_T) \to (X_{T-2}, X_{T-1}) \to \cdots \to (X_1, X_2)$. Then,

$$A_2 - A^* = (A_0 - A^*)(I - 2\gamma X_{T-1}X_{T-1}^\top)(I - 2\gamma X_{T-2}X_{T-2}^\top) + 2\gamma\eta_{T-2}X_{T-2}^\top \\ +2\gamma\eta_{T-1}X_{T-1}^\top(I - 2\gamma X_{T-2}X_{T-2}^\top) \tag{5.4}$$

Now, observe that (X_{T-2}, X_{T-1}) are *independent* of η_{T-1}. Therefore the problematic last term, $2\gamma\eta_{T-1}X_{T-1}^\top(I - 2\gamma X_{T-2}X_{T-2}^\top)$, now has expectation 0. So the updates for *reverse* order SGD would be *unbiased*. This, however, requires us to know all the data points beforehand which is infeasible in the streaming setting. We alleviate this issue by designing $\mathsf{SGD-RER}$, which is the online variant of the above algorithm. $\mathsf{SGD-RER}$ uses a buffer of large enough size to store values of consecutive data points and then performs reverse SGD in each of these buffers and then discards this buffer. Experience replay methods also use such (small) buffers of data, but typically samples point randomly from the buffer instead of the reverse order that we propose. We refer to Figure 6-1 for an illustration of the proposed data processing order.

We present a pseudocode of $\mathsf{SGD-RER}$ in Algorithm 1. Note that the algorithm forms non-overlapping buffers of size $S = B + u$. Here B is the actual size of the buffer while u samples are used to interleave between two buffers so that the buffers are *almost independent* of each other. Now within a buffer, we perform the usual SGD but with samples read in reverse order. Formally, suppose we index our buffers by $t = 0, 1, 2, \cdots$ and let $S = B + u$ be the total samples (including those that were dropped) in the buffers. Let N denote the total number of buffers in horizon T. Within each buffer t, we index the samples as X_i^t where $i = 0, 1, 2, \cdots, S-1$. That is $X_i^t \equiv X_{tS+i}$ is the i-th sample in buffer t. Similarly $\eta_i^t \equiv \eta_{tS+i}$. Further let $X_{-i}^t \equiv X_{(S-1)-i}^t$. Similarly we set $\eta_{-i}^t \equiv \eta_{(S-1)-i}^t$ Then, the algorithm performs the recursion stated in Line 1 of Algorithm 1. Note that the recursion can also be written as,

$$A_{i+1}^{t-1} - A^* = \left(A_i^{t-1} - A^*\right)\left(I - 2\gamma X_{-i}^{t-1}{X_{-i}^{t-1}}^\top\right) + 2\gamma\eta_{-i}^{t-1}X_{-i}^{t-1}. \tag{5.5}$$

for $1 \le t \le N$ and $0 \le i \le B-1$ with $A_0^t = A_B^{t-1}$ and $A_0^0 = A_0$.

We then ignore the iterates corresponding to first a buffers as part of the *burn-in period*, and output average of the remaining iterates ($t > a$) at each step as that step's estimator (see Line 2 of Algorithm 1). That is, we have the tail-averaged iterate:

$$\hat{A}_{a,t} = \frac{1}{t-a}\sum_{\tau=a+1}^{t} A_B^{\tau-1}. \tag{5.6}$$

Algorithm 1: SGD − RER

Input : Streaming data $\{X_\tau\}$, horizon T, buffer size B, buffer gap u, bound R, tail average start: a

Output: Estimate $\hat{A}_{a,t}$, for all $a < t \leq N-1$; $N = T/(B+u)$

begin
  Step-size: $\gamma \leftarrow \frac{1}{8RB}$, Total buffer size: $S \leftarrow B+u$, Number of buffers: $N \leftarrow T/S$
  $A_0^0 = 0$ /*Initialization*/
  for $t \leftarrow 1$ **to** N **do**
    Form buffer $\mathsf{Buf}^{t-1} = \{X_0^{t-1}, \ldots, X_{S-1}^{t-1}\}$, where, $X_i^{t-1} \leftarrow X_{(t-1)\cdot S+i}$
    If $\exists i,\ s.t.,\ \left\|X_i^{t-1}\right\|^2 > R$, then **return** $\hat{A}_{a,t} = 0$
    for $i \leftarrow 0$ **to** $B-1$ **do**
      $A_{i+1}^{t-1} \leftarrow A_i^{t-1} - 2\gamma(A_i^{t-1}X_{S-1-i}^{t-1} - X_{S-i}^{t-1})\left(X_{S-1-i}^{t-1}\right)^\top$
    end
    $A_0^t = A_B^{t-1}$
    If $t > a$, then $\hat{A}_{a,t} \leftarrow \frac{1}{t-a}\sum_{\tau=a+1}^{t} A_B^{\tau-1}$
  end
end

We output the new iterate $\hat{A}_{a,t}$ only at the end of each buffer t. At intermediate steps, $(t-1)B+1 \leq \tau \leq tB$, we output $\hat{A}_{a,t-1}$. Also, note that the tail average can be computed in small space and time complexity, by using a running sum of the tail iterates. The update for each point is rank-one, so can be computed in time linear in number of parameters ($O(d^2)$). In the next section, we show that despite using small buffer size $S = B+u$ (that depends logarithmically on T), and by throwing away a small constant–independent of *any* problem parameter–fraction of points u in each buffer, we are still able to provide error bound similar to that of OLS.

5.3 Main results

We now state our main results with leading order terms. Recall the problem setting, and the covariance matrix $G := \mathbb{E}_{X\sim\pi} XX^\intercal$. Before stating the results, we choose the parameters B, R, α and u as follows, which can be estimated using upper bounds on $\|A^*\|$:

1. $d \leq \mathrm{Poly}(T)$. We use this to bound the norm of covariates in the next item.
2. $\alpha \geq 22$; $R \geq C(\alpha)\frac{\mathrm{Tr}(\Sigma)\log T}{1-\|A^*\|^2} = O(d\tau_{\mathsf{mix}}\log T)$ s.t. $\mathbb{P}\left[\|X_\tau\|^2 \leq R,\ \tau \leq T\right] \geq 1 - \frac{1}{T^\alpha}$. See lemma 5.5.3 in appendix.
3. $u \geq \alpha\frac{\log T}{\log\left(\frac{1}{\|A^*\|}\right)} = O(\tau_{\mathsf{mix}}\log T)$; $B = 10u$

For all the results below, we suppose that Assumptions 1, 2 and 3 hold, the stream of samples X_τ is sampled from $\mathsf{VAR}(A^*, \mu)$ model described in Section 5.1 and that R, B, α and u are chosen as above.

Let $t > a$ and let $\hat{A}_{a,t}$ be the tail averaged output of SGD − RER after buffer $t-1$. Further let $T^{\alpha/2} > cd\kappa(G)$.

Theorem 5.3.1. *Suppose we pick the step size* $\gamma = \min\left(\frac{C}{B\sigma_{\min}(G)}, \frac{1}{8BR}\right)$ *for some constant* C *depending only on* C_μ. *Then, there are constants* $C, c_i > 0$, $0 \le i \le 4$ *such that if* $a > c_0\,(d + \alpha \log T)$ *then with probability at least* $1 - \frac{C}{T^\alpha}$, *we have:*

$$\mathcal{L}_{\mathsf{op}}(\hat{A}_{a,t}, A^*, \mu) \le c_1 \sqrt{\frac{(d + \alpha \log T)\sigma_{\max}(\Sigma)}{(t-a)B\sigma_{\min}(G)}} + \beta_b \left\|A_0 - A^*\right\| + c_4 \frac{T^2}{B^2} \left\|A^{*u}\right\| \tag{5.7}$$

where

$$\beta_b = c_3 \frac{d\kappa(G) \log T}{t-a} e^{-c_2 \frac{a}{d\kappa(G)\log T}} \tag{5.8}$$

The techniques for the proof is developed in Section 5.7 and the Theorem 5.3.1 is proved in Section 5.8.

Theorem 5.3.2. *Let* R, B, u, α *be chosen as in section 5.3. Let* $\gamma = \frac{c}{4RB} \le \frac{1}{2R}$ *for* $0 < c < 1$. *Then there are constants* $c_1, c_2, c_3, c_4 > 0$ *such that for* $T^{\alpha/2} > c_1 \frac{\sqrt{M_4}}{\sigma_{\min}(G)}$ *the expected prediction loss* $\mathcal{L}_{\mathsf{pred}}$ *is bounded as*

$$\begin{aligned}
\mathbb{E}\left[\mathcal{L}_{\mathsf{pred}}(\hat{A}_{a,t}; A^*, \mu)\right] - \operatorname{Tr}(\Sigma) \le c_2 \Bigg[& \frac{d \operatorname{Tr}(\Sigma)}{B(t-a)} + \frac{d^2 \sigma_{\max}(\Sigma)}{B(t-a)} \frac{\sqrt{\kappa(G)}}{B} \Bigg] + \\
c_3 \Bigg[& \frac{d^2 \sigma_{\max}(\Sigma)}{B^2 (t-a)^2} (\kappa(G))^{3/2} dB \log T + \\
& \beta_b \operatorname{Tr}(G) \left\|A_0 - A^*\right\|^2 + \\
& \left(\frac{T^3}{B^3} \left\|A^{*u}\right\| + \frac{d\sigma_{\max}(\Sigma)}{R} \frac{T^2}{B^2} \frac{1}{T^{\alpha/2}} \right) \operatorname{Tr}(G) \Bigg]
\end{aligned} \tag{5.9}$$

where β_b *is defined in* (5.8).

The above theorem is proven only for the case $t = N$. The proof for general t is almost the same. The proof follows by first considering $\mathbb{E}\left[\mathcal{L}_{\mathsf{pred}}(\hat{A}_{a,N}; A^*, \mu) 1\left[\mathcal{D}^{0,N-1}\right]\right]$ ($\mathcal{D}^{0,N-1}$ is defined in 5.5.2) and using theorem 5.10.5 and theorem 5.10.6 along with lemma 5.5.6 in the appendix sections 5.10.1, 5.10.2 and 5.5.3. Then noting that if the norm of any of the covariates X_t exceed $\sqrt{R}$ the algorithm returns the zero matrix we have that $\mathbb{E}\left[\mathcal{L}_{\mathsf{pred}}(\hat{A}_{a,N}; A^*, \mu) 1\left[\mathcal{D}^{0,N-1,C}\right]\right] \le c \left\|A^*\right\| \operatorname{Tr}(G) \frac{1}{T^\alpha}$.

Remark 10.

(1) In theorem 5.3.2 the term $\frac{d^2\sigma_{\max}(\Sigma)}{B(t-a)} \frac{\sqrt{\kappa(G)}}{B}$ *is strictly a lower order term compared to* $\frac{d \operatorname{Tr}(\Sigma)}{B(t-a)}$ *when* $\|A^*\| < c_0 < 1$. *To see this note that* $\sigma_{\max}(G) \le \frac{\sigma_{\max}(\Sigma)}{1 - \|A^*\|^2}$ *and* $\sigma_{\min}(G) \ge \sigma_{\min}(\Sigma)$. *Hence* $\kappa(G) \le \frac{\kappa(\Sigma)}{1-\|A^*\|^2} = O(\tau_{\mathsf{mix}} \kappa(\Sigma))$. *By the choice of* B *in the section 5.3 we see that* $\frac{\sqrt{\kappa(G)}}{B} = o(1)$ *and it* does not depend on condition number of A^*.

(2) If $a = \Omega\left(d\kappa(G) \left(\log T\right)^2\right)$ *the* β_b *is a lower order term. Further choosing* u *and* α *as in section 5.3 we see that the terms depending on* $\|A^{*u}\|$ *and* $\frac{1}{T^{\alpha/2}}$ *are strictly lower order.*

(3) Thus for the choice of a as in the previous remark such that $a < (1+c)t$ (for some $c > 0$), we get minimax optimal rates: $\frac{d\,\mathrm{Tr}(\Sigma)}{Bt}$ for $\mathcal{L}_{\mathsf{pred}}$ and up to log factors, $\sqrt{\frac{d\sigma_{\max}(\Sigma)}{T\sigma_{\min}(G)}}$ for $\mathcal{L}_{\mathsf{op}}$

We now make the following observations:

(1) The dominant term in our bound on $\mathcal{L}_{\mathsf{op}}$ (Theorem 5.3.1) matches the information theoretically optimal bound (up to logarithmic factors) for the $\mathsf{VAR}(A^*, \mu)$ estimation problem [51] as long as $\|A^*\| \leq 1 - \frac{1}{T^\xi}$ for $\xi \in (0, 1/2)$. Note that despite working with dependent data, leading term in our error bound is nearly independent of mixing time τ_{mix}. In contrast, most of the existing streaming/SGD style methods for dependent data have strong dependence on τ_{mix} [66].

(2) SGD for linear regression with *independent* data [65, 176], but with similar problem setting incurs error $O(\frac{d\,\mathrm{Tr}(\Sigma)}{T})$ for $\mathcal{L}_{\mathsf{pred}}$. So our bound for $\mathsf{SGD-RER}$ matches the independent data setting bound in the minimax sense.

(3) The space complexity of our method is $O(Bd + d^2)$ where $B = O(\tau_{\mathsf{mix}} \log T)$ is independent of d and only logarithmically dependent on T.

(4) **Sparse matrices with known support**: Suppose A^* is known to be sparse and *we know the support* (say by running L_1 regularized OLS on a small set of samples). Let s_j denote the sparsity of row j of A^*. Then the $\mathsf{SGD-RER}$ algorithm can be modified to run row by row such that it operates only on the support of row j. That is the covariates can be projected onto the support of each row. Then it can be shown that the prediction error is bounded as $O\left(\sum_{j=1}^d \sigma_j^2 s_j / T\right)$ where σ_j^2 is the j-th diagonal entry of Σ. Note that $\mathsf{SGD-RER}$ requires only $O(|\mathsf{supp}(A^*)|)$ operations per iteration while applying online version of standard OLS would require $O(d^2)$ operations. In the simple case of $\Sigma = \sigma^2 I$, we note that $G \succeq \sigma^2 I$ and hence the bound for $\mathcal{L}_{\mathsf{pred}}$ becomes $O\left(\frac{|\mathsf{supp}(A^*)|}{T}\right)$. We refer to Section 5.11 for a sketch of this extension.

Next, we show that our error bounds are nearly information theoretically optimal. For the lower bound on $\mathcal{L}_{\mathsf{op}}$ we directly use [51, Theorem 2.3].

Theorem 5.3.3. *Let $\rho < 1$ and $\delta \in (0, 1/4)$. Let μ be the distribution $\mathcal{N}(0, \sigma^2 I)$. For any estimator $\hat{A} \in \mathcal{F}$, there exists an matrix $A^* \in \mathbb{R}^{d\times d}$ where $A^* = \rho O$ for some orthogonal matrix O such that $|\sigma_{\max}(A^*)| = \rho$ and we have that with probability at least δ:*

$$\|\hat{A} - A^*\| = \Omega\sqrt{\frac{(d + \log(1/\delta))(1-\rho)}{T}}. \tag{5.10}$$

Notice that in the setting of Theorem 5.3.3, we have $G = \sum_{i=0}^\infty \sigma^2 (A^*)^i (A^*)^{i,\top} = \frac{\sigma^2}{1-\rho^2} I$. Therefore, $\sigma_{\min}(G) = \frac{1}{1-\rho^2} \sim \frac{1}{1-\rho}$. The bound in Theorem 5.3.1 matches the above minimax bound up to logarithmic factors.

Next we consider the prediction loss. We fix dimension d and horizon T and consider the class of VAR models $\mathcal{M}$ such that Assumptions 1, 2, and 3 hold such that $\mathsf{Tr}(\Sigma(\mu)) = \beta \in \mathbb{R}^+$ be fixed. Let $\mathcal{F}$ be the class of all estimators for parameter A^* given data $(Z_0, \ldots, Z_T)$. We want to lower

bound the minimax error:

$$\mathcal{L}_{\mathsf{minmax}}(\mathcal{M}) := \inf_{f \in \mathcal{F}} \sup_{(A^*,\mu)\in\mathcal{M}} \mathbb{E}_{(Z_t)\sim \mathsf{VAR}(A^*,\mu)} \mathcal{L}_{\mathsf{pred}}(f(Z_0,\dots,Z_T); A^*, \mu) - \mathcal{L}_{\mathsf{pred}}(A^*; A^*, \mu).$$

Theorem 5.3.4. *For some universal constant c, we have:*

$$\mathcal{L}_{\mathsf{minmax}}(\mathcal{M}) \geq c\beta(d-1)\min\left(\frac{1}{T}, \frac{1}{d^2}\right), \text{ where } \beta = \mathrm{Tr}(\Sigma(\mu)).$$

Note that the theorem shows that our algorithm is minimax optimal with respect to the prediction loss at stationarity, $\mathcal{L}_{\mathsf{pred}}$. The theorem follows from [122, Theorem 4].

5.3.1 Spectral gap condition

In Assumption 1, we could have used the more general spectral radius condition $\rho(A^*) = \sup_i |\lambda_i(A^*)| < 1$ rather than the one on the operator norm. We have the Gelfand formula for spectral radius which shows that $\lim_{k\to\infty} \|A^{*k}\|^{1/k} = \rho(A^*)$. Now, if A^* is such that $\rho(A^*) < 1$ but $\|A^*\| > 1$ (a case studied by [51]), then we need to make u as large as $Cd\log T$ which would lead to a relatively large buffer size B of $d\log T$. To see this, we verify the proof by [177] (by replacing A with $\frac{A}{\|A\|}$ and $\rho(A)$ with $\frac{\rho}{\|A\|}$ in the proof) to show that $\|A^{*k}\| \leq (2k\|A^*\|)^d \rho^{k-d}$ whenever $k \geq d$. Therefore, in the worst case, we can pick $u = O\big(\left(\log\left(T\sigma_{\max}(G)\right) + d\log d\|A\|\right)/\log 1/\rho\big)$.

In the case of $\rho < 1$ but $\|A^*\| > 1$, $\kappa(G)$ can grow super linearly in d. For instance, consider A^* to be nilpotent of order d (i.e. $A^{*d-1} \neq 0$ but $A^{*d} = 0$). Here $\sigma_{\max}(G)$ can grow like $\|A^*\|^d$. So we need exponentially (in d) many samples for bias decay. However, in many cases of interest (ex: symmetric matrices, normal matrices etc) the spectral radius is the same as the operator norm.

5.4 Idea behind the proofs

In this section, we provide an overview of the key techniques to prove our results. As observed in the discussion following Equation (5.4), when the data is processed in the reverse order within a buffer, it behaves similar to SGD for linear regression with i.i.d. data. Due to the gaps of size u, we can take the buffers to be approximately independent. Therefore, we analyze the algorithm as follows:

1. Analyze reverse order *within* a buffer using the property noted in Equation (5.4).
2. Treat *different* buffers to be i.i.d. due to gap and present an i.i.d data type analysis.

To execute the proposed proof strategy, we introduce the following technical notions:

Coupled Process. For the real data points (X_τ), the points in different buffers are *weakly* dependent. In order to make the analysis straight forward, we introduce the *fictitious* coupled process $\tilde{X}_\tau$ such that $\left\|\tilde{X}_\tau - X_\tau\right\| \lesssim \frac{1}{T^\alpha}$ for large enough α, for every data point X_τ used by $\mathsf{SGD-RER}$. We

have the additional property that the successive buffers are actually independent for this coupled process. We refer to Definition 2 in the appendix for the construction of the coupled process $\tilde{X}_\tau$.

Suppose we run $\mathsf{SGD} - \mathsf{RER}$ with the coupled process $\tilde{X}_\tau$ instead of X_τ to obtain the coupled iterates $\tilde{A}_i^t$. We can then show that $\tilde{A}_i^t \approx A_i^t$. Thus it suffices analyze the coupled iterates $\tilde{A}_i^t$. We refer to Sections 5.5.1 and 5.5.3 for the details.

Bias variance decomposition. We consider the standard bias variance decomposition with individual buffers as the basic unit as opposed to individual data points. We refer to Section 5.6 for the details. We decompose the error in the iterates into the bias part $\left(\tilde{A}_B^{t-1,b} - A^*\right) = (A_0 - A^*)\prod_{s=0}^{t-1}\tilde{H}_{0,B-1}^s$ and the variance part $\left(\tilde{A}_B^{t-1,v}\right) = 2\gamma\sum_{r=1}^{t}\sum_{j=0}^{B-1}\eta_{-j}^{t-r}\tilde{X}_{-j}^{t-r,\top}\tilde{H}_{j+1,B-1}^{t-r}\prod_{s=r-1}^{1}\tilde{H}_{0,B-1}^{t-s}$ where the matrices $\tilde{H}_{0,B-1}^s = \prod_{i=0}^{B-1}\left(I - 2\gamma\tilde{X}_{-i}^s\tilde{X}_{-i}^{s,\top}\right)$ are the independent 'contraction' matrices associated with each buffer s. This result in the geometric decay of the initial distance between $(A_0 - A^*)$. The variance part is due to the inherent noise present in the data. In Section 5.9.2 we first establish the exponential decay of the 'bias'. We then consider the second moment of the variance term. Observe that the distinct terms in the expression for $\left(\tilde{A}_B^{t-1,v}\right)$ are uncorrelated either due to reverse order *within* a buffer as noted in Equation (5.4) or due to independence between the data in distinct buffers (due to coupling). This allows us to split the second moment into diagonal terms with non-zero mean and cross terms with zero mean. Diagonal terms are analyzed via a recursive argument in Claim 10 and the following discussion in order to remove dependence on mixing time factors. The analysis for parameter recovery (the result of Theorem 5.3.2) is similar but we bound the relevant exponential moments using sub-Gaussianity of the noise sequence η_t to obtain high-probability bounds which when combined with standard ϵ-net arguments give us guarantees for the operator norm error $\mathcal{L}_{\mathsf{op}}$.

Averaged iterates. We then combine the bias and variance bounds obtained for individual iterates in Section 5.9.2 to analyze the tail averaged output. Using techniques standard in the analysis of SGD for linear regression, we finally show that this averaging leads error rates of the order $\frac{d^2}{T}$. We refer to Sections 5.8 (for parameter recover) and 5.10 (for prediction error) for the detailed results.

Picking the step sizes and conditioning. Due to the auto-regressive nature of the data generation, the iterates can grow to be of the size $O(\frac{d}{1-\rho})$. The step sizes need to be set small enough so that the $\gamma\|X_\tau X_\tau^\top\| \leq 1$ in order for the $\mathsf{SGD} - \mathsf{RER}$ iterations to not diverge to infinity. In the statement of Theorem 5.3.2, we condition on the event where $\|X_\tau\|^2$ are all bounded by a sufficiently large number R for every τ in order to ensure this property. The relevant events where the norm is bounded are defined in Section 5.5.1. Conditioning on these events results in previously zero mean terms to be not zero mean. Routine calculations using triangle inequality and Cauchy-Schwarz in-

equality ensure that the means are still of the order $\frac{1}{T^\alpha}$ for any fixed constant $\alpha > 0$. Furthermore, we actually require step sizes such that $\gamma \left\| \sum_{\tau \in \text{Buffer}} X_\tau X_\tau^\top \right\| \leq 1$ to show exponential contraction of $\tilde{H}^s_{0,B-1}$ matrices due to the Grammian G as described next.

Probabilistic results. We establish some properties of $\tilde{H}^s_{0,B-1}$, which are products of dependent random matrices in Section 5.7. Specifically we refer to Lemmas 6.9.3, 5.7.3, 6.9.2, and 5.7.5 which establish that $\left\| \prod_{s=0}^{t-1} \tilde{H}^s_{0,B-1} \right\| \lesssim (1 - \gamma B \sigma_{\min}(G))^t$ with high probability.

5.5 Preliminaries for the proofs

5.5.1 Basic lemmata

Since the covariates $\{X_\tau\}_{\tau \leq T}$ are correlated, we will introduce a coupled process such that we have independence across buffers and that Euclidean distance between the covariates of the original process and the coupled process can be controlled.

Remark 11. *Note that the coupled process is imaginary and we do not actually run the algorithm with the coupled process. We construct it to make the analysis simple by first analyzing the algorithm with the imaginary coupled process and then showing that the output of the actual algorithm cannot deviate too much when run with the actual data.*

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

1. *For each buffer t generate, independently of everything else, $\tilde{X}^t_0 \sim \pi$, the stationary distribution of the $\mathsf{VAR}(A^*, \mu)$ model.*
2. *Then, each buffer has the same recursion as eq* (5.1)*:*

$$\tilde{X}^t_{i+1} = A^* \tilde{X}^t_i + \eta^t_i,\ i = 0, 1, \cdots S - 1, \tag{5.11}$$

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

With this definition, we have the following lemma:

Lemma 5.5.1. *For any buffer t, $\|X^t_i - \tilde{X}^t_i\| \leq \|A^{*i}\| \|X^t_0 - \tilde{X}^t_0\|$, a.s.. That is,*

$$\|X^t_i X^{t^T}_i - \tilde{X}^t_i \tilde{X}^{t^T}_i\| \leq 2 \|X\| \|X^t_i - \tilde{X}^t_i\| \leq (2 \|X\|)^2 \|A^{*i}\|. \tag{5.12}$$

Here $\|X\|$ denotes $\sup_{\tau \leq T} \|X_\tau\|$.

Lemma 5.5.2. *Suppose μ obeys Assumption 2 and A^* obeys Assumption 1. Suppose $X \sim \pi$, which is the stationary distribution of $\mathsf{VAR}(A^*, \mu)$. $\langle X, x \rangle$ has mean 0 and is sub-Gaussian with variance proxy $C_\mu x^\top G x$*

Proof. Suppose $\eta_1, \ldots, \eta_n, \ldots$ is a sequence of i.i.d random vectors drawn from the noise distribution μ. We consider the partial sums $\sum_{i=0}^{n} A^{*i}\eta_i$. Call the law of this to be π_n. Clearly π_n converges in distribution to π as $n \to \infty$ since π_n is the law of the $n+1$-th iterate of $\mathsf{VAR}(A^*, \mu)$ chain stated at $X_0 = 0$. By Skorokhod representation theorem, we can define the infinite sequence $X^{(1)}, \ldots, X^{(n)}, \ldots$, and another random variable X such that $X^{(i)} \sim \pi_i$, $X \sim \pi$ and $\lim_{n\to\infty} X^{(n)} = X$ a.s. Define $G_n = \sum_{i=0}^{n} A^{*i}\Sigma(A^{*i})^T$. Clearly, $G_n \preceq G = \sum_{i=0}^{\infty} A^{*i}\Sigma(A^{*i})^T$. A simple evaluation of Chernoff bound for $\langle X^{(n)}, x\rangle$ by decomposing it into the partial sum of noises shows that:

$$\mathbb{E}\exp(\lambda\langle X^{(n)}, x\rangle) \leq \exp\left(\frac{\lambda^2 C_\mu}{2}\langle x, G_n x\rangle\right) \leq \exp\left(\frac{\lambda^2 C_\mu}{2}\langle x, Gx\rangle\right)$$

We now apply Fatou's lemma, since $X^{(n)} \to X$ almost surely, to the inequality above to conclude that:

$$\mathbb{E}\exp(\lambda\langle X, x\rangle) \leq \exp\left(\frac{\lambda^2 C_\mu}{2}\langle x, Gx\rangle\right).$$

□

Hence $\langle x, X_t\rangle$ is subgaussian with mean 0 and variance proxy $C_\mu \sigma_{\max}(G)\|x\|^2$. This will provide uniform variance for all x such that $\|x\|^2 = 1$.

From subgaussianity and standard ϵ-net argument we have the following lemma.

Lemma 5.5.3. *For any $\beta > 0$ there is a constant $c > 0$ such that*

$$\mathbb{P}\left[\exists\tau \leq T \,:\, \|X_\tau\|^2 > c\,\mathrm{Tr}\,G\log T\right] \leq \frac{d}{T^\beta} \tag{5.13}$$

Thus as long as $d < \mathrm{Poly}(T)$, for every $\alpha > 0$ there is a $c > 0$ such that

$$\mathbb{P}\left[\exists\tau \leq T \,:\, \|X_\tau\|^2 > c\,\mathrm{Tr}\,G\log T\right] \leq \frac{1}{T^\alpha} \tag{5.14}$$

5.5.2 Notations

Before we analyze this algorithm, we define some notations. We work in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and all the random elements are defined on this space. We define the following notations:

$$X^t_{-i} = X^t_{(S-1)-i},\ 0 \le i \le S-1, \quad G = \sum_{s=0}^{\infty} A^{*s}\Sigma(A^{*\top})^s, \quad G_t = \sum_{s=0}^{t-1} A^{*s}\Sigma(A^{*\top})^s,$$

$$\tilde{P}^t_i = \left(I - 2\gamma \tilde{X}^t_i \tilde{X}^{t,\top}_i\right), \quad \tilde{H}^t_{i,j} = \begin{cases} \prod_{s=i}^{j} \tilde{P}^t_{-s} & i \le j \\ I & i > j \end{cases},$$

$$\hat{\gamma} = 4\gamma(1-\gamma R), \quad \mathcal{C}^t_{-j} = \left\{\|X^t_{-j}\|^2 \le R\right\}, \quad \tilde{\mathcal{C}}^t_{-j} = \left\{\|\tilde{X}^t_{-j}\|^2 \le R\right\},$$

$$\mathcal{D}^t_{-j} = \left\{\|X^t_{-i}\|^2 \le R :\ j \le i \le B-1\right\} = \bigcap_{i=j}^{B-1} \mathcal{C}^t_{-i},$$

$$\mathcal{D}^{s,t} = \begin{cases} \bigcap_{r=s}^{t} \mathcal{D}^r_{-0} & s \le t \\ \Omega & s > t \end{cases}, \quad \tilde{\mathcal{D}}^t_{-j} = \left\{\|\tilde{X}^t_{-i}\|^2 \le R :\ j \le i \le B-1\right\} = \bigcap_{i=j}^{B-1} \tilde{\mathcal{C}}^t_{-i},$$

$$\tilde{\mathcal{D}}^{s,t} = \begin{cases} \bigcap_{r=s}^{t} \tilde{\mathcal{D}}^r_{-0} & s \le t \\ \Omega & s > t \end{cases}, \quad \hat{\mathcal{D}}^t_{-j} = \mathcal{D}^t_{-j} \cap \tilde{\mathcal{D}}^t_{-j}, \quad \hat{\mathcal{D}}^{s,t} = \mathcal{D}^{s,t} \cap \tilde{\mathcal{D}}^{s,t}.$$

Lastly c and c_i for $i = 0, 1, \cdots$ denote absolute constants that can change from line to line in the proofs.

5.5.3 Initial coupling

We consider the coupled process introduced in Definition 2 and run $\mathsf{SGD} - \mathsf{RER}$ with the fictitious coupled process $\tilde{X}_\tau$ instead of X_τ in order to obtain the iterates $\tilde{A}^t_i$ instead of A^{t-1}_i. Using Lemma 5.5.1, we can show that $\tilde{A}^{t-1}_i \approx A^{t-1}_i$. It is easier to analyze the iterates $\tilde{A}^t_i$ due to buffer independence.

Lemma 5.5.4. *Let $\gamma \le \frac{1}{2R}$. Under the event $\mathcal{D}^{0,N-1}$, for every $t \in [N]$ and $0 \le i \le B-1$ we have:*

$$\|A^{t-1}_i\| \le 2\gamma RT\,.$$

Proof. Consider the $\mathsf{SGD} - \mathsf{RER}$ iteration:

$$\begin{aligned} A^{t-1}_{i+1} &= A^{t-1}_i - 2\gamma(A^{t-1}_i X^{t-1}_{-i} - X^{t-1}_{-(i+1)})X^{t-1,\top}_{-i} \\ &= A^{t-1}_i(I - 2\gamma X^{t-1}_{-i} X^{t-1,\top}_{-i}) + 2\gamma X^{t-1}_{-(i-1))} X^{t-1,\top}_{-(i+1)} \end{aligned} \tag{5.15}$$

Observe that for our choice of γ and under the event $\mathcal{D}^{0,N-1}$, we have $\|(I - 2\gamma X^{t-1}_{-i} X^{t-1,\top}_{-i})\| \le 1$

and $\|X_{-(i+1)}^{t-1} X_{-i}^{t-1,\top}\| \leq R$. Therefore, triangle inequality implies:

$$\|A_{i+1}^{t-1}\| \leq \|A_i^{t-1}\| + 2\gamma R$$

We conclude the bound in the Lemma.

□

Lemma 5.5.5. *Suppose $\gamma < \frac{1}{2R}$. Under the event $\hat{\mathcal{D}}^{0,N-1}$ we have for every $t \in [N]$ and $0 \leq i \leq B-1$. $\|A_i^{t-1} - \tilde{A}_i^{t-1}\| \leq (16\gamma^2 R^2 T^2 + 8\gamma RT) \|A^{*u}\|$*

Proof. We again consider the evolution equation: $\tilde{X}_{-i}^{t-1}$

$$\begin{aligned} A_{i+1}^{t-1} &= A_i^{t-1} - 2\gamma(A_i^{t-1} X_{-i}^{t-1} - X_{-(i+1)}^{t-1}) X_{-i}^{t-1,\top} \\ &= A_i^{t-1} - 2\gamma(A_i^{t-1} \tilde{X}_{-i}^{t-1} - \tilde{X}_{-(i+1)}^{t-1}) \tilde{X}_{-i}^{t-1,\top} + \Delta_{t,i} \end{aligned} \tag{5.16}$$

Where

$$\Delta_{t,i} = 2\gamma A_i^{t-1} \left(\tilde{X}_{-i}^{t-1} \tilde{X}_{-i}^{t-1,\top} - X_{-i}^{t-1} X_{-i}^{t-1,\top} \right) + 2\gamma \left(X_{-(i+1)}^{t-1} X_{-i}^{t-1,\top} - \tilde{X}_{-(i+1)}^{t-1} \tilde{X}_{-i}^{t-1,\top} \right)$$

Using Lemmas 6.8.2 and 5.5.1, we conclude that:

$$\|\Delta_{t,i}\| \leq (16\gamma^2 R^2 T + 8\gamma R) \|A^{*u}\|$$

Using the recursion for $\tilde{A}_i^t$, we conclude:

$$\begin{aligned} & A_{i+1}^{t-1} - \tilde{A}_{i+1}^{t-1} = (A_i^{t-1} - \tilde{A}_i^{t-1}) \tilde{P}_i^t + \Delta_{t,i} \\ \implies & \left\| A_{i+1}^{t-1} - \tilde{A}_{i+1}^{t-1} \right\| \leq \left\| A_i^{t-1} - \tilde{A}_i^{t-1} \right\| \left\| \tilde{P}_i^t \right\| + (16\gamma^2 R^2 T + 8\gamma R) \|A^{*u}\| \\ \implies & \left\| A_{i+1}^{t-1} - \tilde{A}_{i+1}^{t-1} \right\| \leq \left\| A_i^{t-1} - \tilde{A}_i^{t-1} \right\| + (16\gamma^2 R^2 T + 8\gamma R) \|A^{*u}\| \end{aligned} \tag{5.17}$$

In the last step we have used the fact that under the event $\hat{\mathcal{D}}^{0,N-1}$, we must have $\left\| \tilde{P}_i^t \right\| \leq 1$. We conclude the statement of the lemma from Equation (6.37). □

We can now just analyze the iterates $\tilde{A}_i^{t-1}$ and then use Lemma 6.8.3 to infer error bounds for A_i^{t-1}. Henceforth, we will only consider $\tilde{A}_i^{t-1}$.

Lemma 5.5.6. *Consider the algorithmic iterates obtained from the actual process and coupled process*

(A_j^t) *and* $(\tilde{A}_j^t)$*. Then*

$$\mathbb{E}\left[\left(A_j^{t-1}-A^*\right)^\top\left(A_j^{t-1}-A^*\right)1\left[\mathcal{D}^{0,t-1}\right]\right]\preceq\mathbb{E}\left[\left(\tilde{A}_j^{t-1}-A^*\right)^\top\left(\tilde{A}_j^{t-1}-A^*\right)1\left[\tilde{\mathcal{D}}^{0,t-1}\right]\right]$$
$$+c\left(\gamma^3R^3T^3\left\|A^{*u}\right\|+\gamma^2d\sigma_{\max}(\Sigma)RT^2\frac{1}{T^{\alpha/2}}\right)I \tag{5.18}$$

for some constant c*. Furthermore, the same conclusion holds for the average iterates. That is let*

$$\hat{A}_{a,N}=\frac{1}{N-a}\sum_{t=a+1}^{N}A_B^{t-1}$$
$$\hat{\tilde{A}}_{a,N}=\frac{1}{N-a}\sum_{t=a+1}^{N}\tilde{A}_B^{t-1}$$

Then

$$\begin{aligned}&\mathbb{E}\left[\left(\hat{A}_{a,N}-A^*\right)^\top\left(\hat{A}_{a,N}-A^*\right)1\left[\mathcal{D}^{0,N-1}\right]\right]\\&\preceq\mathbb{E}\left[\left(\hat{\tilde{A}}_{a,N}-A^*\right)^\top\left(\hat{\tilde{A}}_{a,N}-A^*\right)1\left[\tilde{\mathcal{D}}^{0,N-1}\right]\right]\\&\quad+c\left(\gamma^3R^3T^3\left\|A^{*u}\right\|+\gamma^2d\sigma_{\max}(\Sigma)RT^2\frac{1}{T^{\alpha/2}}\right)I\end{aligned} \tag{5.19}$$

Remark 12. *The above lemma holds as is when* $A_j^{t-1},\tilde{A}_j^{t-1}$ *is replaced by* $A_j^{t-1,v},\tilde{A}_j^{t-1,v}$ *respectively.*

Proof. First we have

$$\begin{aligned}\mathbb{E}\left[\left(A_j^{t-1}-A^*\right)^\top\left(A_j^{t-1}-A^*\right)1\left[\mathcal{D}^{0,t-1}\right]\right]&\preceq\mathbb{E}\left[\left(A_j^{t-1}-A^*\right)^\top\left(A_j^{t-1}-A^*\right)1\left[\hat{\mathcal{D}}^{0,t-1}\right]\right]\\&\quad+4\gamma^2(Bt)^2R\sqrt{\mu_4}\frac{1}{T^{\alpha/2}}I\\&\preceq\mathbb{E}\left[\left(A_j^{t-1}-A^*\right)^\top\left(A_j^{t-1}-A^*\right)1\left[\hat{\mathcal{D}}^{0,t-1}\right]\right]\\&\quad+c\gamma^2d\sigma_{\max}(\Sigma)RT^2\frac{1}{T^{\alpha/2}}I\end{aligned} \tag{5.20}$$

Next, we have

$$\begin{aligned}&\left\|\left(A_j^{t-1}-A^*\right)^\top\left(A_j^{t-1}-A^*\right)-\left(\tilde{A}_j^{t-1}-A^*\right)^\top\left(\tilde{A}_j^{t-1}-A^*\right)\right\|\\&\leq\left\|A_j^{t-1}-\tilde{A}_j^{t-1}\right\|\left(\left\|\left(A_j^{t-1}-A^*\right)\right\|+\left\|\left(\tilde{A}_j^{t-1}-A^*\right)\right\|\right)\\&\leq\left\|A_j^{t-1}-\tilde{A}_j^{t-1}\right\|\left(2\left\|A^*\right\|+\left\|A_j^{t-1}\right\|+\left\|\tilde{A}_j^{t-1}\right\|\right)\end{aligned} \tag{5.21}$$

Thus on the event $\hat{\mathcal{D}}^{0,t-1}$, using lemma 6.8.3 and lemma 6.8.2 we get

$$\begin{aligned}&\left\|\left(A_j^{t-1}-A^*\right)^\top\left(A_j^{t-1}-A^*\right)-\left(\tilde{A}_j^{t-1}-A^*\right)^\top\left(\tilde{A}_j^{t-1}-A^*\right)\right\|\\&\leq c(\gamma^2R^2T^2+\gamma RT)(\gamma RT+\|A^*\|+\|A_0\|)\left\|A^{*u}\right\|\leq c\gamma^3R^3T^3\left\|A^{*u}\right\|\end{aligned}\tag{5.22}$$

for some constant c. (We have suppressed the dependence on A_0 and A^* since they are constants and γRT grows with T).

The proof follows by combining (5.20) and (5.22).

The proof of (5.19) follows similarly. □

5.6 Bias variance decomposition

Now, we can unroll the recursion in (5.5), but for the coupled iterates $\tilde{A}_i^{t-1}$ as

$$\tilde{A}_B^{t-1}-A^*=\left(\tilde{A}_B^{t-1,b}-A^*\right)+\left(\tilde{A}_B^{t-1,v}\right),\tag{5.23}$$

where

$$\left(\tilde{A}_B^{t-1,b}-A^*\right)=(A_0-A^*)\prod_{s=0}^{t-1}\tilde{H}_{0,B-1}^s\tag{5.24}$$

is the *bias* term, and the *variance* term is given by:

$$\left(\tilde{A}_B^{t-1,v}\right)=2\gamma\sum_{r=1}^{t}\sum_{j=0}^{B-1}\eta_{-j}^{t-r}\tilde{X}_{-j}^{t-r,\top}\tilde{H}_{j+1,B-1}^{t-r}\prod_{s=r-1}^{1}\tilde{H}_{0,B-1}^{t-s}\tag{5.25}$$

Here we use the convention that whenever $r=1$, the product $\prod_{s=r-1}^{1}$ is empty i.e, equal to 1. The 'bias' term is obtained when the noise terms are set to 0, and captures the movement of the algorithm towards the optimal A^* when we set the initial iterate far away from it. The 'variance' term $\left(A_B^{t,v}-A^*\right)$ capture the uncertainty due to the inherent noise in the data. Our main goal is to understand the performance (estimation and prediction) of the tail-averaged iterates output by $\mathsf{SGD-RER}$. Here, we consider just the last iterate, but the same technique applies to all the outputs of $\mathsf{SGD-RER}$. That is, $\hat{\tilde{A}}_{a,N}=\frac{1}{N-a}\sum_{t=a+1}^{N}\tilde{A}_B^{t-1}$, for $a=\lceil\theta N\rceil$ with $0<\theta<1$. We can decompose the above into bias and variance as: $\hat{\tilde{A}}_{a,N}=\hat{\tilde{A}}_{a,N}^v+\hat{\tilde{A}}_{a,N}^b$, with,

$$\hat{\tilde{A}}_{a,N}^v=\frac{1}{N-a}\sum_{t=a+1}^{N}\tilde{A}_B^{t-1,v}\tag{5.26}$$

$$\hat{\tilde{A}}_{a,N}^b=\frac{1}{N-a}\sum_{t=a+1}^{N}\tilde{A}_B^{t-1,b}.\tag{5.27}$$

Similarly, we can decompose the final error into 'bias' and 'variance' as in Lemma 5.6.1 below.

Lemma 5.6.1 (Bias-Variance Decomposition)**.** *We have the following decomposition:*

$$\left(\tilde{A}_B^{t-1} - A^*\right)^\top \left(\tilde{A}_B^{t-1} - A^*\right) \preceq 2\left[\left(\tilde{A}_B^{t-1,b} - A^*\right)^\top \left(\tilde{A}_B^{t-1,b} - A^*\right) + \left(\tilde{A}_B^{t-1,v}\right)^\top \left(\tilde{A}_B^{t-1,v}\right)\right].$$

5.7 Preliminary results for the parameter error

In this section, we develop the concentration inequalities necessary to obtain bounds on $\mathcal{L}_{\mathsf{op}}$. Consider Equation (5.25)

$$\left(\tilde{A}_B^{t-1,v}\right) = 2\gamma \sum_{r=1}^{t} \sum_{j=0}^{B-1} \eta_{-j}^{t-r} \tilde{X}_{-j}^{t-r,\top} \tilde{H}_{j+1,B-1}^{t-r} \prod_{s=r-1}^{1} \tilde{H}_{0,B-1}^{t-s} \tag{5.28}$$

Splitting the sum into $r = 1$ and $r = 2, \ldots, t$, it is easy to show the following recursion:

$$\left(\tilde{A}_B^{t-1,v}\right) = 2\gamma \sum_{j=0}^{B-1} \eta_{-j}^{t-1} \tilde{X}_{-j}^{t-1,\top} \tilde{H}_{j+1,B-1}^{t-1} + \left(\tilde{A}_B^{t-2,v}\right) \tilde{H}_{0,B-1}^{t-1} \tag{5.29}$$

We will consider the matrix $\Delta_{t-1} := 2\gamma \sum_{j=0}^{B-1} \eta_{-j}^{t-1} \tilde{X}_{-j}^{t-1,\top} \tilde{H}_{j+1,B-1}^{t-1}$. Recall the sequence of events $\tilde{\mathcal{D}}_{-j}^{t-1}$ for $j = 0, 1, \ldots, B-1$ as defined in Section 5.5.2. We will pick R as in Section 5.3 so that $\mathbb{P}(\tilde{\mathcal{D}}_{-0}^{t-1})$ is close to 1.

For the sake of clarity, we drop the dependence on t while stating and proving some of the technical results since the events and random variables considered there are identically distributed for every t. That is, consider $\tilde{\mathcal{D}}_{-j}$ instead of $\tilde{\mathcal{D}}_{-j}^{t-1}$ and

$$\Delta := 2\gamma \sum_{j=0}^{B-1} \eta_{-j} \tilde{X}_{-j}^\top \tilde{H}_{j+1,B-1}$$

We will bound the exponential moment generating function of Δ:

Lemma 5.7.1. *Suppose Assumption 2 holds and that $\gamma R < 1$. Let $\lambda \in \mathbb{R}$ and $x, y \in \mathbb{R}^d$ are arbitrary. Then, we have:*

1.

$$\mathbb{E}\left[\exp(\gamma\lambda^2 C_\mu \langle x, \Sigma x\rangle \langle y, \tilde{H}_{0,B-1}^\top \tilde{H}_{0,B-1} y\rangle + \lambda\langle x, \Delta y\rangle) | \tilde{\mathcal{D}}_{-0}\right] \leq \frac{\exp\left(\gamma\lambda^2 C_\mu \langle x, \Sigma x\rangle \|y\|^2\right)}{\mathbb{P}(\tilde{\mathcal{D}}_{-0})}$$

2.

$$\mathbb{E}\left[\exp(\lambda\langle x, \Delta y\rangle)|\tilde{\mathcal{D}}_{-0}\right] \leq \frac{\exp\left(\gamma\lambda^2 C_\mu \langle x, \Sigma x\rangle \|y\|^2\right)}{\mathbb{P}(\tilde{\mathcal{D}}_{-0})}$$

Where C_μ is as given in Assumption 2

Proof. We will just prove item 1 since item 2 follows from it trivially as

$$\gamma\lambda^2 C_\mu \langle x, \Sigma x\rangle \langle y, \tilde{H}_{0,B-1}^\top \tilde{H}_{0,B-1} y\rangle \geq 0\,.$$

For the sake of clarity, we will take:

$$\Xi_0 := \gamma\lambda^2 C_\mu \langle x, \Sigma x\rangle \langle y, \tilde{H}_{0,B-1}^\top \tilde{H}_{0,B-1} y\rangle$$

and more generally,

$$\Xi_k = \gamma\lambda^2 C_\mu \langle x, \Sigma x\rangle \langle y, \tilde{H}_{k,B-1}^\top \tilde{H}_{k,B-1} y\rangle$$

Consider $\Delta_{-k} := 2\gamma \sum_{j=k}^{B-1} \eta_{-j} \tilde{X}_{-j}^\top \tilde{H}_{j+1,B-1}$. We will first prove the following claim before bounding the exponential moment:

Claim 7. *Whenever $\|\tilde{X}_{-k}\|^2 \leq R$ and $\gamma R < 1/2$, we have:*

$$\Xi_k + 2\gamma^2\lambda^2 C_\mu \langle x, \Sigma x\rangle \langle y, \tilde{H}_{k+1,B-1}^\top \tilde{X}_{-k} \tilde{X}_{-k}^\top \tilde{H}_{k+1,B-1} y\rangle \leq \Xi_{k+1}$$

Proof. We use the fact that $\tilde{H}_{k,B-1}^\top \tilde{H}_{k,B-1} = \tilde{H}_{k+1,B-1}^\top (I - 2\gamma \tilde{X}_{-k} \tilde{X}_{-k}^\top)^2 \tilde{H}_{k+1,B-1}$ to conclude that:

$$\begin{aligned}
&\Xi_k + 2\gamma^2\lambda^2 C_\mu \langle x, \Sigma x\rangle \langle y, \tilde{H}_{k+1,B-1}^\top \tilde{X}_{-k} \tilde{X}_{-k}^\top \tilde{H}_{k+1,B-1} y\rangle \\
&= \gamma\lambda^2 C_\mu \langle x, \Sigma x\rangle \langle y, \tilde{H}_{k+1,B-1}^\top \left(I - 2\gamma \tilde{X}_{-k} \tilde{X}_{-k}^\top + 4\gamma^2 \|\tilde{X}_{-k}\|^2 \tilde{X}_{-k} \tilde{X}_{-k}^\top\right) \tilde{H}_{k+1,B-1} y\rangle \\
&\leq \gamma\lambda^2 C_\mu \langle x, \Sigma x\rangle \langle y, \tilde{H}_{k+1,B-1}^\top \tilde{H}_{k+1,B-1} y\rangle = \Xi_{k+1}
\end{aligned} \tag{5.30}$$

In the second step we have used the fact that when $\gamma\|\tilde{X}_{-k}\|^2 \leq 1/2$, we have that

$$I - 2\gamma \tilde{X}_{-k} \tilde{X}_{-k}^\top + 4\gamma^2 \|\tilde{X}_{-k}\|^2 \tilde{X}_{-k} \tilde{X}_{-k}^\top \preceq I$$

□

First note that $\Delta = 2\gamma\eta_0 \tilde{X}_0^\top \tilde{H}_{1,B-1} + \Delta_{-1}$. Now,

$$\begin{aligned}
&\mathbb{E}\left[\exp(\Xi_0 + \lambda\langle x, \Delta y\rangle)|\tilde{\mathcal{D}}_{-0}\right] = \frac{1}{\mathbb{P}(\tilde{\mathcal{D}}_{-0})}\mathbb{E}\left[\exp(\Xi_0 + \lambda\langle x, \Delta y\rangle)\mathbb{1}\left(\tilde{\mathcal{D}}_{-0}\right)\right] \\
&= \frac{1}{\mathbb{P}(\tilde{\mathcal{D}}_{-0})}\mathbb{E}\left[\exp\left(\Xi_0 + 2\lambda\gamma\langle x, \eta_{-0}\rangle\langle \tilde{X}_{-0}, \tilde{H}_{1,B-1}y\rangle + \lambda\langle x, \Delta_{-1}y\rangle\right)\mathbb{1}\left(\tilde{\mathcal{D}}_{-0}\right)\right] \\
&\leq \frac{1}{\mathbb{P}(\tilde{\mathcal{D}}_{-0})}\mathbb{E}\left[\exp\left(\Xi_0 + 2\gamma^2\lambda^2 C_\mu\langle x, \Sigma x\rangle\langle y, \tilde{H}_{1,B-1}^\top \tilde{X}_{-0}\tilde{X}_{-0}^\top \tilde{H}_{1,B-1}y\rangle + \lambda\langle x, \Delta_{-1}y\rangle\right)\mathbb{1}\left(\tilde{\mathcal{D}}_{-0}\right)\right] \\
&\leq \frac{1}{\mathbb{P}(\tilde{\mathcal{D}}_{-0})}\mathbb{E}\left[\exp\left(\Xi_1 + \lambda\langle x, \Delta_{-1}y\rangle\right)\mathbb{1}\left(\tilde{\mathcal{D}}_{-0}\right)\right] \\
&\leq \frac{1}{\mathbb{P}(\tilde{\mathcal{D}}_{-0})}\mathbb{E}\left[\exp\left(\Xi_1 + \lambda\langle x, \Delta_{-1}y\rangle\right)\mathbb{1}\left(\tilde{\mathcal{D}}_{-1}\right)\right] \qquad (5.31)
\end{aligned}$$

In the first step we have used the definition of conditional expectation, in the third step we have used the fact that η_{-0} is independent of $\tilde{\mathcal{D}}_{-0}$, Δ_{-1}, $\tilde{X}_{-0}^\top \tilde{H}_{1,B-1}$, and Δ_{-1} and have applied the sub-Gaussianity from Assumption 2. In the fourth step, using the fact under the event $\tilde{\mathcal{D}}_{-0}$, $\|\tilde{X}_{-0}\|^2 \leq R$ we have applied Claim 7. In the final step, we have used the fact that $\tilde{\mathcal{D}}_{-0} \subseteq \tilde{\mathcal{D}}_{-1}$. We proceed by induction over Equation (5.31) to conclude the result.

□

We now consider the matrix $\tilde{H}_{0,B-1}$ under the event $\tilde{\mathcal{D}}_{-0}$.

Lemma 5.7.2. *Suppose that $\gamma RB < \frac{1}{6}$. Then, under the event $\tilde{\mathcal{D}}_{-0}$, we have:*

$$I - 4\gamma\left(1 + \tfrac{2\gamma BR}{1-4\gamma BR}\right)\sum_{i=0}^{B-1}\tilde{X}_{-i}\tilde{X}_{-i}^\top \preceq \tilde{H}_{0,B-1}^\top \tilde{H}_{0,B-1} \preceq I - 4\gamma\left(1 - \tfrac{2\gamma BR}{1-4\gamma BR}\right)\sum_{i=0}^{B-1}\tilde{X}_{-i}\tilde{X}_{-i}^\top$$

Proof. By definition, we have: $\tilde{H}_{0,B-1} = \prod_{j=0}^{B-1}(I - 2\gamma\tilde{X}_{-j}\tilde{X}_{-j}^\top)$. Expanding out the product, we get an expression of the form:

$$\tilde{H}_{0,B-1}^\top \tilde{H}_{0,B-1} = I - 4\gamma\sum_{i=0}^{B-1}\tilde{X}_{-i}\tilde{X}_{-i}^\top + (2\gamma)^2\sum_{i,j}\tilde{X}_{-i}\tilde{X}_{-i}^\top\tilde{X}_{-j}\tilde{X}_{-j}^\top + \ldots \qquad (5.32)$$

Here, the summation $\sum_{i,j}$ is over all possible combinations possible when the product is expanded and $\ldots$ denotes higher order terms of the form $\tilde{X}_{-i_1}\tilde{X}_{-i_1}^\top \ldots \tilde{X}_{-i_k}\tilde{X}_{-i_k}^\top$

Claim 8. *Assume $k \geq 2$ and $i_1, \ldots, i_k \in \{0, \ldots, B-1\}$. Under the event $\tilde{\mathcal{D}}_{-0}$, for any $x \in \mathbb{R}^d$, we have:*

$$\left|x^\top \tilde{X}_{-i_1}\tilde{X}_{-i_1}^\top \ldots \tilde{X}_{-i_k}\tilde{X}_{-i_k}^\top x\right| \leq \frac{R^{k-1}}{2}\left[x^\top \tilde{X}_{-i_1}\tilde{X}_{-i_1}^\top x + x^\top \tilde{X}_{-i_k}\tilde{X}_{-i_k}^\top x\right]$$

Proof. This follows from an application of AM-GM inequality. It is clear by Cauchy-Schwarz inequality that $|\langle \tilde{X}_{i_l}, \tilde{X}_{i_{l+1}}\rangle| \leq R$, which implies:

$$\left|x^\top \tilde{X}_{-i_1}\tilde{X}_{-i_1}^\top \dots \tilde{X}_{-i_k}\tilde{X}_{-i_k}^\top x\right| \le R^{k-1}\left|\left[x^\top \tilde{X}_{-i_1}\tilde{X}_{-i_k}^\top x\right]\right| \le \frac{R^{k-1}}{2}\left[\langle x, \tilde{X}_{-i_1}\rangle^2 + \langle \tilde{X}_{-i_k}, x\rangle^2\right].$$

Where the last inequality follows from an application of the AM-GM inequality. □

From Claim 8, we conclude that:

$$\sum_{i_1,\dots,i_k} \tilde{X}_{-i_1}\tilde{X}_{-i_1}^\top \dots \tilde{X}_{-i_k}\tilde{X}_{-i_k}^\top \preceq (2B)^{k-1}R^{k-1}\sum_{i=0}^{B-1}\tilde{X}_{-i}\tilde{X}_{-i}^\top$$

Plugging this into Equation (5.32), we have that under the event $\tilde{\mathcal{D}}_{-0}$:

$$\begin{aligned}\tilde{H}_{0,B-1}^\top \tilde{H}_{0,B-1} &\preceq I - 4\gamma\sum_{i=0}^{B-1}\sum_{i=0}^{B-1}\tilde{X}_{-i}\tilde{X}_{-i}^\top + \sum_{k=2}^{2B}(2\gamma)^k(2B)^{k-1}R^{k-1}\sum_{i=0}^{B-1}\tilde{X}_{-i}\tilde{X}_{-i}^\top \\ &\preceq I - 4\gamma\sum_{i=0}^{B-1}\sum_{i=0}^{B-1}\tilde{X}_{-i}\tilde{X}_{-i}^\top + 2\gamma\frac{4\gamma BR}{1-4\gamma BR}\sum_{i=0}^{B-1}\sum_{i=0}^{B-1}\tilde{X}_{-i}\tilde{X}_{-i}^\top \end{aligned} \tag{5.33}$$

Here we have used the fact that $4\gamma BR < 1$ to convert the finite sum to an infinite sum. Using the bound on γ, we conclude the upper bound. The lower bound follows with a similar proof.

□

Lemma 5.7.3. *Suppose $\gamma BR < \frac{1}{6}$. Let $G := \mathbb{E}\tilde{X}_{-i}\tilde{X}_{-i}^\top$ and $M_4 := \mathbb{E}\left\|\tilde{X}_{-i}\right\|^4$. Then, we have:*

$$\begin{aligned}\mathbb{E}\left[\tilde{H}_{0,B-1}^\top\tilde{H}_{0,B-1}\middle|\tilde{\mathcal{D}}_{-0}\right] \preceq I - \frac{4\gamma B}{\mathbb{P}(\tilde{\mathcal{D}}_{-0})}\left(1-\tfrac{2\gamma BR}{1-4\gamma BR}\right)G + \\ \frac{4\gamma B\sqrt{M_4(1-\mathbb{P}(\tilde{\mathcal{D}}_{-0}))}}{\mathbb{P}(\tilde{\mathcal{D}}_{-0})}\left(1-\tfrac{2\gamma BR}{1-4\gamma BR}\right)I\end{aligned}$$

Proof. The result follows from the statement of Lemma 6.9.3, once we show the following inequality via Cauchy Schwarz inequality and the definition of conditional expectation:

$$\mathbb{E}\left[\tilde{X}_{-i}\tilde{X}_{-i}^\top\middle|\tilde{\mathcal{D}}_{-0}\right] \succeq \frac{G}{\mathbb{P}(\tilde{\mathcal{D}}_{-0})} - I\frac{\sqrt{\mathbb{E}\left\|\tilde{X}_{-i}\right\|^4}\sqrt{1-\mathbb{P}(\tilde{\mathcal{D}}_{-0})}}{\mathbb{P}(\tilde{\mathcal{D}}_{-0})}.$$

□

Now we will show that $\tilde{H}_{0,B-1}$ contracts any given vector with probability at-least $p_0 > 0$. For this we will refer to lemma 5.5.2 where it is shown that if $X \sim \pi$ then $\langle X, x\rangle$ has mean 0 and is

sub-Gaussian with variance proxy $C_\mu x^\top Gx$. Using this will show that the matrix $\tilde{H}_{0,B-1}$ operating on a given vector x contracts it with a high enough probability.

Lemma 5.7.4. *Suppose $\gamma RB < \frac{1}{8}$ and that μ obeys Assumption 2. There exists a constant $c_0 > 0$ which depends only on C_μ such that whenever $1 - \mathbb{P}(\tilde{\mathcal{D}}_{-0}) \leq c_0$, then for any arbitrary $x \in \mathbb{R}^2$*

$$\mathbb{P}\left(\|\tilde{H}_{0,B-1}x\|^2 \geq \|x\|^2 - B\gamma x^\top Gx \middle| \tilde{\mathcal{D}}_{-0}\right) \leq 1 - p_0 < 1\,.$$

Where $p_0 > 0$ depends only on C_μ.

Proof. Initially we do not condition on $\tilde{\mathcal{D}}_{-0}$. Consider the quantity: $Y := \sum_{i=0}^{B-1} \langle x, \tilde{X}_{-i}\rangle^2$.

Claim 9.

$$\mathbb{P}\left(Y \geq 1/2Bx^\top Gx\right) \geq q_0$$

where $q_0 > 0$ depends only on sub-Gaussianity parameter C_μ

Proof. We consider the Payley-Zygmund inequality which states that for any positive random variable Y with a finite second moment, we have:

$$\mathbb{P}\left(Y > \tfrac{1}{2}\mathbb{E}Y\right) \geq \frac{1}{4}\frac{(\mathbb{E}Y)^2}{\mathbb{E}Y^2}.$$

Note that $\mathbb{E}Y = Bx^\top Gx$. The statement of the lemma follows once we lower bound the quantity $\frac{(\mathbb{E}Y)^2}{\mathbb{E}Y^2}$. Clearly, $(\mathbb{E}Y)^2 = B^2 x^\top Gx$. Now,

$$\begin{aligned}\mathbb{E}Y^2 &= \sum_{i,j} \mathbb{E}\langle x, X_i\rangle^2 \langle x, X_j\rangle^2 \leq \sum_{i,j} \sqrt{\mathbb{E}\langle x, X_i\rangle^4}\sqrt{\mathbb{E}\langle x, X_j\rangle^2} = B^2\mathbb{E}\langle x, X_i\rangle^4 \\ &\leq B^2 c_1 C_\mu^2 (x^\top Gx)^2 \end{aligned} \tag{5.34}$$

Here, the second step follows from Cauchy-Schwarz inequality. The third step follows from the fact that X_i are all identically distributed. The fourth step follows from Lemma 5.5.2 and Theorem 2.1 from [148]. The statement of the claim follows once we apply Payley-Zygmund inequality. □

Now, by definition of conditional probabililty and Claim 9, we have:

$$\mathbb{P}\left(\sum_{i=0}^{B-1} \langle x, \tilde{X}_{-i}\rangle^2 \leq \frac{B}{2} x^T Gx \middle| \tilde{\mathcal{D}}_{-0}\right) \leq \frac{(1-q_0)}{\mathbb{P}(\tilde{\mathcal{D}}_{-0})}$$

Now the statement of the lemma follows from an application of Lemma 6.9.3 □

Now we want to bound the operator norm of $\prod_{s=a}^{a+b} \tilde{H}^s_{0,B-1}$ with high probability under the event $\cap_{s=a}^{a+b} \tilde{\mathcal{D}}^s_{-0}$.

Lemma 5.7.5. *Suppose the conditions in Lemma 6.9.2 hold. Let $\sigma_{\min}(G)$ denote the smallest eigenvalue of G. We also assume that $\mathbb{P}(\tilde{\mathcal{D}}^{a,b}) > 1/2$. Conditioned on the event $\tilde{\mathcal{D}}^{a,b}$,*

1. *$\|\prod_{s=a}^{b} \tilde{H}^s_{0,B-1}\| \leq 1$ almost surely*
2. *Whenever $b-a+1$ is larger than some constant which depends only on C_μ, we have:*

$$\mathbb{P}\left(\|\prod_{s=a}^{b} \tilde{H}^s_{0,B-1}\| \geq 2(1-\gamma B\sigma_{\min}(G))^{c_4(b-a+1)} \middle| \tilde{\mathcal{D}}^{a,b}\right) \leq \exp(-c_3(b-a+1)+c_5 d)$$

Where c_3, c_4 and c_5 are constants which depend only on C_μ

Proof.

1. The proof follows from an application of Lemma 6.9.3.
2. We will prove this with an ϵ net argument over the sphere in $\mathbb{R}^d$ dimensions.

   Suppose we have arbitrary $x \in \mathbb{R}^d$ such that $\|x\| = 1$. Conditioned on the event $\tilde{\mathcal{D}}^{a,b}$, the matrices $\tilde{H}^s_{0,B-1}$ are all independent for $a \leq s \leq b$. We also note that $\tilde{H}^s_{0,B-1}$ is independent of $\tilde{\mathcal{D}}^t$ for $t \neq s$. Let $K_v := \prod_{s=v}^{b} \tilde{H}^s_{0,B-1}$. When $v \geq b+1$, we take this product to be identity. Consider the set of events $\mathcal{G}_v := \{\|\tilde{H}^v_{0,B-1} K_{v+1} x\|^2 \leq \|K_{v+1}x\|^2(1-\gamma B\sigma_{\min}(G)\}$. From Lemma 6.9.2, we have that whenever $v \in (a,b)$:

$$\mathbb{P}(\mathcal{G}_v^c | \tilde{\mathcal{D}}^v, \tilde{H}^s_{0,B-1} : s \neq v) \leq 1 - p_0 \tag{5.35}$$

   Where p_0 is given in Lemma 6.9.2

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

$$\|\prod_{s=a}^{b} \tilde{H}^s_{0,B-1} x\| \leq (1-\gamma B\sigma_{\min}(G))^{\frac{r}{2}}\,. \tag{5.36}$$

   Therefore, whenever Equation (6.119) is violated, we must have a set $D^c \subseteq \{a, \ldots, b\}$ such that $|D^c| \geq b-a-r$ and the event $\cap_{v \in D^c} \mathcal{G}_v^c$ holds. We will union bound all such events indexed by D^c to obtain an upper bound on the probability that Equation (6.119) is violated. Therefore, using Equation (6.118) along with the union bound, we have:

$$\mathbb{P}\left(\|\prod_{s=a}^{b} \tilde{H}^s_{0,B-1} x\| \geq (1-\gamma B\sigma_{\min}(G))^{\frac{r}{2}} \middle| \tilde{\mathcal{D}}^{a,b}\right) \leq \binom{b-a+1}{b-a-r}(1-p_0)^{b-a-r}$$

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

$$\mathbb{P}\left(\|\prod_{s=a}^{b} \tilde{H}^s_{0,B-1} x\| \geq (1-\gamma B\sigma_{\min}(G))^{\frac{r}{2}} \middle| \tilde{\mathcal{D}}^{a,b}\right) \leq \exp(-c_3(b-a+1))$$

Now, let $\mathcal{N}$ be a 1/2-net of the sphere $\mathcal{S}^{d-1}$. Using Corollary 4.2.13 in [178], we can choose $|\mathcal{N}| \leq 6^d$. By Lemma 4.4.1 in [178] we show that:

$$\|\prod_{s=a}^{b} \tilde{H}^s_{0,B-1}\| \leq 2 \sup_{x \in \mathcal{N}} \|\prod_{s=a}^{b} \tilde{H}^s_{0,B-1} x\| \tag{5.37}$$

By union bounding Equation (6.120) for every $x \in \mathcal{N}$, we conclude that:

$$\begin{aligned}\mathbb{P}\left(\|\prod_{s=a}^{b} \tilde{H}^s_{0,B-1}\| \geq 2(1-\gamma B\sigma_{\min}(G))^{c_4(b-a+1)} \middle| \tilde{\mathcal{D}}^{a,b}\right) &\leq |\mathcal{N}| \exp(-c_3(b-a+1)) \\ = \exp(-c_3(b-a+1)+c_5 d) \end{aligned} \tag{5.38}$$

□

Now we will give a high probability bound for the following operator:

$$Fa, N := \sum_{r=a}^{N-1} \prod_{s=a+1}^{r} \tilde{H}^s_{0,B-1} \tag{5.39}$$

Here, we use the convention that $\prod_{s=a+1}^{a} \tilde{H}^s_{0,B-1} = I$

Lemma 5.7.6. *Suppose $c_4 \gamma B \sigma_{\min}(G) < \frac{1}{4}$ for the constant c_4 as given in Lemma 5.7.5. Suppose all the conditions given in the statement of Lemma 5.7.5 hold. Then, for any $\delta \in (0,1)$, we have:*

$$\mathbb{P}\left(\|F_{a,N}\| \geq C\left(d + \log\frac{N}{\delta} + \frac{1}{\gamma B \sigma_{\min}(G)}\right) \middle| \tilde{\mathcal{D}}^{a,N-1}\right) \leq \delta$$

Where C is a constant which depends only on C_μ

Proof. We consider the triangle inequality: $\|F_{a,N}\| \leq \sum_{t=a}^{N-1} \left\|\prod_{s=a+1}^{t} \tilde{H}^s_{0,B-1}\right\|$. By Lemma 5.7.5, we have that whenever $t - a \geq \frac{c_5 d}{c_3} + \frac{\log\frac{N}{\delta}}{c_3}$:

$$\mathbb{P}\left(\|\prod_{s=a+1}^{t} \tilde{H}^s_{0,B-1}\| \geq 2(1-\gamma B\sigma_{\min}(G))^{c_4(t-a)} \middle| \tilde{\mathcal{D}}^{a,N-1}\right) \leq \frac{\delta}{N}$$

Using union bound, we show that when conditioned on $\tilde{\mathcal{D}}^{a,N-1}$, with probability at least $1-\delta$ the following holds:

1. For all $a \leq t \leq N-1$ such that $t - a \geq \frac{c_5 d}{c_3} + \frac{\log\frac{N}{\delta}}{c_3}$:

$$\|\prod_{s=t}^{N} \tilde{H}^s_{0,B-1}\| \leq 2(1-\gamma B\sigma_{\min}(G))^{c_4(t-a)}$$

2. For all t such that $t - a < \frac{c_5 d}{c_3} + \frac{\log\frac{N}{\delta}}{c_3}$, we have: $\|\prod_{s=t}^{N} \tilde{H}^s_{0,B-1}\| \leq 1$. For this, we use the

almost sure bound given in item 1 of Lemma 5.7.5

Therefore, when conditioned on $\tilde{\mathcal{D}}^{a,N-1}$, with probability at least $1-\delta$ we have:

$$\begin{aligned}
\|F_{a,N}\| &\leq C(d+\log\frac{N}{\delta}) + 2\sum_{j=0}^{\infty}(1-\gamma B\sigma_{\min}(G))^{c_4 j} \\
&\leq C(d+\log\frac{N}{\delta}) + 2\sum_{j=0}^{\infty}\exp(-c_4 j\gamma B\sigma_{\min}(G)) \\
&\leq C(d+\log\frac{N}{\delta}) + \frac{2}{1-\exp(-c_4\gamma B\sigma_{\min}(G))} \\
&\leq C(d+\log\frac{N}{\delta}) + \frac{2}{c_4\gamma B\sigma_{\min}(G) - \frac{c_4^2\gamma^2 B\sigma_{\min}(G)}{2}} \\
&\leq C\left(d+\log\frac{N}{\delta} + \frac{1}{\gamma B\sigma_{\min}(G)}\right) \qquad (5.40)
\end{aligned}$$

In the first step, we have used the event described above to bound the operator norm via. the infinite geometric series. In the second step, we have used the inequality $(1-x)^a \leq \exp(-ax)$ whenever $x \in [0,1]$ and $a > 0$. In the fourth step, we have used the inequality $\exp(-x) \leq 1 - x + \frac{x^2}{2}$ whenever $x \in [0,1]$. In the last step, we have absorbed constants into a single constant C □

We will now consider the averaged iterate of the coupled process as defined in Equation (5.26) with $a = 0$.

$$\hat{\tilde{A}}^v_{0,N} := \frac{1}{N}\sum_{t=1}^{N}\left(\tilde{A}_B^{t-1,v}\right) \qquad (5.41)$$

We recall the definition of Δ_{t-1} from the beginning of the Section 5.7 and the recursion shown in Equation (5.29). We combine these with Equation (5.41) to show:

$$\hat{\tilde{A}}^v_{0,N} = \frac{1}{N}\sum_{t=1}^{N}\Delta_{t-1}F_{t-1,N} \qquad (5.42)$$

Where $F_{a,N}$ is as defined in Equation (6.122). Using the results in Lemma 5.7.1 and a similar proof technique we show the following theorem. We define the following event as considered in Lemma (5.7.6):

$$\tilde{\mathcal{M}}^{t-1} := \left\{\|F_{t-1,N}\| \leq C\left(d+\log\frac{N}{\delta} + \tfrac{1}{\gamma B\sigma_{\min}(G)}\right)\right\}$$

Define the event $\tilde{\mathcal{M}}^{0,N-1} = \cap_{t=0}^{N-1}\tilde{\mathcal{M}}^t$ and recall the definition of the event $\tilde{\mathcal{D}}^{0,N-1}$.

Theorem 5.7.7. *We suppose that the conditions in Lemmas 5.7.1, 5.7.6 and 6.9.3 hold. We also assume that* $\mathbb{P}(\tilde{\mathcal{M}}^{0,N-1}\cap\tilde{\mathcal{D}}^{0,N-1}) \geq \frac{1}{2}$. *Define* $\alpha := C(d+\log\frac{N}{\delta} + \frac{1}{\gamma B\sigma_{\min}(G)})$ *as in the definition of the event* $\tilde{\mathcal{M}}^t$

$$\mathbb{P}\left(\|\hat{\tilde{A}}^v_{0,N}\| > \beta\middle|\tilde{\mathcal{M}}^{0,N-1}\cap\tilde{\mathcal{D}}^{0,N-1}\right) \leq \exp\left(c_1 d - \frac{\beta^2 N}{16\gamma C_\mu\sigma_{\max}(\Sigma)(1+2\alpha)}\right).$$

Proof. Recall the events $\tilde{\mathcal{D}}^{t,N-1}$ and define $\tilde{\mathcal{M}}^{t,N-1} := \cap_{s=t}^{N-1}\tilde{\mathcal{M}}^t$. We recall that Δ_{t-1} is independent of $F_{t-1,N}$ and $\tilde{\mathcal{D}}^{t,N-1}$. Now consider arbitrary $x, y \in \mathbb{R}^d$ such that $\|x\| = \|y\| = 1$. Define $\Gamma_{t-1,N-1} := \frac{1}{N}\sum_{s=t}^{N}\Delta_{s-1}F_{s-1,N}$. For any $\lambda > 0$, consider the following exponential moment:

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\lambda\langle x, (\hat{\tilde{A}}^v_{0,N})y\rangle\right)\middle|\tilde{\mathcal{M}}^{0,N-1}\cap\tilde{\mathcal{D}}^{0,N-1}\right]\\
&= \frac{\mathbb{E}\left[\exp\left(\lambda\langle x, (\hat{\tilde{A}}^v_{0,N})y\rangle\right)\mathbb{1}\left(\tilde{\mathcal{M}}^{0,N-1}\cap\tilde{\mathcal{D}}^{0,N-1}\right)\right]}{\mathbb{P}\left(\tilde{\mathcal{M}}^{0,N-1}\cap\tilde{\mathcal{D}}^{0,N-1}\right)}\\
&= \frac{\mathbb{E}\left[\exp\left(\frac{\lambda}{N}\langle x, \Delta_0 F_{0,N}y\rangle + \lambda\langle x, \Gamma_{1,N-1}y\rangle\right)\mathbb{1}\left(\tilde{\mathcal{M}}^{0,N-1}\cap\tilde{\mathcal{D}}^{0,N-1}\right)\right]}{\mathbb{P}\left(\tilde{\mathcal{M}}^{0,N-1}\cap\tilde{\mathcal{D}}^{0,N-1}\right)}
\end{aligned} \tag{5.43}$$

Here, we note that Δ_0 is independent of $\tilde{\mathcal{M}}^{0,N-1}$, $F_{0,N}$ and $\tilde{D}^{1,N-1}$. We integrate out Δ_0 in Equation (5.43) using item 2 of Lemma 5.7.1 by using the fact that $\tilde{\mathcal{D}}^{0,N-1} = \tilde{\mathcal{D}}^{1,N-1}\cap\tilde{\mathcal{D}}^0_{-0}$ to show:

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\lambda\langle x, (\hat{\tilde{A}}^v_{0,N})y\rangle\right)\middle|\tilde{\mathcal{M}}^{0,N-1}\cap\tilde{\mathcal{D}}^{0,N-1}\right]\\
&\leq \frac{\mathbb{E}\left[\exp\left(\gamma\frac{\lambda^2 C_\mu}{N^2}\langle x, \Sigma x\rangle\|F_{0,N}y\|^2 + \lambda\langle x, \Gamma_{1,N-1}y\rangle\right)\mathbb{1}\left(\tilde{\mathcal{M}}^{0,N-1}\cap\tilde{\mathcal{D}}^{1,N-1}\right)\right]}{\mathbb{P}\left(\tilde{\mathcal{M}}^{0,N-1}\cap\tilde{\mathcal{D}}^{0,N-1}\right)}
\end{aligned} \tag{5.44}$$

We use the fact that $F_{0,N} = I + \tilde{H}^1_{0,B-1}F_{1,N}$ to conclude: $\|F_{0,N}y\|^2 = \|y\|^2 + 2\langle y, \tilde{H}^1_{0,B-1}F_{1,N}y\rangle + \langle y, F^T_{1,N}\tilde{H}^{1,\top}_{0,B-1}\tilde{H}^1_{0,B-1}F_{1,N}y\rangle$. Under the event $\tilde{\mathcal{M}}^{0,N-1}\cap\tilde{\mathcal{D}}^{1,N-1}$, we have: $\|\tilde{H}^1_{0,B-1}\| \leq 1$ and $\|F_{1,N}\| \leq \alpha$. Therefore, $\|F_{0,N}y\|^2 \leq \|y\|^2(1+2\alpha) + \langle y, F^T_{1,N}\tilde{H}^{1,\top}_{0,B-1}\tilde{H}^1_{0,B-1}F_{1,N}y\rangle$. Using this in Equation (5.44), we conclude:

$$\begin{aligned}
&\mathbb{P}\left(\tilde{\mathcal{M}}^{0,N-1}\cap\tilde{\mathcal{D}}^{0,N-1}\right)\mathbb{E}\left[\exp\left(\lambda\langle x, (\hat{\tilde{A}}^v_{0,N})y\rangle\right)\middle|\tilde{\mathcal{M}}^{0,N-1}\cap\tilde{\mathcal{D}}^{0,N-1}\right]\\
&\leq \mathbb{E}\left[\exp\left(\Omega + \lambda\langle x, \Gamma_{1,N-1}y\rangle\right)\mathbb{1}\left(\tilde{\mathcal{M}}^{0,N-1}\cap\tilde{\mathcal{D}}^{1,N-1}\right)\right]\\
&\leq \mathbb{E}\left[\exp\left(\Omega + \lambda\langle x, \Gamma_{1,N-1}y\rangle\right)\mathbb{1}\left(\tilde{\mathcal{M}}^{1,N-1}\cap\tilde{\mathcal{D}}^{1,N-1}\right)\right],
\end{aligned} \tag{5.45}$$

where $\Omega := \gamma\frac{\lambda^2 C_\mu}{N^2}\langle x, \Sigma x\rangle(1+2\alpha)\|y\|^2 + \gamma\frac{\lambda^2 C_\mu}{N^2}\langle x, \Sigma x\rangle\langle y, F^T_{1,N}\tilde{H}^{1,\top}_{0,B-1}\tilde{H}^1_{0,B-1}F_{1,N}y\rangle$. In the last step we have used the fact that $\tilde{\mathcal{M}}^{0,N-1}\cap\tilde{\mathcal{D}}^{1,N-1} \subseteq \tilde{\mathcal{M}}^{1,N-1}\cap\tilde{\mathcal{D}}^{1,N-1}$. We continue just like before but use item 1 of Lemma 5.7.1 instead of item 2 to keep peeling terms of the form $\langle x, \Delta_{t-1}F_{t-1,N}y\rangle$ to

conclude:

$$\mathbb{E}\left[\exp\left(\lambda\langle x,(\hat{\tilde{A}}^v_{0,N})y\rangle\right)\middle|\tilde{\mathcal{M}}^{0,N-1}\cap\tilde{\mathcal{D}}^{0,N-1}\right]\leq 2\exp\left(\gamma\frac{\lambda^2 C_\mu}{N}\langle x,\Sigma x\rangle(1+2\alpha)\|y\|^2\right)$$
$$\leq 2\exp\left(\gamma\frac{\lambda^2 C_\mu}{N}\sigma_{\max}(\Sigma)(1+2\alpha)\right) \tag{5.46}$$

Where $\sigma_{\max}(\Sigma)$ is the maximum eigenvalue of the covariance matrix Σ. Here we have used the assumption that $\mathbb{P}\left(\tilde{\mathcal{M}}^{0,N-1}\cap\tilde{\mathcal{D}}^{0,N-1}\right)\geq\frac{1}{2}$ and the fact that $\|x\|=\|y\|=1$. We apply Chernoff bound to $\langle x,(\hat{\tilde{A}}^v_{0,N})y\rangle$ using Equation (5.46) to conclude that for any $\beta,\lambda\in\mathbb{R}^+$

$$\mathbb{P}\left(\langle x,(\hat{\tilde{A}}^v_{0,N})y\rangle>\beta\middle|\tilde{\mathcal{M}}^{0,N-1}\cap\tilde{\mathcal{D}}^{0,N-1}\right)\leq 2\exp\left(\gamma\frac{\lambda^2 C_\mu}{N}\sigma_{\max}(\Sigma)\rangle(1+2\alpha)-\beta\lambda\right) \tag{5.47}$$

Choose $\lambda=\frac{N\beta}{2\gamma C_\mu\sigma_{\max}(\Sigma)(1+2\alpha)}$ to conclude:

$$\mathbb{P}\left(\langle x,(\hat{\tilde{A}}^v_{0,N})y\rangle>\beta\middle|\tilde{\mathcal{M}}^{0,N-1}\cap\tilde{\mathcal{D}}^{0,N-1}\right)\leq 2\exp\left(-\frac{\beta^2 N}{4\gamma C_\mu\sigma_{\max}(\Sigma)(1+2\alpha)}\right)$$

We now apply an ϵ net argument just like in Lemma 5.7.5. Suppose $\mathcal{N}$ is a 1/4-net of the sphere in $\mathbb{R}^d$. By Corollary 4.2.13 in [178], we can choose $|\mathcal{N}|\leq 12^d$. By Exercise 4.4.3 in [178], we conclude that:

$$\|\hat{\tilde{A}}^v_{0,N}\|\leq 2\sup_{x,y\in\mathcal{N}}\langle x,(\hat{\tilde{A}}^v_{0,N})y\rangle.$$

Therefore,

$$\begin{aligned}
&\mathbb{P}\left(\|\hat{\tilde{A}}^v_{0,N}\|>\beta\middle|\tilde{\mathcal{M}}^{0,N-1}\cap\tilde{\mathcal{D}}^{0,N-1}\right)\\
&\leq\mathbb{P}\left(\sup_{x,y\in\mathcal{N}}\langle x,(\hat{\tilde{A}}^v_{0,N})y\rangle>\frac{\beta}{2}\middle|\tilde{\mathcal{M}}^{0,N-1}\cap\tilde{\mathcal{D}}^{0,N-1}\right)\\
&\leq|\mathcal{N}|^2\sup_{x,y\in\mathcal{N}}\mathbb{P}\left(\langle x,(\hat{\tilde{A}}^v_{0,N})y\rangle>\frac{\beta}{2}\middle|\tilde{\mathcal{M}}^{0,N-1}\cap\tilde{\mathcal{D}}^{0,N-1}\right)\\
&\leq 2(12)^{2d}\exp\left(-\frac{\beta^2 N}{16\gamma C_\mu\sigma_{\max}(\Sigma)(1+2\alpha)}\right)\leq\exp\left(c_1 d-\frac{\beta^2 N}{16\gamma C_\mu\sigma_{\max}(\Sigma)(1+2\alpha)}\right)
\end{aligned}\tag{5.48}$$

□

5.8 Parameter error: Proof of theorem 5.3.1

In this section, we formally prove the bounds on $\mathcal{L}_{\mathsf{op}}(;A^*,\mu)$, by combining several operator norm inequalities that we prove in Section 5.7. As mentioned previously, we will just focus on the algorithmic iterates from the coupled process $(\tilde{A}_j^{t-1})$. Recall the output $\tilde{A}_B^{t-1}$ after the $t-1$-th buffer from Equation (5.23). For any initial buffer index $a\in\{0,1,\ldots,N-1\}$, the tail averaged output of

our algorithm is:

$$\hat{\tilde{A}}_{a,N} := \frac{1}{N-a} \sum_{t=a+1}^{N} \tilde{A}_B^{t-1}.$$

Recall the quantities $\tilde{A}_B^{t-1,v}$ and $\tilde{A}_B^{t-1,b}$ as defined in (5.24) and (5.25). We can use this decomposition to write:

$$\hat{\tilde{A}}_{a,N} - A^* = \hat{\tilde{A}}^b_{a,N} - A^* + \hat{\tilde{A}}^v_{a,N}.$$

Here $\hat{\tilde{A}}^b_{a,N} - A^* := \frac{1}{N-a} \sum_{t=a+1}^{N} \left(\tilde{A}_B^{t-1,b} - A^* \right)$ denotes the bias part and $\hat{\tilde{A}}^v_{a,N} := \frac{1}{N-a} \sum_{t=a+1}^{N} \left(\tilde{A}_B^{t-1,v} \right)$ denotes the variance part.

5.8.1 Variance

Note that

$$\hat{\tilde{A}}^v_{a,N} = \frac{N}{N-a} \left(\hat{\tilde{A}}^v_{0,N} \right) - \frac{a}{N-a} \left(\hat{\tilde{A}}^v_{0,a} \right) \tag{5.49}$$

Now, we apply Theorem 5.7.7 with δ in the definition of $\tilde{\mathcal{M}}^{0,N-1}$ to be $\frac{1}{T^\upsilon}$ for some fixed $\upsilon \geq 1$. We conclude that conditioned on the event $\tilde{\mathcal{M}}^{0,N-1} \cap \tilde{\mathcal{D}}^{0,N-1}$, with probability at least $1 - \frac{1}{T^\upsilon}$, we have:

$$\|\hat{\tilde{A}}^v_{0,N}\| \leq C \sqrt{\frac{\gamma(d + \upsilon \log T)^2 \sigma_{\max}(\Sigma)}{N}} + C \sqrt{\frac{(d + \upsilon \log T)\sigma_{\max}(\Sigma)}{NB\sigma_{\min}(G)}}.$$

Similarly, applying Theorem 5.7.7 with $N = a$ shows that with probability at least $1 - \frac{1}{T^\upsilon}$ conditioned on the event $\tilde{\mathcal{M}}^{0,N-1} \cap \tilde{\mathcal{D}}^{0,N-1}$:

$$\|\hat{\tilde{A}}^v_{0,a}\| \leq C \sqrt{\frac{\gamma(d + \upsilon \log T)^2 \sigma_{\max}(\Sigma)}{a}} + C \sqrt{\frac{(d + \upsilon \log T)\sigma_{\max}(\Sigma)}{aB\sigma_{\min}(G)}}.$$

Here, the constant C depends only on C_μ. We also note that when we pick $\gamma BR \leq C_0$ where $R \gtrsim \mathrm{Tr}(G) + \upsilon \log T$, the first term in the equations above becomes smaller than the second term. Therefore, under this assumption we can simplify the expressions to:

$$\|\hat{\tilde{A}}^v_{0,N}\| \leq C \sqrt{\frac{(d + \upsilon \log T)\sigma_{\max}(\Sigma)}{NB\sigma_{\min}(G)}}. \tag{5.50}$$

$$\|\hat{\tilde{A}}^v_{0,a}\| \leq C \sqrt{\frac{(d + \upsilon \log T)\sigma_{\max}(\Sigma)}{aB\sigma_{\min}(G)}}. \tag{5.51}$$

Applying Equations (5.50) and (5.51) to Equation (5.49) we conclude that conditioned on the

event $\tilde{\mathcal{M}}^{0,N-1} \cap \tilde{\mathcal{D}}^{0,N-1}$, with probability at least $1 - \frac{2}{T^{\upsilon}}$, we have:

$$\begin{aligned}\|\hat{\tilde{A}}^{\upsilon}_{a,N}\| &\leq \frac{N}{N-a}\|\left(\hat{\tilde{A}}^{\upsilon}_{0,N}\right)\| + \frac{a}{N-a}\|\left(\hat{\tilde{A}}^{\upsilon}_{0,a}\right)\| \\ &\leq \frac{CN}{N-a}\sqrt{\frac{(d+\upsilon\log T)\sigma_{\max}(\Sigma)}{NB\sigma_{\min}(G)}} + \frac{Ca}{N-a}\sqrt{\frac{(d+\upsilon\log T)\sigma_{\max}(\Sigma)}{aB\sigma_{\min}(G)}}. \end{aligned} \tag{5.52}$$

Choose $a < N/2$. Since

$$\mathbb{P}\left[\tilde{\mathcal{M}}^{0,N-1} \cap \tilde{\mathcal{D}}^{0,N-1}\right] \geq 1 - (\frac{1}{T^{\upsilon}} + \frac{1}{T^{\alpha}})$$

we have

$$\begin{aligned}&\mathbb{P}\left[\|\hat{\tilde{A}}^{\upsilon}_{a,N}\| > C\sqrt{\frac{(d+\upsilon\log T)\sigma_{\max}(\Sigma)}{(N-a)B\sigma_{\min}(G)}}\right] \\ &\leq \frac{1}{T^{\alpha}} + \frac{3}{T^{\upsilon}}\end{aligned} \tag{5.53}$$

5.8.2 Bias

We now consider the bias term:$\hat{\tilde{A}}^{b}_{a,N} - A^* := \frac{1}{N-a}\sum_{t=a+1}^{N}\left(\tilde{A}_B^{t-1,b} - A^*\right)$. First note that, from equation (5.24), we have

$$\left\|\hat{\tilde{A}}^{b}_{a,N} - A^*\right\| \leq \frac{1}{N-a}\sum_{t=a+1}^{N}\|A_0 - A^*\|\left\|\prod_{s=0}^{t-1}\tilde{H}^{s}_{0,B-1}\right\| \tag{5.54}$$

Now from lemma 5.7.5, if $a > c_1\left(d + \log\frac{N}{\delta}\right)$ then conditional on $\tilde{\mathcal{D}}^{0,N-1}$ with probability at least $1-\delta$, for all $a+1 \leq t \leq N$ we have

$$\left\|\prod_{s=0}^{t-1}\tilde{H}^{s}_{0,B-1}\right\| \leq 2\left(1 - \gamma B\sigma_{\min}(G)\right)^{c_2 t} \tag{5.55}$$

Note that in lemma 5.7.5 we only condition on $\tilde{\mathcal{D}}^{0,t-1}$ but due to buffer independence and that $\mathbb{P}\left[\tilde{\mathcal{D}}^{0,N-1}\right] \geq 1 - \frac{1}{T^{\alpha}}$ we can condition on $\tilde{\mathcal{D}}^{0,N-1}$.

Note that in the proof of lemma 5.7.5 the constant c_2 is actually at most 1 i.e., $0 < c_2 \leq 1$. Hence from Bernoulli's inequality, for $x < 1$

$$(1-x)^{c_2} \leq 1 - c_2 x$$

Thus conditional on $\tilde{\mathcal{D}}^{0,N-1}$ with probability at least $1-\delta$

$$\begin{aligned}\left\|\hat{\tilde{A}}^b_{a,N} - A^*\right\| &\leq \frac{\|A_0 - A^*\|}{N-a} \sum_{t=a+1}^{\infty} 2\left(1-\gamma B\sigma_{\min}(G)\right)^{c_2 t} \\ &= 2\frac{\|A_0 - A^*\|}{N-a} \frac{\left(1-\gamma B\sigma_{\min}(G)\right)^{c_2 a}}{c_2\gamma B\sigma_{\min}(G)} \\ &\leq c_3 \frac{\|A_0 - A^*\|}{N-a} \frac{e^{-c_2 a\gamma B\sigma_{\min}(G)}}{\gamma B\sigma_{\min}(G)} \end{aligned} \tag{5.56}$$

Hence choosing $\delta = \frac{1}{T^\upsilon}$ we have for $a > c_1\left(d + \log\frac{N}{\delta}\right)$

$$\mathbb{P}\left[\left\|\hat{\tilde{A}}^b_{a,N} - A^*\right\| > c_3 \frac{\|A_0 - A^*\|}{N-a} \frac{e^{-c_2 a\gamma B\sigma_{\min}(G)}}{\gamma B\sigma_{\min}(G)}\right] \leq \frac{1}{T^\alpha} + \frac{1}{T^\upsilon} \tag{5.57}$$

Define β_b as

$$\beta_b = c_3 \frac{1}{N-a} \frac{e^{-c_2 a\gamma B\sigma_{\min}(G)}}{\gamma B\sigma_{\min}(G)} \tag{5.58}$$

Thus by union bound and equations (5.53) and (5.57) we get

$$\begin{aligned}&\mathbb{P}\left[\left\|\hat{\tilde{A}}_{a,N} - A^*\right\| > C\sqrt{\frac{(d+\upsilon\log T)\sigma_{\max}(\Sigma)}{(N-a)B\sigma_{\min}(G)}} + \beta_b\|A_0 - A^*\|\right] \\ &\leq \frac{2}{T^\alpha} + \frac{4}{T^\upsilon}\end{aligned} \tag{5.59}$$

Now from lemma 6.8.3 we see that on the event $\hat{\mathcal{D}}^{0,N-1}$

$$\left\|\hat{A}_{a,N} - \hat{\tilde{A}}_{a,N}\right\| \leq c\gamma^2 R^2 T^2 \left\|A^{*u}\right\| \tag{5.60}$$

Since $\mathbb{P}\left[\hat{\mathcal{D}}^{0,N-1}\right] \geq 1 - \frac{1}{T^\alpha}$, we obtain

$$\mathbb{P}\left[\left\|\hat{A}_{a,N} - \hat{\tilde{A}}_{a,N}\right\| \leq c\gamma^2 R^2 T^2 \left\|A^{*u}\right\|\right] \geq 1 - \frac{1}{T^\alpha} \tag{5.61}$$

Therefore choosing $\delta = \frac{1}{T^\upsilon}$ we have for $N/2 > a > c_1\left(d + \log\frac{N}{\delta}\right)$

$$\begin{aligned}&\mathbb{P}\left[\left\|\hat{A}_{a,N} - A^*\right\| > C\sqrt{\frac{(d+\upsilon\log T)\sigma_{\max}(\Sigma)}{(N-a)B\sigma_{\min}(G)}} + \beta_b\|A_0 - A^*\| + c_4\gamma^2 R^2 T^2\left\|A^{*u}\right\|\right] \\ &\leq \frac{3}{T^\alpha} + \frac{4}{T^\upsilon}\end{aligned} \tag{5.62}$$

where β_b is defined in (5.58).

The theorem follows by adjusting the constants (in choosing δ) such the above probability is at most $\frac{3}{T^\alpha} + \frac{1}{2T^\upsilon}$ and then choosing υ such that $\frac{3}{T^\alpha} \leq \frac{1}{2T^\upsilon}$.

5.9 Preliminary results for prediction error

5.9.1 Bias variance analysis of last and average iterates

In this section, our goal is to provide a PSD upper bound on

$$\mathbb{E}\left[\left(\tilde{A}_B^{t-1}-A^*\right)^\top\left(\tilde{A}_B^{t-1}-A^*\right)\right], \mathbb{E}\left[\left(\hat{\tilde{A}}_{a,N}-A^*\right)^\top\left(\hat{\tilde{A}}_{a,N}-A^*\right)\right]$$

using the bias variance decomposition in (5.23) and (5.27). This bound leads to Theorem 5.9.2 which is critical for our parameter error proof (Theorem 5.3.1).

5.9.2 Variance of the last iterate

The goal of this section is to bound error due to $\left(\tilde{A}_B^{t-1,v}\right)$. For brevity, we will introduce the following notation:

$$\tilde{V}_{t-1}=\mathbb{E}\left[\left(\tilde{A}_B^{t-1,v}\right)^\top\left(\tilde{A}_B^{t-1,v}\right)1\left[\tilde{\mathcal{D}}^{0,t-1}\right]\right]. \tag{5.63}$$

The following proposition is the main result of this section.

Proposition 1. *Let $\gamma\le\frac{1}{2R}$. Let the noise covariance be $\mathbb{E}\left[\eta_t\eta_t^T\right]=\Sigma$. Then,*

$$\tilde{V}_{t-1}\preceq\frac{\gamma\operatorname{Tr}(\Sigma)}{1-\gamma R}\left[I-\mathbb{E}\left[\left(\prod_{s=1}^{t}\tilde{H}_{0,B-1}^{t-s,\top}\right)\left(\prod_{s=t}^{1}\tilde{H}_{0,B-1}^{t-s}\right)1\left[\tilde{\mathcal{D}}^{0,t-1}\right]\right]\right]+c_1\gamma^2 d\sigma_{\max}(\Sigma)(Bt)^2\frac{1}{T^{\alpha/2}}I,$$
$$\tilde{V}_{t-1}\succeq\gamma\operatorname{Tr}(\Sigma)\left[I-\mathbb{E}\left[\left(\prod_{s=1}^{t}\tilde{H}_{0,B-1}^{t-s,\top}\right)\left(\prod_{s=t}^{1}\tilde{H}_{0,B-1}^{t-s}\right)1\left[\tilde{\mathcal{D}}^{0,t-1}\right]\right]\right]-c_4\gamma^2 d\sigma_{\max}(\Sigma)(Bt)^2\frac{1}{T^{\alpha/2}}I,$$

for some absolute constants $c_i>0$, $1\le i\le 4$.

We refer to Section 5.9.6 for a full proof. Note that we have, $\frac{1}{1-\gamma\|X\|^2}\le 2$.

Corollary 1. *In the same setting as Proposition 1, we have:*

$$\tilde{V}_{t-1}\preceq c_1\gamma\operatorname{Tr}(\Sigma)I+c_2\gamma^2 d\sigma_{\max}(\Sigma)(Bt)^2\frac{1}{T^{\alpha/2}}I, \tag{5.64}$$

for some constants $c_1,c_2>0$. If $T^{\alpha/2}>T^2$, then $V_{t,1}\preceq c\gamma d\sigma_{\max}I$, for some constant $c>0$.

5.9.3 Variance of the average iterate

In this section we are interested in bounding: $\mathbb{E}\left[\left(\hat{\tilde{A}}_{a,N}^{v}\right)^\top\left(\hat{\tilde{A}}_{a,N}^{v}\right)1\left[\tilde{\mathcal{D}}^{0,N-1}\right]\right]$, for $a=\theta N$ with $0\le\theta<1$, where,

$$\hat{\tilde{A}}_{a,N}^{v}=\frac{1}{N-a}\sum_{t=a+1}^{N}\tilde{A}_B^{t-1,v}, \tag{5.65}$$

and further, recall that $T = N(B+u)$. The main bound in this section is given in Proposition 2. Note that we have,

$$\begin{aligned}
&\mathbb{E}\left[\left(\hat{\tilde{A}}^v_{a,N}\right)^\top \left(\hat{\tilde{A}}^v_{a,N}\right) 1\left[\tilde{\mathcal{D}}^{0,N-1}\right]\right] \\
&= \frac{1}{(N-a)^2}\sum_{t=a+1}^{N} \mathbb{E}\left[\left(\tilde{A}_B^{t-1,v}\right)^\top \left(\tilde{A}_B^{t-1,v}\right) 1\left[\tilde{\mathcal{D}}^{0,N-1}\right]\right] \\
&\quad + \frac{1}{(N-a)^2}\sum_{t_1\neq t_2} \mathbb{E}\left[\left(\tilde{A}_B^{t_1-1,v}\right)^\top \left(\tilde{A}_B^{t_2-1,v}\right) 1\left[\tilde{\mathcal{D}}^{0,N-1}\right]\right] \qquad (5.66)
\end{aligned}$$

Proposition 2. *Let $\gamma \leq \min\{\frac{c}{6RB}\frac{1}{2R}\}$ for $0<c<1$. Then for $\hat{\tilde{A}}^v_{a,N}$ defined in (5.65), there are constants $c_1, c_2 > 0$ such that if $T^{\alpha/2} > c_1\frac{\sqrt{M_4}}{\sigma_{\min}(G)}$, then:*

$$\begin{aligned}
&\mathbb{E}\left[\left(\hat{\tilde{A}}^v_{a,N}\right)^\top \left(\hat{\tilde{A}}^v_{a,N}\right) 1\left[\tilde{\mathcal{D}}^{0,N-1}\right]\right] \\
&\preceq \frac{1}{(N-a)^2}\sum_{t=a+1}^{N}\left[\tilde{V}_{t-1}\left(\sum_{s=0}^{N-t}\mathcal{H}^s\right) + \left(\sum_{s=0}^{N-t}\mathcal{H}^s\right)^\top \tilde{V}_{t-1}\right] + c_2\delta I \qquad (5.67)\\
&= \frac{1}{(N-a)^2}\sum_{t=a+1}^{N}\left[\tilde{V}_{t-1}\left(I-\mathcal{H}\right)^{-1} + \left(I-\mathcal{H}^\top\right)^{-1}\tilde{V}_{t-1}\right] + c_2\delta I + \\
&\quad \frac{1}{(N-a)^2}\sum_{t=a+1}^{N}\left[\tilde{V}_{t-1}\left(I-\mathcal{H}\right)^{-1}\mathcal{H}^{N-t+1} + \left(\mathcal{H}^\top\right)^{N-t+1}\left(I-\mathcal{H}^\top\right)^{-1}\tilde{V}_{t-1}\right] \qquad (5.68)
\end{aligned}$$

and,

$$\delta \equiv \delta(N,B,R) = \gamma^2 T^2 R d \sigma_{\max}(\Sigma)\frac{1}{T^{\alpha/2}} \qquad (5.69)$$

and $\mathcal{H}$ is given by,

$$\mathcal{H} = \mathbb{E}\left[\prod_{j=0}^{B-1}\left(I - 2\gamma\tilde{X}^0_{-j}\tilde{X}^{0,\top}_{-j}\right) 1\left[\cap_{j=0}^{B-1}\left\{\|\tilde{X}^0_{-j}\|^2 \leq R\right\}\right]\right], \qquad (5.70)$$

with $\tilde{X}_0$ sampled from the stationary distribution π and $\tilde{X}_t$ follows the $\mathsf{VAR}(A^*,\mu)$.

See section 5.9.7 for the proof.

5.9.4 Bias of the last iterate

In this we will analyze the bias term of the last iterate. That is we want to bound:

$$\mathbb{E}\left[\left(\tilde{A}_B^{t-1,b} - A^*\right)^\top \left(\tilde{A}_B^{t-1,b} - A^*\right) 1\left[\tilde{\mathcal{D}}^{0,t-1}\right]\right].$$

Where $\left(\tilde{A}_B^{t-1,b} - A^*\right)$ is defined in (5.24).

Theorem 5.9.1. *Let $\gamma RB \leq \frac{c}{6}$ for some $0 < c < 1$ with B such that $\gamma R \leq \frac{1}{2}$. Then there are constants $c_1, c_2, c_3 > 0$ such that if $T^{\alpha/2} > c_1 \frac{\sqrt{M_4}}{\sigma_{\min}(G)}$ (where $M_4 = \mathbb{E}\left[\|\tilde{X}_{-0}^0\|^4\right]$) then*

$$\mathbb{E}\left[\left(\tilde{A}_B^{t-1,b} - A^*\right)^\top \left(\tilde{A}_B^{t-1,b} - A^*\right) 1\left[\tilde{\mathcal{D}}^{0,t-1}\right]\right] \preceq \|A_0 - A^*\|^2 \left(1 - c_2\gamma B\sigma_{\min}(G)\right)^t I \quad (5.71)$$

See section 5.9.9 for the proof.

5.9.5 Bias of the tail-averaged iterate

We define the tail averaged bias as

$$\hat{\tilde{A}}_{a,N}^b = \frac{1}{N-a}\sum_{t=a+1}^{N} \tilde{A}_B^{t-1,b} \quad (5.72)$$

Theorem 5.9.2. *Let $\gamma RB \leq \frac{c}{6}$ for some $0 < c < 1$ and B such that $\gamma R \leq \frac{1}{2}$. There exist constants $c_1, c_2 > 0$ such that if $T = N(B+u)$ satisfies $T^{\alpha/2} > c_1 \frac{\sqrt{M_4}}{\sigma_{\min}(G)}$ then for $a = \theta N$ with $0 < \theta < 1$ we have*

$$\left\|\mathbb{E}\left[\left(\hat{\tilde{A}}_{a,N}^b - A^*\right)^\top \left(\hat{\tilde{A}}_{a,N}^b - A^*\right) 1\left[\tilde{\mathcal{D}}^{0,N-1}\right]\right]\right\| \leq$$
$$c_2 \frac{1}{B(N-a)} \frac{e^{-c_3 B\gamma\sigma_{\min}(G)a}}{\gamma\sigma_{\min}(G)} \|A_0 - A^*\|^2 \quad (5.73)$$

See section 5.9.10 for the proof.

5.9.6 Proof of proposition 1

Proof of proposition 1. First note that

$$\left(\tilde{A}_b^{t-1,v}\right)^\top \left(\tilde{A}_b^{t-1,v}\right) = \sum_{r=1}^{t}\sum_{j=0}^{B-1} \widetilde{\mathrm{Dg}}(t,r,j) + \sum_{r_1,r_2=1}^{t}\sum_{j_1,j_2=0}^{B-1} \widetilde{\mathrm{Cr}}(t,r_1,j_1,r_2,j_2) \quad (5.74)$$

where

$$\widetilde{\mathrm{Dg}}(t,r,j) = 4\gamma^2 \left\|\eta_{-j}^{t-r}\right\|^2 \cdot \left(\prod_{s=1}^{r-1} \tilde{H}_{0,B-1}^{t-s,\top}\right) \tilde{H}_{j+1,B-1}^{t-r,\top} \tilde{X}_{-j}^{t-r} \tilde{X}_{-j}^{t-r,\top} \tilde{H}_{j+1,B-1}^{t-r} \left(\prod_{s=r-1}^{1} \tilde{H}_{0,B-1}^{t-s}\right) \tag{5.75}$$

$$\widetilde{\mathrm{Cr}}(t,r_1,j_1,r_2,j_2) = 4\gamma^2 \left(\eta_{-j_1}^{t-r_1} \tilde{X}_{-j_1}^{t-r_1,\top} \tilde{H}_{j_1+1,B-1}^{t-r_1} \prod_{s=r_1-1}^{1} \tilde{H}_{0,B-1}^{t-s}\right)^{\top} \cdot \left(\eta_{-j_2}^{t-r_2} \tilde{X}_{-j_2}^{t-r_2,\top} \tilde{H}_{j_2+1,B-1}^{t-r_2} \prod_{s=r_2-1}^{1} \tilde{H}_{0,B-1}^{t-s}\right) \tag{5.76}$$

denote the diagonal and cross terms respectively.

We begin by noting the following two facts about $\left(\tilde{A}_b^{t-1,v}\right)$:

- It has zero mean

$$\mathbb{E}\left[\left(\tilde{A}_B^{t-1,v}\right)\right] = 0 \tag{5.77}$$

- Let $(r_1, j_1) \neq (r_2, j_2)$. Then

$$\mathbb{E}\left[\widetilde{\mathrm{Cr}}(t,r_1,j_1,r_2,j_2)\right] = 0 \tag{5.78}$$

This follows because, assuming $r_1 > r_2$, the term $\eta_{-j_1}^{t-r_1} \tilde{X}_{-j_1}^{t-r_1,\top} \tilde{H}_{j_1+1,B-1}^{t-r_1}$ is independent of everything else in that expression, and that $\eta_{-j_1}^{t-r_1}$ is independent of $\tilde{X}_{-j_1}^{t-r_1,\top} \tilde{H}_{j_1+1,B-1}^{t-r_1}$. A similar argument can be made for the case when $r_1 = r_2$ but $j_1 \neq j_2$.

But we are interested in expectation on the event $\tilde{\mathcal{D}}^{0,t-1}$.

We will bound the expectation of cross terms in the following lemma.

Lemma 5.9.3. *We have*

$$\left\|\mathbb{E}\left[\sum_{r_1,r_2}\sum_{j_1,j_2} \widetilde{\mathrm{Cr}}(t,r_1,j_1,r_2,j_2)\right] 1\left[\tilde{\mathcal{D}}^{0,t-1}\right]\right\| \leq 8(Bt)^2\gamma^2 R \operatorname{Tr}(\Sigma) \frac{1}{T^{\alpha/2}} \tag{5.79}$$

Proof. Let

Consider a single cross term: $\widetilde{\mathrm{Cr}}(t,r_1,j_1,r_2,j_2)$ and without loss of generality, assume that either $r_1 > r_2$ or $r_1 = r_2$ but $j_1 < j_2$. In either case, we note that $\eta_{-j_1}^{t-r_1}$ is unconditionally independent of all other terms present in $\widetilde{\mathrm{Cr}}(t,r_1,j_1,r_2,j_2)$. The main problem here is to bound the expectation over the event $\tilde{\mathcal{D}}^{0,t-1}$. For the sake of convenience, only in this proof, we will define the following notation:

$$\widetilde{\mathrm{Cr}}(t, r_1, j_1, r_2, j_2) = E_1 \eta_{-j_1}^{t-r_1,\top} \eta_{-j_2}^{t-r_2} E_2$$

Where E_1 and E_2 are random matrices defined according to the definition of $\widetilde{\mathrm{Cr}}(t, r_1, j_1, r_2, j_2)$ and are unconditionally independent of $\eta_{-j_1}^{t-r_1,\top}$. Let $\mathcal{F}_E = \sigma(E_1, E_2, \eta_{-j_2}^{t-r_2})$. Note that when conditioned on the event $\tilde{\mathcal{D}}^{0,t-1}$, we must have the event $\mathcal{M} := \{\|E_1\| \leq 4\gamma^2\sqrt{R}\} \cap \{\|E_2\| \leq \sqrt{R}\}$ almost surely. Therefore, we conclude:

$$\begin{aligned}
\mathbb{E}\left[\widetilde{\mathrm{Cr}}(t, r_1, j_1, r_2, j_2) 1\left[\tilde{\mathcal{D}}^{0,t-1}\right]\right] &= \mathbb{E}\left[\widetilde{\mathrm{Cr}}(t, r_1, j_1, r_2, j_2) 1\left[\tilde{\mathcal{D}}^{0,t-1}\right] 1\left[\mathcal{M}\right]\right] \\
&= \mathbb{E}\left[1\left[\mathcal{M}\right] E_1 \mathbb{E}\left[\eta_{-j_1}^{t-r_1,\top} 1\left[\tilde{\mathcal{D}}^{0,t-1}\right] \middle| \mathcal{F}_E\right] \eta_{-j_2}^{t-r_2} E_2\right] \\
&\leq \mathbb{E}\left[1\left[\mathcal{M}\right] \|E_1\| \left\|\mathbb{E}\left[\eta_{-j_1}^{t-r_1,\top} 1\left[\tilde{\mathcal{D}}^{0,t-1}\right] \middle| \mathcal{F}_E\right]\right\| \|\eta_{-j_2}^{t-r_2}\| \|E_2\|\right] \\
&\leq 4\gamma^2 R \mathbb{E}\left[\left\|\mathbb{E}\left[\eta_{-j_1}^{t-r_1,\top} 1\left[\tilde{\mathcal{D}}^{0,t-1}\right] \middle| \mathcal{F}_E\right]\right\| \|\eta_{-j_2}^{t-r_2}\|\right] \qquad (5.80)
\end{aligned}$$

In the third step, we have used the fact that under the event $\mathcal{M}$, the norms $\|E_1\|, \|E_2\|$ are bounded. We will now bound $\mathbb{E}\left[\eta_{-j_1}^{t-r_1,\top} 1\left[\tilde{\mathcal{D}}^{0,t-1}\right] \middle| \mathcal{F}_E\right]$. Clearly, due to the unconditional independence, we must have:

$$\begin{aligned}
&\mathbb{E}\left[\eta_{-j_1}^{t-r_1,\top} \middle| \mathcal{F}_E\right] = 0 \\
&\implies \mathbb{E}\left[\eta_{-j_1}^{t-r_1,\top} 1\left[\tilde{\mathcal{D}}^{0,t-1}\right] \middle| \mathcal{F}_E\right] = -\mathbb{E}\left[\eta_{-j_1}^{t-r_1,\top} 1\left[\tilde{\mathcal{D}}^{0,t-1,C}\right] \middle| \mathcal{F}_E\right] \\
&\implies \left\|\mathbb{E}\left[\eta_{-j_1}^{t-r_1,\top} 1\left[\tilde{\mathcal{D}}^{0,t-1}\right] \middle| \mathcal{F}_E\right]\right\| \leq \sqrt{\mathrm{Tr}\,\Sigma} \sqrt{\mathbb{P}\left(\tilde{\mathcal{D}}^{0,t-1,C} \middle| \mathcal{F}_E\right)} \qquad (5.81)
\end{aligned}$$

In the last step, we have used Cauchy Schwarz inequality and the fact that $\eta_{-j_1}^{t-r_1,\top}$ is independent of $\mathcal{F}_E$. We combine the Equation above with Equation (5.80) and apply Jensen's inequality once again to conclude:

$$\left\|\mathbb{E}\left[\widetilde{\mathrm{Cr}}(t, r_1, j_1, r_2, j_2) 1\left[\tilde{\mathcal{D}}^{0,t-1}\right]\right]\right\| \leq 4\gamma^2 R\,\mathrm{Tr}(\Sigma) \sqrt{\mathbb{P}\left[\tilde{\mathcal{D}}^{0,t-1,C}\right]} \leq 4\gamma^2 R \frac{\mathrm{Tr}(\Sigma)}{T^{\alpha/2}} \qquad (5.82)$$

In the last step, we have used Lemma 5.5.3 to bound $\mathbb{P}\left(\tilde{\mathcal{D}}^{0,t-1,C}\right)$. Summing over all the indices (r_1, j_1, r_2, j_2), we conclude the statement of the lemma.

□

Lemma 5.9.4. *We have:*

$$\mathbb{E}\left[\sum_{r=1}^{t}\sum_{j=0}^{B-1}\widetilde{\mathrm{Dg}}(t,r,j)1\left[\tilde{\mathcal{D}}^{0,t-1}\right]\right] \preceq 4\gamma^2\,\mathrm{Tr}(\Sigma)\mathbb{E}\left[\sum_{r=1}^{t}\sum_{j=0}^{B-1}\left(\prod_{s=1}^{r-1}\tilde{H}_{0,B-1}^{t-s,\top}\right)\tilde{H}_{j+1,B-1}^{t-r,\top}\tilde{X}_{-j}^{t-r}\cdot \tilde{X}_{-j}^{t-r,\top}\tilde{H}_{j+1,B-1}^{t-r}\left(\prod_{s=r-1}^{1}\tilde{H}_{0,B-1}^{t-s}\right)1\left[\tilde{\mathcal{D}}^{0,t-1}\right]\right] + \delta_{\mathrm{Dg}}I \tag{5.83}$$

and

$$\mathbb{E}\left[\sum_{r=1}^{t}\sum_{j=0}^{B-1}\widetilde{\mathrm{Dg}}(t,r,j)1\left[\tilde{\mathcal{D}}^{0,t-1}\right]\right] \succeq 4\gamma^2\,\mathrm{Tr}(\Sigma)\mathbb{E}\left[\sum_{r=1}^{t}\sum_{j=0}^{B-1}\left(\prod_{s=1}^{r-1}\tilde{H}_{0,B-1}^{t-s,\top}\right)\tilde{H}_{j+1,B-1}^{t-r,\top}\tilde{X}_{-j}^{t-r}\cdot \tilde{X}_{-j}^{t-r,\top}\tilde{H}_{j+1,B-1}^{t-r}\left(\prod_{s=r-1}^{1}\tilde{H}_{0,B-1}^{t-s}\right)1\left[\tilde{\mathcal{D}}^{0,t-1}\right]\right] - \delta_{\mathrm{Dg}}I \tag{5.84}$$

where

$$\delta_{\mathrm{Dg}} \equiv \delta_{\mathrm{Dg}}(T,\Sigma,R,\mu_4) = 4\gamma^2(Bt)R\sqrt{\mu_4}\frac{1}{T^{\alpha/2}} \tag{5.85}$$

Proof. The evaluation of expectations is clear when there is no indicator $1\left[\tilde{\mathcal{D}}^{0,t-1}\right]$ within the expectation. We will now deal with it just like in the proof of Lemma 5.9.3. Consider $\widetilde{\mathrm{Dg}}(t,r,j)$. For the sake of convenience, only in this proof, we will use the following notation:

$$\widetilde{\mathrm{Dg}}(t,r,j) = 4\gamma^2\left\|\eta_{-j}^{t-r}\right\|^2 E\,.$$

Where the random PSD matrix E is unconditionally independent of η_{-j}^{t-r}. Let $\mathcal{M} = \{\|E\| \le R\}$. Conditioned on the event $\tilde{\mathcal{D}}^{0,t-1}$, the event $\mathcal{M}$ holds almost surely. Let $\mathcal{F}_E = \sigma(E)$.

Now consider:

$$\begin{aligned}\mathbb{E}\left[\widetilde{\mathrm{Dg}}(t,r,j)1\left[\tilde{\mathcal{D}}^{0,t-1}\right]\right] &= \mathbb{E}\left[\widetilde{\mathrm{Dg}}(t,r,j)1\left[\tilde{\mathcal{D}}^{0,t-1}\right]1\left[\mathcal{M}\right]\right]\\ &= 4\gamma^2\mathbb{E}\left[\left\|\eta_{-j}^{t-r}\right\|^2 E1\left[\tilde{\mathcal{D}}^{0,t-1}\right]1\left[\mathcal{M}\right]\right]\\ &= 4\gamma^2\mathbb{E}\left[\mathbb{E}\left[\left\|\eta_{-j}^{t-r}\right\|^2 1\left[\tilde{\mathcal{D}}^{0,t-1}\right]|\mathcal{F}_E\right]E1\left[\mathcal{M}\right]\right]\end{aligned} \tag{5.86}$$

It can be easily shown via similar techniques used in Lemma 5.9.3 that:

$$\mathrm{Tr}(\Sigma) - \sqrt{\mu_4}\sqrt{\mathbb{P}\left(\tilde{\mathcal{D}}^{0,t-1,C}\middle|\mathcal{F}_E\right)} \le \mathbb{E}\left[\left\|\eta_{-j}^{t-r}\right\|^2 1\left[\tilde{\mathcal{D}}^{0,t-1}\right]|\mathcal{F}_E\right] \le \mathrm{Tr}(\Sigma)$$

Using this in Equation (5.86), we conclude:

$$\begin{aligned}
\mathbb{E}\left[\widetilde{\mathrm{Dg}}(t,r,j)1\left[\tilde{\mathcal{D}}^{0,t-1}\right]\right] &\preceq 4\gamma^2 \operatorname{Tr}(\Sigma)\mathbb{E}\left[E1\left[\mathcal{M}\right]\right] \\
&= 4\gamma^2 \operatorname{Tr}(\Sigma)\mathbb{E}\left[E1\left[\mathcal{M}\right]1\left[\tilde{\mathcal{D}}^{0,t-1}\right] + E1\left[\mathcal{M}\right]1\left[\tilde{\mathcal{D}}^{0,t-1,C}\right]\right] \\
&= 4\gamma^2 \operatorname{Tr}(\Sigma)\mathbb{E}\left[E1\left[\tilde{\mathcal{D}}^{0,t-1}\right] + E1\left[\mathcal{M}\right]1\left[\tilde{\mathcal{D}}^{0,t-1,C}\right]\right] \\
&\preceq 4\gamma^2 \operatorname{Tr}\Sigma\mathbb{E}\left[E1\left[\hat{\mathcal{D}}^{0,t-1}\right]\right] + 4\gamma^2 \operatorname{Tr}(\Sigma)R\frac{I}{T^\alpha} \qquad (5.87)
\end{aligned}$$

In the third step, we have used the fact that $\tilde{\mathcal{D}}^{0,t-1} \subseteq \mathcal{M}$. In the last step we have used the fact that E is PSD and over the event $\mathcal{M}$, $E \preceq RI$. We have used Lemma 5.5.3 to bound $\mathbb{P}(\tilde{\mathcal{D}}^{0,t-1,C})$. Using a similar technique as above, we can show that:

$$\mathbb{E}\left[\widetilde{\mathrm{Dg}}(t,r,j)1\left[\tilde{\mathcal{D}}^{0,t-1}\right]\right] \succeq 4\gamma^2 \operatorname{Tr}\Sigma\mathbb{E}\left[E1\left[\tilde{\mathcal{D}}^{0,t-1}\right]\right] - 4\gamma^2\frac{\sqrt{\mu_4}R}{T^{\alpha/2}}I \qquad (5.88)$$

Note that $\frac{\sqrt{\mu_4}R}{T^{\alpha/2}} \geq \frac{\mathrm{Tr}(\Sigma)R}{T^\alpha}$. Summing over r, j and combining Equations (5.88) and (5.87), we conclude the result.

□

For convenience, define $K^s := \sum_{j=0}^{B-1} \tilde{H}^{s,\top}_{j+1,B-1}\tilde{X}^s_{-j}\tilde{X}^{s,\top}_{-j}\tilde{H}^s_{j+1,B-1}$

Claim 10. *Suppose $\gamma < \frac{1}{R}$. Under the event $\tilde{\mathcal{D}}^{0,t-1}$, for every $s \leq t-1$ we must have:*

$$\frac{I - \tilde{H}^{s,\top}_{0,B-1}\tilde{H}^s_{0,B-1}}{4\gamma} \preceq K^s \preceq \frac{I - \tilde{H}^{s,\top}_{0,B-1}\tilde{H}^s_{0,B-1}}{\hat{\gamma}}$$

Where $\hat{\gamma} = 4\gamma(1-\gamma R)$

Proof. In the entire proof, we suppose that the event $\tilde{\mathcal{D}}^{0,t-1}$ holds. Consider:

$$\begin{aligned}
&\tilde{H}^{s,\top}_{j,B-1}\tilde{H}^s_{j,B-1} + 4\gamma\tilde{H}^{s,\top}_{j+1,B-1}\tilde{X}^s_{-j}\tilde{X}^{s,\top}_{-j}\tilde{H}^s_{j+1,B-1} \\
&= \tilde{H}^{s,\top}_{j+1,B-1}\left(I - \left(4\gamma - 4\gamma^2\|\tilde{X}^s_{-j}\|^2\right)\tilde{X}^s_{-j}\tilde{X}^{s,\top}_{-j}\right)\tilde{H}^s_{j+1,B-1} + 4\gamma\tilde{H}^{s,\top}_{j+1,B-1}\tilde{X}^s_{-j}\tilde{X}^{s,\top}_{-j}\tilde{H}^s_{j+1,B-1} \\
&= \tilde{H}^{s,\top}_{j+1,B-1}\left(I + 4\gamma^2\|\tilde{X}^s_{-j}\|^2\tilde{X}^s_{-j}\tilde{X}^{s,\top}_{-j}\right)\tilde{H}^s_{j+1,B-1} \\
&\succeq \tilde{H}^{s,\top}_{j+1,B-1}\tilde{H}^s_{j+1,B-1} \qquad (5.89)
\end{aligned}$$

Using the recursion in Equation (5.89), we show that:

$$\tilde{H}^{s,\top}_{0,B-1}\tilde{H}^s_{0,B-1} + 4\gamma K^s \succeq I\,.$$

This establishes the lower bound. To establish the upper bound, we consider

$$\tilde{H}^{s,\top}_{j,B-1}\tilde{H}^{s}_{j,B-1} + \hat{\gamma}\tilde{H}^{s,\top}_{j+1,B-1}\tilde{X}^{s}_{-j}\tilde{X}^{s,\top}_{-j}\tilde{H}^{s}_{j+1,B-1}\,.$$

Following similar technique used to establish Equation (5.89), using the fact that under the event $\tilde{\mathcal{D}}^{0,t-1}$ we have $\|\tilde{X}^{s}_{-j}\|^2 \leq R$ we show that:

$$\tilde{H}^{s,\top}_{j,B-1}\tilde{H}^{s}_{j,B-1} + \hat{\gamma}\tilde{H}^{s,\top}_{j+1,B-1}\tilde{X}^{s}_{-j}\tilde{X}^{s,\top}_{-j}\tilde{H}^{s}_{j+1,B-1} \preceq \tilde{H}^{s,\top}_{j+1,B-1}\tilde{H}^{s}_{j+1,B-1}\,.$$

Using a similar recursion as before, we establish that:

$$\tilde{H}^{s,\top}_{0,B-1}\tilde{H}^{s}_{0,B-1} + \hat{\gamma}K^{s} \preceq I\,.$$

□

We are now ready to bound the first term in (5.83):

$$\mathbb{E}\left[\sum_{r=1}^{t}\left(\prod_{s=1}^{r-1}\tilde{H}^{t-s,\top}_{0,B-1}\right)K^{t-r}\left(\prod_{s=r-1}^{1}\tilde{H}^{t-s}_{0,B-1}\right)1\left[\tilde{\mathcal{D}}^{0,t-1}\right]\right] \tag{5.90}$$

It is easy to show via. telescoping sum argument that:

$$\sum_{r=1}^{t}\left(\prod_{s=1}^{r-1}\tilde{H}^{t-s,\top}_{0,B-1}\right)\left(I - \tilde{H}^{t-r,\top}_{0,B-1}\tilde{H}^{t-r}_{0,B-1}\right)\left(\prod_{s=r-1}^{1}\tilde{H}^{t-s}_{0,B-1}\right) = I - \left(\prod_{s=1}^{t}\tilde{H}^{t-s,\top}_{0,B-1}\right)\left(\prod_{s=t}^{1}\tilde{H}^{t-s}_{0,B-1}\right) \tag{5.91}$$

We then use Claim 10 to show that under the event $\tilde{\mathcal{D}}^{0,t-1}$, we must have:

$$\frac{I - \left(\prod_{s=1}^{t}\tilde{H}^{t-s,\top}_{0,B-1}\right)\left(\prod_{s=t}^{1}\tilde{H}^{t-s}_{0,B-1}\right)}{4\gamma} \preceq \sum_{r=1}^{t}\left(\prod_{s=1}^{r-1}\tilde{H}^{t-s,\top}_{0,B-1}\right)K^{t-r}\left(\prod_{s=r-1}^{1}\tilde{H}^{t-s}_{0,B-1}\right) \tag{5.92}$$

And:

$$\sum_{r=1}^{t}\left(\prod_{s=1}^{r-1}\tilde{H}^{t-s,\top}_{0,B-1}\right)K^{t-r}\left(\prod_{s=r-1}^{1}\tilde{H}^{t-s}_{0,B-1}\right) \preceq \frac{I - \left(\prod_{s=1}^{t}\tilde{H}^{t-s,\top}_{0,B-1}\right)\left(\prod_{s=t}^{1}\tilde{H}^{t-s}_{0,B-1}\right)}{\hat{\gamma}} \tag{5.93}$$

Finally, combining Lemma 5.9.3, Lemma 5.9.4, claim 10, Equations (5.92), (5.93) and the bound on μ_4 (stated after assumption 3 in section 5.1) along with $\hat{\gamma} = 4\gamma(1-\gamma R)$ we get the statement of the proposition.

□

5.9.7 Proof of proposition 2

Before delving into the proof, we note some useful results below.

Lemma 5.9.5. *For any random matrix $B \in \mathbb{R}^{d\times d}$ we have that*

$$\mathbb{E}\left[B^\top\right]\mathbb{E}\left[B\right] \preceq \mathbb{E}\left[B^\top B\right] \tag{5.94}$$

Hence

$$\|\mathbb{E}\left[B\right]\| \leq \sqrt{\|\mathbb{E}\left[B^\top B\right]\|} \tag{5.95}$$

Proof. Note that for any vector $x \in \mathbb{R}^d$ we have

$$x^\top \mathbb{E}\left[B^\top\right]\mathbb{E}\left[B\right]x = \|\mathbb{E}\left[Bx\right]\|^2 \leq \mathbb{E}\left[\|Bx\|^2\right] = x^\top \mathbb{E}\left[B^\top B\right]x \tag{5.96}$$

□

Lemma 5.9.6. *Let $\gamma RB \leq \frac{c}{6}$ for $0 < c < 1$. The there are constants $c_1, c_2 > 0$ such that for $T^{\alpha/2} > c_1 \frac{\sqrt{M_4}}{\sigma_{\min}(G)}$ we have*

$$\|\mathcal{H}\| \leq \sqrt{1 - c_2\gamma B\sigma_{\min}(G)} \leq 1 - \frac{c_2}{2}\gamma B\sigma_{\min}(G) \tag{5.97}$$

with $1 - c_2\gamma B\sigma_{\min}(G) > 0$.

Proof. Note that $\mathcal{H}$ can be written as $\mathcal{H} = \mathbb{E}\left[\tilde{H}^0_{0,B-1}1[\tilde{\mathcal{D}}^0_{-0}]\right]$. First we use Lemma 5.9.5 to get

$$\|\mathcal{H}\| \leq \sqrt{\left\|\mathbb{E}\left[\tilde{H}^{0,\top}_{0,B-1}\tilde{H}^0_{0,B-1}1[\tilde{\mathcal{D}}^0_{-0}]\right]\right\|} \tag{5.98}$$

Then, from Lemma 5.7.3 we can show that there are constants $c_1, c_2 > 0$ such that

$$\left\|\mathbb{E}\left[\tilde{H}^{0,\top}_{0,B-1}\tilde{H}^0_{0,B-1}1[\tilde{\mathcal{D}}^0_{-0}]\right]\right\| \leq \left(1 - c_1\gamma B\sigma_{\min}(G) + c_2\gamma B\sqrt{M_4}\frac{1}{T^{\alpha/2}}\right) \tag{5.99}$$

Now choosing T such that $T^{\alpha/2} > \frac{c_2\sqrt{M_4}}{2c_1\sigma_{\min}(G)}$ we get

$$\left\|\mathbb{E}\left[\tilde{H}^{0,\top}_{0,B-1}\tilde{H}^0_{0,B-1}1[\tilde{\mathcal{D}}^0_{-0}]\right]\right\| \leq (1 - c_3\gamma B\sigma_{\min}(G)) \tag{5.100}$$

where c_3 is such that the RHS in (5.100) is positive. Hence the claim follows.

□

Proof of Proposition 2. We will prove the proposition only for $a = 0$. The arguments for general a are exactly the same.

For simplicity, we denote

$$\hat{\tilde{A}}_N^v \equiv \left(\hat{\tilde{A}}_{0,N}^v\right) \tag{5.101}$$

From recursion (5.5) we have the following relation between $\left(\tilde{A}_B^{t_2-1,v}\right)$ and $\left(\tilde{A}_B^{t_1-1,v}\right)$ for $t_2 > t_1$

$$\begin{aligned}\left(\tilde{A}_B^{t_2-1,v}\right) \;=\; \left(\tilde{A}_B^{t_1-1,v}\right)\left(\prod_{s=t_2-t_1}^{1}\tilde{H}_{0,B-1}^{t_2-s}\right) + \\ 2\gamma\sum_{r=1}^{t_2-t_1}\sum_{j=0}^{B-1}\eta_{-j}^{t_2-r}\tilde{X}_{-j}^{t_2-r,\top}\tilde{H}_{j+1,B-1}^{t_2-r}\left(\prod_{s=r-1}^{1}\tilde{H}_{0,B-1}^{t_2-s}\right).\end{aligned} \tag{5.102}$$

Hence we have

$$\begin{aligned}\left(\tilde{A}_B^{t_1-1,v}\right)^\top\left(\tilde{A}_B^{t_2-1,v}\right) = \left(\tilde{A}_B^{t_1-1,v}\right)^\top\left(\tilde{A}_B^{t_1-1,v}\right)\left(\prod_{s=t_2-t_1}^{1}\tilde{H}_{0,B-1}^{t_2-s}\right) + \\ 2\gamma\left(\tilde{A}_B^{t_1-1,v}\right)^\top\sum_{r=1}^{t_2-t_1}\sum_{j=0}^{B-1}\eta_{-j}^{t_2-r}\tilde{X}_{-j}^{t_2-r,\top}\tilde{H}_{j+1,B-1}^{t_2-r}\left(\prod_{s=r-1}^{1}\tilde{H}_{0,B-1}^{t_2-s}\right).\end{aligned} \tag{5.103}$$

The second term in (5.103) is bounded in claim 11

The first term in (5.103) can be analyzed using independence as follows.

$$\begin{aligned}&\mathbb{E}\left[\left(\tilde{A}_B^{t_1-1,v}\right)^\top\left(\tilde{A}_B^{t_1-1,v}\right)1\left[\tilde{\mathcal{D}}^{0,t_1-1}\right]\left(\prod_{s=t_2-t_1}^{1}\tilde{H}_{0,B-1}^{t_2-s}\right)1\left[\tilde{\mathcal{D}}^{t_1,N-1}\right]\right]\\
&= \tilde{V}_{t_1-1}\mathbb{E}\left[\left(\prod_{s=t_2-t_1}^{1}\tilde{H}_{0,B-1}^{t_2-s}\right)1\left[\tilde{\mathcal{D}}^{t_1,N-1}\right]\right]\\
&= \tilde{V}_{t_1-1}\mathbb{E}\left[\left(\prod_{s=t_2-t_1}^{1}\tilde{H}_{0,B-1}^{t_2-s}\right)1\left[\tilde{\mathcal{D}}^{t_1,t_2-1}\right]\right]\mathbb{E}\left[1\left[\tilde{\mathcal{D}}^{t_2,N-1}\right]\right]\\
&= \tilde{V}_{t_1-1}\left(\prod_{s=t_2-t_1}^{1}\mathbb{E}\left[\tilde{H}_{0,B-1}^{t_2-s}1\left[\tilde{\mathcal{D}}^{t_1,t_2-1}\right]\right]\right)\mathbb{E}\left[1\left[\tilde{\mathcal{D}}^{t_2,N-1}\right]\right] = \tilde{V}_{t_1-1}\mathcal{H}^{t_2-t_1}\mathbb{E}\left[1\left[\tilde{\mathcal{D}}^{t_2,N-1}\right]\right]\\
&= \tilde{V}_{t_1-1}\mathcal{H}^{t_2-t_1} - \tilde{V}_{t_1-1}\mathcal{H}^{t_2-t_1}\mathbb{E}\left[1\left[\tilde{\mathcal{D}}^{t_2,N-1,C}\right]\right].\end{aligned} \tag{5.104}$$

Note that,

$$\begin{aligned}\left(\tilde{A}_B^{t_1-1,v}\right)^\top\left(\tilde{A}_B^{t_1-1,v}\right) \preceq 4\gamma^2(Bt_1)\sum_{r=1}^{t_1}\sum_{j=0}^{B-1}\left\|\eta_{-j}^{t_1-r}\right\|^2\cdot\\
\left(\prod_{s=1}^{r-1}\tilde{H}_{0,B-1}^{t_1-s,\top}\right)\tilde{H}_{j+1,B-1}^{t_1-r,\top}\tilde{X}_{-j}^{t_1-r}\tilde{X}_{-j}^{t_1-r,\top}\tilde{H}_{j+1,B-1}^{t_1-r}\left(\prod_{s=r-1}^{1}\tilde{H}_{0,B-1}^{t_1-s}\right).\end{aligned} \tag{5.105}$$

From equation (5.105), we have:

$$\left\|\tilde{V}_{t_1-1}\right\| \le c\gamma^2(Bt_1)^2Rd\sigma_{\max}, \tag{5.106}$$

and further, $\|\mathcal{H}\| < 1$ from Lemma 5.9.6. Hence,

$$\left\|\tilde{V}_{t_1-1}\mathcal{H}^{t_2-t_1}\mathbb{E}\left[1\left[\tilde{\mathcal{D}}^{t_2,N-1,C}\right]\right]\right\| \leq \left\|\tilde{V}_{t_1-1}\mathcal{H}^{t_2-t_1}\right\|\frac{1}{T^\alpha} \leq c\gamma^2(Bt_1)^2Rd\sigma_{\max}\frac{1}{T^\alpha}.$$

For brevity, given a matrix $Q \in \mathbb{R}^{d\times d}$, let,

$$\text{Sym}\left(Q\right) = Q + Q^\top. \tag{5.107}$$

Combining everything so far, we have, for $t_2 > t_1$:

$$\begin{aligned}&\text{Sym}\left(\mathbb{E}\left[\left(\tilde{A}_B^{t_1-1,v}\right)^\top\left(\tilde{A}_B^{t_2-1,v}\right)1\left[\tilde{\mathcal{D}}^{0,N-1}\right]\right]\right)\\ \preceq\ &\text{Sym}\left(\tilde{V}_{t_1-1}\mathcal{H}^{t_2-t_1}\right) + c_1\gamma^2(Bt_1)^2Rd\sigma_{\max}\frac{1}{T^\alpha}I +\\ &\left(c_3\gamma^2B^2t_1t_2Rd\sigma_{\max}\frac{1}{T^{\alpha/2}}\right)I\end{aligned} \tag{5.108}$$

Since $Bt_2 \leq T$ we get:

$$\begin{aligned}\text{Sym}\left(\mathbb{E}\left[\left(\tilde{A}_B^{t_1-1,v}\right)^\top\left(\tilde{A}_B^{t_2-1,v}\right)1\left[\tilde{\mathcal{D}}^{0,N-1}\right]\right]\right) \preceq \text{Sym}\left(\tilde{V}_{t_1-1}\mathcal{H}^{t_2-t_1}\right) +\\ c_3\gamma^2T^2Rd\sigma_{\max}\frac{1}{T^{\alpha/2}}I.\end{aligned} \tag{5.109}$$

Therefore we have,

$$\begin{aligned}\frac{1}{N^2}\sum_{t_1\neq t_2}\mathbb{E}\left[\left(\tilde{A}_B^{t_1-1,v}\right)^\top\left(\tilde{A}_B^{t_2-1,v}\right)\right] \preceq \frac{1}{N^2}\sum_{t_1=1}^{N-1}\text{Sym}\left(\tilde{V}_{t_1-1}\left(\sum_{t_2>t_1}\mathcal{H}^{t_2-t_1}\right)\right)\\ +c_3\gamma^2T^2Rd\sigma_{\max}\frac{1}{T^{\alpha/2}}I.\end{aligned}$$

Next observe that,

$$\begin{aligned}&\frac{1}{N^2}\sum_{t=1}^{N}\tilde{V}_{t-1} + \frac{1}{N^2}\sum_{t_1=1}^{N-1}\text{Sym}\left(\tilde{V}_{t_1-1}\left(\sum_{t_2>t_1}\mathcal{H}^{t_2-t_1}\right)\right)\\ =\ &\frac{1}{N^2}\sum_{t=1}^{N}\tilde{V}_{t-1} + \frac{1}{N^2}\sum_{t_1=1}^{N-1}\text{Sym}\left(\tilde{V}_{t_1-1}\left(\sum_{s=1}^{N-t_1}\mathcal{H}^{s}\right)\right)\\ \preceq\ &\frac{1}{N^2}\sum_{t=1}^{N}\text{Sym}\left(\tilde{V}_{t-1}\left(\sum_{s=0}^{N-t}\mathcal{H}^{s}\right)\right).\end{aligned}$$

Hence, substituting in (5.66), we obtain:

$$\mathbb{E}\left[\left(\hat{A}_N^v\right)^\top\left(\hat{A}_N^v\right)1\left[\tilde{\mathcal{D}}^{0,N-1}\right]\right] \preceq \frac{1}{N^2}\sum_{t=1}^{N}\text{Sym}\left(\tilde{V}_{t-1}\left(\sum_{s=0}^{N-t}\mathcal{H}^{s}\right)\right) + \tag{5.110}$$

$$c_3\gamma^2T^2Rd\sigma_{\max}\frac{1}{T^{\alpha/2}}I. \tag{5.111}$$

From Equations (5.110)-(5.111) we obtain (5.67).

Now $\sum_{s=0}^{N-t}\mathcal{H}^s = (I-\mathcal{H})^{-1}(I-\mathcal{H}^{N-t+1})$ since from Lemma 5.9.6 we know that $\|\mathcal{H}\|<1$ for large T. Thus we get (5.68).

□

5.9.8 Claims

Claim 11. *For $\gamma \le \frac{1}{2R}$ we have*

$$\begin{aligned}&\left\|\mathbb{E}\left[2\gamma\left(\tilde{A}_B^{t_1-1,v}\right)^\top \sum_{r=1}^{t_2-t_1}\sum_{j=0}^{B-1}\eta_{-j}^{t_2-r}\tilde{X}_{-j}^{t_2-r,\top}\tilde{H}_{j+1,B-1}^{t_2-r}\left(\prod_{s=r-1}^{1}\tilde{H}_{0,B-1}^{t_2-s}\right)\right]1\left[\tilde{\mathcal{D}}^{0,N-1}\right]\right\| \\ &\le c_1\gamma^2B^2t_1t_2Rd\sigma_{\max}\frac{1}{T^{\alpha/2}} \end{aligned} \tag{5.112}$$

for some constant $c_1>0$.

Proof. The proof is similar to the proof of Lemma 5.9.3.

□

5.9.9 Proof of theorem 5.9.1

Proof of theorem 5.9.1. We start with the following

$$\begin{aligned}\left(\tilde{A}_b^{t-1,b}-A^*\right)^\top\left(\tilde{A}_b^{t-1,b}-A^*\right) &= \left(\prod_{s=1}^{t}\tilde{H}_{0,B-1}^{t-s,\top}\right)(A_0-A^*)^\top(A_0-A)\left(\prod_{s=t}^{1}\tilde{H}_{0,B-1}^{t-s}\right) \\ &\preceq \|A_0-A^*\|^2\left(\prod_{s=1}^{t}\tilde{H}_{0,B-1}^{t-s,\top}\right)\left(\prod_{s=t}^{1}\tilde{H}_{0,B-1}^{t-s}\right)\end{aligned} \tag{5.113}$$

From Lemma 5.7.3 we can show that there are constants $c_1,c_2>0$ such that

$$\begin{aligned}&\left\|\mathbb{E}\left[\left(\prod_{s=1}^{t}\tilde{H}_{0,B-1}^{t-s,\top}\right)\left(\prod_{s=t}^{1}\tilde{H}_{0,B-1}^{t-s}\right)1\left[\tilde{\mathcal{D}}^{0,t-1}\right]\right]\right\| \\ &\le\left(1-c_1\gamma B\sigma_{\min}(G)+c_2\gamma B\sqrt{M_4}\frac{1}{T^{\alpha/2}}\right)^t.\end{aligned} \tag{5.114}$$

Now choosing T such that $T^{\alpha/2}>\frac{c_2\sqrt{M_4}}{2c_1\sigma_{\min}(G)}$ we get,

$$\left\|\mathbb{E}\left[\left(\prod_{s=1}^{t}\hat{H}_{0,B-1}^{t-s,\top}\right)\left(\prod_{s=t}^{1}\hat{H}_{0,B-1}^{t-s}\right)\right]\right\| \le (1-c_3\gamma B\sigma_{\min}(G))^t. \tag{5.115}$$

Thus we get the theorem.

□

5.9.10 Proof of theorem 5.9.2

Proof of theorem 5.9.2. We use the following inequality that is obtained from Lemma 5.9.5

$$\left(\hat{\tilde{A}}^b_{a,N} - A^*\right)^\top \left(\hat{\tilde{A}}^b_{a,N} - A^*\right) \preceq \frac{1}{N-a} \sum_{t=a+1}^{N} \left(\tilde{A}^{t-1,b}_B - A^*\right)^\top \left(\tilde{A}^{t-1,b}_B - A^*\right) \tag{5.116}$$

Therefore

$$\begin{aligned}
&\mathbb{E}\left[\left(\hat{\tilde{A}}^b_{a,N} - A^*\right)^\top \left(\hat{\tilde{A}}^b_{a,N} - A^*\right) 1\left[\tilde{\mathcal{D}}^{0,N-1}\right]\right] \\
&\preceq \frac{1}{N-a} \sum_{t=a+1}^{N} \mathbb{E}\left[\left(\tilde{A}^{t-1,b}_B - A^*\right)^\top \left(\tilde{A}^{t-1,b}_B - A^*\right) 1\left[\tilde{\mathcal{D}}^{0,N-1}\right]\right] \\
&\preceq \frac{1}{N-a} \sum_{t=a+1}^{N} \mathbb{E}\left[\left(\tilde{A}^{t-1,b}_B - A^*\right)^\top \left(\tilde{A}^{t-1,b}_B - A^*\right) 1\left[\tilde{\mathcal{D}}^{0,t-1}\right]\right]
\end{aligned} \tag{5.117}$$

Now using theorem 5.9.1, we get

$$\begin{aligned}
&\mathbb{E}\left[\left(\hat{\tilde{A}}^b_{a,N} - A^*\right)^\top \left(\hat{\tilde{A}}^b_{a,N} - A^*\right) 1\left[\tilde{\mathcal{D}}^{0,N-1}\right]\right] \preceq \\
&\left(\frac{1}{N-a} \frac{(1 - c_1\gamma B \sigma_{\min}(G))^{a+1}}{c_1 \gamma B \sigma_{\min}(G)}\right) \|A_0 - A^*\|^2 I
\end{aligned} \tag{5.118}$$

Hence using $1 - x \leq e^{-x}$ we get

$$\begin{aligned}
&\left\|\mathbb{E}\left[\left(\hat{\tilde{A}}^b_{a,N} - A^*\right)^\top \left(\hat{\tilde{A}}^b_{a,N} - A^*\right) 1\left[\tilde{\mathcal{D}}^{0,N-1}\right]\right]\right\| \\
&\leq c \frac{1}{B(N-a)} \frac{e^{-cB\gamma\sigma_{\min}(G)a}}{\gamma \sigma_{\min}(G)} \|A_0 - A^*\|^2
\end{aligned} \tag{5.119}$$

□

5.10 Prediction error: Proof of theorem 5.3.2

Recall the definition of the prediction error at stationarity.

$$\mathcal{L}_{\mathsf{pred}}(\hat{A}; A^*, \mu) := \mathbb{E}_{X_t \sim \pi} \|X_{t+1} - \hat{A} X_t\|^2 \tag{5.120}$$

where π is the stationary distribution.

Note that the prediction loss is a function of possibly random estimator $\hat{A}$. Hence the expectation in (5.120) is only with respect to the process (X_t) (which is considered independent of $\hat{A}$). Letting

$G = \mathbb{E}\left[X_t X_t^\top\right]$ as the covariance matrix of the process at stationarity, we can write

$$\mathcal{L}_{\mathsf{pred}}(\hat{A}; A^*, \mu) = \mathrm{Tr}(G(\hat{A} - A^*)^\top(\hat{A} - A^*)) + \mathrm{Tr}(\Sigma) \tag{5.121}$$

We are interested in bounding the expected prediction loss of the estimator which is the average iterate $\hat{A}_{a,N}$ of our algorithm $\mathsf{SGD-RER}$ (with $a = \theta N$). Note that $\hat{A}_{a,N} = \hat{A}^b_{a,N} + \hat{A}^v_{a,N}$ where the superscripts b and v correspond to bias and variance respectively (c.f. (5.27))

Hence

$$\begin{aligned}
\mathbb{E}\left[\mathcal{L}_{\mathsf{pred}}(\hat{A}_{a,N}; A^*, \mu)\right] &= \mathrm{Tr}(\Sigma) + \mathrm{Tr}\left(G^{1/2}\mathbb{E}\left[\left(\hat{A}_{a,N} - A^*\right)^\top\left(\hat{A}_{a,N} - A^*\right)\right]G^{1/2}\right) \\
&\leq \mathrm{Tr}(\Sigma) + 2\,\mathrm{Tr}\left(G^{1/2}\mathbb{E}\left[\left(\hat{A}^v_{a,N}\right)^\top\left(\hat{A}^v_{a,N}\right)\right]G^{1/2}\right) \\
&\quad + 2\,\mathrm{Tr}\left(G^{1/2}\mathbb{E}\left[\left(\hat{A}^b_{a,N} - A^*\right)^\top\left(\hat{A}^b_{a,N} - A^*\right)\right]G^{1/2}\right)
\end{aligned} \tag{5.122}$$

But we will only bound $\mathbb{E}\left[\mathcal{L}_{\mathsf{pred}}(\hat{A}_{a,N}; A^*, \mu)1\left[\mathcal{D}^{0,N-1}\right]\right]$ so that we have a tight upper bound on the conditional expectation of $\mathcal{L}_{\mathsf{pred}}$ over a high probability event.

As before we will just focus on the prediction error obtained using the algorithmic iterates from the coupled process, i.e., we will bound $\mathbb{E}\left[\mathcal{L}_{\mathsf{pred}}(\hat{\tilde{A}}_{a,N}; A^*, \mu)1\left[\tilde{\mathcal{D}}^{0,N-1}\right]\right]$

5.10.1 Variance of prediction error

In this section we will focus on analyzing the variance part of the expected prediction loss under the coupled process

$$\tilde{\mathcal{L}}^v = \mathrm{Tr}\left(G^{1/2}\mathbb{E}\left[\left(\hat{\tilde{A}}^v_{a,N}\right)^\top\left(\hat{\tilde{A}}^v_{a,N}\right)1\left[\tilde{\mathcal{D}}^{0,N-1}\right]\right]G^{1/2}\right) \tag{5.123}$$

where $T = N(B+u)$.

We begin with few lemmata which would be useful in bounding $\tilde{\mathcal{L}}^v$. Recall the definition of $\mathcal{H}$

$$\mathcal{H} = \mathbb{E}\left[\prod_{j=0}^{B-1}\left(I - 2\gamma\tilde{X}^0_{-j}\tilde{X}^{0,\top}_{-j}\right)1[\tilde{\mathcal{D}}^0_{-0}]\right] \tag{5.124}$$

with $\tilde{X}_0$ sampled from the stationary distribution π.

Lemma 5.10.1. *Let $\gamma \leq \frac{1}{8RB}$. Then*

$$\mathcal{H} + \mathcal{H}^\top \preceq 2\left(I - \frac{4}{3}\gamma BG\right) + \frac{8}{3}\gamma B\sqrt{M_4}\frac{1}{T^{\alpha/2}}I \tag{5.125}$$

where $M_4 = \mathbb{E}\left[\|\tilde{X}^0_{-0}\|^4\right]$. *For simplicity, we just say that for* $\gamma RB < \frac{c}{4}$ *with* $0 < c < 1$ *then*

$$\mathcal{H} + \mathcal{H}^\top \preceq 2\left(I - c_1\gamma BG\right) + c_2\gamma B\sqrt{M_4}\frac{1}{T^{\alpha/2}}I \tag{5.126}$$

for some absolute constants $c_1, c_2 > 0$.

The proof is similar to the combined proofs of Lemmas 6.9.3 and 5.7.3. We therefore skip it.

Next we will bound $\mathrm{Tr}(G(I-\mathcal{H})^{-1})$.

Lemma 5.10.2. *Let* $\gamma RB < \frac{c_1}{4}$ *with* $0 < c_1 < 1$. *Then for* T *such that* $T^{\alpha/2} > c_2\frac{\sqrt{M_4}}{\sigma_{\min}(G)}$ *we have*

$$\mathrm{Tr}\left(G(I-\mathcal{H})^{-1}\right) \leq c\frac{d}{\gamma B} \tag{5.127}$$

for some absolute constant $c > 0$.

Proof. First note that

$$\begin{aligned}
\mathrm{Tr}\left(G(I-\mathcal{H})^{-1}\right)) &= \mathrm{Tr}\left(G^{1/2}(I-\mathcal{H})^{-1}G^{1/2}\right)) \\
&= \mathrm{Tr}\left(\left(G^{-1} - G^{-1/2}\mathcal{H}G^{-1/2}\right)^{-1}\right) \\
&\leq d\left\|\left(G^{-1} - G^{-1/2}\mathcal{H}G^{-1/2}\right)^{-1}\right\| \\
&= \frac{d}{\sigma_{\min}\left(G^{-1} - G^{-1/2}\mathcal{H}G^{-1/2}\right)}
\end{aligned} \tag{5.128}$$

Let $Q = \left(G^{-1} - G^{-1/2}\mathcal{H}G^{-1/2}\right)$. Let $\mathrm{Sym}\left(Q\right) = Q + Q^\top$. We will relate $\sigma_{\min}(Q)$ with $\sigma_{\min}\left(\frac{\mathrm{Sym}(Q)}{2}\right)$. From AM-GM inequality, for any $\theta > 0$, we have

$$\frac{Q^\top Q}{\theta} + \theta I \succeq \mathrm{Sym}\left(Q\right) \tag{5.129}$$

Also

$$\sigma^2_{\min}(Q) = \inf_{x:\|x\|=1} x^\top Q^\top Q x \tag{5.130}$$

Further, from lemma 5.10.1 we have

$$\begin{aligned}
\mathrm{Sym}\left(Q\right) &= G^{-1} - G^{-1/2}\frac{\mathcal{H} + \mathcal{H}^T}{2}G^{-1/2} \\
&\succeq c_1\gamma BI - c_2\gamma B\sqrt{M_4}\frac{1}{T^{\alpha/2}}G^{-1} \\
&\succeq c_1\gamma BI - c_2\gamma B\sqrt{M_4}\frac{1}{T^{\alpha/2}}\frac{1}{\sigma_{\min}(G)}I
\end{aligned} \tag{5.131}$$

Hence combining equations (5.129), (5.130) and (5.131) we have:

$$\frac{\sigma_{\min}^2(Q)}{\theta} + \theta \succeq c_1\gamma B - c_2\gamma B\sqrt{M_4}\frac{1}{T^{\alpha/2}}\frac{1}{\sigma_{\min}(G)}. \tag{5.132}$$

Now choosing $\theta = \frac{1}{2}c_1\gamma B$ we get:

$$\sigma_{\min}^2(Q) \geq \frac{c_1^2}{4}\gamma^2B^2 - \frac{c_2c_1}{2}\gamma^2B^2\sqrt{M_4}\frac{1}{T^{\alpha/2}}\frac{1}{\sigma_{\min}(G)}. \tag{5.133}$$

Now choose T large enough such that $\frac{c_2c_1}{2}\sqrt{M_4}\frac{1}{T^{\alpha/2}}\frac{1}{\sigma_{\min}(G)} \leq \frac{c_1^2}{8}$. Then, $\sigma_{\min}^2(Q) \geq c_3\gamma^2B^2$, for some constant $c_3 > 0$. Hence from (5.128),

$$\operatorname{Tr}\left(G(I-\mathcal{H})^{-1}\right) \leq c_4\frac{d}{\gamma B}.$$

□

Next we bound $\operatorname{Tr}(\Delta(I-\mathcal{H})^{-1}G)$ for any symmetric matrix Δ. Let $\kappa(G) = \frac{\sigma_{\max(G)}}{\sigma_{\min}(G)}$ denote the condition number of G.

Lemma 5.10.3. *Let $\gamma RB \leq \frac{c_1}{4}$ with $0 < c_1 < 1$. Then for T such that $T^{\alpha/2} > c_2\frac{\sqrt{M_4}}{\sigma_{\min}(G)}$ we have*

$$\left|\operatorname{Tr}\left(\Delta(I-\mathcal{H})^{-1}G\right)\right| \leq c\frac{d}{\gamma B}\left\|\Delta\right\|\sqrt{\kappa(G)} \tag{5.134}$$

for some absolute constant $c > 0$.

Proof. We have

$$\begin{aligned}\left|\operatorname{Tr}\left(\Delta(I-\mathcal{H})^{-1}G\right)\right| &= \left|\operatorname{Tr}\left(G^{1/2}\Delta G^{-1/2}G^{1/2}(I-\mathcal{H})^{-1}G^{1/2}\right)\right| \\ &\leq d\left\|G^{1/2}\Delta G^{-1/2}\right\|\left\|G^{1/2}(I-\mathcal{H})^{-1}G^{1/2}\right\| \\ &\leq d\sqrt{\kappa(G)}\left\|\Delta\right\|\left\|G^{1/2}(I-\mathcal{H})^{-1}G^{1/2}\right\| \end{aligned} \tag{5.135}$$

From the proof of lemma 5.10.2, we know that

$$\left\|G^{1/2}(I-\mathcal{H})^{-1}G^{1/2}\right\| \leq c\frac{1}{\gamma B} \tag{5.136}$$

for T satisfying the condition the statement of the lemma.

Hence:

$$\left|\operatorname{Tr}\left(\Delta(I-\mathcal{H})^{-1}G\right)\right| \leq c\sqrt{\kappa(G)}\left\|\Delta\right\|\frac{d}{\gamma B} \tag{5.137}$$

□

Our goal is to bound $\operatorname{Tr}(\tilde{V}_{t-1}(I-\mathcal{H})^{-1}G)$. From proposition 1 we can decompose $\tilde{V}_{t-1}$ as:

$$\tilde{V}_{t-1} = \gamma \operatorname{Tr}(\Sigma) I + (\tilde{V}_{t-1} - \gamma \operatorname{Tr}(\Sigma) I), \tag{5.138}$$

and hence,

$$\operatorname{Tr}(\tilde{V}_{t-1}(I-\mathcal{H})^{-1}G) = \gamma \operatorname{Tr}(\Sigma) \operatorname{Tr}((I-\mathcal{H})^{-1}G) + \operatorname{Tr}\left((\tilde{V}_{t-1} - \gamma \operatorname{Tr}(\Sigma))(I-\mathcal{H})^{-1}G\right). \tag{5.139}$$

To bound the second term in (5.139) we want to use lemma 5.10.3. Hence we need to bound the norm of $\tilde{V}_{t-1} - \gamma \operatorname{Tr}(\Sigma)$.

Lemma 5.10.4. *Let $\gamma \leq \min\left\{\frac{c}{4RB}, \frac{1}{2R}\right\}$ for $0 < c < 1$. Then there are constants $c_1, c_2, c_3 > 0$ such that for $T^{\alpha/2} > c_1 \frac{\sqrt{M_4}}{\sigma_{\min}(G)}$ we have*

$$\left\|\tilde{V}_{t-1} - \gamma \operatorname{Tr}(\Sigma)\right\| \leq c_2 \gamma d \sigma_{\max} \left[\frac{1}{B} + (1 - c_3 \gamma B \sigma_{\min}(G))^t\right] \tag{5.140}$$

for some constant $c_1 > 0$.

Proof. From proposition 1 we have

$$\begin{aligned}\left\|\tilde{V}_{t-1} - \gamma \operatorname{Tr}(\Sigma) I\right\| \leq{}& \gamma \operatorname{Tr}(\Sigma) \frac{\gamma R}{1-\gamma R} + \\ & c_1 \gamma \operatorname{Tr}(\Sigma) \left\| \mathbb{E}\left[\left(\prod_{s=1}^{t} \tilde{H}_{0,B-1}^{t-s,\top}\right)\left(\prod_{s=t}^{1} \tilde{H}_{0,B-1}^{t-s}\right) 1\left[\tilde{\mathcal{D}}^{0,t-1}\right]\right]\right\| \\ & + c_2 \gamma d \sigma_{\max}(\Sigma) T^2 \frac{1}{T^{\alpha/2}}.\end{aligned} \tag{5.141}$$

From lemma 5.9.6 equation (5.100) we can show that

$$\left\| \mathbb{E}\left[\left(\prod_{s=1}^{t} \tilde{H}_{0,B-1}^{t-s,\top}\right)\left(\prod_{s=t}^{1} \tilde{H}_{0,B-1}^{t-s}\right) 1\left[\tilde{\mathcal{D}}^{0,t-1}\right]\right]\right\| \leq (1 - c_3 \gamma B \sigma_{\min}(G))^t. \tag{5.142}$$

Hence

$$\begin{aligned}\left\|\tilde{V}_{t-1} - \gamma \operatorname{Tr}(\Sigma) I\right\| &\leq c_4 \gamma d \sigma_{\max}(\Sigma) \left[\frac{\gamma R}{1-\gamma R} + (1 - c_3 \gamma B \sigma_{\min}(G))^t\right] \\ \leq c_5 \gamma d \sigma_{\max} \left[\gamma R + (1 - c_3 \gamma B \sigma_{\min}(G))^t\right] &\leq c_6 \gamma d \sigma_{\max} \left[\frac{1}{B} + (1 - c_3 \gamma B \sigma_{\min}(G))^t\right].\end{aligned} \tag{5.143}$$

□

Now we have all required ingredients for the main theorem of this section

Theorem 5.10.5. *Let $\gamma \leq \min\left\{\frac{c}{4RB}, \frac{1}{2R}\right\}$ for $0 < c < 1$. Then there are constants $c_1, c_2, c_3, c_4 > 0$ such that for $T^{\alpha/2} > c_1 \frac{\sqrt{M_4}}{\sigma_{\min}(G)}$ the variance part of the expected prediction loss $\tilde{\mathcal{L}}^v$ (defined in (5.123))*

for $a = \theta N$ is bounded as

$$\tilde{\mathcal{L}}^v \leq c_1 \frac{d \operatorname{Tr}(\Sigma)}{NB(1-\theta)} + c_2 \frac{d^2 \sigma_{\max}(\Sigma)}{NB(1-\theta)} \frac{\sqrt{\kappa(G)}}{B} + c_3 \frac{d^2 \sigma_{\max}(\Sigma)}{(NB)^2(1-\theta)^2} \sqrt{\kappa(G)} \frac{1}{\gamma \sigma_{\min}(G)} + c_4 \gamma^2 R d \sigma_{\max}(\Sigma) T^2 \frac{1}{T^{\alpha/2}} \operatorname{Tr}(G) \tag{5.144}$$

Proof. From (5.123) and proposition 2 equation (5.68) we have

$$\tilde{\mathcal{L}}^v \leq \frac{2}{(N-a)^2} \sum_{t=a+1}^{N} \operatorname{Tr}\left(\tilde{V}_{t-1}(I-\mathcal{H})^{-1} G\right) \tag{5.145}$$

$$+ \frac{2}{(N-a)^2} \sum_{t=a+1}^{N} \operatorname{Tr}\left(\tilde{V}_{t-1}(I-\mathcal{H})^{-1} \mathcal{H}^{N-t+1} G\right) \tag{5.146}$$

$$+ c\delta \operatorname{Tr}(G) \tag{5.147}$$

where $\delta = \gamma^2 T^2 R d \sigma_{\max}(\Sigma) \frac{1}{T^{\alpha/2}}$ as defined in (5.69)

For the first term (5.145) we have from (5.139), lemma 5.10.2, lemma 5.10.3 and lemma 5.10.4

$$\begin{aligned} \operatorname{Tr}\left(\tilde{V}_{t-1}(I-\mathcal{H})^{-1} G\right) &\leq c_1 \gamma \operatorname{Tr}(\Sigma) \frac{d}{\gamma B} + \\ &\quad c_2 \frac{d}{\gamma B} \sqrt{\kappa(G)} \gamma d \sigma_{\max}(\Sigma) \left[\frac{1}{B} + (1 - c_3 \gamma B \sigma_{\min}(G))^t\right] \\ &= c_1 \frac{d \operatorname{Tr}(\Sigma)}{B} + c_2 \frac{d^2 \sigma_{\max}(\Sigma)}{B} \frac{\sqrt{\kappa(G)}}{B} + \\ &\quad c_4 \frac{d^2 \sigma_{\max}(\Sigma)}{B} \sqrt{\kappa(G)} \left(1 - c_3 \gamma B \sigma_{\min}(G)\right)^t \end{aligned} \tag{5.148}$$

Therefore

$$\begin{aligned} \frac{2}{(N-a)^2} \sum_{t=a+1}^{N} \operatorname{Tr}\left(\tilde{V}_{t-1}(I-\mathcal{H})^{-1} G\right) &\leq c_1 \frac{d \operatorname{Tr}(\Sigma)}{NB(1-\theta)} + c_2 \frac{d^2 \sigma_{\max}(\Sigma)}{NB(1-\theta)} \frac{\sqrt{\kappa(G)}}{B} + \\ &\quad c_5 \frac{d^2 \sigma_{\max}(\Sigma)}{N^2 B (1-\theta)^2} \sqrt{\kappa(G)} \frac{\left(1 - c_3 \gamma B \sigma_{\min}(G)\right)^{a+1}}{\gamma B \sigma_{\min}(G)} \end{aligned} \tag{5.149}$$

Similarly, for the second term (5.146), from corollary 1, lemma 5.10.3, lemma 5.9.6 and the fact that $(I-\mathcal{H})^{-1}$ and $\mathcal{H}^{N-t+1}$ commute, we get

$$\begin{aligned} &\left|\operatorname{Tr}\left(\tilde{V}_{t-1}(I-\mathcal{H})^{-1} \mathcal{H}^{N-t+1} G\right)\right| \leq c_1 \frac{d}{\gamma B} \sqrt{\kappa} \|\tilde{V}_{t-1}\| \|\mathcal{H}^{N-t+1}\| \\ &\leq c_2 \frac{d}{\gamma B} \sqrt{\kappa(G)} \gamma d \sigma_{\max}(\Sigma) \left(1 - c_3 \gamma B \sigma_{\min}(G)\right)^{(N-t+1)} \\ &= c_2 \frac{d^2 \sigma_{\max}(\Sigma)}{B} \sqrt{\kappa(G)} \left(1 - c_3 \gamma B \sigma_{\min}(G)\right)^{(N-t+1)} \end{aligned} \tag{5.150}$$

Therefore

$$\left|\frac{2}{(N-a)^2}\sum_{t=a+1}^{N}\operatorname{Tr}\left(\tilde{V}_{t-1}(I-\mathcal{H})^{-1}\mathcal{H}^{N-t+1}G\right)\right| \leq c\frac{d^2\sigma_{\max}(\Sigma)}{N^2B(1-\theta)^2}\sqrt{\kappa(G)}\frac{1}{\gamma B\sigma_{\min}(G)} \tag{5.151}$$

Hence we obtain,

$$\begin{aligned}\tilde{\mathcal{L}}^v \leq\ & c_1\frac{d\operatorname{Tr}(\Sigma)}{NB(1-\theta)} + c_2\frac{d^2\sigma_{\max}(\Sigma)}{NB(1-\theta)}\frac{\sqrt{\kappa(G)}}{B} + \\ & c_3\frac{d^2\sigma_{\max}(\Sigma)}{N^2B^2(1-\theta)^2}\sqrt{\kappa(G)}\frac{1}{\gamma\sigma_{\min}(G)} + c_4\gamma^2Rd\sigma_{\max}(\Sigma)T^2\frac{1}{T^{\alpha/2}}\operatorname{Tr}(G).\end{aligned} \tag{5.152}$$

□

5.10.2 Bias of prediction error

In this section we will focus on analyzing the (tail-averaged) bias part of the expected prediction loss from the coupled process

$$\tilde{\mathcal{L}}^b = \operatorname{Tr}\left(G^{1/2}\mathbb{E}\left[\left(\left(\hat{\tilde{A}}^b_{a,N} - A^*\right)\right)^\top\left(\left(\hat{\tilde{A}}^b_{a,N} - A^*\right)\right)1\left[\tilde{\mathcal{D}}^{0,N-1}\right]\right]G^{1/2}\right) \tag{5.153}$$

where $T = N(B+u)$ and $a = \theta N$ for $0 < \theta < 1$.

Theorem 5.10.6. *Let $\gamma RB \leq \frac{c}{6}$ for some $0 < c < 1$ and B such that $\gamma R \leq \frac{1}{2}$. There exist constants $c_1, c_2, c_3, c_4 > 0$ such that if T satisfies $T^{\alpha/2} > c_1\frac{\sqrt{M_4}}{\sigma_{\min}(G)}$ then for $a = \theta N$ with $0 < \theta < 1$ we have*

$$\tilde{\mathcal{L}}^b \leq c_2\frac{1}{NB(1-\theta)}\frac{\operatorname{Tr}(G)}{\gamma\sigma_{\min}(G)}e^{-c_3NB\gamma\sigma_{\min}(G)\theta}\left\|A_0 - A^*\right\|^2 \tag{5.154}$$

Proof. Proof follows directly from (5.153) and theorem 5.9.2. □

5.10.3 Overall prediction error

Combining theorem 5.10.5 and theorem 5.10.6 along with lemma 5.5.6 we obtain the main theorem on prediction error of $\mathsf{SGD} - \mathsf{RER}$

Theorem 5.10.7. *Let R, B, u, α be chosen as in section 5.3. Let $\gamma = \frac{c}{4RB} \leq \frac{1}{2R}$ for $0 < c < 1$. Then there are constants $c_1, c_2, c_3, c_4 > 0$ such that for $T^{\alpha/2} > c_1\frac{\sqrt{M_4}}{\sigma_{\min}(G)}$ the expected prediction loss*

$\mathcal{L}$ (defined in (5.121)) is bounded as

$$\begin{aligned}\mathbb{E}\left[\mathcal{L}_{\mathsf{pred}}(\hat{A}_{a,N};A^*,\mu)1\left[\mathcal{D}^{0,N-1}\right]\right] \leq c_2\left[\frac{d\operatorname{Tr}(\Sigma)}{B(N-a)}+\frac{d^2\sigma_{\max}(\Sigma)}{B(N-a)}\frac{\sqrt{\kappa(G)}}{B}\right]+\\ c_3\left[\frac{d^2\sigma_{\max}(\Sigma)}{B^2(N-a)^2}\sqrt{\kappa(G)}\frac{1}{\gamma\sigma_{\min}(G)}+\right.\\ \frac{1}{B(N-a)}d\kappa(G)RBe^{-c_4\frac{\sigma_{\min}(G)}{R}a}\left\|A_0-A^*\right\|^2+\\ \left.\left(\frac{T^3}{B^3}\left\|A^{*u}\right\|+\frac{d\sigma_{\max}(\Sigma)}{R}\frac{T^2}{B^2}\frac{1}{T^{\alpha/2}}\right)\operatorname{Tr}(G)\right]\end{aligned} \tag{5.155}$$

Hence, if $\|A^*\|<c_0<1$ then choosing $a\geq C\frac{R\log T}{\sigma_{\min}(G)}$ such that $B(N-a)=\Theta(T)$ and B,u as in section 5.3 we get

$$\mathbb{E}\left[\mathcal{L}_{\mathsf{pred}}(\hat{A}_{a,N};A^*,\mu)1\left[\mathcal{D}^{0,N-1}\right]\right] \leq c_2\frac{d\operatorname{Tr}(\Sigma)}{T}+o\left(\frac{1}{T}\right) \tag{5.156}$$

5.11 Prediction error for sparse systems

In this section we consider the $\mathsf{VAR}(A^*,\mu)$ model with sparse A^* whose sparsity pattern is known. We will present a modification of $\mathsf{SGD}-\mathsf{RER}$ that takes into account the sparsity pattern information. Formally, let $S_l=\{k:A^*_{l,k}\neq 0\}$ be support or sparsity pattern of row l of A^*. Further let $s_l=|S_l|$ denote the sparsity of row j. We assume that S_l is known for each $1\leq l\leq d$. The claim is that the excess expected prediction loss is of order $\frac{\sum_l s_l\sigma_l^2}{T}$. We will present only a sketch of the proof highlighting the main steps. Detailed calculations follow similarly as in sections 5.9.1 and 5.10.

The modification of the $\mathsf{SGD}-\mathsf{RER}$ algorithm to use the sparsity pattern is as follows. Let $a_l^{*,\top}$ denote row l of A^*. The algorithmic iterates are given by (A_j^{t-1}) where row l is $a_{j,l}^{t-1,\top}$. Let $a_{0,l}^0=0\in\mathbb{R}^d$. Let $\{e_l:1\leq l\leq d\}$ denote the standard basis of $\mathbb{R}^d$. Let $P_{S_l}:\mathbb{R}^d\to\mathbb{R}^d$ denote the (self adjoint) orthogonal projection operator onto the subspace spanned by $\{e_l:l\in S_l\}$. Then update for row l is given by

$$a_{j+1,l}^{t-1,\top}=\left[a_{j,l}^{t-1,\top}-2\gamma(a_{j,l}^{t-1,\top}X_{-j}^{t-1}-\langle e_l,X_{-(j-1)}^{t-1}\rangle)X_{-j}^{t-1,\top}\right]P_{S_l} \tag{5.157}$$

and $a_{0,l}^t=a_{B,l}^{t-1}$. Since each iterate above has sparsity pattern S_l by construction, we can rewrite the above as

$$a_{j+1,l}^{t-1,\top}=a_{j,l}^{t-1,\top}-2\gamma(a_{j,l}^{t-1,\top}X_{-j}^{t-1}-\langle e_l,X_{-(j-1)}^{t-1}\rangle)\left(P_{S_l}X_{-j}^{t-1}\right)^\top \tag{5.158}$$

Notice that $a_{j,l}^{t-1,\top} X_{-j}^{t-1} = a_{j,l}^{t-1,\top} P_{S_l} X_{-j}^{t-1}$ and

$$\langle e_l, X_{-(j-1)}^{t-1} \rangle = a_l^{*,\top} X_{-j}^{t-1} + \eta_{-j,l}^{t-1}$$

Thus

$$\left(a_{j+1,l}^{t-1} - a_l^*\right)^\top = \left(a_{j,l}^{t-1} - a_l^*\right)^\top \left(P_{S_l} - 2\gamma \left(P_{S_l} X_{-j}^{t-1}\right)\left(P_{S_l} X_{-j}^{t-1}\right)^\top\right) + 2\gamma \eta_{-j,l}^{t-1} \left(P_{S_l} X_{-j}^{t-1}\right)^\top \tag{5.159}$$

For a vector $v \in \mathbb{R}^d$, let $v_{S_l} \in \mathbb{R}^{s_l}$ be the vector corresponding to the support S_l i.e. entries in v_{S_l} correspond to the entries in v whose indices are in S_l. So we can rewrite (5.159) completely in $\mathbb{R}^{s_l}$ as

$$\left(a_{j+1,l}^{t-1} - a_l^*\right)_{S_l}^\top = \left(a_{j,l}^{t-1} - a_l^*\right)_{S_l}^\top \left(I_{s_l} - 2\gamma \left(X_{-j}^{t-1}\right)_{S_l} \left(X_{-j}^{t-1}\right)_{S_l}^\top\right) + 2\gamma \eta_{-j,l}^{t-1} \left(X_{-j}^{t-1}\right)_{S_l}^\top \tag{5.160}$$

where I_{s_l} is the identity matrix of dimension s_l.

Our goal is to bound the expected prediction error for this modified $\mathsf{SGD-RER}$. To that end, we will make some important observations.

(1) Since we focus on prediction error, the entire analysis can be carried out row by row. To see this, if $\hat{A}$ is any estimator, the

$$\mathcal{L}_{\mathsf{pred}}(\hat{A}; A^*, \mu) - \mathrm{Tr}(\Sigma) = \mathrm{Tr}(G(\hat{A} - A^*)^\top (\hat{A} - A)) = \sum_{l=1}^{d} \mathrm{Tr}(G(\hat{a}_l - a_l^*)(\hat{a}_l - a_l^*)^\top)$$

where $\hat{a}_l^\top$ is the row l of $\hat{A}$.

(2) If $\hat{a}_l$ and a_l^* have sparsity pattern S_l then

$$\begin{aligned} \mathrm{Tr}(G(\hat{a}_l - a_l^*)(\hat{a}_l - a_l^*)^\top) &= \mathrm{Tr}(P_{S_l} G P_{S_l} (\hat{a}_l - a_l^*)(\hat{a}_l - a_l^*)^\top) \\ &= \mathrm{Tr}(G_{S_l} (\hat{a}_l - a_l^*)_{S_l} (\hat{a}_l - a_l^*)_{S_l}^\top) \end{aligned}$$

where $G_{S_l} \in \mathbb{R}^{s_l \times s_l}$ is the submatrix of G obtained by picking rows and columns corresponding to indices in S_l.

(3) Under the stationary measure, we have $\mathbb{E}\left[\left(P_{S_l} X_{-j}^{t-1}\right)\left(P_{S_l} X_{-j}^{t-1}\right)^\top\right] = P_{S_l} G P_{S_l}$. Thus, with high probability $\left\|P_{S_l} X_{-j}^{t-1}\right\|^2 \le c s_l \sigma_{\max}(G) \log T$.

(4) Letting $s_0 = \max_l s_l$, we can set $R = c s_0 \sigma_{\max}(G) \log T$ and use step size $\gamma = O(1/RB)$.

(5) We can perform the same bias-variance decomposition as described in section 5.6 to obtain $a_{B,l}^{t-1,v}$ and $a_{B,l}^{t-1,b}$.

(6) From previous observations, the variance of last iterate corresponding to row l turns out to be

$$\gamma\sigma_l^2(1-o(1))I_{s_l} \preceq \mathbb{E}\left[\left(a_{B,l}^{t-1,v}\right)_{S_l}\left(a_{B,l}^{t-1,v}\right)_{S_l}^\top\right] \preceq \frac{\gamma}{1-\gamma R}\sigma_l^2(1+o(1))I_{s_l}$$

where $\sigma_l^2 = \Sigma_{l,l}$.

(7) Similarly, the variance of the average iterate $\mathbb{E}\left[(\hat{a}_{0,N,l}^v)(\hat{a}_{0,N,l}^v)^\top\right]$ corresponding to row l can be bounded upto leading order by

$$\frac{1}{N^2}\sum_{t=1}^{N}\left[V_{t-1,l}(I_{s_l}-\mathcal{H}_{S_l})^{-1} + (I_{s_l}-\mathcal{H}_{S_l}^\top)^{-1}V_{t-1,l}\right]$$

where $V_{t-1,l} = \mathbb{E}\left[\left(a_{B,l}^{t-1,v}\right)_{S_l}\left(a_{B,l}^{t-1,v}\right)_{S_l}^\top\right]$ and (with abuse of notation) $\mathcal{H}_{S_l}$ is defined as

$$\mathcal{H}_{S_l} = \mathbb{E}\left[\prod_{j=0}^{B-1}\left(I_{s_l} - 2\gamma(\tilde{X}_{-j}^0)_{S_l}(\tilde{X}_{-j}^0)_{S_l}^\top\right)1\left[\cap_{j=0}^{B-1}\left\{\left\|(\tilde{X}_{-j}^0)_{S_l}\right\|^2 \le R\right\}\right]\right]$$

where $\tilde{X}_0^0 \sim \pi$.

(8) Now, similar to lemma 5.10.1 we can bound $\mathcal{H}_{S_l} + \mathcal{H}_{S_l}^\top$ by $2(I_{s_l} - c\gamma BG_{s_l})$ upto leading order.

(9) Thus similar to lemma 5.10.2 we obtain

$$\mathrm{Tr}(G_{S_l}(I-\mathcal{H}_{S_l})^{-1}) \le c\frac{s_l}{\gamma B}$$

(10) Finally as in section 5.10.1 we can bound the variance of prediction error of row l upto leading order by

$$\mathrm{Tr}(G\mathbb{E}\left[(\hat{a}_{0,N,l}^v)(\hat{a}_{0,N,l}^v)^\top\right]) \lesssim \frac{\sigma_l^2 s_l}{T}$$

Thus summing over l we get

$$\mathrm{Tr}\left(G\mathbb{E}\left[(\hat{A}_{0,N}^v)(\hat{A}_{0,N}^v)^\top\right]\right) \lesssim \frac{\sum_l \sigma_l^2 s_l}{T}$$

(11) Bias can also be analyzed in a similar way and it will be of strictly lower order (using suitable tail-averaging).

(12) Thus the excess prediction loss is given bounded as

$$\mathbb{E}\left[\mathcal{L}_{\mathsf{pred}}(\hat{A}_{N/2,N}; A^*, \mu)\right] - \mathrm{Tr}(\Sigma) \lesssim \frac{\sum_l \sigma_l^2 s_l}{T}$$

So the modified $\mathsf{SGD}-\mathsf{RER}$ algorithm effectively utilizes the low dimensional structure in A^*.

Chapter 6

Generalized linear system identification

6.1 Problem statement

Let $\phi : \mathbb{R} \to \mathbb{R}$ be an increasing, 1-Lipschitz function such that $\phi(0) = 0$. Suppose $X_0 \in \mathbb{R}^d$ is a random variable and $A^* \in \mathbb{R}^{d \times d}$. We consider the following non-linear dynamical system (NLDS):

$$X_{t+1} = \phi(A^* X_t) + \eta_t, \tag{6.1}$$

where the noise sequence $\eta_0, \ldots, \eta_T$ is i.i.d random vectors independent of X_0. The noise η_t is such that $\mathbb{E}\eta_t = 0$, $\mathbb{E}\eta_t \eta_t^\top = \sigma^2 I$ for some $\sigma > 0$. We will also assume that $M_4 := \mathbb{E}\|\eta_t\|^4 < \infty$. Let μ be the law of noise η. We denote the model above as $\mathsf{NLDS}(A^*, \mu, \phi)$. Whenever a stationary distribution exists for the process, we will denote it by $\pi(A^*, \mu, \phi)$ or just π when the process is clear from context. We will call the trajectory $X_0, X_1, \ldots, X_T$ 'stationary' if X_0 is distributed according to the measure $\pi(A^*, \mu, \phi)$. Unless specified otherwise, we take $X_0 = 0$ almost surely.

The goal is to estimate A^* given a single trajectory $X_0, X_1, \ldots, X_T$. A natural approach would be to minimize the empirical square loss, i.e, $\mathcal{L}_{\mathsf{sq}}(A; X) := \frac{1}{T}\sum_{t=0}^{T-1} \|\phi(AX_t) - X_{t+1}\|^2$. However, when the link function ϕ is not linear, then this would be non-convex and hard to optimize. Instead, we use a convex proxy loss given by:

$$\mathcal{L}_{\mathsf{prox}}(A; X) = \frac{1}{T}\sum_{t=0}^{T-1}\sum_{i=1}^{d} \bar{\phi}(\langle a_i, X_t\rangle) - \langle e_i, X_{t+1}\rangle\langle a_i, X_t\rangle\,, \tag{6.2}$$

where $\bar{\phi}$ is the indefinite integral of the link function ϕ and a_i is the i-th row of A. Note that the

gradient of $\mathcal{L}_{\mathsf{prox}}(A; X)$ with respect to A is given by:

$$\nabla \mathcal{L}_{\mathsf{prox}}(A; X) = \frac{1}{T} \sum_{t=0}^{T-1} \left(\phi(AX_t) - X_{t+1}\right) X_t^\top \,. \tag{6.3}$$

When the model is clear from context and the stationary distribution exists, we will denote the second moment matrix under the stationary distribution by $G := \mathbb{E}[X_t X_t^\top]$. Note that $G \succeq \mathbb{E}[\eta_t \eta_t^\top] = \sigma^2 I$. Also, the empirical second moment matrix is denoted by $\hat{G} := \frac{1}{T} \sum_{t=0}^{T-1} X_t X_t^\top$.

6.1.1 Assumptions

We now state the assumptions below and use only a subset of the assumptions for each result.

Assumption 4 (Lipschitzness and Uniform Expansivity)**.** *ϕ is 1-Lipschitz and $|\phi(x) - \phi(y)| \geq \zeta|x - y|$, for some $\zeta > 0$.*

Note that when ϕ is only weakly differentiable but satisfy Assumption 4, with a slight abuse of notation, we will write down $\phi(x) - \phi(y) = \phi'(\beta)(x - y)$ for some $\phi'(\beta) \in [\zeta, 1]$.

Assumption 5 (Bounded 2nd Derivative)**.** *ϕ is twice continuously differentiable and $|\phi''|$ is bounded.*

Assumption 6 (Noise Sub-Gaussianity)**.** *For any unit norm vector $x \in \mathbb{R}^d$, we have $\langle \eta_t, x \rangle$ to be sub-Gaussian with variance proxy $C_\eta \sigma^2$.*

Next, we extend the definition of exponential stability in [72] to 'exponential regularity' to allow unstable systems.

Assumption 7 (Exponential Regularity)**.** *Let $X_T = h_{T-1}(X_0, \eta_0, \ldots, \eta_T)$ be the function representation of X_T. We say that $\mathsf{NLDS}(A^*, \mu, \phi)$ is (C_ρ, ρ) exponentially regular if for any choice of $T \in \mathbb{N}$ and $X_0, X_0', \eta_0, \ldots, \eta_T \in \mathbb{R}^d$:*

$$\|h_T(X_0, \eta_0, \ldots, \eta_T) - h_T(X_0', \eta_0, \ldots, \eta_T)\|_2 \leq C_\rho \rho^{T-1} \|X_0 - X_0'\|_2 \,.$$

When $\rho < 1$, we will call the system stable. When $\rho = 1$ we will call it 'possibly marginally stable' and when $\rho > 1$, we will call it 'possibly unstable'

Note that when Assumption 7 holds with $\rho < 1$, the system necessarily mixes and converges to a stationary distribution as $T \to \infty$. Such systems forget their initial conditions in time scales of the order $\tau_{\mathsf{mix}} = O(\frac{1+\log C_\rho}{\log \frac{1}{\rho}}) = O\left(\frac{1+\log C_\rho}{1-\rho}\right)$, and hence we use this as a proxy for the mixing time. In what follows, when we say 'the system does not mix' we either mean that it does not mix within time T or it does not converge to a stationary distribution (ex: $\rho \geq 1$).

Assumption 8 (Norm Boundedness)**.** $\|A^*\|_{\mathsf{op}} = \rho < 1$

Algorithm 2: Quasi Newton Method

Input : Offline data $\{X_0, \ldots, X_T\}$, horizon T, no. of iterations m, link function ϕ, step size γ

Output: Estimate A_m

begin
  $A_0^0 = 0$ /*Initialization*/
  $\hat{G} \leftarrow \frac{1}{T}\sum_{t=0}^{T-1} X_t X_t^\top$; If $\hat{G}$ is not invertible, then **return** $A_m = 0$
  for $i \leftarrow 0$ **to** $m-1$ **do**
    $A_{i+1} \leftarrow A_i - 2\gamma \left(\nabla \mathcal{L}_{\mathsf{prox}}(A_i; X)\right) \hat{G}^{-1}$
  end
end

That is, if A^* satisfies Assumption 8, we have for arbitrary $X, X' \in \mathbb{R}^d$: $\|\phi(A^*X) - \phi(A^*X')\| \leq \rho \|X - X'\|$ and $\left\|(\phi \circ A^*)^k(X)\right\| \leq \rho^k \|X\|$. Hence, for such A^*, NLDS is *necessarily stable*.

6.2 Offline learning with Quasi Newton method

In this section we consider estimating A^* using a single trajectory $(X_1, \ldots, X_T)$ from $\mathsf{NLDS}(A^*, \mu, \phi)$. To this end, we study an offline Quasi Newton Method (Algorithm 2) where the iterates descend in the directions of the gradient of $\mathcal{L}_{\mathsf{prox}}$ normalized by the inverse of the empirical second moment matrix $\hat{G} := \frac{1}{T}\sum_{t=0}^{T-1} X_t X_t^\top$. That is, the iterates follow an approximation of the standard Newton update. Next we present the main result from analysis of algorithm 2.

Theorem 6.2.1 (Learning Without Mixing)**.** *Suppose Assumptions 4 6 and 7 hold with expansivity factor ζ and regularity parameters (C_ρ, ρ). Let $\bar{C}, \bar{C}_3$ be constants depending only on C_η, and let $\delta \in (0, \frac{1}{2})$. Let $R^* := C_\rho^2 C_\eta d\sigma^2 \left(\sum_{t=1}^{T-1} \rho^t\right)^2 \log(\frac{4Td}{\delta})$, and the number samples $T \geq \bar{C}_3 \left(d \log\left(\frac{R^*}{\sigma^2}\right) + \log\frac{1}{\delta}\right)$. Set the step size $\gamma = \frac{1}{4}$ and $m \geq \frac{10}{\zeta} \cdot \log\left(\frac{\|A_0 - A^*\|_{\mathsf{F}}^2 \cdot TR^*}{\sigma^2 d^2}\right)$. Then, the output A_m of Algorithm 2 after m iterations satisfies (w.p. $\geq 1 - \delta$):*

$$\|A_m - A^*\|_{\mathsf{F}}^2 \leq \frac{\bar{C}\sigma^2}{T\zeta^2 \lambda_{\min}(\hat{G})} \left[d^2 \log\left(1 + \frac{R^*}{\sigma^2}\right) + d \log\left(\frac{2d}{\delta}\right)\right], \text{ where } \lambda_{\min}\left(\hat{G}\right) \geq \frac{\sigma^2}{2}.$$

Note that as $\lambda_{\min}(\hat{G}) \geq \sigma^2$, the error rate scales as $\approx d^2/T$, independent of $\tau_{\mathsf{mix}} \approx 1/(1-\rho)$. The theorem also holds for non-mixing or possibly unstable systems as long as $\rho < 1 + \frac{C}{T}$. Furthermore, the error bound above is similar to the *minimax optimal bound* by [51] for the *linear* setting, i.e., when $\phi(x) = x$. Note that as the link function ϕ tends to decrease the information in x, hence intuitively lower bound for linear setting should apply for NLDS as well, which would imply our error rate to be optimal; we leave further investigation into lower bound of NLDS identification for future work. Interestingly, in the linear case whenever the smallest singular value $\sigma_{\min}(A^*) > 1 + \epsilon$, it can be show than $\lambda_{\min}(\hat{G})$ grows exponentially with T, leading to an exponentially small error. It is not clear how to arrive at such a growth lower bound in the non-linear case.

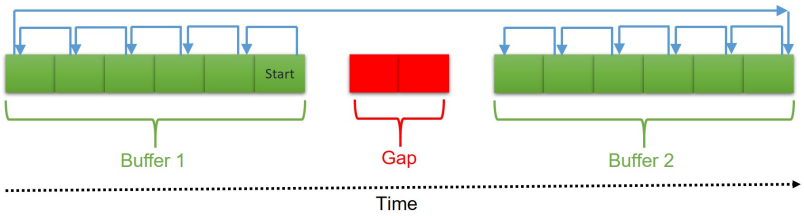

Figure 6-1: Data order in $\mathsf{SGD}-\mathsf{RER}$, where each block represents a data point. Blue arrows indicate the data processing order. The gaps ensure approximate independence between successive buffers.

The computational complexity of the algorithm scales as $m \cdot T$ which depends only logarithmically on τ_{mix}. Interestingly, the algorithm is almost hyperparameter free, and does not require knoweldge of parameters $\sigma, \tau_{\mathsf{mix}}, \zeta$. Also note that the stationary points of Algorithm 2 and GLMtron ([73]) are the same. So, the stronger error rate in the result above compared to the result by [73] is due to a sharper analysis. However, in dynamical systems of the form 1.9, the squared norm of the iterates grow as $\frac{d}{1-\rho}$ even in the stable case. Hence, the GLMtron algorithm requires step sizes to be $\approx \frac{1-\rho}{d}$ which implies significantly slower convergence rate for large $\tau_{\mathsf{mix}} = 1/(1-\rho)$. In contrast, convergence rate for Algorithm 2 depends at most logarithmically on τ_{mix}.

See Section 6.5 for a high-level exposition of the key ideas in the analysis and Section 6.8.1 for a detailed overview of the proof.

6.3 Streaming learning with SGD-RER

In this section, we consider the one-pass, streaming setting, where the data points are presented in a streaming fashion. The goal is to continuously produce better estimates of A^* while also ensuring that the space and the time complexity of the algorithm is small. This disallows approaches that would just store all the observed points and then apply offline Algorithm 2 to produce strong estimation error. Such one-pass streaming algorithms are critical in a variety of settings like large-scale and online time-series analysis [89, 179], TD learning in RL [180], econometrics.

To address this problem, we consider $\mathsf{SGD}-\mathsf{RER}$ (Algorithm 3) which was introduced in [110] in the context of *linear system identification* (LSI). We apply the method for NLDS identification as well. $\mathsf{SGD}-\mathsf{RER}$ uses SGD like updates, but the data is processed in a different order than it is received from the dynamical system. This algorithm is based on the observation made in [110] that for LSI, when SGD is run on the least squares loss in the forward order, there are spurious correlations which prevent the algorithm's convergence to the optimum parameter A^*. Surprisingly, considering the data in the reverse order *exactly* unravels these correlations to resolve the problem. Reverse order traversal of data, even though one pass, does not give a streaming algorithm. Hence, we divide the data into multiple buffers of size B and leave of size u between the buffers (See Figure 6-

Algorithm 3: SGD − RER

Input : Streaming data $\{X_\tau\}$, horizon T, buffer size B, buffer gap u, bound R, tail start: $t_0 \leq N/2$, link function ϕ, step size γ

Output: Estimate $\hat{A}_{t_0,t}$, for all $t_0 \leq t \leq N-1$; $N = T/(B+u)$

begin
  Total buffer size: $S \leftarrow B+u$, Number of buffers: $N \leftarrow T/S$
  $A_0^0 = 0$ /*Initialization*/
  for $t \leftarrow 1$ **to** N **do**
    Form buffer $\mathsf{Buf}^{t-1} = \{X_0^{t-1}, \ldots, X_{S-1}^{t-1}\}$, where, $X_i^{t-1} \leftarrow X_{(t-1)\cdot S+i}$
    If $\exists i,\ s.t.,\ \left\|X_i^{t-1}\right\|^2 > R$, then **return** $\hat{A}_{t_0,t} = 0$
    for $i \leftarrow 0$ **to** $B-1$ **do**
      $A_{i+1}^{t-1} \leftarrow A_i^{t-1} - 2\gamma\left[\phi(A_i^{t-1}X_{S-i-1}^{t-1}) - X_{S-i}^{t-1}\right] X_{S-i-1}^{t-1,\top}$
    end
    $A_0^t = A_B^{t-1}$
    If $t \geq t_0+1$, then $\hat{A}_{t_0,t} \leftarrow \frac{1}{t-t_0}\sum_{\tau=t_0+1}^{t} A_B^{\tau-1}$
  end
end

1). The data *within* each buffer is processed in the reverse order whereas the buffers themselves are processed in the order received. See Figure 6-1 for an illustration of the processing order. The gaps u are set large enough so that the buffers behave approximately independently. Setting $B \geq 10u$ we note that this simple strategy improves the sample efficiency compared to naive data dropping since we use *most* of the samples for estimating A^*. We now present the main result for streaming setting.

Theorem 6.3.1 (Streaming Algorithm)**.** *Suppose Assumptions 4, 5, 6 and 8 hold and that the data points are stationary. Set* $\alpha = 100$, $R = \frac{16(\alpha+2)dC_\eta\sigma^2\log T}{1-\rho}$, $u \geq \frac{2\alpha\log T}{\log(\frac{1}{\rho})}$ *and*

$$B \geq \max\left(\bar{C}_1 \frac{d}{(1-\rho)(1-\rho^2)}\log\left(\frac{d}{1-\rho}\right), 10u\right)$$

for a global constant $\bar{C}_1$ *dependent only* C_η *and* α*. Let* $N = T/(B+u)$ *be the number of buffers. Finally, set step-size* $\gamma = \frac{C}{T^\nu}$ *where* $\nu = 6.5/7$ *and let* T *be large enough such that* $\gamma \leq \min\left(\frac{\zeta}{4BR(1+\zeta)}, \frac{1}{2R}\right)$*. If* $N/2 > t_0 > c_1\frac{\log T}{\zeta\gamma B\lambda_{\min}(G)} = \Theta(T^\nu \log T)$ *for some large enough constant* $c_1 > 0$*, then output* $\hat{A}_{t_0,N}$ *of Algorithm 3 satisfies:*

$$\mathbb{E}\left[\|\hat{A}_{t_0,N} - A^*\|_{\mathsf{F}}^2\right] \leq C\frac{d^2\sigma^2\log T}{T\lambda_{\min}(G)\,\zeta^2} + \textit{ Lower Order Terms} \tag{6.4}$$

where C *is a constant dependent on* C_η*,* α*.*

Remark 13. *The lower order terms are of the order* $\mathrm{Poly}(R, B, \beta, 1/\zeta, 1/\lambda_{\min}, \|\phi''\|)\gamma^{7/2}T^2 + \gamma^2 R\sigma^2 d\frac{1}{T^{\alpha/2-2}} + \|A_0 - A^*\|_{\mathsf{F}}^2\left[\frac{e^{-c_2\zeta\gamma B\lambda_{\min}t_0}}{T\zeta\gamma\lambda_{\min}}\right]$*. We refer to the proof in Section 6.8.13 for details.*

Remark 14. *Although the bound in theorem 6.3.1 is given for the algorithmic iterate at the end of*

the horizon, the proof shows that in fact we can bound the error of the iterates at the end of each buffer after $(1+c)t_0$ *i.e. if* $t \geq (1+c)t_0$ *for some* $c > 0$ *then we obtain*

$$\mathbb{E}\left[\|\hat{A}_{t_0,t} - A^*\|_{\mathsf{F}}^2\right] \leq C\frac{d^2\sigma^2 \log T}{(tB)\lambda_{\min}(G)\,\zeta^2} + \textit{ Lower Order Terms}$$

Note that the estimation error above matches the error by offline method up to log factors (see [105, Theorem 2]). Furthermore, while the method requires NLDS to be mixing, i.e., $\rho < 1$, but the leading term in error rate does not have an explicit dependence on it. Moreover, the space complexity of the method is only $B \cdot d$ which scales as $d^2/((1-\rho)(1-\rho^2))$, i.e., it is $1/((1-\rho)(1-\rho^2)) \sim \tau_{\mathsf{mix}}^2$ factor worse than the obvious lower bound of $\mathrm{O}(d^2)$ to store A. We leave further investigation into space complexity optimization or tightening the lower bound for future work. Also, note that $u \leq B/10$, so $\mathsf{SGD}-\mathsf{RER}$ wastes only about 10% of the samples. Finally, the algorithm requires a reasonable upper bound on ρ to set up various hyperparameters like R, u, B. However, it is not clear how to estimate such an upper bound only using the data, and seems like an interesting open question.

See Section 6.5 for an explanation of the elements involved in the analysis of the algorithm and to Section 6.8.1 for a detailed overview of the proof.

6.4 Exponential lower bounds for non-Expansive link functions

The previous results showed that we can efficiently recover the matrix A^* given that the link function is uniformly expansive. We state a result from [105] that shows that parameter recovery is hard when the link function is non expansive. In particular, even for the case of $\phi = \mathsf{ReLU}$, the noise being $\mathcal{N}(0, I)$, and $\|A^*\| \leq \frac{1}{2}$, the error has an information theoretic lower bound which is exponential in the dimension. We note that this is consistent with Theorem 3 in [73] which too has an exponential dependence on the dimension (since the matrix $K \succeq I$).

Before stating the results, we introduce some notation. Consider any algorithm $\mathcal{A}$, with accepts input $(X_0, \ldots, X_T)$ and outputs an estimate $\hat{A} \in \mathbb{R}^{d\times d}$. For simplicity of calculation, we will assume that $X_0 = 0$ and $X_{t+1} = \mathsf{ReLU}(A^*X_t) + \eta_t$. Since the mixing time is $O(1)$, similar results should hold for stationary sequences. We define the loss $\mathcal{L}(\mathcal{A}, T, A^*) = \mathbb{E}\|\hat{A} - A^*\|_F^2$, where the expectation is over the randomness in the data and the algorithm. By $\Theta(\frac{1}{2})$, we denote all the the elements of $B \in \mathbb{R}^{d\times d}$ such that $\|B\| \leq \frac{1}{2}$. The minimax loss is defined as:

$$\mathcal{L}(\Theta(\tfrac{1}{2}), T) := \inf_{\mathcal{A}} \sup_{A^* \in \Theta(\frac{1}{2})} \mathcal{L}(\mathcal{A}, T, A^*)\,.$$

We have the following result from [105, Theorem 4]

Theorem 6.4.1 (ReLU Lower Bound [105]). *For universal constants $c_0, c_1 > 0$, we have:*

$$\mathcal{L}(\Theta(\tfrac{1}{2}), T) \geq c_0 \min\left(1, \frac{\exp(c_1 d)}{T}\right).$$

6.5 Proof sketch

Modified Newton Method. Let a_i^* be the i-th row of A^* and $a_i(l)$ be the i-th row of A_l both in column vector form. The proof of theorem 6.2.1 follows once we consider the Lyapunov function $\Delta_{l,i} = \|\hat{G}^{1/2}(a_i(l) - a_i^*)\|$ and show that

$$\Delta_{l+1,i} \leq (1 - 2\gamma\zeta)\Delta_l + \gamma\|\hat{G}^{-1/2}\hat{N}_i\| \tag{6.5}$$

Where $\hat{N}_i := \frac{1}{T}\sum_{t=0}^{T-1}\langle e_i, \eta_t\rangle X_t$. In the case of Theorem 6.2.1, we use the sub-Gaussianity of the noise sequence and a martingale argument to obtain a high probability upper bound on $\sum_{i=1}^{d}\|\hat{G}^{-1/2}\hat{N}_i\|^2$ (see Lemma 6.7.1).

SGD-RER. Due to the observations made in Section 6.3, we can split the analysis into the following parts, which are explained in detail below.

1. Analyze the reverse order SGD *within* the buffers.

2. Treat successive buffer as independent samples.

3. Give a bias-variance decomposition similar to the case of linear regression.

4. Use algorithmic stability to control 'spurious' coupling introduced by non-linearity in the bias-variance decomposition.

Coupled process. We deal with the dependence *between* buffers using a fictitious coupled process, constructed just for the sake of analysis (see Definition 4). Leveraging the gap u, this process $(\tilde{X}_\tau)$ is constructed such that $\tilde{X}_\tau \approx X_\tau$ with high probability and the 'coupled buffers' containing data $\tilde{X}$ instead of X are *exactly* independent. Since $\tilde{X}_\tau \approx X_\tau$, the output of $\mathsf{SGD-RER}$ run with the fictitious coupled process should be close output of $\mathsf{SGD-RER}$ run with the actual data points. We then use the strategy outlined above to analyze $\mathsf{SGD-RER}$ with the coupled process. In the analysis given for $\mathsf{SGD-RER}$, all the quantities with $\tilde{}$ involve the coupled process $\tilde{X}$ instead of the real process X.

Non-linear bias variance decomposition. We use the mean value theorem to linearize the non-linear problem. This works effectively when the step size γ is a vanishing function of the horizon T. Observe that the a single SGD/ $\mathsf{SGD}-\mathsf{RER}$ step for a single row can be written as:

$$\begin{aligned} a_i' - a_i^* &= a_i - a_i^* - 2\gamma(\phi(\langle a_i, X_\tau\rangle) - \phi(\langle a_i^*, X_\tau\rangle))X_\tau + 2\gamma\langle \eta_\tau, e_i\rangle X_\tau \\ &= \left(I - 2\gamma\phi'(\beta_\tau)X_\tau X_\tau^\top\right)(a_i - a_i^*) + 2\gamma\langle \eta_\tau, e_i\rangle X_\tau \end{aligned} \tag{6.6}$$

In the second step, we have used the mean value theorem. Equation (6.6) can be interpreted as follows: the matrix $\left(I - 2\gamma\phi'(\beta_\tau)X_\tau X_\tau^\top\right)$ 'contracts' the distance between a_i and a_i^* whereas the noise $2\gamma\langle \eta_\tau, e_i\rangle X_\tau$ is due to the inherent uncertainty. This gives us a bias-variance decomposition similar to the case of SGD with linear regression. We refer to Section 6.8.5 for details on unrolling the recursion in Equation (6.6) to obtain the exact bias-variance decomposition.

Algorithmic stability: Unfortunately, non-linearities result in a 'coupling' between the contraction matrices through the iterates via the first derivative $\phi'(\beta_\tau)$ due to reverse order traversal. This is an important issue since unrolling the recursion in (6.6), we encounter terms such as $\langle \eta_\tau, e_i\rangle(I - 2\gamma\phi'(\beta_{\tau-1})X_{\tau-1}X_{\tau-1}^\top)X_\tau$, which have zero mean in the linear case. However, in the non-linear case, $\beta_{\tau-1}$ depends on η_τ due to reverse order traversal. We show that such dependencies are 'weak' using the idea of algorithmic stability ([181, 182]). In particular, we establish that the output of the algorithm is not affected too much if we re-sample the *entire* data trajectory by independently re-sampling a single noise co-ordinate (η_τ becomes η_τ' and $\beta_{\tau-1}$ becomes $\beta_{\tau-1}'$) when the step size γ is small enough (in other words, the output is stable under small perturbations). Via second derivative arguments, we show that $\beta_{\tau-1} \approx \beta_{\tau-1}'$.

Now observe that resampling noise η_τ does not affect the past value of data i.e, $X_\tau, X_{\tau-1}$ and is independent of $\beta_{\tau-1}'$ by construction. Therefore

$$0 = \mathbb{E}\langle \eta_\tau, e_i\rangle \left(I - 2\gamma\phi'(\beta_{\tau-1}')X_{\tau-1}X_{\tau-1}^\top\right)X_\tau \approx \mathbb{E}\langle \eta_\tau, e_i\rangle \left(I - 2\gamma\phi'(\beta_{\tau-1})X_{\tau-1}X_{\tau-1}^\top\right)X_\tau$$

Such a resampling procedure is also explored in [183] for the analysis of SGD with random reshuffling.

We put together all the ingredients above in order to prove the error bounds given in Theorem 6.3.1

6.6 Preliminaries for the proofs

6.6.1 Concentration under stationary measure

In this section, we will consider the process $\mathsf{NLDS}(A^*, \mu, \phi)$ and the concentration of measure under its stationary distribution. In what follows, we will use the fact that ϕ is 1-Lipschitz as in the

definition of NLDS, even when we don't explicitly use Assumption 4.

Proposition 3. *Under Assumption 7 with $\rho < 1$, the process is exponentially Ergodic and has a stationary distribution π. Suppose $X \sim \pi$ then $\mathbb{E}\|X\|^4 < \infty$.*

The result follows from a technique similar to the one used for Proposition [184] by considering the process in the space of measures endowed with the Wasserstein metric.

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 6.6.1. *Suppose Y is a ν^2 sub-Gaussian random variable with zero mean. Then, for any $\lambda \leq \frac{1}{4\nu^2}$, we have:*

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

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.

$$\mathbb{P}(Z \geq \gamma) = \begin{cases} 1 \text{ if } \gamma \leq 1 \\ \mathbb{P}\left(|Y| \geq \sqrt{\frac{\log(\gamma)}{\lambda}}\right) \text{ if } \gamma > 1 \end{cases} \tag{6.7}$$

Now,

$$\begin{aligned} \mathbb{E}Z &= \int_0^\infty \mathbb{P}(Z \geq \gamma) d\gamma \\ &= \int_0^1 d\gamma + \int_1^\infty \mathbb{P}\left(|Y| \geq \sqrt{\tfrac{\log(\gamma)}{\lambda}}\right) d\gamma \\ &\leq 1 + \int_1^\infty 2\exp\left(-\tfrac{\log(\gamma)}{2\nu^2\lambda}\right) d\gamma \\ &= 1 + 2\int_1^\infty \gamma^{-\frac{1}{2\nu^2\lambda}} d\gamma \\ &= 1 + \frac{4\nu^2\lambda}{1 - 2\nu^2\lambda} \\ &\leq 1 + 8\lambda\nu^2 \end{aligned} \tag{6.8}$$

□

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 6.6.2. *Suppose that Assumptions 6 and 8 hold and ρ be as given in Assumption 8. For any λ such that $0 \leq \lambda \leq \frac{1-\rho}{2\rho C_\eta \sigma^2}$, 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 (1.9), we conclude:

$$\begin{aligned}\|X_{t+1}\|^2 - \|\eta_t\|^2 &= \|\phi(A^*X_t)\|^2 + 2\langle \eta_t, \phi(A^*X_t)\rangle \\ &\leq \rho^2\|X_t\|^2 + 2\langle \eta_t, \phi(A^*X_t)\rangle \end{aligned} \tag{6.9}$$

Let $\mathcal{F}_s = \sigma(X_0, \eta_0, \ldots, \eta_s)$. It is clear that $X_s \in \mathcal{F}_{s-1}$. Using Equation (6.9), we conclude:

$$\begin{aligned}\mathbb{E}\left[\exp(\lambda Z_{t+1})\middle|\mathcal{F}_{t-1}\right] &= \mathbb{E}\left[\exp\left(\lambda\|X_{t+1}\|^2 - \lambda\|\eta_t\|^2\right)\middle|\mathcal{F}_{t-1}\right]\exp\left(-\lambda\sum_{s=0}^{t-1}\rho^{t-s}\|\eta_s\|^2\right) \\ &\leq \mathbb{E}\left[\exp\left(\lambda\rho^2\|X_t\|^2 + 2\lambda\langle\eta_t, \phi(A^*X_t)\rangle\right)\middle|\mathcal{F}_{t-1}\right]\exp\left(-\lambda\sum_{s=0}^{t-1}\rho^{t-s}\|\eta_s\|^2\right) \\ &\leq \exp\left(\lambda\rho^2\|X_t\|^2 + 2\lambda^2 C_\eta\sigma^2\|\phi(A^*X_t)\|^2\right)\exp\left(-\lambda\sum_{s=0}^{t-1}\rho^{t-s}\|\eta_s\|^2\right) \\ &\leq \exp\left(\lambda\rho^2\|X_t\|^2 + 2\lambda^2\rho^2 C_\eta\sigma^2\|X_t\|^2\right)\exp\left(-\lambda\sum_{s=0}^{t-1}\rho^{t-s}\|\eta_s\|^2\right) \\ &\leq \exp\left(\lambda\rho\|X_t\|^2\right)\exp\left(-\lambda\sum_{s=0}^{t-1}\rho^{t-s}\|\eta_s\|^2\right) \\ &= \exp\left(\lambda\rho Z_t\right)\end{aligned} \tag{6.10}$$

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 $\lambda \leq \frac{1-\rho}{2\rho C_\eta\sigma^2}$ to show $\lambda\rho^2 + 2\lambda^2\rho^2 C_\eta\sigma^2 \leq \rho\lambda$. In the last step, we have used the definition of Z_t. We iterate over Equation (6.10) and use the fact that $Z_0 = 0$ almost surely to conclude that whenever $\lambda \leq \frac{1-\rho}{2\rho C_\eta\sigma^2}$, we must have:

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

□

Now, let $Y_{t+1} = \sum_{s=0}^{t}\rho^{t-s}\|\eta_t\|^2$. We will now use Lemma 6.6.1 to bound $\mathbb{E}\exp(\lambda Y_{t+1})$ for $\lambda > 0$ small enough.

Lemma 6.6.3. *Suppose that Assumptions 6 and 8 hold and ρ be as given in Assumption 8. For any λ such that $0 \leq \lambda \leq \frac{1}{4dC_\eta\sigma^2}$, we have:*

$$\mathbb{E}\exp(\lambda Y_{t+1}) \leq \exp\left(8\frac{\lambda dC_\eta\sigma^2}{1-\rho}\right)$$

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

$$\mathbb{E}\exp(\lambda Y_{t+1}) = \prod_{s=0}^{t} N(\rho^{t-s}\lambda) \tag{6.11}$$

For $\beta \leq \frac{1}{4dC_\eta\sigma^2}$

$$\begin{aligned} N(\beta) = \mathbb{E}\exp(\beta\|\eta_s\|^2) &= \mathbb{E}\exp(\beta\sum_{i=1}^{d}\langle e_i, \eta_s\rangle^2) \\ &\leq \frac{1}{d}\sum_{i=1}^{d}\mathbb{E}\exp(\beta d\langle e_i, \eta_s\rangle^2) \leq 1 + 8\beta dC_\eta\sigma^2 \end{aligned} \tag{6.12}$$

In the last step, we have used Jensen's inequality for the function $x \to \exp(x)$ and then invoked the Lemma 6.6.1. Plugging this into Equation (6.11), we conclude:

$$\begin{aligned} \mathbb{E}\exp(\lambda Y_{t+1}) &\leq \prod_{s=0}^{t}\left(1 + 8\lambda d\rho^{t-s}dC_\eta\sigma^2\right) \leq \exp(\sum_{s=0}^{t} 8\lambda d\rho^{t-s}C_\eta\sigma^2) \\ &\leq \exp\left(8\frac{\lambda dC_\eta\sigma^2}{1-\rho}\right) \end{aligned} \tag{6.13}$$

□

Based on Lemmas 6.6.3 and 6.6.2, we will now state the following concentration inequality:

Theorem 6.6.4. *Suppose Assumptions 6 and 8 hold and ρ be as given in Assumption 8. Let X be distributed according π, the stationary distribution of $\mathsf{NLDS}(A^*, \mu, \phi)$. Then, for any $0 < \lambda \leq \lambda^* := \min(\frac{1}{8dC_\eta\sigma^2}, \frac{1-\rho}{4\rho C_\eta\sigma^2})$, we have:*

$$\mathbb{E}\exp\left(\lambda\|X\|^2\right) \leq \exp(\tfrac{8\lambda dC_\eta\sigma^2}{1-\rho})\,.$$

We conclude:

1. *Applying Chernoff bound with $\lambda = \lambda^*$, we conclude:*

$$\mathbb{P}\left(\|X\|^2 > \frac{8dC_\eta\sigma^2}{1-\rho} + \beta\right) \leq \exp(-\lambda^*\beta)\,.$$

2.

$$\mathbb{E}\|X\|^2 \leq \frac{8dC_\eta\sigma^2}{1-\rho}$$

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:

$$\begin{aligned}\mathbb{E}\exp\left(\lambda\|X_{t+1}\|^2\right) &= \mathbb{E}\exp\left(\lambda(Z_{t+1}+Y_{t+1})\right)\\ &\leq \sqrt{\mathbb{E}\exp(2\lambda Z_{t+1})}\sqrt{\mathbb{E}\exp(2\lambda Y_{t+1})}\\ &\leq \exp(\tfrac{8\lambda dC_\eta\sigma^2}{1-\rho}) \end{aligned} \tag{6.14}$$

Here we have used Lemmas 6.6.3 and 6.6.2 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 3, we show that π_t converges weakly to the stationary distribution π. We invoke Skhorokhod representation theorem to show that there exist random variables $\bar{X}_t \sim \pi_t$ and $X \sim \pi$ 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:

$$\mathbb{E}\exp\left(\lambda\|\bar{X}_{t+1}\|^2\right) \leq \exp(\tfrac{8\lambda dC_\eta\sigma^2}{1-\rho})\,.$$

Now, applying Fatou's Lemma to the equation above as $t \to \infty$, we conclude:

$$\mathbb{E}\exp\left(\lambda\|X\|^2\right) \leq \exp(\tfrac{8\lambda dC_\eta\sigma^2}{1-\rho})\,. \tag{6.15}$$

The concentration inequality follows from an application of Chernoff bound and the second moment bound follows from Jensen's inequality to Equation (6.15) (i.e, $\mathbb{E}\exp(Y) \geq \exp(\mathbb{E}Y)$). □

Possibly unstable systems

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

Lemma 6.6.5. *Under Assumption 7, we have:*

$$\|X_t\| \leq C_\rho \sum_{s=0}^{t-1} \rho^{t-s-1}\|\eta_s\|\,. \tag{6.16}$$

No suppose Assumption 6 also holds. Let $\delta \in (0, 1/2)$. Then with probability atleast $1-\delta$, we must have:

$$\sup_{0\leq t\leq T}\|X_t\| \leq CC_\rho\sqrt{C_\eta}S(\rho,T)\sigma\sqrt{d\log(\tfrac{T}{\delta})}\,.$$

Where $S(\rho,T) := \sum_{t=0}^{T-1}\rho^{T-t-1}$ and C is some universal constant.

Proof. We consider the notations established in Assumption 7. We will define the process $X_t^{(s)}$ by $X_0^{(s)} = \cdots = X_s^{(s)} = 0$ and $X_{t+1}^{(s)} = \phi(A^* X_t^{(s)}) + \eta_t$ for $t \geq s$, where η_t is the same noise sequence driving the process $X_0, X_1, \ldots, X_T$. Note that $X_t^{(s)} = h_{t-s}(0, \eta_s, \eta_{s+1}, \ldots, \eta_{t-1})$.

$$\begin{aligned} X_t - 0 &= X_t - X_t^{(1)} + X_t^{(1)} - X_t^{(2)} + \cdots + X_t^{(t)} - 0 \\ \implies \|X_t\| &\leq \sum_{s=0}^{t-1} \|X_t^{(s)} - X_t^{(s+1)}\| \\ &= \sum_{s=0}^{t-1} \|h_{t-s-1}(\eta_s, \ldots, \eta_{t-1}) - h_{t-s-1}(0, \eta_{s+1}, \ldots, \eta_{t-1})\| \\ &\leq \sum_{s=0}^{t-1} C_\rho \rho^{t-s-1} \|\eta_s\| \end{aligned} \tag{6.17}$$

In the last step, we have used Assumption 7. To prove the high probability bound, we note that $\mathbb{P}(\sup_{0 \leq s \leq T-1} \|\eta_s\| > C\sqrt{C_\eta}\sigma\sqrt{d \log(\frac{T}{\delta})}) \leq \delta$ for some universal constant C. □

6.7 Analysis of the Quasi Newton method

In this Section, we give the proof of theorem 6.2.1. 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 3. *Given a matrix $A = [a_1, a_2, \cdots, a_d]^\top$, let $\mathcal{R}(A) = \{a_1, \cdots, a_d\}$ 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 Defintion 3, we will consider the estimation of the i-th row a_i^*. Consider the gradient $\nabla \mathcal{L}_{\mathsf{prox}}^{(i)} : \mathbb{R}^d \to \mathbb{R}^d$ given by:

$$\nabla \mathcal{L}_{\mathsf{prox}}^{(i)}(a) := \frac{1}{T} \sum_{t=0}^{T-1} \left(\phi(\langle a, X_t \rangle) - \langle e_i, X_{t+1} \rangle\right) X_t \,.$$

We can write

$$\begin{aligned} \nabla \mathcal{L}_{\mathsf{prox}}^{(i)}(a) &= \frac{1}{T} \sum_{t=0}^{T-1} \left(\phi(\langle a, X_t \rangle) - \phi(\langle a_i^*, X_t \rangle)\right) X_t - \langle \eta_t, e_i \rangle X_t \\ &= \frac{1}{T} \sum_{t=0}^{T-1} \phi'(\beta_t) \langle a - a_i^*, X_t \rangle) X_t - \langle \eta_t, e_i \rangle X_t \\ &= \hat{K}_{a,i}(a - a_i^*) - \hat{N}_i \end{aligned} \tag{6.18}$$

Where β_t exist because of the mean value theorem. We can make sense of β_t even when ϕ is only weakly differentiable and check that the proof below still follows. Here, $\hat{K}_{a,i} := \frac{1}{T} \sum_{t=0}^{T-1} \phi'(\beta_t) X_t X_t^\top$ and $\hat{N}_i := \frac{1}{T} \sum_{t=0}^{T-1} \langle \eta_t, e_i \rangle X_t$. In the first step we have used the dynamics in Equation (1.9) to write

down X_{t+1} in terms of X_t and η_t.

We now define $\hat{G} \in \mathbb{R}^{d\times d}$ by $\hat{G} := \frac{1}{T}\sum_{t=0}^{T-1} X_t X_t^\top$. From the fact that $\zeta \le \phi'(\beta_t) \le 1$, we note that for every $a \in \mathbb{R}^d$:

$$\hat{G} \succeq \hat{K}_{a,i} \succeq \zeta \hat{G} \tag{6.19}$$

Now consider the Quasi Newton Step given in Algorithm 2.

$$a_i(l+1) = a_i(l) - 2\gamma \hat{G}^{-1} \nabla \mathcal{L}^{(i)}_{\mathsf{prox}}(a_i(l))$$

Denoting $\hat{K}_{a_i(l),i}$ by $\hat{K}_{l,i}$ we use Equation (6.18) to conclude:

$$\begin{aligned} a_i(l+1) - a_i^* &= a_i(l) - a_i^* - 2\gamma \hat{G}^{-1}\hat{K}_{l,i}(a_i(l+1) - a_i^*) + 2\gamma \hat{G}^{-1}\hat{N}_i \\ \implies \sqrt{\hat{G}}(a_i(l+1) - a_i^*) &= (I - 2\gamma \hat{G}^{-1/2}\hat{K}_{l,i}\hat{G}^{-1/2})\sqrt{\hat{G}}(a_i(l) - a_i^*) + 2\gamma \hat{G}^{-1/2}\hat{N}_i \end{aligned} \tag{6.20}$$

Picking $\gamma < \frac{1}{2}$, we conclude from Equation (6.19) that:

$$(1-2\gamma)I \preceq I - 2\gamma \hat{G}^{-1/2}\hat{K}_{l,i}\hat{G}^{-1/2} \preceq (1-2\gamma\zeta)I$$

We use the equation above in Equation (6.20) along with triangle inequality to conclude:

$$\|\sqrt{\hat{G}}(a_i(l+1) - a_i^*)\| \le (1-2\gamma\zeta)\|\sqrt{\hat{G}}(a_i(l) - a_i^*)\| + 2\gamma\|\hat{G}^{-1/2}\hat{N}_i\|$$

Unrolling the recursion above, we obtain that:

$$\begin{aligned} \|\sqrt{\hat{G}}(a_i(m) - a_i^*)\| &\le (1-2\gamma\zeta)^m \|\sqrt{\hat{G}}(a_i(0) - a_i^*)\| + \sum_{l=0}^{m-1} 2\gamma(1-2\gamma\zeta)^l \|\hat{G}^{-1/2}\hat{N}_i\| \\ &\le (1-2\gamma\zeta)^m \|\sqrt{\hat{G}}(a_i(0) - a_i^*)\| + \frac{1}{\zeta}\|\hat{G}^{-1/2}\hat{N}_i\| \end{aligned}$$

Letting A_m be the matrix with rows $a_i(m)$, we conclude:

$$\|A_m - A^*\|_{\mathsf{F}}^2 \le 2\frac{\lambda_{\max}(\hat{G})}{\lambda_{\min}(\hat{G})}(1-2\gamma\zeta)^{2m}\|A_0 - A^*\|_{\mathsf{F}}^2 + \frac{2}{\zeta^2\lambda_{\min}(\hat{G})}\sum_{i=1}^{d}\|\hat{G}^{-1/2}\hat{N}_i\|^2 \tag{6.21}$$

Proof of theorems 6.2.1 follows once we provide high probability bounds for various terms in Equation (6.21). 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. $\mathcal{D}_T(R) := \{\sup_{0\le t\le T}\|X_t\|^2 \le R\}$
2. $\mathcal{E}_T(\kappa) := \{\hat{G} \succeq \frac{\sigma^2 I}{\kappa}\}$

3. $\mathcal{D}_T(R,\kappa) := \mathcal{D}_T(R) \cap \mathcal{E}_T(\kappa)$

Lemma 6.7.1. *Under the Assumptions of Theorem 6.2.1, suppose $\delta \in (0,1/2)$ and take $R = C_\rho^2 C_\eta (S(\rho,T))^2 d\sigma^2 \log(\frac{2T}{\delta})$, $\kappa = 2$, and $T \geq \bar{C}_3 \left(d \log\left(\frac{R}{\sigma^2}\right) + \log\frac{1}{\delta}\right)$*

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

Where $\bar{C}, \bar{C}_3$ are constants depending only on C_η.

The proof of the above lemma follows from the following result.

Lemma 6.7.2. *Let $\delta \in (0, \frac{1}{2})$ Take $R = C_\rho^2 C_\eta (S(\rho,T))^2 d\sigma^2 \log(\frac{2T}{\delta})$, $\kappa = 2$ and suppose $T \geq \bar{C}_3 \left(d \log\left(\frac{R}{\sigma^2}\right) + \log\frac{1}{\delta}\right)$ for some constant $\bar{C}_3$ depending only on C_η. Then, we have:*

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

Proof. From Lemma 6.6.5, we conclude that taking $R \geq C_\rho^2 C_\eta (S(\rho,T))^2 \sigma^2 \log(\frac{2T}{\delta})$ ensures that $\mathbb{P}\left(\mathcal{D}_T(R)\right) \geq 1 - \frac{\delta}{2}$. Only in this proof, we define the following:

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

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

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

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:

$$\mathbb{P}\left(|\langle x, \bar{K}_X x\rangle| \geq 2C_\eta \sigma^2 \lambda x^\top \bar{G} x + \frac{\beta}{T}\middle| \mathcal{D}_T(R)\right) \leq \frac{2}{1-\delta}\exp(-\lambda\beta) \tag{6.22}$$

We will now invoke Theorem 5.39 in [185] to conclude that for some constant $\bar{C}_2$ which depends only on C_η:

$$\mathbb{P}\left(\bar{K}_\eta \preceq \left(1 - \bar{C}_2\left(\sqrt{\tfrac{d}{T}} + \sqrt{\tfrac{\log\frac{1}{\delta}}{T}}\right)\right)\sigma^2 I \middle| \mathcal{D}_T(R)\right) \leq \frac{\delta}{4} \tag{6.23}$$

Consider any ϵ net $\mathcal{N}_\epsilon$ over $\mathcal{S}^{d-1}$. By Corollary 4.2.13 in [185], we can take $|\mathcal{N}_\epsilon| \leq (1+\frac{2}{\epsilon})^d$. From Equations (6.22) and (6.23), we conclude that conditioned on $\mathcal{D}_T(R)$, with probability at-least $1-\frac{\delta}{4}-|\mathcal{N}_\epsilon|\frac{\exp(-\lambda\beta)}{1-\delta}$ we have:

$$\begin{aligned}
\inf_{x\in\mathcal{S}^{d-1}} x^\top \hat{G} x &\geq \inf_{y\in\mathcal{N}_\epsilon} y^\top \hat{G} y - 2\|\hat{G}\|\epsilon \\
&\geq \inf_{y\in\mathcal{N}_\epsilon} y^\top \bar{G} y - |y^\top \bar{K}_X y| + y^\top \bar{K}_\eta y - 2\|\hat{G}\|\epsilon \\
&\geq \inf_{y\in\mathcal{N}_\epsilon} y^\top \bar{G} y - |y^\top \bar{K}_X y| + y^\top \bar{K}_\eta y - 2R\epsilon \\
&\geq \inf_{y\in\mathcal{N}_\epsilon} y^\top \bar{G} y(1-2\lambda\sigma^2 C_\eta) - \frac{\beta}{T} + \sigma^2\left(1-\bar{C}_2\left(\sqrt{\tfrac{d}{T}}+\sqrt{\tfrac{\log\frac{1}{\delta}}{T}}\right)\right) - 2R\epsilon \qquad (6.24)
\end{aligned}$$

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 $\epsilon = \frac{1}{8R\sigma^2}$ 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 $T \geq \bar{C}_3\left(d\log\left(\frac{R}{\sigma^2}\right)+\log\frac{1}{\delta}\right)$ for some constant $\bar{C}_3$ depending only on C_η, 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:

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

We conclude the result from the equation above. □

Now we prove lemma 6.7.1

Proof of lemma 6.7.1. We invoke Theorem 1 in [108] with $S_t = T\hat{N}_i$, $V = T\sigma^2 I$, $\bar{V}_t = V + T\hat{G}$. We know that $\langle \eta, e_i\rangle$ 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 $1-\delta$:

$$\hat{N}_i^\top \bar{V}_t^{-1}\hat{N}_i \leq \frac{2C_\eta\sigma^2}{T^2}\log\left(\frac{\det(\bar{V}_t)^{1/2}\det(V)^{-1/2}}{\delta}\right). \qquad (6.25)$$

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

$$\det(\bar{V}_t)^{1/2}\det(V)^{-1/2} \leq (1+\tfrac{R}{\sigma^2})^{\frac{d}{2}} \qquad (6.26)$$

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

$$\frac{1}{3T}\hat{N}_i^\top \hat{G}^{-1}\hat{N}_i \leq \hat{N}_i^\top \bar{V}_t^{-1}\hat{N}_i \qquad (6.27)$$

Combining Equations (6.25), (6.26) and (6.27) and using Lemma 6.7.2, we conclude that with probability at-least $1-2\delta$, we have:

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

Using union bound, we conclude the result. □

6.7.1 Proof of theorem 6.2.1

Proof. Note that R in Lemma 6.7.1 is the same as R^* in the statement of Theorem 6.2.1. We combine the result of Lemma 6.7.1 with the Equation (6.21). Under the event $\mathcal{D}_T(R^*, \kappa)$, we have $\frac{\lambda_{\max}(\hat{G})}{\lambda_{\min}(\hat{G})} \leq \frac{2R}{\sigma^2}$ almost surely. Hence the result follows.

□

6.8 Analysis of SGD-RER

In this section we consider the following $(X_0, X_1, \ldots, X_T)$ to be a stationary sequence from $\mathsf{NLDS}(A^*, \mu, \phi)$. We make Assumptions 4, 5, 6 and 8. We aim to analyze Algorithm 3 and then prove Theorem 6.3.1.

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 $\mathcal{L}_{\mathsf{prox}}$ in the order described in Section 6.5. Formally, let $X_j^t \equiv X_{tS+j}$ denote the the j-th sample in buffer t. We denote, for $0 \leq i \leq B-1$, $X_{-i}^t \equiv X_{(S-1)-i}^t$ i.e., the i-th processed sample in buffer t. We use similar notation for noise samples i.e., $\eta_j^t \equiv \eta_{tS+j}$ and $\eta_{-j}^t \equiv \eta_{(S-1)-j}^t$.

The algorithm iterates are denoted by the sequence $(A_i^t : 0 \leq t \leq N-1, 0 \leq i \leq B-1)$ where A_i^t 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 $0, 1, \cdots N-1$. Formally

$$A_{i+1}^{t-1} = A_i^{t-1} - 2\gamma \left(\phi(A_i^{t-1} X_{-i}^{t-1}) - X_{-(i-1)}^{t-1} \right) X_{-i}^{t-1,\top} \tag{6.28}$$

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

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

6.8.1 Proof strategy

The proof of Theorem 6.3.1 involves many intricate steps. Therefore, we give a detailed overview about the proof below.

1. In Section 6.8.2 we first construct a fictitious coupled process $\tilde{X}_\tau$ such that for every data point within a buffer t, $\|\tilde{X}_\tau - X_\tau\| \lesssim \frac{1}{T^\alpha}$ for some fixed $\alpha > 0$ chosen arbitrarily beforehand. We then show that the iterates $\tilde{A}_i^t$ which are generated with $\mathsf{SGD-RER}$ is run with the coupled

process $\tilde{X}_\tau$ is very close to the actual iterate A_i^t. The coupled process has the advantage that the data in the successive buffers are independent. We then only deal with the coupled iterates $\tilde{A}_i^t$ and appeal to Lemma 6.8.3 to obtain bounds for A_i^t.

2. In Section 6.8.5, 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 $\tilde{H}_{0,B-1}^s$ defined on the data in buffer s 'contracts' the norm of $A_i^s - A^*$ giving the 'bias term' whereas the noise η_i^s presents the 'variance' term which is due to the inherent uncertaintly in the estimation problem.

3. We refer to Section 6.9.1 where we develop the contraction properties of the matrices $\prod_{s=0}^t \tilde{H}_{0,B-1}^s$ where we show that $\| \prod_{s=0}^t \tilde{H}_{0,B-1}^s \| \lesssim (1 - \zeta\gamma B \lambda_{\min}(G))^t$ in Theorem 6.9.4 after developing some probabilistic results regarding $\mathsf{NLDS}(A^*, \mu, \phi)$. This allows us show exponential decay of the bias.

4. We then turn to the squared variance term in Section 6.8.6. We decompose it into 'diagonal terms' with non-zero expectation and 'cross terms' with a vanishing expectation. Bounding the diagonal term is straight forward using standard recursive arguments and we give the bound in Claim 12.

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 6.5 for a short description). However, we establish 'algorithmic stability' in Section 6.8.7 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 $(\tilde{X}_\tau)$ by re-sampling one noise vector only and show that the iterates of $\mathsf{SGD} - \mathsf{RER}$ are not affected much.

6. We use the 'algorithmic stability' bounds to bound the cross terms in Sections 6.8.8. We then combine the bounds to obtain the bound on the 'variance term'

7. Finally, we analyze the tail averaged output in Sections 6.8.10, 6.8.11 and 6.8.12 and then combine these ingredients to prove Theorem 6.3.1.

6.8.2 Basic notations and coupled process

Definition 4 (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* $\mathsf{NLDS}(A^*, \mu, \phi)$ *model.*

2. Then, each buffer has the same recursion as eq (1.9)*:*

$$\tilde{X}_{i+1}^t = \phi(A^* \tilde{X}_i^t) + \eta_i^t,\ i = 0, 1, \cdots S-1, \tag{6.29}$$

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

Lemma 6.8.1 (Coupling Lemma)**.** *Under Assumption 7, for any buffer* t*, we have* $\left\|X_i^t - \tilde{X}_i^t\right\| \leq C_\rho \rho^i \left\|X_0^t - \tilde{X}_0^t\right\|$*, a.s. Hence*

$$\left\|X_i^t X_i^{t,\top} - \tilde{X}_i^t \tilde{X}_i^{t,\top}\right\| \leq 2(\sup_{\tau \leq T} \|X_\tau\|) \left\|X_i^t - \tilde{X}_i^t\right\| \leq 4 \sup_{\tau \leq T} \|X_\tau\|^2 C_\rho \rho^i \tag{6.30}$$

With the above notation, we can write (6.28) in terms of a generic row (say row r) $a_{i+1}^{t-1,\top}$ of A_{i+1}^{t-1} as follows. Let ε_{-i}^{t-1} denote the element of η_{-i}^{t-1} in row r. Similarly let $a^{*,\top} \equiv (a^*)^\top$ denote the row r of A^*. Then

$$a_{i+1}^{t-1,\top} = a_i^{t-1,\top} - 2\gamma \left(\phi(X_{-i}^{t-1,\top} a_i^{t-1}) - \phi(X_{-i}^{t-1,\top} a^*)\right) X_{-i}^{t-1,\top} + 2\gamma \varepsilon_{-i}^{t-1} X_{-i}^{t-1,\top} \tag{6.31}$$

Now, by the mean value theorem we can write

$$\phi(X_{-i}^{t-1,\top} a_i^{t-1}) - \phi(X_{-i}^{t-1,\top} a^*) = \phi'(\xi_{-i}^{t-1})(a_{-i}^{t-1} - a^*)^\top X_{-i}^{t-1} \tag{6.32}$$

where ξ_{-i}^{t-1} lies between $X_{-i}^{t-1,\top} a_i^{t-1}$ and $X_{-i}^{t-1,\top} a^*$. Hence we obtain

$$(a_{i+1}^{t-1} - a^*)^\top = (a_i^{t-1} - a^*)^\top (I - 2\gamma \phi'(\xi_{-i}^{t-1}) X_{-i}^{t-1} X_{-i}^{t-1,\top}) + 2\gamma \varepsilon_{-i}^{t-1} X_{-i}^{t-1,\top} \tag{6.33}$$

Now we provide a bound on the algorithmic iterates.

Lemma 6.8.2. *Let* $R_{\max} := \sup_{\tau \leq T}(\|X_\tau\|^2, \left\|\tilde{X}\right\|^2)$ *and suppose* $\gamma \leq \frac{1}{2R_{\max}}$*. For every* $t \in [N]$ *and* $i \in [B]$ *we have:*

$$\|a_i^t\| \leq 2\gamma R_{\max} T\,.$$

Proof. Let the row under consideration be the k-th row and e_k be the standard basis vector. Consider the $\mathsf{SGD} - \mathsf{RER}$ iteration:

$$\begin{aligned} a_{i+1}^t &= a_i^t - 2\gamma \left(\phi(\langle a_i^t, X_{-i}^t \rangle) - X_{-(i-1)}^t\right) X_{-i}^t \\ &= (I - 2\gamma \zeta_{t,i} X_{-i}^t X_{-i}^{t,\top}) a_i^t + 2\gamma \langle X_{-(i-1))}^t, e_k \rangle X_{-i}^t \end{aligned} \tag{6.34}$$

Where $\zeta_{t,i} := \frac{\phi(\langle a_i^t, X_{-i}^t \rangle)}{\langle a_i^t, X_{-i}^t \rangle} \in [\zeta, 1]$ exists in a weak sense due to our assumptions on ϕ. Observe that for our choice of γ, we have $\|(I - 2\gamma \zeta_{t,i} X_{-i}^t X_{-i}^{t,\top})\| \leq 1$ and $\|\langle X_{-(i-1)}^t, e_k \rangle X_{-i}^{t,\top}\| \leq R_{\max}$. Therefore,

triangle inequality implies:

$$\|a_{i+1}^t\| \leq \|a_i^t\| + 2\gamma R_{\max}$$

We conclude the bound in the Lemma. □

Definition 5 (Coupled SGD Iteration)**.** *Consider the process described in Defintion 4. We define* $\mathsf{SGD} - \mathsf{RER}$ *iterates run with the coupled process* $(\tilde{X}_i^t)$ *as follows:*

$$\tilde{A}_0^0 = A_0^0$$

$$\tilde{A}_{i+1}^{t-1} = \tilde{A}_i^{t-1} - 2\gamma \left(\phi(\tilde{A}_i^{t-1}\tilde{X}_{-i}^{t-1}) - \tilde{X}_{-(i-1)}^{t-1}\right)\tilde{X}_{-i}^{t-1,\top} \tag{6.35}$$

Using Lemma 6.8.1, 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.

Lemma 6.8.3. *Suppose* $\gamma < \frac{1}{2R_{\max}}$*. we have for every* $t \in [N]$ *and* $i \in [B]$*.*

$$\|a_i^t - \tilde{a}_i^t\| \leq (16\gamma^2 R_{\max}^2 T^2 + 8\gamma R_{\max} T)\rho^u$$

Proof. Let the row under consideration be the k-th row and e_k be the standard basis vector.

$$\begin{aligned} a_{i+1}^t &= a_i^t - 2\gamma(\phi(\langle a_i^t, X_{-i}^t\rangle) - \langle e_k, X_{-(i-1)}^t\rangle)X_{-i}^t \\ &= a_i^t - 2\gamma(\phi(\langle a_i^t, \tilde{X}_{-i}^t\rangle) - \langle e_k, \tilde{X}_{-(i-1)}^t\rangle)\tilde{X}_{-i}^t + \Delta_{t,i} \end{aligned} \tag{6.36}$$

Where

$$\Delta_{t,i} := 2\gamma\left(\phi(\langle a_i^t, \tilde{X}_{-i}^t\rangle)\tilde{X}_{-i}^t - \phi(\langle a_i^t, X_{-i}^t\rangle)X_{-i}^t\right) + 2\gamma\left(\langle X_{-(i-1)}^t, e_k\rangle X_{-i}^t - \langle \tilde{X}_{-(i-1)}^t, e_k\rangle \tilde{X}_{-i}^t\right).$$

Using Lemmas 6.8.2 and 6.8.1, we conclude that:

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

Using the recursion for $\tilde{a}_i^t$, we conclude:

$$\begin{aligned} & a_{i+1}^t - \tilde{a}_{i+1}^t = (I - 2\gamma\tilde{\zeta}_{t,i}\tilde{X}_i^t\tilde{X}_i^{t,\top})(a_i^t - \tilde{a}_i^t) + \Delta_{t,i} \\ \implies & \left\|a_{i+1}^t - \tilde{a}_{i+1}^t\right\| \leq \left\|a_i^t - \tilde{a}_i^t\right\| \left\|(I - 2\gamma\tilde{\zeta}_{t,i}\tilde{X}_i^t\tilde{X}_i^{t,\top})\right\| + (16\gamma^2 R_{\max}^2 T + 8\gamma R_{\max})\rho^u \\ \implies & \left\|a_{i+1}^t - \tilde{a}_{i+1}^t\right\| \leq \left\|a_i^t - \tilde{a}_i^t\right\| + (16\gamma^2 R_{\max}^2 T + 8\gamma R_{\max})\rho^u \end{aligned} \tag{6.37}$$

In the first step, $\tilde{\zeta}_{t,i} := \frac{\phi(\langle a_i^t, \tilde{X}_{-i}^t\rangle) - \phi(\langle \tilde{a}_i^t, \tilde{X}_{-i}^t\rangle}{\langle a_i^t, \tilde{X}_{-i}^t\rangle - \langle \tilde{a}_i^t, \tilde{X}_{-i}^t\rangle} \in [\zeta, 1]$. In the last step we have used the fact that under the conditions on γ, we must have $\left\|(I - 2\gamma\tilde{\zeta}_{t,i}\tilde{X}_i^t\tilde{X}_i^{t,\top})\right\| \leq 1$. We conclude the statement of the lemma from Equation (6.37). $\square$

We note that we can just analyze the iterates $\tilde{A}_i^t$ and then use Lemma 6.8.3 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.

6.8.3 Notations

We define the following notations. Let $R > 0$ to be decided later.

$$X_{-i}^t = X_{(S-1)-i}^t,\ 0 \leq i \leq S-1, \qquad \phi'(\tilde{\xi}_{-i}^t) = \frac{\phi(\tilde{a}_{-i}^{t,\top}\tilde{X}_{-i}^t) - \phi(a^{*,\top}\tilde{X}_{-i}^t)}{\left(\tilde{a}_{-i}^t - a^*\right)^\top \tilde{X}_{-i}^t}$$

$$\tilde{P}_{-i}^t = \left(I - 2\gamma\phi'\left(\tilde{\xi}_{-i}^t\right)\tilde{X}_{-i}^t\tilde{X}_{-i}^{t,\top}\right), \quad \tilde{H}_{i,j}^t = \begin{cases} \prod_{s=i}^j \tilde{P}_{-s}^t & i \leq j \\ I & i > j \end{cases},$$

$$\hat{\gamma} = 4\gamma(1-\gamma R), \quad \mathcal{C}_{-j}^t = \left\{\left\|X_{-j}^t\right\|^2 \leq R\right\}, \quad \tilde{\mathcal{C}}_{-j}^t = \left\{\left\|\tilde{X}_{-j}^t\right\|^2 \leq R\right\},$$

$$\mathcal{D}_{-j}^t = \left\{\left\|X_{-i}^t\right\|^2 \leq R :\ j \leq i \leq B-1\right\} = \bigcap_{i=j}^{B-1} \mathcal{C}_{-i}^t,$$

$$\mathcal{D}^{s,t} = \begin{cases} \bigcap_{r=s}^t \mathcal{D}_{-0}^r & s \leq t \\ \Omega & s > t \end{cases}, \quad \tilde{\mathcal{D}}_{-j}^t = \left\{\left\|\tilde{X}_{-i}^t\right\|^2 \leq R :\ j \leq i \leq B-1\right\} = \bigcap_{i=j}^{B-1} \tilde{\mathcal{C}}_{-i}^t,$$

$$\tilde{\mathcal{D}}^{s,t} = \begin{cases} \bigcap_{r=s}^t \tilde{\mathcal{D}}_{-0}^r & s \leq t \\ \Omega & s > t \end{cases}, \quad \hat{\mathcal{D}}_{-j}^t = \mathcal{D}_{-j}^t \cap \tilde{\mathcal{D}}_{-j}^t, \quad \hat{\mathcal{D}}^{s,t} = \mathcal{D}^{s,t} \cap \tilde{\mathcal{D}}^{s,t}.$$

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

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

6.8.4 Setting the parameter values:

We make Assumptions 4, 5, 6 and 8 throughout. We set the parameters for $\mathsf{SGD-RER}$ as follows for the rest of the analysis. We note that some of these parameter values were set in Section 6.3.

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. $u \geq \frac{2\alpha\log T}{\log(\frac{1}{\rho})} = O(\tau_{\mathsf{mix}}\log T)$

6. $B \geq \max\left(\bar{C}_1 \frac{d}{(1-\rho)(1-\rho^2)}, 10u\right)$ where $\bar{C}_1$ depends only on C_η(see Theorem 6.9.4)

7. $\gamma \leq \min\left(\frac{\zeta}{4BR(1+\zeta)}, 1/2R\right)$ (see Theorem 6.9.4)

From Assumption 6 and Theorem 6.6.4, we conclude that for this choice of R and β, we must have:

$$\mathbb{P}\left[\left(\hat{\mathcal{D}}^{0,N-1} \cap \cap_{r=0}^{N-1} \mathcal{E}^r_{0,B-1}\right)^C\right] \leq \frac{1}{2T^\alpha} \tag{6.38}$$

6.8.5 Bias-variance decomposition

Using the above notation we can unroll the recursion in (6.33) as follows. We will only focus on the algorithmic iterated at the end of each buffer, i.e., we set $i = B-1$ in (6.33).

$$(\tilde{a}_B^{t-1} - a^*)^\top = (a_0 - a^*)^\top \prod_{s=0}^{t-1} \tilde{H}^s_{0,B-1} + 2\gamma \sum_{r=1}^{t}\sum_{j=0}^{B-1} \varepsilon^{t-r}_{-j} \tilde{X}^{t-r,\top}_{-j} \tilde{H}^{t-r}_{j+1,B-1} \prod_{s=r-1}^{1} \tilde{H}^{t-s}_{0,B-1} \tag{6.39}$$

We call the above the *bias-variance* decomposition where

$$(\tilde{a}_B^{t-1,b} - a^*)^\top = (a_0 - a^*)^\top \prod_{s=0}^{t-1} \tilde{H}^s_{0,B-1} \tag{6.40}$$

is the bias, and

$$(\tilde{a}_B^{t-1,v})^\top = 2\gamma \sum_{r=1}^{t}\sum_{j=0}^{B-1} \varepsilon^{t-r}_{-j} \tilde{X}^{t-r,\top}_{-j} \tilde{H}^{t-r}_{j+1,B-1} \prod_{s=r-1}^{1} \tilde{H}^{t-s}_{0,B-1} \tag{6.41}$$

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

Lemma 6.8.4.

$$\left\|\tilde{a}_B^{t-1} - a^*\right\|^2 \preceq 2\left(\left\|\tilde{a}_B^{t-1,b} - a^*\right\|^2 + \left\|\tilde{a}_B^{t-1,v}\right\|^2\right) \tag{6.42}$$

6.8.6 Variance of last iterate - Diagonal Terms

In this section our goal is to decompose $\left\|\tilde{a}_B^{t-1,v}\right\|^2$ 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 6.8.5. *For $k \le t$ define S_k^t as*

$$S_k^t = \sum_{r=k}^{t} \sum_{j=0}^{B-1} \phi'(\tilde{\xi}_{-j}^{t-r}) \left(\prod_{s=1}^{r-1} \tilde{H}_{0,B-1}^{t-s,\top} \right) \tilde{H}_{j+1,B-1}^{t-r,\top} \tilde{X}_{-j}^{t-r} \tilde{X}_{-j}^{t-r,\top} \tilde{H}_{j+1,B-1}^{t-r} \left(\prod_{s=r-1}^{1} \tilde{H}_{0,B-1}^{t-s} \right) \quad (6.43)$$

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

$$S_1^t \preceq \frac{1}{\hat{\gamma}} \left(I - \left(\prod_{s=1}^{t} \tilde{H}_{0,B-1}^{t-s,\top} \right) \left(\prod_{s=t}^{1} \tilde{H}_{0,B-1}^{t-s} \right) \right) \quad (6.44)$$

where $\hat{\gamma} = 4\gamma(1-\gamma R)$

Proof. The proof is similar to that of [110, Claim 1]. □

Next, we write $\left\| \tilde{a}_B^{t-1,v} \right\|^2$ as

$$\left\| \tilde{a}_B^{t-1,v} \right\|^2 = \sum_{r=1}^{t} \sum_{j=0}^{B-1} \mathrm{Dg}(t,r,j) + \sum_{r_1,r_2} \sum_{j_1,j_2} \mathrm{Cr}(t,r_1,r_2,j_1,j_2) \quad (6.45)$$

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

$$\mathrm{Dg}(t,r,j) = 4\gamma^2 |\varepsilon_{-j}^{t-r}|^2 \cdot \tilde{X}_{-j}^{t-r,\top} \tilde{H}_{j+1,B-1}^{t-r} \left(\prod_{s=r-1}^{1} \tilde{H}_{0,B-1}^{t-s} \right) \left(\prod_{s=1}^{r-1} \tilde{H}_{0,B-1}^{t-s,\top} \right) \tilde{H}_{j+1,B-1}^{t-r,\top} \tilde{X}_{-j}^{t-r} \quad (6.46)$$

and

$$\begin{aligned} \mathrm{Cr}(t,r_1,r_2,j_1,j_2) = 4\gamma^2 \varepsilon_{-j_2}^{t-r_2} \tilde{X}_{-j_2}^{t-r_2,\top} \tilde{H}_{j_2+1,B-1}^{t-r_2} \left(\prod_{s=r_2-1}^{1} \tilde{H}_{0,B-1}^{t-s} \right) \cdot \\ \left(\prod_{s=1}^{r_1-1} \tilde{H}_{0,B-1}^{t-s,\top} \right) \tilde{H}_{j_1+1,B-1}^{t-r_1,\top} \tilde{X}_{-j_1}^{t-r_1} \varepsilon_{-j_1}^{t-r_1} \end{aligned} \quad (6.47)$$

Finally, we bound the diagonal term:

Claim 12.

$$\mathbb{E}\left[\sum_{r=1}^{t} \sum_{j=0}^{B-1} \mathrm{Dg}(t,r,j) 1\left[\tilde{\mathcal{D}}^{0,t-1} \right] \right] \le \frac{\gamma d}{\zeta(1-\gamma R)} \beta + 16 C_\eta \sigma^2 \gamma^2 R T \frac{1}{T^{\alpha/2}} \quad (6.48)$$

Proof. Notice that we can write

$$\begin{aligned} \mathrm{Dg}(t,r,j) \le 4\gamma^2 \left(\beta + |\varepsilon_{-j}^{t-r}|^2 1[|\varepsilon_{-j}^{t-r}|^2 > \beta] \right) \cdot \\ \tilde{X}_{-j}^{t-r,\top} \tilde{H}_{j+1,B-1}^{t-r} \left(\prod_{s=r-1}^{1} \tilde{H}_{0,B-1}^{t-s} \right) \left(\prod_{s=1}^{r-1} \tilde{H}_{0,B-1}^{t-s,\top} \right) \tilde{H}_{j+1,B-1}^{t-r,\top} \tilde{X}_{-j}^{t-r} \end{aligned} \quad (6.49)$$

Further

$$\tilde{X}_{-j}^{t-r,\top}\tilde{H}_{j+1,B-1}^{t-r}\left(\prod_{s=r-1}^{1}\tilde{H}_{0,B-1}^{t-s}\right)\left(\prod_{s=1}^{r-1}\tilde{H}_{0,B-1}^{t-s,\top}\right)\tilde{H}_{j+1,B-1}^{t-r,\top}\tilde{X}_{-j}^{t-r}1\left[\tilde{\mathcal{D}}^{0,t-1}\right]\leq R \tag{6.50}$$

Combining the above two we obtain

$$\sum_{r=1}^{t}\sum_{j=0}^{B-1}\mathrm{Dg}(t,r,j)1\left[\tilde{\mathcal{D}}^{0,t-1}\right]\leq 4\gamma^2\beta\frac{\mathrm{Tr}\,S_1^t}{\zeta}+R\sum_{r=1}^{t}\sum_{j=0}^{B-1}|\varepsilon_{-j}^{t-r}|^2 1[|\varepsilon_{-j}^{t-r}|^2>\beta] \tag{6.51}$$

where S_1^t is defined in (6.43). Now taking expectation, and using lemma 6.8.5 and Caucy-Schwarz inequality for the first and second terms, respectively, in (6.51) we obtain the claim. Here we use the fact that $\mathbb{E}\left[|\varepsilon_{-j}^{t-r}|^4\right]\leq 16C_\eta^2\sigma^4$ from [148, Theorem 2.1] □

6.8.7 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 ϕ was identity, then $\mathbb{E}\left[\mathrm{Cr}(t,r_1,r_2,j_1,j_2)\right]$ would vanish. But in the non-linear setting, this does not happen due to dependencies between ε_{-j}^{t-r} and $\tilde{H}_{j+1,B-1}^{t-r}\left(\prod_{s=r-1}^{1}\tilde{H}_{0,B-1}^{t-s}\right)$ through the algorithmic iterates. We can still show that $\mathbb{E}\left[\mathrm{Cr}(t,r_1,r_2,j_1,j_2)\right]\approx 0$ by showing that the iterates depend very weakly on each of the noise co-ordinates ε_{-j}^{t-r}. 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 $(X_\tau)_\tau$ and the coupled process $(\tilde{X}_\tau)_\tau$. Let the corresponding (coupled) algorithmic iterates be $(\tilde{a}_i^s : 0\leq s\leq N-1,\, 0\leq i\leq B-1)$. Now $\tilde{a}_i^s$ are functions of X_0 and noise vectors $\{\eta_{-i}^s : 0\leq s\leq N-1,\, 0\leq i\leq S-1\}$. Suppose we re-sample the noise η_{-j}^r independently of everything else to get $\bar{\eta}_{-j}^r$. So the new noise samples are:

$$\left(\eta_0^0,\eta_1^0,\cdots,\eta_0^r,\cdots,\eta_{(S-1)-(j+1)}^r,\bar{\eta}_{(S-1)-j}^r,\eta_{(S-1)-(j-1)}^r,\cdots\right).$$

We then run the dynamics in Equation (1.9) with the new noise samples to obtain $(\bar{X}_\tau)_\tau$ and the new coupled process $(\bar{\tilde{X}}_\tau)_\tau$ obtained through the new noise sequence (but same stationary renewal given in Definition 4), and they satisfy the following:

$$\bar{X}^s_{-i} = \begin{cases} X^s_{-i}, & s < r,\ 0 \le i \le S-1 \\ X^r_{-i}, & s = r,\ j \le i \le S-1 \end{cases}$$

$$\bar{\tilde{X}}^s_{-i} = \begin{cases} \tilde{X}^s_{-i}, & s \in \{1, \cdots, r-1, r+1, \cdots, N-1\},\ 0 \le i \le S-1 \\ \tilde{X}^r_{-i}, & s = r,\ j \le i \le S-1 \end{cases}$$

We obtain the iterates $\bar{\tilde{a}}^s_i$ by running the update Equation (6.28) with the data $\bar{\tilde{X}}_\tau$ instead of X_τ. Accordingly, the algorithmic iterates change to $(\bar{\tilde{a}}^s_i : 0 \le s \le N-1,\ 0 \le i \le B-1)$ that satisfy

$$\bar{\tilde{a}}^s_i = \tilde{a}^s_i \quad \text{for } s < r,\ 0 \le i \le B-1$$

This is because, resampling η_τ does note change the value of data $\tilde{X}_{\tau'}$ for $\tau' \le \tau$. Under the setting we have the following lemma:

Lemma 6.8.6. *Let $\mathcal{A}^{t-1}$ be the following event*

$$\mathcal{A}^{t-1} = \bigcap_{r=0}^{t-1} \bigcap_{j=0}^{B-1} \left\{ \left\| \tilde{a}^r_j - a^* \right\| \le \left\| a_0 - a^* \right\| + \bar{C} \frac{\sqrt{R\beta}}{\zeta \lambda_{\min}} \right\} \tag{6.52}$$

For some constant $\bar{C}$ depending only on C_η, we have for any $1 \le t \le N$

$$\mathbb{P}\left[\hat{\mathcal{D}}^{0,N-1} \cap \mathcal{A}^{N-1} \cap \cap_{r=0}^{N-1} \mathcal{E}^r_{0,B-1} \right] \ge 1 - \frac{1}{T^\alpha} \tag{6.53}$$

Further more, on the event $\mathcal{E}^r_{0,j} \cap \tilde{\mathcal{D}}^{r,N-1} \cap \mathcal{A}^r$ we have:

$$\bar{\tilde{a}}^s_i = \tilde{a}^s_i,\ 0 \le s < r,\ 0 \le i \le B-1 \tag{6.54}$$

$$\bar{\tilde{a}}^r_0 = \tilde{a}^r_0 \tag{6.55}$$

and for $s \ge r$ we have

$$\left\| \bar{\tilde{a}}^s_i - \tilde{a}^s_i \right\| \le \bar{C}_2 \gamma RB \frac{\sqrt{R\beta}}{\zeta \lambda_{\min}} + 8\gamma RB \left\| a_0 - a^* \right\| \le \bar{C}_2 \gamma RB \frac{\sqrt{R\beta}}{\zeta \lambda_{\min}} \tag{6.56}$$

We give the proof in Section 6.9

Remark 15. *In expression* (6.56)*, we have suppressed the dependence of $\|a_0 - a^*\|$ for the ease of exposition with the rationale being that since $a_0 = 0$, it sould be lower order compared to $\sqrt{R\beta}$.*

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 η^r_{-j}. That is, for any function $f((a_\tau),(X_\tau),(\tilde{X}_\tau))$ we have

$$\mathcal{R}^r_{-j}\left(f((a_\tau),(\tilde{a}_\tau),(X_\tau),(\tilde{X}_\tau))\right) = f((\bar{a}_\tau),(\bar{\tilde{a}}_\tau),(\bar{X}_\tau),(\bar{\tilde{X}}_\tau)) \tag{6.57}$$

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 6.8.7. *On the event $\mathcal{E}^r_{0,j} \cap \tilde{\mathcal{D}}^{0,t-1} \cap \mathcal{A}^{t-1}$, for some constant C depending only on C_η:*

$$\begin{aligned}&\left\|\tilde{H}^{t-r}_{j+1,B-1}\left(\prod_{s=r-1}^{1}\tilde{H}^{t-s}_{0,B-1}\right) - \mathcal{R}^{t-r}_{-j}\tilde{H}^{t-r}_{j+1,B-1}\left(\prod_{s=r-1}^{1}\mathcal{R}^{t-r}_{-j}\tilde{H}^{t-s}_{0,B-1}\right)\right\| \\ &\leq \bar{C}\frac{Bt\left\|\phi''\right\|\gamma^2R^3B\sqrt{\beta}}{\zeta\lambda_{\min}}\end{aligned} \tag{6.58}$$

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

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

$$\left\|\prod_{l=1}^{k}A_l - \prod_{l=1}^{k}\bar{A}_l\right\| \leq \sum_{l=1}^{k}\left\|A_l - \bar{A}_l\right\| \tag{6.59}$$

Therefore, we will just bound each of the component differences $\left\|A_l - \bar{A}_l\right\|$. To this end, consider a typical term $I - 2\gamma\phi'(\tilde{\xi}^{t-s}_{-i})\tilde{X}^{t-s}_{-j}\tilde{X}^{t-s,\top}_{-i}$. We have

$$\begin{aligned}&\left(I - 2\gamma\phi'(\tilde{\xi}^{t-s}_{-i})\tilde{X}^{t-s}_{-i}\tilde{X}^{t-s,\top}_{-i}\right) - \mathcal{R}^{t-r}_{-j}\left(I - 2\gamma\phi'(\tilde{\xi}^{t-s}_{-i})\tilde{X}^{t-s}_{-i}\tilde{X}^{t-s,\top}_{-i}\right) \\ &= 2\gamma(\phi'(\tilde{\xi}^{t-s}_{-i}) - \mathcal{R}^{t-r}_{-j}\phi'(\tilde{\xi}^{t-s}_{-i}))\tilde{X}^{t-s}_{-i}\tilde{X}^{t-s,\top}_{-i}\end{aligned} \tag{6.60}$$

Now

$$\phi'(\tilde{\xi}^{t-s}_{-i}) - \mathcal{R}^{t-r}_{-j}\phi'(\tilde{\xi}^{t-s}_{-i}) = \frac{\phi(\tilde{a}^{t-s,\top}_i\tilde{X}^{t-s}_{-i}) - \phi(a^{*,\top}\tilde{X}^{t-s}_{-i})}{(\tilde{a}^{t-s}_i - a^*)^\top\tilde{X}^{t-s}_{-i}} - \frac{\phi(\bar{\tilde{a}}^{t-s,\top}_i\tilde{X}^{t-s}_{-i}) - \phi(a^{*,\top}\tilde{X}^{t-s}_{-i})}{(\bar{\tilde{a}}^{t-s}_i - a^*)^\top\tilde{X}^{t-s}_{-i}} \tag{6.61}$$

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 ξ such that:

$$g'(x) = \frac{f(x_0) - (f(x) + (x_0 - x)f'(x)}{(x-x_0)^2} = \frac{1}{2}f''(\xi)$$

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 ξ_1 between x and y. Here again we use the mean value theorem in the equality above. Now we will apply this result to ϕ with $x = \tilde{a}_i^{t-s,\top}\tilde{X}_{-i}^{t-s}$, $y = \bar{\tilde{a}}_i^{t-s,\top}\tilde{X}_{-i}^{t-s}$ and $x_0 = a^{*,\top}\tilde{X}_{-i}^{t-s}$ to get

$$\left|\phi'(\tilde{\xi}_{-i}^{t-s}) - \mathcal{R}_{-j}^{t-r}\phi'(\tilde{\xi}_{-i}^{t-s})\right| \leq \frac{1}{2}\|\phi''\|\left\|\tilde{a}_i^{t-s} - \bar{\tilde{a}}_i^{t-s}\right\|\left\|\tilde{X}_{-i}^{t-s}\right\| \tag{6.62}$$

Now we appeal to lemma 6.8.6. In particular, using equation (6.56) we see that, on the event $\mathcal{E}_{0,j}^r \cap \tilde{\mathcal{D}}^{0,t-1} \cap \mathcal{A}^{t-1}$,

$$\begin{aligned}\left|\phi'(\tilde{\xi}_{-i}^{t-s}) - \mathcal{R}_{-j}^{t-r}\phi'(\tilde{\xi}_{-i}^{t-s})\right| &\leq \frac{1}{2}\|\phi''\|\,128C\gamma RB\frac{\sqrt{R\beta}}{\zeta\lambda_{\min}}\sqrt{R} \\ &= 64C\|\phi''\|\,\gamma R^2 B\frac{\sqrt{\beta}}{\zeta\lambda_{\min}} \qquad (6.63)\end{aligned}$$

$$\implies \left\|2\gamma(\phi'(\tilde{\xi}_{-i}^{t-s}) - \mathcal{R}_{-j}^{t-r}\phi'(\tilde{\xi}_{-i}^{t-s}))\tilde{X}_{-i}^{t-s}\tilde{X}_{-i}^{t-s,\top}\right\| \leq 128C\|\phi''\|\,\gamma^2 R^3 B\frac{\sqrt{\beta}}{\zeta\lambda_{\min}} \tag{6.64}$$

We now use Equation (6.59) with $k \leq Bt$ along with Equation (6.64) to conclude the statement of the lemma. □

6.8.8 Bound $\mathrm{Cr}(t, r_1, r_2, j_1, j_2)$

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

Claim 13.

$$\left|\mathbb{E}\left[\sum_{r=1}^{t}\sum_{j_1\neq j_2}\mathrm{Cr}(t,r,r,j_1,j_2)1\left[\tilde{\mathcal{D}}^{0,t-1}\right]\right]\right| \leq \bar{C}\left[\frac{\sigma^2\gamma^2 RB}{T^{\alpha/2-1}} + \frac{\|\phi''\|\,\gamma^4 T^2 R^4 B^2\sigma^2\sqrt{\beta}}{\zeta\lambda_{\min}}\right] \tag{6.65}$$

Where $\bar{C}$ is a constant depending only on C_η

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

Let Cr' denote the resampled version of Cr as defined below

$$\begin{aligned}\mathrm{Cr}'(t,r,r,j_1,j_2) := 4\gamma^2\varepsilon_{-j_1}^{t-r}\varepsilon_{-j_2}^{t-r}\mathcal{R}_{-j_1}^{t-r}\Bigg[\tilde{X}_{-j_2}^{t-r,\top}\tilde{H}_{j_2+1,B-1}^{t-r}\left(\prod_{s=r-1}^{1}\tilde{H}_{0,B-1}^{t-s}\right)\cdot \\ \left(\prod_{s=1}^{r-1}\tilde{H}_{0,B-1}^{t-s,\top}\right)\tilde{H}_{j_1+1,B-1}^{t-r,\top}\tilde{X}_{-j_1}^{t-r}\Bigg] \\ = 4\gamma^2\varepsilon_{-j_1}^{t-r}\varepsilon_{-j_2}^{t-r}\tilde{X}_{-j_2}^{t-r,\top}\mathcal{R}_{-j_1}^{t-r}\left(\tilde{H}_{j_2+1,B-1}^{t-r}\right)\mathcal{R}_{-j_1}^{t-r}\left(\prod_{s=r-1}^{1}\tilde{H}_{0,B-1}^{t-s}\right)\cdot \\ \mathcal{R}_{-j_1}^{t-r}\left(\prod_{s=1}^{r-1}\tilde{H}_{0,B-1}^{t-s,\top}\right)\mathcal{R}_{-j_1}^{t-r}\left(\tilde{H}_{j_1+1,B-1}^{t-r,\top}\right)\tilde{X}_{-j_1}^{t-r} \qquad (6.66)\end{aligned}$$

where we have used the fact that $\mathcal{R}_{-j_1}^{t-r}$ has no effect on the items from the process $(\tilde{X}_\tau)_\tau$ *that appear* in the expression above. Note the this is **not** $\mathcal{R}_{-j_1}^{t-r}\,\mathrm{Cr}$, since in $\mathcal{R}_{-j_1}^{t-r}\,\mathrm{Cr}$ we would have $\bar{\epsilon}_{-j_1}^{t-r}$ instead. Now, since the new algorithmic iterates $(\bar{\tilde{a}}_i^s)$ depend on $\bar{\eta}_{-j_1}^{t-r}$ but not on $\eta_{-j_1}^{t-r}$, it is immediate that

$$\mathbb{E}\left[\mathrm{Cr}'(t,r,r,j_1,j_2)\right] = 0$$

For convenience, we introduce some notation which is only used in this proof. $\mathrm{Cr}'(t,r,r,j_1,j_2)$ can be written in the form $4\gamma^2\varepsilon_{-j_1}^{t-r}\varepsilon_{-j_2}^{t-r}K_1$ for some random variable K_1 independent of $\varepsilon_{-j_1}^{t-r}$. Under the event $\tilde{\mathcal{D}}^{0,t-1}$, we can easily show that $|K_1| \le R$ almost surely. Let $\mathcal{F}_K = \sigma(K_1, \varepsilon_{-j_2}^{t-r})$. Let $\mathcal{M} := \{|K_1| \le R\}$. Clearly, $\tilde{\mathcal{D}}^{0,t-1} \subseteq \mathcal{M}$ and $\varepsilon_{-j_1}^{t-r} \perp \mathcal{F}_K$. We conclude:

$$\begin{aligned}\left|\mathbb{E}\left[\mathrm{Cr}'\,1\left[\tilde{\mathcal{D}}^{0,t-1}\right]\right]\right| &= \left|\mathbb{E}\left[\mathrm{Cr}'\,1\left[\tilde{\mathcal{D}}^{0,t-1}\right]1\left[\mathcal{M}\right]\right]\right| \\ &= 4\gamma^2\left|\mathbb{E}\left[\mathbb{E}\left[\varepsilon_{-j_1}^{t-r}1\left[\tilde{\mathcal{D}}^{0,t-1}\right]\middle|\mathcal{F}_K\right]K_1\varepsilon_{-j_2}^{t-r}1\left[\mathcal{M}\right]\right]\right| \\ &\le 4\gamma^2\mathbb{E}\left[\left|\mathbb{E}\left[\varepsilon_{-j_1}^{t-r}1\left[\tilde{\mathcal{D}}^{0,t-1}\right]\middle|\mathcal{F}_K\right]\right|\cdot|K_1|\cdot|\varepsilon_{-j_2}^{t-r}|1\left[\mathcal{M}\right]\right] \qquad (6.67)\end{aligned}$$

We note that: $\left|\mathbb{E}\left[\varepsilon_{-j_1}^{t-r}1\left[\tilde{\mathcal{D}}^{0,t-1}\right]|\mathcal{F}_K\right]\right| = \left|\mathbb{E}\left[\varepsilon_{-j_1}^{t-r}1\left[\tilde{\mathcal{D}}^{0,t-1,C}\right]|\mathcal{F}_K\right]\right| \le \sigma^2\sqrt{\mathbb{P}\left(1\left[\tilde{\mathcal{D}}^{0,t-1,C}\right]|\mathcal{F}_K\right)}$. Using this in Equation (6.67), and that under event $\mathcal{M}$, $|K_1| \le R$ we apply Cauchy-Schwarz inequality again to conclude:

$$\left|\mathbb{E}\left[\mathrm{Cr}'\,1\left[\tilde{\mathcal{D}}^{0,t-1}\right]\right]\right| \le 4\gamma^2\sigma^2 R\sqrt{\mathbb{P}(\tilde{\mathcal{D}}^{0,t-1,C})} \le \frac{4\gamma^2\sigma^2 R}{T^{\alpha/2}} \qquad (6.68)$$

Using similar technique as lemma 6.8.7, we have that on the event $\mathcal{E}_{0,j_1}^r \cap \tilde{\mathcal{D}}^{0,t-1} \cap \mathcal{A}^{t-1}$,

$$\begin{aligned}&\left\|\tilde{H}_{j_2+1,B-1}^{t-r}\left(\prod_{s=r-1}^{1}\tilde{H}_{0,B-1}^{t-s}\right) - \mathcal{R}_{-j_1}^{t-r}\tilde{H}_{j_2+1,B-1}^{t-r}\left(\prod_{s=r-1}^{1}\mathcal{R}_{-j_2}^{t-r}\tilde{H}_{0,B-1}^{t-s}\right)\right\| \\ &\le \bar{C}\frac{T\,\|\phi''\|\,\gamma^2R^3B\sqrt{\beta}}{\zeta\lambda_{\min}} \qquad (6.69)\end{aligned}$$

Therefore, on the event $\mathcal{E}^r_{0,j_1} \cap \mathcal{E}^r_{0,j_2} \cap \tilde{\mathcal{D}}^{0,t-1} \cap \mathcal{A}^{t-1}$, we have

$$\begin{aligned}
&\left|\mathrm{Cr} - \mathrm{Cr}'\right| \leq \gamma^4 \bar{C} R^4 \left|\varepsilon^{t-r}_{-j_1} \varepsilon^{t-r}_{-j_2}\right| T \left\|\phi''\right\| B \frac{\sqrt{\beta}}{\zeta \lambda_{\min}} \\
&\implies \mathbb{E}\left[\left|\mathrm{Cr} - \mathrm{Cr}'\right| 1\left[\mathcal{E}^r_{0,j_1} \cap \mathcal{E}^r_{0,j_2} \cap \tilde{\mathcal{D}}^{0,t-1} \cap \mathcal{A}^{t-1}\right]\right] \leq \bar{C} \left\|\phi''\right\| \gamma^4 T R^4 B \frac{\sigma^2 \sqrt{\beta}}{\zeta \lambda_{\min}} \qquad (6.70)
\end{aligned}$$

We note that over the event $\tilde{\mathcal{D}}^{0,t-1}$, we must have $|\,\mathrm{Cr} - \mathrm{Cr}'\,| \leq 2R|\varepsilon^{t-r}_{-j_1} \varepsilon^{t-r}_{-j_2}|$. Combining this with Equation (6.70) and noting that $\mathbb{P}\left(\mathcal{E}^r_{0,j_1} \cap \mathcal{E}^r_{0,j_2} \cap \tilde{\mathcal{D}}^{0,t-1} \cap \mathcal{A}^{t-1}\right) \geq 1 - \frac{1}{T^\alpha}$, we conclude:

$$\begin{aligned}
&\mathbb{E}\left[\left|\sum_{r=1}^{t} \sum_{j_1 \neq j_2} \mathrm{Cr}(t, r, r, j_1, j_2) - \mathrm{Cr}'(t, r, r, j_1, j_2)\right| 1\left[\tilde{\mathcal{D}}^{0,t-1}\right]\right] \\
&\leq \bar{C}\left[\left\|\phi''\right\| \gamma^4 T^2 R^4 B^2 \frac{\sigma^2 \sqrt{\beta}}{\zeta \lambda_{\min}} + \sigma^2 \gamma^2 R T B \frac{1}{T^{\alpha/2}}\right] \qquad (6.71)
\end{aligned}$$

Hence combining (6.68) and (6.71) we conclude the statement of the claim. □

Next we want to bound $\mathrm{Cr}(t, r_1, r_2, j_1, j_2)$ for $r_2 > r_1$ and arbitrary j_1 and j_2. Recall the definition of $\tilde{a}^{t-1,v}_B$ from (6.41). Via simple rearrangement of summation, we can express $\sum_{r_2 > r_1} \sum_{j_1, j_2} \mathrm{Cr}(t, r_1, r_2, j_1, j_2)$ in terms of $\tilde{a}^{t-r_1-1,v}_B$ as follows.

Lemma 6.8.8.

$$\begin{aligned}
&\sum_{r_2 > r_1} \sum_{j_1, j_2} \mathrm{Cr}(t, r_1, r_2, j_1, j_2) \\
&= 2\gamma \sum_{r_1=1}^{t-1} \sum_{j_1=0}^{B-1} (\tilde{a}^{t-r_1-1,v}_B)^\top \left(\prod_{s=r_1}^{1} \tilde{H}^{t-s}_{0,B-1}\right) \left(\prod_{s=1}^{r_1-1} \tilde{H}^{t-s,\top}_{0,B-1}\right) \tilde{H}^{t-r_1,\top}_{j_1+1,B-1} \tilde{X}^{t-r_1}_{-j_1} \varepsilon^{t-r_1}_{-j_1} \qquad (6.72)
\end{aligned}$$

Claim 14.

$$\begin{aligned}
&\left|\mathbb{E}\left[\sum_{r_1 \neq r_2} \sum_{j_1, j_2} \mathrm{Cr}(t, r_1, r_2, j_1, j_2) 1\left[\tilde{\mathcal{D}}^{0,t-1}\right]\right]\right| \leq \bar{C} \gamma^2 R (Bt)^2 \sigma^2 \frac{1}{T^{\alpha/2}} + \\
&\bar{C}\left(\left\|\phi''\right\| \gamma^3 T^2 R^3 B \frac{\sqrt{\beta}}{\zeta \lambda_{\min}} + \gamma^2 T R B\right) \sqrt{R\sigma^2} \sqrt{\sup_{s \leq N-1} \mathbb{E}\left[\left\|\tilde{a}^{s,v}_B\right\|^2 1\left[\tilde{\mathcal{D}}^{0,s}\right]\right]}
\end{aligned} \qquad (6.73)$$

The proof of the claim essentially proceeds similar to that of Claim 13 but with additional complications. We refer to Section 6.9 for the proof.

Combining everything in this section we have the following proposition.

Proposition 4. *Let*

$$\tilde{\mathcal{V}}_{t-1} = \mathbb{E}\left[\left\|\tilde{a}^{t-1,v}_B\right\|^2 1\left[\tilde{\mathcal{D}}^{0,t-1}\right]\right] \qquad (6.74)$$

Then for some constant $\bar{C}$ which depends only on C_η:

$$\sup_{s\leq N-1}\tilde{\mathcal{V}}_s \leq \frac{2\gamma d}{\zeta(1-\gamma R)}\beta + \bar{C}\left\|\phi''\right\|\gamma^4T^2R^4B^2\frac{\sigma^2\sqrt{\beta}}{\zeta\lambda_{\min}} + \bar{C}\sigma^2\gamma^2RT^2\frac{1}{T^{\alpha/2}} + \bar{C}R\sigma^2\left(\left\|\phi''\right\|^2\gamma^6T^4R^6B^2\frac{\beta}{\zeta^2\lambda_{\min}^2} + \gamma^4T^2R^2B^2\right) \tag{6.75}$$

Proof. In the whole proof, we will denote any large enough constant depending on C_η by $\bar{C}$. From claims 12, 13 and 14 along with equation (6.45) we have

$$\begin{aligned}&\mathbb{E}\left[\left\|\tilde{a}_B^{t-1,v}\right\|^2 1\left[\tilde{\mathcal{D}}^{0,t-1}\right]\right]\\ &\leq \frac{\gamma d}{\zeta(1-\gamma R)}\beta + \bar{C}\left\|\phi''\right\|\gamma^4T^2R^4B^2\frac{\sigma^2\sqrt{\beta}}{\zeta\lambda_{\min}} + \bar{C}\sigma^2\gamma^2RT\frac{1}{T^{\alpha/2}}(1+B+T) +\\ &\left(\bar{C}\left\|\phi''\right\|\gamma^3T^2R^3B\frac{\sqrt{\beta}}{\zeta\lambda_{\min}} + \gamma^2TRB\right)\sqrt{R\sigma^2}\sqrt{\sup_{s\leq N-1}\mathbb{E}\left[\left\|\tilde{a}_B^{s,v}\right\|^2 1\left[\tilde{\mathcal{D}}^{0,s}\right]\right]}\end{aligned} \tag{6.76}$$

Thus

$$\sup_{s\leq N-1}\tilde{\mathcal{V}}_s \leq \frac{\gamma d}{\zeta(1-\gamma R)}\beta + \bar{C}\left\|\phi''\right\|\gamma^4T^2R^4B^2\frac{\sigma^2\sqrt{\beta}}{\zeta\lambda_{\min}} + \bar{C}\sigma^2\gamma^2RT^2\frac{1}{T^{\alpha/2}} + \bar{C}\left(\left\|\phi''\right\|\gamma^3T^2R^3B\frac{\sqrt{\beta}}{\zeta\lambda_{\min}} + \gamma^2TRB\right)\sqrt{R\sigma^2}\sqrt{\sup_{s\leq N-1}\tilde{\mathcal{V}}_s} \tag{6.77}$$

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

$$x^2 \leq c_1 + c_2x \tag{6.78}$$

then

$$x^2 \leq \frac{1}{4}\left(c_2 + \sqrt{c_2^2+4c_1}\right)^2 \leq c_2^2 + 2c_1 \tag{6.79}$$

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

Thus,

$$\sup_{s\leq N-1}\tilde{\mathcal{V}}_s \leq \frac{2\gamma d}{\zeta(1-\gamma R)}\beta + \bar{C}\left\|\phi''\right\|\gamma^4T^2R^4B^2\frac{\sigma^2\sqrt{\beta}}{\zeta\lambda_{\min}} + \bar{C}\sigma^2\gamma^2RT^2\frac{1}{T^{\alpha/2}} + \bar{C}R\sigma^2\left(\left\|\phi''\right\|^2\gamma^6T^4R^6B^2\frac{\beta}{\zeta^2\lambda_{\min}^2} + \gamma^4T^2R^2B^2\right) \tag{6.80}$$

□

6.8.9 Bias of last iterate

In this part, we will bound the expectation of the bias term $\left\|\tilde{a}_B^{t-1,b} - a^*\right\|^2$.

Theorem 6.8.9. *For some universal constant c_0:*

$$\mathbb{E}\left[\left\|\tilde{a}_B^{t-1,b} - a^*\right\|^2 1\left[\tilde{\mathcal{D}}^{0,t-1}\right]\right] \leq \|a_0 - a^*\|^2 \left(1 - c_0\zeta\gamma B\lambda_{\min}\right)^t \tag{6.81}$$

where $\tilde{a}^{t-1,b} - a^*$ is defined in (6.40)

Proof. Define $x_v = \prod_{s=0}^{v-1} \tilde{H}_{0,B-1}^{s,\top}(a_0 - a^*)$. Here, we consider the event $\mathcal{G}_v$ considered in Claim 17 in the proof of Theorem 6.9.4, and show that for some universal constant $q_0 > 0$,

$$\mathbb{P}(\|\tilde{H}_{0,B-1}^{v,\top} x_v\|^2 \geq (1 - \zeta\gamma\lambda_{\min}(G))\|x_v\|^2 \big| \tilde{\mathcal{D}}^{0,t-1}, x_v) \geq q_0 \tag{6.82}$$

From Theorem 6.9.4 we also note that conditioned on $\tilde{\mathcal{D}}^{0,t-1}$, almost surely:

$$\|\tilde{H}_{0,B-1}^{v,\top} x_v\|^2 \leq 1$$

We let $\mathcal{G}_v$ be the event lower bounded in Equation (6.82).

$$\begin{aligned}
\mathbb{E}\left[\|x_{v+1}\|^2 \big| \tilde{\mathcal{D}}^{0,t-1}\right] &= \mathbb{E}\left[\|\tilde{H}_{0,B-1}^{v,\top} x_v\|^2 \big| \tilde{\mathcal{D}}^{0,t-1}\right] \\
&= \mathbb{E}\left[\|\tilde{H}_{0,B-1}^{v,\top} x_v\|^2 1\left[\mathcal{G}_v\right] + \|\tilde{H}_{0,B-1}^{v,\top} x_v\|^2 1\left[\mathcal{G}_v^C\right] \big| \tilde{\mathcal{D}}^{0,t-1}\right] \\
&\leq \mathbb{E}\left[(1 - \gamma\zeta\lambda_{\min}(G))\|x_v\|^2 1\left[\mathcal{G}_v\right] + \|x_v\|^2 1\left[\mathcal{G}_v^C\right] \big| \tilde{\mathcal{D}}^{0,t-1}\right] \\
&= \mathbb{E}\left[\|x_v\|^2 \left[1 - \gamma\zeta\lambda_{\min}(G)\,\mathbb{P}(\mathcal{G}_v \big| \tilde{\mathcal{D}}^{0,t-1}\, x_v)\right] \big| \tilde{\mathcal{D}}^{0,t-1}\right] \\
&\leq \mathbb{E}\left[\|x_v\|^2 \left[1 - \gamma\zeta\lambda_{\min}(G)\, q_0\right] \big| \tilde{\mathcal{D}}^{0,t-1}\right]
\end{aligned} \tag{6.83}$$

Unrolling the recursion given by Equation (6.83), and noting that $\tilde{a}_B^{t-1,b} - a^* = x_t$, we conclude

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

Hence we have the theorem.

□

6.8.10 Average iterate: Bias-variance decomposition

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

$$\hat{\tilde{a}}_{t_0,N} = \frac{1}{N-t_0}\sum_{t=t_0+1}^{N} \tilde{a}_B^{t-1} \tag{6.84}$$

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

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

$$\hat{\tilde{a}}_{t_0,N} - a^* = (\hat{\tilde{a}}^v_{t_0,N}) + (\hat{\tilde{a}}^b_{t_0,N} - a^*) \tag{6.85}$$

where

$$\hat{\tilde{a}}^v_{t_0,N} = \frac{1}{N-t_0}\sum_{t=t_0+1}^{N} (\tilde{a}_B^{t-1,v}) \tag{6.86}$$

$$\hat{\tilde{a}}^b_{t_0,N} - a^* = \frac{1}{N-t_0}\sum_{t=t_0+1}^{N} (\tilde{a}_B^{t-1,b} - a^*) \tag{6.87}$$

6.8.11 Variance of average iterate

Remark 16. *From now on we will use the following notation:*

$$\sum_{t} \equiv \sum_{t=t_0+1}^{N}$$

$$\sum_{t_1,t_2} \equiv \sum_{t_1,t_2=t_0+1}^{N}$$

$$\sum_{t_1\neq t_2} \equiv \sum_{\substack{t_1,t_2=t_0+1 \\ t_1\neq t_2}}^{N}$$

$$\sum_{t_2>t_1} \equiv \sum_{t_1=t_0+1}^{N-1}\sum_{t_2=t_1+1}^{N}$$

Next we expand $\left\|\hat{\tilde{a}}^v_{t_0,N}\right\|^2$

$$\left\|\hat{\tilde{a}}^v_{t_0,N}\right\|^2 = \frac{1}{(N-t_0)^2}\sum_{t}\left\|\tilde{a}_B^{t-1,v}\right\|^2 + \frac{1}{(N-t_0)^2}\sum_{t_1\neq t_2}(\tilde{a}_B^{t_2-1,v})^\top(\tilde{a}_B^{t_1-1,v}) \tag{6.88}$$

Claim 15. *For $t_2 > t_1$*

$$\left|\mathbb{E}\left[\left[(\tilde{a}_B^{t_2-1,v})^\top\left((\tilde{a}_B^{t_1-1,v}) - \left(\prod_{s=t_2-t_1}^{1}\tilde{H}_{0,B-1}^{t_2-s}\right)(\tilde{a}_B^{t_1-1,v})\right)\right]1\left[\tilde{\mathcal{D}}^{0,N-1}\right]\right]\right|$$
$$\leq C_1\,\mathrm{Poly}(R,B,\beta,1/\zeta,1/\lambda_{\min},\|\phi''\|)\left(\gamma^{7/2}T^2+\gamma^5T^3+\gamma^6T^4\right) \tag{6.89}$$

Proof. From (6.41) we can write

$$(\tilde{a}_B^{t_2-1,v})^\top = (\tilde{a}_B^{t_1-1,v})^\top\left(\prod_{s=t_2-t_1}^{1}\tilde{H}_{0,B-1}^{t_2-s}\right) + 2\gamma\sum_{r=1}^{t_2-t_1}\sum_{j=0}^{B-1}\varepsilon_{-j}^{t_2-r}\tilde{X}_{-j}^{t_2-r,\top}\tilde{H}_{j+1,B-1}^{t_2-r}\left(\prod_{s=r-1}^{1}\tilde{H}_{0,B-1}^{t_2-s}\right) \tag{6.90}$$

Hence

$$(\tilde{a}_B^{t_2-1,v})^\top(\tilde{a}_B^{t_1-1,v}) = (\tilde{a}_B^{t_1-1,v})^\top\left(\prod_{s=t_2-t_1}^{1}\tilde{H}_{0,B-1}^{t_2-s}\right)(\tilde{a}_B^{t_1-1,v}) +$$
$$2\gamma\sum_{r=1}^{t_2-t_1}\sum_{j=0}^{B-1}\varepsilon_{-j}^{t_2-r}\tilde{X}_{-j}^{t_2-r,\top}\tilde{H}_{j+1,B-1}^{t_2-r}\left(\prod_{s=r-1}^{1}\tilde{H}_{0,B-1}^{t_2-s}\right)(\tilde{a}_B^{t_1-1,v}) \tag{6.91}$$

Now recall the noise re-sampling operator $\mathcal{R}_{-j}^{t_2-r}$ from (6.57). It is easy to see that

$$\mathbb{E}\left[2\gamma\sum_{r=1}^{t_2-t_1}\sum_{j=0}^{B-1}\varepsilon_{-j}^{t_2-r}\mathcal{R}_{-j}^{t_2-r}\left[\tilde{X}_{-j}^{t_2-r,\top}\tilde{H}_{j+1,B-1}^{t_2-r}\left(\prod_{s=r-1}^{1}\tilde{H}_{0,B-1}^{t_2-s}\right)(\tilde{a}_B^{t_1-1,v})\right]\right] = 0 \tag{6.92}$$

(Note that $\mathcal{R}_{-j}^{t_2-r}\tilde{X}_{-j}^{t_2-r,\top} = \tilde{X}_{-j}^{t_2-r,\top}$)

Thus, using the decomposition

$$\tilde{\mathcal{D}}^{0,N-1} = \tilde{\mathcal{D}}^{0,t_2-r-1}\cap\tilde{\mathcal{D}}^{t_2-r+1,N-1}\cap\tilde{\mathcal{D}}_{-j}^{t_2-r}\cap\cap_{i=0}^{j-1}\tilde{\mathcal{C}}_{-i}^{t_2-r}$$

we get

$$\left|\mathbb{E}\left[2\gamma\sum_{r=1}^{t_2-t_1}\sum_{j=0}^{B-1}\varepsilon_{-j}^{t_2-r}\tilde{X}_{-j}^{t_2-r,\top}\cdot\right.\right.$$
$$\left.\left.\mathcal{R}_{-j}^{t_2-r}\left[\tilde{H}_{j+1,B-1}^{t_2-r}\left(\prod_{s=r-1}^{1}\tilde{H}_{0,B-1}^{t_2-s}\right)(\tilde{a}_B^{t_1-1,v})\right]1\left[\tilde{\mathcal{D}}^{0,N-1}\right]\right]\right|$$
$$\leq 4\gamma^2R(Bt_1)(B(t_2-t_1))C_\eta\sigma^2\frac{1}{T^{\alpha/2}}$$
$$\leq 4\gamma^2RC_\eta\sigma^2T^2\frac{1}{T^{\alpha/2}} \tag{6.93}$$

Next we need to bound the effect due to noise re-sampling. On the event $\tilde{\mathcal{D}}^{0,N-1} \cap \mathcal{A}^{N-1} \cap \cap_{r=0}^{N-1} \mathcal{E}^r_{0,B-1}$, we have

$$\begin{aligned}
&\left\| \tilde{H}^{t_2-r}_{j+1,B-1} \left(\prod_{s=r-1}^{1} \tilde{H}^{t_2-s}_{0,B-1} \right) - \mathcal{R}^{t_2-r}_{-j} \left[\tilde{H}^{t_2-r}_{j+1,B-1} \left(\prod_{s=r-1}^{1} \tilde{H}^{t_2-s}_{0,B-1} \right) \right] \right\| \\
&\leq C \left\| \phi'' \right\| \gamma^2 T R^3 B \frac{\sqrt{\beta}}{\zeta \lambda_{\min}} \qquad (6.94)
\end{aligned}$$

Thus

$$\begin{aligned}
&2\gamma \left| \mathbb{E} \left[\sum_{r=1}^{t_2-t_1} \sum_{j=0}^{B-1} \varepsilon^{t_2-r}_{-j} \tilde{X}^{t_2-r,\top}_{-j} 1 \left[\tilde{\mathcal{D}}^{0,N-1} \right] \cdot \right. \right. \\
&\left. \left. \left(\tilde{H}^{t_2-r}_{j+1,B-1} \left(\prod_{s=r-1}^{1} \tilde{H}^{t_2-s}_{0,B-1} \right) - \mathcal{R}^{t_2-r}_{-j} \left[\tilde{H}^{t_2-r}_{j+1,B-1} \left(\prod_{s=r-1}^{1} \tilde{H}^{t_2-s}_{0,B-1} \right) \right] \right) (\tilde{a}^{t_1-1,v}_B) \right] \right| \\
&\leq \bar{C}\gamma \left(\left\| \phi'' \right\| \gamma^2 T R^3 B \frac{\sqrt{\beta}}{\zeta \lambda_{\min}} \right) B(t_2 - t_1) \sqrt{R\sigma^2} \mathbb{E} \left[\left\| (\tilde{a}^{t_1-1,v}_B) \right\| 1 \left[\tilde{\mathcal{D}}^{0,t_1-1} \right] \right] \\
&\quad + \bar{C}\gamma^2 (2R) \sigma^2 (Bt_1)(B(t_2 - t_1)) \frac{1}{T^{\alpha/2}} \qquad (6.95)
\end{aligned}$$

Now from proposition 4, there is a constant C_1 such that

$$\begin{aligned}
&\left(\mathbb{E} \left[\left\| \tilde{a}^{t_1-1,v} \right\| 1 \left[\tilde{\mathcal{D}}^{0,t_1-1} \right] \right] \right)^2 \leq \\
&C_1 \left(\frac{\gamma d}{\zeta(1-\gamma R)} \beta + \left\| \phi'' \right\| \gamma^4 T^2 R^4 B^2 \frac{\sigma^2 \sqrt{\beta}}{\zeta \lambda_{\min}} + \sigma^2 \gamma^2 R T^2 \frac{1}{T^{\alpha/2}} + \right. \\
&\left. R\sigma^2 \left(\left\| \phi'' \right\|^2 \gamma^6 T^4 R^6 B^2 \frac{\beta}{\zeta^2 \lambda^2_{\min}} + \gamma^4 T^2 R^2 B^2 \right) \right) \qquad (6.96)
\end{aligned}$$

So

$$\begin{aligned}
&2\gamma \left| \mathbb{E} \left[\sum_{r=1}^{t_2-t_1} \sum_{j=0}^{B-1} \varepsilon^{t_2-r}_{-j} \tilde{X}^{t_2-r,\top}_{-j} 1 \left[\tilde{\mathcal{D}}^{0,N-1} \right] \cdot \right. \right. \\
&\left. \left. \left(\tilde{H}^{t_2-r}_{j+1,B-1} \left(\prod_{s=r-1}^{1} \tilde{H}^{t_2-s}_{0,B-1} \right) - \mathcal{R}^{t_2-r}_{-j} \left[\tilde{H}^{t_2-r}_{j+1,B-1} \left(\prod_{s=r-1}^{1} \tilde{H}^{t_2-s}_{0,B-1} \right) \right] \right) (\tilde{a}^{t_1-1,v}_B) \right] \right| \\
&\leq \text{Poly}(R, B, \beta, 1/\zeta, 1/\lambda_{\min}, \left\| \phi'' \right\|) \left(\gamma^{7/2} T^2 + \gamma^5 T^3 + \gamma^6 T^4 \right) \qquad (6.97)
\end{aligned}$$

where we absorbed terms involving $\frac{1}{T^\alpha}$ since α is taken to be large.

□

Claim 16.

$$\left|\mathbb{E}\left[(\tilde{a}_B^{t_1-1,v})^\top \sum_{t_2>t_1}\left(\prod_{s=t_2-t_1}^{1}\tilde{H}_{0,B-1}^{t_2-s}\right)(\tilde{a}_B^{t_1-1,v})1\left[\tilde{\mathcal{D}}^{0,N-1}\right]\right]\right|$$
$$\leq \mathcal{V}_{t_1-1}\frac{C}{\zeta\gamma B\lambda_{\min}} + 16(N-t_1)\gamma^2 RC_\eta\sigma^2T^2\frac{1}{T^{\alpha/2}} \tag{6.98}$$

where $\mathcal{V}_{t_1-1}$ is defined in (6.74).

Proof. Note that

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

From theorem 6.9.5, there is a universal constant C such that with $\delta = \frac{1}{2T^\alpha}$ we have

$$\mathbb{P}\left[\left\|\sum_{t_2>t_1}\prod_{s=t_1}^{t_2-1}\tilde{H}_{0,B-1}^{s}\right\| > C\left(d+\log\frac{N}{\delta}+\frac{1}{\zeta\gamma B\lambda_{\min}}\right)|\tilde{\mathcal{D}}^{t_1,N-1}\right] \leq \frac{1}{2T^\alpha} \tag{6.99}$$

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

$$\mathbb{P}\left[\left\|\sum_{t_2>t_1}\prod_{s=t_1}^{t_2-1}\tilde{H}_{0,B-1}^{s}\right\| > C\left(d+\log\frac{N}{\delta}+\frac{1}{\zeta\gamma B\lambda_{\min}}\right)\right] \leq \frac{1}{T^\alpha} \tag{6.100}$$

Choosing γ such that

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

we get

$$\mathbb{P}\left[\left\|\sum_{t_2>t_1}\prod_{s=t_1}^{t_2-1}\tilde{H}_{0,B-1}^{s}\right\| > \frac{C}{\zeta\gamma B\lambda_{\min}}\right] \leq \frac{1}{T^\alpha} \tag{6.101}$$

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

$$\left|\mathbb{E}\left[(\tilde{a}_B^{t_1-1,v})^\top \sum_{t_2>t_1}\left(\prod_{s=t_2-t_1}^{1}\tilde{H}_{0,B-1}^{t_2-s}\right)(\tilde{a}_B^{t_1-1,v})1\left[\tilde{\mathcal{D}}^{0,N-1}\right]\right]\right|$$
$$\leq \mathbb{E}\left[\left\|\tilde{a}_B^{t_1-1,v}\right\|^2 1\left[\tilde{\mathcal{D}}^{0,t-1}\right]\right]\frac{C}{\zeta\gamma B\lambda_{\min}} + (N-t_1)4\gamma^2R(4C_\eta\sigma^2)(Bt_1)^2\frac{1}{T^{\alpha/2}}$$
$$\leq \mathcal{V}_{t_1-1}\frac{C}{\zeta\gamma B\lambda_{\min}} + 16(N-t_1)\gamma^2RC_\eta\sigma^2T^2\frac{1}{T^{\alpha/2}} \tag{6.102}$$

□

Thus combining everything we have the following theorem

Theorem 6.8.10. *Suppose $\gamma \gtrsim \frac{1}{T}$. Then*

$$\mathbb{E}\left[\left\|\hat{\tilde{a}}^{v}_{t_0,N}\right\|^2 1\left[\tilde{\mathcal{D}}^{0,N-1}\right]\right] \leq C_1 \frac{d\beta}{\zeta^2 \lambda_{\min} B(N-t_0)} + \bar{P}\cdot\left(\gamma^{7/2}T^2 + \gamma^6 T^4\right) \tag{6.103}$$

$\bar{P} = \mathrm{Poly}(R, B, \beta, 1/\zeta, 1/\lambda_{\min}, \|\phi''\|, C_\eta)$ *and $C_1 > 0$ is some constant.*

6.8.12 Bias of average iterate

Theorem 6.8.11. *There are constants C, c_1, c_2 such that :*

$$\mathbb{E}\left[\left\|\hat{\tilde{a}}^{b}_{t_0,N} - a^*\right\|^2 1\left[\tilde{\mathcal{D}}^{0,N-1}\right]\right] \leq \quad C\left\|a_0 - a\right\|^2 \left[e^{-c_2\zeta\gamma B\lambda_{\min} t_0} \min\left\{1, \frac{1}{(N-t_0)\zeta\gamma B\lambda_{\min}}\right\}\right] \tag{6.104}$$

The proof follows from an application of Theorem 6.8.9.

6.8.13 Proof of theorem 6.3.1

Proof. Let $\bar{P}$ denote the polynomial in Theorem 6.8.10. Theorems 6.8.11 and 6.8.10, imply for every row of the coupled iterate:

$$\begin{aligned}
\mathbb{E}\left[\|\hat{\tilde{a}}_{t_0,N} - a^*\|^2 1\left[\tilde{\mathcal{D}}^{0,N-1}\right]\right] &\leq 2\mathbb{E}\left[\|\hat{\tilde{a}}^{v}_{t_0,N}\|^2 1\left[\tilde{\mathcal{D}}^{0,N-1}\right]\right] + 2\mathbb{E}\left[\|\hat{\tilde{a}}^{b}_{t_0,N} - a^*\|^2 1\left[\tilde{\mathcal{D}}^{0,N-1}\right]\right] \\
&\leq C_2 \frac{d\beta}{\zeta^2 \lambda_{\min}(G) B(N-t_0)} + \bar{P}\cdot\left(\gamma^{7/2}T^2 + \gamma^6 T^4\right) \\
&\quad + C\left\|a_0 - a\right\|^2 \left[e^{-c_2\zeta\gamma B\lambda_{\min} t_0} \min\left\{1, \frac{1}{(N-t_0)\zeta\gamma B\lambda_{\min}}\right\}\right]
\end{aligned} \tag{6.105}$$

Thus for the actual process we can use the following decomposition

$$\begin{aligned}
\mathbb{E}\left[\|\hat{a}_{t_0,N} - a^*\|^2 1\left[\mathcal{D}^{0,N-1}\right]\right] &\leq \mathbb{E}\left[\|\hat{a}_{t_0,N} - a^*\|^2 1\left[\hat{\mathcal{D}}^{0,N-1}\right]\right] + \\
&\quad \mathbb{E}\left[\|\hat{a}_{t_0,N} - a^*\|^2 1\left[\mathcal{D}^{0,N-1}\right] 1\left[\tilde{\mathcal{D}}^{0,N-1,C}\right]\right] \\
&\leq \mathbb{E}\left[\|\hat{a}_{t_0,N} - a^*\|^2 1\left[\hat{\mathcal{D}}^{0,N-1}\right]\right] + C\gamma^2 T^2 R C_\eta \sigma^2 \frac{1}{T^{\alpha/2}} \\
&\quad + 2\left\|a_0 - a^*\right\|^2 \frac{1}{T^\alpha}
\end{aligned}$$

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

$$\left\|\hat{a}_{t_0,N} - a^*\right\|^2 \leq \frac{1}{N-t_0} \sum_{t=t_0+1}^{N} \left(2\left\|a_0 - a^*\right\|^2 + 2(Bt)(4\gamma^2 R \sum_{r=1}^{t}\sum_{j=0}^{B-1} |\varepsilon_{-j}^{t-r}|^2)\right) \tag{6.106}$$

and then used Cauchy-Schwarz inequality for the expectation over $\tilde{\mathcal{D}}^{0,N-1,C}$.

Now using lemma 6.8.3 we get

$$\mathbb{E}\left[\|\hat{a}_{t_0,N} - a^*\|^2 1\left[\hat{\mathcal{D}}^{0,N-1}\right]\right] \leq \mathbb{E}\left[\left\|\hat{\tilde{a}}_{t_0,N} - a^*\right\|^2 1\left[\hat{\mathcal{D}}^{0,N-1}\right]\right] + C\gamma^2 R^2 T^2 \frac{1}{T^\alpha} \tag{6.107}$$

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

$$\mathbb{E}\left[\|\hat{a}_{t_0,N} - a^*\|^2 1\left[\mathcal{D}^{0,N-1}\right]\right] \leq \mathbb{E}\left[\left\|\hat{\tilde{a}}_{t_0,N} - a^*\right\|^2 1\left[\tilde{\mathcal{D}}^{0,N-1}\right]\right] + C\gamma^2 R\sigma^2 \frac{1}{T^{\alpha/2-2}} \tag{6.108}$$

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

$$\begin{aligned}\mathbb{E}\left[\|\hat{\tilde{a}}_{t_0,N} - a^*\|^2 1\left[\tilde{\mathcal{D}}^{0,N-1}\right]\right] \leq\ & C_1 \frac{d\beta}{\zeta^2 \lambda_{\min}(G)\, B(N-t_0)} + \bar{P} \cdot \left(\gamma^{7/2} T^2 + \gamma^6 T^4\right) \\ & + C_2 \|a_0 - a\|^2 \left[\frac{e^{-c_2 \zeta \gamma B \lambda_{\min} t_0}}{(N-t_0)\zeta\gamma B \lambda_{\min}}\right] \\ & + C_3 \gamma^2 R \sigma^2 \frac{1}{T^{\alpha/2-2}}\end{aligned} \tag{6.109}$$

Summing over all the rows we get a bound on the Frobenius norm. Lastly if the event $\mathcal{D}^{0,N-1}$ does not occur, the $\hat{A}_{t_0,N}$ is the zero matrix and hence

$$\mathbb{E}\left[\|\hat{A}_{t_0,N} - A^*\|_{\mathsf{F}}^2 1\left[\mathcal{D}^{0,N-1,C}\right]\right] \leq \|A^*\|^2 \frac{1}{T^\alpha}. \tag{6.110}$$

Therefore:

$$\begin{aligned}\mathbb{E}\left[\|\hat{A}_{t_0,N} - A^*\|_{\mathsf{F}}^2\right] \leq\ & \bar{C} \frac{d^2\beta}{\zeta^2 \lambda_{\min}(G)\, B(N-t_0)} + \bar{P} \cdot d \left(\gamma^{7/2} T^2 + \gamma^6 T^4\right) \\ & + \bar{C} \|A_0 - A^*\|_{\mathsf{F}}^2 \left[\frac{e^{-c_2 \zeta \gamma B \lambda_{\min} t_0}}{(N-t_0)\zeta\gamma B \lambda_{\min}}\right] \\ & + \bar{C} \gamma^2 R \sigma^2 d \frac{1}{T^{\alpha/2-2}}\end{aligned} \tag{6.111}$$

□

6.9 Technical results

6.9.1 Well conditioned second moment matrices

In this section we will consider a stationary sequence $X_0, \ldots, X_T$ derived from the process $\mathsf{NLDS}(A^*, \mu, \phi)$, with the corresponding noise sequence $\eta_0, \ldots, \eta_T$. We want to show that the matrix $\frac{1}{B}\sum_{t=0}^{B-1} X_t X_t^\top$

behaves similar to $G := \mathbb{E}X_t X_t^\top$. To do this, we will first to control the quantity: $\mathbb{E}\langle X_t, x\rangle^2\langle X_s, x\rangle^2$ for arbitrary fixed vector $x \in \mathbb{R}^d$. Clearly, $\mathbb{E}\langle X_t, x\rangle^2 = x^\top G x$.

Lemma 6.9.1. *Without loss of generality, we suppose that $t > s$. Suppose $X_0, \ldots, X_T$ be a stationary sequence from $\mathsf{NLDS}(A^*, \mu, \phi)$. Suppose Assumptions 6 and 8 hold. Then we have:*

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

Where $\bar{C}_1$ depends only on C_η.

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:

$$\begin{aligned}\langle X_t, x\rangle^2\langle X_s, x\rangle^2 &= \langle X_t - \tilde{X}_t + \tilde{X}_t, x\rangle^2\langle X_s, x\rangle^2 \\ &\leq 2\langle \tilde{X}_t, x\rangle^2\langle X_s, x\rangle^2 + 2\langle X_t - \tilde{X}_t, x\rangle^2\langle X_s, x\rangle^2\end{aligned}$$

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

$$\mathbb{E}\langle X_t, x\rangle^2\langle X_s, x\rangle^2 \leq 2(x^\top G x)^2 + 2\mathbb{E}\langle X_t - \tilde{X}_t, x\rangle^2\langle X_s, x\rangle^2 \tag{6.112}$$

By Assumption 7, 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 (6.112), we conclude:

$$\begin{aligned}\mathbb{E}\langle X_t, x\rangle^2\langle X_s, x\rangle^2 &\leq 2(x^\top G x)^2 + 2\mathbb{E}\|x\|^2 C_\rho^2\rho^{2(t-s)}\|X_s - \tilde{X}_s\|^2\langle X_s, x\rangle^2 \\ &\leq 2(x^\top G x)^2 + 4\|x\|^2 C_\rho^2\rho^{2(t-s)}\mathbb{E}\left(\|X_s\|^2 + \|\tilde{X}_s\|^2\right)\langle X_s, x\rangle^2 \\ &= 2(x^\top G x)^2 + 4\|x\|^2 C_\rho^2\rho^{2(t-s)}\left[\mathbb{E}\|X_s\|^2\langle X_s, x\rangle^2 + x^\top G x\mathbb{E}\|\tilde{X}_s\|^2\right] \end{aligned}\tag{6.113}$$

We can evaluate $\mathbb{E}\|\tilde{X}_s\|^2$ from Theorem 6.6.4. Consider the Sub-Gaussian setting with $\|A^*\| = \rho < 1$ and $C_\rho = 1$. Fix $R > 0$. We will use the notation from Theorem 6.6.4 below. We can then write,

$$\begin{aligned}\mathbb{E}\|X_s\|^2\langle X_s, x\rangle^2 &= \mathbb{E}\|X_s\|^2\langle X_s, x\rangle^2\mathbb{1}(\|X_s\|^2 \leq R) + \mathbb{E}\|X_s\|^2\langle X_s, x\rangle^2\mathbb{1}(\|X_s\|^2 > R) \\ &\leq \mathbb{E}R\langle X_s, x\rangle^2\mathbb{1}(\|X_s\|^2 \leq R) + \mathbb{E}\|X_s\|^4\mathbb{1}(\|X_s\|^2 > R) \\ &\leq Rx^\top G x + \mathbb{E}\|X_s\|^4\mathbb{1}(\|X_s\|^2 > R) \\ &\leq Rx^\top G x + \sqrt{\mathbb{E}\|X_s\|^8}\sqrt{\mathbb{P}(\|X_s\|^2 > R)}\end{aligned}\tag{6.114}$$

From Theorem 6.6.4 and Proposition 2.7.1 in [186], we show that $\mathbb{E}\|X_s\|^8 \leq C\left(\frac{dC_\eta\sigma^2}{1-\rho}\right)^4$ for some universal constant C. Again, taking $R = \frac{8dC_\eta\sigma^2}{1-\rho} + \frac{2\log\frac{1}{\delta}}{\lambda^*} \leq \frac{24dC_\eta\sigma^2\log(\frac{1}{\delta})}{1-\rho}$, we have: $\sqrt{\mathbb{P}(\|X_s\|^2 > R)} \leq$

δ. We plug this into Equation (6.114), take $\delta = \frac{(1-\rho)x^\top Gx}{d\sigma^2}$ after noting that $x^\top Gx \geq \sigma^2\|x\|^2$ to show that:

$$\mathbb{E}\|X_s\|^2\langle X_s, x\rangle^2 \leq \bar{C}\frac{d\sigma^2}{1-\rho}x^\top Gx \log\left(\frac{d}{1-\rho}\right) \tag{6.115}$$

Where $\bar{C}$ is a constant which depends only on C_η. Using Equation (6.115) in Equation (6.113), we conclude that:

$$\mathbb{E}\langle X_t, x\rangle^2\langle X_s, x\rangle^2 \leq 2(x^\top Gx)^2 + \bar{C}_1 C_\rho^2 \rho^{2(t-s)}\frac{d\sigma^2}{1-\rho}x^\top Gx \log\left(\frac{d}{1-\rho}\right) \tag{6.116}$$

□

Now, consider the random matrix $\hat{G}_B := \frac{1}{B}\sum_{t=0}^{B-1} X_t X_t^\top$. Clearly, $\hat{G}_B \succeq 0$ and because of stationarity, $\mathbb{E}\hat{G}_B = G$. We write down the following lemma:

Lemma 6.9.2. *Suppose $X_0, \ldots, X_T$ be a stationary sequence from $\mathsf{NLDS}(A^*, \mu, \phi)$. Suppose Assumptions 6 and 8 hold. Let $\bar{C}_1$ be as in Lemma 6.9.1. 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 $x \in \mathbb{R}^d$,*

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

Where p_0 is a universal constant which can be taken to be $\frac{1}{16}$. Furthermore, for any event $\mathcal{A}$ such that $\mathbb{P}(\mathcal{A}) > 1 - p_0$, we must have:

$$\mathbb{P}\left(x^\top \hat{G}_B x \geq \frac{1}{2}x^\top Gx \middle| \mathcal{A}\right) \geq q_0 := \frac{p_0 - \mathbb{P}(\mathcal{A}^c)}{\mathbb{P}(\mathcal{A})} > 0\,.$$

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 $\theta = \frac{1}{2}$.

The simple calculation shows that:

$$\begin{aligned}
\mathbb{P}\left(x^\top \hat{G}_B x \geq \frac{1}{2} x^\top G x\right) &\geq \frac{1}{4} \frac{B^2 (x^\top G x)^2}{\sum_{s,t=0}^{B-1} \mathbb{E}\langle X_t, x\rangle^2 \langle X_s, x\rangle^2} \\
&\geq \frac{1}{4} \frac{B^2 (x^\top G x)^2}{2B^2 (x^\top G x)^2 + \sum_{s,t=0}^{B-1} \bar{C}_1 \rho^{2|t-s|} \frac{d\sigma^2}{1-\rho} x^\top G x \log\left(\frac{d}{1-\rho}\right)} \\
&\geq \frac{1}{4} \frac{B^2 (x^\top G x)^2}{2B^2 (x^\top G x)^2 + 2\sum_{t=0}^{B-1} \bar{C}_1 \frac{d\sigma^2}{(1-\rho)(1-\rho^2)} x^\top G x \log\left(\frac{d}{1-\rho}\right)} \\
&= \frac{1}{4} \frac{B^2 (x^\top G x)^2}{2B^2 (x^\top G x)^2 + 2B\bar{C}_1 \frac{d\sigma^2}{(1-\rho)(1-\rho^2)} x^\top G x \log\left(\frac{d}{1-\rho}\right)} \\
&= \frac{1}{8} \frac{1}{1+\tau_B} \qquad (6.117)
\end{aligned}$$

Here, $\tau_B := \frac{\bar{C}_1}{x^\top G x} \frac{d\sigma^2}{B(1-\rho)(1-\rho^2)} \log\left(\frac{d}{1-\rho}\right)$. In the second step we have used item 1 of Lemma 6.9.1. In the third step, we have summed the infinite series $\sum_{s\geq t} \rho^{2(t-s)}$. 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. □

We will now follow the method used to prove [110, Lemma 31]. We now consider the matrix $\tilde{H}^s_{0,B-1}$ under the event $\tilde{\mathcal{D}}^s_{-0}$ in order to prove Theorem 6.9.4, where the terms are as defined in Section 6.8.3. For the sake of clarity, we will drop the superscript s.

Remark 17. *We prove the results below for $\tilde{H}^s_{0,B-1}$ but they hold unchanged when the matrices are all replaced with $\tilde{H}^{s,\top}_{0,B-1}$ given that we reverse the order of taking products whenever they are encountered.*

Lemma 6.9.3. *Suppose Assumption 4 holds. Suppose that $\gamma RB < \frac{1}{4}$. Then, for any buffer s, under the event $\tilde{\mathcal{D}}^s_{-0}$, we have:*

$$I - 4\gamma\left(1 + \tfrac{2\gamma BR}{1-4\gamma BR}\right)\sum_{i=0}^{B-1} \tilde{X}^s_{-i}\tilde{X}^{s,\top}_{-i} \preceq \tilde{H}^s_{0,B-1}\tilde{H}^{s,\top}_{0,B-1} \preceq I - 4\gamma\left(\zeta - \tfrac{2\gamma BR}{1-4\gamma BR}\right)\sum_{i=0}^{B-1} \tilde{X}^s_{-i}\tilde{X}^{s,\top}_{-i}$$

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

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

Proof. The proof follows from the proof of [110, Lemma 28] with minor modifications to account for the fact that $\phi'(\beta) \in [\zeta, 1]$. □

Combining Lemma 6.9.3 with Lemma 6.9.1 we will show that $\tilde{H}^s_{0,B-1}$ contracts any given vector with probability at-least $p_0 > 0$.

Theorem 6.9.4. *Suppose Assumptions 4, 6 and 8 hold. Assume that B and γ are such that: $B \geq \bar{C}_1 \frac{d}{(1-\rho)(1-\rho^2)} \log\left(\frac{d}{1-\rho}\right)$ and $\gamma BR \leq \frac{\zeta}{4(1+\zeta)}$ where $\bar{C}_1$ is as given in Lemma 6.9.2. We also assume that $\mathbb{P}(\hat{\mathcal{D}}^{b,a}) > \max(\frac{1}{2}, 1-\frac{p_0}{2})$, where p_0 is as given in Lemma 6.9.2. Let $a \geq b$. Let $\lambda_{\min}(G)$ denote the smallest eigenvalue of G. Conditioned on the event $\tilde{\mathcal{D}}^{b,a}$,*

(1) $\| \prod_{s=a}^{b} \tilde{H}_{0,B-1}^{s,\top} \| \leq 1$ *almost surely*

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

$$\mathbb{P}\left(\| \prod_{s=a}^{b} \tilde{H}_{0,B-1}^{s,\top} \| \geq 2(1-\zeta\gamma B\lambda_{\min}(G))^{c_4(a-b+1)} \middle| \tilde{\mathcal{D}}^{b,a} \right) \leq \exp(-c_3(a-b+1)+c_5 d)$$

Where c_3, c_4 and c_5 are universal constants.

Proof. The proof of (1) above follows from an application of Lemma 6.9.3. So we will just prove (2). We will prove this with an ϵ net argument over the unit ℓ^2 sphere in $\mathbb{R}^d$.

Suppose we have arbitrary $x \in \mathbb{R}^d$ such that $\|x\| = 1$. Let $K_v := \prod_{s=v}^{b} \tilde{H}_{0,B-1}^{s,\top}$. When $v \leq b$, we take this product to be identity. Now, define $\hat{G}_B^v := \frac{1}{B}\sum_{j=0}^{B-1} X_j^v X_j^{v,\top}$

Consider the class of events indexed by v: $\mathcal{G}_v := \{\|\tilde{H}_{0,B-1}^{v,\top} K_{v-1} x\|^2 \leq \|K_{v-1}x\|^2(1-\gamma\zeta B\lambda_{\min}(G)\}$. From Lemma 6.9.2, we will prove the following claim:

Claim 17. *Whenever $v \in [b,a] \cap \mathbb{Z}$:*

$$\mathbb{P}(\mathcal{G}_v^c | \tilde{\mathcal{D}}^{b,a}, \tilde{H}_{0,B-1}^{s,\top} : s < v) \leq 1 - q_0 \tag{6.118}$$

Where $q_0 > 0$ is as given in Lemma 6.9.2 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 $\tilde{H}_{0,B-1}^{s,\top}$ for $s < v$, x_v is fixed. Using Lemma 6.9.3, we note that:

$$\mathbb{P}(\mathcal{G}_v^c | \tilde{\mathcal{D}}^{b,a}, \tilde{H}_{0,B-1}^{s,\top} : s < v) \leq \mathbb{P}(x_v^\top \hat{G}_B^v x_v < \tfrac{1}{2} x_v^\top G x_v | \tilde{\mathcal{D}}^{b,a}, \tilde{H}_{0,B-1}^{s,\top} : s < v)$$

We note that $\hat{G}_B^v$ is independent of $\tilde{H}_{0,B-1}^{s,\top}$ for $s \leq v$ (eventhough $\tilde{H}_{0,B-1}^{v}$ is not necessarily). Now we also note that $\hat{G}_B^v$ is independent of $\tilde{\mathcal{D}}^s$ for $s \neq v$. Therefore, we can apply Lemma 6.9.2 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 $\cap_{v \in D} \mathcal{G}_v$ holds, we have:

$$\| \prod_{s=a}^{b} \tilde{H}_{0,B-1}^{s,\top} x \| \leq (1-\gamma B\lambda_{\min}(G))^{\frac{r}{2}}. \tag{6.119}$$

Therefore, whenever Equation (6.119) is violated, we must have a set $D^c \subseteq \{b, \dots, a\}$ such that $|D^c| \geq b - a - r$ and the event $\cap_{v \in D^c} \mathcal{G}_v^c$ holds. We will union bound all such events indexed by D^c to obtain an upper bound on the probability that Equation (6.119) is violated. Therefore, using Equation (6.118) along with the union bound, we have:

$$\mathbb{P}\left(\| \prod_{s=a}^{b} \tilde{H}_{0,B-1}^{s,\top} x \| \geq (1 - \gamma B \lambda_{\min}(G))^{\frac{r}{2}} \middle| \tilde{\mathcal{D}}^{b,a} \right) \leq \binom{a-b+1}{a-b-r} (1-q_0)^{a-b-r}$$

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:

$$\mathbb{P}\left(\| \prod_{s=a}^{b} \tilde{H}_{0,B-1}^{s,\top} x \| \geq (1 - \gamma B \lambda_{\min}(G))^{\frac{r}{2}} \middle| \tilde{\mathcal{D}}^{b,a} \right) \leq \exp(-c_3(b - a + 1))$$

Now, let $\mathcal{N}$ be a 1/2-net of the sphere $\mathcal{S}^{d-1}$. Using Corollary 4.2.13 in [186], we can choose $|\mathcal{N}| \leq 6^d$. By Lemma 4.4.1 in [186] we show that:

$$\| \prod_{s=a}^{b} \tilde{H}_{0,B-1}^{s,\top} \| \leq 2 \sup_{x \in \mathcal{N}} \| \prod_{s=a}^{b} \tilde{H}_{0,B-1}^{s,\top} x \| \tag{6.120}$$

By union bounding Equation (6.120) for every $x \in \mathcal{N}$, we conclude that:

$$\begin{aligned} &\mathbb{P}\left(\| \prod_{s=a}^{b} \tilde{H}_{0,B-1}^{s,\top} \| \geq 2(1 - \zeta\gamma B \lambda_{\min}(G))^{c_4(b-a+1)} \middle| \tilde{\mathcal{D}}^{b,a} \right) \leq |\mathcal{N}| \exp(-c_3(a - b + 1)) \\ &= \exp(-c_3(a - b + 1) + c_5 d) \end{aligned} \tag{6.121}$$

□

We will now state the equivalent of [110, Lemma 32]. The proof proceeds similarly, but using Theorem 6.9.4 instead. Consider the following operator:

$$Fa, N := \sum_{t=a}^{N-1} \prod_{s=t}^{a+1} \tilde{H}_{0,B-1}^{s,\top} \tag{6.122}$$

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

Theorem 6.9.5. *Suppose all the conditions in Theorem 6.9.4 hold. Then, for any $\delta \in (0, 1)$, we have:*

$$\mathbb{P}\left(\|F_{a,N}\| \geq C \left(d + \log \frac{N}{\delta} + \frac{1}{\zeta\gamma B \lambda_{\min}(G)} \right) \middle| \tilde{\mathcal{D}}^{a,N} \right) \leq \delta$$

Where C is a universal constant.

6.9.2 Proof of lemma 6.8.6

First, we will obtain a crude upper bound on $\left\|\tilde{a}_j^{t-1} - a^*\right\|$ using Theorem 6.9.4. That is, we want to show that $\left\|\tilde{a}_j^{t-1} - a^*\right\|$ does not grow too large with high probability.

Proposition 5. *Let $\lambda_{\min} \equiv \lambda_{\min}(G)$. Conditional on $\tilde{\mathcal{D}}^{0,t-1} \cap \cap_{r=0}^{t-1} \mathcal{E}_{0,B-1}^r$, with probability at least $1 - N\delta$, for all $1 \leq t \leq N$, all $1 \leq j \leq B$ we have*

$$\left\|\tilde{a}_j^{t-1} - a^*\right\| \leq \|a_0 - a\| + 2\gamma B\sqrt{R\beta}C\left(d + \log\frac{N}{\delta} + \frac{1}{\zeta\gamma B\lambda_{\min}}\right) \tag{6.123}$$

where C is constant depending only on C_η.

Proof. Let us start with the expression for $\tilde{a}_j^{t-1} - a^*$

$$\begin{aligned}(\tilde{a}_j^{t-1} - a^*)^\top = (a_0 - a^*)^\top \left(\prod_{s=0}^{t-2} \tilde{H}_{0,B-1}^s\right) \tilde{H}_{0,j-1}^{t-1} + 2\gamma \sum_{i=0}^{j-1} \phi'(\tilde{\xi}_{-i}^{t-1})\varepsilon_{-i}^{t-1} \tilde{X}_{-i}^{t-1,\top} \tilde{H}_{i+1,j-1}^{t-1} \\ + 2\gamma \sum_{r=2}^{t} \sum_{i=0}^{B-1} \phi'(\tilde{\xi}_{-i}^{t-r})\varepsilon_{-i}^{t-r} \tilde{X}_{-i}^{t-r,\top} \tilde{H}_{i+1,B-1}^{t-r} \left(\prod_{s=r-1}^{1} \tilde{H}_{0,B-1}^{t-s}\right)\end{aligned} \tag{6.124}$$

We will work on the event $\tilde{\mathcal{D}}^{0,t-1} \cap \cap_{r=0}^{t-1} \mathcal{E}_{0,B-1}^r$. It is clear from Equation (6.124) that:

$$\left\|\tilde{a}_j^{t-1} - a^*\right\| \leq \|a_0 - a^*\| + 2\gamma B\sqrt{R\beta} + 2\gamma\sqrt{R\beta}B \sum_{r=2}^{t} \left\|\left(\prod_{s=r-1}^{1} \tilde{H}_{0,B-1}^{t-s}\right)\right\|$$

We use Theorem 6.9.5 (with appropriate constant $C > 1$ to account for minor differences in indexing) to show that conditional on $\tilde{\mathcal{D}}^{0,t-1} \cap \cap_{r=0}^{t-1} \mathcal{E}_{0,B-1}^r$, for fixed t, with probability at least $1 - \delta$, for all $1 \leq j \leq B$

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

Thus taking union bound we get that conditional on $\tilde{\mathcal{D}}^{0,t-1} \cap \cap_{r=0}^{t-1} \mathcal{E}_{0,B-1}^r$ with probability at least $1 - N\delta$, for all $1 \leq t \leq N-1$ and all $1 \leq j \leq B$

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

□

Proof of Lemma 6.8.6. On the event $\mathcal{E}_{0,j}^r \cap \tilde{\mathcal{D}}^{r,N-1}$, we note the following inequalities

$$\bar{\tilde{a}}_i^s = \tilde{a}_i^s \; 0 \leq s < r, \; 0 \leq i \leq B-1 \tag{6.125}$$

$$\bar{\tilde{a}}_0^r = \tilde{a}_0^r \tag{6.126}$$

$$\|\bar{\tilde{a}}_i^s - \tilde{a}_i^s\| \leq \begin{cases} 4i\gamma\sqrt{R\beta} + \sum_{k=0}^{i-1} 4\gamma R \left\|\tilde{a}_k^r - a^*\right\|, & s = r,\ 1 \leq i \leq j \\ 4(j+1)\gamma\sqrt{R\beta} + \sum_{k=0}^{j-1} 4\gamma R \left\|\tilde{a}_k^r - a^*\right\|, s = r, & j+1 \leq i \leq B-1 \\ 4(j+1)\gamma\sqrt{R\beta} + \sum_{k=0}^{j-1} 4\gamma R \left\|\tilde{a}_k^r - a^*\right\|, r < s, & 0 \leq i \leq B-1 \end{cases} \tag{6.127}$$

The result then follows from an application of Proposition 5 with δ chosen as in 6.8.4 □

6.9.3 Proof of claim 14

Proof. Let $r_2 > r_1$. As in proof of Claim 13, let Cr' denote the resampled version of Cr obtained by re-sampling $\eta_{-j_1}^{t-r_1}$ i.e.,

$$\begin{aligned} \mathrm{Cr}'(t, r_1, r_2, j_1, j_2) := 4\gamma^2 \varepsilon_{-j_1}^{t-r_1} \varepsilon_{-j_2}^{t-r_2} \mathcal{R}_{-j_1}^{t-r_1} \Bigg[\tilde{X}_{-j_2}^{t-r_2,\top} \tilde{H}_{j_2+1,B-1}^{t-r_2} \left(\prod_{s=r_2-1}^{1} \tilde{H}_{0,B-1}^{t-s} \right) \cdot \\ \left(\prod_{s=1}^{r_1-1} \tilde{H}_{0,B-1}^{t-s,\top} \right) \tilde{H}_{j_1+1,B-1}^{t-r_1,\top} \tilde{X}_{-j_1}^{t-r_1} \Bigg] \\ = 4\gamma^2 \varepsilon_{-j_1}^{t-r_1} \varepsilon_{-j_2}^{t-r_2} \tilde{X}_{-j_2}^{t-r_2,\top} \left(\tilde{H}_{j_2+1,B-1}^{t-r_2} \right) \left(\prod_{s=r_2-1}^{r_1+1} \tilde{H}_{0,B-1}^{t-s} \right) \cdot \\ \mathcal{R}_{-j_1}^{t-r_1} \left(\prod_{s=r_1}^{1} \tilde{H}_{0,B-1}^{t-s} \right) \mathcal{R}_{-j_1}^{t-r_1} \left(\prod_{s=1}^{r_1-1} \tilde{H}_{0,B-1}^{t-s,\top} \right) \mathcal{R}_{-j_1}^{t-r_1} \left(\tilde{H}_{j_1+1,B-1}^{t-r_1,\top} \right) \tilde{X}_{-j_1}^{t-r_1} \end{aligned} \tag{6.128}$$

Here we have used the fact that $\mathcal{R}_{-j_1}^{t-r_1}$ does not affect the buffers up to $t - r_1 - 1$ and only $\tilde{X}$s that are affected are in the term $\tilde{H}_{0,j_1-1}^{t-r_1}$. Like in Claim 13, notice that

$$\mathbb{E}\left[\sum_{r_2 > r_1} \sum_{j_1, j_2} \mathrm{Cr}'(t, r_1, r_2, j_1, j_2) \right] = 0$$

Applying Lemma 6.8.8, we conclude that:

$$\begin{aligned} \sum_{r_2 > r_1} \sum_{j_1, j_2} \mathrm{Cr}'(t, r_1, r_2, j_1, j_2) \\ = 2\gamma \sum_{r_1=1}^{t-1} \sum_{j_1=0}^{B-1} (\tilde{a}_B^{t-r_1-1,v})^\top \mathcal{R}_{-j_1}^{t-r_1} \left(\tilde{H}_{0,B-1}^{t-r_1} \right) \mathcal{R}_{-j_1}^{t-r_1} \left(\prod_{s=r_1-1}^{1} \tilde{H}_{0,B-1}^{t-s} \right) \cdot \\ \mathcal{R}_{-j_1}^{t-r_1} \left(\prod_{s=1}^{r_1-1} \tilde{H}_{0,B-1}^{t-s,\top} \right) \mathcal{R}_{-j_1}^{t-r_1} \left(\tilde{H}_{j_1+1,B-1}^{t-r_1,\top} \right) \tilde{X}_{-j_1}^{t-r_1} \varepsilon_{-j_1}^{t-r_1} \end{aligned} \tag{6.129}$$

We cannot continue our analysis like in Claim 13 because due to resampling of $\varepsilon_{-j_1}^{t-r_1}$, $\tilde{H}_{0,B-1}^{t-r_1}$ changes not just because of the iterates $\tilde{a}_i^{t-r_1}$ but also due to $\tilde{X} \to \bar{\tilde{X}}$.

Further

$$\mathbb{E}\left[\sum_{r_2>r_1}\sum_{j_1,j_2}\mathrm{Cr}'(t,r_1,r_2,j_1,j_2)1\left[\tilde{\mathcal{D}}^{0,t-r_1-1}\right]1\left[\tilde{\mathcal{D}}^{t-r_1+1,t-1}\right]1\left[\tilde{\mathcal{D}}^{t-r_1}_{-j_1}\right]\right]=0 \tag{6.130}$$

Next we have simple lemma

Lemma 6.9.6. *Consider for each (r_1,j_1), the re-sampling operator $\mathcal{R}^{t-r_1}_{-j_1}$*

$$\begin{aligned}&\left|\mathbb{E}\left[\sum_{r_2>r_1}\sum_{j_1,j_2}\mathrm{Cr}(t,r_1,r_2,j_1,j_2)1\left[\tilde{\mathcal{D}}^{0,t-1}\right]\right]\right|\leq 4\gamma^2R\frac{(Bt)^2}{2}C_\eta\sigma^2\frac{1}{T^{\alpha/2}}+\\&\left|\mathbb{E}\left[\sum_{r_2>r_1}\sum_{j_1,j_2}\mathrm{Cr}(t,r_1,r_2,j_1,j_2)1\left[\tilde{\mathcal{D}}^{0,t-1}\right]\mathcal{R}^{t-r_1}_{-j_1}1\left[\tilde{\mathcal{D}}^{t-r_1}_{-0}\right]\right]\right|\end{aligned} \tag{6.131}$$

Proof. We have

$$1\left[\tilde{\mathcal{D}}^{0,t-1}\right]=1\left[\tilde{\mathcal{D}}^{0,t-1}\right]\mathcal{R}^{t-r_1}_{-j_1}1\left[\tilde{\mathcal{D}}^{t-r_1}_{-0}\right]+1\left[\tilde{\mathcal{D}}^{0,t-1}\right]\mathcal{R}^{t-r_1}_{-j_1}1\left[\tilde{\mathcal{D}}^{t-r_1,C}_{-0}\right] \tag{6.132}$$

Hence

$$\begin{aligned}&\left|\mathbb{E}\left[\sum_{r_2>r_1}\sum_{j_1,j_2}\mathrm{Cr}(t,r_1,r_2,j_1,j_2)1\left[\tilde{\mathcal{D}}^{0,t-1}\right]\right]\right|\\\leq&\left|\mathbb{E}\left[\sum_{r_2>r_1}\sum_{j_1,j_2}\mathrm{Cr}(t,r_1,r_2,j_1,j_2)1\left[\tilde{\mathcal{D}}^{0,t-1}\right]\mathcal{R}^{t-r_1}_{-j_1}1\left[\tilde{\mathcal{D}}^{t-r_1}_{-0}\right]\right]\right|+\\&4\gamma^2R\frac{(Bt)^2}{2}C_\eta\sigma^2\frac{1}{T^{\alpha/2}}\end{aligned} \tag{6.133}$$

where we used $\mathcal{R}^{t-r_1}_{-j_1}1\left[\tilde{\mathcal{D}}^{t-r_1,C}_{-0}\right]$ is identically distributed as $1\left[\tilde{\mathcal{D}}^{t-r_1,C}_{-0}\right]$ and hence $\mathbb{E}\left[\mathcal{R}^{t-r_1}_{-j_1}1\left[\tilde{\mathcal{D}}^{t-r_1,C}_{-0}\right]\right]\leq\frac{1}{T^\alpha}$

□

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

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

Now notice that

$$\begin{aligned}
&\mathbb{E}\Bigg[\sum_{r_2>r_1}\sum_{j_1,j_2}\mathrm{Cr}'(t,r_1,r_2,j_1,j_2)\cdot \\
&\quad 1\left[\tilde{\mathcal{D}}^{0,t-r_1-1}\right]1\left[\tilde{\mathcal{D}}^{t-r_1+1,t-1}\right]1\left[\tilde{\mathcal{D}}^{t-r_1}_{-j_1}\right]\mathcal{R}^{t-r_1}_{-j_1}1\left[\tilde{\mathcal{D}}^{t-r_1}_{-0}\right]\Bigg] \\
&= 0
\end{aligned} \tag{6.134}$$

Hence

$$\begin{aligned}
&\mathbb{E}\left[\sum_{r_2>r_1}\sum_{j_1,j_2}\mathrm{Cr}'(t,r_1,r_2,j_1,j_2)1\left[\tilde{\mathcal{D}}^{0,t-1}\right]\mathcal{R}^{t-r_1}_{-j_1}1\left[\tilde{\mathcal{D}}^{t-r_1}_{-0}\right]\right] = 0 - \\
&\mathbb{E}\Bigg[\sum_{r_2>r_1}\sum_{j_1,j_2}\mathrm{Cr}'(t,r_1,r_2,j_1,j_2)1\left[\tilde{\mathcal{D}}^{0,t-r_1-1}\right]1\left[\tilde{\mathcal{D}}^{t-r_1+1,t-1}\right]\cdot \\
&\quad 1\left[\tilde{\mathcal{D}}^{t-r_1}_{-j_1}\right]1\left[\cup_{i=0}^{j_1-1}\tilde{\mathcal{C}}^{t-r,C}_{-i}\right]\mathcal{R}^{t-r_1}_{-j_1}1\left[\tilde{\mathcal{D}}^{t-r_1}_{-0}\right]\Bigg]
\end{aligned} \tag{6.135}$$

Thus

$$\left|\mathbb{E}\left[\sum_{r_2>r_1}\sum_{j_1,j_2}\mathrm{Cr}'(t,r_1,r_2,j_1,j_2)1\left[\tilde{\mathcal{D}}^{0,t-1}\right]\mathcal{R}^{t-r_1}_{-j_1}1\left[\tilde{\mathcal{D}}^{t-r_1}_{-0}\right]\right]\right| \leq 2\gamma^2 R\frac{(Bt)^2}{2}C_\eta\sigma^2\frac{1}{T^{\alpha/2}} \tag{6.136}$$

Now, similar to lemma 6.8.7, on the event $\mathcal{E}^{r_1}_{0,j_1}\cap\tilde{\mathcal{D}}^{0,t-1}\cap\mathcal{A}^{t-1}$ we have:

$$\left\|\left(\prod_{s=r_1-1}^{1}\tilde{H}^{t-s}_{0,B-1}\right)-\mathcal{R}^{t-r_1}_{-j_1}\left(\prod_{s=r_1-1}^{1}\tilde{H}^{t-s}_{0,B-1}\right)\right\| \leq CBt\left\|\phi''\right\|\gamma^2R^3B\frac{\sqrt{\beta}}{\zeta\lambda_{\min}} \tag{6.137}$$

Next, similar to lemma 6.8.7 for $\gamma R\leq\frac{1}{2}$, on the event $\tilde{\mathcal{D}}^{t-r_1}_{-0}\cap\cap_{i=0}^{B-1}\left\{\left\|\mathcal{R}^{t-r_1}_{j_1}\tilde{X}^{t-r_1}_{-i}\right\|^2\leq R\right\}$ we have

$$\left\|\tilde{H}^{t-r_1}_{0,B-1}-\mathcal{R}^{t-r_1}_{-j_1}\left(\tilde{H}^{t-r_1}_{0,B-1}\right)\right\| \leq 4\gamma RB \tag{6.138}$$

Finally we can bound the norm of the expected difference of sums of Cr and Cr$'$ using lemma 6.8.8

and (6.129) as

$$\begin{aligned}
&\left|\mathbb{E}\left[\sum_{r_2>r_1}\sum_{j_1,j_2}(\mathrm{Cr}-\mathrm{Cr}')1\left[\tilde{\mathcal{D}}^{0,t-1}\right]\mathcal{R}_{-j_1}^{t-r_1}1\left[\tilde{\mathcal{D}}_{-0}^{t-r_1}\right]1\left[\cap_{s=0}^{t-1}\cap_{i=0}^{B-1}\mathcal{E}_i^s\right]1\left[\mathcal{A}^{t-1}\right]\right]\right| \\
&\leq 2\gamma\mathbb{E}\left[\sum_{r_1=1}^{t-1}\sum_{j_1}\sqrt{R}|\varepsilon_{-j_1}^{t-r_1}|\left[\left\|\tilde{a}_B^{t-r_1-1,v}\right\|1\left[\tilde{\mathcal{D}}^{0,t-r_1-1}\right]\right]\cdot\right. \\
&\quad\left.\left(C\left\|\phi''\right\|\gamma^2TR^3B\frac{\sqrt{\beta}}{\zeta\lambda_{\min}}+C\gamma RB\right)\right] \\
&\leq \left(C\left\|\phi''\right\|\gamma^3T^2R^3B\frac{\sqrt{\beta}}{\zeta\lambda_{\min}}+C\gamma^2TRB\right)\sqrt{RC_\eta\sigma^2}\sqrt{\sup_{s\leq N-1}\mathbb{E}\left[\left\|\tilde{a}_B^{s,v}\right\|^2 1\left[\tilde{\mathcal{D}}^{0,s}\right]\right]}
\end{aligned} \tag{6.139}$$

Thus

$$\begin{aligned}
&\left|\mathbb{E}\left[\sum_{r_2>r_1}\sum_{j_1,j_2}(\mathrm{Cr}-\mathrm{Cr}')1\left[\tilde{\mathcal{D}}^{0,t-1}\right]\mathcal{R}_{-j_1}^{t-r_1}1\left[\tilde{\mathcal{D}}_{-0}^{t-r_1}\right]\right]\right| \\
&\leq C\left(\left\|\phi''\right\|\gamma^3T^2R^3B\frac{\sqrt{\beta}}{\zeta\lambda_{\min}}+\gamma^2TRB\right)\sqrt{RC_\eta\sigma^2}\sqrt{\sup_{s\leq N-1}\mathbb{E}\left[\left\|\tilde{a}_B^{s,v}\right\|^2 1\left[\tilde{\mathcal{D}}^{0,s}\right]\right]} \\
&\quad+C(Bt)^2\left[\gamma^2RC_\eta\sigma^2\left(\sqrt{\mathbb{P}\left[\cup_{s=0}^{N-1}\cup_{i=0}^{B-1}\mathcal{E}_i^{s,C}\right]}+\sqrt{\mathbb{P}\left[\mathcal{A}^{t-1,C}\right]}\right)\right] \\
&\leq C\left(\left\|\phi''\right\|\gamma^3T^2R^3B\frac{\sqrt{\beta}}{\zeta\lambda_{\min}}+\gamma^2TRB\right)\sqrt{RC_\eta\sigma^2}\sqrt{\sup_{s\leq N-1}\mathbb{E}\left[\left\|\tilde{a}_B^{s,v}\right\|^2 1\left[\tilde{\mathcal{D}}^{0,s}\right]\right]} \\
&\quad+C(Bt)^2\gamma^2RC_\eta\sigma^2\frac{1}{T^{\alpha/2}}
\end{aligned} \tag{6.140}$$

Combining everything we conclude the claim.

□

Chapter 7

Numerical evaluation

7.1 Linear system identification

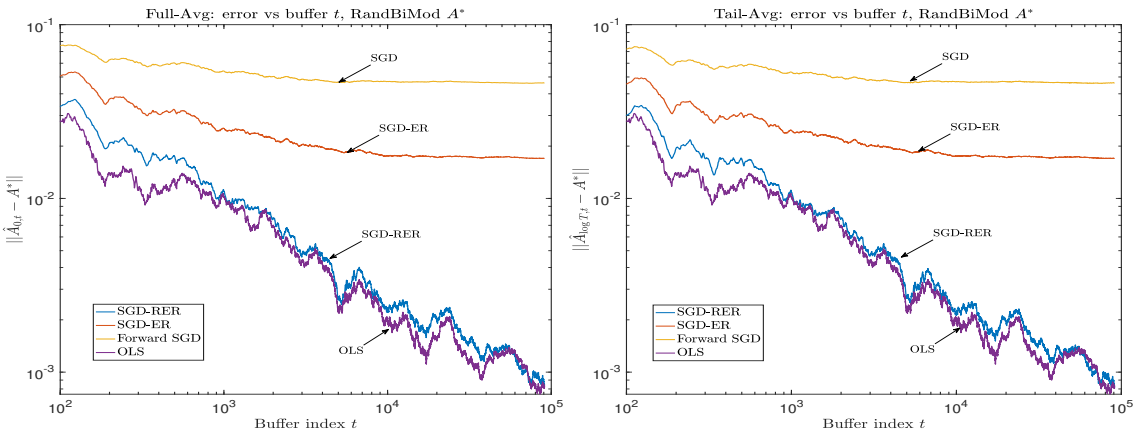

Figure 7-1: Gaussian $\mathsf{VAR}(A^*, \mu)$: Parameter error for tail averaged and full average iterates of $\mathsf{SGD}-\mathsf{RER}$ and baselines. $\mathsf{SGD}-\mathsf{RER}$ and OLS incur similar parameter error, while error incurred by SGD and $\mathsf{SGD}-\mathsf{ER}$ saturate at significantly higher level, indicating non-zero bias. The parameters used are $\rho = 0.9$, $d = 5$, $T = 10^7$, $B = 100$, $u = 10$. R is estimated and $\gamma = 1/2R$.

In this section, we compare performance of our $\mathsf{SGD}-\mathsf{RER}$ method on synthetic data generated from a linear dynamical system against the performance of standard baselines OLS and SGD, along with $\mathsf{SGD}-\mathsf{ER}$ method that applies standard experience replay technique, but where points from a buffer are sampled *randomly*.

Synthetic data: We sample data from $\mathsf{VAR}(A^*, \mu)$ with $X_0 = 0$, $\mu \sim \mathcal{N}(0, \sigma^2 I)$ and $A^* \in \mathbb{R}^{d\times d}$ is generated from the "RandBiMod" distribution. That is, $A^* = U\Lambda U^\top$ with random orthogonal U, and Λ is diagonal with $\lceil d/2 \rceil$ entries on diagonal being ρ and the remaining diagonal entries are set to $\rho/3$. We set $d = 5$, $\rho = 0.9$ and $\sigma^2 = 1$. We fix a horizon $T = 10^7$ and set the buffer size as $B = 100$ and $u = 10$. To estimate R from the data, we use the first $\lfloor 2 \log T \rfloor = 32$ samples and set

R as the sum of the norms of these samples. We let the stepsize to be $\gamma = \frac{1}{2R}$ which is *aggressive* compared to our theorems. We start the $\mathsf{SGD-RER}$ and other *SGD*-like algorithms from the second buffer onward.

For tail averaging, as described in algorithm 1, we ignore the first $\lfloor \log T \rfloor = 16$ buffers, and maintain a running tail average at the end of each of the subsequent buffers. In figure 7-1, we plot the parameter errors $\left\| \hat{A}_{\log T,t} - A^* \right\|$ and $\left\| \hat{A}_{0,t} - A^* \right\|$ versus the buffer index t as the algorithm runs for horizon T. For OLS, we include samples in the first buffer as well (which were used for estimating R). Clearly, $\mathsf{SGD-RER}$ has very similar performance as that of OLS whereas $\mathsf{SGD-ER}$ and SGD seem to display residual bias for the chosen step-size (which is logarithmic in the horizon T) and buffer lengths. We also observe a similar behavior when we choose $A^* = \rho I$.

7.2 Generalized linear system identification

In this section, we compare performance of our methods $\mathsf{SGD-RER}$ and Quasi Newton method on synthetic data generated from a generalized linear dynamical system against the performance of standard baselines SGD (called 'Forward SGD' here), GLMtron, along with $\mathsf{SGD-ER}$ method that applies standard experience replay technique i.e, the points from a buffer are sampled *randomly* instead of the reverse order. Since GLMtron and Quasi Newton Method are offline and $\mathsf{SGD-RER}$, SGD and $\mathsf{SGD-ER}$ are streaming, we compare the algorithms by plotting parameter error measured by the Frobenius norm with respect to the compute time. We also compare error vs. number of iterations for the streaming algorithms.

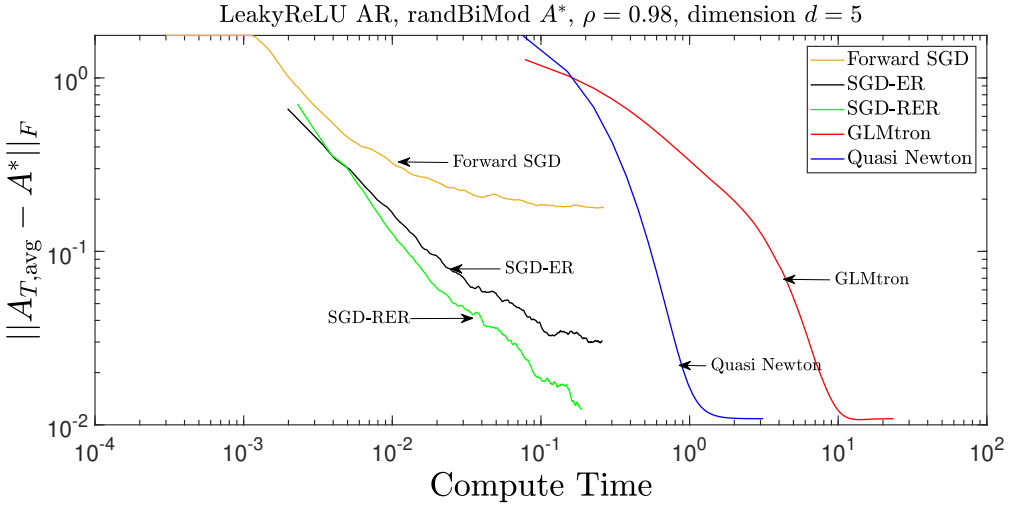

Figure 7-2: Error vs. Computation

Synthetic data: We sample data from $\mathsf{NLDS}(A^*, \mu, \phi)$ where $\mu \sim \mathcal{N}(0, \sigma^2 I)$ and $A^* \in \mathbb{R}^{d \times d}$ is generated from the "RandBiMod" distribution. That is, $A^* = U \Lambda U^\top$ with random orthogonal U,

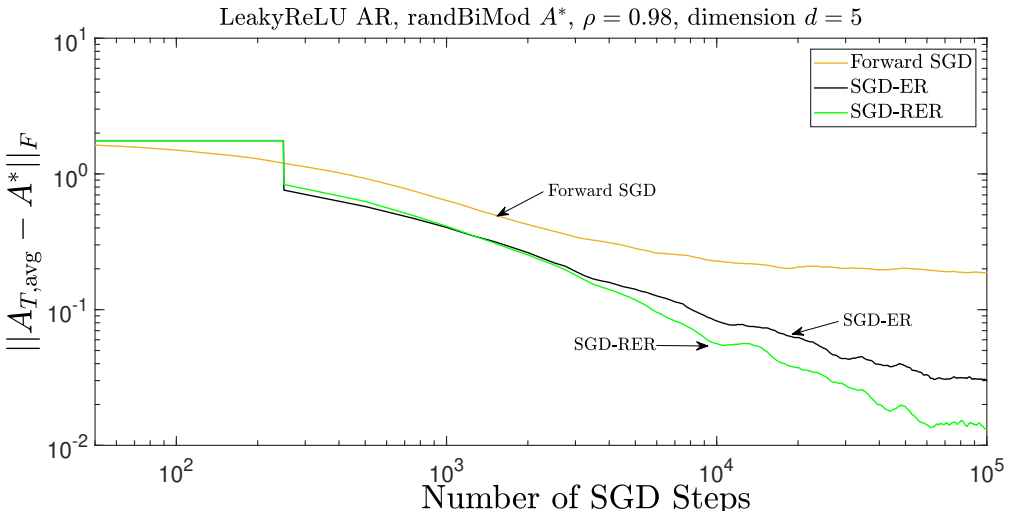

Figure 7-3: Error vs. SGD updates

Figure 7-4: Performance of various algorithms for the case of $\phi = \mathsf{LeakyReLU}$

and Λ is diagonal with $\lceil d/2 \rceil$ entries on diagonal being ρ and the remaining diagonal entries are set to $\rho/3$. ϕ is the leaky ReLU function given by $\phi(x) = 0.5x\mathbb{1}(x < 0) + x\mathbb{1}(x \geq 0)$. We set $d = 5$, $\rho = 0.98$ and $\sigma^2 = 1$. We set a horizon of $T = 10^5$.

Algorithm Parameters We set $B = 240$ and $u = 10$ for the buffer size and gap size respectively for both $\mathsf{SGD} - \mathsf{RER}$ and $\mathsf{SGD} - \mathsf{ER}$ and use full averaging (i.e, $\theta = 0$ in Algorithm 3). We set the step size $\gamma = \frac{5 \log T}{T}$ for SGD, $\mathsf{SGD} - \mathsf{RER}$, and $\mathsf{SGD} - \mathsf{ER}$ and $\gamma_{\mathsf{newton}} = 0.2$ and $\gamma_{\mathsf{GLMtron}} = 0.017$.

From Figure 7-4 observe that $\mathsf{SGD}-\mathsf{ER}$ and SGD obtain sub-optimal results compared $\mathsf{SGD}-\mathsf{RER}$, Quasi Newton Method and GLMtron. After a single pass, the performance of $\mathsf{SGD} - \mathsf{RER}$ almost matches that of the offline algorithms. The step sizes for GLMtron have to be chosen to be small in-order to ensure that the algorithm does not diverge as noted in Section 6.2, which slows down its convergence time compared to the Quasi Newton method.

Bibliography

[1] X. Chen, D. W. K. Ng, W. Yu, E. G. Larsson, N. Al-Dhahir, and R. Schober, "Massive access for 5g and beyond," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 3, pp. 615–637, 2020.

[2] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[3] Y. Polyanskiy, "A perspective on massive random-access," in *2017 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2017, pp. 2523–2527.

[4] O. Ordentlich and Y. Polyanskiy, "Low complexity schemes for the random access gaussian channel," in *2017 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2017, pp. 2528–2532.

[5] V. K. Amalladinne, J.-F. Chamberland, and K. R. Narayanan, "A coded compressed sensing scheme for unsourced multiple access," *IEEE Transactions on Information Theory*, vol. 66, no. 10, pp. 6509–6533, 2020.

[6] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Minimum energy to send k bits through the Gaussian channel with and without feedback," *IEEE Transactions on Information Theory*, vol. 57, no. 8, pp. 4880–4902, 2011.

[7] V. K. Amalladinne, J. R. Ebert, J.-F. Chamberland, and K. R. Narayanan, "An enhanced decoding algorithm for coded compressed sensing with applications to unsourced random access," *arXiv preprint arXiv:2112.00270*, 2021.

[8] A. Fengler, P. Jung, and G. Caire, "SPARCs for Unsourced Random Access," *IEEE Transactions on Information Theory*, vol. 67, no. 10, pp. 6894–6915, 2021.

[9] A. Fengler, S. Haghighatshoar, P. Jung, and G. Caire, "Non-Bayesian activity detection, large-scale fading coefficient estimation, and unsourced random access with a massive MIMO receiver," *IEEE Transactions on Information Theory*, vol. 67, no. 5, pp. 2925–2951, 2021.

[10] A. Fengler, P. Jung, and G. Caire, "Pilot-based unsourced random access with a massive mimo receiver in the quasi-static fading regime," in *2021 IEEE 22nd International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE, 2021, pp. 356–360.

[11] A. K. Pradhan, V. K. Amalladinne, K. R. Narayanan, and J.-F. Chamberland, "Polar coding and random spreading for unsourced multiple access," in *ICC 2020-2020 IEEE International Conference on Communications (ICC)*. IEEE, 2020, pp. 1–6.

[12] ——, "Ldpc codes with soft interference cancellation for uncoordinated unsourced multiple access," *arXiv preprint arXiv:2105.13985*, 2021.

[13] X. Chen, T.-Y. Chen, and D. Guo, "Capacity of Gaussian Many-Access Channels." *IEEE Trans. Information Theory*, vol. 63, no. 6, pp. 3516–3539, 2017.

[14] I. Zadik, Y. Polyanskiy, and C. Thrampoulidis, "Improved bounds on Gaussian MAC and sparse regression via Gaussian inequalities," in *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2019.

[15] F. Wei, Y. Wu, W. Chen, W. Yang, and G. Caire, "On the Fundamental Limits of MIMO Massive Multiple Access Channels," *arXiv preprint arXiv:1807.05553*, 2018.

[16] Y. Polyanskiy, "Information theoretic perspective on massive multiple-access," in *Short Course (slides)*, Skoltech Inst. of Tech., Moscow, Russia, Jul. 2018. [Online]. Available: https://people.lids.mit.edu/yp/homepage/data/SkolTech18-MAC-lectures.pdf

[17] S. Verdú, "Spectral efficiency in the wideband regime," *IEEE Transactions on Information Theory*, vol. 48, no. 6, pp. 1319–1343, 2002.

[18] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Minimum energy to send k bits with and without feedback," in *2010 IEEE International Symposium on Information Theory*. IEEE, 2010, pp. 221–225.

[19] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Quasi-static multiple-antenna fading channels at finite blocklength," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 4232–4265, 2014.

[20] M. J. Wainwright, "Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting," *IEEE Transactions on Information Theory*, vol. 55, no. 12, pp. 5728–5741, 2009.

[21] G. Reeves and M. Gastpar, "A note on optimal support recovery in compressed sensing," in *Signals, Systems and Computers, 2009 Conference Record of the Forty-Third Asilomar Conference on*. IEEE, 2009, pp. 1576–1580.

[22] ——, "The sampling rate-distortion tradeoff for sparsity pattern recovery in compressed sensing," *IEEE Transactions on Information Theory*, vol. 58, no. 5, pp. 3065–3092, 2012.

[23] I. Bettesh and S. Shamai, "Outages, expected rates and delays in multiple-users fading channels," in *Proceedings of the 2000 Conference on Information Science and Systems*, vol. 1, 2000.

[24] S. S. Kowshik, K. Andreev, A. Frolov, and Y. Polyanskiy, "Energy efficient random access for the quasi-static fading MAC," in *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2019.

[25] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proceedings of the National Academy of Sciences*, vol. 106, no. 45, pp. 18 914–18 919, 2009.

[26] A. Javanmard and A. Montanari, "State evolution for general approximate message passing algorithms, with applications to spatial coupling," *Information and Inference: A Journal of the IMA*, vol. 2, no. 2, pp. 115–144, 2013.

[27] K. Hsieh, C. Rush, and R. Venkataramanan, "Near-optimal coding for massive multiple access," *arXiv preprint arXiv:2102.04730*, 2021.

[28] M. Bayati and A. Montanari, "The dynamics of message passing on dense graphs, with applications to compressed sensing," *IEEE Transactions on Information Theory*, vol. 57, no. 2, pp. 764–785, 2011.

[29] P. R. Kumar and P. Varaiya, *Stochastic systems: Estimation, identification, and adaptive control*. SIAM, 2015.

[30] B. Açıkmeşe, J. M. Carson, and L. Blackmore, "Lossless convexification of nonconvex control bound and pointing constraints of the soft landing optimal control problem," *IEEE Transactions on Control Systems Technology*, vol. 21, no. 6, pp. 2104–2113, 2013.

[31] J. D. Hamilton, *Time series analysis*. Princeton university press, 2020.

[32] A. Fujita, J. R. Sato, H. M. Garay-Malpartida, R. Yamaguchi, S. Miyano, M. C. Sogayar, and C. E. Ferreira, "Modeling gene expression regulatory networks with the sparse vector autoregressive model," *BMC Systems Biology*, vol. 1, p. 39, 2007.

[33] H. B. Mann and A. Wald, "On the statistical treatment of linear stochastic difference equations," *Econometrica, Journal of the Econometric Society*, pp. 173–220, 1943.

[34] T. W. Anderson, "On asymptotic distributions of estimates of parameters of stochastic difference equations," *The Annals of Mathematical Statistics*, pp. 676–687, 1959.

[35] T. Lai and C. Wei, "Asymptotic properties of general autoregressive models and strong consistency of least-squares estimates of their parameters," *Journal of multivariate analysis*, vol. 13, no. 1, pp. 1–23, 1983.

[36] M. Duflo, R. Senoussi, and A. Touati, "Propriétés asymptotiques presque sûres de l'estimateur des moindres carrés d'un modèle autoregressif vectoriel," in *Annales de l'IHP Probabilités et statistiques*, vol. 27, no. 1, 1991, pp. 1–25.

[37] L. Ljung, *System identification: theory for the user*. Prentice Hall, Englewood Cliffs, NJ, 1987.

[38] B. Nielsen, "Singular vector autoregressions with deterministic terms: Strong consistency and lag order determination." 2008.

[39] H. Robbins and S. Monro, "A stochastic approximation method," *The annals of mathematical statistics*, pp. 400–407, 1951.

[40] Y. Z. Tsypkin, "Optimality in identification of linear plants," *International Journal of Systems Science*, vol. 14, no. 1, pp. 59–74, 1983.

[41] Y. Z. Tsypkin and B. Polyak, "Optimal recurrent algorithms for identification of nonstationary plants," *Computers & electrical engineering*, vol. 18, no. 5, pp. 365–371, 1992.

[42] L. Ljung, "Analysis of recursive stochastic algorithms," *IEEE transactions on automatic control*, vol. 22, no. 4, pp. 551–575, 1977.

[43] L. Ljung and T. Söderström, *Theory and practice of recursive identification*, ser. The MIT press series in signal processing, optimization, and control. United States: MIT Press, 1983, vol. 4.

[44] M. B. Nevel'son and R. Z. Has' minskii, *Stochastic approximation and recursive estimation*. American Mathematical Soc., 1976, vol. 47.

[45] M. C. Campi and E. Weyer, "Finite sample properties of system identification methods," *IEEE Transactions on Automatic Control*, vol. 47, no. 8, pp. 1329–1334, 2002.

[46] M. Vidyasagar and R. L. Karandikar, "A learning theory approach to system identification and stochastic adaptive control," in *Probabilistic and randomized methods for design under uncertainty*. Springer, 2006, pp. 265–302.

[47] V. Kuznetsov and M. Mohri, "Generalization bounds for non-stationary mixing processes," *Machine Learning*, vol. 106, no. 1, pp. 93–117, 2017.

[48] ——, "Learning Theory and Algorithms for Forecasting Non-stationary Time Series," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28. Curran Associates, Inc., 2015. [Online]. Available: https://proceedings.neurips.cc/paper/2015/file/41f1f19176d383480afa65d325c06ed0-Paper.pdf

[49] ——, "Theory and algorithms for forecasting time series," *arXiv preprint arXiv:1803.05814*, 2018.

[50] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, "Finite time identification in unstable linear systems," *Automatica*, vol. 96, pp. 342–353, 2018.

[51] M. Simchowitz, H. Mania, S. Tu, M. I. Jordan, and B. Recht, "Learning without mixing: Towards a sharp analysis of linear system identification," in *Conference On Learning Theory*. PMLR, 2018, pp. 439–473.

[52] T. Sarkar and A. Rakhlin, "Near optimal finite time identification of arbitrary linear dynamical systems," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5610–5618.

[53] Y. Jedra and A. Proutiere, "Finite-time Identification of Stable Linear Systems Optimality of the Least-Squares Estimator," in *2020 59th IEEE Conference on Decision and Control (CDC)*. IEEE, 2020, pp. 996–1001.

[54] ——, "Sample complexity lower bounds for linear system identification," in *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, 2019, pp. 2676–2681.

[55] A. Tsiamis and G. J. Pappas, "Linear systems can be hard to learn," *arXiv preprint arXiv:2104.01120*, 2021.

[56] S. Basu, G. Michailidis *et al.*, "Regularized estimation in sparse high-dimensional time series models," *The Annals of Statistics*, vol. 43, no. 4, pp. 1535–1567, 2015.

[57] S. Basu, X. Li, and G. Michailidis, "Low Rank and Structured Modeling of High-Dimensional Vector Autoregressions," *IEEE Transactions on Signal Processing*, vol. 67, no. 5, p. 1207–1222, Mar 2019.

[58] C. Hanck, M. Arnold, A. Gerber, and M. Schmelzer, "Introduction to econometrics with r," *University of Duisburg-Essen*, 2019.

[59] Y. Zheng, B. Tang, W. Ding, and H. Zhou, "A neural autoregressive approach to collaborative filtering," in *International Conference on Machine Learning*. PMLR, 2016, pp. 764–773.

[60] M. Hardt, T. Ma, and B. Recht, "Gradient descent learns linear dynamical systems," *arXiv preprint arXiv:1609.05191*, 2016.

[61] E. Hazan, K. Singh, and C. Zhang, "Learning linear dynamical systems via spectral filtering," *Advances in Neural Information Processing Systems*, vol. 30, pp. 6702–6712, 2017.

[62] E. Hazan, H. Lee, K. Singh, C. Zhang, and Y. Zhang, "Spectral filtering for general linear dynamical systems," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 4639–4648.

[63] P. Jain, S. S. Kowshik, D. Nagaraj, and P. Netrapalli, "Near-optimal Offline and Streaming Algorithms for Learning Non-Linear Dynamical Systems," *arXiv preprint arXiv:2105.11558*, 2021.

[64] N. Agarwal, S. Chaudhuri, P. Jain, D. Nagaraj, and P. Netrapalli, "Online target q-learning with reverse experience replay: Efficiently finding the optimal policy for linear mdps," *arXiv preprint arXiv:2110.08440*, 2021.

[65] P. Jain, P. Netrapalli, S. M. Kakade, R. Kidambi, and A. Sidford, "Parallelizing stochastic gradient descent for least squares regression: mini-batching, averaging, and model misspecification," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 8258–8299, 2017.

[66] D. Nagaraj, X. Wu, G. Bresler, P. Jain, and P. Netrapalli, "Least Squares Regression with Markovian Data: Fundamental Limits and Algorithms," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[67] L. Györfi and H. Walk, "On the averaged stochastic approximation for linear regression," *SIAM Journal on Control and Optimization*, vol. 34, no. 1, pp. 31–61, 1996.

[68] E. Rotinov, "Reverse Experience Replay," *arXiv preprint arXiv:1910.08780*, 2019.

[69] R. E. Ambrose, B. E. Pfeiffer, and D. J. Foster, "Reverse replay of hippocampal place cells is uniquely modulated by changing reward," *Neuron*, vol. 91, no. 5, pp. 1124–1136, 2016.

[70] M. T. Whelan, T. J. Prescott, and E. Vasilaki, "A robotic model of hippocampal reverse replay for reinforcement learning," *arXiv preprint arXiv:2102.11914*, 2021.

[71] S. Oymak and N. Ozay, "Non-asymptotic identification of lti systems from a single trajectory," in *2019 American Control Conference (ACC)*. IEEE, 2019, pp. 5655–5661.

[72] Y. Sattar and S. Oymak, "Non-asymptotic and accurate learning of nonlinear dynamical systems," *arXiv preprint arXiv:2002.08538*, 2020.

[73] D. Foster, T. Sarkar, and A. Rakhlin, "Learning nonlinear dynamical systems from a single trajectory," in *Learning for Dynamics and Control*. PMLR, 2020, pp. 851–861.

[74] M. Hardt, T. Ma, and B. Recht, "Gradient descent learns linear dynamical systems," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 1025–1068, 2018.

[75] A. Tsiamis and G. J. Pappas, "Finite sample analysis of stochastic system identification," in *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, 2019, pp. 3648–3654.

[76] T. Sarkar, A. Rakhlin, and M. A. Dahleh, "Finite Time LTI System Identification." *J. Mach. Learn. Res.*, vol. 22, pp. 26–1, 2021.

[77] S. Lale, K. Azizzadenesheli, B. Hassibi, and A. Anandkumar, "Logarithmic regret bound in partially observable linear dynamical systems," *arXiv preprint arXiv:2003.11227*, 2020.

[78] H. Lee, "Improved rates for identification of partially observed linear dynamical systems," *arXiv preprint arXiv:2011.10006*, 2020.

[79] B. Lee and A. Lamperski, "Non-asymptotic closed-loop system identification using autoregressive processes and hankel model reduction," in *2020 59th IEEE Conference on Decision and Control (CDC)*. IEEE, 2020, pp. 3419–3424.

[80] A. Cohen, A. Hasidim, T. Koren, N. Lazic, Y. Mansour, and K. Talwar, "Online linear quadratic control," in *International Conference on Machine Learning*. PMLR, 2018, pp. 1029–1038.

[81] N. Agarwal, B. Bullins, E. Hazan, S. Kakade, and K. Singh, "Online control with adversarial disturbances," in *International Conference on Machine Learning*. PMLR, 2019, pp. 111–119.

[82] E. Hazan, S. Kakade, and K. Singh, "The nonstochastic control problem," in *Algorithmic Learning Theory*. PMLR, 2020, pp. 408–421.

[83] X. Chen and E. Hazan, "Black-box control for linear dynamical systems," *arXiv preprint arXiv:2007.06650*, 2020.

[84] U. Ghai, H. Lee, K. Singh, C. Zhang, and Y. Zhang, "No-regret prediction in marginally stable systems," in *Conference on Learning Theory*. PMLR, 2020, pp. 1714–1757.

[85] P. Rashidinejad, J. Jiao, and S. Russell, "SLIP: Learning to predict in unknown dynamical systems with long-term memory," *arXiv preprint arXiv:2010.05899*, 2020.

[86] M. Kozdoba, J. Marecek, T. Tchrakian, and S. Mannor, "On-line learning of linear dynamical systems: Exponential forgetting in kalman filters," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 4098–4105.

[87] A. Tsiamis, N. Matni, and G. Pappas, "Sample complexity of kalman filtering for unknown systems," in *Learning for Dynamics and Control*. PMLR, 2020, pp. 435–444.

[88] A. Tsiamis and G. Pappas, "Online learning of the kalman filter with logarithmic regret," *arXiv preprint arXiv:2002.05141*, 2020.

[89] V. Kuznetsov and M. Mohri, "Time series prediction and online learning," in *Conference on Learning Theory*. PMLR, 2016, pp. 1190–1213.

[90] P. Alquier, K. Bertin, P. Doukhan, and R. Garnier, "High-dimensional VAR with low-rank transition," *Statistics and Computing*, vol. 30, no. 4, pp. 1139–1153, 2020.

[91] Y. Zheng and G. Cheng, "Finite-time analysis of vector autoregressive models under linear restrictions," *Biometrika*, vol. 108, no. 2, pp. 469–489, 2021.

[92] I. Wilms, S. Basu, J. Bien, and D. S. Matteson, "Sparse identification and estimation of large-scale vector autoregressive moving averages," *Journal of the American Statistical Association*, pp. 1–12, 2021.

[93] X. Lv, W. Cui, and Y. Liu, "Linear Convergence of Gradient Methods for Estimating Structured Transition Matrices in High-dimensional Vector Autoregressive Models," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[94] A. Modi, M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, "Joint Learning of Linear Time-Invariant Dynamical Systems," *arXiv preprint arXiv:2112.10955*, 2021.

[95] D. Wang and R. S. Tsay, "Robust Estimation of High-Dimensional Vector Autoregressive Models," *arXiv preprint arXiv:2107.11002*, 2021.

[96] L. Liu and D. Zhang, "Robust estimation of high-dimensional non-gaussian autoregressive models," 2021.

[97] S. Chen and S. A. Billings, "Representations of non-linear systems: the NARMAX model," *International journal of control*, vol. 49, no. 3, pp. 1013–1032, 1989.

[98] S. Chen, S. A. Billings, and P. Grant, "Non-linear system identification using neural networks," *International journal of control*, vol. 51, no. 6, pp. 1191–1214, 1990.

[99] S. N. Kumpati, P. Kannan *et al.*, "Identification and control of dynamical systems using neural networks," *IEEE Transactions on neural networks*, vol. 1, no. 1, pp. 4–27, 1990.

[100] L. M. Aguilar-Lobo, J. Loo-Yau, S. Ortega-Cisneros, P. Moreno, and J. Reynoso-Hernández, "Experimental study of the capabilities of the Real-Valued NARX neural network for behavioral modeling of multi-standard RF power amplifier," in *2015 IEEE MTT-S International Microwave Symposium*. IEEE, 2015, pp. 1–4.

[101] J. M. P. Menezes Jr and G. A. Barreto, "Long-term time series prediction with the narx network: An empirical evaluation," *Neurocomputing*, vol. 71, no. 16-18, pp. 3335–3343, 2008.

[102] L. Ljung, "System identification," *Wiley encyclopedia of electrical and electronics engineering*, pp. 1–19, 1999.

[103] Åström, Karl Johan and Eykhoff, Peter, "System identification—a survey," *Automatica*, vol. 7, no. 2, pp. 123–162, 1971.

[104] E. C. Hall, G. Raskutti, and R. Willett, "Inference of high-dimensional autoregressive generalized linear models," *arXiv preprint arXiv:1605.02693*, 2016.

[105] S. Kowshik, D. Nagaraj, P. Jain, and P. Netrapalli, "Near-optimal offline and streaming algorithms for learning non-linear dynamical systems," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 8518–8531. [Online]. Available: https://proceedings.neurips.cc/paper/2021/file/47a658229eb2368a99f1d032c8848542-Paper.pdf

[106] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan, "Stochastic Convex Optimization." in *COLT*, 2009.

[107] D. Paulin *et al.*, "Concentration inequalities for Markov chains by Marton couplings and spectral methods," *Electronic Journal of Probability*, vol. 20, 2015.

[108] Abbasi-Yadkori, Yasin and Pál, Dávid and Szepesvári, Csaba, "Online least squares estimation with self-normalized processes: An application to bandit problems," *arXiv preprint arXiv:1102.2670*, 2011.

[109] V. H. Peña, T. L. Lai, and Q.-M. Shao, *Self-normalized processes: Limit theory and Statistical Applications*. Springer Science & Business Media, 2008.

[110] P. Jain, S. S. Kowshik, D. Nagaraj, and P. Netrapalli, "Streaming Linear System Identification with Reverse Experience Replay," *arXiv preprint arXiv:2103.05896*, 2021.

[111] Z. Allen-Zhu, Y. Li, and Z. Song, "On the convergence rate of training recurrent neural networks," *arXiv preprint arXiv:1810.12065*, 2018.

[112] S. Bahmani and J. Romberg, "Convex programming for estimation in nonlinear recurrent models," *arXiv preprint arXiv:1908.09915*, 2019.

[113] S. Oymak, "Stochastic gradient descent learns state equations with nonlinear activations," in *Conference on Learning Theory*. PMLR, 2019, pp. 2551–2579.

[114] J. Miller and M. Hardt, "Stable recurrent models," *arXiv preprint arXiv:1805.10369*, 2018.

[115] H. Mania, M. I. Jordan, and B. Recht, "Active learning for nonlinear system identification with guarantees," *arXiv preprint arXiv:2006.10277*, 2020.

[116] A. Sarker, J. E. Gaudio, and A. M. Annaswamy, "Parameter Estimation Bounds Based on the Theory of Spectral Lines," *arXiv preprint arXiv:2006.12687*, 2020.

[117] Y. Mao, N. Hovakimyan, P. Voulgaris, and L. Sha, "Finite-Time Model Inference From A Single Noisy Trajectory," *arXiv preprint arXiv:2010.06616*, 2020.

[118] S. Kakade, A. T. Kalai, V. Kanade, and O. Shamir, "Efficient learning of generalized linear and single index models with isotonic regression," *arXiv preprint arXiv:1104.2018*, 2011.

[119] I. Diakonikolas, S. Goel, S. Karmalkar, A. R. Klivans, and M. Soltanolkotabi, "Approximation schemes for relu regression," in *Conference on Learning Theory*. PMLR, 2020, pp. 1452–1485.

[120] M. Simchowitz, R. Boczar, and B. Recht, "Learning linear dynamical systems with semi-parametric least squares," in *Conference on Learning Theory*. PMLR, 2019, pp. 2714–2802.

[121] T. Sarkar, A. Rakhlin, and M. A. Dahleh, "Finite-time system identification for partially observed lti systems of unknown order," *arXiv preprint arXiv:1902.01848*, 2019.

[122] S. Kowshik, D. Nagaraj, P. Jain, and P. Netrapalli, "Streaming linear system identification with reverse experience replay," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 30 140–30 152. [Online]. Available: https://proceedings.neurips.cc/paper/2021/file/fd2c5e4680d9a01dba3aada5ece22270-Paper.pdf

[123] A. T. Kalai and R. Sastry, "The Isotron Algorithm: High-Dimensional Isotonic Regression." in *COLT*. Citeseer, 2009.

[124] N. M. Boffi, S. Tu, and J.-J. E. Slotine, "The Reflectron: Exploiting geometry for learning generalized linear models," *arXiv preprint arXiv:2006.08575*, 2020.

[125] E. Hazan, A. Agarwal, and S. Kale, "Logarithmic regret algorithms for online convex optimization," *Machine Learning*, vol. 69, no. 2-3, pp. 169–192, 2007.

[126] Schraudolph, Nicol N and Yu, Jin and Günter, Simon, "A stochastic quasi-Newton method for online convex optimization," in *Artificial intelligence and statistics*. PMLR, 2007, pp. 436–443.

[127] S.-i. Amari, H. Park, and K. Fukumizu, "Adaptive method of realizing natural gradient learning for multilayer perceptrons," *Neural computation*, vol. 12, no. 6, pp. 1399–1409, 2000.

[128] L.-J. Lin, "Self-improving reactive agents based on reinforcement learning, planning and teaching," *Machine learning*, vol. 8, no. 3-4, pp. 293–321, 1992.

[129] Kowshik, Suhas S. and Polyanskiy, Yury, "Fundamental Limits of Many-User MAC With Finite Payloads and Fading," *IEEE Transactions on Information Theory*, vol. 67, no. 9, pp. 5853–5884, 2021.

[130] S. S. Kowshik, "Improved bounds for the many-user MAC," *arXiv preprint arXiv:2201.00866*, 2022.

[131] S. Aeron, V. Saligrama, and M. Zhao, "Information theoretic bounds for compressed sensing," *IEEE Transactions on Information Theory*, vol. 56, no. 10, pp. 5111–5130, 2010.

[132] Y. Jin, Y.-H. Kim, and B. D. Rao, "Limits on support recovery of sparse signals via multiple-access communication techniques," *IEEE Transactions on Information Theory*, vol. 57, no. 12, pp. 7877–7892, 2011.

[133] S. S. Kowshik, K. Andreev, A. Frolov, and Y. Polyanskiy, "Energy efficient coded random access for the wireless uplink," *IEEE Transactions on Communications*, vol. 68, no. 8, pp. 4694–4708, 2020.

[134] E. Biglieri, J. Proakis, and S. Shamai, "Fading channels: Information-theoretic and communications aspects," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2619–2692, 1998.

[135] T. S. Han, "An information-spectrum approach to capacity theorems for the general multiple-access channel," *IEEE Transactions on Information Theory*, vol. 44, no. 7, pp. 2773–2795, 1998.

[136] D. Guo and S. Verdú, "Randomly spread cdma: Asymptotics via statistical physics," *IEEE Transactions on Information Theory*, vol. 51, no. 6, pp. 1983–2010, 2005.

[137] G. Reeves and M. C. Gastpar, "Approximate sparsity pattern recovery: Information-theoretic lower bounds," *IEEE Transactions on Information Theory*, vol. 59, no. 6, pp. 3451–3465, 2013.

[138] A. Joseph and A. R. Barron, "Least squares superposition codes of moderate dictionary size are reliable at rates up to capacity," *IEEE Transactions on Information Theory*, vol. 58, no. 5, pp. 2541–2557, 2012.

[139] R. Venkataramanan, S. Tatikonda, and A. Barron, "Sparse regression codes," *Foundations and Trends® in Communications and Information Theory*, vol. 15, no. 1-2, pp. 1–195, 2019. [Online]. Available: http://dx.doi.org/10.1561/0100000092

[140] C. Rush and R. Venkataramanan, "The error probability of sparse superposition codes with approximate message passing decoding," *IEEE Transactions on Information Theory*, vol. 65, no. 5, pp. 3278–3303, 2018.

[141] C. Rush, K. Hsieh, and R. Venkataramanan, "Capacity-achieving Spatially Coupled Sparse Superposition Codes with AMP Decoding," *arXiv preprint arXiv:2002.07844*, 2020.

[142] J. Barbier and F. Krzakala, "Approximate message-passing decoder and capacity achieving sparse superposition codes," *IEEE Transactions on Information Theory*, vol. 63, no. 8, pp. 4894–4927, 2017.

[143] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2307–2359, 2010.

[144] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Quasi-static SIMO fading channels at finite blocklength," in *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*. IEEE, 2013, pp. 1531–1535.

[145] R. G. Gallager, *Information Theory and Reliable Communication*. New York, NY, USA: John Wiley & Sons, Inc., 1968.

[146] L. Birgé, "An alternative point of view on Lepski's method," *Lecture Notes-Monograph Series*, pp. 113–133, 2001.

[147] R. R. Bahadur, "A note on quantiles in large samples," *The Annals of Mathematical Statistics*, vol. 37, no. 3, pp. 577–580, 1966.

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

[149] X. Meng, S. Wu, L. Kuang, and J. Lu, "Concise derivation of complex bayesian approximate message passing via expectation propagation," *arXiv preprint arXiv:1509.08658*, 2015.

[150] Q. Zou, H. Zhang, C.-K. Wen, S. Jin, and R. Yu, "Concise derivation for generalized approximate message passing using expectation propagation," *IEEE Signal Processing Letters*, vol. 25, no. 12, pp. 1835–1839, 2018.

[151] A. Maleki, L. Anitori, Z. Yang, and R. G. Baraniuk, "Asymptotic analysis of complex lasso via complex approximate message passing (camp)," *IEEE Transactions on Information Theory*, vol. 59, no. 7, pp. 4290–4308, 2013.

[152] J. Barbier, "Statistical physics and approximate message-passing algorithms for sparse linear estimation problems in signal processing and coding theory," *arXiv preprint arXiv:1511.01650*, 2015.

[153] D. L. Donoho, A. Javanmard, and A. Montanari, "Information-theoretically optimal compressed sensing via spatial coupling and approximate message passing," *IEEE transactions on information theory*, vol. 59, no. 11, pp. 7434–7464, 2013.

[154] K. Hsieh, "Spatially Coupled Sparse Regression Codes for Single-and Multi-user Communications," Ph.D. dissertation, University of Cambridge, 2021.

[155] W. Van Zwet, "A strong law for linear functions of order statistics," *The Annals of Probability*, pp. 986–990, 1980.

[156] J. S. Rosenthal, *A first look at rigorous probability theory*. World Scientific Publishing Company, 2006.

[157] S. Verdú and S. Shamai, "Spectral efficiency of CDMA with random spreading," *IEEE Transactions on Information theory*, vol. 45, no. 2, pp. 622–640, 1999.

[158] J. Salo, D. Seethaler, and A. Skupch, "On the asymptotic geometric mean of MIMO channel eigenvalues," in *2006 IEEE International Symposium on Information Theory*. IEEE, 2006, pp. 2109–2113.

[159] R. Zamir and M. Feder, "A generalization of the entropy power inequality with applications," *IEEE transactions on information theory*, vol. 39, no. 5, pp. 1723–1728, 1993.

[160] W. Li and Y. Chen, "Some new two-sided bounds for determinants of diagonally dominant matrices," *Journal of Inequalities and Applications*, vol. 2012, no. 1, p. 61, 2012.

[161] S. Verdú, "Total variation distance and the distribution of relative information," in *2014 Information Theory and Applications Workshop (ITA)*. IEEE, 2014, pp. 1–3.

[162] Y. Polyanskiy and Y. Wu, "Lecture notes on Information Theory," 2017. [Online]. Available: http://people.lids.mit.edu/yp/homepage/data/itlectures_v5.pdf

[163] A. Collins, G. Durisi, T. Erseghe, V. Kostina, J. Östman, Y. Polyanskiy, I. Tal, and W. Yang, "SPECTRE: Short Packet Communication Toolbox," https://github.com/yp-mit/spectre, 2016.

[164] L. A. Shepp and I. Olkin, "Entropy of the sum of independent bernoulli random variables and of the multinomial distribution," in *Contributions to probability*. Elsevier, 1981, pp. 201–206.

[165] C. Knessl, "Integral representations and asymptotic expansions for Shannon and Renyi entropies," *Applied mathematics letters*, vol. 11, no. 2, pp. 69–74, 1998.

[166] B. Lindström, "On a combinatorial problem in number theory," *Canadian Mathematical Bulletin*, vol. 8, no. 4, pp. 477–490, 1965.

[167] D. G. Cantor and W. Mills, "Determination of a subset from certain combinatorial properties," *Canadian Journal of Mathematics*, vol. 18, pp. 42–48, 1966.

[168] G. H. Khachatrian and S. S. Martirossian, "A new approach to the design of codes for synchronous-CDMA systems," *IEEE Transactions on Information Theory*, vol. 41, no. 5, pp. 1503–1506, 1995.

[169] G. Reeves and H. D. Pfister, "The replica-symmetric prediction for random linear estimation with gaussian matrices is exact," *IEEE Transactions on Information Theory*, vol. 65, no. 4, pp. 2252–2283, 2019.

[170] J. Barbier, M. Dia, N. Macris, and F. Krzakala, "The mutual information in random linear estimation," in *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2016, pp. 625–632.

[171] J. Barbier, F. Krzakala, M. Mézard, and L. Zdeborová, "Compressed sensing of approximately-sparse signals: Phase transitions and optimal reconstruction," in *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2012, pp. 800–807.

[172] R. Vicente, D. Saad, and Y. Kabashima, "Low-density parity-check codes: statistical physics perspective," in *Advances in imaging and electron physics*. Elsevier, 2003, vol. 125, pp. 231–353.

[173] S. Kirkpatrick and B. Selman, "Critical behavior in the satisfiability of random boolean expressions," *Science*, vol. 264, no. 5163, pp. 1297–1301, 1994.

[174] G. Reeves, J. Xu, and I. Zadik, "The all-or-nothing phenomenon in sparse linear regression," *arXiv preprint arXiv:1903.05046*, 2019.

[175] J. C. Duchi, A. Agarwal, M. Johansson, and M. I. Jordan, "Ergodic Mirror Descent," *SIAM Journal on Optimization*, vol. 22, no. 4, pp. 1549–1578, 2012. [Online]. Available: https://doi.org/10.1137/110836043

[176] A. Défossez and F. Bach, "Averaged least-mean-squares: Bias-variance trade-offs and optimal sampling distributions," in *Artificial Intelligence and Statistics*. PMLR, 2015, pp. 205–213.

[177] F. Petrov, "Non-asympototic version of Gelfand's formula," MathOverflow, 2016. [Online]. Available: https://mathoverflow.net/q/228561

[178] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press, 2018, vol. 47.

[179] J. D. Hamilton, *Time series analysis*. Princeton university press, 1994.

[180] R. S. Sutton, A. G. Barto *et al.*, *Introduction to reinforcement learning*. MIT press Cambridge, 1998, vol. 135.

[181] Bousquet, Olivier and Elisseeff, André, "Stability and generalization," *The Journal of Machine Learning Research*, vol. 2, pp. 499–526, 2002.

[182] M. Hardt, B. Recht, and Y. Singer, "Train faster, generalize better: Stability of stochastic gradient descent," in *International Conference on Machine Learning*. PMLR, 2016, pp. 1225–1234.

[183] D. Nagaraj, P. Jain, and P. Netrapalli, "SGD without replacement: Sharper rates for general smooth convex functions," in *International Conference on Machine Learning*. PMLR, 2019, pp. 4703–4711.

[184] A. Dieuleveut, A. Durmus, F. Bach *et al.*, "Bridging the gap between constant step size stochastic gradient descent and markov chains," *Annals of Statistics*, vol. 48, no. 3, pp. 1348–1382, 2020.

[185] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," *arXiv preprint arXiv:1011.3027*, 2010.

[186] ——, "High-dimensional probability," 2019.