From Spectral Theorem to Spectral Statistics of Large Random Matrices with Spatio-Temporal Dependencies

by

Muhammad Abdullah Naeem

Department of Electrical and Computer Engineering
Duke University

Date: _____

Approved:

_____
Miroslav Pajic, Supervisor

_____
Michael Zavlanos, Co-Chair

_____
Henri Gavin

_____
Leila Bridgeman

_____
Galen Reeves

Dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Department of Electrical and Computer Engineering
in the Graduate School of
Duke University

2023

ABSTRACT

From Spectral Theorem to Spectral Statistics of Large Random Matrices with Spatio-Temporal Dependencies

by

Muhammad Abdullah Naeem

Department of Electrical and Computer Engineering
Duke University

Date: _____

Approved:

_____
Miroslav Pajic, Supervisor

_____
Michael Zavlanos, Co-Chair

_____
Henri Gavin

_____
Leila Bridgeman

_____
Galen Reeves

An abstract of a dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Department of Electrical and Computer Engineering
in the Graduate School of
Duke University

2023

Abstract

Overarching theme of this thesis is new theoretical results on spectral theorem for non-Hermitian operators, non-asymptotic behavior of high dimensional dynamical systems , which we incorporate with the work of Talagrand on concentration of measure phenomenon to better understand spectral behavior of the structured random matrices(data matrix) and subsequently the performance of regression based algorithms with dependent data. All the estimates in this thesis are explicit in n, dimension of the state space and N, length of the simulated trajectory. Non-asymptotic spectral analysis developed in this thesis have lead following interesting findings:

1. Learning for dynamical systems: least squares estimation for even stable dynamical systems can have non-vanishing error in high dimensions.

2. Functional inequalities: there exists a stable ARMA model with process level Talagrands' inequality exponential in dimension of the underlying state space

This thesis is based upon 4 articles written by the author:

- Naeem, Muhammad Abdullah. "Concentration Phenomenon for Random Dynamical Systems: An Operator Theoretic Approach." Learning for Dynamics and Control Conference. PMLR, 2023.

- Naeem, Muhammad Abdullah, and Miroslav Pajic. "Transportation-inequalities, lyapunov stability and sampling for dynamical systems on continuous state space." Learning for Dynamics and Control Conference. PMLR, 2023.

- Naeem, Muhammad Abdullah, Amir Khazraei, and Miroslav Pajic. "From Spectral Theorem to Statistical Independence with Application to System Identification." arXiv preprint arXiv:2310.10523 (2023).

- Naeem, Muhammad Abdullah, and Miroslav Pajic. "Spectral Statistics of the Sample Covariance Matrix for High Dimensional Linear Gaussians." arXiv preprint arXiv:2312.05794. 2023 Dec 10.

Contents

List of Figures

Acknowledgements

With regard to my time at Duke University, I am grateful to my adviser Miroslav Pajic: for the freedom he allowed me to work on the problem of my interest and his patience with me. I would like to offer my thanks to my committee members, ECE administration and the faculty that I have had the opportunity to take a course with and my lab colleagues.

My deepest gratitude goes out to my parents, Fatima and Naeem, and my siblings: Khadija, Hamza and Ibrahim for their unconditional love and support. I will always be indebted to my guide, Sir Asif Ameen for his wisdom and spiritual insights. I have been extremely fortunate to have friends like: Imman, Nourang, Waris and Nomi, who have assisted and supported me throughout my life trajectory.

Chapter 1

Introduction

High dimensional dynamical systems are ubiquitous, including-but not limited to-cyber-physical systems, daily return on different stocks of S&P 1500 and velocity profile of interacting particle systems around McKeanVlasov limit. A quantitative analyst trying to come up with investment strategies to maximize profit in the stock market and an engineer trying to land a rocket on the moon while minimizing some certain predefined cost, heavily rely on availability of accurate system models. This brings us to initial part of this thesis; we are concerned with performance of ordinary least squares(OLS) method for the *estimation of high dimensional state transition matrix* A from a single noisy observed trajectory of the linear time invariant(LTI)[1] system $(x_0, x_1, \ldots, x_N)$, i.e.,

$$x_{t+1} = Ax_t + w_t, \quad \text{where } w_t \sim N(0, I_n). \tag{1.1}$$

We work with the standard assumption of spectral radius of state transition matrix:$\rho(A)$ being strictly inside the unit circle so as to ensure stationarity(atleast eventually). This is a standard time series learning problem and it is known by the name of autoregressive-moving-average model(ARMA), common to fields like econometrics and finance. Advances in high dimensional statistics have lead to a recent surge in non-asymptotic analysis of least squares regression on Markovian data(see e.g., [1],[2],[3]) and their performance is essentially dictated by spectral properties of the data matrix $X_- = [x_0, x_1, \ldots, x_{(N-1)}]$ and Gaussian ensemble $E = [w_0, w_1, \ldots, w_{N-1}]$,

[1]Linear Gaussian (LG) in Markov chain literature

where X_- is itself a function of E (dictated by $(A^k)_{k=0}^N$). Existing literature classify dynamical systems merely based on spectral radius and even base sample complexity of various learning algorithms solely based on it. However, along with temporal correlations between $N-$ observed samples-each of length n, high dimensionality of space allows for various choices of spatial interactions dictated by *algebraic and geometric spectral content of* A and consequently non-trivial interactions between nN-one dimensional observed covariates. Major goal of this thesis is to ensure all the errors, system parameters e.t.c do not hide any dependencies in n or N(i.e., non-asymptotic analysis in truest spirit). All the theory in the first part of the thesis can be easily followed by any one with decent exposure to linear algebra and knowledge of Talagrands inequality. Instead of tedious probabilistic methods we focus on operator theoretic methods which are extremely concise and have no whatsoever susceptibility to confusion.

1.1 Summary of the chapters

Main themes discussed in the thesis are as follows:

- **Chapter one:** We review various related work in the literature followed by notations and preliminaries: where we provide a brief introduction to basis-free approach to Linear Transformation. After introducing an inner product structure on domain and image space of Linear transformation, adjoint operator is also introduced. Followed by probabilistic side of preliminaries, emphasis is put towards familiarizing the reader to concentration of measure of phenomenon; particularly important is *dimension independent tensorization of Talagarands' inequality* for norm stable diagonalizable dynamical systems with Gaussian excitations. We will eventually show that exact, elementwise error in OLS is a

weighted sum of standard normals and its' concentration behavior is studied in literature under the well-known Littlewood-Offord problem. Therefore, we briefly highlight their findings in particular how concentration behvaior can depend on structure of weigths

- **Chapter two:** Here we raise the following question: given dimensions of the underlying state space n and a constant $\rho \in (0, 1)$, provide uniform bounds w.r.t dimension of underlying state space for the smallest $\Gamma(n, \rho) \in \mathbb{N}$ such that any linear transformation $A \in \mathbb{C}^{n\times n}$ with spectral radius $\rho(A)$ being equal to ρ satisfy $\|A^k\| < 1$ for all $k > \Gamma(n, \rho)$. It allows us to link the spectrum of A to spatial and statistical dependence between the rows of X_-.

- **Chapter three:** We begin with showing that least squares performance is intimately linked to dependence between the basis(on which we project our observations). In the dynamic version of OLS these basis are provided by the rows of the data matrix X_- and A^* is estimated by projecting the observations $X_+ := [x_1, x_2, \ldots, x_N]$ onto the rows of the data matrix. We provide geometric and spectral approximations(based on extreme singular values of the data matrix) to estimation error in Frobenius norm. As suspected, even when dynamics are stable but spatially inseparable, OLS seems to be inconsistent in high dimension($n \geqslant 10$, as increasing iterations only worsen the error as shown in simulations). We conclude this chapter by showing that elementwise estimation error is essentially a higher degree variant of the famous Littlewood-Offord problem.

- **Chapter four:** We demystify the order of various spectral statistics(mentioned in chapter three) which approximates estimation error, explicitly in dimensionality of the state space n and samples N in the data matrix. It is important

to mention that smallest singular value is essential for upper bounding the estimation error whereas the behavior of largest singular value is imperative to conclude when OLS contains non-vanishing error. This chapter is at the confluence of concentration of measure phenomenon and spectral theory. In this chapter, answer two of the previously open problems is provided: i) Exponential sample complexity for learning state transition matrix for data data generated from purely a single Jordan block corresponding to any stable eigenvalue, raised in[4] and ii) dimension+iteration dependent tensorization of Talagrands' inequality for stable ARMA model, see e.g., ([5] and [6]).

- **Chapter five**: is abstract in nature and aims at understanding space-time decay of correlations for non-linear random dynamical systems, we also propose Lyapunov type conditions that dictate convergence to stationary distribution for underling Markov Chain.

1.2 Summary of the contributions

Highlight contribution of this thesis is on non-asymptotic analysis of

- **Quantiative handle on decay of $\|A^k\|$:** We provide first quantiative handle on decay of $\|A^k\|$ using using spectral theorem for non-Hermitian finite dimensional linear transformations and equipping underlying space with inner product structure. Position of the eigenvalues and characteristic polynomial only correspond to algebraic structure of Linear operator; Geometric structure follows from the span of associated eigenvectors. Given each distinct eigenvalue of A, discrepancy between its' algebraic and geometric multiplicity leads to an $A-$ invariant subspace with dimension equal to size of the discrepancy. By re-

stricting A onto each invariant subspace via projection operator and computing the norm of the powers of A on the underlying subspace, allows us to control the overall norm of the powers of A. We would like to highlight that our result negates the common misconception of operator norm being sensitive to choice of basis.

- **Spectral theorem to statistical independence:** The mentions of invariant subspaces and orthogonal projections must have already hinted a high dimensional statistician that using Gaussian projection lemma, small discrepancies imply that original trajectory essentially comprises of multiple *lower dimensional random dynamical systems living on A invariant subspaces and are statistically independent of each other*: this result in itself has far reaching consequences in design and analysis of large scale, highly inter-connected dynamical systems. We show that $\Gamma(n,\rho) = O\left(\frac{(n-1)\ln n}{\ln \frac{1}{\rho}}\right)$, and in fact our bounds are tight up to $\ln n$ factor. As a simple corollary, mixing time of (1.1) (viewed as a Markov chain) in Wasserstein distance turns out to be indeed $O\left(\frac{(n-1)\ln n}{\ln \frac{1}{\rho}}\right)$. It is shown that when a stable dynamical system has only one distinct eigenvalue and discrepancy of $n-1$: $\|A\|$ has a dependence on n, resulting dynamics are *spatially inseparable*

- **Spectral statistics of the sample covariance matrix**: Performance of OLS estimator heavily rely on negative moments of the sample covariance matrix: $(X_-X_-^*) = \sum_{i=0}^{N-1} x_i x_i^*$ and singular values of EX_-^*, where E is a rectangular Gaussian ensemble $E = [w_0, \ldots, w_{N-1}]$. Negative moments requires sharp estimates on all the eigenvalues $\lambda_1\big(X_-X_-^*\big) \geqslant \ldots \geqslant \lambda_n\big(X_-X_-^*\big) \geqslant 0$. Leveraging upon recent results on spectral theorem for non-Hermitian operators in [7]), along with concentration of measure phenomenon and perturba-

tion theory(Gershgorins' and Cauchys' interlacing theorem) we show that only when $A = A^*$, typical order of $\lambda_j\big(X_-X_-^*\big) \in \big[N - n\sqrt{N}, N + n\sqrt{N}\big]$ for all $j \in [n]$. However, in *high dimensions* when A has only one distinct eigenvalue λ with geometric multiplicity of one, then as soon as eigenvalue leaves *complex half unit disc*, largest eigenvalue suffers from curse of dimensionality: $\lambda_1\big(X_-X_-^*\big) = \Omega\big(\lfloor\frac{N}{n}\rfloor e^{\alpha_\lambda n}\big)$, while smallest eigenvalue $\lambda_n\big(X_-X_-^*\big) \in (0, N+\sqrt{N}]$. Consequently, OLS estimator incurs a *phase transition* and becomes *transient: increasing iteration only worsens estimation error*, all of this happening when the dynamics are generated from stable systems.

- **Exponential in underlying dimension, tensorization of Talagrands inequality and non-vanishing estimation error:** Dimension and iteration dependence of Talgrands' inequality for for $n-$ dimensional stable ARMA models had been a subject to various speculations, [5] and [6]. We show that for sufficiently large length N of simulated trajectory, $n-$ dimensional stable ARMA trajectory corresponding to spatially inseparable case satisfies process level Talagrands' inequality with constant $\Theta(e^n)$. This is simply follows by realizing that Talagrands' constant is equal to the ratio of Frobenious norm of data matrix X_- and Gaussian ensemble E. It is well know that $\|E\|_F = \Theta(\sqrt{nN})$ but as we have already shown $\sigma_1(X_-) = \Theta(\sqrt{N-n+1}e^n)$. A simple corollary now reveals trace of sample covariance matrix, $=\Theta(e^n\sqrt{Nn})$ and we finally show that estimation error in Frobenius norms is

$$\Omega\bigg(\frac{(\sqrt{N}-\sqrt{n-1})\sqrt{nN}}{(N-n+1)}\bigg) > 0. \tag{1.2}$$

1.3 Previous work

Successful completion of any control task, relies heavily on availability of accurate models for underlying dynamical system. Owing to advances in machine learning and high dimensional statistics, when physics based models are too complicated or do not really exist to begin with as in the case of S & P 1500, one attempts to use some sort of regression models on an observed trajectory of underlying dynamical system of length N and provide a non-asymptotic high probability guarantees on learning its' correct models. Said in another way, given an $n \times N$ dimensional data matrix which captures evolution over time of $n-$ dimensional random dynamical system, goal of system identification is to extract 'useful' information about underlying system.

To name a few recent results: [8] provides non-asymptotic analysis on learning system parameters (state transition matrix, input-to-state transformation, etc) of a linear time invariant (LTI) system from a single observed trajectory of input-output data. [3],[9] and [10] provides statistical analysis on learning the state transition matrix via Ordinary Least Squares (OLS) estimator on an observed trajectory $(x_t)_{t=1}^N$, where dynamics follow $x_{k+1} = Ax_k + w_k$, dimension of the underlying state space is n and w_k is isotropic Gaussian, [11] provides non-asymptotic analysis on learning value function corresponding to closed loop stable policy for a Linear System and [12] provides non-asymptotic analysis for online Kalman filtering.

Surprisingly, their has been no reference to how dimensionality of underlying space plays out in learning tasks. As additive randomness is usually assumed to be isotropic, how does spectrum of underlying linear transform translates into spatio-temeperal correlations of the covariates represented in the data matrix. The gap might have stemmed from the fact that these results classify dynamical systems only based on the magnitude of their eigenvalues, which suffices if the underlying

state-transition matrix is Hermitian. To fully characterize any linear transformations whether or not it is stable in control theoretic sense, one needs to take into account geometric part of the spectrum as well i.e., span of eigenvectors. Furthermore, as high probability guarantees in majority of recent work requires sufficient length of simulated trajectory to be lower bounded by various Grammians(essentially higher degree polynomial of A) and prior to this work uncertaintly loomed over *sensitivity of operator norm to similarity transforms*-so it does not come as much of a surprise why these dependencies on dimensionality had been missing and spatial interactions were never before even a part of discussion.

On the quantiative side of things, regardless of some minor technicalities, starting point in all these problems assume that via some oracle we are given dimension of the underlying state space n and spectral radius $\rho(A) < 1$ (i.e., underlying dynamical system is stable) then a regression algorithm is proposed and if the length of simulated trajectory is of order $\log\left(\frac{1}{\rho(A)}\right)$ then the desired accuracy is achieved. A common theme in all these works is controlling decay of power of $\|A^k\|$ via Gelfands' formula. [8] learns systems Markov Parameters up till time T such that $\sum_{\tau=T}^{\infty} \|A^\tau\| = O(1)$ and then estimate system parameters using Ho-Kalman algorithm [13]. Pretty much exclusively, all these result argue along the same lines [3, 8, 11]: if $\rho(A) < 1$ then $\|A^k\| \leqslant \text{poly}(k)\rho(A)^k$ ($\text{poly}(k)$ is used to denote polynomial function of k), and using tools from high dimensional statistic derive bounds on the length of the trajectory in terms of spectral radius, various Gramians related to system dynamics for the desired accuracy as discussed in preceding paragraph. However, Gelfands' formula is in fact true for infinite dimensional Hilbert spaces as well, and we are dealing with finite dimensional linear transformations which leaves a lot of room for imporvement in decay rate of A^k.

To the best of authors' knowledge, only very recently, similar questions appeared

in [4] and [14], where they posed the question of what kind of linear systems are hard to learn and control and also provided with real world examples. However they could not mathematically classify such systems and complexity of learning was given itself as a function of operator norm of the the linear transform. Unfortunately, their analysis only works under the assumption that operator norm is independent of dimensionality of the state space, which we will show is not true for spatially inseparable stable systems. On the quantitative side of things, in finite sample analysis of temporal difference learning for value function corresponding to a closed loop stable policy given in [11] quantifies the decay of the operator norm via $\mathcal{H}_\infty$ norm of the resolvent of state transition matrix, without any interpretability: define $\Phi_A(z) := (zI - A)^{-1}$ for complex numbers z then their result claims for any $\rho \in (\rho(A), 1)$ and for all $k \in \mathbb{N}$, $\|A^k\| \leqslant \sup_{z\in\mathbb{C}:|z|=1} \|\Phi_{\rho^{-1}A}(z)\|\rho^k$, but using Neumann series expansion of resolvent $\|\Phi_{\rho^{-1}A}(z)\| = \|\frac{1}{z}\sum_{l=0}^{\infty}(\frac{A}{z\rho})^l\|$ so their result essentially translates to $\|A^k\| \leqslant \sup_{z\in\mathbb{C}:|z|=1} \|\frac{1}{z}\sum_{l=0}^{\infty}\left(\frac{A}{z\rho}\right)^l\|\rho^k$ i.e., their quantitative handle controls norm of $\|A^k\|$ by a polynomial of all the finite powers of A weighted by some constant factor. [3] provides control on norm of $\|A^k\|$ by computing norm of the associated Jordan block but these estimates require norm bounds on associated similarity transformation and its' inverse(we will show is not necessary) and do not offer any interpretation into how size of the Jordan blocks translate into operator norm of powers of A.

For various variants of regression problem related to dynamical systems, existing literature focuses exclusively on upper bounding operator norm by some martingale term and showing least singular value of the data matrix is lower bounded with high probability. Concentration behavior of martingale terms and various quadratic forms in literature are studied using *Hanson Wright inequality*, which shows deviation of quadratic form based on Frobenius and operator norm of the weights defining it and

high probability estimates are given as a function of these norms. As we show in Section 4.3.1, while studying the concentration behavior of distance between a fixed $n-1$ dimensional subspace of $\mathbb{R}^N$ and a trajectory of length N from one dimensional ARMA model, that distance function is essentially a quadratic form with Frobenius and operator norm of its weights having potential dependence on N (number of iterations) and n(dimension of underlying state space). Hidden dependence on dimensionality of state space and number of iterations may suggest existing upper bounds are vacuous. Furthermore, lower bounding the error would require quantifying largest singular value of the data matrix, which had not been explored before this work.

These limitations suggest we break down finite sample analysis problem into two components, dynamics part and probabilistic part. So, we take an initial step in developing a non-asymptotic, geometric and interpret-able version of systems theory that can explain dynamical evolution of high dimensional systems by studying (i) quantitative decay of A^k, with explicit dependence in dimensionality of the state space and the number of iterations, k. In order to get this handle we first had to qualitatively differentiate linear transformations i.e., ones' with strong or weak spatial couplings. We then combine our results with findings of high dimensional geometry/ statistics to collect (ii) various estimates(explicit in n and N) that will be useful for complete understanding of least squares regression when trying to estimate high dimensional LTI systems from a single observed trajectory and in fact show qualitatively different behavior of LTI systems(with same eigenvalues) lead to qualitative different behavior of regression.

Instead of naively applying results from high dimensional statistics and random matrix theory(which were developed to cater i.i.d random variable setting), we initially focused on developing non-asymptotic version of mathematical systems the-

ory and then combined it with results from high dimensional geometry/statistics to conclude previously unknown result like: *Spectral Theorem combined with Gaussian projection lemma implies that data matrix can be decomposed into low dimensional random dynamical systems which are statistically independent of each other(modulo spatially inseparable systems).*

1.4 Notation and Preliminaries

Notation

We use $I_n \in \mathbb{R}^{n\times n}$ to denote the n dimensional identity matrix. $B_\alpha^n := \{x \in \mathbb{R}^n : \|x\| := \|x\|_2 < \alpha\}$ is the open α-ball in $\mathbb{R}^n$. Similarly, $\mathcal{S}^{n-1} := \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$ is the unit sphere in $\mathbb{R}^n$ and $\mathbb{T}$ denotes unit circle in complex plane. $\rho(A)$, $\|A\|_2$, $\|A\|_F$, $det(A)$, $tr(A)$ and $\sigma(A)$ represent the spectral radius, matrix 2-norm, Frobenius norm, determinant, trace and set of eigenvalues(spectrum) of A, respectively. When, subscript under norm is not specified, automatically matrix 2-norm is assumed. For a positive definite matrix A, largest and smallest eigenvaues are denoted by $\lambda_{max}(A)$ and $\lambda_{min}(A)$, respectively. Associated with every rectangular matrix $F \in \mathbb{R}^{N\times n}$ are its' singular values $\sigma_1(F) \geqslant \sigma_2(F), \ldots, \sigma_n(F) \geqslant 0$, where without loss of generality we assume that $N > n$. A variational characterization of each singular values $\sigma_k(F)$ follows from Courant-Fischer:

$$\begin{aligned}\sigma_k(F) &= \max_{V\subset\mathbb{R}^n:dim(V)=k} \min_{x\in V\cap\mathcal{S}^{n-1}} \|Fx\| \\ &= \min_{V\subset\mathbb{R}^n:dim(V)=n-k+1} \max_{x\in V\cap\mathcal{S}^{n-1}} \|Fx\|\end{aligned}$$

A function $g : \mathbb{R}^n \to \mathbb{R}^p$ is Lipschitz with constant L if for every $x, y \in \mathbb{R}^n$, $\|g(x) - g(y)\| \leqslant L\|x - y\|$. Notations like O, Θ and Ω will be used to highlight the

dependence (*non-asymptotically*) w.r.t number of iterations N and dimensionality of the state space n. If a statement is only asymptotically true we will highlight it separately.

Space of probability measure on $\mathcal{X}$(continuous space) is denoted by $\mathcal{P}(\mathcal{X})$ and space of its Borel subsets is represented by $\mathbb{B}\big(\mathcal{P}(\mathcal{X})\big)$. On a metric space $(\mathcal{X}, d)$, for $\mu, \nu \in \mathcal{P}(\mathcal{X})$, we define Wasserstein metric of order $p \in [1, \infty)$ as

$$W_p^d(\nu, \mu) = \left(\inf_{(X,Y)\in\Gamma(\nu,\mu)} \mathbb{E}\ d^p(X, Y) \right)^{\frac{1}{p}}; \tag{1.3}$$

here, $\Gamma(\nu, \mu) \in P(\mathcal{X}^2)$, and $(X, Y) \in \Gamma(\nu, \mu)$ implies that random variables (X, Y) follow some probability distributions on $P(\mathcal{X}^2)$ with marginals ν and μ. Another way of comparing two probability distributions on $\mathcal{X}$ is via relative entropy, which is defined as

$$H(\nu||\mu) = \begin{cases} \int \log\left(\frac{d\nu}{d\mu}\right) d\nu, & \text{if } \nu << \mu, \\ +\infty, & \text{otherwise.} \end{cases} \tag{1.4}$$

1.4.1 Finite dimensional linear transformations: basis-free approach

Given two finite dimensional vector spaces $\mathcal{V}$ and $\mathcal{W}$ with field $\mathbb{F}$ (can be $\mathbb{R}$ or $\mathbb{C}$ depending on the specific problem), $A \in L(\mathcal{V}, \mathcal{W})$ means that A is linear transformation from $\mathcal{V}$ to $\mathcal{W}$. Given set of vectors $[v_i]_{i=1}^k \in \mathcal{V}$, we say that $v \in$span$[v_1, v_2, \ldots, v_m]$ if there exists constants $[a_i]_{i=1}^k$ such that:

$$v = a_1 v_1 + a_2 v_2 + \ldots + a_k v_k, \ \ , \textit{where each } a_i \in \mathbb{F}. \tag{1.5}$$

As in [15] chapter 1, on a finite dimensional vector space $\mathcal{V}$, we define its *dimension*, denoted by $dim(\mathcal{V})$, as the smallest number n, such that there exists vectors $[v_i]_{i=1}^n$ such that:

$$span[v_1, v_2, \ldots, v_n] = \mathcal{V}. \tag{1.6}$$

If this is true, then $[v_i]_{i=1}^n$ is a *basis* for $\mathcal{V}$, and automatically satisfy *linear independence*:

$$a_1 v_1 + a_2 v_2 + \ldots + a_n v_n = 0, \tag{1.7}$$

then $a_i = 0$ for all $i \in [1, 2, \ldots, n]$. If $\mathcal{V}$ and $\mathcal{W}$ are finite dimensional vector space with complex field, they can be can be equipped with an *inner product structure* i.e.,

$$\begin{aligned}
&\|v\|_{\mathcal{V}} := \langle v, v\rangle_{\mathcal{V}} \geqslant 0, \textit{ for all, } v \in \mathcal{V} \textit{ and } 0 \iff v = 0 \\
&\langle x_1 + x_2, y\rangle_{\mathcal{V}} = \langle x_1, y\rangle_{\mathcal{V}} + \langle x_2, y\rangle_{\mathcal{V}}. \\
&\langle x, \alpha y\rangle = \overline{\alpha}\langle x, y\rangle_{\mathcal{V}}, \textit{ where } \overline{\alpha} \textit{ means complex conjugate of } \alpha \\
&\overline{\langle x, y\rangle_{\mathcal{V}}} = \langle y, x\rangle_{\mathcal{V}}.
\end{aligned} \tag{1.8}$$

After defining a similar inner product structure on $\mathcal{W}$, one can define an *adjoint transformation* $A^* \in L(\mathcal{W}, \mathcal{V})$ such that.

$$\langle Av, w\rangle_{\mathcal{W}} = \langle v, A^* w\rangle_{\mathcal{V}}, \tag{1.9}$$

$$\|A\|_{L(\mathcal{V},\mathcal{W})} := \sup_{\|v\|_{\mathcal{V}}=1} \|Av\|_{\mathcal{W}} = \|A^*\|_{L(\mathcal{W},\mathcal{V})}, \tag{1.10}$$

$$\|AA^*\|_{L(\mathcal{W},\mathcal{W})} = \|A^*A\|_{L(\mathcal{V},\mathcal{V})} = \|A\|^2_{L(\mathcal{V},\mathcal{W})}, \tag{1.11}$$

see e.g., chapter 2 in [16]. Moreover, A can be viewed as the following map(we will

discuss the formal procedure in Section 2.1):

$$A : \underbrace{\underbrace{N(A) \oplus N(A)^{\perp}}_{\mathcal{V}}}_{} \rightarrow \underbrace{\underbrace{Im(A) \oplus Im(A)^{\perp}}_{\mathcal{W}}}_{}, \tag{1.12}$$

where $A : N(A)^{\perp} \rightarrow Im(A) = \{w \in \mathcal{W} : \exists v \in \mathcal{V} : Av = w\}$ is bijective and $N(A) := \{v \in \mathcal{V} : Av = 0\}$ is the null space of A. Similarly, $A^* : Im(A) \rightarrow N(A)^{\perp}$ is bijective and $N(A^*) = Im(A)^{\perp}$.

Definition 1. *We say a linear transformation A, from vector space $\mathcal{V}$ to $\mathcal{W}$ is* full rank *if $Im(A) = \mathcal{W}$.*

This brings us to a very important observation that will help us in improving learning rate for state transition matrices via OLS:

Lemma 1. *Given a* full rank *$A \in L(\mathcal{V}, \mathcal{W})$, where $1 \leqslant dim[\mathcal{W}] = n' < dim[\mathcal{V}] = n$. Then:*

1. $\left(AA^*\right)^{-1} \in L(\mathcal{W}, \mathcal{W})$,
2. $dim[N(A)^{\perp}] = n'$, $dim[N(A)] = n - n'$
3. $A^*\left(AA^*\right)^{-1} \in L(\mathcal{W}, N(A)^{\perp})$, *is a bijection and given $N(A)$ one can construct a matrix version of $A^*\left(AA^*\right)^{-1}$ by padding $n - n'$ rows of zeros to an $n' \times n'$ matrix*

Proof. First two results are obvious from preceding discussion. Let $[v_1, v_2, \ldots, v_{(n-n')}]$ be linearly independent vectors that span $N(A)$. We know that image space of $A^*\left(AA^*\right)^{-1}$ should be orthogonal to $N(A)$,, so we can come up with vectors $[a_1 \ldots a_{n'}]$ whose span is $Im[A^*\left(AA^*\right)^{-1}]$ with only restriction of orthogonality w.r.t v_i's i.e. for

each $j \in [1, 2, \ldots, n']$

$$\langle v_i, a_j \rangle = 0 \ \ \forall i \in [1, n - n']. \tag{1.13}$$

Let $a_j(n'+1) = \ldots a_j(n) = 0$, then we have n' equations in (1.13) with n' unknowns and the claim follows. □

1.4.2 Concentration of Measure and Littlewood-Offord problem

Let $(x_i)_{i=1}^N$ be centered and bounded size on average i.e., variance of $O(1)$, then one would expect $S_N := \sum_{i=1}^N x_i$ to vary in an interval of $O(N)$. However, *under sufficient independence assumption* between each individual components $x_1, x_2, \ldots, x_N$, the sum concentrates in a much narrower interval of size $O(\sqrt{N})$ i.e., $P(|S_N| \geqslant \delta\sqrt{N}) = O(\frac{1}{\delta^2})$. This is because each individual variable to collectively vary in a way to produce deviation of $O(N)$ becomes more and more less likely with increase in variables and this remarkable phenomenon is called *concentration of measure*

In fact, trajectory from *one-dimensional stable ARMA* model, with $x_0 = 0$:

$$x_{i+1} = \lambda x_i + w_i, \ w_i \sim N(0, 1) \tag{1.14}$$

for some $|\lambda| < 1$, w_i and w_j are independent for $i \neq j$, also satisfies this phenomenon, precisely:

$$P\left(|S_N| \geqslant \delta\sqrt{\frac{N - \sum_{i=1}^N |\lambda|^{2i}}{1 - |\lambda|^2}}\right) = O\left(\frac{1}{\delta^2}\right) \tag{1.15}$$

This phenomenon is not just limited to deviations of sums and can be extended to smooth functions of 'sufficiently independent' random variables via Talagrands' Inequality:

Definition 2. *We say that the probability measure μ satisfies the L_p-transportation cost inequality on $(\mathcal{X}, d)$ if there is some constant $C > 0$ such that for any probability measure ν*

$$\mathcal{W}_p^d(\nu, \mu) \leqslant \sqrt{2CH(\nu||\mu)}, \tag{1.16}$$

and we use $\mu \in T_p^d(C)$ as a shorthand notation.

Since for $p > 1$, $T_p^d(C)$ implies $T_1^d(C)$ which is related to concentration of Lipschitz functions:

Remark 1. *[Theorem 1.1 in [6]: $T_1^d(C)$ is equivalent to Lipschitz function concentration] Given any Lipschitz function f on random variable x with underlying distribution $\mu \in T_1^d(C)$, we have:*

$$\mathbb{P}\left[\left|f(x) - < f >_\mu \right| > \epsilon\right] \leqslant 2\exp\left(-\frac{\epsilon^2}{2C\|f\|^2_{L(d)}}\right). \tag{1.17}$$

As we will be interested in deviations and typical behaviors of random processes, like trajectories following some Markovian dynamics e.t.c, it is important to understand how it varies with number of iterations and dimension of the underlying space, a useful result is:

Theorem 2. *[Theorem 1.1 in [17]:Dimension independent tensorization of Gaussian Measure] Standard normal on any finite dimensional conventional metric space $\mathbb{R}^n$ satisfies $T_2(1)$. Moreover, for an ℓ_2 additive metric*

$$d^2_{(N)}(x^N, y^N) := \sqrt{\sum_{i=1}^N d^2(x_i, y_i)}, \tag{1.18}$$

on product space $\mathbb{R}^{n^{\otimes N}}$, isotropic Gaussian satsfies $T_1^{d^2_{(N)}}(1)$.

From now on throughout the paper we will use ℓ_2 metric and $d = d^2_{(N)}(x^N, y^N)$. A

remarkable advantage of preceding result combined with Lipschitz function concentration result is that given a process $x = (x_1, x_2, \ldots, x_N) \sim \mu_N = N\Big(0, \Sigma_N\Big)$, then $\mu_N \in T_1\left(\left\|\Sigma_N^{\frac{1}{2}}\right\|^2\right)$ and even though process might have temporal dependencies but as long as one can prove $\left\|\Sigma_N^{\frac{1}{2}}\right\|^2 = O(1)$, dimension independent tensorization follows. This brings us to *norm-stable* ARMA models

Theorem 3. *[Proposition 4.1 in [5]:Dimension-Independent Tensorization of Talagrands' Inequality for norm-stable ARMA models] Consider the following model:*

$$x_{t+1} = Ax_t + w_t, \quad \textit{and i.i.d} \;\; w_t \sim \mathcal{N}(0, \mathcal{I}_n). \tag{1.19}$$

If system transition matrix, A satisfies $\|A\| < 1$, then the process level law of $(x_1 \ldots x_N)$ satisfies $T_1\left(\frac{1}{(1-\|A\|)^2}\right)$, if $\|A\| = 1$ we have $T_1\Big(N(N+1)\Big)$ and for $\|A\| > 1$, $T_1\big(\|A\|^N N\big)$

.

A curious reader might wonder what about the case $\rho(A) < 1$, unfortunately dimension independent tensorization is not guaranteed. However, one can find a sub-trajectory that satisfies dimension independent Talagrands' inequality, see e.g., [18]. A fundamental limitation of existing results, stem from the fact that they base their analysis only on the extreme singular values of the data matrix which do not offer much of geometric insight. This brings us to an extremely important, although simple to prove equality that relates the distance between different rows (whose concentration we understand well because of preceding theorem 3) of the data matrix to all of its' singular values.

Theorem 4 (Lemma A.4 in [19] Negative second moment)**.** *Let $1 \leqslant d \leqslant p$ and $Y \in \mathbb{R}^{d\times p}$ be a full rank matrix with singular values, $\sigma_1(Y) \geqslant \sigma_2(Y) \ldots \geqslant \sigma_d(Y)$. Let v_j be the hyperplane generated by all the rows of Y-except the $j-th$: i.e., span of*

$y_1, y_2, \ldots, y_{j-1}, y_{j+1}, \ldots, y_d$ *for* $1 \leqslant j \leqslant d$, $(e_j)_{j=1}^d$ *be the canonical basis of* $\mathbb{R}^d$, *then:*

$$\sum_{j=1}^{d} \sigma_j^{-2}(Y) = \sum_{j=1}^{d} \left\langle \left(YY^*\right)^{-1} e_j, e_j \right\rangle = \sum_{j=1}^{d} d_j^{-2}, \tag{1.20}$$

where $d_j^{-2} := d^{-2}(y_j, v_j)$, *distance between* y_j *and the point closest to it in the subspace* v_j

An important observation about orthogonal projections which will be extremely useful in providing partial information on inverse sample covariance matrix and consequently, pseudo-inverse of the data matrix in system identification problem is:

Corollary 1. *Let* $x_j \in \mathcal{S}^{p-1}$ *such that it is orthogonal to subspace* v_j *and* $P_{v_j^\perp}$ *be the orthogonal projection onto subspace orthogonal to* v_j *then using properties of projections and cauchy-schwarz:*

$$\left|\langle y_j, x_j \rangle\right| = \left|\langle P_{v_j^\perp}(y_j), x_j \rangle\right| \leqslant \left\|P_{v_j^\perp}(y_j)\right\| = d(y_j, v_j). \tag{1.21}$$

Our novel analysis on OLS estimation error will boil down to studying sum of N identically distributed independent variables weighted by columns of pseudoinverse of the data matrix. It turns out that this problem is related to the well known

Littlewood-Offord Problem: while working on random polynomials, Littlewood and Offord posed the following question: Let $a := [a_1, ..., a_N]$ be nonzero integers, and consider the function $f(x_1, \ldots, x_N) = \langle x, a \rangle$ How many solutions can $f(x) = r$ have with $x_i \in \{+1, -1\}^N$? which amounts to studying probability of maximal atom w.r.t non-zero constants a, defined as

$$p(a) := \max_{r \in \mathbb{R}} P\big(f(x) = r\big). \tag{1.22}$$

and turns out to depend heavily on the structure of co-efficients a

1. If $a = (1, 1, \ldots, 1)$ and N is even:

$$p(a) = \binom{N}{\frac{N}{2}} 2^{-N} = O\left(\frac{1}{\sqrt{N}}\right), \tag{1.23}$$

   which is essentially probability of equal heads and tails in N trials.

2. If $a = (2^0, 2^1, 2^2, \ldots, 2^{N-1})$

$$p(a) = \frac{1}{2^N} \tag{1.24}$$

3. If $a = (1, 2, \ldots, N)$

$$p(a) = \Theta\left(\frac{1}{N^{\frac{3}{2}}}\right), \tag{1.25}$$

see e.g., [20] also note that O and Θ are asymptotically true in this case.

Chapter 2

Linear time invariant systems

A common belief is given a stable state transition matrix A, $\|A^k\|$ converges exponentially fast to 0 and the rate is dependent on spectral radius. However, such arguments can hide dependency on the size of underlying state space. Limitation of existing bounds arise from ignoring the fact: given oracle bound on spectral radius $\rho < 1$ and underlying dimension of the state space n, there exists a class of linear transformations that satisfy spectral radius restriction but elements of this class can have their $k-$ th powers' norm behave very differently from each other; an $n \times n$ matrix with diagonal element of ρ is similar to n one dimensional systems and $\|A^k\| = \rho^k$, on the other hand we will show that there exists state transition matrix with same spectral radius, but its' resulting dynamics have strong spatial correlations and $\|A^k\| \geqslant 1$ for $k = n - 1$. Main results from this chapter are as follows:

- Using spectral theorem for non-hermitian matrices we show magnitude and discrepancy between algebraic and geometric multiplicity of distinct eigenvalues controls the decay rate of the norm associated with finite powers of a stable linear transformation
- A simple corollary of handle on $\|A^k\|$ leads to quantitave handle on mixing time of LG in Wasserstein metric.

2.1 Spectral theorem via invariant subspaces: non-hermitian case

Model under consideration is a following $n-$ dimensional stable LTI-system with isotropic Gaussian noise(interchangeably called $n-$ dimensional stable ARMA model or $n-$ dimensional stable LG).

$$x_{t+1} = Ax_t + w_t, \quad \rho(A) < 1 \quad \text{and i.i.d} \quad w_t \sim \mathcal{N}(0, \mathcal{I}_n). \tag{2.1}$$

It mixes to stationary distribution $\mu_\infty \sim \mathcal{N}(0, P_\infty)$, where P_∞ is the unique positive definite solution of the following Lyapunov equation:

$$A^* P_\infty A - P_\infty + I_n = 0. \tag{2.2}$$

Position or magnitude of eigenvalues associated to a linear operator A only provides partial information about its' properties (for the ease of exposition, throughout this paper we will assume that A does not have any non-trivial null space). In this paper, we will study operators and matrices via their actions on associated invariant subspaces and concepts like generalized eigenevectors. Topic in itself can take a semester of work and we refer the reader to [21,22]. However, we will try here to give the reader a quick intuition of this approach, often at the cost of rigor and thoroughness. One of the advantage of taking this approach is: *control on the $k-th$ power of matrix norm, independent of the basis structure* . Roughly speaking, algebraic multiplicity of eigenvalues follow from chatacteristic polynomial of the matirx.

$$det(zI - A) = \prod_{i=1}^{K} (z - \lambda_i)^{m_i}, \tag{2.3}$$

where λ_i are distinct with multiplicity m_i and $\sum_{i=1}^{K} m_i = n$. Since algebraic multiplicity of eigenvalue λ_i is m_i , we denote it by $AM(\lambda_i) = m_i$. Similarly with each $\lambda_i \in \sigma(A)$, their is an associated set of eigenvectors and dimension of their span corresponds to geometric multiplicity of λ_i , which we denote by $GM(\lambda_i) = dim[N(A - \lambda_i I)]$. Recall, from linear algebra:

Lemma 2. *Let $A \in \mathbb{C}^{n\times n}$. Then A is diagonalizable iff there is a set of n linearly independent vectors, each of which is an eigenvector of A.*

So, in a situation where $GM(\lambda_i) < AM(\lambda_i)$, eigenvectors do not span $\mathbb{C}^n$ and one resorts with spanning the underlying state space by direct sum decomposition of $A-$ invariant subspaces (which might be spanned by more that one linearly independent vector, comprising of eigenvector and generalized eigenvectors).

Definition 3. *Given a matrix $A \in \mathbb{C}^{n\times n}$ and a subspace $M \subset \mathbb{C}^n$, we say that M is an $A-$ invariant subspace if $AM \subset M$.*

Indeed, preceding philosophy is underlying principal of Jordan canonical forms, but its' geometric intricacies can not be ignored when dealing with High Dimensional estimation and control problems.

Proposition 1. *We can decompose the underlying state space as a direct sum decomposition of $A-$ invariant subspaces $[M_{\lambda_i}]_{i=1}^{K}$ denoted by:*

$$\mathbb{C}^n = M_{\lambda_1} \oplus M_{\lambda_2} \oplus \ldots \oplus M_{\lambda_K}, \tag{2.4}$$

*where $M_{\lambda_i} = N(A - \lambda_i I)^{m_i}$ is called the **Generalized eigenspace** associated with eiegenvalue λ_i(see e.g., theorem 3.11 in [21]).*

Direct sum decomposition of the state space via projections onto $A-$ invariant subspace Furthermore, one can define orthogonal projection matrices $[P_{\lambda_i}]_{i=1}^{K}$ associated to these invariant subspaces and the identity matrix can be written as:

$$I_n = P_{\lambda_1} \oplus P_{\lambda_2} \oplus \ldots \oplus P_{\lambda_K}. \tag{2.5}$$

Consequently, every $x \in \mathbb{R}^n$ can be written uniquely as $\oplus_{i=1}^{K} x_{\lambda_i}$, where x_{λ_i} is an element of subspace M_{λ_i}. How ? just for the intuition assume we are given a k dimensional subspace $\mathcal{V} \subset \mathbb{R}^n$ with its basis vectors $(v_1, v_2, \ldots, v_k)$, then orthogonal projection of an arbitray vector $x \in \mathbb{R}^n$ onto $\mathcal{V}$ is uniquely represented by:

$$P_{\mathcal{V}}(x) = a_1(x)v_1 + a_2(x)v_2 + \ldots a_k(x)v_k, \tag{2.6}$$

where $[a_i(x)]_{i=1}^{k}$ are minimizers of the following optimization problem:

$$\min_{a_1, a_2, \ldots, a_k} \|x - (a_1 v_1 + a_2 v_2 + \ldots a_k v_k)\|. \tag{2.7}$$

Let $V := [v_1, v_2, \ldots, v_k]$ that is $[v_i]_{i=1}^{k}$ be the columns of matrix V, then the matrix version of orthogonal projection of a given vector x on subspace $\mathcal{V}$ and associated coefficients are:

$$P_{\mathcal{V}}(x) = V(V^*V)^{-1}V^*x \quad \textit{and} \quad a(x) = (V^*V)^{-1}V^*x, \tag{2.8}$$

respectively.

Now we can re-arrange the basis according to (2.4) via projection operators and have the following diagonal blocks(corresponding to $A-$ invariant sub-spaces) of ma-

trices:

$$\hat{A} = \begin{bmatrix} A_{\lambda_1} & & \\ & \ddots & \\ & & A_{\lambda_k}, \end{bmatrix}$$

where superscript was used to signify this representation corresponds to change of basis according to proposition 1. Consider the following instructive example:

Example 1. Assume that given $A \in \mathbb{C}^{n\times n}$, comprises of only two distinct eigenvalues: λ and ρ , such that $AM(\lambda) = n_1$ and $GM(\lambda) = 1$. Similarly, $AM(\rho) = n_2$ (also $n_1 + n_2 = n$) and $GM(\rho) = 1$. Given an eigenvector v_1 such that $Av_1 = \lambda v_1$, we generate generalized eigenvector $v_2, v_3, \ldots, v_{n_1}$ recursively as $(A - \lambda I)v_2 = v_1$, $(A - \lambda I)v_3 = v_2$ and so on up to $(A - \lambda I)v_{n_1} = v_{n_1-1}$. We also have the following k -th step iteration:

$$\begin{aligned} A^k v_1 &= \lambda^k v_1 \\ A^k v_2 &= \lambda^k v_2 + \binom{k}{1}\lambda^{k-1} v_1 \\ A^k v_3 &= \lambda^k v_3 + \binom{k}{1}\lambda^{k-1} v_2 + \binom{k}{2}\lambda^{k-2} v_1 \\ \ldots &= \quad \ldots \\ A^k v_{n_1} &= \lambda^k v_{n_1} + \binom{k}{1}\lambda^{k-1} v_{n_1-1} + \ldots \binom{k}{n_1-1} v_1. \end{aligned} \tag{2.9}$$

Similarly given $Aw_1 = \rho w_1$, we recursively generate generalized eigenvectors up to $(A - \rho I)w_{n_2} = w_{n_2-1}$. Now under new basis: $[v_1, \ldots, v_{n_1-1}, v_{n_1}, w_1, \ldots, w_{n_2}]$, A can be represent as:

$$\hat{A} = \begin{bmatrix} A_\lambda & \\ & A_\rho \end{bmatrix}$$

$A_\lambda = \lambda I_{n_1} - N_{n_1}$ and $A_\rho = \rho I_{n_2} - N_{n_2}$ Where, N_{n_1} and N_{n_2} are n_1 and n_2 dimensional Nilpotent matrices, respectively. Two $A-$ invariant subspace are $M_\lambda = span[v_1, \ldots, v_{n_1}]$ and $M_\rho = span[w_1, \ldots, w_{n_2}]$.

2.2 Control on $\|A^k\|$

Assumption 1. *We assume here that each distinct eigenvalue λ_i has geometric multiplicity of 1 and $m_i > 1$. This is merely for the ease of exposition; as we can always replace a Jordan block by a diagonal element.*

Theorem 5. *Upper bound on the norm of the $k-th$ iteration associated to action of matrix A on invariant subspace M_{λ_i}, with block size m_i is precisely given as:*

$$\|A^k_{\lambda_i}\|_2 \leqslant |\lambda_i|^k k^{m_i-1} \sum_{m=0}^{m_i-1} \frac{1}{|\lambda_i|^m}, \tag{2.10}$$

Moreover, if $|\lambda_i| \in (0,1)$ then:

$$\|A^k_{\lambda_i}\|_2 \leqslant k^{m_i-1}|\lambda_i|^{k+1-m_i} \tag{2.11}$$

and for any $k \in \mathbb{N}$ such that

$$k > \frac{\ln(m_i)}{\ln\left(\frac{1}{|\lambda_i|}\right)} + \frac{[m_i - 1]\ln(k)}{\ln\left(\frac{1}{|\lambda_i|}\right)} + (m_i - 1) \tag{2.12}$$

we have that $\|A^k_{\lambda_i}\|_2 < 1$.

Proof.

$$\|A_{\lambda_i}^k\|_2 = \|(\lambda_i I_{m_i} + N_{m_i})^k\| = \|\sum_{m=0}^{k} \binom{k}{m} N_{m_i}^m \lambda_i^{k-m}\| \tag{2.13}$$
$$\leqslant \sum_{m=0}^{m_i-1} \binom{k}{m} |\lambda_i|^{k-m} \leqslant |\lambda_i|^k k^{m_i-1} \sum_{m=0}^{m_i-1} \frac{1}{|\lambda_i|^m}.$$

For the first two equalities and first inequality we have used the fact that any Jordan block of size m_i can be written as $(\lambda I_{m_i} + N_{m_i})$, where N_{m_i} is a Nilpotent matrix i.e., for any $m \geqslant m_i$, $N_{m_i}^m = 0$ and for $m < m_i, \|N_{m_i}^m\| = 1$. In the last inequality we have used the fact $\binom{k}{m} \leqslant k^{m_i-1}$ for $m \in [0, \ldots, m_i - 1]$. Given $|\lambda_i| \in (0, 1)$, $\sum_{m=0}^{m_i-1} \frac{1}{|\lambda_i|^m} \leqslant m_i |\lambda_i|^{1-m_i}$ and the results follows. □

Although our upper bound on the ℓ_2 norm and consequently lower bound for k that suffices for $\|A_{\lambda_i}^k\|_2 < 1$ is non-trivial, but using finite sum of geometric series, can be improved to

$$\|A_{\lambda_i}^k\|_2 \leqslant k^{m_i-1} |\lambda_i|^k \left(\frac{1 - |\lambda_i|}{1 - |\lambda_i|^{m_i}}\right), \quad \textit{and} \quad k > \frac{[m_i - 1]\ln(k)}{\ln\left(\frac{1}{|\lambda_i|}\right)} + \frac{\ln\left[\frac{1-|\lambda_i|^{m_i}}{1-|\lambda_i|}\right]}{\ln\left(\frac{1}{|\lambda_i|}\right)} \tag{2.14}$$
$$\textit{implies } \|A_{\lambda_i}^k\|_2 < 1.$$

Taylors' expansion can be used to show

$$\ln\left[\frac{1 - |\lambda_i|^{m_i}}{1 - |\lambda_i|}\right] \leqslant \left(\frac{1}{m_i - 1}\right) \ln\left(\frac{1}{1 - |\lambda|^{m_i}}\right) + \frac{|\lambda_i|}{1 - |\lambda_i|}, \tag{2.15}$$

and get a tighter lower bound on sufficient condition for k such that $\|A_{\lambda_i}^k\| < 1$.

Theorem 6. *Given a full-rank stable matrix $A \in \mathbb{C}^{n\times n}$, with distinct eigenvalues $[\lambda_i]_{i=1}^K$. Assume that geometric multiplicity of each distinct eigenvalue is 1 while*

algebraic multiplicity is $AM(\lambda_i) = m_i$*, then for any* k *greater than*

$$\hat{k} = \min\Big[k \in \mathbb{N} : k \geqslant \max_{i\in 1,\dots,K}\left(\frac{4[m_i-1]\ln m_i}{\ln\frac{1}{|\lambda_i|}}\right)\Big],$$
$$\|A^k\|_2 < 1$$

Proof. Since, generalized eigenvectors corresponding to distinct eigenvalues are linearly independent(see e.g., 8.13 in [22]), $[P_{\lambda_i}]_{i=1}^K$ are orthogonal projections. Consequently,

$$\begin{aligned}\|A^k x\|_2 &= \left\|A^k\left[\sum_{i=1}^K P_{\lambda_i}x\right]\right\| = \left\|\sum_{i=1}^K A^k_{\lambda_i}x_{\lambda_i}\right\| \\ &= \sqrt{\sum_{i=1}^K\left[\langle A^k_{\lambda_i}x_{\lambda_i}, A^k_{\lambda_i}x_{\lambda_i}\rangle + 2\sum_{j>i}^K\langle A^k_{\lambda_i}x_{\lambda_i}, A^k_{\lambda_j}x_{\lambda_j}\rangle\right]} \\ &= \sqrt{\sum_{i=1}^K\langle A^k_{\lambda_i}x_{\lambda_i}, A^k_{\lambda_i}x_{\lambda_i}\rangle} \leqslant \sqrt{\sum_{i=1}^K\left\|A^k_{\lambda_i}\right\|^2\left\|x_{\lambda_i}\right\|^2}\sqrt{\sum_{i=1}^K\left\|x_{\lambda_i}\right\|^2} = \|x\|. \qquad (2.16)\end{aligned}$$

where $x_{\lambda_i} := P_{\lambda_i}x$, where first equality in (2.16) follows from the fact that $N(A^*_{\lambda_i}) = \oplus_{j\neq i}^K M_{\lambda_j}$ and $\left\langle x_{\lambda_i}, (A^k_{\lambda_i})^* A^k_{\lambda_i}x_{\lambda_i}\right\rangle \leqslant \left\|A^k_{\lambda_i}\right\|^2\left\|x_{\lambda_i}\right\|^2$. Strict inequality in (2.16) follows from the fact that by hypothesis $k > \hat{k}$ and $\frac{\ln(m_i)}{\ln\left(\frac{1}{|\lambda_i|}\right)} + \frac{[m_i-1]\ln(k)}{\ln\left(\frac{1}{|\lambda_i|}\right)} + (m_i-1) < \left(\frac{4[m_i-1]\ln m_i}{\ln\frac{1}{|\lambda_i|}}\right)$ for all $i \in [1,2,\dots,K]$. □

Remark 7. *Since for a stable state transition matrix* $|\lambda_i| \leqslant \rho(A)$ *and* $m_i \leqslant n$ *for each* $i \in [1,2,\dots,K]$*, therefore* $\Gamma(n,\rho) = O\left(\frac{[n-1]\ln n}{\ln\frac{1}{\rho}}\right)$.

As any diagonalisable linear transformation A with ρ on diagonals satisfy $\|A^k\| = \rho^k < 1$ for all $k \in \mathbb{N}$, one might question tightness of preceding bound, which brings us to following result:

Definition 4 (Stable with Strong Spatial Correlations, S-w-SSCs)**.** *Consider the*

following example of a stable state transition matrix with only one distinct eigenvalue of $\rho \in (0,1)$

$$J_n(\rho) := \begin{bmatrix} \rho & 1 & 0 & \cdots & 0 & 0 \\ 0 & \rho & 1 & \ddots & 0 & 0 \\ 0 & 0 & \rho & \ddots & 0 & 0 \\ 0 & 0 & 0 & \ddots & 1 & 0 \\ \vdots & \ddots & \ddots & \ddots & \rho & 1 \\ 0 & 0 & 0 & \cdots & 0 & \rho \end{bmatrix} \tag{2.17}$$

with algebraic multiplicity of n *but only one linearly independent eigenvector.*

Theorem 8. *Given any* $n \in \mathbb{N}$ *and spectral radius* $\rho \in (0,1)$*, there exists a linear transformation* $A \in \mathbb{R}^{n\times n}$ *with* $\rho(A) = \rho$ *and* $\|A^{n-1}\| > 1$*.*

Proof. Let $A := J_n(\rho)$ and $[e_i]_{i=1}^n$ be canonical basis of $\mathbb{C}^n$, i.e., 1 at $i-$th position and 0 everywhere else, then notice for $x_0 = e_n$:

$$A^{n-1}e_n = \sum_{m=0}^{n-1} \binom{n-1}{m} \rho^{(n-1)-m} e_{n-m} = \sum_{m=0}^{n-2} \binom{n-1}{m} \rho^{(n-1)-m} e_{n-m} + e_1.$$

Therefore,

$$\|A^{n-1}\|_2 := \sup_{x\in\mathcal{S}^{n-1}} \|A^{n-1}x\| \geqslant \|A^{n-1}e_n\| > \|e_1\| = 1. \tag{2.18}$$

□

Remark 9. *Now we are in a position to interpret consequences of bounds on the operator norm of the powers of linear transformations that we have derived after a painstaking process of general spectral decomposition using projections onto invariant subspaces of a linear transformation. From dynamical systems point of view, in contrast with hermitian or diagonalizable setting where original dynamical system can be*

decomposed into n one dimensional dynamical systems with no spatial correlations, for non-Hermitian case we can only 'hope' to decompose the original dynamical system into multiple lower dimensional dynamical systems that are independent of each other(even in statistical sense as we will see this in subsection 3.0.3 of Section 3.1). But, when a linear transformation has only one distinct eigenvalue and span of corresponding eigenvectors is only one dimensional as in S-w-SSCs, dynamics are stable but spatially inseparable*. Consequently, if the underlying state space is high dimensional and dynamical system follows state transitions from S-w-SSCs, there exists an initial condition on S^{n-1} such that it can take a considerable amount of time for the system to be inside open unit ball and stay there forever as shown by* (2.18)*. Let us emphasise that property of being strongly* spatially correlated *is related to* large discrepancy between algebraic and geometric multiplicity of eigenvalues*, and only one Jordan block corresponding to n dimensional dynamics as in* (2.17) *is merely a graphical representation of this phenomenon.*

In the process we also discovered a quantitative handle on the decay of $\|A^k\|$, preciesely:

Theorem 10. *For a stable matrix A with K distinct eigenvalues, let discrepancy related to eigenvalue λ_i be $D_{\lambda_i} := AM(\lambda_i) - GM(\lambda_i)$, then*

$$\|A^k\| \leqslant \max_{1 \leqslant i \leqslant K} k^{D_{\lambda_i}} |\lambda_i|^k \left(\frac{1 - |\lambda_i|}{1 - |\lambda_i|^{D_{\lambda_i}+1}} \right) \tag{2.19}$$

Proof. Recall that $\|P_{\lambda_i}(x)\| = d(x, P_{\lambda_i^\perp})$ and for all $x \in S^{n-1}$ we have that

$$\sum_{i=1}^{K} d^2(x, P_{\lambda_i^\perp}) = \|x\|_2^2 \tag{2.20}$$

Furthermore,

$$\|A^k x\|_2 \leqslant \sqrt{\sum_{i=1}^{K} \left\|A_{\lambda_i}^k\right\|^2 \left\|P_{\lambda_i}(x)\right\|^2} = \sqrt{\sum_{i=1}^{K} \left\|A_{\lambda_i}^k\right\|^2 d^2(x, P_{\lambda_i^\perp})} \tag{2.21}$$

$$\leqslant \sqrt{\sum_{i=1}^{K} k^{2(m_i-1)} |\lambda_i|^{2k} \left(\frac{1-|\lambda_i|}{1-|\lambda_i|^{m_i}}\right)^2 d^2(x, P_{\lambda_i^\perp})}$$

where (2.21) follows from using tighter upper bound from (2.14), now without loss of generality assume that

$$j = \arg\max_{1 \leqslant i \leqslant K} \left[k^{(m_i-1)} |\lambda_i|^k \left(\frac{1-|\lambda_i|}{1-|\lambda_i|^{m_i}}\right)\right]. \tag{2.22}$$

Then a simple convexity argument reveals:

$$\sup_{x \in S^{n-1}} \sum_{i=1}^{K} k^{2(m_i-1)} |\lambda_i|^{2k} \left(\frac{1-|\lambda_i|}{1-|\lambda_i|^{m_i}}\right)^2 d^2(x, P_{\lambda_i^\perp}) = k^{2(m_j-1)} |\lambda_j|^{2k} \left(\frac{1-|\lambda_j|}{1-|\lambda_j|^{m_j}}\right)^2,$$

where equality follows by picking any $x \in S^{n-1}$ with $d^2(x, P_{\lambda_j^\perp}) = 1$. □

A trivial corollary reveals an upper bound on operator norm of a finite dimensional linear transformation with all eigenvalues strictly inside unit circle

Corollary 2. *Given a stable state transition matrix A with K distinct eigenvalues $[\lambda_i]_{i=1}^k$ along with their associated discrepancy D_{λ_i}, operator norm can be upper bounded as:*

$$\|A\|_2 \leqslant \max_{1 \leqslant i \leqslant K} |\lambda_i| \left(\frac{1-|\lambda_i|}{1-|\lambda_i|^{D_{\lambda_i}+1}}\right) \tag{2.23}$$

Remark 11. *Our analysis reveal that operator norm of k-th power of stable state transition matrix A is a maximum over the operator norm of k-th power of lower dimensional linear transformations corresponding to restriction of A to its' invariant subspaces. Furthermore, operator norm of k-th power of A restricted to its' invariant*

subspace with eigenvalue λ is at most a product of (i) polynomial in k with degree equal to size of the invariant subspace minus one(i.e., discrepancy of eigenvalue λ) (ii) an exponential which decays with increase in k and the rate is determined by magnitude of λ, modulo a constant dependent on magnitude of λ and size of the invariant subspace. Therefore, as opposed to common belief, there is no direct causal relationship between spectral radius and decay rate of $\|A^k\|$ for finite powers of A

Our analysis is first of a kind where we translate the knowledge on entire spectrum of non-Hermitian, finite dimensional linear transformation into quantitative behavior of its powers, including evolution of norm of its' powers. One of the frequently used quantitative handle on $\|A^k\|$, is provided by [4]. In an attempt to circumvent appearance of the condition number of similarity transform associated with Jordan blocks, they employed Schur form with assumptions of upper bound on $\|A\|_2$, i.e., $\|A\|_2 \leqslant M$ and all the eigenvalues of A are on or inside the unit circle in complex plane, then quantitative handle is precisely $\|A^k\|_2 \leqslant (ek)^{n-1} \max(M^n, 1)$. To begin with, our analysis shows operator norm is independent of basis choice and $\|A\|_2 \leqslant \max_{1\leqslant i\leqslant K} |\lambda_i| \left(\frac{1-|\lambda_i|}{1-|\lambda_i|^{D_{\lambda_i}+1}} \right)$ when eigenvalues are strictly inside the unit circle. The worst case for eigenvalue on unit circle will be discrepancy of $n-1$ and can be bounded simply as $\left\|A^k\right\| = \left\|(\lambda I + N)^k\right\| = \left\| \sum_{m=0}^{k} \binom{k}{m} N^m \lambda^{k-m} \right\| \leqslant \sum_{m=0}^{k\wedge(n-1)} \binom{k}{m} |\lambda|^{k-m} = \sum_{m=0}^{k} \binom{k}{m} = 2^k$ for $k < n$. A general purpose upper bound for $k \geqslant n$ is $\|A^k\| \leqslant \sum_{m=0}^{n-1} \binom{k}{m} = 2^k - \sum_{m=n}^{k} \binom{k}{m}$, playing around with binomial expansions/ stirling approximation and picking k as a function of n can often lead to good quantitative results; if we chose $k = 2n$ we get $\|A^{2n}\| \leqslant 2^n - n$.

2.3 Wasserstein-mixing time

When dealing with Linear Gaussians on unbounded state space, instead of total variation distance it is more natural to to use Wasserstein metric(this will become evident shorty)

Lemma 3. *(Wassersetin distance between two Gaussians)*

$$W_2^2\big(\mathcal{N}(\mu_1,\Sigma_1),\mathcal{N}(\mu_2,\Sigma_2)\big) = \|m_1-m_2\|^2 + \|\Sigma_1^{\frac{1}{2}} - \Sigma_2^{\frac{1}{2}}\|_F^2 \tag{2.24}$$

Ergodicity in Wasserstein metric Ergodicity is a notion of how fast does the Markov chain forgets its initial condition (and eventually converges to stationary distribution). It turns out to be dependent on decay of $\|A^k\|$ for Linear Gaussians in Wasserstein metric.

Definition 5. *Given $\epsilon \in (0,1]$, we define ϵ- Wasserstein mixing time of the Markov chain with transition kernel-invariant measure $(P, \mu_{[\infty,P]})$ as:*

$$\tau^{wass}_{(P,\mu_{[\infty,P]})}(\epsilon) := \min\big[k \in \mathbb{N} : W_2(\mu P^k, \mu_{[\infty,P]}) < \epsilon W_2(\mu, \mu_{[\infty,P]})\big] \tag{2.25}$$

Proposition 2. *Given dimension of the state space n and $\rho \in (0,1)$, with every linear transformation $A \in \mathbb{R}^{n\times n}$ with spectral radius $\rho(A) = \rho$ we define a Markov chain with transition kernel $P_x \sim N(Ax, I_n)$ for every $x \in \mathbb{R}^n$, then max-min mixing time:*

$$\max_{P_x\sim N(Ax,I_n):\rho(A)=\rho} \tau^{wass}_{(P,\mu_{[\infty,P]})}(\epsilon) = O\left(\frac{\max\big([n-1]\ln(n), \ln(\frac{1}{\epsilon})\big)}{\ln(\frac{1}{\rho})}\right) \tag{2.26}$$

Proof. It suffices to consider transition kernels for $n-$ dimensional S-w-SSCs, we have

for $x \neq y$:

$$W_2^2\big(P_x^k, P_y^k\big) = \big\|A^k(x-y)\big\|^2 = O\bigg(k^{2(n-1)}\rho^{2k}\big(\frac{1-\rho}{1-\rho^n}\big)^2\bigg)\|x-y\|^2.$$

This implies that:

$$W_2^2\big(\mu P^k, \nu P^k\big) = O\bigg(k^{2(n-1)}\rho^{2k}\big(\frac{1-\rho}{1-\rho^n}\big)^2\bigg)W_2^2(\mu,\nu) \tag{2.27}$$

Now let μ_∞ be the steady state distribution, and if $x_0 \sim \mu$ then divergence between distribution of x_k and steady state follows by letting $\nu := \mu_\infty$ in the previous equation and noting that stationary implies $\mu_\infty P^k = \mu_\infty$

$$W_2^2\big(\mu P^k, \mu_\infty\big) = O\bigg(k^{2(n-1)}\rho^{2k}\big(\frac{1-\rho}{1-\rho^n}\big)^2\bigg)W_2^2(\mu,\mu_\infty),$$

and result follows. □

Remark 12. *We know from Chapter 1 that these mixing bounds are tight upto* $\ln(n)$ *factor so convergence to stationarity as believed in previous works [12], [3] and references there in to be* $\ln(\frac{1}{\rho})$ *should be changed to* $\frac{(n-1)\ln(n)}{\ln(\frac{1}{\rho})}$. *So behavior towards stationarity is again dictated by not just magnitude of distinct eigenvalues but also discrepancy between their algebraic and geometric multiplicities.*

Chapter 3

System Identification via single trajectory

This chapter sets the stage for error analysis for estimating linear time-invariant(LTI) systems from a single observed trajectory via methods of ordinary least squares(OLS). Although, system identification via ordinary least squares regression had been a hot topic of research for last few years see e.g., [9], [3], and [10]. However, recently it was noted in [23] and [4] that there exists example of stable dynamical systems in dimension $\geqslant 10$ where OLS contains non-vanishing error, pointing out gaps in existing analysis. Something that has been left un-noticed: least squares/ regression related problem are of geometric nature for example with minimal effort we manage to point out its' deterioration with possible dependence between rows of the data matrix, which happens in case of large discrepancy between algebraic and geometric multiplicity of eigenvalues associated with state-transition matrix., We begin Section 3.0.1, with general least squares estimation problem and link the estimation error with linear dependence of the given basis. Subsequently, OLS set up for Linear Dynamical systems with Gaussian noise is introduced. Main results from this chapter are as follows:

- Two different error bounds in estimation are introduced, a geometric one based on controlling distance between a given row and conjugate hyperplane and another one based on extreme singular values of the data matrix. Using inner product structure of the state space and sample space we manage to get an almost 'closed' form expression for element-wise least squares error and turns out to be a scalar weigthed random walk of standard Gaussians with weights defined by the columns of pseudo-inverse of the data matrix, where columns of pseudo-

inverse are constrained to be orthonormal to the rows of the data-matrix and hence capture structural properties of the entire trajectory show qualitatively different behavior of pseudo-inverse in the presence of strong spatial correlations versus only temporal correlations. Error is at most a polynomial function of all the Gaussian excitations of the dynamical system weighted by powers of A, which is a higher degree variant of *Littlewood-Offord problem*, a work in progress and requires an entire paper of its' own. Furthermore, $i-$th diagonal entry of the inverse sample covariance matrix (which dictates the ℓ_2 norm of the $i-$ th column of pseudo-inverse) is inverse distance squared between the $i-$th row of the data matrix and hyperplane spanned by all the other rows of the data matrix

- We show how spectral theorem combined with Gaussian projection lemma allows us to spatially decouple the data matrix into lower dimensional statistically independent random dynamical systems. Various other estimates that will assist with a conclusive remark on estimation error in OLS are also collected.As a first step towards demystifying these intricate dependencies

3.0.1 Ordinary Least Squares estimator

In this section we analyse the problem of OLS estimation for system transition matrix A from single observed (as in [9], [3], [4]) trajectory of $(x_0, x_1, \ldots, x_N)$ satisyfing:

$$x_{t+1} = Ax_t + w_t, \quad \text{where } w_t \sim N(0, I). \tag{3.1}$$

Before delving into solution of the estimation problem, we would like to give a brief overview into working of general OLS regression along with potential limitations: A

priori you are given $k-$ basis functions of an $n-$ dimensional vector space, where $n > k$ and one can form an $n \times k$ matrix $X := [v_1, v_2, \ldots, v_k] \in \mathbb{R}^{n \times k}$. Now one observes y :

$$y = X\beta^* + \epsilon, \ \epsilon \sim N(0, I_n) \tag{3.2}$$

and tries to estimate $\beta := \beta(X, y) \in \mathbb{R}^{k \times 1}$ by projecting observation y onto the span of X as in (2.8), we get $\beta = (X^*X)^{-1}X^*y$ and the expected error in ℓ_2 norm is:

$$\mathbb{E}\|\beta - \beta^*\|^2 = Tr([X^*X]^{-1}) = \sum_{j=1}^{k} d_j^{-2}, \tag{3.3}$$

where last equality follows from negative second moment identity from Theorem 4 and one can immediately conclude: if any column of X gets closer in terms of ℓ_2 distance on $\mathbb{R}^n$ to the span of remaining $k-1$ columns, expected squared error in OLS estimation will increase. Vaguely speaking, linear dependence between basis of the data matrix deteriorates the performance of standard OLS regression.

OLS solution for identification of LTI system from a single observed trajectiry is:

$$\hat{A} = \arg\min_{B \in \mathbb{R}^{n \times n}} \sum_{t=0}^{N-1} \|x_{t+1} - Bx_t\|^2. \tag{3.4}$$

Recall, $X_+ = [x_1, x_2, \ldots, x_N]$ and $X_- = [x_0, x_1, \ldots, x_{(N-1)}]$, noise covariates $E = [w_0, w_1, \ldots, w_{N-1}]$, y_j be the rows of X_- and v_j be the hyperplane as defined in theorem 4. Also notice that conditioned on $x_0 = 0$ state at time i can be represented in terms of powers of A and noise covariates as:

$$x_i = \sum_{t=1}^{i} A^{i-t} w_{t-1}, \ \textit{and graphical representation} \tag{3.5}$$

$$A^* := \begin{bmatrix} | & | & | & | & | \\ b_1 & b_2 & | & b_{n-1} & b_n \\ | & | & | & | & | \end{bmatrix}, X_-^* = \begin{bmatrix} | & | & | & | & | \\ | & | & | & | & | \\ | & | & | & | & | \\ y_1 & y_2 & | & y_{n-1} & z_n \\ | & | & | & | & | \\ | & | & | & | & | \\ | & | & | & | & | \end{bmatrix}$$

In this dynamical version of OLS, at most n basis are provided by columns of X_-^*. Then $i-$ th column of A^* corresponds to co-effiecients estimated, by orthogonally projecting observation z_i onto span of X_-^*

$$\hat{b}_i = \left(X_- X_-^*\right)^{-1} X_- z_i \tag{3.6}$$

Then the closed form expression for Least squares solution and error are:

$$\hat{A} = X_+ X_-^\dagger, \ \textit{where}\, X_-^\dagger := X_-^*(X_- X_-^*)^{-1} \tag{3.7}$$

$$\left\|A - \hat{A}\right\|_F = \left\|E X_-^\dagger\right\|_F \tag{3.8}$$

3.0.2 Geometric and spectral approaches to error analysis

Spectral statistics of a random matrix of i.i.d Gaussian ensembles E (each entry of the matrix is normally distributed with mean 0 and variance 1) has been very well studied in an attempt to leverage upon this information, authors in [23] managed to bound the error in terms of distances between row vectors and span of remaining rows. We reiterate there argument, by definition $\left\|A - \hat{A}\right\|_F^2 = \sum_{k=1}^{n} \sigma_k^2\left(E X_-^\dagger\right)$, using

Courant-Fischer:

$$\sigma_n\big(E_{N(X_-)^\perp}\big)\sigma_k\big(X_-^\dagger\big) \leqslant \sigma_k\big(EX_-^\dagger\big) \leqslant \sigma_1\big(E_{N(X_-)^\perp}\big)\sigma_k\big(X_-^\dagger\big) \tag{3.9}$$

Recall from Lemma 1, $X_-^\dagger$ is a bijection from $Im(X_-) = \mathbb{R}^n$ to $N(X_-)^\perp$ (subspace orthogonal to null space of X_-) and $dim[N(X_-)^\perp] = n$, so given $N(X_-)$ one can construct $N(X_-)^\perp$ such that $E_{N(X_-)^\perp} := EX_-^\dagger = \text{span}[w_0, w_1, \ldots, w_{n-1}]$, i.e., it should be viewed as an $n \times n$ square matrix of i.i.d Gaussian ensembles. Furthermore,

$$\big\|X_-^\dagger\big\|_F^2 = Tr\big(X_-^\dagger[X_-^\dagger]^*\big) = Tr\big((X_-X_-^*)^{-1}\big) = \sum_{j=1}^{n} \sigma_j^{-2}(X_-).$$

using negative second moment identity:

$$\sigma_n(E^\perp)\bigg(\sum_{j=1}^{n} d_j^{-2}\bigg)^{\frac{1}{2}} \leqslant \big\|A - \hat{A}\big\|_F \leqslant \sigma_1(E^\perp)\bigg(\sum_{j=1}^{n} d_j^{-2}\bigg)^{\frac{1}{2}}, \tag{3.10}$$

where we have used $E^\perp$ as shorthand for $E_{N(X_-)^\perp}$. Preceding bounds on error analysis seems favorable when underlying dynamics are generated from hermitian state transition matrix, as analysis of $\left(\sum_{j=1}^{n} d_j^{-2}\right)^{\frac{1}{2}}$ can potentially be decoupled into analysis of individual d_j with some cost, because rows of the data matrix can be considered independent of each other.

Estimation error is independent of basis function choice Recall that if $A = A^*$, then there exists a unitary matrix $U \in \mathbb{C}^{n\times n}$ and a real diagonal matrix $\Omega \in \mathbb{R}^{n\times n}$ such that $A = U^*\Omega U$ and a coordinate transform for (3.1), i.e., $z_{t+1} = \Omega z_t + Uw_t$, where $z_t := Ux_t$ and Uw_t is again an isotropic Gaussian. Now let $W := UE$, and notice that rows of $Z_- := UX_-$ are independent one dimensional Linear Gaussian dynamics with growth/ decay parameter defined by corresponding diagonal entery in

Ω. As unitary by definition $U^*U = UU^* = I$, implies $U^{-1} = U^*$, we have

$$\begin{aligned}\left\|WZ_-^*\left(Z_-Z_-^*\right)^{-1}\right\|_F^2 &= \left\|UE(UX_-)^*(UX_-X_-^*U^*)^{-1}\right\|_F^2 \\ &= Tr(EX_-^*(X_-X_-^*)^{-2}X_-E^*) = \|EX_-^*(X_-X_-^*)^{-1}\|_F^2.\end{aligned}$$

So a first step towards rigorous analysis along bound in (3.10) would require quantifying distance between a random one dimensional ARMA trajectory of length N and a fixed $n-1$ dimensional subspace of $\mathbb{R}^N$ with high probability and we study it in Section 4.3.1. Roughly speaking, analysis provided in [3] and [9] focuses on upper bounding the estimation error in operator norm by decomposing the error into a sample covariance term $\left(X_-X_-^*\right)^{-\frac{1}{2}}$ and a martingale difference term $EX_-\left(X_-X_-^*\right)^{-\frac{1}{2}}$. Again using the Courant-Fischer argument as in (3.9), we can now instead bound the Frobenius norm error in terms of the martingale difference term and extreme singular values of the data matrix:

$$\frac{\left\|EX_-^*\left(X_-X_-^*\right)^{-\frac{1}{2}}\right\|_F}{\sigma_1(X_-)} \leqslant \left\|A-\hat{A}\right\|_F \leqslant \frac{\left\|EX_-^*\left(X_-X_-^*\right)^{-\frac{1}{2}}\right\|_F}{\sigma_n(X_-)}. \tag{3.11}$$

Work of [2] and [3], focused on bounding the right hand side of (3.12) in operator norm, which along with controlling martingale term required showing a high probability lower bound on $\sigma_n(X_-)$. However, in order to ensure error convergence or divergence, error formulation in (3.12) would require understanding the behavior of largest singular value of the data matrix or said in another way largest eigenvalue of sample covariance matrix as $\sigma_1(X_-) = \sqrt{\lambda_{\max}(X_-X_-^*)}$. In fact, [2] shows inconsistency of explosive system with state transition matrix $A = 1.1I_n$ by showing largest singular value grows exponentially, see proposition 19.1 of [9]. Ubiquity of error formulation based on a martingale term as in (3.12), for various system identification

problems (stochastic system identification in [12], Ho-Kalman algorithm in [8]) is justified in literature by arguments like E and X_- are dependent and this gives optimal bounds. However we find it easier to consider a somewhat similar error formulation

$$\frac{\left\|EX_-^*\right\|_F}{\sigma_1^2(X_-)} \leqslant \left\|A-\hat{A}\right\|_F \leqslant \frac{\left\|EX_-^*\right\|_F}{\sigma_n^2(X_-)}, \tag{3.12}$$

to show that for high dimensional spatially inseparable dynamics OLS error stays bounded away from zero.

Remark 13. *But this will have a huge gap when order of the smallest and largest singular values is different. So instead we will now try*

$$\begin{aligned}
&\sigma_n(EX_-^*)\left\|(X_-X_-^*)^{-1}\right\|_F \leqslant \left\|A-\hat{A}\right\|_F \leqslant \sigma_1(EX_-^*)\left\|(X_-X_-^*)^{-1}\right\|_F \\
&\sigma_n(EX_-^*)\left(\sum_{j=1}^{n}\sigma_j^{-4}(X_-)\right)^{\frac{1}{2}} \leqslant \left\|A-\hat{A}\right\|_F \leqslant \sigma_1(EX_-^*)\left(\sum_{j=1}^{n}\sigma_j^{-4}(X_-)\right)^{\frac{1}{2}},
\end{aligned} \tag{3.13}$$

where recall that $\sigma_j^4(X_-) = \lambda_j^2(X_-X_-^*)$, eigenvalue of the sample covariance matrix.

3.0.3 Spatially independent lower dimensional random dynamical systems

Notice that using the description of linear transformation in Theorem 6, $i-$ the column of the data matrix can be decomposed into row blocks that are statistically independent of each other:

$$x_i = \sum_{t=1}^{i} A^{i-t}w_{t-1} = \sum_{t=1}^{i} A^{i-t}\left(\sum_{m=1}^{K} P_{\lambda_m}\right)w_{t-1} = \sum_{m=1}^{K}\underbrace{\sum_{t=1}^{i} A_{\lambda_m}^{i-t}w_{t-1}^m}_{:=B_{\lambda_m}(i)}. \tag{3.14}$$

Now recall that D_{λ_m} was used to denote discrepancy between algebraic and geometric multiplicity of eigenvalue λ_m where for fixed t notice that $\mathbb{E}\|w_{t-1}^m\|^2 = \mathbb{E}\big\langle w_{t-1}, P_{\lambda_m} w_{t-1}\big\rangle = Tr(P_{\lambda_m}) = D_{\lambda_m} + 1$. So w_{t-1}^m is standard normal on $A-$ invariant subspace M_{λ_m} of dimension $D_{\lambda_m} + 1$. Since for all $t \in \mathbb{N}$ and $m, m^{'} \in [K]$ such that $m \neq m^{'}$, w_t^m and $w_t^{m^{'}}$ are independent: $\mathbb{E}\big[w_t^m (w_t^{m^{'}})^*\big] = \mathbb{E}\big[P_{\lambda_m} w_t (P_{\lambda_{m^{'}}} w_t)^*\big] = \mathbb{E}\big[P_{\lambda_m} w_t w_t^* P_{\lambda_{m^{'}}}\big] = P_{\lambda_m} P_{\lambda_{m^{'}}} = \delta_m(m^{'}) I_n$, combined with the fact that $A_m^* A_{m'} = 0$ for $m \neq m'$ (because $N(A_{\lambda_m}^*) = \oplus_{j \neq m}^K M_{\lambda_j}$) implies that $B_{\lambda_m}(i)$ and $B_{\lambda_{m'}}(i)$ are independent of each other for all i. Therefore, data matrix essentially contains time realization of K-low dimensional dynamical systems, which are statistically independent of each other. As we showed in the previous section that estimation error is independent of the choice of underlying basis, w.l.o.g we can take canonical basis implying rows in data matrix comprises of independent blocks, where each block is a time realization of trajectory generated via canonical form of linear transformation with specified eigenvalue and size of the block equals $D_{\lambda_m} + 1$. Only situation where dynamics are spatially inseparable corresponds to covariates being generated from S-w-SSCs and we will now focus on interactions between rows of the data matrix in that case. In order to unravel spatio-temporal correlations, it is important to represent elements of the data matrix in a compact form which one can do with inner products; $i-$ th column(time index) and $j-$ th row(space index) of the data matrix can be represented as

$$\big[X_-\big]_{j,i} = \big\langle x_i, e_j \big\rangle = \sum_{t=1}^{i} \big\langle A^{i-t} w_{t-1}, e_j \big\rangle.$$

What makes the case of n dimensional S-w-SSCs so peculiar is that covariate in j th row is a function of row $[j, \ldots, n]$ of Gaussian ensemble E, precisely expressed:

Proposition 3. *$A = J_n(\lambda)$ i.e., underlying dynamics are S-w-SSCs, then element*

corresponding to $j-$ th row and $i-$th column can be concisely expressed as:

$$[X_-]_{j,i} = \sum_{t=1}^{i} \sum_{m=0}^{(i-t)\wedge(n-j)} \binom{i-t}{m} \lambda^{i-t-m} \big\langle w_{t-1}, e_{m+j} \big\rangle \tag{3.15}$$

Proof. This is where inner product representation really helps along with noticing that nilpotent matrix is a shift operator, as:

$$\begin{aligned} \sum_{t=1}^{i} \big\langle A^{i-t} w_{t-1}, e_j \big\rangle &= \sum_{t=1}^{i} \big\langle (\lambda I + N_n)^{i-t} w_{t-1}, e_j \big\rangle \\ &= \left\langle \sum_{t=1}^{i} \sum_{m=0}^{(i-t)\wedge n} \binom{i-t}{m} N_n^m \lambda^{i-t-m} w_{t-1}, e_j \right\rangle \end{aligned}$$

Now a simple observation reveals that $N_n^{m*} e_j = e_{j+m}$ for $m \leqslant n-j$, and 0 otherwise. Hence,

$$[X_-]_{j,i} = \sum_{t=1}^{i} \sum_{m=0}^{(i-t)\wedge(n-j)} \binom{i-t}{m} \lambda^{i-t-m} \left\langle w_{t-1}, e_{j+m} \right\rangle. \tag{3.16}$$

□

As we have convinced ourselves that data matrix can be divided into row blocks that are statistically independent except for the S-w-SSCs case, along with the compact representation for individual elements inside the data matrix, we now aim at understanding elementwise estimation error.

3.1 Elementwise estimation error

Although, bounds on estimation error discussed in preceding section provide a good intuition, but existing analysis based on martingales e.t.c does not reveal explicit dependencies on number of iterations and state space dimensions. Surprisingly, it

had been left unnoticed that pseudo-inverse is constrained with respect to the data matrix which we discussed is itself a function of Gaussian ensemble E and element-wise estimation error is an inner product between Gaussian ensemble and pseudo-inverse.

Theorem 14. *[Construction of Inverse Sample Covariance Matrix] Let* $[e_j]_{j=1}^n$ *be canonical basis of* $\mathbb{C}^n$ *and* X_- *be the data matrix of stable linear dynamical system with isotropic Gaussian noise as in* (3.1)*, Given, the rows of data matrix* $[y_j]_{j=1}^n$*, inverse sample covariance matrix* $\left(X_-X_-^*\right)^{-1} = [v_{j,k}]_{j,k=1}^n$ *satisfies following constraints:*

1. *For fixed* j *in* $[1,\ldots,n]$, $\sum_{k\neq j} v_{j,k}\big\langle y_k, y_j\big\rangle = 1 - v_{j,j}\|y_j\|^2$*, precisely said:*

$$\begin{aligned}&\sum_{k\neq j} v_{j,k} \sum_{i=1}^{N-1}\sum_{t=1}^{i}\big\langle A^{i-t}w_{t-1}, e_k\big\rangle\overline{\sum_{t=1}^{i}\big\langle A^{i-t}w_{t-1}, e_j\big\rangle} \\ &= 1 - v_{j,j}\sum_{i=1}^{N-1}\left|\sum_{t=1}^{i}\big\langle A^{i-t}w_{t-1}, e_j\big\rangle\right|^2\end{aligned} \tag{3.17}$$

2. *and for all* $l \neq j$, $\sum_{k=1}^n v_{j,k}\big\langle y_k, y_l\big\rangle = 0$*, precisely said:*

$$\begin{aligned}&\sum_{k=1}^{n} v_{j,k} \sum_{i=1}^{N-1}\sum_{t=1}^{i}\big\langle A^{i-t}w_{t-1}, e_k\big\rangle\overline{\sum_{t=1}^{i}\big\langle A^{i-t}w_{t-1}, e_l\big\rangle} \\ &= 0\end{aligned} \tag{3.18}$$

3. *diagonal entries of inverse sample covariance matrix are inverse distance squared between row and conjugate hyperplane, to be precise:*

$$\left\langle (\sum_{t=0}^{N-1} x_t x_t^*)^{-1} e_j, e_j \right\rangle = \frac{1}{d^2(y_j, n_j)} = v_{j,j}, \tag{3.19}$$

Proof. We follow the direction given in [19], let $v_j := \left(X_-X_-^*\right)^{-1}e_j$, in columns format

we have $X_-^* = [y_1, \ldots, y_n]$, $X_-^* e_j = y_j$, $n_j = \text{span}[y_i]_{i \neq j}$.

$$\left\langle X_-^* \big(X_- X_-^*\big)^{-1} e_j, X_-^* e_k \right\rangle = \delta_j(k) \tag{3.20}$$

Notice that inner product of $X_-^* \big(X_- X_-^*\big)^{-1} e_j$ with $X_-^* e_k =: y_k$ for $k \neq j$ is zero; so for each $j \in [n]$, $X_-^* \big(X_- X_-^*\big)^{-1} e_j$ is orthogonal to n_j. Since,

$$X_-^* v_j = v_{j,1} y_1 + \ldots + v_{j,n} y_n, \tag{3.21}$$

(3.17) and (3.18) readily follows. As $X_-^* \big(X_- X_-^*\big)^{-1} e_j$ is orthogonal to n_j, according to Corollary 1

$$\big\langle X_-^* \big(X_- X_-^*\big)^{-1} e_j, X_-^* e_j \big\rangle = \big\langle X_-^* \big(X_- X_-^*\big)^{-1} e_j, P_{n_j^\perp}(X_-^* e_j) \big\rangle \tag{3.22}$$

Furthermore, using (3.20) and properties of orthogonal projection:

$$\begin{aligned}
1 &= \big\langle X_-^* \big(X_- X_-^*\big)^{-1} e_j, X_-^* e_j \big\rangle \\
&= \big\langle X_-^* \big(X_- X_-^*\big)^{-1} e_j, P_{n_j^\perp}(X_-^* e_j) \big\rangle \big\langle P_{n_j^\perp}\big[X_-^* \big(X_- X_-^*\big)^{-1} e_j\big], X_-^* e_j \big\rangle \\
&= \sum_{i \neq j}^{n} v_{j,i} \big\langle \underbrace{P_{n_j^\perp} y_i, y_j}_{=0} \big\rangle + v_{j,j} \langle P_{n_j^\perp} y_j, y_j \rangle \\
&= v_{j,j} \langle P_{n_j^\perp} y_j, P_{n_j^\perp} y_j \rangle = v_{j,j} \| P_{n_j^\perp} y_j \|^2 = v_{j,j} d^2(y_j, n_j).
\end{aligned}$$

Therefore, $v_{j,j} = \frac{1}{d^2(y_j, n_j)}$ □

Remark 15. *Notice that total number of unknowns $v_{j,k}$ are $\frac{n(n+1)}{2}$ which equals distinct interaction potentials of observations$\langle y_j, y_k \rangle$. A remarkable advantage of enlisting these constraints is when we do not have independence between the rows as in*

Hermitian case, we still have:

$$d_j^{-2} = \frac{1 - \sum_{k \neq j} v_{j,k} \langle y_k, y_j \rangle}{\|y_j\|^2}, \text{ where } \sum_{k \neq j} v_{j,k} \langle y_k, y_j \rangle \leqslant 0 \text{ and}$$
$$\sum_{j=1}^{n} d_j^{-2} = \frac{\sum_{j=1}^{n} \prod_{l \neq j} \|y_l\|^2 \big(1 - \sum_{k \neq j} v_{j,k} \langle y_k, y_j \rangle\big)}{\prod_{j=1}^{n} \|y_j\|^2} \tag{3.23}$$

Spectrum of the precision matrix

$$\|(X_- X_-)^{-1}\|_F = \sqrt{Tr\big((X_- X_-^*)^{-2}\big)} = \sqrt{\sum_{j=1}^{n} d_j^{-4} + 2 \sum_{k>j} |v_{j,k}|^2} \tag{3.24}$$

As we will see shortly afterwards, elementwise estimation error is an inner product between the rows of the Gaussian ensemble and and columns of the pseudo-inverse, it will be helpful to translate preceding constraints into constraints on columns of pseudo-inverse.

Corollary 3. *[Exact Error in Frobenius norm] Let $\hat{e}_i$ be the canonical basis of $\mathbb{C}^N$*

$$\|A - \hat{A}\|_F = \sqrt{\sum_{j,k=1}^{n} \Big| \sum_{i=1}^{N} \langle w_i, e_j \rangle \langle c_k, \hat{e}_i \rangle \Big|^2}, \tag{3.25}$$

where for each $k \in [1, 2, \ldots, n]$, $c_k \in \mathbb{C}^N$, $\|c_k\|^2 = \frac{1}{d^2(y_k, n_k)}$ and satisfies:

$$\sum_{i=1}^{N} \Big[\sum_{t=1}^{i} \langle A^{i-t} w_{t-1}, e_j \rangle \Big] \langle c_k, \hat{e}_i \rangle = \delta_j(k), \tag{3.26}$$

for every $j \in [1, 2, \ldots, n]$

Proof. Consider the pseudo-inverse

$$X_-^*(X_-X_-^*)^{-1} = \begin{bmatrix} | & | & | & | & | \\ | & | & | & | & | \\ c_1 & c_2 & | & c_{n-1} & c_n \\ | & | & | & | & | \\ | & | & | & | & | \end{bmatrix}$$

as inner product of $X_-^*\big(X_-X_-^*\big)^{-1}e_j$ with $X_-^*e_k =: y_k$ for $k \neq j$ is zero; so for each $j \in [n]$, $c_j := X_-^*\big(X_-X_-^*\big)^{-1}e_j$ is orthogonal to n_j

$$\sum_{i=1}^{N}[X_-]_{j,i}\langle c_k, \hat{e}_i\rangle = \delta_j(k), \quad \textit{implies} \quad \sum_{i=1}^{N}\Big[\sum_{t=1}^{i}\langle A^{i-t}w_{t-1}, e_j\rangle\Big]\langle c_k, \hat{e}_i\rangle = \delta_j(k),$$

Also notice that $c_k := X_-^*(X_-X_-^*)^{-1}e_k$ and

$$\big\|c_k\big\|^2 = \big\langle X_-^*(X_-X_-^*)^{-1}e_k, X_-^*(X_-X_-^*)^{-1}e_k\big\rangle = \frac{1}{d^2(y_k, n_k)}. \tag{3.27}$$

□

3.1.1 Higher degree variant of Littlewood-Offord problem

Now we simply notice that elementwise error is an inner product between rows of Gaussian ensemble and a constrained random variable.

Remark 16. *Element-wise estimation error of A,* $[EX_-^*(X_-X_-^*)^{-1}]_{j,k}$ *for each* $j, k \in [n]$ *is an inner product, hence a scalar weighted random walk with weight functions*

coming from pseudo-inverse.

$$\sum_{i=1}^{N}[E]_{j,i}\langle c_k, \hat{e}_i\rangle = \underbrace{\sum_{i=1}^{N}\langle w_i, e_j\rangle\langle c_k, \hat{e}_i\rangle}_{\textit{Littlewood-Offord problem}} \ , \tag{3.28}$$

Furthermore, total error in Frobenius norm is a random polynomial:

$$\sum_{j,k=1}^{n}\sum_{i=1}^{N}\left[|\langle w_i, e_j\rangle\langle c_k, \hat{e}_i\rangle|^2 + \langle w_i, e_j\rangle\langle c_k, \hat{e}_i\rangle\sum_{l\neq i}\overline{\langle w_l, e_j\rangle\langle c_k, \hat{e}_l\rangle}\right]$$

In a situation where a solution of c_k is independent of the $j-$th row of Gaussian ensemlbe E, after conditioning $\sum_{i=1}^{N}\langle w_i, e_j\rangle\langle c_k, \hat{e}_i\rangle$ is similar to a scalar random walk, which is well studied under the name of *Littlewood Offord problem* and its' typical behavior is sensitive to the structure of constants c_k, as discussed in the section 1.4.2. On the other hand if c_k is itself some function of $[\langle w_i, e_j\rangle]_{j=1}^{N}$, bounding the error is atleast a *Quadratic variant of the Littlewood Offord problem*(see e.g., [24]). So understanding dependence of constants on covariates of E will be imperative in getting deep insights into working of OLS.

Example 2. Now we consider two extreme cases of stable dynamical systems and show how contraints on pseudoinverse have distinct dependencies on the rows of Gaussian ensemble in each case:

- $A = D$, diagonal block with only one eigenvalue λ, fix $k \in [n]$ then constraints on pseudo-inverse from (3.26) imply:

$$\sum_{i=1}^{N}[X_-]_{j,i}\langle c_k, \hat{e}_i\rangle = \sum_{i=1}^{N}\sum_{t=1}^{i}\lambda^{i-t}\langle w_{t-1}, e_j\rangle\langle c_k, \hat{e}_i\rangle = \delta_j(k)$$

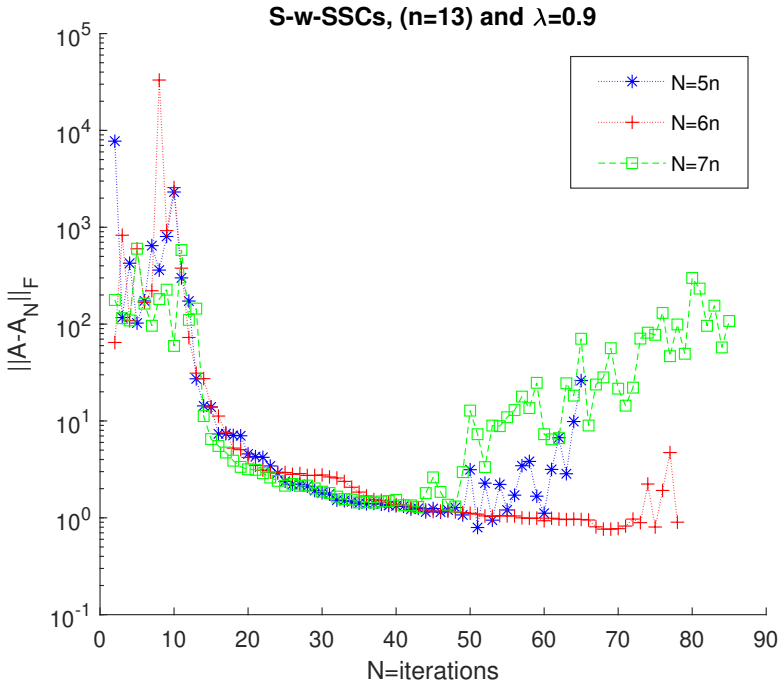

Figure 3.1: Estimation error worsening with increase in iterations

Notice that for fixed k, preceding constraint is only a function of the $j-$th row of Gaussian ensemble, on the other hand when

- $A = J_n(\lambda)$, constraint is a function of rows $[j, j+1, \ldots, n]$, precisely:

$$\delta_j(k) = \sum_{i=1}^{N} \sum_{t=1}^{i} \sum_{m=0}^{(i-t)\wedge(n-j)} \binom{i-t}{m} \lambda^{i-t-m} \langle w_{t-1}, e_{j+m} \rangle \langle c_k, \hat{e}_i \rangle \tag{3.29}$$

Chapter 4

Spectral statistics of structured random matrices

Most of the work in this chapter is from authors work in [25]. Performance of least squares method for learning system parameters by a single observed trajectory of unknown dynamical system has been extensively studied, see e.g., [1],[2],[3] and references therein. A fundamental limitation of these existing work stems from the fact that sample complexity is in terms of eigenvalues of various Grammians: which is expected value of the sample covariance matrix. Apart from insufficiency of expectation to characterize typical order of the relevant quantity, Grammian eigenvalues hide dependence on length of the simulated trajectory N and dimension of the underlying dynamical systems n. These issues severely limit our understanding of regression on dependent data and turns out that we can non-asymptoticlly learn correct system parameters under stability assumption for low dimensional(essentially scalar) dynamical system and in high dimensions when underlying state-transition matrix is Hermitian(which is essentially multiple independent scalar dynamical systems). Throughout this paper we will work under high dimensional framework.

Formalizing the approach, conclusions and main results: The biggest obstruction in our current understanding of regression on dependent data is using eigenvalues of the naive' variant of sample covariance matrix i.e., Grammian matrix instead of the random eigenvalues of the sample covariance matrix to conclude OLS performance. In fact this task is challenging: take for example $N \times n$ rectangular Gaussian ensemble E with its elements i.i.d standard normals. Let $\sigma_1(E) \geqslant \sigma_2(E) \geqslant \ldots \sigma_n(E)$

be the singular values of the Gaussian ensemble, even though all the enteries are independent but singular values are highly correlated,[26]. So it comes as no surprise why no attempt had been made towards investigating spectral statistics $\left[\sigma_j^2(X_-)\right]_{j\in[n]}$, where recall that $\sigma_j^2(X_-) = \lambda_j\big(X_-X_{-*}\big)$. Let $[y_j]_{j\in[n]}$ be the rows of the data matrix X_-, then notice that sample covariance matrix is essentially

$$X_-X_-^* := \begin{bmatrix} \langle y_1, y_1\rangle & \dots & \langle y_1, y_{k+1}\rangle & \dots & \langle y_1, y_n\rangle \\ \vdots & \dots & \vdots & \dots & \vdots \\ \langle y_{k+1}, y_1\rangle & \dots & \langle y_{k+1}, y_{k+1}\rangle & \dots & \langle y_{k+1}, y_n\rangle \\ \vdots & \dots & \vdots & \dots & \vdots \\ \langle y_n, y_1\rangle & \dots & \langle y_n, y_{k+1}\rangle & \dots & \langle y_n, y_n\rangle. \end{bmatrix}$$

In order to get accurate insights into least squares on Markovian data, we need not just accurate estimates of the edge of the spectrum $\lambda_1\big(X_-X_{-*}\big), \lambda_n\big(X_-X_{-*}\big)$ but that of the bulk as well. Now if we new for the *concentration behavior or* $\left[\big\langle y_j, y_k\big\rangle\right]_{j,k\in[n]}$ then we could use tools from perturbation theory e.g., Gershgorins' theorem: *all the eigenvalues of* $X_-X_-^*$ *lies inside disc centered at* $\langle y_j, y_j\rangle$ *with radius* $\sum_{k\neq j}^n |\langle y_j, y_k\rangle|$ *for some* $j \in [n]$ and Cauchy's interlacing theorem: *smallest eigenvalue is upper bounded by typical size of the smallest row and bulk eigenvalues can be interlaced by the eigenvalues of a lower dimensional sample covariance matrix* to get sharp estimates of spectral statistics of the sample covariance matrix *explicit in* N *and* n. Diagonal entries of the sample covariance matrix $\big\langle y_j, y_j\big\rangle$ are essentially typical size of each row and off-diagonal entries $\big\langle y_j, y_k\big\rangle_{j\neq k}$ *are correlation/interaction between the rows.* Mere assumption of spectral radius strictly inside unit disc is insufficient to characterize these interactions and row variances. As observed recently in the work of [7], when state transition is hermitian rows of the data matrix are essentially independent: spatial correlations are minor compared to the typical size of the row.

However when state-transition matrix is not Hermitian and distinct eigenvalues of A have *large discrepancy between their algebraic and geometric multiplicity then spatial correlation are strong.* Analytic display of strong spatial correlations appear in off diagonal terms: that can become large, leading to transience of OLS even under control theoretic stability assumption. Therefore, along with stability we need to characterize dynamical system based on spatial dependencies of its generated trajectory. We provide spectral statistics for the two extreme cases with same spectral radius but one spatially independent(Hermitian case) and other spatially inseparable,S-w-SSCs: only one distinct eigenvalue with algebraic and geometric multiplicity having a difference of $n-1$, which includes single Jordan block of size n. Combining the results from spectral theorem of non-Hermitian operators, perturbation theory and concentration of measure phenomenon: we manage to get first quantitative handle on various spectral statistics of the sample covariance matrix explicit in dimension of the state space n and trajectory length N as shown in:

Typical Size	**S-w-SSCs**	**Hermitian**
Largest Eigenvalue: $\lambda_1\big(X_-X_-^*\big)$	$\Omega\Bigg(\bigg\lfloor\frac{N}{n}\bigg\rfloor e^{\alpha_\lambda n}\Bigg)$	$\mathcal{O}\Big(N+n\sqrt{N}\Big)$
Smallest Eigenvalue: $\lambda_n\big(X_-X_-^*\big)$	$\mathcal{O}\Big(N+\sqrt{N}\Big)$	$\mathcal{O}\Big(N+\sqrt{N}\Big)$
Eigenvalue Spread	$\mathcal{O}\big(e^{\alpha_\lambda n}\big)$	$\mathcal{O}\big(n\sqrt{N}\big)$
Temporal interaction: $\langle y_j, y_j\rangle$	same order as spatial	$N \pm \mathcal{O}\sqrt{N}$
Spatial interaction: $\langle y_j, y_k\rangle$	same order as temporal	$\pm\mathcal{O}\sqrt{N}$
Largest singular value: $\sigma_1(EX_-^*)$	$\Omega\Bigg(\sqrt{n\bigg\lfloor\frac{N}{n}\bigg\rfloor e^{\alpha_\lambda n}}\Bigg)$	$\mathcal{O}\Big(\sqrt{nN}+n\Big)$

Error Statistics via negative moments of Sample Covariance matrix: Let $\sigma_1(X_-) \geqslant \sigma_2(X_-) \geqslant \ldots \sigma_n(X_-) > 0$, be the singular values of the data matrix. One can consider following error approximations for least squares:

1.

$$\sigma_n(EX_-^*)\sqrt{\sum_{j=1}^{n}\frac{1}{\sigma_j^4(X_-)}} \leqslant \left\|A-\hat{A}\right\|_F \leqslant \sigma_1(EX_-^*)\sqrt{\sum_{j=1}^{n}\frac{1}{\sigma_j^4(X_-)}}, \tag{4.1}$$

2.

$$\sqrt{\sum_{j=1}^{n}\frac{1}{n\sigma_j^2(X_-)}} \leqslant \left\|A-\hat{A}\right\|_F \leqslant \sqrt{\sum_{j=1}^{n}\frac{n}{\sigma_j^2(X_-)}} \tag{4.2}$$

Most important takeaway is that we need to compute negative moments of the sample covariance matrix to really understand performance of least squares. Which is a very difficult task, but if we can somehow get really tight estimate of all the singular values then there is hope. We will first focus on the extreme singular values of X_- and then come to bulk and in the end we will analyse extreme singular values of EX_-^*.

4.1 Statistics of the Largest Eigenvalue

Extreme singular values(largest and smallest singular values) of the data matrix can act as a sanity check for the performance of OLS. They also play a crucial role in numerical linear algebra(see e.g., [27] for more details). For rectangular matrices with i.i.d entries(centered and unit variance) everywhere as in Gaussian ensemble E, statistics of $\sigma_1(E)$ and $\sigma_n(E)$ are well known and tend to concentrate around $\sqrt{N}+\sqrt{n}$ and $\sqrt{N}-\sqrt{n}$ respeectively. Understanding upper tail behavior of $\sigma_1(E)$ requires discretization of S^{n-1}.

Definition 6. *(epsilon net of S^{n-1}) Finite subset, $\mathcal{N}_\epsilon(S^{n-1})$ called epsilon net of n dimensional sphere- with the property that given any $a \in S^{n-1}$ there exists $\hat{a} \in \mathcal{N}_\epsilon(S^{n-1})$ such that $\|a-\hat{a}\|_2 \leqslant \epsilon$*

Lemma 4. *Metric entropy of sphere is exponential w.r.t dimension of the underlying state space, precisely for any* $\epsilon \in (0, 1]$*:*

$$|\mathcal{N}_\epsilon(S^{n-1})| \leqslant \left(\frac{3}{\epsilon}\right)^n. \tag{4.3}$$

Recall, $\sigma_1(E) = \sup_{a\in S^{n-1}} \|E^*a\|$ but E is random and one does not know in advance for which $a^{'} \in S^{n-1}$ supremum is achieved, neither we can upper bound $P(\sup_{a\in S^{n-1}} \|E^*a\| > \delta)$ via union bound as the S^{n-1} is not countable. To circumvent this issue, one can discretize S^{n-1} into finite subset, $\mathcal{N}_\epsilon(S^{n-1})$ when combined with triangle inequality, reveals:

$$\sigma_1(E) \leqslant \left(\frac{1}{1-\epsilon}\right) \sup_{a\in\mathcal{N}_\epsilon(S^{n-1})} \|E^*a\|. \tag{4.4}$$

Therefore,

$$\begin{aligned}
&\mathbb{P}\bigg(\sup_{a\in S^{n-1}} \|E^*a\| \geqslant \delta\bigg) \leqslant \mathbb{P}\bigg(\sup_{a\in\mathcal{N}_\epsilon(S^{n-1})} \|E^*a\| \geqslant (1-\epsilon)\delta\bigg) \\
&\leqslant \sum_{a\in\mathcal{N}_\epsilon(S^{n-1})} \mathbb{P}\bigg(\|E^*a\| \geqslant (1-\epsilon)\delta\bigg) \leqslant \left(\frac{3}{\epsilon}\right)^n \mathbb{P}\bigg(\|z_N\|^2 \geqslant (1-\epsilon)^2\delta^2\bigg),
\end{aligned} \tag{4.5}$$

where, first inequality in (4.5) follows from union bound and second from metric entropy of n dimensional unit sphere. Notice that: *regardless of exact realization of 'a' on n dimensional unit sphere,* $E^*a = z_N \sim N(0, I_N)$ and a simple exponential moment calculation on the last expression of (4.5) will reveal the correct behavior $\sigma_1(E)$. Similar discretization approaches have been considered in literature for understanding statistics of extreme singular values of the data matrix. However, as suggested by authors work in [7] on spectral theorem for non-Hermitian linear operators, discrepancy between algebraic and geometric multiplicities of distinct eigenvalues of A add

structure to data matrix and naive discretization will be wasteful; except for Hermitian and all the same eigenvalues. Furthermore, rows of the data matrix X_- can be divided into blocks, independent of each other and concentration of $\|X_-^* a\|_2$ for some $a \in S^{n-1}$ can be better understood by also restricting a onto invariant subspaces $[M_{\lambda_i}]_{i\in[K]}$ of A, which we now explore in further detail.

4.1.1 Lower bound via typical size of rows

Notice that,

$$\mathbb{P}\bigg(\|X_-^* a\| \geqslant \delta\bigg) = \mathbb{P}\bigg(\Big\| \sum_{m=1}^{K} X_-^* P_{\lambda_m} a\Big\|^2 \geqslant \delta^2\bigg), \tag{4.6}$$

where $[X_-^* P_{\lambda_m} a]_{m=1}^K$ are independent of each other: orthogonal projections of Gaussians are independent. Furthermore, OLS error in Frobenius norm is independent of choice of the basis, so w.l.o.g we can assume $X_-^* P_{\lambda_m} a \in \mathbb{R}^{N^{\otimes D^+_{\lambda_m}}}$, where we use shorthand $D^+_{\lambda_m} := (D_{\lambda_m} + 1)$ and recall $\sum_{m=1}^K D^+_{\lambda_m} = n.$, for each $m \in [K]$ let $S_m(\lambda) := \sum_{i=1}^{m-1} D^+_{\lambda_i}$

$$P_{\lambda_m} a := \big(0, \ldots, 0, a_{S_m(\lambda)+1}, \ldots, a_{S_m(\lambda)+D^+_{\lambda_m}}, 0, \ldots, 0\big) \tag{4.7}$$

Recall that $\sum_{i=1}^N [X_-]_{j,i} = \sum_{i=1}^N \bigg[\sum_{t=1}^i \langle A^{i-t} w_{t-1}, e_j\rangle\bigg]$, therefore:

$$X_-^* P_{\lambda_m} a = \bigg[a_{S_m(\lambda)+1} \overline{[X_-]}_{[S_m(\lambda)+1,:]}, a_{S_m(\lambda)+2} \overline{[X_-]}_{[S_m(\lambda)+2,:]}, \\ \ldots, a_{S_m(\lambda)+D^+_{\lambda_m}} \overline{[X_-]}_{[S_m(\lambda)+D^+_{\lambda_m},:]} \bigg],$$

where we used : to denote all the columns of the data matrix. For each time $t \in \mathbb{N}$, $w_t^m \in N\big(0, I_{D^+_{\lambda_m}}\big)$ i.e., i.i.d isotropic Gaussian of dimension $D^+_{\lambda_m}$. Hence, for $i \in [N]$

and $l \in [D^+_{\lambda_m}]$:

$$\overline{[X_-]}_{[S_m(\lambda)+l,i]} = \sum_{t=1}^{i} \overline{\langle A^{i-t}_{\lambda_m} w^m_{t-1}, e_l \rangle} = \sum_{t=1}^{i} \langle e_l, A^{i-t}_{\lambda_m} w^m_{t-1} \rangle = \sum_{t=1}^{i} \langle e_l, (\lambda_m I + N)^{i-t} w^m_{t-1} \rangle$$
$$= \sum_{t=1}^{i} \sum_{p=0}^{i-t} \binom{i-t}{p} \overline{\lambda_m^{i-t-p}} \langle e_{l+p}, w^m_{t-1} \rangle = \sum_{t=1}^{i} \sum_{p=0}^{(i-t)\wedge(D^+_{\lambda_m}-l)} \binom{i-t}{p} \overline{\lambda_m^{i-t-p}} \langle e_{l+p}, w^m_{t-1} \rangle \tag{4.8}$$

where $A^{i-t}_{\lambda_m}$ is a Linear transformation from and onto an $D^+_{\lambda_m}$ dimensional A- invariant subspace, i.e., Jordan block of size $D^+_{\lambda_m}$ with eigenvalue λ_m Consequently,

$$\|X^*_- P_{\lambda_m} a\|^2 = \sum_{i=1}^{N} \left(\sum_{l=1}^{D^+_{\lambda_m}} \sum_{t=1}^{i} a_{S_m(\lambda)+l} \langle e_l, A^{i-t}_{\lambda_m} w^m_{t-1} \rangle \right)^2 \tag{4.9}$$
$$= \sum_{i=1}^{N} \left(\sum_{l=1}^{D^+_{\lambda_m}} a_{S_m(\lambda)+l} \sum_{t=1}^{i} \sum_{p=0}^{(i-t)\wedge(D^+_{\lambda_m}-l)} \binom{i-t}{p} \overline{\lambda_m^{i-t-p}} \langle e_{l+p}, w^m_{t-1} \rangle \right)^2. \tag{4.10}$$

Therefore,

$$\mathbb{P}\left(\left\| X^*_- a \right\| \geqslant \delta \right) = \mathbb{P}\left(\sum_{m=1}^{K} \left\| X^*_- P_{\lambda_m} a \right\|^2 \geqslant \delta^2 \right)$$
$$\leqslant e^{-s\delta^2} \prod_{m=1}^{K} \mathbb{E}\left[e^{s \sum_{i=1}^{N} \left(\sum_{l=1}^{D^+_{\lambda_m}} a_{S_m(\lambda)+l} \sum_{t=1}^{i} \sum_{p=0}^{(i-t)\wedge(D^+_{\lambda_m}-l)} \binom{i-t}{p} \overline{\lambda_m^{i-t-p}} \langle e_{l+p}, w^m_{t-1} \rangle \right)^2} \right], \tag{4.11}$$

where inequality (4.11) follows from Markov inequality combined with independence of $[X^*_- P_{\lambda_m} a]$ for each $m \in [K]$(spatial independence between row blocks of the data matrix). Now notice that if $[\lambda_m]_{m\in[K]}$ are all positive, then for each $m \in [K]$, and $i \in [N]$ sufficiently large with l as variable,

$$a_{S_m(\lambda)+l} \sum_{t=1}^{i} \sum_{p=0}^{(i-t)\wedge(D^+_{\lambda_m}-l)} \binom{i-t}{p} \overline{\lambda_m^{i-t-p}} \langle e_{l+p}, w^m_{t-1} \rangle, \tag{4.12}$$

will have the largest typical size for $l = 1$, so supremum over all $a \in S^{D_{\lambda_m}}$ for $\big(\sum_{l=1}^{D^+_{\lambda_m}} a_{S_m(\lambda)+l} \sum_{t=1}^{i} \sum_{p=0}^{(i-t)\wedge(D^+_{\lambda_m}-l)} \binom{i-t}{p} \overline{\lambda_m^{i-t-p}} \langle e_{l+p}, w^m_{t-1}\rangle\big)^2$, would be in fact

$$\left(\sum_{t=1}^{i} \sum_{p=0}^{(i-t)\wedge(D^+_{\lambda_m}-l)} \binom{i-t}{p} \overline{\lambda_m^{i-t-p}} \langle e_{l+p}, w^m_{t-1}\rangle\right)^2. \tag{4.13}$$

So when discrepancies of eigenvalues are large typical size of the row should be a good approximate of $\sigma_1(X_-)$. Thus, at least when rows are correlated we do not need to discretize the unit sphere. Also notice that

$$\begin{aligned} P\bigg(\sigma_1(X_-) \leqslant \delta\bigg) = P\bigg(\sup_{a\in S^{n-1}} \|X_-^* a\| \leqslant \delta\bigg) = P\bigg(\sup_{a\in S^{n-1}} \Big\|\sum_{j=1}^{n} a_j y_j\Big\| \leqslant \delta\bigg) \\ \leqslant P\bigg(\max_{j\in[n]} \|y_j\|_2 \leqslant \delta\bigg), \end{aligned} \tag{4.14}$$

where $[y_j]_{j=1}^n$ are the rows of the data matrix and if $k = \arg\max_{j\in[n]} \|y_j\|$ then last inequality follows by setting $a_k = 1$ and $a_j = 0$ for $j \neq k$.

Remark 17. *If the typical size of any row of the data matrix is 'large' then so is the magnitude of $\sigma_1(X_-)$. Typical size of the row is essentially dictated by non-Hermitian variant of the spectral theorem which define the weighted constants on standard normals that populate any given row of the data matrix.*

4.1.2 Covariates from stable dynamics can suffer from the curse of dimensionality

It was noted in [7], for the case of dynamics correspodning to S-w-SSCs, each covariate in $j-$ th row of the data matrix is a weighted sum of i.i.d standard normals that make up rows $[j, \ldots, n]$ of Gaussian Ensemble E. So as one would suspect that first row is most susceptible to having the largest typical size. Which turns out to be exponential in dimension of the state space

Proposition 4. *$N > n$ and $\lambda \in (\frac{1}{2}, 1)$ then typical order of first row of the data matrix*

corresponding to dynamics generated from n dimensional S-w-SSCs is exponential in n, precisely said

$$\|y_1\| \geqslant \Omega\left(\left\lfloor \frac{N}{n} \right\rfloor^{\frac{1}{2}} e^{\frac{\alpha_\lambda n}{2}}\right) \tag{4.15}$$

where α_λ entirely depends on λ.

Proof. Let $[\hat{e}_i]_{i\in N}$ be the canonical basis of $\mathbb{R}^N$. As, $j-$ th row and $i-$ th column of the data matrix corresponding to S-w-SSCs case can be compactly written as:

$$\langle y_j, \hat{e}_i\rangle = \sum_{t=1}^{i} \sum_{m=0}^{(i-t)\wedge(n-j)} \binom{i-t}{m} \lambda^{i-t-m} \langle w_{t-1}, e_{m+j}\rangle. \tag{4.16}$$

Thus revealing: $\langle y_1, \hat{e}_n\rangle$ is a weighted sum of independent $\frac{n(n+1)}{2}$ standard normals, more precisely:

$$\langle y_1, \hat{e}_n\rangle = \sum_{t=1}^{n} \sum_{m=0}^{(n-t)} \binom{n-t}{m} \lambda^{(n-t-m)} \langle w_{t-1}, e_{m+1}\rangle$$

which is centered with variance:

$$\mathbb{E}\big(\langle y_1, \hat{e}_n\rangle^2\big) = \sum_{k=1}^{n} \sum_{m=0}^{n-k} \binom{n-k}{m}^2 \lambda^{2(n-k-m)} \tag{4.17}$$

upper and lower bounded via Stirling type approximation:

$$4^n\lambda^{2n} \sum_{k=1}^{n} \frac{1}{4^k\lambda^{2k}\sqrt{\pi(n-l+\frac{1}{3})}} \leqslant \mathbb{E}\big(\langle y_1, \hat{e}_n\rangle^2\big) \leqslant 4^n\lambda^{2n} \sum_{k=1}^{n} \frac{1}{4^k\lambda^{2k}\sqrt{\pi(n-l+\frac{1}{4})}} \tag{4.18}$$

Let $L_{\lambda,n}(1) := \sum_{k=1}^{n} \frac{1}{4^k\lambda^{2k}\sqrt{\pi(n-k+\frac{1}{3})}}$ and $U_{\lambda,n}(1) := \sum_{k=1}^{n} \frac{1}{4^k\lambda^{2k}\sqrt{\pi(n-k+\frac{1}{4})}}$ So as soon as $\lambda > \frac{1}{2}$, $n-$ th element of first row has a typical size which is exponential in

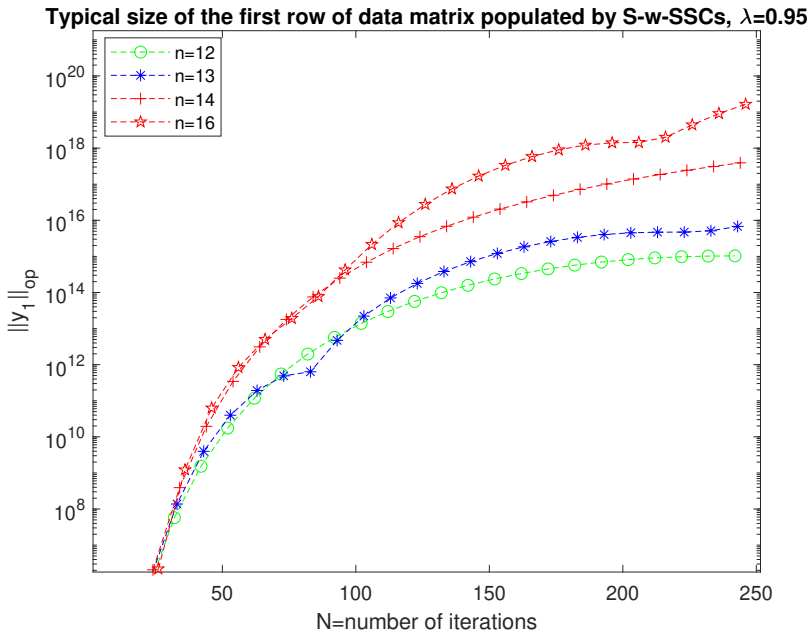

Figure 4.1: ℓ_2 norm of the data matrixs' first row populated by $n-$ dimensional S-w-SSCs suffers from curse of dimensionality

dimension of the state space. But this trend is periodic: with period n, so lower bound follows. □

Corollary 4. *Given $\lambda \in (\frac{1}{2}, 1)$, almost surely largest singular of data matrix generated from $n-$ dimensional S-w-SSCs with eigenvalue λ suffers from curse of dimnensionality, more precisely:*

$$\sigma_1(X_-) \geqslant \Omega\left(\left\lfloor \frac{N}{n} \right\rfloor^{\frac{1}{2}} e^{\frac{\alpha_\lambda n}{2}}\right) \tag{4.19}$$

We believe that we can take away the complicated floor function with a linear function of N and n

Proposition 5. *$N > n$ and $\lambda \in (\frac{1}{2}, 1)$ then typical order of first row of the data matrix corresponding to dynamics generated from n dimensional S-w-SSCs is exponential in*

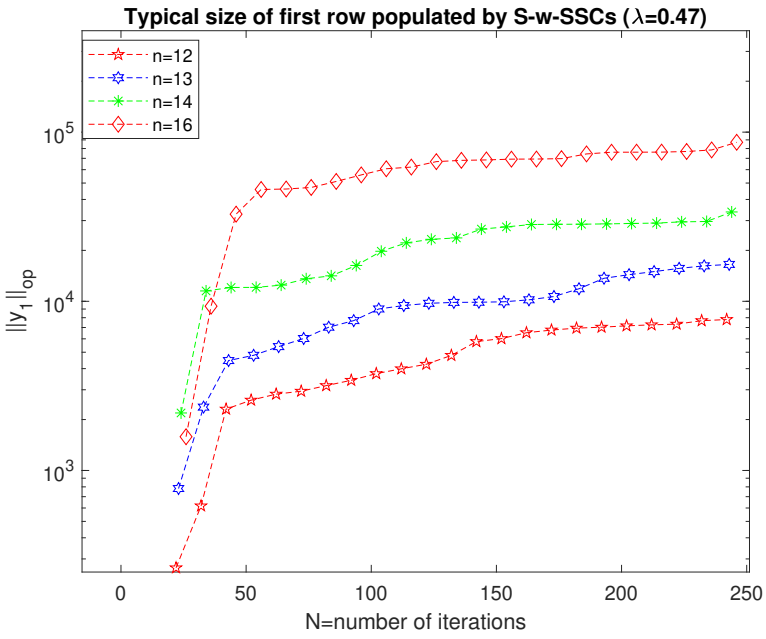

Figure 4.2: Estimates in Proposition 5 are optimal: S-w-SSCs(with $\lambda = 0.47$) do not suffer from curse of dimensionality

n, precisely said

$$(N-n+1)\Omega(4^n\lambda^{2n}) + \sum_{i>n}^{N}\sum_{m=0}^{n-1}\lambda^{2(i-m)}\sum_{l=1}^{i-n}\zeta_{m,l}(i,\lambda)\exp\left(-O\left(2\frac{m^3}{(i-l)^2}\right)\right)$$
$$\preccurlyeq Var(y_1) \preccurlyeq \sum_{i\geqslant n}^{N}\sum_{m=0}^{n-1}\lambda^{2(i-n+1)}\sum_{l=1}^{i-n}\zeta_{m,l}(i,\lambda) + (N-n+1)O_\lambda(4^n) \tag{4.20}$$

where $\zeta_{m,l}(N,\lambda) := \frac{1}{m!m!}\frac{(N-l)^{2m}}{\lambda^{2l}}\exp\left(-\frac{m(m-1)}{N-l}\right)$

Proof. We begin with investigation of the typical size of $N-$ th covariate in the first row of data matrix, $[X_-]_{1,N}$. Let $\lambda \in (0,1)$ and notice that $[X_-]_{1,N}$ is normally

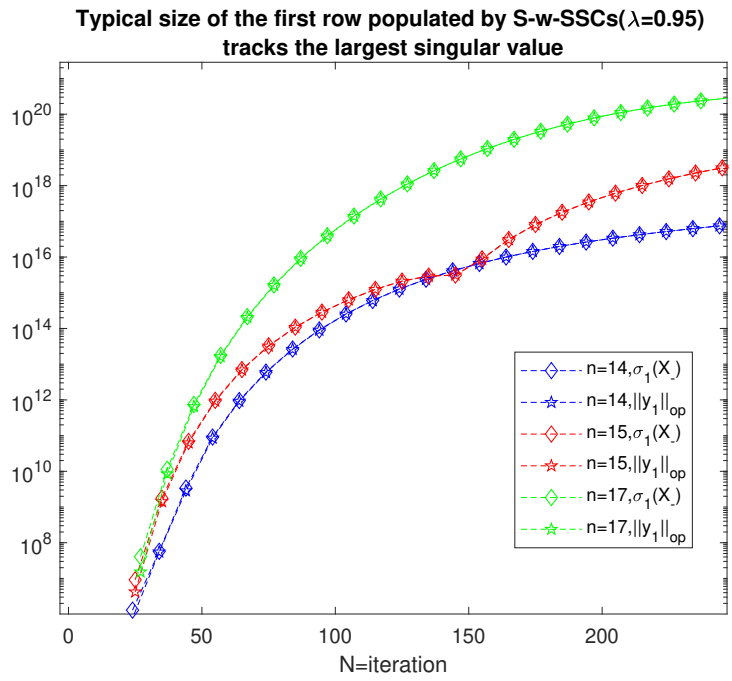

Figure 4.3: Largest singular value and ℓ_2 norm of the data matrixs' first row populated by $n-$ dimensional S-w-SSCs are the same

distributed with variance:

$$\sum_{l=1}^{N}\lambda^{2(N-l)}\sum_{m=0}^{(N-l)\wedge(n-1)}\binom{N-l}{m}^2\lambda^{-2m}\geqslant\sum_{l=1}^{N-n}\lambda^{2(N-l)}\sum_{m=0}^{n-1}\binom{N-l}{m}^2\lambda^{-2m}$$

$$+\sum_{l=N-n+1}^{N}\lambda^{2(N-l)}\sum_{m=0}^{N-l}\binom{N-l}{m}^2\geqslant\sum_{l=1}^{N-n}\lambda^{2(N-l)}\sum_{m=0}^{n-1}\binom{N-l}{m}^2\lambda^{-2m}$$

$$+4^N\lambda^{2N}\sum_{l=N-n+1}^{N}\frac{1}{\sqrt{16^l\lambda^{4l}\pi(N-l+\frac{1}{3})}}=\underbrace{\sum_{l=1}^{N-n}\lambda^{2(N-l)}\sum_{m=0}^{n-1}\binom{N-l}{m}^2\lambda^{-2m}}_{C_\lambda(N,n)}$$

$$+\frac{4^N\lambda^{2N}}{\sqrt{16^{(N-n)}\lambda^{4(N-n)}}}\underbrace{\sum_{l=1}^{n}\frac{1}{\sqrt{16^l\lambda^{4l}\pi(n-l+\frac{1}{3})}}}_{R_\lambda(n)}=\sum_{l=1}^{N-n}\lambda^{2(N-l)}\sum_{m=0}^{n-1}\binom{N-l}{m}^2\lambda^{-2m}$$

$$+ 4^n \lambda^{2n} \sum_{l=1}^{n} \frac{1}{\sqrt{16^l \lambda^{4l} \pi (n-l+\frac{1}{3})}} \geqslant \sum_{l=1}^{N-n} \lambda^{2(N-l)} \sum_{m=0}^{n-1} \binom{N-l}{m}^2 \lambda^{-2m} + \frac{4^n}{\lambda^2} \lambda^{2n} \sum_{l=1}^{n} \frac{1}{\sqrt{16^l \pi (n-l+\frac{1}{3})}} \tag{4.21}$$

Therefore, for every $\lambda \in (\frac{1}{2}, 1)$ typical behavior(standard deviation) of $[X_-]_{1,N}$ is greater than $\Omega_\lambda(e^n) + C_\lambda(N,n)$, where $C_\lambda(N,n)$ can be lower bounded as:

$$\begin{aligned}
C_\lambda(N,n) &= \lambda^{2N} \sum_{l=1}^{N-n} \sum_{m=0}^{n-1} \binom{N-l}{m}^2 \lambda^{-2(l+m)} \\
&= \lambda^{2N} \sum_{l=1}^{N-n} \sum_{m=0}^{n-1} \frac{\left(N-l\right)^{2m} \prod_{p=1}^{m-1} \left(1-\frac{p}{N-l}\right)^2}{\lambda^{2(l+m)} m! m!} \\
&= \lambda^{2N} \sum_{m=0}^{n-1} \frac{1}{\lambda^{2m} m! m!} \left(\frac{\left(N-1\right)^{2m} \prod_{p=1}^{m-1} \left(1-\frac{p}{N-1}\right)^2}{\lambda^{2(1)}} \right. \\
&\qquad \left. + \ldots + \frac{\left(N-n+1\right)^{2m} \prod_{p=1}^{m-1} \left(1-\frac{p}{N-n+1}\right)^2}{\lambda^{2(N-n+1)}} \right) \\
&\geqslant \lambda^{2N} \sum_{m=0}^{n-1} \frac{1}{\lambda^{2m} m! m!} \sum_{l=1}^{N-n} \frac{\left(N-l\right)^{2m}}{\lambda^{2l}} \exp\left(-\frac{m(m-1)}{N-l} \right) \\
&\qquad \exp\left(-O\left(2\frac{m^3}{(N-l)^2} \right) \right).
\end{aligned} \tag{4.22}$$

Now we upper bound the typical size of the covariate $[X_-]_{1,N}$ which is normally

distributed with variance:

$$\begin{aligned}
&\sum_{l=1}^{N}\lambda^{2(N-l)}\sum_{m=0}^{(N-l)\wedge(n-1)}\binom{N-l}{m}^2\lambda^{-2m}=\sum_{l=1}^{N-n}\lambda^{2(N-l)}\sum_{m=0}^{n-1}\binom{N-l}{m}^2\lambda^{-2m}\\
&+\sum_{l=N-n+1}^{N}\lambda^{2(N-l)}\sum_{m=0}^{N-l}\binom{N-l}{m}^2\lambda^{-2m}\\
&\leqslant\lambda^{2(N-n+1)}\sum_{l=1}^{N-n}\frac{1}{\lambda^{2l}}\sum_{m=0}^{n-1}\frac{\big(N-l\big)^{2m}\prod_{p=1}^{m-1}\big(1-\frac{p}{N-l}\big)^2}{m!m!}\\
&+\frac{4^N}{\lambda^2}\sum_{l=N-n+1}^{N}\frac{1}{\sqrt{16^l\pi(N-l+\frac{1}{4})}}\\
&\leqslant\lambda^{2(N-n+1)}\sum_{m=0}^{n-1}\frac{1}{m!m!}\sum_{l=1}^{N-n}\frac{(N-l)^{2m}}{\lambda^{2l}}\exp\bigg(-\frac{m(m-1)}{N-l}\bigg)\\
&+\frac{4^n}{\lambda^2}\sum_{l=1}^{n}\frac{1}{\sqrt{16^l\pi(n-l+\frac{1}{4})}}.
\end{aligned}\tag{4.23}$$

Therefore for all $N\geqslant n$ and $\lambda\in(\frac{1}{2},1)$

$$C_\lambda(N,n)+\Omega(4^n\lambda^{2n})(4^n\lambda^{2n})\leqslant[X_-]_{1,N}\leqslant C_\lambda(N,n)+O_\lambda(4^n),\tag{4.24}$$

combined with analysis related to birthday paradox problem we have:

$$\begin{aligned}
&\lambda^{2N}\sum_{m=0}^{n-1}\lambda^{-2m}\sum_{l=1}^{N-n}\frac{1}{m!m!}\frac{\big(N-l\big)^{2m}}{\lambda^{2l}}\exp\bigg(-\frac{m(m-1)}{N-l}\bigg)\exp\bigg(-O\Big(2\frac{m^3}{(N-l)^2}\Big)\bigg)\\
&+O(4^n\lambda^{2n})\leqslant[X_-]_{1,N}\leqslant\\
&\qquad\lambda^{2(N-n+1)}\sum_{m=0}^{n-1}\sum_{l=1}^{N-n}\underbrace{\frac{1}{m!m!}\frac{(N-l)^{2m}}{\lambda^{2l}}\exp\bigg(-\frac{m(m-1)}{N-l}\bigg)}_{\zeta_{m,l}(N,\lambda)}+O_\lambda(4^n)
\end{aligned}$$

Therefore variance of first row is:

$$\sum_{i>n}^{N}[X_-]_{1,i} \preccurlyeq \sum_{i>n}^{N}\sum_{m=0}^{n-1}\lambda^{2(i-n+1)}\sum_{l=1}^{i-n}\zeta_{m,l}(i,\lambda) + (N-n+1)O_\lambda(4^n)$$

$$\sum_{i>n}^{N}[X_-]_{1,i} \succcurlyeq \sum_{i>n}^{N}\sum_{m=0}^{n-1}\lambda^{2(i-m)}\sum_{l=1}^{i-n}\zeta_{m,l}(i,\lambda)\exp\left(-O\left(2\frac{m^3}{(i-l)^2}\right)\right)$$
$$+ (N-n+1)\Omega(4^n\lambda^{2n})$$

□

4.2 Eigenvalue Statistics of the Sample covariance Matrix

Overwhelming challenge in unraveling the typical order of the eigenvalues associated to the sample covariance matrix(which in turn unravels the fate of OLS) is first establishing the concentration behavior of each individual elements of the sample covariance matrix.

$$X_-X_-^* := \begin{bmatrix} \langle y_1, y_1\rangle & \dots & \langle y_1, y_{k+1}\rangle & \dots & \langle y_1, y_n\rangle \\ \vdots & \dots & \vdots & \dots & \vdots \\ \langle y_{k+1}, y_1\rangle & \dots & \langle y_{k+1}, y_{k+1}\rangle & \dots & \langle y_{k+1}, y_n\rangle \\ \vdots & \dots & \vdots & \dots & \vdots \\ \langle y_n, y_1\rangle & \dots & \langle y_n, y_{k+1}\rangle & \dots & \langle y_n, y_n\rangle \end{bmatrix}$$

which we will later combine with tools from the perturbation theory to study the localization of all the eigenvalues of the sample covariance matrix.

4.2.1 On measure concentration of Sample Covariance matrix

Hermitian Case:

Proposition 6. *When $A = A^*$, stable and w.l.o.g let λ,ρ be the eigenvalues and w_t, s_t be the standard normals corresponding to rows k and j, respectively. Then $\langle y_k, y_j \rangle$ is centered (i.e., expectated value is 0) and there exists positive constants $c_{2,\lambda}, c_{4,\lambda}, c_{\lambda,\rho}, d_{\lambda,\rho}, d_\lambda$ and d_ρ such that:*

$$\mathbb{E}(\langle y_k, y_j \rangle^2) = c_{\lambda,\rho}(N-1) + c_{\lambda,\rho}\Bigg[d_{\lambda,\rho}\big(1 - o_{(\lambda,\rho);N}(1)\big) - d_\lambda\big(1 - o_{\lambda;N}(1)\big) - d_\rho\big(1 - o_{\rho;N}(1)\big)\Bigg]$$

$$\mathbb{E}(\langle y_k, y_k \rangle^2) = 3c_{2,\lambda}^2\Bigg[(N-1) + c_{4,\lambda}\big(1 - o_{\lambda;N}^2(1)\big) - 2c_{2,\lambda}\big(1 - o_{\lambda;N}(1)\big)\Bigg].$$

where given $|\lambda| < 1$, $o_{\lambda;N}(1)$ is used to denote a term vanishing to 0 as N approaches infinity.

Proof. Simply notice that:

$$\langle y_k, y_j \rangle = \sum_{i=1}^{N-1}\left(\sum_{t=1}^{i}\lambda^{i-t}w_{t-1}\right)\left(\sum_{t=1}^{i}\rho^{i-t}s_{t-1}\right), \qquad \textit{Furthermore,} \qquad \mathbb{E}\left(\langle y_k, y_j \rangle^2\right)$$
$$= \sum_{i=1}^{N-1}\left(\sum_{t=1}^{i}\lambda^{2(i-t)}\right)\left(\sum_{t=1}^{i}\rho^{2(i-t)}\right) = \frac{\sum_{i=1}^{N-1}\big(1-\lambda^{2i}\big)\big(1-\rho^{2i}\big)}{(1-\lambda^2)(1-\rho^2)}$$
$$= \frac{N - \sum_{i=1}^{N-1}(\lambda^{2i} + \rho^{2i}) + (\lambda\rho)^{2i}}{(1-\lambda^2)(1-\rho^2)}.$$

Diagonal elements are centered at:

$$\mathbb{E}\langle y_k, y_k\rangle = \sum_{i=1}^{N-1} \mathbb{E}\left[\left(\sum_{t=1}^{i} \lambda^{i-t} w_{t-1}\right)^2\right] = \sum_{i=1}^{N-1}\sum_{t=1}^{i} \lambda^{2(i-t)} = c_{2,\lambda}(N-1) - c_{2,\lambda}^2\big(1 - o_{\lambda;N}(1)\big)$$

□

Now simply recall that:

$$\mathbb{E}(\langle y_k, y_k\rangle^2) = \frac{3}{(1-\lambda^2)^2}\sum_{i=1}^{N-1}(1-\lambda^{2i})^2 = \frac{3(N-1)}{(1-\lambda^2)^2} + \frac{3(1-\lambda^{4N})}{(1-\lambda^2)^2(1-\lambda^4)} - \frac{6(1-\lambda^{2N})}{(1-\lambda^2)^2(1-\lambda^2)}$$

Therefore variance of $\langle y_k, y_k\rangle$

$$3(N-1)c_{2,\lambda}^2 + 3c_{2,\lambda}^2 c_{4,\lambda}\big(1 - o_{\lambda;N}^2(1)\big) - 6c_{2,\lambda}^3\big(1 - o_{\lambda;N}(1)\big) - c_{2,\lambda}^2(N-1)^2 - c_{2,\lambda}^4\big(1 - o_{\lambda;N}(1)\big)^2 + 2c_{2,\lambda}^3(N-1)\big(1 - o_{\lambda;N}(1)\big)$$

Remark 18. *This implies that $\langle y_k, y_k\rangle$ is centered at $c_{2,\lambda}(N-1) - c_{2,\lambda}^2\big(1 - o_{\lambda;N}(1)\big)$ with fluctuation of size:*

$$\Big(3(N-1)c_{2,\lambda}^2 + 3c_{2,\lambda}^2 c_{4,\lambda}\big(1 - o_{\lambda;N}^2(1)\big) - 6c_{2,\lambda}^3\big(1 - o_{\lambda;N}(1)\big) - c_{2,\lambda}^2(N-1)^2 - c_{2,\lambda}^4\big(1 - o_{\lambda;N}(1)\big)^2 + 2c_{2,\lambda}^3(N-1)\big(1 - o_{\lambda;N}(1)\big)\Big)^{\frac{1}{2}}$$

On the other hand, now ignoring constants related to eigenvalues λ, ρ e.t.c)off-diagonal entries of the sample covariance matrix $(X\_X_-^*)$, $\langle y_k, y_j\rangle \in [-\sqrt{N}, \sqrt{N}]$ w.p 1. and $\langle y_k, y_k\rangle \in [N - \sqrt{N}, N + \sqrt{N}]$.

Theorem 19. *For Hermitian stable case, ignoring all the other parameters besides n and N, we have that for all $j \in [n]$:*

$$N - n\sqrt{N} \leqslant \lambda_j\big(X_- X_-^*\big) \leqslant N + n\sqrt{N} \tag{4.25}$$

Proof. According to Gershgorins' theorem all the eigenvalues of $X_- X_-^*$ lies inside disc centered at $\langle y_j, y_j \rangle$ with radius $\sum_{k \neq j}^{n} |\langle y_j, y_k \rangle|$ for some $j \in [n]$. Result follows from proposition 6 and preceding remark. □

Non-Hermitian case: Now we will study concentration for an extreme non Hermitian case S-w-SSCs: which allows for many intricate spatial and temporal correlations between elements of the data matrix. Evident from the previous section, giving typical size to the elements of the sample covariance matrix will require painstaking attention to time evolution of the dynamics in every row. Since the rows have a causal structure: i.e., y_k is itself a function of $[y_{k+1}, \ldots, y_n]$ so we will start from y_n. In this working draft, we managed to extract concentration behavior for the last two rows and it becomes evident how off diagonal terms are quiet large which will lead to transient behavior of OLS, but recursive framework of time evolution of one row in terms of rows beneath it will be extremely helpful in getting concentration for all the entries. Writing the rows of the data in compact form when dynamics correspond to $S - w - SSCs$:

$$\langle y_j, \hat{e}_i \rangle = \sum_{t=1}^{i} \sum_{m=0}^{(i-t)\wedge(n-j)} \binom{i-t}{m} \lambda^{i-t-m} \big\langle w_{t-1}, e_{m+j} \big\rangle$$

$$\textit{Furthermore,} \quad \langle y_{n-1}, \hat{e}_{i+1} \rangle = \lambda \langle y_{n-1}, \hat{e}_i \rangle + \langle w_i, e_{n-1} \rangle + \langle y_n, \hat{e}_i \rangle \tag{4.26}$$

$$\textit{Let,} \qquad s_i := \sum_{t=1}^{i} \lambda^{i-t} w_{t-1}$$

In order to understand the cross correlation between the rows of the data matrix, we leverage upon (4.26)

$$\begin{aligned}
\langle y_{n-1}, y_n\rangle &= \sum_{i=1}^{N-1}\Big(\sum_{t=1}^{i}\lambda^{i-t}\langle w_{t-1}, e_{n-1}\rangle + \sum_{t=1}^{i-1}\lambda^{i-t-1}\langle w_{t-1}, e_n\rangle\Big)\Big(\sum_{t=1}^{i}\lambda^{i-t}\langle w_{t-1}, e_n\rangle\Big)\\
&= \sum_{i=1}^{N-1}\big(\langle s_i, e_{n-1}\rangle + \lambda^{-1}\langle s_{i-1}, e_n\rangle\big)\langle s_i, e_n\rangle\\
&= \sum_{i=1}^{N-1}\big(\langle s_i, e_{n-1}\rangle + \lambda^{-1}\langle s_{i-1}, e_n\rangle\big)\big(\lambda\langle s_{i-1}, e_n\rangle + \langle w_{i-1}, e_n\rangle\big)\\
&= \sum_{i=1}^{N-1}\Bigg[\sum_{t=1}^{i-1}\lambda^{2(i-t)-1}\langle w_{t-1}, e_n\rangle^2 + \langle s_i, e_{n-1}\rangle\langle s_i, e_n\rangle\\
&\qquad + \langle w_{i-1}, e_n\rangle\sum_{t=1}^{i}\lambda^{i-t}\langle w_{t-1}, e_{n-1}\rangle\Bigg]
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}\langle y_{n-1}, y_n\rangle &= \lambda^{-2}\sum_{i=1}^{N-1}\sum_{t=1}^{i-1}\lambda^{4(i-t)} = \frac{1}{\lambda^{-2}(1-\lambda^4)}\sum_{i=1}^{N-1}\big(1-\lambda^{4i}\big) \qquad (4.27)\\
&= \lambda^{-2}c_{4,\lambda}\Big[(N-1) - c_{4,\lambda}\big(1-o^2_{\lambda;N}(1)\big)\Big]\\
\big(\mathbb{E}\langle y_{n-1}, y_n\rangle\big)^2 &= \lambda^{-4}c^2_{4,\lambda}\Big[(N-1) - c_{4,\lambda}\big(1-o^2_{\lambda;N}(1)\big)\Big]^2.
\end{aligned}$$

Also notice that for $i' > i$, $s_{i'} = \lambda^{(i'-i)}s_i + \sum_{t=i+1}^{i'}\lambda^{i'-t}w_{t-1}$, because:

$$s_{i'} = \sum_{t=1}^{i'}\lambda^{i'-t}w_{t-1} = \sum_{t=1}^{i}\lambda^{i'-t}w_{t-1} + \sum_{t=i+1}^{i'}\lambda^{i'-t}w_{t-1} = \lambda^{(i'-i)}\sum_{t=1}^{i}\lambda^{i-t}w_{t-1} + \sum_{t=i+1}^{i'}\lambda^{i'-t}w_{t-1} \qquad (4.28)$$

Now consider:

$$\mathbb{E}\Big(\sum_{i=1}^{N-1}\langle s_i, e_{n-1}\rangle\langle s_i, e_n\rangle\Big)^2 = \sum_{i=1}^{N-1}\mathbb{E}\big[\langle s_i, e_{n-1}\rangle^2\langle s_i, e_n\rangle^2 + 2\sum_{i'>i}\langle s_i, e_{n-1}\rangle\langle s_i, e_n\rangle\langle s_{i'}, e_{n-1}\rangle\langle s_{i'}, e_n\rangle\big],$$

where:

$$\sum_{i'>i}\big\langle\lambda^{(i'-i)}s_i + \sum_{t=i+1}^{i'}\lambda^{i'-t}w_{t-1}, e_{n-1}\big\rangle\big\langle\lambda^{(i'-i)}s_i + \sum_{t=i+1}^{i'}\lambda^{i'-t}w_{t-1}, e_n\big\rangle$$
$$= \sum_{i'>i}(\big\langle\lambda^{(i'-i)}s_i, e_{n-1}\big\rangle + \big\langle\sum_{t=i+1}^{i'}\lambda^{i'-t}w_{t-1}, e_{n-1}\big\rangle)(\big\langle\lambda^{(i'-i)}s_i, e_n\big\rangle$$
$$+\big\langle\sum_{t=i+1}^{i'}\lambda^{i'-t}w_{t-1}, e_n\big\rangle).$$

Therefore,

$$2\sum_{i=1}^{N-1}\mathbb{E}\langle s_i, e_{n-1}\rangle\langle s_i, e_n\rangle\sum_{i'>i}(\langle\lambda^{(i'-i)}s_i, e_{n-1}\rangle\langle\lambda^{(i'-i)}s_i, e_n\rangle + \langle\lambda^{(i'-i)}s_i, e_{n-1}\rangle$$
$$\big\langle\sum_{t=i+1}^{i'}\lambda^{i'-t}w_{t-1}, e_n\big\rangle + \big\langle\sum_{t=i+1}^{i'}\lambda^{i'-t}w_{t-1}, e_{n-1}\big\rangle\big\langle\lambda^{(i'-i)}s_i, e_n\big\rangle$$
$$+\big\langle\sum_{t=i+1}^{i'}\lambda^{i'-t}w_{t-1}, e_{n-1}\big\rangle\big\langle\sum_{t=i+1}^{i'}\lambda^{i'-t}w_{t-1}, e_n\big\rangle)$$

Simplifies to,

$$
\begin{aligned}
&2\sum_{i=1}^{N-1}\sum_{i'>i}^{N-1}\lambda^{2(i'-i)}\mathbb{E}\langle s_i, e_{n-1}\rangle^2\mathbb{E}\langle s_i, e_n\rangle^2 = 2\sum_{i=1}^{N-1}\sum_{i'>i}^{N-1}\lambda^{2(i'-i)}\bigg(\sum_{t=1}^{i}\lambda^{2(i-t)}\bigg)^2 \\
&= 2c_{2,\lambda}^2\sum_{i=1}^{N-1}\big(1-\lambda^{2i}\big)^2\sum_{i'>i}^{N-1}\lambda^{2(i'-i)} = 2c_{2,\lambda}^2\sum_{i=1}^{N-1}\big(1-\lambda^{2i}\big)^2\sum_{s=1}^{N-1-i}\lambda^{2s} \\
&= 2c_{2,\lambda}^3\sum_{i=1}^{N-1}\big(1-\lambda^{2i}\big)^2\big(1-\lambda^{2(N-i)}\big) \\
&= 2c_{2,\lambda}^3\bigg[(N-1)+c_{4,\lambda}\big(1-o_{\lambda;N}^2(1)\big)-2c_{2,\lambda}\big(1-o_{\lambda;N}(1)\big)\bigg] \\
&\quad -2c_{2,\lambda}^3\sum_{i=1}^{N-1}\big(1-\lambda^{2i}\big)^2\lambda^{2(N-i)},
\end{aligned}
$$

where we can further simplify

$$
\begin{aligned}
&2c_{2,\lambda}^3\sum_{i=1}^{N-1}\big(1-\lambda^{2i}\big)^2\lambda^{2(N-i)} = 2c_{2,\lambda}^3\sum_{i=1}^{N-1}\big(\lambda^{2(N-i)}+\lambda^{2(N+i)}-2\lambda^{2N}\big) \\
&= -4c_{2,\lambda}^3\lambda^{2N}\big(N-1\big)+2c_{2,\lambda}^3\big(1+\lambda^{2N}\big)\sum_{i=1}^{N-1}\lambda^{2i}.
\end{aligned}
$$

Furthermore, consider the following term:

$$
\sum_{i=1}^{N-1}\sum_{t=1}^{i-1}\lambda^{2(i-t)-1}\langle w_{t-1}, e_n\rangle^2 = \lambda^{-2}\sum_{t=1}^{N-2}\bigg(\sum_{l=1}^{N-t}\lambda^{2l}\bigg)\langle w_{t-1}, e_n\rangle^2,
$$

and

$$
\begin{aligned}
&\mathbb{E}\bigg[\lambda^{-2}\sum_{t=1}^{N-2}\bigg(\sum_{l=1}^{N-t}\lambda^{2l}\bigg)\langle w_{t-1}, e_n\rangle^2\bigg]^2 \\
&= 3\lambda^{-4}c_{2,\lambda}^4\bigg[(N-2)+c_{4,\lambda}\lambda^4\big(\lambda^4-o_{\lambda;N}^2(1)\big)\bigg]+2\lambda^{-4}\sum_{t=1}^{N-2}\sum_{t'>t}^{N-2}\sum_{l=1}^{N-t}\sum_{l'=1}^{N-t'}\lambda^{2(l'+l)},
\end{aligned}
$$

where

$$2\lambda^{-4}\sum_{t=1}^{N-2}\sum_{t'>t}^{N-2}\sum_{l=1}^{N-t}\sum_{l'=1}^{N-t'}\lambda^{2(l'+l)} = 2c_{2,\lambda}\lambda^{-4}\sum_{t=1}^{N-2}\sum_{t'>t}^{N-2}\sum_{l=1}^{N-t}\lambda^{2l}\big(1+\lambda^{2(N-t'+1)}\big)$$

$$2c_{2,\lambda}\lambda^{-4}\sum_{t=1}^{N-2}\sum_{t'>t}^{N-2}\Big(\sum_{l=1}^{N-t}\lambda^{2l}+\lambda^{2(N-t'+1)}\sum_{l=1}^{N-t}\lambda^{2l}\Big) = 2c_{2,\lambda}^2\lambda^{-4}\sum_{t=1}^{N-2}\sum_{t'>t}^{N-2}\Big(\big[1+\lambda^{2(N-t+1)}\big]$$

$$+\lambda^{2(N-t'+1)}\big[1+\lambda^{2(N-t+1)}\big]\Big) = c_{2,\lambda}^2\lambda^{-4}\Bigg[\bigg((N-2)+\sum_{t=1}^{N-2}\lambda^{2(N-t+1)}\bigg)^2$$

$$-\sum_{t=1}^{N-2}\big(1+\lambda^{2(N-t+1)}\big)^2\Bigg]$$

$$= c_{2,\lambda}^2\lambda^{-4}\bigg(N-2+c_{2,\lambda}\lambda^6\big(1+\lambda^{2(N-2)}\big)\bigg)^2$$

$$-c_{2,\lambda}^2\lambda^{-4}\Big[(N-2)+c_{4,\lambda}\lambda^{4(3)}\big(1+\lambda^{4(N-2)}\big)+2c_{2,\lambda}\lambda^{2(3)}\big(1+\lambda^{2(N-2)}\big)\Big]$$

$$= c_{2,\lambda}^2\lambda^{-4}\Bigg[(N-2)^2+c_{2,\lambda}^2\lambda^{2(6)}\big(1+\lambda^{2(N-2)}\big)^2-(N-2)-c_{4,\lambda}\lambda^{4(3)}\big(1+\lambda^{4(N-2)}\big)$$

$$-2c_{2,\lambda}\lambda^{2(3)}\big(1+\lambda^{2(N-2)}\big)\Bigg]$$

Now we are only left with the task of bounding

$$\mathbb{E}\bigg(\sum_{i=1}^{N-1}\sum_{t=1}^{i}\lambda^{i-t}\langle w_{i-1},e_n\rangle\langle w_{t-1},e_{n-1}\rangle\bigg)^2 = \sum_{i=1}^{N-1}\mathbb{E}\Bigg[\Big(\sum_{t=1}^{i}\lambda^{i-t}\langle w_{i-1},e_n\rangle\langle w_{t-1},e_{n-1}\rangle\Big)^2$$

$$+2\sum_{i'>i}\big(\langle s_i,e_{n-1}\rangle\langle w_{i-1},e_n\rangle\big)\Big(\langle\lambda^{(i'-i)}s_i+\sum_{t=i+1}^{i'}\lambda^{i'-t}w_{t-1},e_{n-1}\rangle\langle w_{i'-1},e_n\rangle\Big)\Bigg]$$

$$= \sum_{i=1}^{N-1}\mathbb{E}\big(\langle\sum_{t=1}^{i}\lambda^{i-t}w_{t-1},e_{n-1}\rangle\langle w_{i-1},e_n\rangle\big)^2 = \sum_{i=1}^{N-1}\mathbb{E}\big(\langle s_i,e_{n-1}\rangle\langle w_{i-1},e_n\rangle\big)^2$$

$$= \sum_{i=1}^{N-1}\mathbb{E}(\langle s_i,e_{n-1}\rangle^2) = \sum_{i=1}^{N-1}\sum_{t=1}^{i}\lambda^{2(i-t)} = c_{2,\lambda}(N-1)-c_{2,\lambda}^2\big(1-o_{\lambda;N}(1)\big)$$

Theorem 20. *Standard deviation of* $\langle y_{n-1},y_n\rangle$, $SD\big(\langle y_{n-1},y_n\rangle\big) = O(N-2)$ *and*

there exists positive constants independent of N, $\kappa_{\lambda,1}$ and $\kappa_{\lambda,2}$ such that

$$\langle y_{n-1}, y_n \rangle \in \left[\kappa_{\lambda,1}, 2N - \kappa_{\lambda,2}\right] \tag{4.29}$$

Proof. Proof follows the general computational procedure, we know from (4.27) analytic expression of $\mathbb{E}\langle y_n, y_{n-1}\rangle$

$$\begin{aligned}
\mathbb{E}\big|\langle y_{n-1}, y_n\rangle - \mathbb{E}\langle y_{n-1}, y_n\rangle\big|^2 &= 2c_{2,\lambda}^3\Big[(N-1) + c_{4,\lambda}\big(1 - o_{\lambda;N}^2(1)\big) - 2c_{2,\lambda}\big(1 - o_{\lambda;N}(1)\big)\Big] \\
&+ c_{2,\lambda}(N-1) - c_{2,\lambda}^2\big(1 - o_{\lambda;N}(1)\big) - \lambda^{-4}c_{4,\lambda}^2\Big[(N-1) - c_{4,\lambda}\big(1 - o_{\lambda;N}^2(1)\big)\Big]^2 \\
&+ 4c_{2,\lambda}^3\lambda^{2N}\big(N-1\big) - 2c_{2,\lambda}^4\big(1 + \lambda^{2N}\big)^2 + 3\lambda^{-4}c_{2,\lambda}^4\Big[(N-2) + c_{4,\lambda}\lambda^4\big(\lambda^4 - o_{\lambda;N}^2(1)\big)\Big] \\
&+ c_{2,\lambda}^2\lambda^{-4}\Bigg[(N-2)^2 + c_{2,\lambda}^2\lambda^{2(6)}\big(1 + \lambda^{2(N-2)}\big)^2 - (N-2) \\
&\qquad - c_{4,\lambda}\lambda^{4(3)}\big(1 + \lambda^{4(N-2)}\big) - 2c_{2,\lambda}\lambda^{2(3)}\big(1 + \lambda^{2(N-2)}\big)\Bigg].
\end{aligned}$$

From (4.27) we know that $\mathbb{E}\langle y_n, y_{n-1}\rangle = O(N-1)$ and result follows. □

Remark 21. *This is where things start getting tricky, compared to at most typical size of $O(\sqrt{N})$ in Hermitian case for off-diagonal terms from remark 18,rows with typical size of $N \pm \sqrt{N}$ and polynomial in N can have a typical size of correlation $\big(N-1\big) \pm O\big(N-2\big)$. As the higher rows will have higher typical size and off-diagonal terms can potentially be $O(N^{n-1})$, which we will study in future work.*

Typical size of the rows As we will shortly see afterwards, in order to get typical order of the bulk eigenvalues, one will need typical order of all the rows, which brings us to:

Proposition 7. *Sequnce of random variables* $\left\langle y_j, \hat{e}_{n-(j-1)} \right\rangle_{j\in[n]}$ *are centered at* 0 *and:*

$$O\Bigg(\sum_{k=j}^{n-(j-1)} \sum_{m=0}^{n-k} \binom{n-k}{m}^2 \lambda^{2(n-k-m)} \Bigg)^{\frac{1}{2}} \tag{4.30}$$

$$\langle y_j, \hat{e}_i \rangle = \sum_{t=1}^{i} \sum_{m=0}^{(i-t)\wedge(n-j)} \binom{i-t}{m} \lambda^{i-t-m} \big\langle w_{t-1}, e_{m+j} \big\rangle, \tag{4.31}$$

$$\langle y_2, \hat{e}_{n-1} \rangle = \sum_{t=1}^{n-1} \sum_{m=0}^{n-1-t} \binom{n-1-t}{m} \lambda^{n-1-t-m} \langle w_{t-1}, e_{m+2} \rangle, \tag{4.32}$$

is centered at 0, weighted sum of independent $\frac{n(n-1)}{2}$ standard normals with variance:

$$4^n \lambda^{2n} L_{\lambda,n}(2) \leqslant \mathbb{E}\big(\langle y_2, \hat{e}_{n-1} \rangle^2 \big) = \sum_{k=2}^{n} \sum_{m=0}^{n-k} \binom{n-k}{m}^2 \lambda^{2(n-k-m)} \leqslant 4^n \lambda^{2n} U_{\lambda,n}(2), \tag{4.33}$$

and repeating the recursion we have for $k \in [0, \ldots, n-1]$,

$$4^n \lambda^{2n} L_{\lambda,n}(k+1) \leqslant \mathbb{E}(\langle y_{k+1}, \hat{e}_{n-k} \rangle^2) \leqslant 4^n \lambda^{2n} U_{\lambda,n}(k+1) \tag{4.34}$$

4.2.2 Interlacing the eigenvalues of Sample Covariance matrix

We will use Courant-Fischer for the distribution of eigenvalues corresponding to the sample covariance matrix, $\Sigma_n := X_- X_-^*$

$$\lambda_1(\Sigma_n) \geqslant \lambda_2(\Sigma_n) \geqslant \ldots \lambda_n(\Sigma_n) > 0. \tag{4.35}$$

$$\Sigma_n := \begin{bmatrix} \langle y_1, y_1 \rangle & \dots & \langle y_1, y_{k+1} \rangle & \dots & \langle y_1, y_n \rangle \\ \vdots & \dots & \vdots & \dots & \vdots \\ \langle y_{k+1}, y_1 \rangle & \dots & \langle y_{k+1}, y_{k+1} \rangle & \dots & \langle y_{k+1}, y_n \rangle \\ \vdots & \dots & \vdots & \dots & \vdots \\ \langle y_n, y_1 \rangle & \dots & \langle y_n, y_{k+1} \rangle & \dots & \langle y_n, y_n \rangle \end{bmatrix}$$

Σ_{n-k} be the modified sample covariance matrix with first $k-$ columns and rows removed

$$\Sigma_{n-k} := \begin{bmatrix} \langle y_{k+1}, y_{k+1} \rangle & \dots & \langle y_{k+1}, y_n \rangle \\ \vdots & \dots & \vdots \\ \langle y_n, y_{k+1} \rangle & \dots & \langle y_n, y_n \rangle \end{bmatrix}$$

Now we have a task of comparing the eigenvalues of both matrices, letting V_i denote $i-$ th dimensional subspace:

$$\lambda_i(\Sigma_n) := \sup_{V_i \subset \mathbb{R}^n} \inf_{y \in S^{n-1} \cap V_i} \left\| \Sigma_n y \right\|, \ \textit{and} \ \lambda_i\big(\Sigma_{n-k}\big) := \sup_{V_i \subset \mathbb{R}^{n-k}} \inf_{z \in S^{n-k-1} \cap V_i} \left\| \Sigma_{n-k} z \right\| \tag{4.36}$$

Not hard to realize that for $i = [1, \dots, n-k]$

$$\lambda_i(\Sigma_n) \geqslant \lambda_i(\Sigma_{n-k}) \geqslant \lambda_{k+i}(\Sigma_n). \tag{4.37}$$

Just to be hands on, fix $n = 3, k = 1$ and vary i from 1 to 2 and notice:

$$\begin{aligned} &\lambda_1(\Sigma_3) \geqslant \lambda_1(\Sigma_2) \geqslant \lambda_2(\Sigma_3) \geqslant \lambda_2(\Sigma_2) \geqslant \lambda_3(\Sigma_3), \ \textit{Now for: } n = 4, k = 1, i \in [3] \\ &\lambda_1(\Sigma_4) \geqslant \lambda_1(\Sigma_3) \geqslant \lambda_2(\Sigma_4) \geqslant \lambda_2(\Sigma_3) \geqslant \lambda_3(\Sigma_4) \geqslant \lambda_3(\Sigma_3) \geqslant \lambda_4(\Sigma_4), \end{aligned} \tag{4.38}$$

and one trivially notices that if dynamics follow S-w-SSCs and one increase the dimension,

4.3 Smallest eigenvalue

will always be upper bounded by $\lambda_1(\Sigma_1) = \langle y_n, y_n \rangle = \|y_n\|^2$, where recall from (4.16) that:

$$\langle y_n, \hat{e}_i \rangle = \sum_{t=1}^{i} \lambda^{i-t} \langle w_{t-1}, e_n \rangle \tag{4.39}$$

i.e., typical size of standard scalar ARMA trajectory of length N. Furthermore:

Theorem 22 (Upper bounding the smallest singular value for all stable dynamics)**.** *Least singular value of the data matrix is essential upper bounded by the typical size of the scalar stable, length N ARMA trajectory. Ignoring constant that can potentially only depend on eigenvalue of A,*

$$\sigma_n^2(X_-) \in (0, N + \sqrt{N}], \;\; w.p. \;\; 1. \tag{4.40}$$

Proof. From interlacing theorem we have that:

$$\mathbb{P}\big(\sigma_n^2(X_-) \geqslant \delta\big) = \mathbb{P}\big(\lambda_n(\Sigma_n) \geqslant \delta\big) \leqslant \mathbb{P}\big(\lambda_1(\Sigma_1)\big) = \mathbb{P}\big(\|y_n\|^2 \geqslant \delta\big)$$
$$\frac{\sigma_n^2(X_-)}{N} \longrightarrow (0, 1].$$

and as discussed in Remark 18, $\langle y_n, y_n \rangle \in [N - \sqrt{N}, N + \sqrt{N}]$ w.p 1 □

Fix a row of the data matrix X_-, in order to get lower bound and better estimates on the least singular value, one needs to quantify its' distance from the span of remaining rows of data matrix. As we can take conjugate transpose and write least

singular value of the data matrix as:

$$\sigma_n(X_-) := \inf_{a \in S^{n-1}} \|X_-^* a\|_2 \tag{4.41}$$

Leveraging upon the Hilbertian structure(orthogonal projections/inner products), given any subspace $V \subset \mathbb{R}^n$, we have that:

$$\|X_-^* a\|_2^2 = \|P_V(X_-^* a)\|_2^2 + \|P_{V^\perp}(X_-^* a)\|_2^2 \geqslant \|P_{V^\perp}(X_-^* a)\|_2^2. \tag{4.42}$$

Therefore,

$$\left\|X_-^* a\right\|_2 \geqslant |a_i| d(y_i, n_i), \ \textit{for all}\, i \in [n]. \tag{4.43}$$

Now if $a \in S^{n-1}$, then there exists an $i \in [n]$ such that $|a_i| \geqslant \frac{1}{\sqrt{n}}$. Subsequently,

$$P\left(\inf_{a \in S^{n-1}} \|X_-^* a\| \leqslant \delta \right) = P\left(\exists\, a \in S^{n-1} :\, \|X_-^* a\| \leqslant \delta \right) \leqslant P\left(\exists\, j \in [n] :\, \frac{d(y_j, n_j)}{\sqrt{n}} \leqslant \delta \right) \tag{4.44}$$

1. when we have independent rows union bound is not terribly wasteful, so:

$$\mathbb{P}\big(\sigma_n(X_-) \leqslant \delta\big) \leqslant \sum_{j=1}^{n} \mathbb{P}\left(d(y_j, n_j) \leqslant \delta\sqrt{n} \right) \tag{4.45}$$

   Let $w_j \in S^{N-1}$ be orthogonal unit vector to subspace n_j(not necessarily unique), if the rows are independent then w_j can be chosen independently of y_j. Therefore:

$$\langle w_j, y_j \rangle = \langle w_j, P_{n_j^\perp} y_j \rangle \leqslant \left\| P_{n_j^\perp} y_j \right\| = d(y_j, n_j), \tag{4.46}$$

where inequality follows from Cauchy-Schwarz. Consequently:

$$\mathbb{P}\bigg(d(y_j, n_j) \leqslant \delta\sqrt{n}\bigg) \leqslant \mathbb{P}\bigg(\langle w_j, y_j\rangle \leqslant \delta\sqrt{n}\bigg), \tag{4.47}$$

where controlling the last term is commonly studied under the name of *anti-concentration or small ball probability.*

2. rows with strong correlation(S-w-SSC-s), then union bound will essentially conclude vacuous estimates so its' better to bound

$$\mathbb{P}\big(\sigma_n(X_-) \leqslant \delta\big) \leqslant \mathbb{P}\bigg(\min_{j\in[n]} d(y_j, n_j) \leqslant \delta\sqrt{n}\bigg) \tag{4.48}$$

It turns out that distance estimates between a length N random vector(that represents the scalar ARMA trajectory) and $N - n + 1$ dimensional subspace as in subsection 4.3.3, play a crucial role in dictating the least singular value of the data matrix and consequently provide an upper bound on the estimation error related to OLS for identification of LTI systems as in (3.12). As we can take conjugate transpose and write least singular value of the data matrix as:

$$\sigma_n(X_-) := \inf_{a\in S^{n-1}} \|X_-^* a\|_2 \tag{4.49}$$

Leveraging upon the Hilbertian structure(orthogonal projections/inner products), given any subspace $V \subset \mathbb{R}^n$, we have that:

$$\|X_-^* a\|_2^2 = \|P_V(X_-^* a)\|_2^2 + \|P_{V^\perp}(X_-^* a)\|_2^2 \geqslant \|P_{V^\perp}(X_-^* a)\|_2^2. \tag{4.50}$$

Therefore,

$$\left\|X_-^* a\right\|_2 \geqslant |a_i| d(y_i, n_i), \ \textit{for all}\, i \in [n]. \tag{4.51}$$

Now if $a \in S^{n-1}$, then there exists an $i \in [n]$ such that $|a_i| \geqslant \frac{1}{\sqrt{n}}$. Subsequently,

$$\begin{aligned} P\bigg(\inf_{a \in S^{n-1}} \|X_-^* a\| \leqslant \delta\bigg) &= P\bigg(\exists\, a \in S^{n-1} :\, \|X_-^* a\| \leqslant \delta\bigg) \\ &\leqslant P\bigg(\exists\, j \in [n] :\, \frac{d(y_j, n_j)}{\sqrt{n}} \leqslant \delta\bigg) \end{aligned}$$

- when we have independent rows union bound is not terribly wasteful, so:

$$\mathbb{P}\big(\sigma_n(X_-) \leqslant \delta\big) \leqslant \sum_{j=1}^{n} \mathbb{P}\bigg(d(y_j, n_j) \leqslant \delta\sqrt{n}\bigg) \tag{4.52}$$

  Let $w_j \in S^{N-1}$ be orthogonal unit vector to subspace n_j(not necessarily unique), if the rows are independent then w_j can be chosen independently of y_j. Therefore:

$$\langle w_j, y_j\rangle = \langle w_j, P_{n_j^\perp} y_j\rangle \leqslant \left\|P_{n_j^\perp} y_j\right\| = d(y_j, n_j), \tag{4.53}$$

  where inequality follows from Cauchy-Schwarz. Consequently:

$$\mathbb{P}\bigg(d(y_j, n_j) \leqslant \delta\sqrt{n}\bigg) \leqslant \mathbb{P}\bigg(\langle w_j, y_j\rangle \leqslant \delta\sqrt{n}\bigg), \tag{4.54}$$

  where controlling the last term is commonly studied under the name of *anti-concentration or small ball probability.*

- rows with strong correlation(S-w-SSC-s), then union bound will essentially con-

clude vacuous estimates so its' better to bound

$$\mathbb{P}\big(\sigma_n(X_-) \leqslant \delta\big) \leqslant \mathbb{P}\bigg(\min_{j\in[n]} d(y_j, n_j) \leqslant \delta\sqrt{n}\bigg) \tag{4.55}$$

4.3.1 Decomposition of sphere based on compressibility

Ideas and notation in this subsection heavily borrows from the work of [26], [28] and [29]. For a better estimate of least singular value, one has to take infimum separatley over compressibe and incompressible elements of $n-$ dimensional unit sphere.

Definition 7. *We define the support of a vector $x \in \mathbb{R}^n$ as*

$$supp(x) := [j \in [n] : \langle x, e_j\rangle \neq 0]. \tag{4.56}$$

For $J \subset [n]$, we define S^J as set of unit vectors supported on J. Given $\epsilon > 0$ and $V \subset \mathbb{R}^n$, set of vectors with in ϵ euclidean distance from V are denoted by V_ϵ. Compressible vectors with parameters (δ, ϵ) are unit vectors which are within ϵ euclidean distance from some vector supported on at most δn coordinates and mathematically defined as:

$$Comp(\delta, \epsilon) := S^{n-1} \bigcap \bigcup_{J\in\binom{\delta n}{n}} \big(S^J\big)_\epsilon. \tag{4.57}$$

and complementary set of incompressible vectors:

$$Incomp(\delta, \epsilon) := S^{n-1} \backslash Comp(\delta, \epsilon) \tag{4.58}$$

Lemma 5. *For a square random matrix M with i.i.d enteries*

$$\mathbb{P}\bigg(\inf_{u\in Incomp(\delta,\epsilon)} \big\|Mu\big\| \leqslant \frac{t}{\sqrt{n}}\bigg) \leqslant \frac{1}{\delta n}\sum_{i=1}^{n} \mathbb{P}\bigg(d(y_i, n_i)\bigg) \tag{4.59}$$

Lemma 6. *Fix $\delta, \epsilon \in (0,1)$ and let $v \in Incomp(\delta, \epsilon)$, there is a set $P^+ \subset [n]$ with $|P^+| \geqslant \delta n$ and $|v_j| \geqslant \frac{\epsilon}{\sqrt{n}}$ for all $j \in P^+$. Moreover, for all $\lambda \geqslant 1$ there is a set $P \subset [n]$ with $|P| \geqslant (1 - \frac{1}{\lambda^2})\delta n$ such that for all $j \in P$*

$$\frac{\epsilon}{\sqrt{n}} \leqslant |v_j| \leqslant \frac{\lambda}{\sqrt{\delta n}} \tag{4.60}$$

Proof. Take $P^+ := \{j : |v_j| \geqslant \frac{\epsilon}{\sqrt{n}}\}$, as v is at least distance ϵ from any unit vector with support at most δn, so $1 = \sum_{j \in ([n] \setminus P^+)} |v_j|^2 + \sum_{j \in P^+} |v_j|^2 \geqslant \sum_{j \in ([n] \setminus P^+)} |v_j|^2 + \frac{\epsilon^2}{n}|P^+|$ □

Let $0 < \nu_1 := (1 - \frac{1}{\lambda^2})\delta$, $\nu_2 := \epsilon < 1$ and $\nu_3 := \frac{\lambda}{\sqrt{\delta}} > 1$. It is imperative to notice that regardless of $x \in Incomp(\delta, \epsilon)$, (ν_1, ν_2, ν_3) are only dependent on (δ, ϵ).

Controlling the infimum over the incompressible vectors. Let $p := \mathbb{P}\big(d(y_k, n_k) < \epsilon_1\big)$. Then notice that under the assumption of independence between the rows,

$$\mathbb{E}\big|\{k : d(y_k, n_k) < \epsilon\}\big| = \sum_{k=1}^{n} k\mathbb{P}\big(B = k\big) = \sum_{k=1}^{n} k\binom{n}{k}p^k(1-p)^{n-k} = np. \tag{4.61}$$

If we denote by U the event that the cardinality of the set, $\sigma_1 := \{k : d(y_k, n_k) \geqslant \epsilon_1\}$ contains more that $(1-\nu_1)n$ elements. Then by Chebyshevs' inequality we have that:

$$\mathbb{P}(U^c) \leqslant \frac{p}{\nu_1} \tag{4.62}$$

and eventually:

$$\mathbb{P}\left(\inf_{a \in Incomp(\delta,\epsilon)} \left\|X_-^* a\right\| < \frac{\epsilon_1 \nu_2}{\sqrt{n}}\right) \leqslant \mathbb{P}(U^c) \leqslant \frac{\mathbb{P}\big(d(y_k, n_k) < \epsilon_1\big)}{\nu_1} \tag{4.63}$$

It is not too difficult to realize now, when dynamics are generated from Hermitian linear transformation, we can use concentration of measure to understand the behavior

of least singular value.

4.3.2 Distance of a random vector to fixed subspace

These distance estimates have been an integral part of recent breakthrough progress in Random Matrix Theory. However, the problem under dynamical systems setting is much more complicated as: observations in every row are correlated and different rows are themselves correlated. We will first focus on dealing the former problem. A good starting point is quantifying distance between a random row of length N and an $n-1$ dimensional subspace of $\mathbb{R}^N$.

Theorem 23. *Let $\mathcal{V}$ be a fixed $n-1$ dimensional subspace of R^N, and random vector $x \sim N(0, I_N)$ where we assme that $N > n$, with probability 1 distance function concentrates around $[(N-n+1) - \sqrt{2(N-n+1)}, (N-n+1) + \sqrt{2(N-n+1)}]$.*

Proof. Let P be the projection matrix associated with the subspace $\mathcal{V}$. By the definintion of projection matrix we have that $P \in \mathbb{R}^{N\times N}$–if this is confusing, think of P as SVD where it acts as an identity on $n-1$ basis vectors in $\mathbb{R}^N$ that define $\mathcal{V}$-, $Tr(P) = \sum_{i=1}^{N} p_{ii} = n-1$ and $P = P^* = P^2$. Let $A = P-$ Diag$(p_{11}, \ldots, p_{NN})$, so diagonal elements of A: $(a_{ii})_{i=1}^{N}$ are zeros and non-diagonal are same as projection matrix. Pythagoreous theorem reveals:

$$d^2(x, \mathcal{V}) = \|x\|^2 - \langle x, Px\rangle = \sum_{i=1}^{N}(1-p_{ii})x_i^2 - \sum_{j,k=1}^{N} a_{jk}x_k x_j$$

Taking expectation, while keeping in mind $\mathbb{E}[x_k x_j] = 0$ for $k \neq j$ and trace equality, reveals:

$$\mathbb{E}[d^2(x, \mathcal{V})] = N - n + 1 \tag{4.64}$$

However, in order to use Chebyshev inequality that will lead to concentration

result for $d^2(x,V)$, we need to compute:

$$Var[d^2(x,V)] = \mathbb{E}d^4(x,V) - (N-n+1)^2 = \mathbb{E}\Big[\big(\sum_{i=1}^{N}(1-p_{ii})x_i^2 - \sum_{j,k=1}^{N} a_{jk}x_kx_j\big)^2\Big]$$
$$- (N-n+1)^2 = \sum_{i=1}^{N}\Big[3(1-p_{ii})^2 + 2(1-p_{ii})\sum_{j>i}(1-p_{jj})\Big] - (N-n+1)^2 + \mathbb{E}[Y^2]. \tag{4.65}$$

Where (4.65) follows from *Stein's Lemma*; if $w \sim \mathcal{N}(0,\sigma^2)$, then for any odd power k, $\mathbb{E}w^k = 0$ and for all $k \in \mathbb{N}$ we have higher moments $\mathbb{E}w^{2k} = \sigma^{2k}\prod_{l=1}^{k}(2l-1)$. Let $Y := \sum_{j,k=1}^{N} a_{jk}x_kx_j$ and notice that $\mathbb{E}(Y^2) = 2Tr(A^2)$, where:

$$Tr(A^2) = Tr(P^2) - \sum_{i=1}^{N} p_{ii}^2 = \sum_{i=1}^{N}\sum_{j=1}^{N} p_{ij}^2 - \sum_{i=1}^{N} p_{ii}^2 = Tr(P) - \sum_{i=1}^{N} p_{ii}^2$$

The remaining terms in (4.65),

$$\sum_{i=1}^{N}[3(1-p_{ii})^2 + 2(1-p_{ii})\sum_{j>i}(1-p_{jj})] = \sum_{i=1}^{N} 2(1-p_{ii})^2 + [Tr(I-P)]^2$$
$$= \sum_{i=1}^{N} 2(1-p_{ii})^2 + (N-n+1)^2.$$

$$\textit{But, } \sum_{i=1}^{N} 2(1-p_{ii})^2 + 2Tr(A^2) = 2[N-n+1]. \tag{4.66}$$
$$\textit{Therefore, } \quad Var[d^2(x,\mathcal{V})] = 2(N-n+1)$$

Consequently, Tchebyshev implies with probability 1, $d^2(x,\mathcal{V}) \in [(N-n+1) - \sqrt{2(N-n+1)}, (N-n+1) + \sqrt{2(N-n+1)}]$ □

$$\Sigma_{N,\lambda} := \begin{pmatrix} 1 & \lambda & \lambda^2 & \dots & \dots \\ \lambda & 1+\lambda^2 & \lambda^3+\lambda & \dots & \dots \\ \lambda^2 & \lambda^3+\lambda & 1+\lambda^2+\lambda^4 & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \lambda^{N-1} & \lambda^N+\lambda^{N-1} & \dots & \dots & \lambda^{2(N-1)}+\lambda^{2(N-2)}+\ldots+1 \end{pmatrix}$$

4.3.3 Covariance estimate via moment method

Now instead of each element of the row being independent, it now corresponds to one dimensional ARMA trajectory of length N as in (1.14) and its' distance from a fixed subspace V is:

$$d^2(x,\mathcal{V}) = \sum_{i=1}^{N}(1-p_{ii})x_i^2 - 2\sum_{j=1}^{N}\sum_{k>j}^{N} p_{jk}x_k x_j.$$

Exploiting the Markovian structure of the dynamics for $k > j$:

$$x_k = \lambda^{[k-j]}x_j + \lambda^{[k-j]-1}w_j + \lambda^{[k-j]-2}w_{j+1} + \ldots + w_{k-1}$$
$$x_k x_j = \lambda^{[k-j]}x_j^2 + \big(\lambda^{[k-j]-1}w_j + \lambda^{[k-j]-2}w_{j+1} + \ldots + w_{k-1}\big)x_j.$$

Trace and elements of the covariance matrix $\Sigma_{N,\lambda}[k,j] := \mathbb{E}[x_k x_j]$ are

$$Tr(\Sigma_{N,\lambda}) = \sum_{j=1}^{N}\Sigma_{N,\lambda}[j,j] = \mathbb{E}\big[\sum_{i=1}^{N}x_i^2\big] = \sum_{i=1}^{N} i\lambda^{2(N-i)} \tag{4.67}$$

$$\Sigma_{N,\lambda}[k,j] = \lambda^{[k-j]}\mathbb{E}[x_j^2] = \lambda^{[k-j]}(1+\lambda^2+\ldots+\lambda^{2[k-j]-1}), \tag{4.68}$$

In the presence of temporal correlation between elements of the row vectors we need a tight control on the spectrum of the covariance matrix, shown in 4.3.3.

Theorem 24. *There exists positive constants $c_{\lambda,1}, c_{\lambda,2}$ and $c_{\lambda,3}$ such that for all $N \in \mathbb{N}$*

$$c_{\lambda,1}N - c_{\lambda,2} \leqslant Tr(\Sigma_{N,\lambda}) \leqslant c_{\lambda,1}N - c_{\lambda,2} + c_{\lambda,1}\left(1 - \frac{c_{\lambda,3}}{N}\right). \tag{4.69}$$

Where $c_{\lambda,1} = \frac{|\lambda|^{-2}}{\ln|\lambda|^{-2}}$,$c_{\lambda,2} = \frac{|\lambda|^{-2}}{[\ln|\lambda|^{-2}]^2}$ and $c_{\lambda,3} = \frac{1}{\ln|\lambda|^{-2}}$. Consequently,

$$\Omega\big(1\big) \leqslant \big\|\Sigma_{N,\lambda}^{\frac{1}{2}}\big\|_2 \leqslant O(N^{\frac{1}{2}}).$$

Moreover, there exists a constant $d_{\lambda,1}$ and positive constant $d_{\lambda,2}$ such that:

$$\big\|\Sigma_{N,\lambda}\big\|_F^2 = Tr(\Sigma_{N,\lambda}^2) = d_{\lambda,2}N + o_{\lambda;N}(1) + d_{\lambda,1}, \tag{4.70}$$

where given $|\lambda| < 1$, we use notation $o_{\lambda;N}(1)$ to represent a term dependent on λ but going to zero as $N \to \infty$. Furthermore, (4.70) *implies even better control on $\|\Sigma_{N,\lambda}^{\frac{1}{2}}\|_2$:*

$$\Omega(1) \leqslant \|\Sigma_{N,\lambda}^{\frac{1}{2}}\|_2 \leqslant O(N^{\frac{1}{4}}). \tag{4.71}$$

Proof. Recall from equation (4.67), $Tr(\Sigma_{N,\lambda}) = \mathbb{E}\big[\sum_{i=1}^N x_i^2\big] = \sum_{i=1}^N i\lambda^{2(N-i)}$. Now we can control the value of trace by lower bounding it by area under an appropriate curve, similarly upper bounding it by area of some curve. Let $a := \ln|\lambda|^{-1}$, then:

$$\int_0^N xe^{2ax}dx \leqslant \sum_{i=1}^N i\lambda^{-2i} \leqslant \int_0^N (x+1)e^{2a(x+1)}dx \tag{4.72}$$

Frobenius norm estimates follows by noticing:

$$\big\|\Sigma_{N,\lambda}\big\|_F^2 = \sum_{j=1}^{N}|\mathbb{E}[x_j^2]|^2 + 2\sum_{k>j}^{N}|\mathbb{E}[x_j x_k]|^2 = 2\sum_{j=1}^{N}\big[1+\sum_{k>j}^{N}|\lambda|^{2[k-j]}\big]|\mathbb{E}[x_j^2]|^2 - \sum_{j=1}^{N}|\mathbb{E}[x_j^2]|^2$$

$$= \frac{4N}{[1-|\lambda|^2]^3}|\lambda|^{2(N+1)} - 2|\lambda|^{2(N+1)}\sum_{j=1}^{N}\big[|\lambda|^{2j}+|\lambda|^{-2j}\big] + \frac{1}{[1-|\lambda|^2]^2}\big[\frac{2}{[1-|\lambda|^2]}-1\big]$$

$$\sum_{j=1}^{N}\big[1-|\lambda|^{2j}\big]^2 = \frac{4N}{[1-|\lambda|^2]^3}|\lambda|^{2(N+1)} - 2\frac{\big[1-|\lambda|^{2N}\big]|\lambda|^2}{1-|\lambda|^2}\big[1+|\lambda|^{2(N+1)}\big]$$

$$+\frac{1}{[1-|\lambda|^2]^2}\big[\frac{2}{[1-|\lambda|^2]}-1\big]\big[N+\frac{|\lambda|^4\big(1-|\lambda|^{4N}\big)}{1-|\lambda|^4}+\frac{|\lambda|^2\big(1-|\lambda|^{2N}\big)}{1-|\lambda|^2}\big]$$

□

Remark 25 (The moment method:)**.** *notice that how computing trace of higher powers of covariance matrix leads to better estimates of operator norm of covariance matrix. Let* $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \lambda_N > 0$ *be the eigenvalues of* $\Sigma_{N,\lambda}$*, as trace of a matrix equals sum of its eigenvalues and operator norm of positive definite matrix equals largest positive eigenvalue , we have for any* $k \in \mathbb{N}$*:*

$$\big\|\Sigma_{N,\lambda}\big\|_2^k = \lambda_1^k \leqslant \sum_{i=1}^{N}\lambda_i^k = Tr(\Sigma_{N,\lambda}^k) \leqslant N\lambda_1^k = N\big\|\Sigma_{N,\lambda}\big\|_2^k$$

$$\textit{implying,} \quad \frac{Tr(\Sigma_{N,\lambda}^k)^{\frac{1}{k}}}{N^{\frac{1}{k}}} \leqslant \big\|\Sigma_{N,\lambda}\big\|_2 \leqslant Tr(\Sigma_{N,\lambda}^k)^{\frac{1}{k}}. \tag{4.73}$$

That is by computing trace of higher powers of $\Sigma_{N,\lambda}$*, one can get a tight control on estimate of* $\big\|\Sigma_{N,\lambda}\big\|_2$ *which will reveal order of the deviation of underlying* quadratic form *of interest in terms of state space dimensions and number of iterations. It is important to point out that we believe that* $\Omega(1) \leqslant \|\Sigma_{N,\lambda}^{\frac{1}{2}}\|_2 \leqslant O(N^{\frac{1}{4}})$ *in* (4.71) *can be improved to* $\|\Sigma_{N,\lambda}^{\frac{1}{2}}\|_2 = \Theta(1)$ *as suggested by dimension independent tensorization of Talagrands' inequality in Theorem 2, by computing higher powers of trace.*

Theorem 26. *Let $x = (x_1, x_2, \ldots, x_N)$ be the trajectory of length N from one dimensional ARMA model (1.14), with stable eigenvalue λ and V be an $n-1$ dimensional subspace of $\mathbb{R}^N$, then:*

$$\begin{aligned}\mathbb{E}d^2(x,V) = Tr(\Sigma_{N,\lambda}P_{V^\perp}) &\leqslant \|\Sigma_{N,\lambda}\|_F\|P_{V^\perp}\|_F \\ &= O\bigg(\sqrt{\big[d_{\lambda,2}N + o_{\lambda;N}(1) + d_{\lambda,1}\big](N-n+1)}\bigg).\end{aligned} \tag{4.74}$$

Furthermore, there exists positive constants c_λ and c such that for all $\delta > 0$:

$$P\bigg(|d^2(x,V) - Tr(\Sigma_{N,\lambda}P_{V^\perp})| \geqslant \delta c_\lambda N^{\frac{1}{2}}(N-n+1)^{\frac{1}{4}}\bigg) \leqslant \exp\left(-c\delta^2\right), \tag{4.75}$$

implying fluctuation of d^2 around its mean, are of order $O_\lambda(N^{\frac{1}{2}}(N-n+1)^{\frac{1}{4}})$

Proof. We know from application of C-S, $Tr(\Sigma_{N,\lambda}P_{V^\perp}) \leqslant \|\Sigma_{N,\lambda}\|_F\|P_{V^\perp}\|_F$. Orthogonal projections satisfy, $\|P_{V^\perp}\|_F := \sqrt{Tr(P_{V^\perp}P_{V^\perp})} = \sqrt{N-n+1}$ and first result in (4.74) follows from Frobenius norm control of $\Sigma_{N,\lambda}$ given in (4.70).
From Talagrands' inequality, we know that:

$$P\bigg(\big|d(x,V) - \mathbb{E}d(x,V)\big| \geqslant \delta\big\|\Sigma_{N,\lambda}^{\frac{1}{2}}\big\|\bigg) \leqslant \exp\big(-\delta^2\big). \tag{4.76}$$

Therefore, fluctuations of d^2 around its mean are at most of order $O\big(\big\|\Sigma_{N,\lambda}^{\frac{1}{2}}\big\|\mathbb{E}d(x,V)\big)$, result now follows by using Jensens' inequality and ℓ_2 norm estimate of $\Sigma_{N,\lambda}^{\frac{1}{2}}$ given in (4.71). □

Remark 27. *Notice that $d^2(x,V) = \left\langle z_N, \Sigma_{N,\lambda}^{\frac{1}{2}}P_{V^\perp}\Sigma_{N,\lambda}^{\frac{1}{2}}z_N \right\rangle$ for $z_N \sim N(0, I_N)$, so it is a quadratic form and we just saw its' deviation around mean is at most of order $N^{\frac{1}{2}}(N-n+1)^{\frac{1}{4}}$. Most of the literature on sample complexity of system identification type problems, study quadratic forms using Hanson-Wright inequality([30]) which*

gives deviation in terms of $\|\Sigma_{N,\lambda}^{\frac{1}{2}} P_{V^\perp} \Sigma_{N,\lambda}^{\frac{1}{2}}\|_F, \|\Sigma_{N,\lambda}^{\frac{1}{2}} P_{V^\perp} \Sigma_{N,\lambda}^{\frac{1}{2}}\|_2$*; our result suggest using moment methods and approximation techniques (we used in preceding theorems) to unravel potential hidden dependencies on dimensionality and iterations while using Hanson-Wright.*

Proposition 8. *Few tail estimates that will be really helpful understanding least singular value of the data matrix(Hermitian case)*

$$Tr(\Sigma_{N,\lambda} P_{V^\perp}) \leqslant 4\|P_{V^\perp} \Sigma_{N,\lambda}^{\frac{1}{2}}\|^2 + \Big(\mathbb{E}\big\|P_{V^\perp} \Sigma_{N,\lambda}^{\frac{1}{2}} z_N\big\|\Big)^2$$
$$Tr(\Sigma_{N,\lambda}) \leqslant 4\|\Sigma_{N,\lambda}^{\frac{1}{2}}\|^2 + \Big(\mathbb{E}\big\|\Sigma_{N,\lambda}^{\frac{1}{2}} z_N\big\|\Big)^2,\ z_N \sim N\big(0, I_N\big)$$

Proof. Proof is standard Fubini type argument as in Appendix A.6 of [31],

$$\mathbb{E}d^2 - (Ed)^2 = \mathbb{E}\big|d(x,V) - Ed(x,V)\big|^2 = \int_0^\infty 2tP\Big(\big|d(x,V) - \mathbb{E}d(x,V)\big| > t\Big)dt \tag{4.77}$$
$$\leqslant \int_0^\infty 2t \times 2\exp\left(\frac{-t^2}{2\|P_{V^\perp} \Sigma_{N,\lambda}^{\frac{1}{2}}\|^2}\right)dt = 4\|P_{V^\perp} \Sigma_{N,\lambda}^{\frac{1}{2}}\|^2$$

□

In fact, we know much more about the projection matrix that can be helpful if instead of scalar Gaussian random variables system was excited by Bernoulli or some other random variable where we did not apriori know about its' higher dimensional moments

Corollary 5. *An application of Cauchy-Schwarz and* $P = P^2$ *as in Lemma 2.2 of*

[32] reveals:

$$\sum_{i=1}^{N} p_{ii}^2 \geqslant \frac{(\sum_{i=1}^{N} p_{ii})^2}{N} = \frac{(n-1)^2}{N}, \tag{4.78}$$

and sum of the squared $\frac{N(N-1)}{2}$ *distinct off-diagonal elements of the projection matrix can be upper bounded as:*

$$2\sum_{i=1}^{N}\sum_{j>i} p_{ij}^2 \leqslant (n-1) - \frac{(n-1)^2}{N} \leqslant \min[(n-1), N-n+1]. \tag{4.79}$$

Theorem 28. *From proposition 8 in Chapter 1, we know that:*

$$\sqrt{Tr(\Sigma_{N,\lambda}P_V^{\perp}) - 4\|P_V^{\perp}\Sigma_{N,\lambda}^{\frac{1}{2}}\|} \leqslant \mathbb{E}d(x,V) \leqslant \sqrt{Tr(\Sigma_{N,\lambda}P_V^{\perp})}. \tag{4.80}$$

We know from symmetric nature of Lipschitz' function concentration,

$$\mathbb{P}\bigg(d(x,V) \leqslant \mathbb{E}d(x,V) - \delta\|P_V^{\perp}\Sigma_{N,\lambda}^{\frac{1}{2}}\|\bigg) \leqslant \exp\bigg(-\frac{\delta^2}{2}\bigg) \tag{4.81}$$

Therefore,

$$\begin{aligned}
&\mathbb{P}\bigg(d(x,V) \leqslant \sqrt{Tr(\Sigma_{N,\lambda}P_V^{\perp}) - 4c_\lambda N^{\frac{1}{4}}} - \delta c_\lambda N^{\frac{1}{4}}\bigg) \\
&\leqslant \mathbb{P}\bigg(d(x,V) \leqslant \sqrt{Tr(\Sigma_{N,\lambda}P_V^{\perp}) - 4\|P_V^{\perp}\Sigma_{N,\lambda}^{\frac{1}{2}}\|} - \delta\|P_V^{\perp}\Sigma_{N,\lambda}^{\frac{1}{2}}\|\bigg) \\
&\leqslant \mathbb{P}\bigg(d(x,V) \leqslant \mathbb{E}d(x,V) - \delta\|P_V^{\perp}\Sigma_{N,\lambda}^{\frac{1}{2}}\|\bigg) \leqslant \exp\bigg(-\frac{\delta^2}{2}\bigg).
\end{aligned}$$

Only thing remaining is lower bounding $Tr(\Sigma_{N,\lambda}P_V^{\perp})$ which is actually:

$$Tr(\Sigma_{N,\lambda}P_V^{\perp}) = \sum_{i=1}^{N}\mathbb{E}[x_i^2](1-p_{ii}) - 2\sum_{j>i}\mathbb{E}[x_i x_j]p_{ij}, \tag{4.82}$$

although we have not been able to show accurate lower bound on $Tr(\Sigma_{N,\lambda}$, but lower bound on off-diagonal entries of projection matrix given in 5 from Chapter 2 and Section 4.3.1 can lead to correct results.

4.4 Statistics of bulk eigenvalues

One strategy for upper bounding the bulk eigenvalues follows naturally from interlacing given in (4.37), i.e., for $k \in [n]$

$$\lambda_k(\Sigma_n) \leqslant \lambda_1(\Sigma_{n-k+1}), \tag{4.83}$$

when combined with typical size of all the rows from Theorem (4.34) translates into

$$\sqrt{4^n\lambda^{2n}L_{\lambda,n}(k+1)\left\lfloor\frac{N}{n-k+1}\right\rfloor} \leqslant \sqrt{\lambda_1(\Sigma_{n-k+1})} \leqslant \sqrt{\big(N-n+\big)4^n\lambda^{2n}U_{\lambda,n}(k+1)}, \tag{4.84}$$

and consequently a lower bound on estimation error:

$$\left\|A-\hat{A}\right\|_F \geqslant \sigma_n(EX_-^*)\sqrt{\sum_{j=1}^{n}\frac{1}{\lambda_1^2\big(\Sigma_{n-k+1}\big)}}.$$

In fact, when we have typical size of all the entries of sample covariance matrix we can probably get even better estimates on the bulk via tools from Quadratic constrained optimization. For example, using min-max characterization of the eigenvalues:

$$\lambda_1(\Sigma_2) = \max\left(\sqrt{\langle y_{n-1}, y_{n-1}\rangle^2 + \langle y_{n-1}, y_n\rangle^2}, \sqrt{\langle y_n, y_n\rangle^2 + \langle y_{n-1}, y_n\rangle^2}\right) \tag{4.85}$$

So in the case of S-w-SSCs, $\lambda_1(\Sigma_2) = \langle y_{n-1}, y_{n-1}\rangle^2 + \langle y_{n-1}, y_n\rangle^2$ and

$$\lambda_2(\Sigma_2) = \inf_{a\in S^1} \sqrt{(a_1\langle y_{n-1}, y_{n-1}\rangle + a_2\langle y_n, y_{n-1}\rangle)^2 + (a_1\langle y_{n-1}, y_n\rangle + a_2\langle y_n, y_n\rangle)^2} \tag{4.86}$$

which is a quadratic constrained optimization problem. Again using the preceding argument:

$$\begin{aligned}\lambda_1^2(\Sigma_3) =& \left(\|y_{n-2}\|^2 + \langle y_{n-2}, y_{n-1}\rangle^2 + \langle y_{n-2}, y_n\rangle^2\right) \vee \left(\langle y_{n-2}, y_{n-1}\rangle^2 + \|y_{n-1}\|^2\right. \\ &\left. + \langle y_{n-1}, y_n\rangle^2\right) \vee \left(\|y_n\|^2 + \langle y_{n-2}y_n\rangle^2 + \langle y_{n-1}, y_n\rangle^2\right)\end{aligned}$$

For S-w-SSCs case: guess $\lambda_1(\Sigma_3) = \sqrt{\langle y_{n-2}, y_{n-2}\rangle^2 + \langle y_{n-2}, y_{n-1}\rangle^2 + \langle y_{n-2}, y_n\rangle^2}$, and

$$\begin{aligned}\lambda_2^2(\Sigma_3) = \inf_{a\in S^1} \Big[&(a_1\langle y_{n-2}, y_{n-2}\rangle + a_2\langle y_{n-2}, y_{n-1}\rangle)^2 + (a_1\langle y_{n-1}, y_{n-2}\rangle + a_2\langle y_{n-1}, y_{n-1}\rangle)^2 \\ &+ (a_1\langle y_n, y_{n-2}\rangle + a_2\langle y_n, y_{n-1}\rangle)^2\Big],\end{aligned}$$

which we may be able to explicitly compute using optimization principles and is a part of future plan of the authors' research.

4.4.1 Soft edge statistics of the martingale transform

Another evident quantity that appears in estimation error of least squares is edge spectral statistics of the commonly known martingale transform term EX_-^*. For a quick insight into the largest singular value of martingale transform assume $A = A^*$

with only one eigenvalue λ, so

$$EX_-^* = \sum_{i=1}^{N-1}\sum_{t=1}^{i}\begin{bmatrix} \langle w_i, e_1\rangle\langle w_{t-1}, e_1\rangle & \dots & \langle w_i, e_1\rangle\langle w_{t-1}, e_n\rangle \\ \vdots & \ddots & \vdots \\ \langle w_i, e_n\rangle\langle w_{t-1}, e_1\rangle & \dots & \langle w_i, e_n\rangle\langle w_{t-1}, e_n\rangle \end{bmatrix}$$

now pick $(1, 0, \dots, 0) =: a \in S^{n-1}$ then , then denote $s_i := \sum_{t=1}^{i} w_{t-1}$ and notice $s_{i+m} = \sum_{t=1}^{i+m} w_{t-1} = \underbrace{\sum_{t=1}^{i} w_{t-1}}_{s_i} + \sum_{t=i+1}^{i+m} w_{t-1}$

$$\mathbb{P}\Big(\big\|EX_-^* a\big\| \geqslant \delta\Big) \leqslant \delta^{-2}\sum_{j=1}^{n}\mathbb{E}\Big(\sum_{i=1}^{N-1}\langle w_i, e_j\rangle\langle\underbrace{\sum_{t=1}^{i} w_{t-1}}_{=:s_i}, e_1\rangle\Big)^2$$

$$\delta^{-2}\sum_{j=1}^{n}\sum_{i=1}^{N-1}\mathbb{E}(\langle w_i, e_j\rangle\langle s_i, e_1\rangle)^2 + 2\langle w_i, e_j\rangle\langle s_i, e_1\rangle\sum_{m=1}^{N-1-i}\langle w_{i+m}, e_j\rangle\langle s_{i+m}, e_1\rangle$$

$$\delta^{-2}\sum_{j=1}^{n}\sum_{i=1}^{N-1}\mathbb{E}\big[(\langle w_i, e_j\rangle\langle s_i, e_1\rangle)^2 + 2\langle w_i, e_j\rangle\langle s_i, e_1\rangle\sum_{m=1}^{N-1-i}(\langle w_{i+m}, e_j\rangle\langle s_i, e_1\rangle$$

$$+\langle w_{i+m}, e_j\rangle\langle\sum_{t=i+1}^{i+m} w_{t-1}, e_1\rangle)\big]$$

$$= \delta^{-2}\sum_{j=1}^{n}\sum_{i=1}^{N-1}\mathbb{E}\big[(\langle w_i, e_j\rangle\langle s_i, e_1\rangle)^2 + 2\langle w_i, e_j\rangle\langle s_i, e_1\rangle\sum_{m=1}^{N-1-i}\langle w_{i+m}, e_j\rangle\langle w_i, e_1\rangle$$

$$+\langle w_{i+m}, e_j\rangle\sum_{t'=i+2}^{i+m}\langle w_{t'-1}, e_1\rangle\big]$$

$$= \delta^{-2}\sum_{j=1}^{n}\sum_{i=1}^{N-1}\mathbb{E}\big[\langle w_i, e_j\rangle\langle w_i, e_j\rangle\langle s_i, e_1\rangle\langle s_i, e_1\rangle\big]$$

$$= \delta^{-2}\sum_{j=1}^{n}\sum_{i=1}^{N-1}\mathbb{E}[(\langle w_i, e_j\rangle)^2]\mathbb{E}[(\langle s_i, e_1\rangle)^2] = \delta^{-2}\sum_{j=1}^{n}\sum_{i=1}^{N-1} i = \frac{n(N-1)N}{\delta^2}.$$

Now one can trivially see that if $a = e_j \in S^{n-1}$, regardless of $j \in [n]$, $\|EX_-^* a\| = \frac{n(N-1)N}{\delta^2}$. Therefore, $\sigma_1(EX_-^*) = O\big(\frac{n(N-1)N}{\delta^2}\big)$. Now for Hermitian stable case, with only eigenvalue λ,

$$\mathbb{E}\bigg(\sum_{i=1}^{N-1}\langle w_i, e_j\rangle\underbrace{\langle\sum_{t=1}^{i}\lambda^{i-t}w_{t-1}, e_1\rangle}_{=:s_i(\lambda)}\bigg)^2 = \sum_{j=1}^{n}\sum_{i=1}^{N-1}\mathbb{E}[\langle s_i(\lambda), e_1\rangle^2] = \sum_{j=1}^{n}\sum_{i=1}^{N-1}\sum_{t=1}^{i}\lambda^{2(i-t)}$$

Therefore, $$\sigma_1(EX_-^*) = O\bigg(\frac{n(N-\sum_{i=0}^{N-1}\lambda^{2i})}{1-\lambda^2}\bigg).$$

For the peculiar case of S-w-SSCs with $\lambda \in \big(\frac{1}{2}, 1\big)$, notice that

$$EX_-^* = \sum_{i=1}^{N-1} w_i x_i^* = EX_-^* = \sum_{i=1}^{N-1} w_i\bigg(\sum_{t=1}^{i}\sum_{m=0}^{(i-t)\wedge n}\binom{i-t}{m}N_n^m\lambda^{i-t-m}w_{t-1}\bigg)^* \tag{4.87}$$

Now for $j, k \in [n]$

$$\begin{aligned}
[EX_-^*]_{j,k} &= \sum_{i=1}^{N-1}\sum_{t=1}^{i}\sum_{m=0}^{(i-t)\wedge n}\binom{i-t}{m}\lambda^{i-t-m}\big\langle w_i w_{t-1}^* N_n^{m*} e_k, e_j\big\rangle \\
&= \sum_{i=1}^{N-1}\sum_{t=1}^{i}\sum_{m=0}^{(i-t)\wedge(n-k)}\binom{i-t}{m}\lambda^{i-t-m}\big\langle w_i w_{t-1}^* e_{k+m}, e_j\big\rangle \\
&= \sum_{i=1}^{N-1}\sum_{t=1}^{i}\sum_{m=0}^{(i-t)\wedge(n-k)}\binom{i-t}{m}\lambda^{i-t-m}\langle w_i, e_j\rangle\langle w_{t-1}, e_{k+m}\rangle
\end{aligned}$$

Now in order to find the order of $\sigma_1(EX_-^*)$, we need to compute

$$\begin{aligned}
&\max_{k\in[n]}\sum_{j=1}^{n}\mathbb{E}\Bigg(\sum_{i=1}^{N-1}\langle w_i,e_j\rangle\sum_{t=1}^{i}\sum_{m=0}^{(i-t)\wedge(n-k)}\binom{i-t}{m}\lambda^{i-t-m}\langle w_{t-1},e_{k+m}\rangle\Bigg)^2\\
=\;&\max_{k\in[n]}\sum_{j=1}^{n}\sum_{i=1}^{N-1}\mathbb{E}\Bigg(\sum_{t=1}^{i}\sum_{m=0}^{(i-t)\wedge(n-k)}\binom{i-t}{m}\lambda^{i-t-m}\langle w_{t-1},e_{k+m}\rangle\Bigg)^2\\
=\;&\sum_{j=1}^{n}\sum_{i=1}^{N-1}\mathbb{E}\Bigg(\sum_{t=1}^{i}\sum_{m=0}^{(i-t)\wedge(n-1)}\binom{i-t}{m}\lambda^{i-t-m}\langle w_{t-1},e_{1+m}\rangle\Bigg)^2=\sum_{j=1}^{n}\mathbb{E}\big\|y_1\big\|^2=n\mathbb{E}\|y_1\|^2,
\end{aligned}\tag{4.88}$$

where (4.88) follows along the similar argument in deducing $\sigma_1(X_-)$ of $S-w-SSCs$; first row has the largest typical size. This bring us to a remarkable observation

Theorem 29. *Largest singular value of martingale term is upper bounded by $\sqrt{n}$ times of largest singular value of the data matrix*

$$\sigma_1(EX_-^*)\leqslant\sqrt{n}\sigma_1(X_-),$$

In fact this was first mentioned in [23] by using Courant-Fischer that $\sigma_1(EX_-^*)\leqslant\sigma_1(E_{Im(X_-^*)})\sigma_1(X_-)$ where $E_{Im(X_-^*)}$ is essentially $n\times n$ Gaussian ensemble, known to have largest singular value of order $\sqrt{n}$.

4.5 Non-vanishing error for stable-spatially inseparable case via tensorization of Talagrands' inequality: Heurestics

Dimension+iteration independent tensorization of Talagrands' inequality would require sufficient independence between the covariates. In fact tensorization for stable ARMA processes had been of much interest to people working in the field of *Functional Inequalities* in early 2000s. [6] was able to show dimension+iteration

independent tensorization under the assumption of $\|A\|_2 < 1$. [5] concludes dimension+iteration independent tensorization under the assumption of spectral radius being strictly inside unit circle, which turns out to be incorrect in high dimensions.

Definition 8 (Dimension + Iteration free Tensorization)**.** *Let* $(x_1, x_2, \ldots, x_N) \in \mathbb{R}^{n^{\otimes N}}$ *be a length* $N-$ *trajectory of an* $n-$ *dimensional linear Gaussian as in* (3.1)*, and* $\mu_n^N := Law(x_1, x_2, \ldots, x_N)$ *be its' distribution. We say that the process* $(x_1, x_2, \ldots, x_N)$ *satisfies dimension+iteration independent Talagrands' inequality, if for all probability measures* ν *on* $\mathbb{R}^{n^{\otimes N}}$ *which are absolutely constant w.r.t* μ_n^N*(* $\nu << \mu_n^N$*), we have that*

$$W_1^d\big(\mu_n^N, \nu\big) \leqslant \sqrt{2CH\big(\nu||\mu_n^N\big)}, \tag{4.89}$$

where C *is independent of* N *and* n*.*

Question: Now given that dynamics are generated from stable $n-$ dimensional linear Gaussian (i.e., $\rho(A) < 1$) can we conclude dimension+iteration independent tensorization of Talagrands' inequality?

The slickest way to conclusion is a realization that $(x_1, x_2, \ldots, x_N)$ is a function of the Gaussian ensemble $(w_0, w_1, \ldots, w_{N-1})$ as $x_i = \sum_{t=1}^{i} A^{i-t} w_{t-1}$. Let $F(w_0, \ldots, w_{N-1})$ $:= \sqrt{\sum_{i=1}^{N} \|x_i\|^2}$. As isotropic Gaussians satisfy dimension+iteration independent tensorization (Theorem 2), when combined with equivalent notion of Lipschitz concentration(Remark 1) implies for all $\delta > 0$:

$$\mathbb{P}\bigg(\big|F(w_0, w_1, \ldots, w_{N-1}) - \mathbb{E}F(w_0, w_1, \ldots, w_{N-1})\big| > \delta\bigg) \leqslant 2\exp\bigg(-\frac{\delta^2}{2\|F\|_L^2}\bigg), \tag{4.90}$$

but the map $(x_1, x_2, \ldots, x_N) \mapsto \sqrt{\sum_{i=1}^{N} \|x_i\|^2}$ is one-Lipschitz in ℓ_2 additive metric on $\mathbb{R}^{n^{\otimes N}}$ and Remark 1 implies that $\mu_n^N \in T_1\big(\|F\|_L^2\big)$. We first need to collect some estimates that will lead to conclusive answer to tensorization of Talagrands' inequality

for stable $n-$ dimensional ARMA model.

Proposition 9. *Frobenius norm of the data matrix populated by $n-$ dimensional S-w-SSCs is precisely:*

$$\left\|X_-\right\|_F^2 = \sum_{i=1}^{N}\sum_{j=1}^{n}\sum_{s,t=1}^{i}\sum_{m=0}^{(i-t)\wedge(n-j)}\sum_{m'=0}^{(i-s)\wedge(n-j)}\binom{i-t}{m}\binom{i-s}{m'} \lambda^{i-t-m}\overline{\lambda^{i-s-m'}}\langle w_{t-1}, e_{j+m}\rangle\overline{\langle w_{s-1}, e_{j+m'}\rangle}$$

Proof. We again leverage upon the inner product structure to conclude.

$$\begin{aligned}
\sum_{i=1}^{N}\left\|x_i\right\|^2 &= \sum_{i=1}^{N}\sum_{s,t=1}^{i}\sum_{m=0}^{(i-t)\wedge n}\sum_{m'=0}^{(i-s)\wedge n}\binom{i-t}{m}\binom{i-s}{m'}\lambda^{i-t-m}\overline{\lambda^{i-s-m'}}\langle N_n^m w_{t-1}, N_n^{m'} w_{s-1}\rangle \\
&= \sum_{i=1}^{N}\sum_{s,t=1}^{i}\sum_{m=0}^{(i-t)\wedge n}\sum_{m'=0}^{(i-s)\wedge n}\binom{i-t}{m}\binom{i-s}{m'}\lambda^{i-t-m}\overline{\lambda^{i-s-m'}}\sum_{j=1}^{m\wedge m'}\langle w_{t-1}, e_{j+m}\rangle\overline{\langle w_{s-1}, e_{j+m'}\rangle} \\
&= \sum_{i=1}^{N}\sum_{s,t=1}^{i}\sum_{m=0}^{(i-t)\wedge n}\sum_{m'=0}^{(i-s)\wedge n}\sum_{j=1}^{m\wedge m'}\binom{i-t}{m}\binom{i-s}{m'}\lambda^{i-t-m}\overline{\lambda^{i-s-m'}}\langle w_{t-1}, e_{j+m}\rangle\overline{\langle w_{s-1}, e_{j+m'}\rangle} \\
&= \sum_{i=1}^{N}\sum_{j=1}^{n}\sum_{s,t=1}^{i}\sum_{m=0}^{(i-t)\wedge(n-j)}\sum_{m'=0}^{(i-s)\wedge(n-j)}\binom{i-t}{m}\binom{i-s}{m'} \\
&\quad \lambda^{i-t-m}\overline{\lambda^{i-s-m'}}\langle w_{t-1}, e_{j+m}\rangle\overline{\langle w_{s-1}, e_{j+m'}\rangle}
\end{aligned}$$

□

Theorem 30. *$\|F\|_L$ is the smallest positive constant $L_{n,N}$ such that:*

$$\left\|X_-\right\|_F = \sqrt{\sum_{i=1}^{N}\Big\langle\sum_{t=1}^{i}A^{i-t}w_{t-1}, \sum_{s=1}^{i}A^{i-s}w_{s-1}\Big\rangle} \leqslant L_{n,N}\sqrt{\sum_{i=0}^{N-1}\left\|w_i\right\|^2} = L_{n,N}\left\|E\right\|_F, \tag{4.91}$$

in the limit of large number of iterations we have,

- *Hermitian and stable case $L_{n,N} = \Theta(1)$: i.e, independent of dimension+iteration*

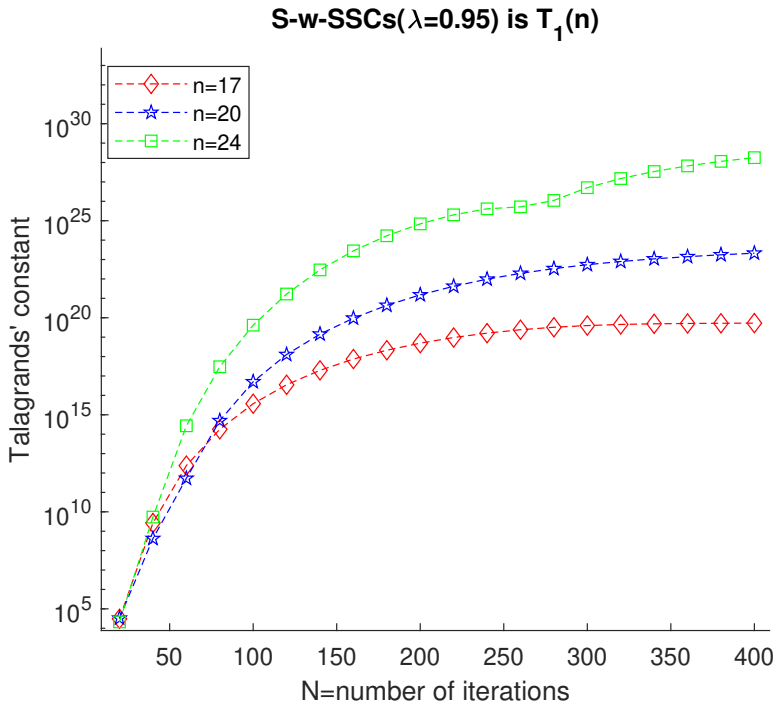

Figure 4.4: Tensorization is exponential in $n-$ and independent of N for large N. Verified for S-w-SSCs(with $\lambda = 0.95$)

- *Stable but spatially inseparable case with* $\lambda \in (\frac{1}{2}, 1)$, *then* $L_{n,N} = \Theta(e^n)$*: i.e., linear in underlying dimension of the state space and independent of iterations.*

Proof. of Hermitian and stable case is obvious after recalling spectral theorem(i.e., all the rows are essentially one-dimensional scalar stable ARMA). Let $[\lambda_j]_{j=1}^n$ be the eigenvalues of A which are assumed to be strictly inside the unit circle.

$$\sum_{i=1}^{N-1}\sum_{j=1}^{n}\Big|\sum_{t=1}^{i}\langle A^{i-t}w_{t-1}, e_j\rangle\Big|^2 = \sum_{i=1}^{N-1}\sum_{j=1}^{n}\Big|\sum_{t=1}^{i}\langle \lambda_j^{i-t}w_{t-1}, e_j\rangle\Big|^2 \leqslant L_{n,N}^2\sum_{i=0}^{N-1}\|w_i\|^2$$

$$\textit{where, } L_{n,N} \leqslant \min_{\epsilon>0}\max_{j\in[n]}\sqrt{(1+\epsilon^{-1})\sum_{i=0}^{N-1}\big((1+\epsilon)|\lambda_j|^2\big)^k} \leqslant \frac{1}{1-\rho}, \tag{4.92}$$

where proof of (4.92) is given in Proposition 4.1 of [5]. However, most difficult but

most interesting case is for trajectory generated from $n-$ dimensional $S-w-SSCs$ and $\lambda \in (\frac{1}{2}, 1)$: and an immediate daunting approach suggested by Prooposition 9 would be a tight estimate on $L_{n,N}$ such that:

$$\sum_{i=1}^{N}\sum_{j=1}^{n}\sum_{s,t=1}^{i}\sum_{m=0}^{(i-t)\wedge(n-j)}\sum_{m'=0}^{(i-s)\wedge(n-j)}\binom{i-t}{m}\binom{i-s}{m'}\lambda^{i-t-m}\overline{\lambda^{i-s-m'}}\langle w_{t-1}, e_{j+m}\rangle\overline{\langle w_{s-1}, e_{j+m'}\rangle} \leqslant L_{n,N}^2\sum_{i=0}^{N-1}\left\|w_i\right\|^2.$$

Recall from corollary 4 that $\sigma_1(X_-)$ has a typical size exponential in n.Now recall that $\left\|E\right\|_F = \sqrt{\sum_{i=0}^{N-1}\left\|w_i\right\|^2}$ has a typical size of $\Theta\big(\sqrt{nN}\big)$. Notice that, for S-w-SSCs Talagrands' constant is $\Theta(e^n)$, thus:

$$L_{n,N} = \frac{\sqrt{\sum_{i=1}^{n}\sigma_i^2(X_-)}}{\sqrt{\sum_{i=1}^{n}\sigma_i^2(E)}} = \frac{\sqrt{\sum_{i=2}^{n}\sigma_i^2(X_-) + \Theta(\sqrt{N-n+1}e^n)^2}}{\Theta(\sqrt{nN})}, \tag{4.93}$$

and the result follows. □

Corollary 6. *In high dimensions, OLS estimation on data corresponding to S-w-SSCs has non-vanishing error in Frobenius norm.*

Proof. Using Courant-Fischer and the fact that $\sigma_n(E) = \sqrt{N} - \sqrt{n-1}$ with high probability(see e.g., [26]) we conclude:

$$\begin{aligned}\left\|A - \hat{A}\right\|_F = \left\|EX_-^*(X_-X_-^*)^{-1}\right\|_F &\geqslant \frac{\left\|EX_-^*\right\|_F}{\sigma_1^2(X_-)} \geqslant \frac{\sigma_n(E)\|X_-\|_F}{\sigma_1^2(X_-)} \\ &= \frac{(\sqrt{N}-\sqrt{n-1})\sqrt{nN}e^n}{(N-n+1)e^n} = \frac{(\sqrt{N}-\sqrt{n-1})\sqrt{nN}}{(N-n+1)} > 0\end{aligned}$$

□

Chapter 5

Concentration of measure phenomonen for Random Dynamical Systems an operator theoretic approach

Via operator theoretic methods, we formalize the concentration phenomenon for a given observable 'r' of a discrete time Markov chain with 'μ_π' as invariant ergodic measure, possibly having support on an unbounded state space. The main contribution of this paper is circumventing tedious probabilistic methods with a study of a composition of the Markov transition operator P followed by a multiplication operator defined by e^r. It turns out that even if the observable/ reward function is unbounded, but for some for some $q > 2$, $\|e^r\|_{q\to 2} \propto \exp\left(\mu_\pi(r) + \frac{2q}{q-2}\right)$ and P is hyperbounded with norm control $\|P\|_{2\to q} < e^{\frac{1}{2}[\frac{1}{2}-\frac{1}{q}]}$, sharp non-asymptotic concentration bounds follow. *Transport-entropy* inequality ensures the aforementioned upper bound on multiplication operator for all $q > 2$. Also, the role of *reversibility* in concentration phenomenon is demystified. These results are particularly useful for the reinforcement learning and controls communities by allowing for concentration inequalities w.r.t standard unbounded obersvables/reward functions where exact knowledge of the system is not available, let alone the reversibility of stationary measure.

Motivation. With the ubiquity of Machine Learning over last two decades, we have seen a tremendous surge in research activity on non-asymptotic analysis of system identification and reinforcement learning for linear time-invariant dynamical systems (e.g., [2, 3, 11, 33–36]). Analysis developed in this paper is a continuation of, [37]. Initially, motivated by extending non-asymptotic analysis of average reward based

optimal control shown for LDS to SLDS where expected value of the reward w.r.t stationary distribution of a Markov chain is approximated by its' empirical averages, with high probability. Although, sample complexity for control/dynamical systems on continuous state space has been extremely popular recently, but sharp results are limited to stable LDS with Gaussian noise, for system identification point of view see ([11]) and average reward based context via uniform Transportaion-entropy inequality (T-E) is discussed in section 2.4 of ([37]); current understanding of the world outside of the linear Gaussian case is limited to non-existent. We previously leveraged upon *concentration of measure phenomenon* for process level law of Markov chain via uniform T-E inequalities, to prove sharp concentration for Lipschitz observables (for a good monograph see [6, 38]). T-E approach is compact, concise and clear: even leading to exponential concentration for unbounded reward function of the form $r(x) := \|x\|$ for (LDS). However, uniform T-E inequality is not valid for Harris ergodic Markov chains (HEMCs) and one might resort to brute force analysis which becomes intractable.

In their previous work ([37]), under the existence of exponential-type Lyapunov function, could only show sharp deviation inequalities for i.i.d samples from the stationary distribution of HEMCs. However, an important realization was not used: exponential type Lyapunov function implies exponential decay of correlation between two time separated observations of any given Lipshitz observable (Theorem 3.3 of [37]): alluding to room for improvement that we will explore in this paper. Preceding phenomenon is related to *Poincare/ Spectral gap Inequality* which is well studied in a parallel line of research in Physics and Functional analysis: the study of convergence Of Markov transition operators and its' variants via spectral theory (see e.g., [39, 40]). This work is a continuation of concentration phenomenon for nonlinear random dynamical system, through the lens of Operator and Spectral theory, when-

ever possible we will attempt to come up with easily verifiable Lyapunov conditions that are equivalent to some necessary or sufficient phenomenon for concentration in Operator theoretic terms.

Further implications of the analysis: estimating steady state correlations via sample covariance matrices. Let us assume that we are observing i.i.d samples $(x_i)_{i=1}^N$ from a centered distribution P_∞ on $\mathbb{R}^n$, where n is high dimensional. For example, in finance, each x_i could be return on n stocks and N the number of trading days [41]. One wishes to estimate P_∞ via $\Sigma_N := \frac{1}{N}\sum_{i=0}^{N-1} x_i x_i^T$; in order to understand most significant factors and spatial correlations between the n variables. It is well known in statistics literature that for $N >> n$, empirical covariance matrix Σ_N is a good approximation of P_∞. Now, consider covariates $(x_i)_{i=1}^N$ as realizations of a random dynamical system in steady state; good estimate of P_∞ is imperative for designing reduced order models, after performing principal component analysis. However, samples are now temporally correlated, along with spatial correlations. Thus, one would expect sample complexity to be worse than in the i.i.d case. Intuitively, one needs to quantify decay of temporal correlation/spectral gaps (as we do in Section 5.2), and then average only over 'distant' samples to get an accurate estimation of the stationary covariance matrix. Analysis developed in this paper can be extended, as part of future work, to estimation of correlations where the observable would be instead $f(x) = xx^T$.

Related Literature. Techniques used to prove concentration of measure include martingale methods, exchangeable pairs e.g., (see [42]) and functional inequalities: transportation-entropy, logarithmic Sobolev/ hypercontractivity and Poincare, [43]. However, exploring measure concentration for dependent random variables as in

Markov chains on unbounded spaces limit the application of martingale methods and exchangeable pairs. RL and system identification community, pretty much exclusively, have used *independent block technique* [11, 34] that seems to work when observables are *bounded*. SLDS fall into the category of Markov chains admitting a Lyapunov function with geometric drift condition. Following [44], by using martingale and independent block techniques, authors in [45] were able to bound the mean-squared error between empirical average and expected reward (unbounded) from stationary distribution of SLDS but probabilistic methods employed are tedious, opaque and leads to weak results. Stein methods are well suited towards discrete setting in statistical physics but have limitations for model under consideration. Only recently, [46] started a formal, functional analytic study for concentration inequalities in discrete-time setting using transport-information(T-I) inequality. Verification of T-I is plausible either in discrete state space setting, or when Markov chain is reversible and posseses a spectral gap in space of square integrable functions w.r.t its' stationary distribution. Although mathematically sound and compact results exist in [46], they bury under them profound insights into when concentration phenomenon may or may not work ; these insights are necessary to abridge the gap between theory and application in RL and controls domain. Spectral gap assumption for entire space of square integrable functions is very strong, as even a naive i.i.d oriented mindset can realize that one may only need exponential decay of correlation w.r.t observables that we aim to show concentrate sharply as ergodic averages around their stationary mean. Also, verifiable conditions like existence of Lyapunov function which lead to concentration inequality is restricted to the reversible case. An alternative line of research is taken by [47], where under the assumption of a spectral gap for Markov transition operator he uses perturbation theory to show spectral gap for Feynman-Kac operator associated to observable under consideration which has to be bounded

for analysis to work. As we will see in concentration for nonlinear random dynamical systems, spectral gaps might only exist in Wasserstein sense. Reversibility assumption originates from study of Langevin-type stochastic differential equations : used to model physical phenomenons in the nature; however, the scope of this paper is not limited to reversibility assumption. [5] and [6] have studied concentration phenomenon for stable LDS via process level T-E inequalities ; their approach become intractable for nonlinear random dynamical systems, partly because dynamics are not necessarily contractive in trivial euclidean metric.

Space of probability measure on $\mathcal{X}$ is denoted by $\mathcal{P}(\mathcal{X})$ and space of its Borel subsets is represented by $\mathbb{B}\big(\mathcal{P}(\mathcal{X})\big)$. For a function r and $\mu \in \mathcal{P}(\mathcal{X})$, we use $< r >_\mu$ to denote expectation of r w.r.t μ. Finally, for a set $\mathcal{K} \subseteq \{1, ..., M\}$, its complement is $\mathcal{K}^{\complement} := \{1, ..., M\} \backslash \mathcal{K}$.

5.0.1 Problem Statement

Under the action of some state dependent policy π, we consider a closed loop random dynamical system of the form:

$$x_{k+1} = F\big(x_k, \pi(x_k), \epsilon_k\big), \quad \epsilon_k \sim \mathcal{N}(0, I_k) \quad i.i.d, \tag{5.1}$$ [1]

where $x_k \in \mathbb{R}^n$ for all $k \in \mathbb{N}$ and $F : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$. In probabilistic language, closed loop dynamical system is a *Harris Ergodic Markov chain.* Assume, transition kernel converges to some stationary distribution μ_π under Wasserstein metric, $\mathcal{W}_{d_{\beta *}}$ equipped with some distance function $d_{\beta *}$. If we have access to empirical averages of some unbounded reward function $r(x)$, we explore the following questions:

[1] For the sake of brevity, from now on we will exclude reference to π in state update equations as state dependent policy imlplies there exists some function G such that $F\big(x_k, \pi(x_k), \epsilon_k\big) = G(x_k, \epsilon_k)$.

- *Concentration from simulating a single trajectory*

  When, how and why can we provide something similar to following exponential concentrations ?

$$\mu^N\left[\left|\frac{1}{N}\sum_{i=1}^{N} r(x_i) - < r >_{\mu_\pi} \right| > \epsilon\right] \leqslant 2\exp\left(-\frac{N\epsilon^2}{K_{sys}(r)}\right), \tag{5.2}$$

  where r can be some unbounded function, in control theoretic or (RL) framework $r(x) := \|x\|$. $K_{sys}(r)$ is a constant dependent on system properties and 'smoothness' of r (related to Lipschitz constant)

- What sort of *regularity assumptions* on Markov transition operator are required to ensure concentration?

- How can we ensure *exponential-integrability* of the observables under the stationary measure?

- To come up with easily verifiable *Lyapunov conditions*, that ensures the aforementioned objectives.

- What is the role of the omnipresent *reversibility assumption* in concentration phenomenon?

5.0.2 Contribution and Main Results

- This paper's fundamental contribution is reducing concentration phenomenon for Harris Ergodic Markov chain to a problem of bounding operator $e^r P$: $L^2(\mu_\pi) \longrightarrow L^2(\mu_\pi)$, a composition of Markov operator P and a multiplication operator e^r.

- Under the easily verifiable condition of *exponential-type Lyapunov function*, a *central limit theorem* is proven w.r.t centered unbounded observable.

- A detailed mathematical and pedagogical overview of *Poincare and Spectral gap inequality* is provided, as understanding the causal events for the converse inequality reveals where things may go wrong. For example, it is found that *periodic* behavior of HEMCs can be a hindrance to sharp concentration phenomenon .

- Even though multiplication operator e^r, associated with standard reward function (observable) used in continuous control and RL problems is *unbounded*, but *exponential-type Lyapunov function* ensures that for all $q > 2$, $e^r : L^q(\mu_\pi) \longrightarrow L^2(\mu_\pi)$ is bounded. Combined with *hyperboundedness* of Markov kernel, concentration phenomenon is achieved.

- Consequently, given a possibly unbounded-although Lipschitz observable 'r' in some metric $d_{\beta*}$, a sufficient condition for concentration phenomenon is invariant measure 'μ_π' satisfies transport-entropy inequality with underlying metric $d_{\beta*}$ and Markov transition kernel is Hyperbounded.

- If the *Range space* of $(I - P)$ is closed for an *ergodic and aperioidic Markov chain*, non-asymptotic uncorrelation (Poincare inequality) follows. If dimension of *Null space* of $(I-P^*)$ is finite, then *Range space* of $(I-P)$ is closed. However, for reversible chain *Null space* of $(I - P^*)$ is *Null space* of $(I - P)$ and P is *ergodic*: only constant functions belong to *Null space* of $(I - P)$ and dimension is 1. So, *non-reversible* case requires some alternative knowledge to know *Range space* of $(I - P)$ is closed.

Outline of paper. In Section 2, we give a brief recap of how uniform Transport-Entropy inequality and contractivity in Wasserstein metric of the underlying Markov chain (LDS with i.i.d Gaussian noise) lead to sharp concentration of ergodic averages of unbounded Lipschitz observables. Section 3 begins with formulation of Harris Ergodic Markov chains as defined by, [48] via Lyapunov drift and minorization condition and a *central limit theorem* is proven for Lipschitz observables under the assumption of *exponential type Lyapunov function*. In Section 4, essentials of ergodic theory, spectral theory and non-asymptotic uncorrelation phenomenon are studied via operator theoretic convergence of associated Markov transition operators and a detailed discussion on necessity and implications of Poincare and spectral gap inequalities. Susceptibility of HEMCs to periodic behavior, which can be seen as a potential hindrance to concentration phenomenon, is discussed with an example. In section 5, Large deviation principle for Markov Chains is discussed and concentration inequalities are derived under the hyperboundedness and norm condition of multiplication operator. In section 6, we discuss correct rate function for occupation measure of Markov chain and associated functional inequality for concentration is provided with a Lyapunov condition. We conclude the paper in section 7, with discussion and problems for future consideration.

Study plan: We suggest that after finishing section 2, reader who is not very familiar with functional analytic framework for Markov chains; should skip to section 4 and read upto (but not including) subsection 4.3, before coming back to section 3.

5.0.3 Contractivity and Uniform Transport Constants

As one would wonder, when does the T-E for process level law of Markov chain, increases at worse linearly with dimension (in sample term)? Sufficient conditions

(see e.g., [6, 49]) are

$$(i) \quad P(x,\cdot) \in T_1^d(C), \qquad \text{for all } x \in \mathcal{X}, \text{and some } C > 0, \tag{5.3}$$

$$(ii) \quad \mathcal{W}_d(P(x,\cdot), P(y,\cdot)) \leqslant \hat{\lambda} d(x,y), \quad \text{for all } (x,y) \in \mathcal{X}^2 \text{and some } \hat{\lambda} \in [0,1). \tag{5.4}$$

Property (5.3) is often referred to as existence of a uniform transportation constant and (5.4) represents contractivity of the Markov Chain in the Wasserstein metric / *spectral gap in the Wasserstein sense*. Now, the following result holds.

Lemma 7. *If* (5.3) *and* (5.4) *hold, process level distribution of samples from a Markov chain* $(x_1, \ldots, x_N)$*, which we will denote as* $Law(x_1, \ldots, x_N)$*, denoted by* μ^N *satisfies* $T_1^{d_{(N)}}\left(\frac{CN}{(1-\hat{\lambda})^2}\right)$*, for all* $N \in \mathbb{N}$.

Proof. See Theorem 2.5 of [6] for a detailed proof. □

5.0.4 Sharp Deviation Inequalities for Average-Reward Based Optimal Control

In optimal control and average reward reinforcement learning for *continuing task*, it is often the case that under the action of some state dependent policy, the resulting closed-loop dynamical system under consideration, $(x_i)_{i=1}^N$, mixes to some stationary distribution $\mu_\pi \in P(\mathcal{X})$. When exact system parameters and state are unknown but time-averages of the reward function $r(x) := \|x\|$, although unbounded, are available, sharp concentration bounds of $\frac{1}{N}\sum_{i=1}^N r(x_i)$ around $< r >_{\mu_\pi}$ are of utmost importance to ensure policy update algorithm converges in finite time: as knowledge of $< r >_{\mu_\pi}$ is required in actor-critic algorithm for average reward case, see e.g.,[50]. Sufficient conditions for concentration of empirical averages of $r(x) := \|x\|$ or any Lipschitz function essentially boils down to showing that the process level law of the Markov chain $\mu^N \in \mathcal{T}_1^{d_{(N)}}\big(\mathcal{O}(N)\big)$.

Theorem 31. *Assume that under a metric d, a random dynamical system uniformly satisfies T-E inequality with constant $C > 0$ and is contractive in Wasserstein sense with constant $\gamma \in (0,1)$. If we start deterministically with $x_0 := x$, we have for any 1 Lipschitz function r, the following deviation inequality:*

$$\mu^N\left[\left|\frac{1}{N}\sum_{i=1}^{N} r(x_i) - < r >_{\mu_\pi} \right| > \frac{\mathcal{W}_d(P(x.\cdot),\mu_\pi)}{N(1-\gamma)} + \epsilon\right] \leqslant 2\exp\left(-\frac{N\epsilon^2(1-\gamma)^2}{2C}\right). \tag{5.5}$$

Proof. Since the theorem assumptions satisfy the claim in (7), it holds that $\mu^N \in \mathcal{T}_1^{d_{(N)}}\left(\frac{CN}{(1-\gamma)^2}\right)$. Let $(y_i)_{i=1}^N$ be i.i.d samples from μ_π and assume that the Markov chain starts deterministicly with $x_1 = x$. Then, we have for all $\epsilon > 0$ it holds

$$\begin{aligned}
&\mu^N\left[\left|\frac{1}{N}\sum_{i=1}^{N} r(x_i) - < r >_{\mu_\pi} \right| > \frac{\mathcal{W}_d(P(x,\cdot),\mu_\pi)}{N(1-\gamma)} + \epsilon\right] \\
&\leqslant \mu^N\left[\left|\frac{1}{N}\sum_{i=1}^{N} r(x_i) - \mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N} r(x_i)\right)\right| + \left|\mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N} r(y_i)\right) - \mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N} r(x_i)\right)\right| \right. \\
&\qquad\qquad \left. > \frac{\mathcal{W}_d(P(x.\cdot),\mu_\pi)}{N(1-\gamma)} + \epsilon\right] \\
&\leqslant \mu^N\left[\left|\frac{1}{N}\sum_{i=1}^{N} r(x_i) - \mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N} r(x_i)\right)\right| > \epsilon\right] \qquad (5.6) \\
&\leqslant 2\exp\left(-\frac{N\epsilon^2(1-\gamma)^2}{2C}\right), \qquad (5.7)
\end{aligned}$$

where (5.6) follows from contractive dynamics in Wasserstein distance. □

Remark 32. *The structure of the aforementioned concentration inequality is inspired by the work in [51] on bounding deviations of empirical measure formed by interacting particle systems from their infinite particle limit (Mckean-Vlasov diffusion); see e.g., problem section in [52] for an involved discussion on this matter. However, there is a subtle difference in our approach, as we work with an ℓ^2 inspired metric on $\mathcal{X}^N$ as*

in [51], i.e.,

$$d^2_{(N)}(x^N, y^N) := \sqrt{\sum_{i=1}^N d^2(x_i, y_i)}. \tag{5.8}$$

Then, $\Phi(x^N) := \frac{1}{N}\sum_{i=1}^N r(x_i)$ *w.r.t* ℓ^2 *metric is only* $\frac{1}{\sqrt{N}}$ *Lipschitz — i.e.,* $\|\Phi\|_{L(d^2_{(N)})} \leqslant \frac{1}{\sqrt{N}}$ *and any hope for concentration would require* $\mu^N \in \mathcal{T}_1^{d^2_{(N)}}\big(\mathcal{O}(1)\big)$ *(i.e.,* dimension free concentration*), which is very difficult to check; this is feasible only in Markov diffusion processes with uniformly convex external potentials and symmetric interaction potentials, see e.g., [51].*

Decay of correlation. By combining conditions from (5.3) and (5.4), with Taylor's expansion for small λ (terms of order up to λ^2) appearing on both sides in Bobkov-Gotze dual form (5.37), for all $x \in \mathcal{X}$ it holds that

$$|Cov_{P_x}[f(x_n), f(x_{n+k})]| \leqslant \frac{\hat{\lambda}^k}{1-\hat{\lambda}^2} C\|f\|^2_{L(d)}. \tag{5.9}$$

Sharp Concentration via Single Trajectory of norm-Stable/ contractive Linear Dynamical Systems: Consider a discrete-time linear dynamical system (LDS) of the form

$$y_{t+1} = Ay_t + \epsilon_t, \quad \|A\|_2 = \hat{\lambda} < 1 \quad \text{and i.i.d} \quad \epsilon_t \sim \mathcal{N}(0, \mathcal{I}_n). \tag{5.10}$$

With the trivial euclidean metric $d(x, y) := \|x - y\|$, the transition kernel from (5.10) satisfies $P(x, \cdot) \in T_1^d(1)$, $\forall x \in \mathcal{X}$ [17] and $\mathcal{W}_d^2(P(x, \cdot), P(y, \cdot)) = \|Ax - Ay\| \leqslant \hat{\lambda} d(x, y)$ (see e.g., [53]). Now, an application of Jensens' inequality reveals $\mathcal{W}_d(P(x, \cdot), P(y, \cdot)) \leqslant \mathcal{W}_d^2(P(x, \cdot), P(y, \cdot))$ and contractivity follows. As conditions (5.3) and (5.4) are satisfied, we can use coupling technique inspired by [38] to prove that $\mathbb{P}^N :=$Law$(y_1, \ldots, y_N)$ of the LDS satisfies $T_1^{d_{(N)}}(\frac{N}{(1-\hat{\lambda})^2})$; and if ν_π is the invariant

measure corresponding to (5.10), we have

$$\mathbb{P}^N\left[\left|\frac{1}{N}\sum_{i=1}^{N}\|y_i\| - < \|y\| >_{\nu_\pi}\right| > \frac{\mathcal{W}_d(P(x.\cdot),\nu_\pi)}{N(1-\hat{\lambda})} + \epsilon\right] \leqslant$$
$$\leqslant \mathbb{P}^N\left[\left|\frac{1}{N}\sum_{i=1}^{N}\|y_i\| - \mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N}\|y_i\|\right)\right| > \epsilon\right] \leqslant 2\exp\left(-\frac{N\epsilon^2(1-\hat{\lambda})^2}{2}\right). \tag{5.11}$$

5.1 Concentration for Nonlinear Random Dynamical Systems: The Case of Harris Chains

The case of nonlinear random dynamical systems suffers from a lack of uniform transport-entropy constant related to a contractive metric. However, if we can show that the invariant measure satisfies transport-entropy inequality and Markov chain converges exponentially fast to its stationary distribution: it is plausible to run independent simulations of Markov chain and sample the averages after some burn-in period.

Since, the results developed in this paper are aimed at facilitating the RL and controls community. In the absence of exact dynamics, we develop easily verifiable/realistic conditions that ensures exponential integrability of invariant measure. This brings us to the weighted transportaion-inequalities introduced in [49] and plays an integral role when studying concentration phenomenon for nonlinear random dynamical systems. Their work allows for adding different weigths to the underlying distance function, precisely said:

Lemma 8. *Let ϕ be a non-negative integrable function, such that $\int e^{\phi(x)^2}\mu_\pi(dx) < \infty$- we get an upper bound on weighted total variation distance*

$$\|\phi(\mu_\pi - \nu)\|_{TV} \leqslant \sqrt{2}\left(1 + \log\int e^{\phi(x)^2}\mu_\pi(dx)\right)^{\frac{1}{2}}\sqrt{Ent(\nu||\mu_\pi)}. \tag{5.12}$$

A remarkable advantage of this formulation is an Lyapunov condition for underlying Markov chain so as to ensure that at least its' stationary measure satisfies T-E inequality.

Exponential Lyapunov function. Inspired by the assumption made in (particular case 14 of [49]), we proposed an exponential Lyapunov condition:

1. There exists $\hat{\alpha} > 0$, $\beta > 0$ and $C > 0$ such that $\beta < \hat{\alpha}$ and:

$$\int e^{\hat{\alpha}\|y\|^2} P(x, dy) \leqslant C e^{\beta\|x\|^2}, \quad \text{for all} \quad x \in \mathcal{X}.$$

Theorem 33. *If the exponential Lyapunov condition is satisfied, define* $W_{\hat{\alpha}}(x) := e^{\hat{\alpha}\|x\|^2}$*, then* $n-$ *th step transition kernel* $P^n(x, \cdot)$ *satisfy the following transport entropy inequality:*

$$\mathcal{W}_d(P^n(x, \cdot), \nu) \leqslant \sqrt{2}\left(\frac{1 + \log P^n W_{\hat{\alpha}}(x)}{\hat{\alpha}}\right)^{\frac{1}{2}} \sqrt{Ent(\nu||P^n(x, \cdot))} \tag{5.13}$$

Moreover, if an ergodic invariant measure μ_π *exists: then exists also a finite positive constant* $L_{\hat{\alpha},\beta,C}$ *such that:*

$$\mathcal{W}_d(\mu_\pi, \nu) \leqslant \sqrt{2L_{\hat{\alpha},\beta,C} Ent(\nu||\mu_\pi)} \tag{5.14}$$

Proof. We will only prove the result for the invariant measure as the result for $n-$ th step will follow the same argument. Since the condition in hypothesis can also be written as:

$$\int e^{\hat{\alpha}\|y\|^2} P(x, dy) \leqslant C e^{(\beta-\hat{\alpha})\|x\|^2} e^{\hat{\alpha}\|x\|^2} \tag{5.15}$$

As $(\beta - \hat{\alpha}) < 0$, we can find $\eta_{\hat{\alpha}} \in (0,1)$ and $\hat{C}_{\hat{\alpha}} < \infty$ such that $W_{\hat{\alpha}}(x)$ satisfy:

$$PW_{\hat{\alpha}}(x) \leqslant \eta_{\hat{\alpha}} W_{\hat{\alpha}}(x) + \hat{C}_{\hat{\alpha}}, \text{ and consequently via recursion } \int e^{\hat{\alpha}\|x\|^2} \mu_\pi(dx) \leqslant \frac{\hat{C}_{\hat{\alpha}}}{1 - \eta_{\hat{\alpha}}}. \tag{5.16}$$

and by defining $\phi(x) = \sqrt{\hat{\alpha}}\|x\|$, upper bound on weighted total variation from Lemma 8 implies that :

$$\|\phi(\mu_\pi - \nu)\|_{TV} \leqslant \sqrt{2}\Big(1 + \log\Big[\frac{\hat{C}_{\hat{\alpha}}}{1 - \eta_{\hat{\alpha}}}\Big]\Big)^{\frac{1}{2}} \sqrt{Ent(\nu||\mu_\pi)}. \tag{5.17}$$

Since, Wasserstein distance is upper bounded by weighted total-variation with weight $\|x\|$, after scaling we conclude that $\mu_\pi \in \mathcal{T}_1^d\left(\frac{1+\log\left[\frac{\hat{C}_{\hat{\alpha}}}{1-\eta_{\hat{\alpha}}}\right]}{\hat{\alpha}}\right)$ and the result for n-th step transition kernel follows via same argument. □

Harris chains. As the notion of running multiple independent trajectories is plausible when burn-in period is negligible : $n-$th step transition kernel of Markov chain converges exponentially fast to an invariant measure in Wasserstein metric, this brings us to Harris ergodic Markov chains that by definition satisfy following conditions:

Lyapunov condition with geometric drift: There exists a Lyapunov function $V : \mathcal{X} \to [0, +\infty)$, which satisfies:

$$PV(x) \leqslant \hat{\gamma} V(x) + K, \quad \text{for some } \hat{\gamma} \in (0,1)\ ,\ K < \infty \text{ and} \tag{5.18}$$

minorization condition: A sufficiently large level set of V (ironically it is called 'small set'), satisfies the minorization condition: i.e., there exists a set $\mathcal{S} := \{x \in \mathcal{X} : V(x) \leqslant R\}$ for some $R > \frac{2K}{1-\hat{\gamma}}$, $\beta \in (0,1)$ and $\hat{\nu} \in \mathcal{P}(\mathcal{X})$ such that:

$$\mathcal{P}(x,\cdot) \geqslant \beta\chi_{\mathcal{S}}(x)\hat{\nu}(\cdot). \tag{5.19}$$

Under these conditions it was shown by [48] that for some $\beta^* > 0$HEMC is contractive in Wasserstein metric $\mathcal{W}_d$ with distance function: $d(x,y) := (2 + \beta^* V(x) + \beta^* V(y))\chi_{x\neq y}$, . A unique ergodic invariant measure μ_π exists and for some finite C and $\kappa \in (0,1)$

$$\mathcal{W}_d\big(P^n(x,\cdot),\mu_\pi\big) \leqslant C\kappa^n \mathcal{W}_d(P(x,\cdot),\mu_\pi). \tag{5.20}$$

5.1.1 Application to Concentration for SLDSs

Model specifications. We consider a discrete-time SLDS of the form

$$x_{t+1} = \sum\nolimits_{j=1}^{M} (A_j x_t + w_t^j)\chi_{\mathcal{M}_j}(x_t). \tag{5.21}$$

Here, $x_t \in \mathbb{R}^n$ denote the system's state and $A_j \in \mathbb{R}^{n\times n}$ for $j = 1, ..., M$ capture system dynamics in each of the M Borel measurable regions that decompose the state-space and are pairwise disjoint satisfying $\bigcup_{j=1}^{M} \mathcal{M}_j = \mathbb{R}^n$. In addition, for a fixed region j, noise vectors w_t^j are i.i.d, and satisfy $w_t^j \sim \mathcal{N}(0, I_n)$ and $Cov(w_t^j, w_s^k) = 0$, for all $t, s \geqslant 0$ and $j \neq k \in \{1, 2, ..., M\}$.

Lemma 9. *Assume that there exists $\varrho < \infty$ such that for all $l \in \mathcal{K}_{bdd} := \{k \mid (1 \leqslant k \leqslant M)$ such that $\mathcal{M}_k \subsetneq \mathcal{B}_\varrho^n\}$,it holds that $\|A_l\|_2 \leqslant L < \infty$ and $\forall j \in (\mathcal{K}_{bdd})^{\complement}$, $\|A_j\|_2 \leqslant \gamma < 1$. Then, the system (5.21) mixes geometrically to a unique ergodic invariant distribution μ_π.*

Proof. Consider function $V(x) = \|x\|_2$. From (5.21), we have transition kernel $\sum_{j=1}^{M} P_j(x,\mathcal{A})\chi_{\mathcal{M}_j}(x)$, where $P_j(x,\cdot) \sim \mathcal{N}(A_j x, I_n)$. Assuming the initial state $x_0 :=$

$x \in \mathcal{M}_k$ for some $k \in \mathcal{K}_{bdd}$, then:

$$PV^2(x) = \mathbb{E}_{y\sim\mathcal{N}(A_k x, I_n)}\|y\|_2^2 = \mathbb{E}_{z\sim\mathcal{N}(0,I_n)}\|z\|_2^2 + \|A_k x\|_2^2 \leqslant (n + L^2\varrho^2). \tag{5.22}$$

However, if the initial state is $x_0 := x \in \mathcal{M}_j$ such that $j \in (\mathcal{K}_{bdd})^{\complement}$, then:

$$PV^2(x) = \mathbb{E}_{y\sim\mathcal{N}(A_j x, I_n)}\|y\|_2^2 = \mathbb{E}_{z\sim\mathcal{N}(0,I_n)}\|z\|_2^2 + \|A_j x\|_2^2 \leqslant n + \gamma^2\|x\|_2^2 = (n + \gamma^2 V^2(x)). \tag{5.23}$$

Therefore, starting from any initial condition in $\mathbb{R}^n$, from (5.22) and (5.23) it holds that $PV^2(x) \leqslant \gamma^2 V^2(x) + (n + L^2\varrho^2)$ and a trivial application of Jensen inequality reveals

$$PV(x) \leqslant \gamma V(x) + \underbrace{\sqrt{n + L^2\varrho^2}}_{K}. \tag{5.24}$$

Minorization condition can be verified from [45] and the result follows. □

Theorem 34. *For any $\hat{\alpha} \in (0, \frac{1-\gamma^2}{2})$, we have $\int e^{\hat{\alpha}\|x\|^2}\mu_\pi(dx) < \infty$ and consequently there exists a finite positive constant L_γ such that the invariant measure of SLDS, $\mu_\pi \in \mathcal{T}_1^d\big(L_\gamma\big)$.*

Proof. An application of Stein's lemma on transition kernel of (5.21), reveal:

$$\int e^{\alpha\|y\|^2}P(x,dy) = \frac{1}{(1-2\alpha)^{\frac{n}{2}}}e^{\|A_j x\|^2(\alpha + \frac{2\alpha^2}{(1-2\alpha)})} = \frac{1}{(1-2\alpha)^{\frac{n}{2}}}e^{\|A_j x\|^2\frac{\alpha}{(1-2\alpha)}}, \alpha < \frac{1}{2}, x \in \mathcal{M}_j$$

A simple linear algebra exercise reveals existence of a $\beta < \hat{\alpha}$ and C when $\hat{\alpha} \in (0, \frac{1-\gamma^2}{2})$ as mentioned in Theorem 33 and conclusion follows. □

5.1.2 Gaussian Tail Inequality for Stationary Distribution of SLDS/Harris Chain with Exponential Lyapunov Function and its Consequences for Sampling

Assume that we have access to sampling $(y_i)_{i=1}^N$ i.i.d from μ_π. Since, $r(x) := \|x\|$ is $1-$ Lipschitz w.r.t $d(x,y) := \|x-y\|_2$ and $\mu_\pi \in \mathcal{T}_1^d\left(\frac{1+\log\left[\frac{\hat{C}_{\hat{\alpha}}}{1-\eta_{\hat{\alpha}}}\right]}{\hat{\alpha}}\right)$, we have that

$$\mathbb{P}\left[\left|\frac{1}{N}\sum_{i=1}^N r(y_i) - < r >_{\mu_\pi}\right| > \epsilon\right] \leqslant 2\exp\left(-\frac{N\epsilon^2\hat{\alpha}}{2+2\log\left[\frac{\hat{C}_{\hat{\alpha}}}{1-\eta_{\hat{\alpha}}}\right]}\right). \tag{5.25}$$

As any valid $\hat{\alpha}$ can be written down in the form of $\frac{1-\gamma^2}{2n}$ for $n > 1$, comparing with Linear Gaussian case (5.11) it is reassuring to see how deviations for SLDS and LDS have similar dependence in terms of the norm of stable system matrix.

Remark 35. *Although, we tried our best to show concentration for process level law of HEMCs under exponential-type Lyapunov condition, via weigthed T-E inequality but it got intractable due to non-uniform transport constants.*

Definition 9. *Kernel representation of Markov operator: we say that a Markov transition operator P has a kernel representation w.r.t its' stationary measure μ_π if $Pf(x) = \int f(y)k(x,y)\mu_\pi(dy)$ and $\int k(x,y)\mu_\pi(dy) = \int k(x,y)\mu_\pi(dx) = 1$ a.e. x and y w.r.t μ_π, respectively. If $P(x,\cdot) << \mu_\pi$ almost all x w.r.t stationary measure, then for each fixed x, $k(x,\cdot) = \frac{dP_x}{d\mu_\pi}(\cdot)$, i.e. Radon-Nikodym derivative of $P(x,\cdot)$ w.r.t μ_π.*

5.1.3 Compact operator associated to minorization

Let's define an integral operator on $L^2(\mu_\pi)$:

$$Q_\pi f(x) := \int f(y)\beta\chi_{\mathcal{S}}(x)\frac{d\hat{\nu}}{d\mu_\pi}(y)\mu_\pi(dy), \tag{5.26}$$

which is compact if $\left\|\frac{d\hat{\nu}}{d\mu_\pi}\right\|_{L^2(\mu_\pi)} < \infty$, also:

Remark 36. *Minorization operator Q_π, when viewed as an element of $\mathcal{B}\big(L^2(\mu_\pi)\big)$ is bounded above by $\beta\left\|\frac{d\hat{\nu}}{d\mu_\pi}\right\|_2\sqrt{\mu_\pi(\mathcal{S})}$.*

Proof. Given any $f \in L^2(\mu_\pi)$:

$$\begin{aligned}\sqrt{\langle Q_\pi f, Q_\pi f\rangle} &= \sqrt{\big(\hat{\nu}(f)\big)^2\beta^2\mu_\pi(\mathcal{S})} \\ &\leqslant \beta\left\|\frac{d\hat{\nu}}{d\mu_\pi}\right\|_2\sqrt{\mu_\pi(\mathcal{S})}\|f\|_2, \end{aligned} \tag{5.27}$$

where the last inequality follows from Cauchy-Schwarz. □

Remark 37. *Owing to time and space limitations we will not explore the consequences of preceding result in this paper, but we invite other researchers to do so by pointing out to them following interpretation: loosely speaking, we can view P as sum of a compact operator with a kernel (in the sense $\hat{\nu} << \mu_\pi$) and some added perturbation 'S' i.e., $P = S + Q_\pi$. One possible way to prove spectral gap for P is by first ensuring spectral radius of $Q_\pi > 0$ and if the norm of perturbation $\|S\|$ is small enough (correct scaling can be found in [54]): spectral gap for P follows. [55], uses a similar idea but with Fredholm type perturbation analysis to show spectral gap under the assumption of* uniform integrability of P.

Exponential-type Lyapunov function implies a lower bound on $L^1 - L^q$ Hyperboundedness, for some $q > 1$

Theorem 38. *Assume there exists $\hat{\alpha}$ and β such that $\beta < \hat{\alpha}$ and $\int e^{\hat{\alpha}\|y\|^2}\mathcal{P}(x, dy) \leqslant C(\hat{\alpha})e^{\beta\|x\|^2}$. Then for some $q > 1$, $\|P\|_{L^1\to L^q} \geqslant C(\hat{\alpha})$.*

Proof. We know from theorem 34, $\|W_{\hat{\alpha}}\|_{L^1(\mu_\pi)} < \infty$, w.l.o.g assume that for any $\eta > \hat{\alpha}$, W_η is not integrable: i.e., $W_{\hat{\alpha}}$ doesn't belong to any L^p for $p > 1$. Now notice that

$PW_{\hat{\alpha}} \leqslant C(\hat{\alpha})\big(W_{\hat{\alpha}}\big)^{\frac{\beta}{\hat{\alpha}}}$. So there exists a $q > 1$, such that $\big(PW_{\hat{\alpha}}\big)^q \leqslant C(\hat{\alpha})^q W_{\hat{\alpha}}$ and eventually we have: $\|PW_{\hat{\alpha}}\|_{L^q(\mu_\pi)} \leqslant C(\hat{\alpha})\|W_{\hat{\alpha}}\|_{L^1(\mu_\pi)}$ □

5.2 Spectral Gaps, Ergodic Theorems and Poincare Inequality

To mathematically express the phenomenon that *time-distant samples, although temporally dependent, behave as iid* , we will have to go to fundamentals of ergodic theory. Let us use $\mathbb{D}$ to denote the unit disc in the complex plane and $\mathbb{T}$ represents the unit circle in the plane. Consider the Banach space, with complex field, $L^p(\mu_\pi)$ of $p \in [1, \infty]$ integrable functions; i.e., all measurable functions f with $\int |f(x)|^p \mu_\pi(dx) < \infty$. Conjugate power of p, $c(p)$ is defined as $c(p) \geqslant 1$ that satisfies $\frac{1}{c(p)} + \frac{1}{p} = 1$. Every $g \in L^{c(p)}$ corresponds to a bounded linear functional on $f \in L^p$ and its action is captured by $\langle f, g\rangle_{\mu_\pi} := \int f(x)\overline{g(x)}d\mu_\pi(x)$.

We consider linearity in first argument of the inner product and the boundedness follows by Cauchy-Schwarz: $|\langle f, g\rangle_{\mu_\pi}| \leqslant \|f\|_p \|g\|_{c(p)}$. Markov transition operator $P : L^p \to L^p$ is defined as $\|Pf\|_{L^p} = \left(\int |Pf(x)|^p \mu_\pi(dx)\right)^{\frac{1}{p}} = \left(\int |\int f(y)p(x, dy)|^p \mu_\pi(dx)\right)^{\frac{1}{p}}$ which is $\leqslant \|f\|_{L^p}$, where the last inequality follows from Jensen inequality. Note that P acts as an identity on constant functions, i.e., $P1 = 1$, and it is positive, i.e., $Pf \geqslant 0$ if $f \geqslant 0$.

A Markov chain is *reversible or satisfies detailed balance condition* if Markov transition operator P when viewed as an operator on Hilbert space $L^2(\mu_\pi)$ is equal to its adjoint P^*: $\langle Pf, g\rangle := \langle f, P^*g\rangle_{\mu_\pi} = \langle f, Pg\rangle$. A simple observation reveals that P, P^*, PP^* and P^*P are all Markov operators with μ_π as invariant measure. Consider a sequence $(A_n)_{n\in\mathbb{N}}$ of bounded operators on some Banach space, there are different topologies (uniform, strong, and weak) to study its convergence (see e.g., Chapter

6 of [16] for exact definitions). For the reader unfamiliar with spectral properties of operators and associated notations (particularly Hyperboundedness and Fredholm theory):

Spectral Decomposition

Preliminaries. Let A be a bounded operator on a Banach space $\mathcal{X}$, denoted by $A \in \mathcal{B}(\mathcal{X})$. $\lambda \in \mathbb{C}$ is said to be in resolvent of A, denoted by $\lambda \in RS(A)$, if $\lambda I - A : \mathcal{X} \to \mathcal{X}$ is bijective. Bounded inverse theorem implies that for $\lambda \in RS(A)$, $R_\lambda(A) := (\lambda I - A)^{-1} \in \mathcal{B}(\mathcal{X})$. The spectrum of A is denoted by $\sigma(A) := \mathbb{C} \backslash RS(A)$. Spectral radius of A is denoted by $\rho(A) := \sup |\lambda| : \lambda \in \mathbb{C}$. Consequently, if there exists a $v \in \mathcal{X}$ and $\lambda \in \mathbb{C}$ such that $Av = \lambda v$ (i.e., λ is an eigenvalue with the corresponding eigenfunction v) then $\lambda I - A$ is not injective, which implies $\lambda \in \sigma(A)$. However, the spectrum of A is not limited to eigenvalues; for a detailed monograph of spectral theory see e.g., [16].

Definition 10. *Let A be a bounded operator on some Banach space $\mathcal{X}$. It is called* Fredholm *if: (a) $dim[N(A)] < \infty$, (b) $dim[\mathcal{X} \backslash Im(A)] < \infty$, and (c) $Im(A)$ is closed. Here, $dim[N(A)]$ denotes the dimension of the null space of A and $Im(A)$ means the range of A; $dim[\mathcal{X} \backslash Im(A)]$ is often read as the dimension of* cokernel *of A.*

Note that the condition (c) is redundant; it follows from (b).

Definition 11. *The index of a Fredholm operator A is defined as $ind(A) = dim[N(A)] - dim[X \backslash Im(A)]$, and for some small perturbations $\Delta(A)$ to the operator $ind(A + \Delta(A)) = ind(A)$.*

Definition 12. *$\lambda \in \sigma(A)$ is said to be in essential spectrum of A, ($\lambda \in \sigma_{ess}(A)$) if and only if $\lambda I - A$ is not a Fredholm operator. $\lambda \in \sigma(A) \backslash \sigma_{ess}(A)$ is said to be in*

discrete spectrum, ($\lambda \in \sigma_{disc}(A)$) *if and only if (a)* λ *is an isolated point of* $\sigma(A)$ *and (b)* $\{\psi \in \mathcal{X} : A\psi = \lambda\psi\}$ *is finite dimensional.*

Definition 13. *Markov operator is called* hyperbounded*, if for some* $1 \leqslant p < q \leqslant \infty$*, P is a bounded operator from* $L^p(\mu_\pi)$ *to* $L^q(\mu_\pi)$ *(we will often call it* $L^p - L^q$ *hyperboundedness). More stringent requirement of* hypercontractivity *requires P to be hyperbounded with* $\|P\|_{L^p \to L^q} \leqslant 1$*.*

Remark 39. *It follows from Jensen's inequality that for all* $r \geqslant 1$, $\|P\|_{L^r(\mu_\pi)} \leqslant 1$*, and Riesz-Thorin interpolation implies that for all* $1 < p < q$*, P is* $L^p - L^q$ *hyperbounded.*

5.2.1 Ergodic Theorems and Consequences

Definition 14. *In functional analytic framework, the pair* (P, μ_π) *is said to be ergodic if for any* $f \in L^\infty(\mu_\pi)$ *satisfies* $Pf = f$*, then f is a constant.*

In probabilistic language, an ergodic Markov chain has a unique invariant stationary distribution.

Definition 15. *The pair* (P, μ_π) *is called aperiodic if for all* $\lambda \in \mathbb{T}\backslash\{1\}$, $dim[N(\lambda I - P)] = 0$*.*

Theorem 40. *Birkhoff Pointwise Ergodic theorem: Let* $(x_n)_{n\in\mathbb{N}}$*, be the samples from an ergodic Markov chain. Then,* $\frac{1}{N}\sum_{n=0}^{N-1} f(x_n) \longrightarrow \mu_\pi(f)$*, almost every initial condition* $x_0 = x$ *w.r.t* μ_π*, and f integrable w.r.t* μ_π*.*

This is reminiscent of the strong law of large numbers for iid sampling from distribution μ_π. However, a very natural requirement for having a concentration similar to iid setting is sharp decay of correlation – i.e., for some $C < \infty$ and $\eta \in (0, 1)$

$$|Cov_{\mu_\pi}[f(x_n), f(x_{n+m})]| \leqslant C\eta^m Var_{\mu_\pi}(f), \quad \forall f \in L^2(\mu_\pi). \tag{5.28}$$

Now, the first step towards formalizing the preceding phenomenon is a concept related to *uniform ergodicity*.

Theorem 41. *If the pair (P, μ_π) is ergodic, let U be the orthogonal projection on $\{f \in L^2(\mu_\pi) : Pf = f\}$. Then,*

$$\|\frac{1}{N}\sum_{n=0}^{N-1} P^n - U\|_{L^2(\mu_\pi)} \longrightarrow 0; \tag{5.29}$$

i.e., the operator, $\frac{1}{N}\sum_{n=0}^{N-1} P^n$ converges to a bounded linear projection U in the uniform operator topology.

Proof. If the pair (P, μ_π) is ergodic then so is the pair (P^*, μ_π). We first show that $Im(I-P)$ is closed using *Fredholm argument* – because the dimension of *co-kernel* is finite: $dim[L^2(\mu_\pi) \backslash Im(I-P)] = dim[N(I-P^*)] = dim[N(I-P)] = 1$. Now, consider the following decomposition:

$$(I-P) : N(I-P) \bigoplus N(I-P)^\perp \to Im(I-P) \bigoplus Im(I-P)^\perp. \tag{5.30}$$

Therefore, $I-P : N(I-P)^\perp \to Im(I-P)$ is bijective and by inverse mapping theorem $(I-P)^{-1} \in \mathcal{B}(Im(I-P), N(I-P)^\perp)$ so there exists a $K < \infty$ such that $\|(I-P)^{-1}\|_{Im(I-P)\to N(I-P)^\perp} \leqslant K$ implying that the pair (P, μ_π) is *uniformly ergodic* because: given any $f \in Im(I-P)$ there exists a unique $\hat{g} \in N(I-P)^\perp$ such that $f = (I-P)\hat{g}$ and $\|\frac{1}{N}\sum_{n=0}^{N-1} P^n f\| = \|\frac{1}{N}\sum_{n=0}^{N-1} P^n (I-P)\hat{g}\| = \frac{\|(I-P^N)\hat{g}\|}{N} = \frac{\|(I-P^N)(I-P)^{-1}f\|}{N} \leqslant 2\frac{\|(I-P)^{-1}f\|}{N} \leqslant \frac{2K}{N}\|f\|$. So $\frac{1}{N}\sum_{n=0}^{N-1} P^n$ converges to 0 at a uniform rate on $Im(I-P)$. Notice that, when $Im(I-P)$ is closed: $Im(I-P) = N(I-P)^\perp$, because regardless of $Im(I-P)$ being closed : $Im(I-P) \subset N(I-P^*)^\perp$ and $Im(I-P)^\perp = N(I-P^*)$. Therefore, $Im(I-P)^\perp$ corresponds to $N(I-P)$ which by ergodicity assumption only comprises of constant function and $\frac{1}{N}\sum_{n=0}^{N-1} P^n_{|_{N(I-P)}} = I$. □

Remark 42. *The above theorem effectively provide a stronger version of the mean ergodic theorem – i.e., 'L^2- uniform ergodicity'.*

Remark 43. *It is worth noting the following:*

1. *Under ergodicity and reversibility assumption, $I-P$ is an* Fredholm operator *with index 0; effectively, it means* 1 *is an isolated eigenvalue of P – i.e., for some $\epsilon > 0$, $\sigma(P) = [-1, 1-\epsilon] \cup \{1\}$ and the dimension of space of eigenfunctions associated to* 1*, is* 1 *.*

2. *We define $S := P_{|Im(I-P)}$ (when $Im(I-P)$ is closed), and since $Im(I-P)$ is invariant under P, implying $S : Im(I-P) \to Im(I-P)$. Because $(I-P)^{-1} \in \mathcal{B}(Im(I-P))$, $(I-S)^{-1}$ exists as well so $1 \notin \sigma(S)$ and from previous bullet point as S will contain subset of spectrum of P, indeed $\sigma(S) \subset [-1, 1-\epsilon]$.*

Theorem 44. *If the pair (P, μ_π) is reversible, ergodic and aperiodic, then $\|P^n - U\|_{L^2(\mu_\pi)} \longrightarrow 0$, in uniform operator topology.*

Proof. Consider the operator $I+P$, as a map

$$I + P : N(I+P) \bigoplus N(I+P)^\perp \to Im(I+P) \bigoplus Im(I+P)^\perp. \tag{5.31}$$

$N(I+P) = \{f \in L^2(\mu_\pi) : Pf = -f\}$, which corresponds to *periodic* behavior but by hypothesis $dim[N(I+P)] = 0$ and $dim[L^2(\mu_\pi) \backslash Im(I+P)] = dim[N(I+P)] = 0$ (thanks to reversibility) so $I+P$ is a Fredholm operator implying that $-1 \notin \sigma_{ess}(P)$ (see e.g., [40]). Moreover, $-1 \notin \sigma(P)$ either. Since the spectrum of a bounded operator is always closed, there exists $\delta > 0$ such that $\sigma(S) \subset [-1+\delta, 1-\epsilon]$. Consequently $\rho(S)$, the spectral radius of S, satisfies $\rho(S) \leqslant max(|-1+\delta|, |1-\epsilon|) < 1$ and $P^n = S^n \bigoplus I_{N(I-P)}$, but $\rho(S) < 1$ implies $S^n \to 0$, because by Gelfand's formula for any $\rho \in (\rho(S), 1)$ there exists $M(\rho) < \infty$ such that $\|S^n\| \leqslant M(\rho)\rho^n$. □

Spectral analysis of non-reversible P is more involved as it can lies inside a unit disc and this is where *hyperboundedness* comes into play.

Theorem 45. *Hyperboundedness(see definition 13 in Appendix) of $L^2 - L^p$, where $p = 2 + \epsilon$ and $\epsilon > 0$ implies that (a) for all $\lambda \in \mathbb{T}$, $dim[N(\lambda I - P)_2] < \infty$ and (b) $dim[N(\lambda I - P^*)_2] < \infty$. (b) independently implies $Im(\lambda I - P)_2$ is closed. Consequently:*

1. *$\sigma(P) \cap \mathbb{T}$ may only comprise of a finite number of distinct eigenvalues and each distinct eigenvalue has a finite-dimensional eigenspan.*

2. *$\sigma_{ess}(P)$ is contained inside some closed disc of radius $\alpha < 1$ in the complex plane.*

3. *$\sigma(P)$ is 'gapped' – i.e., it comprises of two disjoint sets: $\sigma(P) = \{\sigma(P) \cap \mathbb{T}\} \cup \{\sigma(P) \cap \alpha\mathbb{D}\}$.*

Proof. Our proof is based on a result of [56]: infinite dimensional L^p spaces are not isomorphic for $p \neq q$. If $f \in N[\lambda I - P]$, for $\lambda \in \mathbb{T}$, it must satisfy $|Pf(x)| = |f(x)|$. P is $L^2 - L^{p:=2+\epsilon}$ hyperbounded, for some $\epsilon > 0$. Then duality implies that P^* is $L^{c(p)} - L^2$ hyperbounded, where $c(p)$ is the conjugate power of p. Now recall, $P \in \mathcal{B}(L^p)$ and $dim[L^p \backslash Im(\lambda I - P)_p] = dim[N(\overline{\lambda} I - P^*)_{c(p)}]$. If $dim[N(\overline{\lambda} I - P^*)_{c(p)}] = \infty$, strict inclusion implies there exists a $\hat{g} \in L^{c(p)} \backslash L^2$ such that $P^* \hat{g} = \overline{\lambda} \hat{g}$ and $|P^* \hat{g}| = |\hat{g}|$ *which contradicts hyperboundedness.* Therefore, $dim[L^p \backslash Im(\lambda I - P)_p]$ is finite which implies $Im(\lambda I - P)$ is closed in $L^p$and as $L^2 \subsetneq L^{c(p)}$ we have $dim[L^2 \backslash Im(\lambda I - P)_2] = dim[N(\overline{\lambda} I - P^*)_2] < dim[N(\overline{\lambda} I - P^*)_{c(p)}] < \infty$ and we have that $Im(\lambda I - P)$ is closed when viewed as a map in L^2. Argument for $dim[N(\lambda I - P)_2]$ being finite follows the same hyperboundedness argument. Consequences follow from the disjoint nature of essential and discrete spectrum (see Theorem 7.9-7.11 of [16]). □

Theorem 46. *If the pair (P, μ_π) is ergodic, $L^2 - L^p$ hyperbounded for $p = 3, 4$ with norm condition $\|P\|_{L^2-L^p} < 2^{\frac{1}{2}-\frac{1}{p}}$, then*

$$\|P^n - U\|_{L^2(\mu_\pi)} \longrightarrow 0, \text{ in a uniform operator topology.} \tag{5.32}$$

Proof. The norm condition $\|P\|_{L^2\to L^4} < 2^{\frac{1}{2}-\frac{1}{4}}$ or $\|P\|_{L^2\to L^3} < 2^{\frac{1}{2}-\frac{1}{3}}$ ensures aperiodicity (see Theorem 5.1 and 5.3 in [57]); only $\{1\} \in \sigma(P) \cap \mathbb{T}$ and corresponds to the eigenspan of constant functions. Following the argument of Theorem 44: $P^n = S^n \bigoplus I_{N(I-P)}$, where $S = P_{|Im(I-P)}$ and $\sigma(S) \subset \alpha\mathbb{D}$ for some $\alpha \in (0, 1)$. Consequently, $\rho(S) \leqslant \alpha$, and for all $\rho \in (\alpha, 1)$ there exists an $M(\rho) < \infty$ such that $\|S^n\| \leqslant M(\rho)\rho^n$; and thus, the result follows. □

Remark 47. *A trivial corollary of $\|P^n - U\|_{L^2(\mu_\pi)} \longrightarrow 0$ in uniform operator topology is that for some $\rho \in (0, 1)$, $C < \infty$ and for all $n \in \mathbb{N}$, it holds that*

$$\|P^n\big(f - \mu_\pi(f)\big)\|_{L^2} \leqslant C\rho^n \|f - \mu_\pi(f)\|_{L^2}, \tag{5.33}$$

which is equivalent to sharp decay of correlation in (5.46).

5.3 Hyperboundedness and Transport-Entropy Inequality Implies Concentration

After ensuring uncorrelation for time-distant samples, it is safe to relate to the iid setting and remind ourselves that *sharp concentrations heavily rely on the ability to take exponential moments of the observable w.r.t underlying measure.* An attempt to analyse the uncorrelation and exponential integrability via a single linear operator brings us to the following discussion.

Feynman-Kac semigroup plays an integral role in the study of fluctuations of time additive quantities of diffusion process in continuous time, [58]. The reader is referred to [46] for a detailed exposition on this topic in discrete-time setting. We identify with the observable/reward function r, a Feynman-Kac semigroup, which is a composition of the Markov transition operator followed by an exponentiated multiplication operator w.r.t observable – i.e., $e^r P : D_{(\cdot)} \subseteq L^2(\mu_\pi) \rightarrow L^2(\mu_\pi)$, such that for all $g \in D_{(\cdot)}$ (the domain on which semigroup operators can be viewed as bounded operator) it holds

$$(e^r P)^n g(x) := \mathbb{E}_x[g(x_N) e^{\sum_{i=0}^{n-1} r(x_i)}] \ , \forall n \in \mathbb{N}. \tag{5.34}$$

Assume that $x_0 \sim \beta$ and $\beta << \mu_\pi$. We have the following upper bound on the deviation of the observable:

$$\mathbb{P}_\beta\left(\frac{1}{N}\sum_{i=0}^{N-1} r(x_i) - \mu_\pi(r) \geqslant \epsilon\right) \leqslant \left\|\frac{d\beta}{d\mu_\pi}\right\|_2 \inf_{s>0} \|(e^{sr}P)^N\|_{L^2-L^2} e^{-sN(\mu_\pi(r)+\epsilon)}. \tag{5.35}$$

Thus, the task at hand is bounding the operator $e^r P$ in a meaningful way to get a favorable concentration result. P in itself is a positive contraction but exponentiated multiplication operator defined by unbounded observable require a short detour into its spectral analysis and transport-entropy inequality.

Multiplication operator defined by e^r. The operator $\big(e^r, D(e^r)_p\big)$ is closed and densely defined, see e.g., Proposition 3.10 from Chapter 1 in [59]. The essential range of e^r with μ_π as an underlying measure is defined as $e^r_{ess}(\mu_\pi) := \Big\{\lambda \in [1,\infty) : \mu_\pi\big(|e^r - \lambda| < \epsilon\big) \neq 0, \forall\, \epsilon > 0\Big\}$, and the essential norm of e^r, $\|e^r\|_\infty := \sup\{\lambda \in e^r_{ess}(\mu_\pi)\}$. Multiplication operator e^r is bounded in some and hence all L^p, iff $\|e^r\|_\infty < \infty$. Consequently, $D(e^r)(p) = L^p$ and $\|e^r\|_{L^p \rightarrow L^p} = \|e^r\|_\infty$. If the stationary measure μ_π is

not compactly supported and is absolutely continuous w.r.t Lebesgue measure then $\|e^r\|_{L^p \to L^p} = \infty$. However, not all is lost, as suggested in the previous section that we need some notion of hyperboundedness for uncorrelation. So if for some $p > 2$, $P : L^2 \to L^p$, in order to ensure $e^r P \in \mathcal{B}(L^2)$ it is sufficient to prove $e^r : L^p \to L^2$. This is where the *transport-entropy inequality* comes into play.

Definition 16. *Consider metric space $(\mathcal{X}, d)$ and reference probability measure $\mu \in \mathcal{P}(\mathcal{X})$. Then we say that μ satisfies* Transport-Entropy *(T-E) inequlaity with constant C or to be concise $\mu \in \mathcal{T}_1^d(C)$ for some $C > 0$ if for all $\nu \in \mathcal{P}(\mathcal{X})$ and $\nu << \mu$, it holds that*

$$\mathcal{W}_d(\mu, \nu) \leqslant \sqrt{2CEnt(\nu||\mu)}. \tag{5.36}$$

Lemma 10 ([60])**.** *μ satisfies $\mathcal{T}_1^d(C)$ if and only if for all Lipschitz function f with $< f >_\mu := \mathbb{E}_\mu f$, it holds that*

$$\int e^{\lambda(f - < f >_\mu)} d\mu \leqslant \exp(\frac{\lambda^2}{2} C \|f\|^2_{L(d)}), \quad \textit{where} \quad \|f\|_{L(d)} := \sup_{x \neq y} \frac{|f(x) - f(y)|}{d(x, y)}. \tag{5.37}$$

Theorem 48. *If $\mu_\pi \in \mathcal{T}_1^d(C_\pi)$ for some $C_\pi > 0$ and r is Lipschitz w.r.t metric d then we have for all $p > 2$, $e^r : L^p \longrightarrow L^2$ is a bounded operator with norm:*

$$\|e^r\|_{L^p \to L^2} \leqslant \exp\left(\mu_\pi(r) + \frac{2pC_\pi \|r\|^2_{L(d)}}{(p-2)2}\right). \tag{5.38}$$

Proof. The proof is a simple application of Cauchy-Schwarz and the exponential moment inequality for Lipschitz functions under distribution μ_π satisfying the transport-entropy inequality. □

Theorem 49. *Without any assumption of reversibility, if the invariant measure $\mu_\pi \in \mathcal{T}_1^d(C_\pi)$ for some $C_\pi > 0$ and r is Lipschitz w.r.t metric d and for some $q > 2$, $P : L^2(\mu_\pi) \longrightarrow L^q(\mu_\pi)$ is hyperbounded with norm $\|P\|_{L^2 \to L^q} < e^{\frac{1}{2}\left[\frac{1}{2} - \frac{1}{q}\right]}$. Given*

$n \in \mathbb{Z}^+$ and $\delta \in (0,1)$ and an initial distribution $\beta << \mu_\pi$, if $N \geqslant \ln\left(\left\|\frac{d\beta}{\mu_\pi}\right\|_2 \frac{1}{1-\delta}\right)$ $\left(\frac{4n^2 q}{(q-2)-4n^2 q \ln \|P\|_{L^2 \to L^q}}\right)$, we have that

$$\mathbb{P}_\beta\left(\frac{1}{N}\sum_{i=0}^{N-1} r(x_i) - \mu_\pi(r) \geqslant \frac{\sqrt{C_\pi}\|r\|_{L(d)}}{n}\right) < 1-\delta. \tag{5.39}$$

Proof. Since the theorem assumptions ensure $e^r P \in \mathcal{B}(L^2(\mu_\pi))$, we have that $\|(e^r P)^N\|$ $\leqslant \|e^r P\|^N$. Let $\epsilon(n) := \frac{\sqrt{C_\pi}\|r\|_{L(d)}}{n}$, where $\sqrt{C_\pi}\|r\|_{L(d)}$ is proportional to the square root of variance of r under distribution μ_π. Notice that as opposed to standard (ϵ, δ) probability arguments, here we have scaled $\epsilon = \epsilon(n)$ with a control variable n that we can increase to make ϵ arbitrarily small. From the preceding discussion, it holds that

$$\|(e^{sr}P)^N\|_{L^2-L^2} e^{-sN(\mu_\pi(r)+\frac{\sqrt{C_\pi}\|r\|_{L(d)}}{n})} \leqslant \|P\|_{2-q}^N e^{N\left(\left(\frac{2q}{q-2}\right)\frac{s^2}{2}C_\pi\|r\|^2_{L(d)} - s\sqrt{C_\pi}\frac{\|r\|_{L(d)}}{n}\right)}, \tag{5.40}$$

and a trivial calculation reveals

$$\inf_{s>0} \|P\|_{2-q}^N e^{N\left(\left(\frac{2q}{q-2}\right)\frac{s^2}{2}C_\pi\|r\|^2_{L(d)} - s\sqrt{C_\pi}\frac{\|r\|_{L(d)}}{n}\right)} = e^{N\left(ln\|P\|_{2-q} - \frac{1}{2n^2}\left[\frac{q-2}{2q}\right]\right)}, \tag{5.41}$$

which we should be able to decrease as N increase; for every $n \in \mathbb{Z}^+$, which happens to be the case if $\|P\|_{L^2 \to L^q} < e^{\frac{1}{2}\left[\frac{1}{2}-\frac{1}{q}\right]}$ and all other results follow. □

Remark 50. *Since our norm control $\|P\|_{L^2 \to L^q} < e^{\frac{1}{2}\left[\frac{1}{2}-\frac{1}{q}\right]}$ is in harmony with the upper bound for $q = 3, 4$ given by [57], which implies implies aperiodicity and consequently Poincare' L^2- Spectral gap for Markov transition operator. We conjecture that for $q \in (2,3)$ our upper bound might imply aperiodicity and hence L^2- Spectral gap.*

Assuming hypercontractivity, a stronger theoretical result similar to [46] can be directly deduced.

Corollary 7. *Without any reversibility assumption on the pair (P, μ_π), if the stationary distribution satisfies T-E inequality i.e., $\mu_\pi \in \mathcal{T}_1^d(C_\pi)$ and for some $p > 2$ associated transition kernel is hypercontractive, i.e., $\|P\|_{2\to p} \leqslant 1$, then for any initial distribution $\beta << \mu_\pi$ and $N \in \mathbb{N}$ it holds*

$$\mathbb{P}_\beta\left(\frac{1}{N}\sum_{i=0}^{N-1} r(x_i) - \mu_\pi(r) \geqslant \epsilon\right) \leqslant \left\|\frac{d\beta}{d\mu_\pi}\right\|_2 \exp\left(-\frac{N\epsilon^2(p-2)}{4C_\pi\|r\|^2_{L(d)}p}\right). \tag{5.42}$$

Consequently, given $\epsilon > 0$ and $\delta \in (0,1)$, it holds that for all $N \geqslant \frac{\ln\left(\left\|\frac{d\beta}{\mu_\pi}\right\|_2 \frac{1}{1-\delta}\right) 4C_\pi\|r\|^2_{L(d)}p}{\epsilon^2(p-2)}$ *the chain satisfies* $\mathbb{P}_\beta\left(\frac{1}{N}\sum_{i=0}^{N-1} r(x_i) - \mu_\pi(r) \geqslant \epsilon\right) \leqslant 1-\delta$.

Decay of correlation and the 'dilemma of duality': $\|P^n - U\|_{L^2(\mu_\pi)} \longrightarrow 0$, in uniform operator topology, implies if $x_0 \sim \mu_\pi$ and x_n evolves according to the Markov transition operator P:realization of any $f \in L^2(\mu_\pi)$, $f(x_n)$ are asymptotically uncorrelated. In the *reversible* case, information on P is equivalent to information on time reversed chain represented by P^*. Therefore, *ergodicity and aperiodicity implies asymptotic uncorellation.* In *non-reversible* setting, practitioner doesn't have access to simulating P^*. So in order to ensure time distant samples are becoming uncorrelated, a prior knowledge of hyperboundedness of P with the discussed norm control makes up for reversibility assumption. *Does this ascertain the phillosophical argument: we cannot fully understand one side of the dual nature of something without understanding the opposing side ?* From an application point of view: *Can we simulate P^* by adding some modifications to the simulation of P ?*

Remark 51. *A trivial corollary of $\|P^n - U\|_{L^2(\mu_\pi)} \longrightarrow 0$, in uniform operator topology*

is that for some $\rho \in (0,1)$, $C < \infty$ *and for all* $n \in \mathbb{N}$*:*

$$\|P^n\big(f - \mu_\pi(f)\big)\|_{L^2} \leqslant C\rho^n \|f - \mu_\pi(f)\|_{L^2}. \tag{5.43}$$

This brings us to a more restrictive functional counterpart of preceding inequality:

5.3.1 Poincare inequality and its' converse

Correct functional formulation of *Poincare inequality* (see e.g., [61]) for discrete time Markov chain is as follows:

Definition 17. *The Dirichlet form* $\mathcal{E}_{PP*}$ *for a Markov semigroup* PP^**, w.r.t its' stationary distribution* μ_π *is a bilinear form* $\mathcal{E}_{PP*}(f,g) := \langle (I - P^*P)f, g\rangle_{L^2_{\mu_\pi}}$*. The Markov process is said to satisfy Poincare' inequality with constant* $C > 1$ *if:*

$$Var_{\mu_\pi}(f) \leqslant C\mathcal{E}_{PP*}(f,f). \tag{5.44}$$

Corollary 8. *A trivial consequence of* (5.44) *reveals, exponential decay of the variance of* $(P^n f)_{n\in\mathbb{N}}$ *with initial distribution taken as the staionary distribution, to be precise:*

$$Var_{\mu_\pi}(P^n f) \leqslant \left(\frac{C-1}{C}\right)^N Var_{\mu_\pi}(f), \quad \forall f \in \mathcal{L}^2(\mu_\pi). \tag{5.45}$$

and simple application of stationarity and Cauchy-Schwarz imply exponential decay of correlation:

$$|Cov_{\mu_\pi}[f(x_n), f(x_{n+m})]| \leqslant \left(\frac{C-1}{C}\right)^m Var_{\mu_\pi}(f), \quad \forall f \in \mathcal{L}^2(\mu_\pi). \tag{5.46}$$

Should be obvious from the mathematical formulation and its' consequences: Poincare inequality reveals the speed of convergence of a Markov chain to equilib-

rium and ensures applicability of Markov chain Monte Carlo methods like Metropolis-Hasting ; where one tries to simulate samples from a target distribution by instead generating a Markov kernel whose invariant distribution is the target distribution i.e., if Poincare' inequality holds, Markov chain converges exponentially fast to the stationary distribution(in total variation), correlation of samples decrease exponentially fast and one manages to get 'almost' i.i.d samples from the target distribution, for a detailed exposition on this topic see e.g., [62]. If the Markov chain lies inside a finite state space, [63] and [64] have shown to use conductance methods and graph theory to get finite dimensional spectral analysis of Markov transition kernel (matrix eigenvalue/eigenvector analysis in) which can then leads to Poincare' inequality. Poincare' inequality and spectral gap in L^2 are often used interchangeably, with origins in continuous time Markov process where they are indeed mathematically equivalent but for discrete-time case it is a dangerous misnomer: Spectral gap shown for reversible chain in [65] is around 1, but -1 can belong to $\sigma(P)$ when dealing with reversible HEMCs and can have a debilitating affect on attempt to provide concentration inequality as we will show shortly in example below. *Notion of spectral gap as we mentioned in theorem 45 and aperiodicity are equivalent to Poincare (modulo a constant factor).* Discrete-time variants on unbounded state space e.g., (Polish spaces) always have symmetry assumption [66]. Even though [67] managed to prove L^2 spectral gap and aperiodicity inequality for HEMCs under minimal assumptions but reversibility assumption was crucial. [55] used perturbation theory to prove spectral gap for *positive Markov transition operator* under uniform integrability condition, but how to come up with easily verifialble conditions that imply Uniform Integrability is not obvious. Continuous time variant [68], links Lyapunov condition to different functional inequalities but apriori makes the assumption on the structure of invariant measure and lyapunov condition is equivalent to Poincare only under

reversibility assumption.

Periodic behavior leading to failure of Poincare inequality

Example 1. *We consider an illuminating example from [57], μ_π is a lebesgue measure on $[0,1]$ for $k(x,y) := 2\chi_{[0,0.5]\times[0.5,1]}(x,y) + 2\chi_{[0.5,1]\times[0,0.5]}(x,y)$, define a Markov transition operator as $Pf(x) := \int f(y)k(x,y)d\mu_\pi(y)$. Markov chain is reversible, ergodic and even hyperbounded. Let $g(x) := \chi_{[0,0.5]}(x) - \chi_{[0.5,1]}(x)$, then $Pg(x) = -g(x)$ and $\|P^n\big(g - \mu_\pi(g)\big)\|_{L^2(\mu_\pi)} = \|(-1)^n g - \mu_\pi(g)\|_{L^2(\mu_\pi)} = 1$ for all $n \in \mathbb{N}$ and $Cov_{x_0\sim\mu_\pi}[g(x_n)g(x_{n+m})]$ can only be upper bounded by 1.*

Remark 52. *The preceding example urges us to ensure that HEMC doesn't posses such a periodic behavior w.r.t observables: that we aim to show, as ergodic averages concentrate sharply around their mean w.r.t ergodic invariant distribution. Here comes the existence of exponential-type Lyapunov function into play.*

5.3.2 Exponential decay of correlation on Lipschitzs observables of HEMCS and Spectral gaps

A curious reader can identify that for HEMCs, an exponential-type Lyapunov function, implies an exponential decay of correlation as in (5.46) on a subset of $L^2(\mu_\pi)$. Therefore, an attempt to discover $L^2(\mu_\pi)$ spectral gap for HEMCs with exponential-type Lyapunov function, should take initial domain of analysis as the space of Lipschitz function w.r.t contractive metric $d_{\beta*}$ for Markovian operator P that we will denote by $L(d_{\beta*})$.

Theorem 53. *Markov transition operator for HEMCs with exponential type Lyapunov function,is a strict contraction on the domain of Lipschitz functions as sub-*

space of Banach spaces $L^p(\mu_\pi)$, with the following anaytical bounds $n \in \mathbb{N}$:

$$\|P^n f - \mu_\pi(f)\|_{L^p} \leqslant 3(\lambda_{\beta*})^n \|f\|_{L(d_{\beta*})} p^{\frac{1}{2}} \sqrt{\frac{C_\pi}{2}} \tag{5.47}$$

$$\|f - \mu_\pi(f)\|_{L^p} \leqslant 3\|f\|_{L(d_{\beta*})} p^{\frac{1}{2}} \sqrt{\frac{C_\pi}{2}}. \tag{5.48}$$

Proof. Since, exponential-type Lyapunov function implies Transportation inequality and $P^n f$ is Lipschitz with norm $(\lambda_{\beta*})^n \|f\|_{L(d_{\beta*})}$, $\mu_\pi \in \mathcal{T}_1^{d_{\beta*}}(C_\pi)$, using (**??**) for $N = 1$ we can get an exponential bound on deviation of $P^n f$ subject to $x_0 \sim \mu_\pi$: to be precise $\mathbb{P}(|P^n f - \mu_\pi(f)| > t) \leqslant 2\exp\big(-\frac{t^2}{2C_\pi \|P^n f\|^2_{L(d_{\beta*})}}\big)$. Then we can calculate p-th moment via integral identity; step by step shown in the proof of Proposition 2.5.2, [69]

$$\begin{aligned}
\|P^n f - \mu_\pi(f)\|_{L^p} &= \int_0^\infty \mathbb{P}(|P^n f - \mu_\pi(f)|^p > s) ds \\
&= \int_0^\infty p t^{p-1} \mathbb{P}(|P^n f - \mu_\pi(f)| > t) dt \qquad (5.49) \\
&\leqslant \int_0^\infty p t^{p-1} 2\exp\bigg(-\frac{t^2}{2C_\pi \|P^n f\|^2_{L(d_{\beta*})}}\bigg) \\
&\leqslant (\lambda_{\beta*})^n \|f\|_{L(d_{\beta*})} \underbrace{3p^{\frac{1}{2}} \sqrt{\frac{C_\pi}{2}}}_{C_{(p,\mu_\pi)}} \qquad (5.50)
\end{aligned}$$

where (5.49) follows from change of variable $t := s^{\frac{1}{p}}$ and (5.50) follows from property of gamma function $\Gamma(a) \leqslant 3a^a$ for all $a \geqslant \frac{1}{2}$. □

Theorem 54. *Functions corresponding to periodic behavior (if they exist) w.r.t HEMC do not belong to class of Lipschitz function and for any two Lipschitz function f and g, we have the following decay of correlation:*

$$Cov_{x_0 \sim \mu_\pi}\big[f(x_n) g(x_{n+m})\big] \leqslant 9C_\pi \big(\|f\|_{L(d_{\beta*})} \vee \|g\|_{L(d_{\beta*})}\big)^2 (\lambda_{\beta*})^m \tag{5.51}$$

Proof. First claim follows from discussion in example 1 and (5.51) follows from stationarity and Cauchy-Schwarz inequality along with upper bounds given in (5.48) and (5.47). □

Any attempt to strengthen Theorem 53 into likes of *Poincare inequality*, would first require us to prove denseness of Lipshchitz functions in the space of L^p functions.

Theorem 55. *Indeed, we have something similar to spectral gap in terms of $L^p(\mu_\pi)$, i.e., for any $f \in L^p(\mu_\pi)$ there exists a constant $C_{(p,\mu_\pi)}(f)$ only dependent on the function f, $p-th$ power of Banach space under consideration and the $T-E$ constant of the stationary distribution μ_π such that $\forall$, $n \in \mathbb{N}$:*

$$\|P^n f - \mu_\pi(f)\|_{L^p} \leqslant (\lambda_{\beta*})^n C_{(p,\mu_\pi)}(f) \tag{5.52}$$

Proof. It suffices to show that $L_{d_{\beta*}}$ is dense in $L_p(\mu_\pi)$. Although, not so evident from the first look at (**??**), it turns out that indicator functions are indeed $1-$ Lipschitz w.r.t metric introduced by [48]. We can trivially pick sequence of sets $(A_n)_{n\in\mathbb{N}}$, such that $\mu_\pi(A_n) \searrow 0$. Notice that $x \mapsto P\chi_{A_n}(x)$ is again Lipschitz w.r.t $d_{\beta*}$. Assume that the subspace $L_{d_{\beta*}} \cap L_\infty(\mu_\pi)$ is not dense in $L_p(\mu_\pi)$, a Corrolary of Hahn-Banach theorem reveals that there exists a $g \in L_q(\mu_\pi) \neq 0$, where q is a conjugate power of p i.e., $\frac{1}{p} + \frac{1}{q} = 1$ and $\langle f, g\rangle_{\mu_\pi} = 0, \forall f \in L_{d_{\beta*}} \cap L_\infty(\mu_\pi)$. As discussed in the begining of this section, recall $P^* : L_q(\mu_\pi) \to L_q(\mu_\pi)$ is a Markov kernel. Exploiting the invariance and duality structure:

$$0 = \mu_\pi(g) = \int P^* g(x) d\mu_\pi(x) = \langle P\chi_{A_n}, g\rangle = \langle \chi_{A_n}, P^* g\rangle = \int P^* g(x)\chi_{A_n}(x) d\mu_\pi(x) \searrow 0, \tag{5.53}$$

which implies $P^*g(x) = 0$, μ_π a.e. x but P^* is Markov kernel implying $g = 0$ which is in contradiction with hypothesis. □

Theorem 56. *For a reversible HEMC, existence of an exponential-type Lyapunov function implies $L^2(\mu_\pi)$ 'spectral gap and aperiodicity = poincare' i.e., for some $\hat{\beta} \in (0, 1)$:*

$$\sup_{f \in L^2(\mu_\pi)} \frac{\|Pf - \mu_\pi(f)\|_2}{\|f - \mu_\pi(f)\|_2} < \hat{\beta}. \tag{5.54}$$

Proof. With the inequality (5.47) in hand and denseness of $L_{d_{\beta^*}}$ in $L^2(\mu_\pi)$, proof follows from Lemma 2.8 and 2.9 of [67] with $\hat{\beta} := \lambda_{\beta^*}$. □

Remark 57. *For a reversible ergodic Markov chain,* Poincare inequality *is merely a validation of* aperiodic *behavior.*

5.4 Large Deviations and Transport-Information Inequality

A parallel line of research for concentration inequalities is based on Large deviation principle for Markov chains developed by Donsker and Varadhan (see e.g., [70]), which roughly implies that for any $\mathcal{B} \in \mathbb{B}\big(\mathcal{P}(\mathcal{X})\big)$, occupation measure of Markov chain, $\frac{1}{N}\sum_{i=1}^{N} \delta_{x_i}$, with stationary measure μ_π satisfies:

$$\mathbb{P}(\frac{1}{N}\sum_{i=1}^{N} \delta_{x_i} \in \mathcal{B}) \simeq e^{-N \inf_{\nu \in \mathcal{B}} I(v|P,\mu_\pi)}, \tag{5.55}$$

where P is the transition operator associated with Markov chain having a stationary measure μ_π. Donsker-Varadhan information of a measure ν w.r.t Markov chain is represented by:

$$I(v|P, \mu_\pi) := \inf_{Q \in \mathcal{P}(\mathcal{X}\times\mathcal{X}): \pi_1 Q = \pi_2 Q = \nu} Ent(Q||\mu_\pi \otimes P), \tag{5.56}$$

where $\pi_1 Q$ and $\pi_2 Q$ denotes first and second marginals of Q and $\mu_\pi \otimes P = \mu_\pi(x)P(x,y)$. In the language of large deviation, Donsker-Varadhan information is the rate function for occupation measure of Markov chain.

Definition 18. *We say the pair (P, μ_π) of Markov chain satisfies Transport Information inequality (T-I) for some $C_\pi > 0$ if for all $\nu \in \mathcal{P}(X)$:*

$$\mathcal{W}_d^2(\nu, \mu_\pi) \leqslant 2C_\pi I(v|P, \mu_\pi). \tag{5.57}$$

.

Feynman-Kac semigroup and stationary chain. Feynman-Kac semigroups play an integral role in the study of fluctuations of time additive quantities of diffusion process in continuous time, [58]. The reader is referred to [46], for a detailed exposition on this topic in discrete time setting. As in [46], we identify with reward function r, three diffrent Feynman-Kac semigroup for the chain as operators $e^r P, P^r$ and $Pe^r : D_{(.)} \subseteq L^2(\mu_\pi) \to L^2(\mu_\pi)$ such that for all $g \in D_{(.)}$ (domain on which semigroup operators can be viewed as bounded operator):

$$(e^r P)^N g(x) := \mathbb{E}_x[g(x_N) e^{\sum_{i=0}^{N-1} r(x_i)}] \tag{5.58}$$

$$(P^r)^N g(x) := \mathbb{E}_x[g(x_N) e^{\sum_{i=1}^{N-1} r(x_i) + \frac{r(x_0)+r(x_N)}{2}}] \tag{5.59}$$

$$(Pe^r)^N g(x) := \mathbb{E}_x[g(x_N) e^{\sum_{i=1}^{N} r(x_i)}]. \tag{5.60}$$

Remark 58. *Notice that $\|Pe^r - P\| = \|P \bullet (e^r - 1) \cdot \|$, a composition of Markov Transition operator and multiplication operator. Therefore, $\|Pe^r - P\| \leqslant \|P\| \|e^r - 1\|_{L^\infty(\mu_\pi)}$: provided that the stationary distribution is concentrated on a compact set or observable is bounded and spectral gap exists for P, perturbation analysis developed*

by [54] (or Fredholm theory) can be used to show spectral gap for Feynman-Kac operators and consequently exponential concentration inequalities.

Theorem 59. *[46], assume that Markov chain is reversible then the following conditions are equivalent:*

$$(i)\mathcal{W}_d^2(\nu,\mu_\pi) \leqslant 2C_\pi I(v|P,\mu_\pi), \quad \textit{for all } \nu \in \mathcal{P}(\mathcal{X}). \tag{5.61}$$

$$(ii)\|(P^{sr})^N\|_2 \leqslant e^{N\left(s\ \mu_\pi(r)+\frac{C_\pi s^2}{2}\right)}, \textit{for all } s>0, N\in\mathbb{N} \textit{ and } r \textit{ s.t } \|r\|_{L(d)} \leqslant 1.$$

$$\begin{aligned}(iii)&\mathbb{P}_\beta\left(\frac{1}{N}\Big(\sum_{i=1}^{N-1} r(x_i)+\frac{r(x_0)+r(x_N)}{2}\Big)-\mu_\pi(r)\geqslant\epsilon\right)\\ &\leqslant \left\|\frac{d\beta}{d\mu_\pi}\right\|_2 \exp\left(\frac{-N\epsilon^2}{2C_\pi\|r\|^2_{L(d)}}\right),\ \forall\beta << \mu_\pi.\end{aligned} \tag{5.62}$$

Remark 60. *Although, results from the Theorem 59 have a great theoretical significance, assumption of reversibility is not easy to check and its validity is somewhat questionable for Harris Ergodic chains. We will take a heuristic approach to explore concentration phenomenon and then come back to LDP for non-reversible chains.*

5.4.1 Hyper-contractivity/ boundedness and exponential Lyapunov function Implies Concentration

Example 2. *Let's improvise by considering one dimensional linear Gaussian dynamical system with $|\alpha|<1$:*

$$x_{n+1} = \alpha x_n + w_n. \tag{5.63}$$

An easy check reveals stationary distribution of (5.63) is $\gamma_{1,\alpha} := \mathcal{N}\left(0, \frac{1}{1-\alpha^2}\right)$

Theorem 61. *Given α such that $|\alpha|<1$, there exitst $p:=p(\alpha)>2$ such that $\|P\|_{2\to p,\gamma_{1,\alpha}} \leqslant 1$ (Hypercontractivity) and $\left\|\left(e^r P\right)^N\right\| \leqslant e^{N\left(\gamma_{1,\alpha}(r)+\left(\frac{2p}{p-2}\right)\frac{\|r\|^2_{L(d)}}{2}\right)}$.*

Proof. A trivial application of change of variable and Stein's lemma reveals:

$$\|P\|_{2\to p} \leqslant \frac{1}{(1-\alpha^4)^{\frac{1}{4}}} \frac{1}{\left(1-\frac{\alpha^2 p}{(1+\alpha^2)}\right)^{\frac{1}{2p}}} \frac{1}{e^{\frac{2}{p}\left(1-\frac{\alpha^2 p}{(1+\alpha^2)}\right)}} \tag{5.64}$$

Recall, $\gamma_{1,\alpha} \in \mathcal{T}_1^d(1)$ i.e., satisfies T-E inequality with constant 1. It follows from previous equation (5.64) that for any $g \in L^2\big(\gamma_{1,\alpha}\big)$ such that $\|g\|_2 \leqslant 1$, there exists a $p > 2$ such that $\|Pg\|_p \leqslant 1$ by applying Cauchy-Schwarz to the powers $\frac{p}{2}$ and its' conjugate number $\frac{p}{p-2}$ we get:

$$\|\big(e^r P\big)g\|_2 \leqslant \|Pg\|_p \bigg(\int e^{\frac{2p}{p-2} r(x)} d\gamma_{1,\alpha}(x) \bigg)^{\frac{p-2}{2p}} \leqslant \|P\|_{2\to p} \bigg(e^{\gamma_{1,\alpha}(r)} e^{\big(\frac{2p}{p-2}\big)\frac{\|r\|^2_{L(d)}}{2}} \bigg), \tag{5.65}$$

where (5.65) follows from hypercontractivity of transition operator and the fact that stationary distribution satisfies T-E inequality with $C = 1$ and d is the trivial metric. Conclusion follows from the trivial inequality $\|\big(e^r P\big)^N\| \leqslant \|\big(e^r P\big)\|^N$ and definition of the operator norm. □

Corollary 9. *Without any reversibility assumption on the pair (P, μ_π), if the stationary distribution satisfies T-E inequality i.e., $\mu_\pi \in \mathcal{T}_1^d(C_\pi)$ and for some $p > 2$ associated transition kernel is hypercontractive i.e., $\|P\|_{2\to p} \leqslant 1$ then for any initial distribution $\beta << \mu_\pi$ and $N \in \mathbb{N}$:*

$$\mathbb{P}_\beta \bigg(\frac{1}{N} \sum_{i=0}^{N-1} r(x_i) - \mu_\pi(r) \geqslant \epsilon \bigg) \leqslant \bigg\| \frac{d\beta}{d\mu_\pi} \bigg\|_2 \exp\bigg(- \frac{N\epsilon^2 (p-2)}{4C_\pi \|r\|^2_{L(d)} p} \bigg). \tag{5.66}$$

Consequently, given $\epsilon > 0$ and $\delta \in (0,1)$ we have that for all $N \geqslant \frac{\big\|\frac{d\beta}{d\mu_\pi}\big\|_2 ln(\frac{1}{1-\delta}) 4C_\pi \|r\|^2_{L(d)} p}{\epsilon^2 (p-2)}$ stationary chain satisfies $\mathbb{P}_{\mu_\pi}\bigg(\frac{1}{N}\sum_{i=0}^{N-1} r(x_i) - \mu_\pi(r) \geqslant \epsilon \bigg) \leqslant 1 - \delta$.

Similar theoretical result can be interpolated from functional inequalities in [46],

but here we would like to illuminate the effectiveness of this approach when system dynamics are unknown and give insight into why concentration phenomenon happens.

Remark 62. *We can verify T-E inequality for stationary distribution via existence of exponential Lyapunov function so this approach is pertinent to problems in reinforcement learning and control where one does not know a priori stationary distribution of the Markov chain.*

Theorem 63. *Without any assumption of reversibility, if there exists an exponential-type Lyapunov function for the HEMC under consideration: and for some $q > 2$, $P : L^2(\mu_\pi) \longrightarrow L^q(\mu_\pi)$ is hyperbounded with norm $\|P\|_{L^2 \to L^q} < e^{\frac{1}{2}-\frac{1}{q}}$. Given $n \in \mathbb{Z}^+$ and $\delta \in (0,1)$ and an initial distribution $\beta << \mu_\pi$,*
if $N \geqslant \left\|\frac{d\beta}{\mu_\pi}\right\|_2 \ln(\frac{1}{1-\delta})\left(\frac{2n^2q}{(q-2)-2n^2q\ln\|P\|_{L^2\to L^q}}\right)$, we have that

$$\mathbb{P}_{\mu_\pi}\left(\frac{1}{N}\sum_{i=0}^{N-1} r(x_i) - \mu_\pi(r) \geqslant \frac{\sqrt{C_\pi}\|r\|_{L(d_{\beta^*})}}{n}\right) < 1-\delta. \tag{5.67}$$

Proof. Since, assumptions made in Theorem ensure $e^r P \in \mathcal{B}(L^2(\mu_\pi))$, we have that $\|(e^r P)^N\| \leqslant \|e^r P\|^N$. Let $\epsilon(n) := \frac{\sqrt{2C_\pi}\|r\|_{L(d)}}{n}$, and notice that $\sqrt{2C_\pi}\|r\|_{L(d)}$ is proportional to the square root of asymptotic variance. In order to upper bound the following probability:

$$\mathbb{P}_{\mu_\pi}\left(\frac{1}{N}\sum_{i=0}^{N-1} r(x_i) - \mu_\pi(r) \geqslant \frac{\sqrt{C_\pi}\|r\|_{L(d_{\beta^*})}}{n}\right), \tag{5.68}$$

□

we should be able to decrease the term $e^{\left(N\left[\ln\|P\|_{2\to q} - \frac{(q-2)}{n^2 2q}\right]\right)}$ as N increase; for every $n \in \mathbb{Z}^+$, $\|P\|_{L^2\to L^q} < e^{\frac{1}{2}-\frac{1}{q}}$ suffices.

Multiplication operator defined by e^r In order to understand the sufficiency of Hyperboundedness in the preceding concentration inequality, one needs to realize that Feynman-Kac Semigroup defined by $e^r P$ is nothing but a composition of the Markov transition operator P followed by the multiplication operator e^r. Therefore, it is imperative to take into consideration the the spectral analysis for multiplication operator that acts on $L^p(\mu_\pi)$ with domain, $D(e^r)(p) := \{f \in L^p : e^r f \in L^p\}$. The operator $\left(e^r, D(e^r)_p\right)$ is closed and densely defined, see e.g., Proposition 3.10 from Chapter 1 in [59]. Essential range of e^r with μ_π as an underlying measure is defined as $e^r_{ess}(\mu_\pi) := \left\{\lambda \in [1,\infty) : \mu_\pi\big(|e^r - \lambda| < \epsilon\big) \neq 0,\ \forall\, \epsilon > 0\right\}$ and essential norm of e^r, $\|e^r\|_\infty := \sup\{\lambda \in e^r_{ess}(\mu_\pi)\}$.

Remark 64. *Multiplication operator e^r is bounded in some and hence all L^p, iff $\|e^r\|_\infty < \infty$. Consequently, $D(e^r)(p) = L^p$ and $\|e^r\|_{L^p \to L^p} = \|e^r\|_\infty$. If the stationary measure μ_π is not compactly supported and is absolutely continuous w.r.t Lebesgue measure then $\|e^r\|_{L^p \to L^p} = \infty$ and hyperboundedness / hypercontractivity should be sufficient under the existence of exponential type Lyapunov function, precisely said:*

Theorem 65. *Exponential type Lyapunov function ensures that for every $q > 2$, $e^r : L^q \longrightarrow L^2$ is a bounded operator with norm:*

$$\|e^r\|_{L^q \to L^2} \leqslant \exp\left(\mu_\pi(r) + \frac{2qC_\pi \|r\|^2_{L(d)}}{(q-2)2}\right). \tag{5.69}$$

Proof. Proof is a simple application of Cauchy-Schwarz and exponential moment inequality for Lipschitz functions under distribution μ_π satisfying Transport-entropy inequality. □

5.5 Functional Inequality via rate function of chain associated to PP^*

As the correct formulation of *Poincare inequality* in 17, required verifying a functional inequality w.r.t PP^*–same is the case for concentration of Lipschitz observables from the Harris chain.

Definition 19. *We say the pair (P, μ_π) satisfies symmetrized transport-information inequality with constant $\hat{C}_\pi > 0$, if:*

$$\mathcal{W}_d^2(\nu, \mu_\pi) \leqslant \hat{C}_\pi I(\nu|PP^*, \mu_\pi), \ \ \forall \nu \in \mathcal{P}(X). \tag{5.70}$$

Lemma 11. *Symmetrized transport-information inequality* (5.61) *is equivalent to following concentration inequality:*

$$\mathbb{P}_\beta\bigg(\frac{1}{N}\sum_{i=0}^{N-1} r(x_i) - \mu_\pi(r) \geqslant \epsilon\bigg) \leqslant \bigg\|\frac{d\beta}{d\mu_\pi}\bigg\|_2 \exp\bigg(-\frac{N\epsilon^2}{2\hat{C}_\pi \|r\|^2_{L(d)}}\bigg), \tag{5.71}$$

see Theorem 2.12 and Remark 2.13 of [46].

The following equation can be verified from [46]:

$$I(\nu|PP^*, \mu_\pi) \leqslant 2I(\nu|P, \mu_\pi), \ \ \forall \nu \in \mathcal{P}(X). \tag{5.72}$$

Remark 66. *This implies from practical point of view that verifying T-I is not useful as symmetrized transport-information inequality implies T-I inequality (with a different constant) but concentration won't follow without a priori knowledge of reversibility as mentioned in theorem 59. Therefore, from now on our emphasis would be centered around analysis of symmetrized transport-information inequality.*

A simple corollary follows by relating lemma 11 to corollary 9:

Corollary 10. *If the stationary distribution satisfies T-E inequality with constant $C_\pi > 0$ and for some $p > 2$ associated transition kernel is hypercontractive, then the pair (P, μ_π) satisfies symmetrized transport-information inequality with constant $\hat{C}_\pi = C_\pi\big(\frac{2p}{p-2}\big)$.*

Remark 67. *Existence of exponential-type Lyapunov function verifies T-E but verifying hypercontractivity without explicit knowledge of stationary distribution can be very difficult. Therefore, we would like to look for other verifiable conditions that can lead to or are equivalent to symmetrized transport-information inequality, which brings us to the following sufficient condition:*

Theorem 68. *If there exist positive $\hat{\alpha}$, β and $C(\hat{\alpha})$ such that $\beta < \hat{\alpha}$ and $PP^* W_{\hat{\alpha}} \leqslant C(\hat{\alpha}) W_\beta$. Poincare inequality i.e. $Var_{\mu_\pi}(f) \leqslant c_p \mathcal{E}_{PP^*}(f)$, holds: then with metric $d(x,y) := \sqrt{\hat{\alpha} - \beta}\|x - y\|$, the pair (P, μ_π) satisfies symmetrized transport-information inequality*

Proof. Proof follows the same line of argument as in Theorem 2.33 of [46]. □

Chapter 6

Conclusion and future work

We have managed to make significant progress in giving typical order explicit in N and n of various spectral statistics that determine performance of OLS. Although OLS is transient in spatially inseparable case but their does exist a sweet spot on length of the simulated trajectory so that the error is relatively small. In order to find that regime we will need to extend the typical order analysis in section 4.1 to higher typical sized rows. From functional inequalities side, we narrowed down the concentration phenomenon for Harris ergodic Markov chains to a study of composition of the Markov transition operator followed by a an exponentiated multiplication operator defined by observable under consideration. Hyperboundedness and exponential-type Lyapunov function suffices for concentration phenomen. A major contribution of this work is the conclusive remark on assumption of reversibility, which we reiterate: $Im(I - P)$ is closed under reversibility assumption. However, there are still unanswered questions that needs further exploration: *Does exponential-type Lyapunov function imlpy hyperboundedness?* If so is the case then $P(x, \cdot) << \mu_\pi$ for almost all x w.r.t μ_π, i.e., information about the structure of the invariant measure can be extracted; which is extremely useful when exact system dynamics are unknown. *aperiodicity and hyperboundedness*: we believe that via continuity of *fredholm index* we can extend the norm control on $\|P\|_{L^2-L^p}$ that implies *aperiodicity*. Currently, only result for $p = 3, 4$ is available. Instead of *hyperboundedness* or *reversibility*, is there any easily verifiable assumption that ensures $Im(I - P)$ is closed ?

Bibliography

[1] 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, pp. 16 666–16 676, 2020.

[2] 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.

[3] 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*, 2018, pp. 439–473.

[4] A. Tsiamis and G. J. Pappas, "Linear systems can be hard to learn," in *2021 60th IEEE Conference on Decision and Control (CDC)*. IEEE, 2021, pp. 2903–2910.

[5] G. Blower and F. Bolley, "Concentration inequalities on product spaces with applications to markov processes," *arXiv preprint math/0505536*, 2005.

[6] H. Djellout, A. Guillin, L. Wu *et al.*, "Transportation cost-information inequalities and applications to random dynamical systems and diffusions," *Annals of Probability*, vol. 32, no. 3B, pp. 2702–2732, 2004.

[7] M. A. Naeem, A. Khazraei, and M. Pajic, "From spectral theorem to statistical independence with application to system identification," *arXiv preprint arXiv:2310.10523*, 2023.

[8] S. Oymak and N. Ozay, "Revisiting ho–kalman-based system identification: Robustness and finite-sample analysis," *IEEE Transactions on Automatic Control*, vol. 67, no. 4, pp. 1914–1928, 2021.

[9] 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.

[10] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, "Finite time identification in unstable linear systems," *Automatica*, vol. 96, pp. 342–353, 2018.

[11] S. Tu and B. Recht, "Least-squares temporal difference learning for the linear quadratic regulator," in *International Conference on Machine Learning*, 2018, pp. 5005–5014.

[12] A. Tsiamis and G. J. Pappas, "Online learning of the kalman filter with logarithmic regret," *IEEE Transactions on Automatic Control*, 2022.

[13] B. Ho and R. E. Kálmán, "Effective construction of linear state-variable models from input/output functions: Die konstruktion von linearen modeilen in der darstellung durch zustandsvariable aus den beziehungen für ein-und ausgangsgrößen," *at-Automatisierungstechnik*, vol. 14, no. 1-12, pp. 545–548, 1966.

[14] A. Tsiamis, I. M. Ziemann, M. Morari, N. Matni, and G. J. Pappas, "Learning to control linear systems can be hard," in *Conference on Learning Theory*. PMLR, 2022, pp. 3820–3857.

[15] G. E. Dullerud and F. Paganini, *A course in robust control theory: a convex approach*. Springer Science & Business Media, 2013, vol. 36.

[16] M. Reed and B. Simon, *I: Functional analysis*. Gulf Professional Publishing, 1980, vol. 1.

[17] M. Talagrand, "Transportation cost for gaussian and other product measures," *Geometric & Functional Analysis GAFA*, vol. 6, no. 3, pp. 587–600, 1996.

[18] M. A. Naeem, "Learning and concentration for high dimensional linear gaussians: an invariant subspace approach," *arXiv preprint arXiv:2304.01708*, 2023.

[19] T. Tao, V. Vu, and M. Krishnapur, "Random matrices: Universality of esds and the circular law," 2010.

[20] A. Ferber, "When combinatorics meets littlewood-offord theory."

[21] S. Axler, "Down with determinants!" *The American mathematical monthly*, vol. 102, no. 2, pp. 139–154, 1995.

[22] ——, *Linear algebra done right*. Springer Science & Business Media, 1997.

[23] M. A. Naeem and M. Pajic, "High dimensional geometry and limitations in system identification," *arXiv preprint arXiv:2305.12083*, 2023.

[24] K. P. Costello, T. Tao, and V. Vu, "Random symmetric matrices are almost surely nonsingular," 2006.

[25] M. Abdullah Naeem and M. Pajic, "Spectral statistics of the sample covariance matrix for high dimensional linear gaussians," *arXiv e-prints*, pp. arXiv–2312, 2023.

[26] M. Rudelson, "Recent developments in non-asymptotic theory of random matrices," *Modern aspects of random matrix theory*, vol. 72, p. 83, 2014.

[27] V. H. Vu and T. Tao, "The condition number of a randomly perturbed matrix," in *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, 2007, pp. 248–255.

[28] M. Rudelson and R. Vershynin, "Smallest singular value of a random rectangular matrix," *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, vol. 62, no. 12, pp. 1707–1739, 2009.

[29] N. A. Cook, *Spectral properties of non-Hermitian random matrices*. University of California, Los Angeles, 2016.

[30] A. Tsiamis, I. Ziemann, N. Matni, and G. J. Pappas, "Statistical learning theory for control: A finite sample perspective," *arXiv preprint arXiv:2209.05423*, 2022.

[31] M. Talagrand, *Mean field models for spin glasses: Volume I: Basic examples*. Springer Science & Business Media, 2010, vol. 54.

[32] T. Tao and V. Vu, "On random pm 1 matrices: singularity and determinant," in *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, 2005, pp. 431–440.

[33] B. Hao, N. Lazic, Y. Abbasi-Yadkori, P. Joulani, and C. Szepesvari, "Provably efficient adaptive approximate policy iteration," *arXiv preprint arXiv:2002.03069*, 2020.

[34] S. Oymak, "Stochastic gradient descent learns state equations with nonlinear activations," in *Conference on Learning Theory*, 2019, pp. 2551–2579.

[35] M. Fazel, R. Ge, S. Kakade, and M. Mesbahi, "Global convergence of policy gradient methods for the linear quadratic regulator," in *International Conference on Machine Learning*, 2018, pp. 1467–1476.

[36] T. Zahavy, A. Cohen, H. Kaplan, and Y. Mansour, "Average reward reinforcement learning with unknown mixing times," *arXiv preprint arXiv:1905.09704*, 2019.

[37] M. A. Naeem and M. Pajic, "Transportation-inequalities, lyapunov stability and sampling for dynamical systems on continuous state space," *arXiv preprint arXiv:2205.12448*, 2022.

[38] K. Marton *et al.*, "Measure concentration for euclidean distance in the case of dependent random variables," *Annals of probability*, vol. 32, no. 3B, pp. 2526–2544, 2004.

[39] M. Lin, "On the uniform ergodic theorem. ii," *Proceedings of the American Mathematical Society*, vol. 46, no. 2, pp. 217–225, 1974.

[40] L. Wu, "Essential spectral radius for markov semigroups (i): discrete time case," *Probability theory and related fields*, vol. 128, no. 2, pp. 255–321, 2004.

[41] J. Bun, J.-P. Bouchaud, and M. Potters, "Cleaning large correlation matrices: tools from random matrix theory," *Physics Reports*, vol. 666, pp. 1–109, 2017.

[42] S. Chatterjee, "Stein's method for concentration inequalities," *Probability theory and related fields*, vol. 138, no. 1-2, pp. 305–321, 2007.

[43] M. Ledoux, *The concentration of measure phenomenon*. American Mathematical Soc., 2001, no. 89.

[44] K. Łatuszyński, B. Miasojedow, W. Niemiro *et al.*, "Nonasymptotic bounds on the estimation error of mcmc algorithms," *Bernoulli*, vol. 19, no. 5A, pp. 2033–2066, 2013.

[45] M. A. Naeem and M. Pajic, "Learning expected reward for switched linear control systems: A non-asymptotic view," *arXiv preprint arXiv:2006.08105*, 2020.

[46] N.-Y. Wang and L. Wu, "Transport-information inequalities for markov chains," *The Annals of Applied Probability*, vol. 30, no. 3, pp. 1276–1320, 2020.

[47] B. Kloeckner, "Effective berry–esseen and concentration bounds for markov chains with a spectral gap," *The Annals of Applied Probability*, vol. 29, no. 3, pp. 1778–1807, 2019.

[48] M. Hairer and J. C. Mattingly, "Yet another look at harris' ergodic theorem for markov chains," in *Seminar on Stochastic Analysis, Random Fields and Applications VI*. Springer, 2011, pp. 109–117.

[49] F. Bolley and C. Villani, "Weighted csiszár-kullback-pinsker inequalities and applications to transportation inequalities," in *Annales de la Faculté des sciences de Toulouse: Mathématiques*, vol. 14, no. 3, 2005, pp. 331–352.

[50] Y. Zhang and K. W. Ross, "On-policy deep reinforcement learning for the average-reward criterion," in *International Conference on Machine Learning*. PMLR, 2021, pp. 12 535–12 545.

[51] F. Malrieu, "Logarithmic sobolev inequalities for some nonlinear pde's," *Stochastic processes and their applications*, vol. 95, no. 1, pp. 109–132, 2001.

[52] C. Villani, *Topics in optimal transportation*. American Mathematical Soc., 2003, no. 58.

[53] C. R. Givens and R. M. Shortt, "A class of wasserstein metrics for probability distributions." *Michigan Mathematical Journal*, vol. 31, no. 2, pp. 231–240, 1984.

[54] B. Kloeckner, "Effective limit theorems for markov chains with a spectral gap," *arXiv preprint arXiv:1703.09623*, 2017.

[55] L. Wu, "Uniformly integrable operators and large deviations for markov processes," *Journal of Functional Analysis*, vol. 172, no. 2, pp. 301–376, 2000.

[56] J. Glück, "Spectral gaps for hyperbounded operators," *Advances in Mathematics*, vol. 362, p. 106958, 2020.

[57] G. Cohen *et al.*, "l2-quasi-compact and hyperbounded markov operators," *arXiv preprint arXiv:2206.08003*, 2022.

[58] H. Touchette, "Introduction to dynamical large deviations of markov processes," *Physica A: Statistical Mechanics and its Applications*, vol. 504, pp. 5–19, 2018.

[59] K.-J. Engel and R. Nagel, *A short course on operator semigroups*. Springer Science & Business Media, 2006.

[60] S. G. Bobkov and F. Götze, "Exponential integrability and transportation cost related to logarithmic sobolev inequalities," *Journal of Functional Analysis*, vol. 163, no. 1, pp. 1–28, 1999.

[61] P. Cattiaux, "Long time behavior of markov processes," in *ESAIM: Proceedings*, vol. 44. EDP Sciences, 2014, pp. 110–128.

[62] C. P. Robert, G. Casella, and G. Casella, *Monte Carlo statistical methods*. Springer, 1999, vol. 2.

[63] D. A. Levin and Y. Peres, *Markov chains and mixing times*. American Mathematical Soc., 2017, vol. 107.

[64] R. Montenegro, P. Tetali *et al.*, "Mathematical aspects of mixing times in markov chains," *Foundations and Trends® in Theoretical Computer Science*, vol. 1, no. 3, pp. 237–354, 2006.

[65] L. Miclo, "On hyperboundedness and spectrum of markov operators," *Inventiones mathematicae*, vol. 200, no. 1, pp. 311–343, 2015.

[66] A. Taghvaei and P. G. Mehta, "On the lyapunov foster criterion and poincaré inequality for reversible markov chains," *IEEE Transactions on Automatic Control*, 2021.

[67] M. Hairer, A. M. Stuart, and S. J. Vollmer, "Spectral gaps for a metropolis–hastings algorithm in infinite dimensions," 2014.

[68] P. Cattiaux and A. Guillin, "Hitting times, functional inequalities, lyapunov conditions and uniform ergodicity," *Journal of Functional Analysis*, vol. 272, no. 6, pp. 2361–2391, 2017.

[69] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*. Cambridge university press, 2018, vol. 47.

[70] M. D. Donsker and S. S. Varadhan, "Asymptotic evaluation of certain markov process expectations for large time, i," *Communications on Pure and Applied Mathematics*, vol. 28, no. 1, pp. 1–47, 1975.