Streaming PCA for Markovian Data

Syamantak Kumar*[1] and Purnamrita Sarkar †[2]

[1]Department of Computer Science, University of Texas at Austin
[2]Department of Statistics and Data Sciences, University of Texas at Austin

May 5, 2023

Abstract

Since its inception in Erikki Oja's seminal paper in 1982, Oja's algorithm has become an established method for streaming principle component analysis (PCA). We study the problem of streaming PCA, where the data-points are sampled from an irreducible, aperiodic, and reversible Markov chain. Our goal is to estimate the top eigenvector of the unknown covariance matrix of the stationary distribution. This setting has implications in situations where data can only be sampled from a Markov Chain Monte Carlo (MCMC) type algorithm, and the goal is to do inference for parameters of the stationary distribution of this chain. Most convergence guarantees for Oja's algorithm in the literature assume that the data-points are sampled IID. For data streams with Markovian dependence, one typically downsamples the data to get a "nearly" independent data stream. In this paper, we obtain the first sharp rate for Oja's algorithm on the entire data, where we remove the logarithmic dependence on n resulting from throwing data away in downsampling strategies.

1 Introduction

Principal Component Analysis (PCA), invented by Karl Pearson in 1901, is a well-established dimensionality reduction technique that can be used to extract linearly uncorrelated features from high-dimensional datasets. The many applications of PCA include image processing, visualization, and dictionary learning [16]. Mathematically, PCA involves the computation of the principal eigenvectors of the unknown covariance matrix derived from the dataset. This is typically done by extracting principal eigenvectors of the sample covariance matrix. For very large dimensionality, it becomes more memory efficient to process one data-point at a time and keep updating the estimated principal component, reducing memory use from quadratic to linear in dimensionality. This method, also known as streaming PCA, has a rich history in Computer Science and Statistics. The problem of streaming PCA updates the estimated principal component one data-point at a time. One of the most popular algorithms for streaming PCA was introduced by Erikki Oja in 1982 [28, 29]. The "Oja" update has roots in the Hebbian principle put forward by Donald Hebb, a psychologist, in his 1949 book "Organization of Behavior" [11].

We consider the streaming PCA problem where the data are sampled from an irreducible, aperiodic, and reversible Markov Chain. In many applications, the data-points are not sampled IID but from an MCMC process which is converging to a target stationary distribution. Consider, for example, a set of machines, each hosting an arbitrary fraction of data-points or features. The machines can communicate with each other using a fixed graph topology that is connected. The goal is to design a streaming algorithm that respects this topology for communicating between machines and returns the principal component of the whole dataset. One way to achieve this would be to design

*syamantak@utexas.edu
†purna.sarkar@utexas.edu

a Metropolis-Hastings scheme that uses local information to design the transition matrix of a Markov chain with any desired stationary distribution. Governed by this Markov chain, a random walker then travels the network of machines and samples one data-point at a time from the current machine, and computes the update.

However, even when the data stream has reached the stationary distribution, the data-points are dependent, which deviates from the IID setup. Our goal is to obtain sharp error bounds for the $\sin^2$ error of the estimate from the streaming PCA algorithm and the true top eigenvector of the unknown covariance matrix.

Estimating the first principal component with streaming PCA : Let X_t be a mean zero d dimensional vector with covariance matrix Σ, and let η_t be a decaying learning rate. The update rule of Oja's algorithm is given as -

$$w_t \leftarrow (I + \eta_t X_t X_t^T) w_{t-1}, \quad w_t \leftarrow \frac{w_t}{\|w_t\|_2} \tag{1}$$

where w_t is the estimate of v_1 and η_t is the step-size at timestep t. We aim to analyse the $\sin^2$ error of Oja's iterate at timestep t, defined as $1 - \langle w_t, v_1 \rangle^2$, where v_1 is the top eigenvector of Σ.

Streaming PCA in the IID setting: For an IID data stream with $\mathbb{E}[X_i] = 0$ and $\mathbb{E}\left[X_i X_i^T\right] = \Sigma$, there has been a lot of work on determining the non-asymptotic convergence rates for Oja's algorithm and its various adaptations [14, 1, 3, 37, 12, 13, 25, 20, 24]. Amongst these, [14], [1] and [13] match the optimal offline sample complexity bound, suggested by the independent and identically distributed (IID) version of Theorem 1 (See Theorem 1.1 in [14]). We consider Oja's algorithm in a Markovian

Paper	Data Model	Online?	$\sin^2$ error rate	Sample Complexity
Jain et al. [14]	IID	Y	$O\left(\frac{\mathcal{V}}{\text{gap}^2}\frac{1}{n}\right)$	$O\left(\frac{\mathcal{V}}{\text{gap}^2}\frac{1}{\epsilon}\right)$
		N	$O\left(\frac{\mathcal{V}\log(d)}{\text{gap}^2}\frac{1}{n}\right)$	$O\left(\frac{\mathcal{V}\log(d)}{\text{gap}^2}\frac{1}{\epsilon}\right)$
Chen et al. [3]	Markov	Y	-	$O\left(\frac{\mathcal{V}}{\text{gap}^2}\frac{1}{\epsilon}\ln^2\left(\frac{\mathcal{V}}{\text{gap}^2}\frac{1}{\epsilon}\right)\right)$
Neeman et al. [27]	Markov	N	$O\left(\frac{\mathcal{V}\log\left(d^{2-\frac{\pi}{4}}\right)}{(1-\lvert\lambda_2(P)\rvert)\,\text{gap}^2}\frac{1}{n}\right)$	$O\left(\frac{\mathcal{V}\log\left(d^{2-\frac{\pi}{4}}\right)}{(1-\lvert\lambda_2(P)\rvert)\,\text{gap}^2}\frac{1}{\epsilon}\right)$
Theorem 1	Markov	Y	$O\left(\frac{\mathcal{V}}{(1-\lvert\lambda_2(P)\rvert)\,\text{gap}^2}\frac{1}{n}\right)$	$O\left(\frac{\mathcal{V}}{(1-\lvert\lambda_2(P)\rvert)\,\text{gap}^2}\frac{1}{\epsilon}\right)$

Table 1: Comparison of $\sin^2$ error rates and sample complexities using different data models and algorithms. Here $\text{gap} := (\lambda_1 - \lambda_2)$, where λ_1, λ_2 are the top 2 eigenvalues of Σ and the sample complexities represent the number of samples required to achieve $\sin^2$ error at most ϵ. We note that [1] and [13] also match the online sample complexity bound provided in [14]. Further, for the offline algorithm with IID data, [15] removes the $\log(d)$ factor in exchange for a constant probability of success for large enough n.

data setting where the data is generated from a reversible Markov chain with stationary distribution π. In this setting our goal is to estimate the principal eigenvector of $\mathbb{E}_\pi\left[X_i X_i^T\right]$. The challenge is that the data, even when it reaches stationarity, is dependent. Here the degree of dependence is captured by the second eigenvalue in magnitude of the transition matrix P of the Markov chain. We denote this quantity by $|\lambda_2(P)|$. This is closely related to the mixing time of a Markov chain (see 8), denoted as τ_{mix}, which is simply the time after which the conditional distribution of a state is close in total variational distance to its stationary distribution, π (See Section 2.1 for a quantitative description). Intuitively this means that samples which are τ_{mix} apart are "nearly" independent of each other and follow the distribution π.

Our contribution: Using a series of approximations, we obtain an optimal error rate for the $\sin^2$ error, which is worse by a factor of $1/(1 - |\lambda_2(P)|)$ from the corresponding error rate of the IID case.

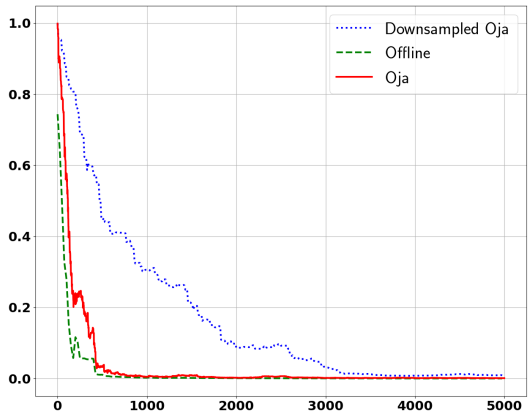


Figure 1: Comparison of Oja's algorithm with and without downsampling along with the offline baseline for Bernoulli state distributions. The X axis represents the sample size and the Y axis represents the $\sin^2$ error of each algorithm's estimate of the leading eigenvector. The experimental setup is available in Section 6. Observe that downsampling performs considerably worse compared to Oja's algorithm on the entire dataset.

Previous work [3] has established rates worse by a poly-logarithmic factor. We *break this logarithmic barrier* by considering a series of approximations of finer granularity which uses reversibility of the Markov chain and standard mixing conditions of irreducible and aperiodic Markov chains. Our rates are comparable to the recent work of [27] (Proposition 1) that establishes an offline error analysis for estimating the principal component of the empirical covariance matrix of Markovian data by using a Matrix Bernstein inequality. Our results also imply a linearly convergent decentralized algorithm for streaming PCA in a distributed setting. As a simple byproduct of our theoretical result, we also obtain a rate for Oja's algorithm applied on downsampled data, which is worse by a factor of $\log n$, as shown in Figure 1. To our knowledge, this is the first work that analyzes the Markovian streaming PCA problem without any downsampling that matches the error of the offline algorithm.

The crux of our analysis uses the mixing properties of the Markov chain. Strong mixing intuitively says that the conditional distribution of a state s in timestep k given the starting state is exponentially close to the stationary distribution of s, the closeness being measured using the total variation distance. All previous work on Markovian data exploits this property by conditioning on states many time steps before. However, it is crucial to a) adaptively find how far to look back and b) bound the error of the sequence of matrices we ignore between the current state and the state we are conditioning on. Observe that these two components are not independent of each other. Looking back too far makes the dependence very small but increases the error resulting from approximating a larger matrix product of intermediate matrices. We present a fine analysis that balances these two parts and then uses spectral theory to bound the second part within a factor of a variance parameter that characterizes the variability of the matrices and shows up in the analysis of [14, 27].

Related work on streaming PCA and online matrix decomposition on Markovian data: Amongst recent work, [3] is very relevant to our setting, since it analyzes Oja's algorithm with Markovian Data samples. Inspired by the ideas of [8], the authors propose a downsampled version of Oja's algorithm to reduce dependence amongst samples and provide a Stochastic Differential Equation (SDE) based analysis to achieve a sample complexity of $O\left(\frac{\mathcal{V}}{(\lambda_1-\lambda_2)^2}\frac{1}{\epsilon}\ln^2\left(\frac{\mathcal{V}}{(\lambda_1-\lambda_2)^2}\frac{1}{\epsilon}\right)\right)$ for $\sin^2$ error smaller than ϵ. We obtain a similar rate in Corollary 1 through our techniques. However, comparing with Theorem 1, we observe that downsampling leads to an extra $O\left(\ln\left(n\right)\right)$ factor. It is important to point out that [3] provides an analysis for estimating top k principal components, whereas this paper focuses on obtaining a sharp rate for the first principal component.

Another related work is [21], where the authors analyze online non-negative matrix factorization

for Markovian data and show that a suitably defined error converges to zero. Their analysis also uses the mixing properties of the Markov chain by conditioning on the distant past. While this work considers a harder problem, it does not provide a rate of convergence.

Stochastic Optimization with Markovian Data : In numerous applications such as Reinforcement Learning [2, 5] and Linear Dynamic Systems [9, 30, 4], a commonly-used approach for modeling data dependencies is to assume that a Markovian process produces the data and therefore there has been a lot of focus in statistics, optimization, and control theory to investigate and develop techniques for learning and modeling Markovian data. A class of methods has focused on establishing asymptotic bounds [34, 18, 23]. [8] provided one of the first non-asymptotic analyses of stochastic gradient descent (SGD) methods for general convex functions with Markovian data. Since then, there has been extensive work on SGD algorithms for both convex and non-convex problems [31, 6, 7, 10, 39, 33]. The convergence rates (sample complexities) obtained in these works apply to more general problems but do not exploit the matrix product structure inherent to Oja's algorithm. In this work, we develop novel techniques to show that a sharper analysis is possible for the PCA objective.

2 Problem Setup and Preliminaries

This section presents the problem setup and outlines important properties of the Markov chain that will be utilized subsequently. We assume that the Markov chain is irreducible, aperiodic, reversible, and starts in stationarity, with state distribution π [1]. Such a Markov chain can arise in various situations, for eg., while performing random walks on expander graphs which are used extensively in fields such as computer networks, error-correcting codes, and pseudorandom generators. Each state s of the Markov chain is associated with a distribution $D(s)$ over d-dimensional vectors with mean $\mu_s \in \mathbb{R}^d$ and covariance matrix $\Sigma_s \in \mathbb{R}^{d\times d}$.

For a random walk $s_1, s_2, \cdots s_t$ on C, we define the sequence of random variables $X_1, X_2 \cdots X_t$, where $X_i \sim D(s_i)$ is drawn from the distribution corresponding to the state s_i. We represent the total mean as $\mu := \mathbb{E}_{s\in\pi}[\mu_s]$ and the total covariance matrix as $\Sigma \in \mathbb{R}^{d\times d}$, which can be expressed as -

$$\begin{aligned}\Sigma &:= \mathbb{E}_{s\in\pi}\mathbb{E}_{D(s)}\left[(X-\mu)(X-\mu)^T\right] = \mathbb{E}_{s\in\pi}\mathbb{E}_{D(s)}\left[XX^T\right] - \mu\mu^T \\ &= \mathbb{E}_{s\in\pi}\mathbb{E}_{D(s)}\left[(X-\mu_s)(X-\mu_s)^T + \mu_s\mu_s^T\right] - \mu\mu^T \\ &= \mathbb{E}_{s\in\pi}\mathbb{E}_{D(s)}\left[\Sigma_s\right] + \mathbb{E}_{s\in\pi}\mathbb{E}_{D(s)}\left[\mu_s\mu_s^T\right] - \mu\mu^T\end{aligned}$$

In this work, we assume $\mu = 0$ i.e, the data-points are zero-mean with respect to π, which is a common assumption in the IID setting (see [1]) [2]. Therefore, $\Sigma = \mathbb{E}_{s\in\pi}\mathbb{E}_{D(s)}\left[XX^T\right]$.

Let the eigenvalues of Σ be denoted as $\lambda_1 > \lambda_2 \geq \lambda_3 \cdots \lambda_d$. Let v_1 denote the leading eigenvector of Σ and $V_\perp$ denote the $\mathbb{R}^{d\times(d-1)}$ matrix with the remaining eigenvectors as columns. For $s \in \Omega$, let $X \sim D(s)$. We proceed under the following standard assumptions (see for eg. [13]) -
For all states $s \in \Omega$,

Assumption 1. $\|\mathbb{E}_{s\in\pi}\left(\Sigma_s + \mu_s\mu_s^T - \Sigma\right)^2\|_2 \leq \|\mathbb{E}_{s\in\pi}\mathbb{E}_{D(s)}[(XX^T-\Sigma)^2]\|_2 \leq \mathcal{V}$

Assumption 2. $\|XX^T - \Sigma\|_2 \leq \mathcal{M}$ *with probability 1*

Assumption 2 also implies $\|\Sigma_s + \mu_s\mu_s^T - \Sigma\|_2 \leq \mathcal{M}$ with probability 1. Without loss of generality, $\mathcal{M} + \lambda_1 \geq 1$. We will use $\mathbb{E}[.] := \mathbb{E}_{s\in\pi}\mathbb{E}_{D(s)}[.]$ to denote the expectation over both the states as they are drawn from the stationary distribution π and over the state-specific distributions $D(.)$, unless otherwise specified.

[1] This assumption may be eliminated by observing an initial burn-in period of τ_{mix}.

[2] [38] extends Oja's algorithm to handle non-zero mean IID samples. We believe it's possible to generalize our result to this setting as well.

Define the matrix product

$$B_t := \left(I + \eta_t X_t X_t^T\right)\left(I + \eta_t X_{t-1} X_{t-1}^T\right) \ldots \left(I + \eta_1 X_1 X_1^T\right) \tag{2}$$

Unrolling the recursion in 1, the output of Oja's algorithm at timestep t is given as -

$$w_t := \frac{B_t w_0}{\|B_t w_0\|_2} \tag{3}$$

In this work, $\|.\|_2$ denotes the Euclidean L_2 norm for vectors and the operator norm for matrices unless otherwise specified.

2.1 Markov chain mixing times

Now we will discuss some well-known properties of an irreducible, aperiodic, and reversible Markov chain. We refer the reader to Chapter 4 in [19] for detailed proofs of these results. The second largest absolute eigenvalue of the Markov chain is denoted by $|\lambda_2(P)|$. We denote the state-distribution of the Markov chain at timestep t with $X_1 = x$ as $P^t(x, .)$. For any two probability distributions ν_1 and ν_2, the total variational distance between them is defined as :

$$TV(\nu_1, \nu_2) := \|\nu_1 - \nu_2\|_{TV} := \frac{1}{2}\sum_{x \in \Omega} |\nu_1(x) - \nu_2(x)|$$

The distance from stationarity at the t^{th} timestep is defined as :

$$d_{\text{mix}}(t) := \sup_{x \in \Omega} TV(P^t(x, .), \pi) \tag{4}$$

For irreducible and aperiodic Markov chains, by Theorem 4.9 in [19], we have

$$d_{\text{mix}}(t) \leq C \exp(-ct) \text{ for some } C, c > 0$$

The mixing time of the Markov chain, $\tau_{\text{mix}}(\epsilon)$ is defined as :

$$\tau_{\text{mix}}(\epsilon) := \inf\{t : d_{\text{mix}}(t) \leq \epsilon\}$$

and we will denote $\tau_{\text{mix}} := \tau_{\text{mix}}\left(\frac{1}{4}\right)$. Then, we have

$$\tau_{\text{mix}}(\epsilon) \leq \left\lceil \log_2\left(\frac{1}{\epsilon}\right) \right\rceil \tau_{\text{mix}}. \tag{5}$$

It is worth mentioning the useful relationship between $d_{\text{mix}}(.)$ and τ_{mix}, given as

$$d_{\text{mix}}(l\tau_{\text{mix}}) \leq 2^{-l} \qquad \forall l \in \mathbb{N}_0. \tag{6}$$

These results about mixing time are valid for general irreducible and aperiodic Markov chains. A Markov chain is said to be reversible if it satisfies, $\forall\, x, y \in \Omega$,

$$\pi(x) P(x, y) = \pi(y) P(y, x) \tag{7}$$

Let $\pi_{\text{min}} := \min_{x \in \Omega} \pi(x)$. For a reversible, irreducible, and aperiodic Markov chain, using Theorems 12.4 and 12.5 from [19], we note the following bound on the mixing time, $\tau_{\text{mix}}(\epsilon)$, involving the eigengap of the transition matrix, $|\lambda_2(P)|$ -

$$\frac{|\lambda_2(P)|}{1 - |\lambda_2(P)|} \ln\left(\frac{1}{2\epsilon}\right) \leq \tau_{\text{mix}}(\epsilon) \leq \frac{1}{1 - |\lambda_2(P)|} \ln\left(\frac{1}{\epsilon \pi_{\text{min}}}\right) \tag{8}$$

The gap $1 - |\lambda_2(P)|$, therefore, determines how quickly the chain mixes. Lemma 1 talks about additional properties of a reversible, irreducible, and aperiodic Markov chain.

3 Main Results

In this section, we present our main result, a near-optimal convergence rate for Oja's algorithm on Markovian data. As a corollary, we also establish a rate of convergence for Oja's algorithm applied on downsampled data, where every k^{th} data-point is considered. Supplement S.5 contains comprehensive proofs of Theorem 1 and Corollary 1 while the proof of Proposition 1 can be found in Supplement Section S.2.

Theorem 1. *Fix a $\delta \in (0,1)$ and let the step-sizes be $\eta_i := \frac{\alpha}{(\lambda_1-\lambda_2)(\beta+i)}$ with $\eta_0 \leq \frac{1}{e}, \alpha > 2$. For sufficiently large number of samples n such that $\frac{n}{\ln\left(\frac{1}{\eta_n}\right)} > \frac{\beta}{\ln\left(\frac{1}{\eta_0}\right)}$ and*

$$\beta := \frac{1000\alpha^2 \max\left\{\tau_{mix} \ln\left(\frac{1}{\eta_0}\right)(\mathcal{M}+\lambda_1)^2, \frac{\left(\frac{\mathcal{V}}{1-|\lambda_2(P)|}+\lambda_1^2\right)}{100}\right\}}{(\lambda_1-\lambda_2)^2 \ln\left(1+\frac{\delta}{200}\right)}$$

the output w_n of Oja's algorithm (1) satisfies

$$1-\left(w_n^T v_1\right)^2 \leq \frac{C\log\left(\frac{1}{\delta}\right)}{\delta^2}\left[d\left(\frac{2\beta}{n}\right)^{2\alpha} + \frac{C_1\mathcal{V}}{(\lambda_1-\lambda_2)^2(1-|\lambda_2(P)|)}\frac{1}{n} + \frac{C_2\mathcal{M}(\mathcal{M}+\lambda_1)^2}{(\lambda_1-\lambda_2)^3}\frac{\tau_{mix}\left(\eta_n^2\right)^2}{n^2}\right]$$

with probability atleast $(1-\delta)$. Here C is an absolute constant and

$$C_1 := \frac{\alpha^2(3+7|\lambda_2(P)|)}{2\alpha-1}, \;\; C_2 := \frac{35\alpha^3}{\alpha-1}$$

Next, we compare the rate of convergence proposed in Theorem 1 with the offline algorithm having access to the entire dataset $\{X_i\}_{i=1}^n$ using a recent result from [27]. Here, the authors extend the Matrix Bernstein inequality [35, 32], to Markovian random matrices. Their setup is much like ours except that the matrix at any state is fixed, i.e., there is no data distribution $D(s)$ as in our setup. However, it is easy to extend their result to our setting by observing that conditioned on the state sequence, the matrices $X_iX_i^T, i \in [n]$ are independent under our model, and we can push in the expectation over the state-specific distributions, $D(s)$, whenever required. Therefore, we have the following result -

Proposition 1 (Theorem 2.2 of [27]+Wedin's theorem)**.** *Fix $\delta \in (0,1)$. Under assumptions 1, 2, with probability $1-\delta$, the leading eigenvector $\hat{v}$ of $\hat{\Sigma} := \frac{1}{n}\sum_{i=1}^n X_iX_i^T$ satisfies*

$$1-\left(\hat{v}^T v_1\right)^2 \leq C_1'\frac{\mathcal{V}\log\left(\frac{d^{2-\frac{\pi}{4}}}{\delta}\right)}{(\lambda_1-\lambda_2)^2}\left(\frac{1+|\lambda_2(P)|}{1-|\lambda_2(P)|}\right).\frac{1}{n} + C_2'\left(\frac{\mathcal{M}\log\left(\frac{d^{2-\frac{\pi}{4}}}{\delta}\right)}{(\lambda_1-\lambda_2)(1-|\lambda_2(P)|)}\right)^2.\frac{1}{n^2} \tag{9}$$

for absolute constants C_1' and C_2'.

Observe that Theorem 1 matches the leading term $\frac{\mathcal{V}}{(\lambda_1-\lambda_2)^2(1-|\lambda_2(P)|)}$ in 1 except the $\log(d)$ term. We believe, much like the IID case (also see the remark in [14]), this logarithmic term in [27]'s result is removable for large n and a constant probability of success.

Remark 1. *(Comparison with IID algorithm) Fix a $\delta \in (0,1)$. If the data-points $\{X_i\}_{i=1}^n$ are sampled IID from the stationary distribution π, then using Theorem 4.1 from [14], we have that the output w_n of Oja's algorithm 1 satisfies -*

$$1-\left(w_n^T v_1\right)^2 \leq \frac{C\log\left(\frac{1}{\delta}\right)}{\delta^2}\left[d\left(\frac{\beta'}{n}\right)^{2\alpha} + \frac{\alpha'^2\mathcal{V}}{(2\alpha'-1)(\lambda_1-\lambda_2)^2}\frac{1}{n}\right] \tag{10}$$

We note that the leading term of Theorem 1 is worse by a factor of $\frac{1}{1-|\lambda_2(P)|}$. Further, it has an additive lower order error term $O\left(\frac{\ln^2(n)}{n^2}\right)$ which is due to the correlations between data samples in the Markov case.

Corollary 1. *(Downsampled Oja's algorithm) Fix a $\delta \in (0,1)$. If Oja's algorithm is applied on the downsampled data-stream with every k^{th} data-point, where $k := \tau_{mix}\left(\eta_n^2\right)$ then under the conditions of Theorem 1 with appropriately modified α and β, the output w_n satisfies*

$$1 - \left(w_n^T v_1\right)^2 \leq \frac{C\log\left(\frac{1}{\delta}\right)}{\delta^2}\left[d\left(\frac{2\beta\tau_{mix}\ln(n)}{n}\right)^{2\alpha} + \frac{C_1\mathcal{V}\tau_{mix}}{\left(\lambda_1-\lambda_2\right)^2}\frac{\ln(n)}{n} + \frac{C_2\mathcal{M}\left(\mathcal{M}+\lambda_1\right)^2}{\left(\lambda_1-\lambda_2\right)^3}\frac{\ln^2(n)\,\tau_{mix}\left(\eta_n^2\right)^2}{n^2}\right]$$

with probability atleast $(1-\delta)$. Here C is an absolute constant and $C_1 := \frac{30\alpha^2}{2\alpha-1}$, $C_2 := \frac{35\alpha^3}{\alpha-1}$.

Remark 2. *Data downsampling to reduce dependence amongst samples has been suggested in recent work [26, 22, 3]. In Corollary 1, we establish that the rate obtained is sub-optimal compared to Theorem 1 by a $\ln(n)$ factor. We prove this by a simple yet elegant observation: the downsampled data stream can be considered to be drawn from a Markov chain with transition kernel $P^k(.,.)$ since each data-point is k steps away from the previous one. For sufficiently large k, this implies that the mixing time of this chain is $\Theta(1)$. These new parameters are used to select the modified values of α, β according to Lemma S.12 in the Supplement.*

The proof of Theorem 1 follows the same general recipe as in [14] for obtaining a bound on the $\sin^2$ error. However, the key is to do a refined analysis for each step under the Markovian data model since the original proof technique heavily relies on the IID setting. The first step involves obtaining a high-probability bound on the $\sin^2$ error, by noting that 3 can be viewed as a single iteration of the power method on B_n. Therefore, fixing a $\delta \in (0,1)$ using Lemma 3.1 from [14], we have with probability at least $(1-\delta)$,

$$\sin^2(w_n, v_1) \leq \frac{C\log\left(\frac{1}{\delta}\right)}{\delta}\frac{\mathrm{Tr}\left(V_\perp^T B_n B_n^T V_\perp\right)}{v_1^T B_n B_n^T v_1}, \tag{11}$$

where C is an absolute constant. The numerator is bounded by first bounding its expectation (see Theorem 3) and then using Markov's inequality. To bound the denominator, similar to [14], we will use Chebyshev's inequality. Theorem 4 provides a lower bound for the expectation $\mathbb{E}\left[v_1^T B_n B_n^T v_1\right]$. Chebyshev's inequality also requires upper-bounding the variance of $\mathbb{E}\left[v_1^T B_n B_n^T v_1\right]$, which requires us to bound $\mathbb{E}\left[\left(v_1^T B_n B_n^T v_1\right)^2\right]$ (see Theorem 5).

4 Proof Idea

In this section, we provide an intuitive sketch of our proof.

4.1 Warm-up with downsampled Oja's algorithm

Let us start with the simple downsampled Oja's algorithm to build intuition. Here, one applies Oja's update rule (Eq 1) to every k^{th} data-point, for a suitably chosen k. For $k = \lceil L\tau_{\text{mix}}\log n\rceil$, the total variation distance between any consecutive data-points in the downsampled data stream is $O(n^{-L})$. As we show in Corollary 1, the error of this algorithm is similar to the error of Oja's algorithm applied to n/k data-points in the IID setting, i.e., $O(\mathcal{V}\tau_{\text{mix}}\log n/n)$.

4.2 Oja's algorithm on the entire dataset

We will take $\mathbb{E}\left[v_1^T B_n B_n^T v_1\right]$ as an example. Let us introduce some notation.

$$B_{j,i} := \left(I + \eta_j X_j X_j^T\right)\left(I + \eta_{j-1} X_{j-1} X_{j-1}^T\right)\ldots\left(I + \eta_i X_i X_i^T\right) \tag{12}$$

We will start by peeling this quantity one matrix at a time from the inside. Note that for a reversible Markov chain, standard results imply (see Lemma 1) that the mixing conditions apply to the conditional distribution of a state given another state k steps in the "future". The proof can be found in Supplement section S.3.

Lemma 1. *Consider a reversible, irreducible and aperiodic Markov chain started from the stationary distribution. Then,*

$$\frac{1}{2}\sup_{t\in\Omega}\sum_{s}\left|\mathbb{P}\left(Z_t=s|Z_{t+k}=t\right)-\pi\left(s\right)\right|=d_{mix}\left(k\right)$$

It will be helpful to explain our analysis by comparing it with the IID setting. For this reason, we will use $\mathbb{E}_{\text{IID}}[.]$ to denote the expectation under the IID data model.

$$\begin{aligned}\alpha_{n,1}:=\mathbb{E}\left[v_1^TB_nB_n^Tv_1\right]&=\mathbb{E}\left[v_1^TB_{n,2}\left(I+\eta_1\Sigma+\eta_1(X_1X_1^T-\Sigma)\right)\left(I+\eta_1\Sigma+\eta_1(X_1X_1^T-\Sigma)\right)^TB_{n,2}^Tv_1\right]\\&=\underbrace{\mathbb{E}\left[v_1^TB_{n,2}\left(I+\eta_1\Sigma\right)^2B_{n,2}^Tv_1\right]}_{\leq(1+\eta_1\lambda_1)^2\alpha_{n,2}}+2\eta_1\underbrace{\mathbb{E}\left[v_1^TB_{n,2}\left(I+\eta_1\Sigma\right)\left(X_1X_1^T-\Sigma\right)B_{n,2}^Tv_1\right]}_{T_1}\\&+\eta_1^2\underbrace{\mathbb{E}\left[v_1^TB_{n,2}\left(X_1X_1^T-\Sigma\right)^2B_{n,2}^Tv_1\right]}_{T_2}\end{aligned}\tag{13}$$

For the IID setting, the *second term is zero*, and the third term can be bounded as follows:

$$\mathbb{E}_{\text{IID}}\left[v_1^TB_{n,2}\left(X_1X_1^T-\Sigma\right)^2B_{n,2}^Tv_1\right]=\mathbb{E}_{\text{IID}}\left[v_1^TB_{n,2}\mathbb{E}\left[\left(X_1X_1^T-\Sigma\right)^2\right]B_{n,2}^Tv_1\right]\leq\mathcal{V}\mathbb{E}_{\text{IID}}\left[v_1^TB_{n,2}B_{n,2}^Tv_1\right]$$

Let us denote the IID version of $\alpha_{n,i}$ by $\alpha_{n,i}^{\text{IID}}=\mathbb{E}_{\text{IID}}[v_1^TB_{n,i}B_{n,i}^Tv_1]$. The final recursion for the IID case becomes:

$$\alpha_{n,1}^{\text{IID}}\leq\left(1+2\eta_1\lambda_1+\eta_1^2\left(\lambda_1+\mathcal{V}\right)\right)\alpha_{n,1}^{\text{IID}}$$

So, for our Markovian data model, the hope is that the cross term T_1 (which has a multiplicative factor of η_1) is $O(\eta_1)$ and T_2 is $O(\eta_1^2)$. We will start with the T_1 term, which is zero in the IID setting.

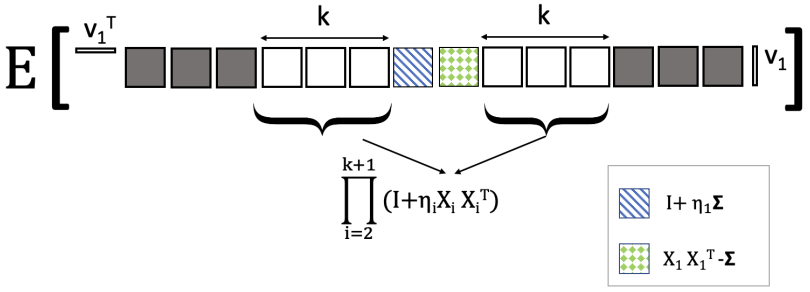


Figure 2: If the identity matrix could replace the intermediate products (white matrices), we would be able to use the fact that the conditional expectation of the noise matrix $X_1X_1^T-\Sigma$ conditioned on the grey matrices is nearly zero.

4.3 Approximation - take one

We hope to reduce the product $B_{n,2}(X_1X_1^T-\Sigma)$ into a product of nearly independent matrices. One hope is that if instead of $B_{n,2}$, we had $B_{n,2+k}$ for some suitably large integer k, then using (reverse) mixing properties of the Markov chain, we could argue using Lemma 1 that $\mathbb{E}[X_1X_1^T-\Sigma|s_{1+k},\ldots,s_n]$ is very close to zero. See Figure 2. The first question we start with is: if we can

replace $\prod_{i=2}^{k+1}(I+\eta_i X_i X_i^T)$ by the identity matrix. Our first approximation shows that this is indeed possible *if* $k\eta_1(\mathcal{M}+\lambda_1)$ is small enough.

This leads to the question: how big should k be? Let us expand out $\prod_{i=2}^{k+1}(I+\eta_i X_i X_i^T)$. We have:

$$\prod_{i=2}^{k+1}(I+\eta_i X_i X_i^T) = I + \sum_{i=2}^{k+1}\eta_i X_i X_i^T + \sum_{2\le i<j\le k+1}\eta_i\eta_j (X_i X_i^T)(X_j X_j^T) + \dots \tag{14}$$

Note that the operator norm of the j^{th} term in this expansion can be bounded by $O\left((k\eta_1(\mathcal{M}+\lambda_1))^j\right)$ (I being the 0^{th} term) using assumption 2. We are also using the fact that we will use a decaying learning rate, i.e. $\eta_1 \ge \eta_2 \ge \dots$. Thus, as long as $\eta_1 k(\mathcal{M}+\lambda_1)$ is sufficiently small, indeed,

$$\left|\prod_{i=2}^{k+1}(I+\eta_i X_i X_i^T) - I\right| = O\left(\eta_1 k(\mathcal{M}+\lambda_1)\right)$$

Observe that, Eq 15 of Lemma 2 establishes that the $\prod_{i=2}^{k+1}(I+\eta_i X_i X_i^T)$ can be approximated by I, as long as $\eta_1 k$ is small. Since this is a recursive argument, we would need $\eta_i k$ to be small for $i = 1, \dots n$, which is satisfied by the strong condition $\eta_1 k$ is small. However, this seems wasteful. If all we need is the former condition, should k not also be chosen adaptively? We set $k_i = d_{\text{mix}}(\eta_i^2)$ (see definition in Eq 4). This particular choice of k_i allows us to push the error resulting from the total variation distance to a smaller order term (see Supplement Section S.4 for the detailed proof). Now we present the lemma which formally bounds the deviation of the $k-$ length matrix product from identity.

Lemma 2. *If* $\forall i \in [n], \eta_i k_i (\mathcal{M}+\lambda_1) \le \epsilon, \epsilon \in (0,1)$ *and* η_i *forms a non-increasing sequence then* $\forall\, m \le n - k_n$,

$$\left\|B_{m+k_m-1,m} - I\right\|_2 \le (1+\epsilon)\, k_m \eta_m (\mathcal{M}+\lambda_1) \text{ and} \tag{15}$$

$$\left\|B_{m+k_m-1,m} - I - \sum_{t=m}^{m+k_m-1}\eta_t X_t X_t^T\right\|_2 \le k_m^2\eta_m^2 (\mathcal{M}+\lambda_1)^2 \tag{16}$$

Lemma 2 bounds the norm of the matrix product $B_{t+k_t-1,t}$ at two levels. The first result provides a coarse bound, approximating linear and higher-order terms. The second result provides a finer bound, preserving the linear term and approximating quadratic and higher-order terms. The proofs involve a straightforward combinatorial expansion of $B_{t+k_t-1,t}$ and are deferred to the Supplement. Both of these results play an important role at different stages.

However, unfortunately, this approximation gives us the *sub-optimal rate* of downsampled Oja's algorithm. In hindsight, this is not unexpected since this analysis completely removes all the in-between matrices. The question is how bad is the approximation in Lemma 2 Eq 15. This investigation brings us to the second take.

$$\begin{aligned}
T_1 :=& \mathbb{E}\left[v_1^T B_{n,2}\left(I+\eta_1\Sigma\right)\left(X_1X_1^T-\Sigma\right)B_{n,2}^T v_1\right] \\
&\le \mathbb{E}\left[v_1^T B_{n,k+2}\left(I+\sum_{j=2}^{k+1}\eta_i X_j X_j^T\right)(I+\eta_1\Sigma)\left(X_1X_1^T-\Sigma\right)\left(I+\sum_{j=2}^{k+1}\eta_j X_j X_j^T\right)B_{n,k+2}^T v_1\right] \\
&\quad + O(\eta_1^2 k_1^2)\alpha_{n,k+2} \\
&= \sum_{j=2}^{k+1}\mathbb{E}\left[v_1^T B_{n,k+2}\left(\eta_j X_j X_j^T\right)(I+\eta_1\Sigma)\left(X_1X_1-\Sigma\right)B_{n,k+2}^T v_1\right] + O(\eta_1^2k_1^2)\alpha_{n,k+2}
\end{aligned} \tag{17}$$

$$= \sum_{i=2}^{k+1} \mathbb{E}\left[v_1^T B_{n,k+2} \left(\eta_j X_j X_j^T\right) (I + \eta_1 \Sigma) (X_1 X_1 - \Sigma) B_{n,k+2}^T v_1\right] + O(\eta_1^2 k_1^2)\alpha_{n,k+2}$$

$$= \sum_{j=2}^{k+1} \eta_j \mathbb{E}\left[v_1^T B_{n,k+2} \underbrace{\mathbb{E}\left[\left(X_j X_j^T\right)(I + \eta_1 \Sigma)(X_1 X_1 - \Sigma) | X_{k+2}, \ldots, X_n\right]}_{T_{1,j}} B_{n,k+2}^T v_1\right]$$
$$+ O(\eta_1^2 k_1^2)\alpha_{n,k+2} \tag{18}$$

Now the question we ask is "Is the $T_{1,j}$ term really $O(1)$?" Indeed, naively it can be bounded as $O(1)$. But we will use a delicate analysis of this to show that it is much smaller.

4.4 Approximation: take two

Recall that, approximating the matrix product in Eq 14 by the 0^{th} term, i.e. the identity incurs a $O(\eta_1 k_1)$ error. However, if we approximate it by the identity plus the linear term, the error would be $O(\eta_1 k_1)^2$. The question is, by including the linear term, can we get a sharper bound?

In the following lemma, we will establish that, indeed, $T_{1,j}$ has a much smaller norm. The novelty of our bound is not just in using the mixing properties of the Markov chain but also in teasing out the variance parameter $\mathcal{V}$. We will state the lemma, in a slightly more general form as -

Lemma 3. *For $i < j \leq i + k_i$,*

$$\left\|\mathbb{E}\left[\left(X_i X_i^T - \Sigma\right) S X_j X_j^T | s_{i+k_i}, \ldots s_n\right]\right\|_2 \leq \left(|\lambda_2(P)|^{j-i} \mathcal{V} + 8\eta_i^2 \mathcal{M}(\mathcal{M} + \lambda_1)\right) \|S\|_2$$

where k_i is as defined in Lemma S.12 and S is a constant symmetric positive semi-definite matrix.

Lemma 3 bounds the norm of the covariance between matrices $\left(X_i X_i^T - \Sigma\right) S$ and $X_j X_j^T$. In particular, this implies that the norm of $T_{1,j}$ decays as $|\lambda_2(P)|^{j-1}$.

4.4.1 Proof sketch of Lemma 3 with $S = I$ and $k_i = k$

Here we give a short sketch of the proof with $S = I$ for simplicity of exposition. Define

$$G(\ell) := \mathbb{E}\left[X_i X_i^T - \Sigma | s_i = \ell\right] = \Sigma_l + \mu_l \mu_l^T - \Sigma$$

If X_i and X_j were sampled IID from the stationary distribution π then, since $\mathbb{E}_{\text{IID}}\left[X_i X_i^T\right] = \Sigma$,

$$\mathbb{E}_{\text{IID}}\left[(X_i X_i^T - \Sigma) X_j X_j^T | s_{i+k_i}, \ldots s_n\right] = \mathbb{E}_{\text{IID}}\left[(X_i X_i^T - \Sigma) X_j X_j^T\right] = \mathbb{E}_{\text{IID}}\left[X_i X_i^T - \Sigma\right] \mathbb{E}_{\text{IID}}\left[X_j X_j^T\right] = 0$$

For the Markov case, intuitively from Eq 6, we know that X_i and X_j are approximately independent if sampled at distant timesteps in the Markovian data stream. This notion is formalized in Lemma 3, which shows that the covariance norm decreases geometrically with $(j - i)$. We sketch a proof here and defer the details to Supplement section S.4. Note that for a Markov chain, the Markov property holds in reverse (see Lemma S.6 in Supplement) i.e $P(s_t | s_{t+1}, s_{t+2} \ldots s_n) = P(s_t | s_{t+1})$. Therefore,

$$\mathbb{E}\left[(X_i X_i^T - \Sigma) X_j X_j^T | s_{i+k, \ldots s_n}\right] = \mathbb{E}\left[(X_i X_i^T - \Sigma)(X_j X_j^T - \Sigma) | s_{i+k}\right] + \mathbb{E}\left[X_i X_i^T - \Sigma | s_{i+k}\right] \Sigma.$$

Then,

$$\mathbb{E}\left[X_i X_i^T - \Sigma | s_{i+k_i} = x_0\right] = \sum_x \pi(x) G(x) + \sum_{x \in \Omega} \left(P^{k_i}(x_0, x) - \pi(x)\right) G(x)$$

and

$$\mathbb{E}\left[(X_i X_i^T - \Sigma)(X_j X_j^T - \Sigma) | s_{i+k_i} = x_0\right] = \mathbb{E}\left[\left(X_i X_i^T - \Sigma\right)\left(X_j X_j^T - \Sigma\right)\right]$$
$$+ \sum_{x,y \in \Omega} \left(P^{j-i}(x, y) - \pi(y)\right)\left(P^{i+k_i-j}(x_0, x) - \pi(x)\right) G(x) G(y)$$

Since the Markov chain starts in stationarity, $\mathbb{E}\left[X_i X_i^T - \Sigma\right] = 0$. For $\mathbb{E}\left[\left(X_i X_i^T - \Sigma\right)\left(X_j X_j^T - \Sigma\right)\right]$, we note that conditioned on the states $s_i = x, s_j = y$, the expectations over of the state-specific distributions, $D(.)$ can be pushed inside.

$$\begin{aligned}\mathbb{E}\left[\left(X_i X_i^T - \Sigma\right)\left(X_j X_j^T - \Sigma\right)\right] &= \sum_{x,y\in\Omega} \left(P^{j-i}(y,x) - \pi(x)\right)\pi(y)\, B(x)B(y) \\ &= \sum_{x,y\in\Omega} \underbrace{\frac{\left(P^{j-i}(y,x) - \pi(x)\right)}{\sqrt{\pi(x)}}\sqrt{\pi(y)}}_{Q_{yx}} \left(\sqrt{\pi(x)}B(x)\right)\left(\sqrt{\pi(y)}B(y)^T\right)\end{aligned}$$

Let $\Pi := \mathrm{diag}(\pi) \in \mathbb{R}^{\Omega\times\Omega}$. Using spectral theory, we can show that the matrix $Q := \Pi^{\frac{1}{2}}\left(P^t - \mathbb{1}\mathbb{1}^T\Pi\right)\Pi^{-\frac{1}{2}}$ has operator norm bounded by $|\lambda_2(P)|^{j-i}$. Consequently we can show that

$$\mathbb{E}\left[\left(X_i X_i^T - \Sigma\right)\left(X_j X_j^T - \Sigma\right)\right] \leq \mathcal{V}|\lambda_2(P)|^{j-i}$$

Note that for $i = j$, this boils down to $\mathbb{E}\left[\left(X_i X_i^T - \Sigma\right)^2\right] \leq \mathcal{V}$, which is ensured by Assumption 1. It can be shown that the remaining terms involving $\left(P^t(.,x) - \pi(x)\right)$ are of a smaller order. This is achieved by using the mixing properties of the Markov chain, specifically using bounds on the total variation distance (4), the uniform norm bound (2) on $\left(X_i X_i^T - \Sigma\right)$ and the fact that $\|\Sigma\|_2 \leq \lambda_1$ and the details can be found in Supplement section S.4.

Let $\{c_1, c_2, c_3, c_4\}$ be positive constants for ease of notation. Coming back to Eq 13, we can bound T_1 as follows:

$$T_1 \leq \alpha_{n,k+2}\left(\eta_1 \frac{c_1|\lambda_2(P)|\mathcal{V}}{1-|\lambda_2(P)|} + c_2\eta_1^2 k_1^2\right),$$

A similar argument can be applied to bound T_2 as follows -

$$T_2 \leq \alpha_{n,k+2}\left(\mathcal{V} + c_3\eta_1 k_1^2\right)$$

Putting everything together in 13, we have

$$\alpha_{n,1} \leq \underbrace{\left((1+\eta_t\lambda_1)^2 + \mathcal{V}\right)\alpha_{n,2}}_{\text{Recursion for IID setting}} + \underbrace{\left(\frac{c_1|\lambda_2(P)|}{1-|\lambda_2(P)|}\right)\mathcal{V}\eta_t^2\alpha_{n,k+2}}_{\text{Error due to Markovian dependence}} + \underbrace{c_4\eta_1^3 k_1^2\alpha_{n,k+2}}_{\text{Error due to approximation of matrix product}} \tag{19}$$

Recursing on this inequality gives us our bound on $\mathbb{E}\left[v_1^T B_n B_n^T v_1\right]$, stated formally in Theorem 2. With the intuition of our proof techniques built in this section, we are now ready to present all our accompanying theorems.

5 Convergence Analysis of Oja's Algorithm for Markovian Data

In this section, we present our accompanying theorems which are used to obtain the main result in Theorem 1. But before doing so, we will need to establish some notation. Let $k_i := \tau_{\text{mix}}\left(\eta_i^2\right)$, and the step-sizes be set as $\eta_i := \frac{\alpha}{(\lambda_1-\lambda_2)(\beta+i)}$ with α, β as defined in Theorem 1. Let $\epsilon := \frac{1}{100}$. As shown in Lemma S.12 in Supplement Section S.3 our choice of step-sizes satisfy, $\forall i \in [n]$,

1. $\eta_i k_i\left(\mathcal{M} + \lambda_1\right) \leq \epsilon$
2. $\eta_i \leq \eta_{i-k_i} \leq (1+2\epsilon)\eta_i \leq 2\eta_i$ (slow-decay)

Further, we define scalar variables -

$$r := 2\left(1+\epsilon\right)k_n\eta_n\left(\mathcal{M}+\lambda_1\right), \qquad \zeta_{k,t} := 40k_{t+1}\left(\mathcal{M}+\lambda_1\right)^2$$
$$\psi_{k,t} := 6\mathcal{M}\left[1+3k_{t+1}^2\left(\mathcal{M}+\lambda_1\right)^2\right], \qquad \mathcal{V}' := \frac{1+\left(3+4\epsilon\right)\left|\lambda_2\left(P\right)\right|}{1-\left|\lambda_2\left(P\right)\right|}\mathcal{V} \tag{20}$$

and recall the definitions of B_t and $B_{j,i}$ in Eqs 2 and 12, respectively. Then, under assumptions 1 and 2, we state our results below, and draw parallels with the results in the IID setting proved in [14] (Lemmas 5.1,5.2,5.3 and 5.4).

We are now ready to present the theoretical results needed to prove our main result. For simplicity of notation, we present versions of the results by using $\eta_i := \frac{\alpha}{(\lambda_1-\lambda_2)(\beta+i)}$ with α, β as defined in Theorem 1. However, these theorems are in fact valid under more general step size schedules. We state and prove the more general version in the Supplement Section S.4.

Theorem 2.

$$\mathbb{E}\left[v_1^T B_n B_n^T v_1\right] \leq \left(1+r\right)^2 \exp\left(\sum_{t=1}^{n-k_n}\left(2\eta_t\lambda_1 + \eta_t^2\left(\mathcal{V}'+\lambda_1^2\right) + \eta_t^3\psi_{k,t}\right)\right)$$

We notice three primary differences with the IID case here. The first is the $\left(1+r\right)^2$ term. This occurs since the recursion in 19 leaves out the last k_n terms which are bounded by $\left(1+r\right)^2$. The second is the presence of a factor of $\frac{1}{1-\left|\lambda_2(P)\right|}$ with $\mathcal{V}$ which occurs due to the Markovian dependence between terms and the use of mixing properties of the chain. Finally, we notice an extra lower order term $\eta_t^3\psi_{k,t}$. This is a result of our approximation of the matrix product while conditioning and using mixing.

Theorem 3. *Let* $u := \min\left\{t : t \in [n], t-k_t \geq 0\right\}$*, then,*

$$\mathbb{E}\left[\mathrm{Tr}\left(V_\perp^T B_n B_n^T V_\perp\right)\right] \leq \left(1+5\epsilon\right)\exp\left(\sum_{t=u+1}^{n} 2\eta_t\lambda_2 + \eta_{t-k_t}^2\left(\mathcal{V}'+\lambda_1^2\right) + \eta_{t-k_t}^3\psi_{k,t}\right)$$
$$\times\left(d + \sum_{t=u+1}^{n}\left(\mathcal{V}'+\eta_t\psi_{k,t}\right)C'_{k,t}\eta_{t-k_t}^2\exp\left(\sum_{i=u+1}^{t} 2\eta_i\left(\lambda_1-\lambda_2\right)\right)\right)$$

where $C'_{k,t} := \left(1+\frac{\delta}{200}\right)\exp\left(2\lambda_1\sum_{i=1}^{u}\eta_j\right)$

In this case, the first point of difference is the variable $u := \min\left\{t : t \in [n], t-k_t \geq 0\right\}$. This is again a result of conditioning and the constraints of our recursion. The difference from Theorem 2 arises because Theorem 3 uses conditioning on the past, whereas the former uses conditioning on the future. The factor $\left(1+5\epsilon\right)$ represents the approximation of the first u terms. The other differences regarding $\mathcal{V}'$ and $\psi_{k,t}$ remain the same as in the case of Theorem 2.

Theorem 4.

$$\mathbb{E}\left[v_1^T B_n B_n^T v_1\right] \geq \left(1-s\right)\exp\left(\sum_{t=1}^{n-k_n} 2\eta_t\lambda_1 - \sum_{t=1}^{n-k_n} 4\eta_t^2\lambda_1^2\right)$$

where $s := 2r + \frac{\delta}{1000}$

Here the bound differs by a multiplicative factor of $\left(1-s\right)$ from its IID counterpart. Furthermore, the sums go up to $\left(n-k_n\right)$ again because of the constraints of the recursion and conditioning. Note that for sufficiently large n as is proved in the Supplement Lemma S.13, $r = O\left(\frac{\ln(n)}{n}\right) \to 0$ and $\delta \in (0,1)$. Therefore, $\left(1-s\right) \approx 1$ as n increases.

Theorem 5.

$$\mathbb{E}\left[\left(v_1^T B_n B_n^T v_1\right)^2\right] \leq \left(1+r\right)^4\exp\left(\sum_{t=1}^{n-k_n} 4\eta_t\lambda_1 + \sum_{t=1}^{n-k_n}\eta_t^2\zeta_{k,t}\right)$$

Finally, this bound again exhibits the same patterns as the previous theorems involving v_1. An interesting point of difference with the IID counterpart is the absence of $\mathcal{V}$ in the bound. This is due to the use of the coarse approximation discussed in Section 4.3, which suffices in this case for our main result in Theorem 1.

Having established these results, the final step is to plug them into Eq 11 and follow the proof recipe described earlier. This step is straightforward but requires significant calculations, and is therefore deferred to the Supplement Section S.5.

6 Experimental Validation

In this section, we present experiments conducted to validate the results presented in Section 3. We design a Markov chain with $|\Omega| = 10$ states, where each state has a probability of $(1-\rho)$ to remain in the same state and $\frac{\rho}{|\Omega|-1}$ probability of transitioning to any of the remaining states, for $\rho \in (0,1)$. The parameter ρ controls the ease of exploration. Smaller values of ρ make it harder to explore new states leading to a longer time needed to mix. It can be verified that the stationary distribution $\pi = \mathcal{U}(\Omega)$ is uniform over the state-space and $|\lambda_2(P)| \approx (1-\rho)$. We set $\rho = 0.2$ for figures 1, 3a and 3b, and vary it in figure 3c.

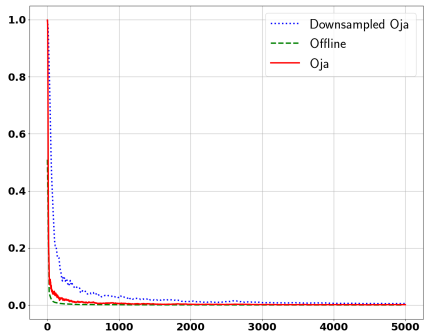
Each state $s \in \Omega$ is associated with a zero-mean distribution $D(s)$ over $d = 1000$ dimensional data-points having a covariance matrix Σ_s with $\Sigma_s(i,j) = \exp(-|i-j|c_s)\,\sigma_i\sigma_j$ where $c_s := 1 + 9\left(\frac{s-1}{|\Omega|-1}\right), \sigma_i := 5i^{-\beta}$. The eigengap of the true covariance matrix $\Sigma = \sum_{i\in\Omega}\pi(i)\,\Sigma_i = \frac{1}{|\Omega|}\sum_{i\in\Omega}\Sigma_i$ for this construction, with $\beta = 1.0$, is $\lambda_1 - \lambda_2 \approx 20$. We set $\beta = 1.0$ for figures 1, 3a and 3c, and vary it in figure 3b.

During the random-walk, for each state s_i, we draw IID samples $Z_i \in \mathbb{R}^d$ from either the Bernoulli distribution (parameter p, figures 1 and 3c) or the Uniform distribution ($\mathcal{U}\left(-\sqrt{3},\sqrt{3}\right)$, figure 3a). The parameter, p, of the Bernoulli distribution is fixed at the beginning of the experiment as $p \sim \mathcal{U}(0, 0.05)$. We normalize Z_i such that all components have zero mean and unit variance, if not already so. We then generate the sample data-point for PCA as $X_i = \Sigma_i^{\frac{1}{2}} Z_i$. By construction, $\mathbb{E}_{D(s_i)}[X_i] = 0^d$ and $\mathbb{E}_{D(s_i)}\left[X_i X_i^T\right] = \Sigma_i$. The step sizes for Oja's algorithm are set as $\eta_i = \frac{\alpha}{(\beta+i)(\lambda_1-\lambda_2)}$ for $\alpha = 5, \beta = \frac{5}{1-|\lambda_2(P)|}$. For downsampled Oja's algorithm, every 10^{th} data-point is considered and β is accordingly divided by 10. Each plot shows the average over 20 random walks.

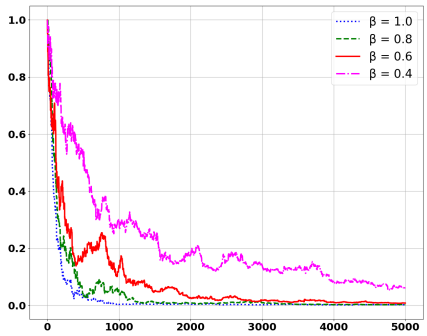
Figures 1, and 3a provide a comparison of the performance of different algorithms for various state distributions. Here, we are checking if the results obtained in Theorem 1, Proposition 1, and Corollary 1 are reflected in the experiments. The experimental results demonstrate that Oja's algorithm performs significantly better than the downsampled version, consistent with the theoretical results. Figure 1 shows that data downsampling can in fact be much worse. Figure 3a shows a faster rate of convergence for the case of the uniform distribution since the Bernoulli distribution generates fewer non-zero data-points. Figure 3b compares the performance of Oja's algorithm for different covariance matrices. Smaller values of β decrease the eigengap, $\lambda_1 - \lambda_2$, and hence lead to a slower convergence. Figure 3c confirms that smaller values of ρ (larger values of $|\lambda_2(P)|$) make the problem more challenging.

7 Conclusion

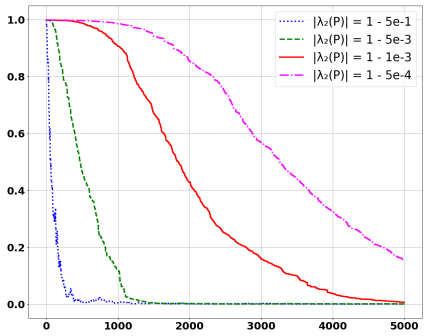
We have considered the problem of streaming PCA for Markovian data, which has implications in various settings like decentralized optimization, reinforcement learning, etc. The analysis of streaming algorithms in such settings has seen a renewed surge of interest in recent years. However, the dependence between data-points makes it difficult to obtain sharp bounds. We provide, to our knowledge, the first sharp bound for obtaining the first principal component from a Markovian data stream that breaks the logarithmic barrier present in the analysis done for downsampled data. We believe that the theoretical tools that we have developed in this paper would enable one to obtain



(a) Uniform state distributions



(b) Variation of convergence rate with β



(c) Variation of convergence rate with $|\lambda_2(P)|$

Figure 3: Experimental validation of main results: X axis represents the sample size, and Y axis represents the $\sin^2$ error.

sharp bounds for other dependent data settings, learning top k principal components, and online inference algorithms with updates involving products of matrices.

8 Acknowledgements

We gratefully acknowledge NSF grants 2217069 and DMS 2109155. We are also grateful to Rachel Ward and Bobby Shi for valuable discussions.

References

[1] Zeyuan Allen-Zhu and Yuanzhi Li. First efficient convergence for streaming k-pca: a global, gap-free, and near-optimal rate. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 487–492. IEEE, 2017.

[2] Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. *CoRR*, abs/1806.02450, 2018.

[3] Minshuo Chen, Lin Yang, Mengdi Wang, and Tuo Zhao. Dimensionality reduction for stationary time series via stochastic nonconvex optimization. *Advances in Neural Information Processing Systems*, 31, 2018.

[4] Shuhang Chen, Adithya Devraj, Ana Busic, and Sean Meyn. Explicit mean-square error bounds for monte-carlo and linear stochastic approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 4173–4183. PMLR, 2020.

[5] Thinh T. Doan, Lam M. Nguyen, Nhan H. Pham, and Justin Romberg. Convergence rates of accelerated markov gradient descent with applications in reinforcement learning, 2020.

[6] Thinh T. Doan, Lam M. Nguyen, Nhan H. Pham, and Justin Romberg. Finite-time analysis of stochastic gradient descent under markov randomness, 2020.

[7] Ron Dorfman and Kfir Yehuda Levy. Adapting to mixing time in stochastic optimization with markovian data. In *International Conference on Machine Learning*, pages 5429–5446. PMLR, 2022.

[8] John C. Duchi, Alekh Agarwal, Mikael Johansson, and Michael I. Jordan. Ergodic mirror descent. *SIAM J. Optim.*, 22(4):1549–1578, 2012.

[9] Alain Durmus, Eric Moulines, Alexey Naumov, Sergey Samsonov, and Hoi-To Wai. On the stability of random matrix product with markovian noise: Application to linear stochastic approximation and td learning. In *Conference on Learning Theory*, pages 1711–1752. PMLR, 2021.

[10] Mathieu Even. Stochastic gradient descent under markovian sampling schemes, 2023.

[11] Donald O. Hebb. *The organization of behavior: A neuropsychological theory*. Wiley, New York, June 1949.

[12] Amelia Henriksen and Rachel Ward. AdaOja: Adaptive Learning Rates for Streaming PCA. *arXiv e-prints*, page arXiv:1905.12115, May 2019.

[13] De Huang, Jonathan Niles-Weed, and Rachel Ward. Streaming k-pca: Efficient guarantees for oja's algorithm, beyond rank-one updates. *CoRR*, abs/2102.03646, 2021.

[14] Prateek Jain, Chi Jin, Sham Kakade, Praneeth Netrapalli, and Aaron Sidford. Streaming pca: Matching matrix bernstein and near-optimal finite sample guarantees for oja's algorithm. In *Proceedings of The 29th Conference on Learning Theory (COLT)*, June 2016.

[15] Chi Jin, Sham M Kakade, Cameron Musco, Praneeth Netrapalli, and Aaron Sidford. Robust shift-and-invert preconditioning: Faster and more sample efficient algorithms for eigenvector computation. *arXiv preprint arXiv:1510.08896*, 2015.

[16] I Jolliffe. Myilibrary, principal component analysis, vol. 2, 2002.

[17] László Kozma. Inequalities cheat sheet, 2018. PDF file.

[18] Harold J Kushner and G George Yin. Applications in signal processing, communications, and adaptive control. *Stochastic Approximation and Recursive Algorithms and Applications*, pages 63–93, 2003.

[19] David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

[20] Robert Lunde, Purnamrita Sarkar, and Rachel Ward. Bootstrapping the error of oja's algorithm. *Advances in Neural Information Processing Systems*, 34:6240–6252, 2021.

[21] Hanbaek Lyu, Deanna Needell, and Laura Balzano. Online matrix factorization for Markovian data and applications to Network Dictionary Learning. *arXiv e-prints*, page arXiv:1911.01931, November 2019.

[22] Shaocong Ma, Ziyi Chen, Yi Zhou, Kaiyi Ji, and Yingbin Liang. Data sampling affects the complexity of online sgd over dependent data. In *Uncertainty in Artificial Intelligence*, pages 1296–1305. PMLR, 2022.

[23] Abdelkader Mokkadem. Mixing properties of arma processes. *Stochastic Processes and their Applications*, 29(2):309–315, 1988.

[24] Jean-Marie Monnez. Stochastic approximation of eigenvectors and eigenvalues of the q-symmetric expectation of a random matrix. *Communications in Statistics-Theory and Methods*, pages 1–15, 2022.

[25] Nikos Mouzakis and Eric Price. Spectral guarantees for adversarial streaming pca, 2022.

[26] Dheeraj Nagaraj, Xian Wu, Guy Bresler, Prateek Jain, and Praneeth Netrapalli. Least squares regression with markovian data: Fundamental limits and algorithms. *Advances in neural information processing systems*, 33:16666–16676, 2020.

[27] Joe Neeman, Bobby Shi, and Rachel Ward. Concentration inequalities for sums of markov dependent random matrices, 2023.

[28] Erkki Oja. Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15(3):267–273, November 1982.

[29] Erkki Oja and Juha Karhunen. On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix. *Journal of mathematical analysis and applications*, 106(1):69–84, 1985.

[30] Rayadurgam Srikant and Lei Ying. Finite-time error bounds for linear stochastic approximation andtd learning. In *Conference on Learning Theory*, pages 2803–2830. PMLR, 2019.

[31] Tao Sun, Yuejiao Sun, and Wotao Yin. On markov chain gradient descent. *Advances in neural information processing systems*, 31, 2018.

[32] Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12:389–434, 2012.

[33] Lan V Truong. Generalization error bounds on deep learning with markov datasets. *Advances in Neural Information Processing Systems*, 35:23452–23462, 2022.

[34] John Tsitsiklis and Benjamin Van Roy. Analysis of temporal-diffference learning with function approximation. In M.C. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9. MIT Press, 1996.

[35] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[36] Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12:99–111, 1972.

[37] Puyudi Yang, Cho-Jui Hsieh, and Jane-Ling Wang. History pca: A new algorithm for streaming pca. *arXiv preprint arXiv:1802.05447*, 2018.

[38] Siyun Zhou and Yanqin Bai. Convergence analysis of oja's iteration for solving online pca with nonzero-mean samples. *Science China Mathematics*, 64:849–868, 2021.

[39] Ingvar Ziemann and Stephen Tu. Learning with little mixing. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 4626–4637. Curran Associates, Inc., 2022.

S.1 Notation

For conciseness, we define the stochastic function $A : \Omega \to \mathbb{R}^{d\times d}$ which maps each state variable of the Markov chain to a $(d \times d)$ positive semi-definite symmetric matrix as

$$A\left(s_t\right) := X_t X_t^T$$

where $X_t \sim D\left(s_t\right)$ is drawn from the the distribution corresponding to the state at timestep s_t.

S.2 Offline PCA with Markovian Data

In this section, we prove Proposition 1. We note that [27] considers $F_j\left(s_j\right)$ to be random only with respect to the states. Therefore, we first show that their results generalize to our setting as well, using $F_j\left(s_j\right) := A\left(s_j\right) - \Sigma$. From Eq (5) in [27], we have

$$\begin{aligned}\left\|\prod_{j=1}^{n}\exp\left(\frac{\theta}{2}\left(A\left(s_j\right)-\Sigma\right)\right)\right\|_F^2 &= \operatorname{Tr}\left(\prod_{j=1}^{n}\exp\left(\frac{\theta}{2}\left(A\left(s_j\right)-\Sigma\right)\right)\prod_{j=n}^{1}\exp\left(\frac{\theta}{2}\left(A\left(s_j\right)-\Sigma\right)\right)\right)\\ &= \operatorname{vec}\left(I_d\right)^T\left(\prod_{j=1}^{n}\exp\left(\theta H\left(s_j\right)\right)\right)\operatorname{vec}\left(I_d\right)\end{aligned}$$

where $H\left(s_j\right) := \frac{1}{2}\left[\left(A\left(s_j\right)-\Sigma\right)\otimes I_d + I_d \otimes \left(A\left(s_j\right)-\Sigma\right)\right]$. Noting that conditioned on the state sequence, the matrices $A\left(s_i\right), i \in [n]$ are independent under our model, we can push in the expectation over the state-specific distributions inside. Let $\mathbb{E}_\pi$ denote the expectation over the stationary state-sequence of the Markov chain, and $\mathbb{E}_D$ denote the distribution over states. Therefore,

$$\mathbb{E}_\pi\mathbb{E}_D\left[\left\|\prod_{j=1}^{n}\exp\left(\frac{\theta}{2}\left(A\left(s_j\right)-\Sigma\right)\right)\right\|_F^2\right] = \mathbb{E}_\pi\left[\operatorname{vec}\left(I_d\right)^T\left(\prod_{j=1}^{n}\mathbb{E}_{D(s_j)}\left[\exp\left(\theta H\left(s_j\right)\right)\right]\right)\operatorname{vec}\left(I_d\right)\right]$$

Defining the multiplication operator $\left(E_j^\theta \mathbf{h}\right)(x) = \mathbb{E}_{D(x)}\left[\exp\left(\theta H_j\left(x\right)\right)\right]\mathbf{h}\left(x\right)$ for any vector-valued function $\mathbf{h}$, we note that Eq (8) from [27] holds for our case as well.

Next, we adapt Proposition 5.3 from [27] for our setting. Specifically, we have the following lemma -

Lemma S.1. *Consider the operator* $H\left(x\right) := \frac{1}{2}\left[\left(A\left(x\right)-\Sigma\right)\otimes I_d + I_d\otimes\left(A\left(x\right)-\Sigma\right)\right]$. *Then, under assumptions 2 and 1 and the definition of* Σ*, we have,*

1. $\mathbb{E}_\pi\mathbb{E}_{D(x)}\left[H\left(x\right)\right] = 0$
2. $H\left(x\right) \preceq \mathcal{M}I$
3. $\left\|\mathbb{E}_\pi\mathbb{E}_{D(x)}\left[H\left(x\right)^2\right]\right\|_2 \leq \mathcal{V}$

Proof. The proof follows by using the same arguments as Proposition 5.3 from [27] and using the expectation $\mathbb{E}_\pi\mathbb{E}_{D(x)}$ over both the state sequence and the distribution over states, along with assumptions 2 and 1. □

Finally, to prove Bernstein's inequality, we prove that Lemma 6.7 from [27] holds for our case. To

note this, we start with equation (57) in their work. We have, using Lemma S.1,

$$\begin{aligned}
\left|\left\langle v_2, \mathbb{E}_\pi \mathbb{E}_{D(x)} \left[\exp\left(\theta H\left(x\right)\right)\right] v_1 \right\rangle\right| &= \left|\left\langle v_2, \mathbb{E}_\pi \mathbb{E}_{D(x)} \left[\exp\left(\theta H\left(x\right)\right)\right] v_1 \right\rangle\right| \\
&= \left|\left\langle v_2, \left(I + \mathbb{E}_\pi \mathbb{E}_{D(x)}\left[H\left(x\right)\right] + \sum_{k=2}^{\infty} \frac{\theta^k}{k!} \mathbb{E}_\pi \mathbb{E}_{D(x)} \left[H\left(x\right)^k\right]\right) v_1 \right\rangle\right| \\
&= \left|\left\langle v_2, v_1 \right\rangle + \left\langle v_2 \left(\sum_{k=2}^{\infty} \frac{\theta^k}{k!} \mathbb{E}_\pi \mathbb{E}_{D(x)} \left[H\left(x\right)^k\right]\right) v_1 \right\rangle\right| \\
&\leq \left|\left\langle v_2, v_1 \right\rangle\right| \left(1 + \mathcal{V}\left(\sum_{k=2}^{\infty} \frac{\theta^k}{k!} \mathcal{M}^{k-2}\right)\right)
\end{aligned}$$

Therefore, Eq (60) from [27] follows. The other bounds in the proof of Lemma 6.7 from [27] follow similarly. Therefore, we have the following version of Theorem 2.2 from [27] -

Proposition S.1. *Under assumptions 1 and 2, we have*

$$P\left(\left\|\frac{1}{n}\sum_{j=1}^{n} A\left(s_j\right) - \Sigma\right\|_2 \geq t\right) \leq d^{2-\frac{\pi}{4}} \exp\left(\frac{t^2/\frac{32}{\pi^2}}{\frac{1+|\lambda_2(P)|}{1-|\lambda_2(P)|} n\mathcal{V} + \frac{8/\pi}{1-|\lambda_2(P)|}\mathcal{M}t}\right)$$

The proof of Proposition 1 now follows by converting the tail bound into a high probability bound and using Wedin's theorem [36]. See proof of Theorem 1.1 in [14] for details.

S.3 Useful Results

This section presents some useful lemmas and their proofs that are subsequently used in our proofs.

Lemma S.2. *(Reverse mixing) Consider a reversible, irreducible, and aperiodic Markov chain started from the stationary distribution. Then,*

$$\frac{1}{2}\sup_{t\in\Omega}\sum_{s} \left|\mathbb{P}\left(Z_t = s | Z_{t+k} = t\right) - \pi\left(s\right)\right| = d_{mix}\left(k\right)$$

Proof. Let the transition probabilities of the Markov chain be represented as $P(x|y) := P(Z_{t+1} = x|Z_t = y)$. Consider the time-reversed chain $Y_i := Z_{n-i+1}$ for $i = 1, 2, \ldots n$. Then,

$$\begin{aligned}
&\mathbb{P}\left(Y_l = s_l | Y_{l-1} = s_{l-1}, Y_{l-2} = s_{l-2} \ldots Y_1 = s_1\right) \\
&\quad = \mathbb{P}\left(Z_{n-l+1} = s_l | Z_{n-l+2} = s_{l-1}, Z_{n-l+3} = s_{l-2}, \ldots Z_n = s_1\right) \\
&\quad = \mathbb{P}\left(Z_{n-l+1} = s_l | Z_{n-l+2} = s_{l-1}\right) \quad \text{using Lemma S.6} \\
&\quad = \frac{\mathbb{P}\left(Z_{n-l+1} = s_l, Z_{n-l+2} = s_{l-1}\right)}{\mathbb{P}\left(Z_{n-l+2} = s_{l-1}\right)} \\
&\quad = \frac{\pi\left(s_l\right) P(s_{l-1}|s_l)}{\pi\left(s_{l-1}\right)} \\
&\quad = P(s_l|s_{l-1}) \quad \text{using reversibility}
\end{aligned}$$

This proves that Y_n is an irreducible Markov chain with the same transition probabilities as the original Markov chain. The irreducibility of Y_n follows from the original Markov chain being irreducible. Therefore,

$$\mathbb{P}\left(Z_t = s_1 | Z_{t+k} = s_2\right) = \mathbb{P}\left(Y_{n+1-t} = s_1 | Y_{n+1-t-k} = s_2\right) \tag{21}$$

Then,

$$\frac{1}{2}\sup_{t\in\Omega}\sum_{s} \left|\mathbb{P}\left(Z_t = s | Z_{t+k} = t\right) - \pi\left(s\right)\right| = \frac{1}{2}\sup_{t\in\Omega}\sum_{s} \left|\mathbb{P}\left(Y_{n+1-t} = s | Y_{n+1-t-k} = t\right) - \pi\left(s\right)\right| = d_{\text{mix}}\left(k\right)$$

where the last inequality follows from the forward mixing properties of the Markov chain. □

Lemma S.3. *Let $C_{j,i} = \prod_{t=j}^{i}(I + Z_t)$ for $i \leq j \leq n$, where $Z_t \in \mathbb{R}^{d\times d}$ are symmetric PSD matrices. Let $U \in \mathbb{R}^{d\times d'}$. Then,*

$$\mathrm{Tr}\left(U^T C_{j,i+1} C_{j,i+1}^T U\right) \leq \mathrm{Tr}\left(U^T C_{j,i} C_{j,i}^T U\right)$$

Proof.

$$\begin{aligned}\mathrm{Tr}\left(U^T C_{j,i} C_{j,i}^T U\right) &= \mathrm{Tr}\left(U^T C_{j,i+1}(I + 2Z_i + Z_i^2) C_{j,i+1}^T U\right) \\ &= \mathrm{Tr}\left(U^T C_{j,i+1} C_{j,i+1}^T U\right) + \mathrm{Tr}\left(U^T C_{j,i+1}(2Z_i + Z_i^2) C_{j,i+1}^T U\right)\end{aligned}$$

Since Z_i and Z_i^2 are both PSD, the second term on the RHS is always positive. This yields the proof. □

Lemma S.4. *Let $B_t = \prod_{i=t}^{1}(I + Z_i)$, where $Z_i \in \mathbb{R}^{d\times d}$ are symmetric PSD matrices.*

$$\mathrm{Tr}\left(B_{n-1} B_{n-1}^T\right) \leq \mathrm{Tr}\left(B_n B_n^T\right)$$

Proof.

$$\begin{aligned}\mathrm{Tr}\left(B_n B_n^T\right) &= \mathrm{Tr}\left((I + Z_n) B_{n-1} B_{n-1}^T (I + Z_n)\right) \\ &= \mathrm{Tr}\left(B_{n-1} B_{n-1}^T\right) + \mathrm{Tr}\left(Z_n B_{n-1} B_{n-1}^T\right) + \mathrm{Tr}\left(B_{n-1} B_{n-1}^T Z_n\right) + \mathrm{Tr}\left(Z_n B_{n-1} B_{n-1}^T Z_n\right) \\ &= \mathrm{Tr}\left(B_{n-1} B_{n-1}^T\right) + 2\,\mathrm{Tr}\left(B_{n-1}^T Z_n B_{n-1}\right) + \mathrm{Tr}\left(B_{n-1}^T Z_n^2 B_{n-1}\right)\end{aligned}$$

Since Z_n and Z_n^2 are both PSD, the last two terms on the RHS are always positive. This yields the proof. □

Lemma S.5. *Consider matrices $X \in \mathbb{R}^{d\times d'}$ and $A \in \mathbb{R}^{d\times d}$. Then,*

$$\left|\mathrm{Tr}\left(X^T A X\right)\right| \leq \|A\|_2 \mathrm{Tr}\left(X^T X\right)$$

Proof. For a matrix $Z \in \mathbb{R}^{d\times d}$, let the singular values be denoted as :

$$\sigma_{max}(Z) = \sigma_1(Z) \geq \sigma_2(Z) \ldots \geq \sigma_d(Z)$$

Using Von-Neumann's trace inequality, we have

$$\begin{aligned}\left|\mathrm{Tr}\left(X^T A X\right)\right| &= \left|\mathrm{Tr}\left(A X X^T\right)\right| \\ &\leq \sum_{i=1}^{d} \sigma_i(A)\,\sigma_i\left(X X^T\right) \\ &\leq \sigma_{max}(A) \sum_{i=1}^{d} \sigma_i\left(X X^T\right) \\ &= \|A\|_2 \mathrm{Tr}\left(X X^T\right) \\ &= \|A\|_2 \mathrm{Tr}\left(X^T X\right)\end{aligned}$$

□

Lemma S.6. *Given the Markov property in a Markov chain, the reverse Markov property holds, i.e*

$$P\left(Z_t = s | Z_{t+1} = w, Z_{t+2} = s_{t+2} \ldots Z_n = s_n\right) = P\left(Z_t = s | Z_{t+1} = w\right)$$

Proof.

$$
\begin{aligned}
&P\left(Z_t = s | Z_{t+1} = w, Z_{t+2} = s_{t+2} \dots Z_n = s_n\right) \\
&= \frac{P\left(Z_t = s, Z_{t+1} = w, Z_{t+2} = s_{t+2} \dots Z_n = s_n\right)}{P\left(Z_{t+1} = t, Z_{t+2} = s_{t+2} \dots Z_n = s_n\right)} \\
&= \frac{P\left(Z_t = s, Z_{t+1} = w\right) P\left(Z_{t+2} = s_{t+2} \dots Z_n = s_n | Z_t = s, Z_{t+1} = w\right)}{P\left(Z_{t+1} = w\right) P\left(Z_{t+2} = s_{t+2} \dots Z_n = s_n | Z_{t+1} = w\right)} \\
&= \frac{P\left(Z_t = s, Z_{t+1} = w\right) P\left(Z_{t+2} = s_{t+2} \dots Z_n = s_n | Z_{t+1} = w\right)}{P\left(Z_{t+1} = w\right) P\left(Z_{t+2} = s_{t+2} \dots Z_n = s_n | Z_{t+1} = w\right)} \\
&= \frac{P\left(Z_t = s, Z_{t+1} = w\right)}{P\left(Z_{t+1} = w\right)} \\
&= P\left(Z_t = s | Z_{t+1} = w\right)
\end{aligned}
$$

□

S.3.1 Proof of Lemma 2

Now we are ready to provide a proof of Lemma 2.

Proof of Lemma 2. Without loss of generality, we prove the statement for $m = 1$. For convenience of notation, we denote $k := k_1$. Note that,

$$
B_{k,1} = \sum_{r=0}^{k} \sum_{(i_1, i_2 \dots i_r) \in G_r} \prod_{j=1}^{r} \eta_{i_j} A(s_{i_j}),\ G_r = \{(i_1, \dots, i_r) \in \{1, \dots, N\}^r : i_1 < \dots < i_r\}
$$

with the convention that $\prod_{\phi} = I$. Therefore, since η_i forms a non-increasing sequence and $|G_r| = \binom{k}{r}$, we have,

$$
\begin{aligned}
\left\|B_{k,1} - I\right\|_2 &= \left\| \sum_{r=1}^{k} \sum_{(i_1, i_2 \dots i_r) \in G_r} \prod_{j=1}^{r} \eta_{i_j} A(s_{i_j}) \right\|_2 \\
&\leq \sum_{r=1}^{k} \sum_{(i_1, i_2 \dots i_r) \in G_r} \left\| \prod_{j=1}^{r} \eta_{i_j} A(s_{i_j}) \right\|_2 \\
&\leq \sum_{r=1}^{k} \binom{k}{r} \left(\prod_{i=1}^{r} \eta_i \right) \left(\mathcal{M} + \lambda_1\right)^r \\
&\leq \sum_{r=1}^{k} \frac{k^r}{r!} \left(\prod_{i=1}^{r} \eta_i \right) \left(\mathcal{M} + \lambda_1\right)^r \\
&\leq \sum_{r=1}^{k} \frac{k^r}{r!} \eta_1^r \left(\mathcal{M} + \lambda_1\right)^r \\
&\leq \exp\left(k \eta_1 \left(\mathcal{M} + \lambda_1\right)\right) - 1 \\
&\leq k \eta_1 \left(\mathcal{M} + \lambda_1\right) \left(1 + k \eta_1 \left(\mathcal{M} + \lambda_1\right)\right) \text{ using 23} \\
&\leq \left(1 + \epsilon\right) k \eta_1 \left(\mathcal{M} + \lambda_1\right) \qquad (22)
\end{aligned}
$$

where we have used the assumptions that $\|A(s)\|_2 \leq \|A(s) - \Sigma\| + \|\Sigma\|_2 = \left(\mathcal{M} + \lambda_1\right)$, $k \eta_1 \left(\mathcal{M} + \lambda_1\right) < 1$ and the useful result that

$$
e^x \leq 1 + x + x^2, x \in [0, 1.79] \qquad (23)
$$

This completes the proof for (a).

For part (b), we have

$$\begin{aligned}
\left\| B_{k,1} - I - \sum_{t=1}^{k} \eta_t A\left(s_t\right) \right\|_2 &= \left\| \sum_{r=2}^{k} \sum_{(i_1, i_2 \ldots i_r) \in G_r} \prod_{j=1}^{r} \eta_{i_j} A(s_{i_j}) \right\|_2 \\
&\leq \sum_{r=2}^{k} \sum_{(i_1, i_2 \ldots i_r) \in G_r} \left\| \prod_{j=2}^{r} \eta_{i_j} A(s_{i_j}) \right\|_2 \\
&\leq \sum_{r=2}^{k} \binom{k}{r} \left(\prod_{i=2}^{r} \eta_i \right) \left(\mathcal{M} + \lambda_1\right)^r \\
&\leq \sum_{r=2}^{k} \frac{k^r}{r!} \left(\prod_{i=2}^{r} \eta_i \right) \left(\mathcal{M} + \lambda_1\right)^r \\
&\leq \sum_{r=2}^{k} \frac{k^r}{r!} \eta_1^r \left(\mathcal{M} + \lambda_1\right)^r \\
&\leq \exp\left(k\eta_1 \left(\mathcal{M} + \lambda_1\right)\right) - 1 - k\eta_1 \left(\mathcal{M} + \lambda_1\right) \\
&\leq k^2 \eta_1^2 \left(\mathcal{M} + \lambda_1\right)^2 \text{ using 23 along with } k\eta_1 \left(\mathcal{M} + \lambda_1\right) < 1 \quad (24)
\end{aligned}$$

which completes the proof. □

S.3.2 Proof of Lemma 3

Before proving Lemma 3, we will need the following lemma.

Lemma S.7. *For arbitrary matrices $M_i \in \mathbb{R}^{d \times d}, i \in [n]$ and $Q \in \mathbb{R}^{n \times n}$, we have*

$$\left\| \sum_{x,y \in [n]} Q\left(x, y\right) M_x M_y^T \right\|_2 \leq \|Q\|_2 \left\| \sum_{x \in [n]} M_x M_x^T \right\|_2$$

where $\|.\|_2$ denotes the spectral norm.

Proof. Define matrix $X \in \mathbb{R}^{d \times nd}$ as $X := \begin{bmatrix} M_1 & M_2 & \ldots & M_n \end{bmatrix}$. We note that

$$\begin{aligned}
\|X\|_2 &= \sqrt{\lambda_{max}\left(XX^T\right)} \\
&= \sqrt{\lambda_{max}\left(\sum_{x \in [n]} M_x M_x^T \right)} \\
&= \sqrt{\left\| \sum_{x \in [n]} M_x M_x^T \right\|_2} \text{ since } \sum_{x \in [n]} M_x M_x^T \text{ is a symmetric matrix}
\end{aligned}$$

Then, we have,

$$\begin{aligned}
\sum_{x,y \in [n]} Q\left(x, y\right) M_x M_y^T &= X\left(Q \otimes I_{d \times d}\right) X^T, \text{ where } \otimes \text{ denotes the kronecker product} \\
&\leq \|X\|_2^2 \left\|Q \otimes I_{d \times d}\right\|_2 \text{ using submultiplicativity of the spectral norm} \\
&= \|X\|_2^2 \|Q\|_2 \text{ since } \|A \otimes B\|_2 = \|A\|_2 \|B\|_2
\end{aligned}$$

which completes our proof. □

Proof of Lemma 3. We denote $k_i := k$ for convenience of notation. By using reversibility (see 21), we know that the time-reversed process is also a Markov chain with the same transition probabilities. Then, for $i < j \leq i + k$ and any m,

$$\begin{aligned} P(s_i = s, s_j = t | s_{i+k} = u) &= P(s_i = s | s_j = t) P(s_j = t | s_{i+k} = u) \\ &\overset{(i)}{=} P^{j-i}(t, s) P^{i+k-j}(u, t) \\ &= P(s_m = s | s_{m-j+i} = t) P(s_{m-j+i} = t | s_{m-k} = u) \\ &= P(s_m = s, s_{m-j+i} = t | s_{m-k} = u) \end{aligned} \tag{25}$$

Step (i) uses reversibility. Therefore,

$$\begin{aligned} \mathbb{E}\left[\left(A\left(s_i\right) - \Sigma\right) S A\left(s_j\right) | s_{i+k}, \ldots s_n\right] &= \sum_{s,t} \left(\Sigma_s + \mu_s \mu_s^T - \Sigma\right) S \left(\Sigma_t + \mu_t \mu_t^T\right) P(s_i = s, s_j = t | s_{i+k}, \ldots s_n) \\ \text{using Lemma S.6} \ &= \sum_{s,t} \left(\Sigma_s + \mu_s \mu_s^T - \Sigma\right) S \left(\Sigma_t + \mu_t \mu_t^T\right) P(s_i = s, s_j = t | s_{i+k}) \\ \text{using Eq 25} \ &= \sum_{s,t} \left(\Sigma_s + \mu_s \mu_s^T - \Sigma\right) S \left(\Sigma_t + \mu_t \mu_t^T\right) P(s_m = s, s_{m-j+i} = t | s_{m-k} = u) \\ &= \mathbb{E}\left[\left(A\left(s_m\right) - \Sigma\right) S A\left(s_{m-j+i}\right) | s_{m-k}\right] \\ &= \mathbb{E}\left[\left(A\left(s_j\right) - \Sigma\right) S A\left(s_i\right) | s_{j-k}\right] \ \text{setting } m := j \end{aligned}$$

Therefore, without loss of generality, we proceed with the second form.

$$\begin{aligned} &\left\|\mathbb{E}\left[\left(A\left(s_j\right) - \Sigma\right) S A\left(s_i\right) | s_{j-k} = x_0\right]\right\|_2 \\ &\leq \underbrace{\left\|\mathbb{E}\left[\left(A\left(s_j\right) - \Sigma\right) S \Sigma | s_{j-k} = x_0\right]\right\|_2}_{T_1} + \underbrace{\left\|\mathbb{E}\left[\left(A\left(s_j\right) - \Sigma\right) S \left(A\left(s_i\right) - \Sigma\right) | s_{j-k} = x_0\right]\right\|_2}_{T_2} \end{aligned}$$

$$\begin{aligned} T_1 &:= \left\|\mathbb{E}\left[\left(A\left(s_j\right) - \Sigma\right) S \Sigma | s_{j-k} = x_0\right]\right\|_2 \\ &= \left\|\mathbb{E}\left[\mathbb{E}_{D(s_j)}\left[\left(A\left(s_j\right) - \Sigma\right)\right] | s_{j-k} = x_0\right] S \Sigma\right\|_2 \\ &= \left\|\mathbb{E}\left[\left(\Sigma_{s_j} + \mu_{s_j} \mu_{s_j}^T - \Sigma\right) | s_{j-k} = x_0\right] S \Sigma\right\|_2 \\ &= \left\|\sum_{s \in \Omega} P^k(s_{j-k}, s) \left(\Sigma_s + \mu_s \mu_s^T - \Sigma\right) S \Sigma\right\|_2 \\ &\leq \left\|\sum_{s \in \Omega} \left(P^k(s_{j-k}, s) - \pi\left(s\right)\right) \left(\Sigma_s + \mu_s \mu_s^T - \Sigma\right) + \underbrace{\mathbb{E}_\pi\left[\left(\Sigma_s + \mu_s \mu_s^T - \Sigma\right)\right]}_{=0}\right\|_2 \|S\|_2 \|\Sigma\|_2 \\ &= \lambda_1 \|S\|_2 \left(\left\|\sum_{s \in \Omega} \left(P^k(s_{j-k}, s) - \pi\left(s\right)\right) \left(\Sigma_s + \mu_s \mu_s^T - \Sigma\right)\right\|_2\right) \\ &\leq \lambda_1 \|S\|_2 \mathcal{M} \sum_{s \in \Omega} \left|P^k(s_{j-k}, s) - \pi\left(s\right)\right| \\ &\leq 2\lambda_1 \|S\|_2 \mathcal{M} d_{\text{mix}}\left(k_{i+1}\right) \\ &\leq 2\eta_i^2 \mathcal{M} \lambda_1 \|S\|_2 \end{aligned} \tag{26}$$

$$
\begin{aligned}
T_2 &= \left\|\mathbb{E}\left[\left(A\left(s_j\right)-\Sigma\right)S\left(A\left(s_i\right)-\Sigma\right)|s_{j-k}=x_0\right]\right\|_2 \\
&= \left\|\sum_{x,y\in\Omega}\mathbb{P}\left(s_j=x,s_i=y|s_{j-k}=x_0\right)\mathbb{E}_{D(x)}\left[A\left(x\right)-\Sigma\right]S\mathbb{E}_{D(y)}\left[A\left(y\right)-\Sigma\right]\right\|_2 \text{ using independence of} \\
&\quad D\left(x\right) \text{ and } D\left(y\right) \text{ conditioned on } x \text{ and } y \\
&= \left\|\sum_{x,y\in\Omega}\mathbb{P}\left(s_j=x,s_i=y|s_{j-k}=x_0\right)\underbrace{\left(\Sigma_x+\mu_x\mu_x^T-\Sigma\right)S^{\frac{1}{2}}}_{W_x}\underbrace{S^{\frac{1}{2}}\left(\Sigma_y+\mu_y\mu_y^T-\Sigma\right)}_{W_y^T}\right\|_2 \\
&= \left\|\sum_{x,y\in\Omega}\mathbb{P}\left(s_j=x|s_i=y\right)\mathbb{P}\left(s_i=y|s_{j-k}=x_0\right)W_xW_y^T\right\|_2 \text{ using the Markov property} \\
&= \left\|\sum_{x,y\in\Omega}P^{j-i}\left(y,x\right)P^{i-j+k}\left(x_0,y\right)W_xW_y^T\right\|_2 \\
&= \left\|\sum_{x,y\in\Omega}\left(P^{j-i}\left(y,x\right)-\pi\left(x\right)\right)P^{i-j+k}\left(x_0,y\right)W_xW_y^T+\sum_{x,y\in\Omega}\pi\left(x\right)P^{i-j+k}\left(x_0,y\right)W_xW_y^T\right\|_2 \\
&= \left\|\sum_{x,y\in\Omega}\left(P^{j-i}\left(y,x\right)-\pi\left(x\right)\right)P^{i-j+k}\left(x_0,y\right)W_xW_y^T+\underbrace{\sum_{x\in\Omega}\pi\left(x\right)W_x}_{=0}\sum_{y\in\Omega}P^{i-j+k}\left(x_0,y\right)W_y^T\right\|_2 \\
&= \left\|\sum_{x,y\in\Omega}\left(P^{j-i}\left(y,x\right)-\pi\left(x\right)\right)P^{i-j+k}\left(x_0,y\right)W_xW_y^T\right\|_2 \\
&\leq \underbrace{\left\|\sum_{x,y\in\Omega}\left(P^{j-i}\left(y,x\right)-\pi\left(x\right)\right)\left(P^{j-i+k}\left(x_0,y\right)-\pi\left(y\right)\right)W_xW_y^T\right\|_2}_{T_{21}}+\underbrace{\left\|\sum_{x,y\in\Omega}\left(P^{j-i}\left(y,x\right)-\pi\left(x\right)\right)\pi\left(y\right)W_xW_y^T\right\|_2}_{T_{22}}
\end{aligned}
\tag{27}
$$

For T_{21}, we have,

$$
\begin{aligned}
T_{21} &\leq \sum_{x,y\in\Omega}\left|P^{j-i}\left(y,x\right)-\pi\left(x\right)\right|\left|P^{i-j+k}\left(x_0,y\right)-\pi\left(y\right)\right|\left\|W_xW_y^T\right\|_2 \\
&\leq \left\|S\right\|_2\mathcal{M}^2\sum_{y\in\Omega}\left|P^{i-j+k}\left(x_0,y\right)-\pi\left(y\right)\right|\sum_{x\in\Omega}\left|P^{j-i}\left(y,x\right)-\pi\left(x\right)\right| \\
&\leq 2\left\|S\right\|_2\mathcal{M}^2d_{\text{mix}}\left(j-i\right)\sum_{y\in\Omega}\left|P^{i-j+k}\left(x_0,y\right)-\pi\left(y\right)\right| \\
&\leq 4\left\|S\right\|_2\mathcal{M}^2d_{\text{mix}}\left(j-i\right)d_{\text{mix}}\left(i-j+k\right) \\
&\leq 4\left\|S\right\|_2\mathcal{M}^2 2^{-\left\lfloor\frac{j-i}{\tau_{\text{mix}}}\right\rfloor}2^{-\left\lfloor\frac{i-j+k}{\tau_{\text{mix}}}\right\rfloor} \\
&\leq 8\left\|S\right\|_2\mathcal{M}^2 2^{-\left\lfloor\frac{j-i+i-j+k}{\tau_{\text{mix}}}\right\rfloor} \text{ since } \forall a,b\ \ \lfloor a\rfloor+\lfloor b\rfloor\geq\lfloor a+b\rfloor-1 \\
&\leq 8\left\|S\right\|_2\mathcal{M}^2 2^{-\left\lfloor\frac{k}{\tau_{\text{mix}}}\right\rfloor}\leq 8\left\|S\right\|_2\mathcal{M}^2d_{\text{mix}}\left(k\right)\leq 8\eta_i^2\mathcal{M}^2\left\|S\right\|_2
\end{aligned}
\tag{28}
$$

For T_{22}, we have,

$$
\begin{aligned}
T_{22} &= \left\| \sum_{x,y\in\Omega} \left(P^{j-i}(y,x) - \pi(x)\right) \pi(y) W_x W_y^T \right\|_2 \\
&= \left\| \sum_{x,y\in\Omega} \frac{\left(P^{j-i}(y,x) - \pi(x)\right)}{\sqrt{\pi(x)}} \sqrt{\pi(y)} \left(\sqrt{\pi(x)} W_x\right) \left(\sqrt{\pi(y)} W_y^T\right) \right\|_2 \\
&= \left\| \sum_{x,y\in\Omega} \frac{\left(P^{j-i}(y,x) - \pi(x)\right)}{\sqrt{\pi(x)}} \sqrt{\pi(y)} \left(\sqrt{\pi(x)} \left(\Sigma_x + \mu_x\mu_x^T - \Sigma\right) S^{\frac{1}{2}}\right) \left(\sqrt{\pi(y)} S^{\frac{1}{2}} \left(\Sigma_y + \mu_y\mu_y^T - \Sigma\right)\right) \right\|_2 \\
&\overset{(i)}{\leq} \|Q\|_2 \left\| \sum_{x\in\Omega} \pi(x) \left(\Sigma_x + \mu_x\mu_x^T - \Sigma\right) S \left(\Sigma_x + \mu_x\mu_x^T - \Sigma\right) \right\|_2 \\
&= \|Q\|_2 \left\| \mathbb{E}_\pi \left[\left(\Sigma_x + \mu_x\mu_x^T - \Sigma\right) S \left(\Sigma_x + \mu_x\mu_x^T - \Sigma\right) \right] \right\|_2 \\
&\leq \|Q\|_2 \|S\|_2 \left\| \mathbb{E}_\pi \left[\left(\Sigma_x + \mu_x\mu_x^T - \Sigma\right)^2 \right] \right\|_2 \\
&\leq \mathcal{V} \|Q\|_2 \|S\|_2
\end{aligned} \tag{29}
$$

Step (i) uses Lemma S.7 with $Q(y,x) := \frac{\left(P^{i-j}(y,x) - \pi(x)\right)}{\sqrt{\pi(x)}} \sqrt{\pi(y)}$ and $M_x = \sqrt{\pi(x)} \left(\Sigma_x + \mu_x\mu_x^T - \Sigma\right) S^{\frac{1}{2}}$.
Let's now bound $\|Q\|_2$. Let $\Pi := \operatorname{diag}(\pi) \in \mathbb{R}^{\Omega\times\Omega}$ and $t := j - i$. Then, we have

$$
\begin{aligned}
Q &= \Pi^{\frac{1}{2}} \left(P^t - \mathbb{1}\mathbb{1}^T\Pi\right) \Pi^{-\frac{1}{2}} \\
&= \Pi^{\frac{1}{2}} P^t \Pi^{-\frac{1}{2}} - \Pi^{\frac{1}{2}} \mathbb{1}\mathbb{1}^T \Pi^{\frac{1}{2}}
\end{aligned}
$$

Now, since we have a reversible Markov chain, $\Pi P = P^T \Pi$. Therefore,

$$
\begin{aligned}
\Pi^{\frac{1}{2}} P \Pi^{-\frac{1}{2}} &= \Pi^{\frac{1}{2}} \Pi^{-1} P^T \Pi \Pi^{-\frac{1}{2}} \\
&= \Pi^{-\frac{1}{2}} P^T \Pi^{\frac{1}{2}}
\end{aligned}
$$

Therefore, P is similar to the self-adjoint matrix $\Pi^{\frac{1}{2}} P \Pi^{-\frac{1}{2}}$ and their eigenvalues are real and the same. Further note that $\Pi^{\frac{1}{2}} \mathbb{1}$ is the leading eigenvector of $\Pi^{\frac{1}{2}} P \Pi^{-\frac{1}{2}}$ with eigenvalue 1 since

$$
\begin{aligned}
\Pi^{\frac{1}{2}} P \Pi^{-\frac{1}{2}} \Pi^{\frac{1}{2}} \mathbb{1} &= \Pi^{\frac{1}{2}} P \mathbb{1} \\
&= \Pi^{\frac{1}{2}} \mathbb{1} \text{ since } P \text{ is a stochastic matrix}
\end{aligned}
$$

Now,

$$
\begin{aligned}
\|Q\|_2 &= \left\| \Pi^{\frac{1}{2}} P^t \Pi^{-\frac{1}{2}} - \Pi^{\frac{1}{2}} \mathbb{1}\mathbb{1}^T \Pi^{\frac{1}{2}} \right\|_2 \\
&= \left\| \left(\Pi^{\frac{1}{2}} P \Pi^{-\frac{1}{2}}\right)^t - \Pi^{\frac{1}{2}} \mathbb{1}\mathbb{1}^T \Pi^{\frac{1}{2}} \right\|_2 \\
&\leq \left| \lambda_2 \left(\Pi^{\frac{1}{2}} P \Pi^{-\frac{1}{2}}\right) \right|^t \\
&= \left|\lambda_2(P)\right|^t
\end{aligned}
$$

where $|\lambda_2(.)|$ denotes the second-largest eigenvalue in magnitude. Therefore, using 26, 28 and 29, we have

$$
\begin{aligned}
\mathbb{E}\left[\left(A(s_i) - \Sigma\right) S A(s_j) \,|s_{i+k}, \ldots s_n\right] &\leq \left(\left|\lambda_2(P)\right|^{j-i} \mathcal{V} + 8\eta_i^2 \mathcal{M}^2 + 2\eta_i^2 \mathcal{M}\lambda_1\right) \|S\|_2 \\
&\leq \left(\left|\lambda_2(P)\right|^{j-i} \mathcal{V} + 8\eta_i^2 \mathcal{M}\left(\mathcal{M} + \lambda_1\right)\right) \|S\|_2
\end{aligned}
$$

Hence proved. □

Lemma S.8. *Let $\forall i \in [n], \eta_i k_i (\mathcal{M} + \lambda_1) \leq \epsilon, \epsilon \in (0,1)$ and η_i forms a non-increasing sequence. Set $k_i := \tau_{mix}\left(\gamma \eta_i^2\right), \gamma \in (0,1]$. Then for constant matrix $U \in \mathbb{R}^{d \times d'}$, and constant positive semi-definite matrix $G \in \mathbb{R}^{d \times d}$, $i \leq j \leq n,\ j - i \geq k_i$, we have*

$$\begin{aligned}
&\left|\mathbb{E}\left[\mathrm{Tr}\left(U^T B_{j,i+1} G \left(A_i - \Sigma\right) B_{j,i+1}^T U\right)\right]\right| \\
&\quad \leq \eta_{i+1}\|G\|_2 \left(\frac{2\mathcal{V}\left|\lambda_2(P)\right|}{1 - \left|\lambda_2(P)\right|} + \eta_{i+1}\mathcal{M}\left(2\gamma(1+8\epsilon) + \left(2 + (1+\epsilon)^2\right) k_{i+1}^2 \left(\mathcal{M} + \lambda_1\right)^2\right)\right) \\
&\qquad \times \mathbb{E}\left[\mathrm{Tr}\left(U^T B_{j,i+k_{i+1}} B_{j,i+k_{i+1}}^T U\right)\right]
\end{aligned}$$

where $B_{j,i}$ is defined in 12.

Proof. For the convenience of notation, we denote $k_{i+1} := k$. Let $B_{j,i+1} = B_{j,i+k}(I + R)$, then

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{Tr}\left(U^T B_{j,i+1} G \left(A_i - \Sigma\right) B_{j,i+1}^T U\right)\right] = \\
&\quad \mathbb{E}\left[\underbrace{\mathrm{Tr}\left(U^T B_{j,i+k} G \left(A_i - \Sigma\right) B_{j,i+k}^T U\right)}_{T_1}\right] + \mathbb{E}\left[\underbrace{\mathrm{Tr}\left(U^T B_{j,i+k} G \left(A_i - \Sigma\right) R^T B_{j,i+k}^T U\right)}_{T_2}\right] + \\
&\quad \mathbb{E}\left[\underbrace{\mathrm{Tr}\left(U^T B_{j,i+k} R G \left(A_i - \Sigma\right) B_{j,i+k}^T U\right)}_{T_3}\right] + \mathbb{E}\left[\underbrace{\mathrm{Tr}\left(U^T B_{j,i+k} R G \left(A_i - \Sigma\right) R^T B_{j,i+k}^T U\right)}_{T_4}\right]
\end{aligned} \tag{30}$$

We will now bound each of the terms $\mathbb{E}[T_1], \mathbb{E}[T_2], \mathbb{E}[T_3]$ and $\mathbb{E}[T_4]$.

$$\begin{aligned}
\mathbb{E}[T_1] &= \mathbb{E}\left[\mathrm{Tr}\left(U^T B_{j,i+k} G \left(A_i - \Sigma\right) B_{j,i+k}^T U\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathrm{Tr}\left(U^T B_{j,i+k} G \left(A_i - \Sigma\right) B_{j,i+k}^T U\right) \middle| s_{i+k}, \ldots s_{j-1}, s_j\right]\right] \\
&= \mathbb{E}\left[\mathrm{Tr}\left(U^T B_{j,i+k} G\, \mathbb{E}\left[\left(A_i - \Sigma\right) \middle| s_{i+k}, \ldots s_{j-1}, s_j\right] B_{j,i+k}^T U\right)\right] \\
&= \mathbb{E}\left[\mathrm{Tr}\left(U^T B_{j,i+k} G\, \mathbb{E}\left[\left(A_i - \Sigma\right) \middle| s_{i+k}\right] B_{j,i+k}^T U\right)\right] \text{ using Lemma S.6}
\end{aligned}$$

Now, using Lemma 1, we have,

$$\begin{aligned}
\left\|\mathbb{E}\left[\left(A_i - \Sigma\right) \middle| s_{i+k}\right]\right\|_2 &= \left\|\sum_{s \in \Omega} P^k(s_{i+k}, s)\left(A_i - \Sigma\right)\right\|_2 \\
&= \left\|\sum_{s \in \Omega}\left(P^k(s_{i+k}, s) - \pi(s)\right)\left(A_i - \Sigma\right) + \underbrace{\mathbb{E}_\pi\left[\left(A_i - \Sigma\right)\right]}_{=0}\right\|_2 \\
&= \left\|\sum_{s \in \Omega}\left(P^k(s_{i+k}, s) - \pi(s)\right)\left(A_i - \Sigma\right)\right\|_2 \\
&\leq \mathcal{M} \sum_{s \in \Omega}\left|P^k(s_{i+k}, s) - \pi(s)\right| \\
&\leq 2\mathcal{M} d_{\text{mix}}(k_{i+1}) \\
&\leq 2\gamma\eta_{i+1}^2\mathcal{M}
\end{aligned} \tag{31}$$

where we have used Lemma S.5. Therefore,

$$\left|\mathbb{E}[T_1]\right| \leq \gamma\eta_{i+1}^2\mathcal{M}\|G\|_2 \mathbb{E}\left[\mathrm{Tr}\left(U^T B_{j,i+k} B_{j,i+k}^T U\right)\right] \tag{32}$$

We will now bound $\mathbb{E}\left[T_2\right]$. Let $R_0 := \sum_{\ell=i+1}^{i+k-1} \eta_\ell A_\ell$. Using Lemma 2 we have

$$\|R - R_0\|_2 \leq \eta_{i+1}^2 k_{i+1}^2 \left(\mathcal{M} + \lambda_1\right)^2$$

Then,

$$\begin{aligned}
\mathbb{E}[T_2] &= \mathbb{E}\left[\mathrm{Tr}\left(U^T B_{j,i+k} G \left(A_i - \Sigma\right) R^T B_{j,i+k}^T U\right)\right] \\
&= \mathbb{E}\left[\mathrm{Tr}\left(U^T B_{j,i+k} G \left(A_i - \Sigma\right) R_0^T B_{j,i+k}^T U\right)\right] + \mathbb{E}\left[\mathrm{Tr}\left(U^T B_{j,i+k} G \left(A_i - \Sigma\right) \left(R - R_0\right)^T B_{j,i+k}^T U\right)\right] \\
&= \mathbb{E}\left[\mathrm{Tr}\left(U^T B_{j,i+k} G \mathbb{E}[\left(A_i - \Sigma\right) R_0^T | s_{i+k}, \ldots s_{j-1}, s_j] B_{j,i+k}^T U\right)\right] + \\
&\qquad\qquad \mathbb{E}\left[\mathrm{Tr}\left(U^T B_{j,i+k} G \left(A_i - \Sigma\right) \left(R - R_0\right)^T B_{j,i+k}^T U\right)\right]
\end{aligned}$$

Using Lemma 3 with $S := I$ we have,

$$\begin{aligned}
\|\mathbb{E}\left[\left(A_i - \Sigma\right) R_0^T | s_{i+k}, \ldots, s_j\right]\|_2 &\leq \sum_{\ell=i+1}^{i+k-1} \eta_\ell \left(\left|\lambda_2\left(P\right)\right|^{\ell-i} \mathcal{V} + 8\gamma\eta_{i+1}^2 \mathcal{M}\left(\mathcal{M} + \lambda_1\right)\right) \\
&\leq \eta_{i+1} \mathcal{V} \frac{\left|\lambda_2\left(P\right)\right|}{1 - \left|\lambda_2\left(P\right)\right|} + 8\gamma\eta_{i+1}^3 k_{i+1} \mathcal{M}\left(\mathcal{M} + \lambda_1\right) \qquad (33)
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\left|\mathbb{E}\left[T_2\right]\right| \\
&\leq \|G\|_2 \left(\eta_{i+1} \frac{\mathcal{V}\left|\lambda_2\left(P\right)\right|}{1 - \left|\lambda_2\left(P\right)\right|} + 8\gamma\eta_{i+1}^3 k_{i+1} \mathcal{M}\left(\mathcal{M} + \lambda_1\right) + \eta_{i+1}^2 k_{i+1}^2 \mathcal{M}\left(\mathcal{M} + \lambda_1\right)^2\right) \mathbb{E}\left[\mathrm{Tr}\left(U^T B_{j,i+k} B_{j,i+k}^T U\right)\right] \\
&= \eta_{i+1} \|G\|_2 \left(\frac{\mathcal{V}\left|\lambda_2\left(P\right)\right|}{1 - \left|\lambda_2\left(P\right)\right|} + 8\gamma\eta_{i+1}^2 k_{i+1} \mathcal{M}\left(\mathcal{M} + \lambda_1\right) + \eta_{i+1} k_{i+1}^2 \mathcal{M}\left(\mathcal{M} + \lambda_1\right)^2\right) \mathbb{E}\left[\mathrm{Tr}\left(U^T B_{j,i+k} B_{j,i+k}^T U\right)\right] \qquad (34)
\end{aligned}$$

Similarly using Lemma 3 with $S := G$,

$$\left|\mathbb{E}\left[T_3\right]\right| \leq \eta_{i+1} \|G\|_2 \left(\frac{\mathcal{V}\left|\lambda_2\left(P\right)\right|}{1 - \left|\lambda_2\left(P\right)\right|} + 8\gamma\eta_{i+1}^2 k_{i+1} \mathcal{M}\left(\mathcal{M} + \lambda_1\right) + \eta_{i+1} k_{i+1}^2 \mathcal{M}\left(\mathcal{M} + \lambda_1\right)^2\right) \mathbb{E}\left[\mathrm{Tr}\left(U^T B_{j,i+k} B_{j,i+k}^T U\right)\right] \qquad (35)$$

Finally,

$$\begin{aligned}
\left|\mathbb{E}\left[T_4\right]\right| &\leq \mathcal{M} \|G\|_2 \|R\|_2^2 \mathbb{E}\left[\mathrm{Tr}\left(U^T B_{j,i+k} B_{j,i+k}^T U\right)\right] \\
&\leq \left(1 + \epsilon\right)^2 \eta_{i+1}^2 k_{i+1}^2 \mathcal{M}\left(\mathcal{M} + \lambda_1\right)^2 \|G\|_2 \mathbb{E}\left[\mathrm{Tr}\left(U^T B_{j,i+k} B_{j,i+k}^T U\right)\right] \text{ using Lemma 2} \qquad (36)
\end{aligned}$$

Therefore, using Eqs 32, 34, 35, 36 along with 30, we have

$$\begin{aligned}
&\left|\mathbb{E}\left[\mathrm{Tr}\left(U^T B_{j,i+1} G \left(A_i - \Sigma\right) B_{j,i+1}^T U\right)\right]\right| \\
&\leq \eta_{i+1} \|G\|_2 \left(\frac{2\mathcal{V}\left|\lambda_2\left(P\right)\right|}{1 - \left|\lambda_2\left(P\right)\right|} + \eta_{i+1} \mathcal{M}\left(2\gamma + 16\gamma\eta_{i+1} k_{i+1}\left(\mathcal{M} + \lambda_1\right) + \left(2 + \left(1 + \epsilon\right)^2\right) k_{i+1}^2 \left(\mathcal{M} + \lambda_1\right)^2\right)\right) \\
&\qquad\qquad \times \mathbb{E}\left[\mathrm{Tr}\left(U^T B_{j,i+k} B_{j,i+k}^T U\right)\right] \\
&\leq \eta_{i+1} \|G\|_2 \left(\frac{2\mathcal{V}\left|\lambda_2\left(P\right)\right|}{1 - \left|\lambda_2\left(P\right)\right|} + \eta_{i+1} \mathcal{M}\left(2\gamma\left(1 + 8\epsilon\right) + \left(2 + \left(1 + \epsilon\right)^2\right) k_{i+1}^2 \left(\mathcal{M} + \lambda_1\right)^2\right)\right) \\
&\qquad\qquad \times \mathbb{E}\left[\mathrm{Tr}\left(U^T B_{j,i+k} B_{j,i+k}^T U\right)\right]
\end{aligned}$$

where in the last line we used $\eta_{i+1} k_{i+1}\left(\mathcal{M} + \lambda_1\right) \leq \epsilon$. Hence proved. □

Lemma S.9. *Let $\forall i \in [n], \eta_i k_i (\mathcal{M} + \lambda_1) \leq \epsilon, \epsilon \in (0,1)$ and η_i forms a non-increasing sequence. Set $k_i := \tau_{mix}\left(\gamma \eta_i^2\right), \gamma \in (0,1]$. Then for constant matrices $U \in \mathbb{R}^{d \times d'}$, $G \in \mathbb{R}^{d \times d}$, $i \leq j \leq n,\ j - i \geq k_i$, we have*

$$\begin{aligned}
&\left|\mathbb{E}\left[\mathrm{Tr}\left(U^T B_{j,i+1} G \left(A_i - \Sigma\right)^2 B_{j,i+1}^T U\right)\right]\right| \\
&\qquad \leq \left(\mathcal{V} + \eta_{i+1}\mathcal{M}^2\left(2\gamma\eta_{i+1} + (1+\epsilon)\left(2 + \epsilon\left(1+\epsilon\right)\right) k_{i+1}\left(\mathcal{M} + \lambda_1\right)\right)\right) \|G\|_2 \, \mathbb{E}\left[\mathrm{Tr}\left(U^T B_{j,i+k_{i+1}} B_{j,i+k_{i+1}}^T U\right)\right]
\end{aligned}$$

where $B_{j,i}$ is defined in 12.

Proof. For convenience of notation, we denote $k_{i+1} := k$. Let $B_{j,i+1} = B_{j,i+k}\left(I + R\right)$, then

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{Tr}\left(U^T B_{j,i+1} G \left(A_i - \Sigma\right)^2 B_{j,i+1}^T U\right)\right] = \\
&\quad \mathbb{E}\left[\underbrace{\mathrm{Tr}\left(U^T B_{j,i+k} G \left(A_i - \Sigma\right)^2 B_{j,i+k}^T U\right)}_{T_1}\right] + \mathbb{E}\left[\underbrace{\mathrm{Tr}\left(U^T B_{j,i+k} G \left(A_i - \Sigma\right)^2 R^T B_{j,i+k}^T U\right)}_{T_2}\right] + \\
&\quad \mathbb{E}\left[\underbrace{\mathrm{Tr}\left(U^T B_{j,i+k} R G \left(A_i - \Sigma\right)^2 B_{j,i+k}^T U\right)}_{T_3}\right] + \mathbb{E}\left[\underbrace{\mathrm{Tr}\left(U^T B_{j,i+k} R G \left(A_i - \Sigma\right)^2 R^T B_{j,i+k}^T U\right)}_{T_4}\right]
\end{aligned}$$

We will now bound each of the terms $\mathbb{E}\left[T_1\right], \mathbb{E}\left[T_2\right], \mathbb{E}\left[T_3\right]$ and $\mathbb{E}\left[T_4\right]$.

Since $\left\|\mathbb{E}_\pi\left[\left(A_t - \Sigma\right)^2\right]\right\|_2 \leq \mathcal{V}$, therefore

$$\begin{aligned}
\mathbb{E}\left[T_1\right] &= \mathbb{E}\left[\mathrm{Tr}\left(U^T B_{j,i+k} G \left(A_i - \Sigma\right)^2 B_{j,i+k}^T U\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathrm{Tr}\left(U^T B_{j,i+k} G \left(A_i - \Sigma\right)^2 B_{j,i+k}^T U\right) \middle| s_{i+k}, \ldots s_{j-1}, s_j\right]\right] \\
&= \mathbb{E}\left[\mathrm{Tr}\left(U^T B_{j,i+k} G\, \mathbb{E}\left[\left(A_i - \Sigma\right)^2 \middle| s_{i+k}, \ldots s_{j-1}, s_j\right] B_{j,i+k}^T U\right)\right] \\
&= \mathbb{E}\left[\mathrm{Tr}\left(U^T B_{j,i+k} G\, \mathbb{E}\left[\left(A_i - \Sigma\right)^2 \middle| s_{i+k}\right] B_{j,i+k}^T U\right)\right] \text{ using Lemma S.6} \\
&\overset{(i)}{\leq} \left(\mathcal{V} + 2 d_{\text{mix}}\left(k\right)\mathcal{M}^2\right)\|G\|_2 \, \mathbb{E}\left[\mathrm{Tr}\left(U^T B_{j,i+k}\, B_{j,i+k}^T U\right)\right]
\end{aligned}$$

where in (i), we used similar steps as 31 to get

$$\left\|\mathbb{E}\left[\left(A_i - \Sigma\right)^2 \middle| s_{i+k}\right]\right\|_2 \leq \left\|\mathbb{E}_\pi\left[\left(A_i - \Sigma\right)^2\right]\right\|_2 + 2 d_{\text{mix}}\left(k\right)\mathcal{M}^2 \tag{37}$$

Next, using Lemma 2 we have that

$$\|R\|_2 \leq \left(1+\epsilon\right) k_{i+1}\eta_{i+1}\left(\mathcal{M} + \lambda_1\right). \tag{38}$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[T_2\right] &= \mathbb{E}\left[\mathrm{Tr}\left(U^T B_{j,i+k} G \left(A_i - \Sigma\right)^2 R^T B_{j,i+k}^T U\right)\right] \\
&\leq \left(1+\epsilon\right) k_{i+1}\eta_{i+1}\mathcal{M}^2\left(\mathcal{M} + \lambda_1\right)\|G\|_2 \, \mathbb{E}\left[\mathrm{Tr}\left(U^T B_{j,i+k} B_{j,i+k}^T U\right)\right]
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathbb{E}\left[T_3\right] &= \mathbb{E}\left[\mathrm{Tr}\left(U^T B_{j,i+k} R G \left(A_i - \Sigma\right)^2 B_{j,i+k}^T U\right)\right] \\
&\leq \left(1+\epsilon\right) k_{i+1}\eta_{i+1}\mathcal{M}^2\left(\mathcal{M} + \lambda_1\right)\|G\|_2 \, \mathbb{E}\left[\mathrm{Tr}\left(U^T B_{j,i+k} B_{j,i+k}^T U\right)\right]
\end{aligned}$$

Finally, using the bound on $\|R\|_2$ from Eq 38, we have:

$$\begin{aligned}
\mathbb{E}\left[T_4\right] &= \mathbb{E}\left[\mathrm{Tr}\left(U^T B_{j,i+k} R G \left(A_i - \Sigma\right)^2 R^T B_{j,i+k}^T U\right)\right] \\
&\leq \left(1+\epsilon\right)^2 k_{i+1}^2 \eta_{i+1}^2 \mathcal{M}^2 \left(\mathcal{M}+\lambda_1\right)^2 \|G\|_2 \, \mathbb{E}\left[\mathrm{Tr}\left(U^T B_{j,i+k} B_{j,i+k}^T U\right)\right] \\
&\leq \epsilon \left(1+\epsilon\right)^2 k_{i+1} \eta_{i+1} \mathcal{M}^2 \left(\mathcal{M}+\lambda_1\right) \|G\|_2 \, \mathbb{E}\left[\mathrm{Tr}\left(U^T B_{j,i+k} B_{j,i+k}^T U\right)\right] \text{ using } \forall i,\ \eta_i k_i \left(M+\lambda_1\right) \leq c
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\left|\mathbb{E}\left[\mathrm{Tr}\left(U^T B_{j,i+1} G \left(A_i - \Sigma\right)^2 B_{j,i+1}^T U\right)\right]\right| \\
&\overset{(i)}{\leq} \left(\mathcal{V} + \eta_{i+1}\left(2\gamma\eta_{i+1}\mathcal{M}^2 + \left(1+\epsilon\right)\left(2+\epsilon\left(1+\epsilon\right)\right) k_{i+1} \mathcal{M}^2 \left(\mathcal{M}+\lambda_1\right)\right)\right) \|G\|_2 \, \mathbb{E}\left[\mathrm{Tr}\left(U^T B_{j,i+k} B_{j,i+k}^T U\right)\right] \\
&= \left(\mathcal{V} + \eta_{i+1}\mathcal{M}^2\left(2\gamma\eta_{i+1} + \left(1+\epsilon\right)\left(2+\epsilon\left(1+\epsilon\right)\right) k_{i+1} \left(\mathcal{M}+\lambda_1\right)\right)\right) \|G\|_2 \, \mathbb{E}\left[\mathrm{Tr}\left(U^T B_{j,i+k} B_{j,i+k}^T U\right)\right]
\end{aligned}$$

where in (i), we used $d_{\text{mix}}\left(k\right) = d_{\text{mix}}\left(k_{i+1}\right) \leq \gamma\eta_{i+1}^2$. Hence proved. □

Lemma S.10. *Let $\forall i \in [n], \eta_i k_i \left(\mathcal{M}+\lambda_1\right) \leq \epsilon, \epsilon \in (0,1)$ and step-sizes η_i forms a non-increasing sequence. Further, let the step-sizes follow a slow-decay property, i.e, $\forall i, \eta_i \leq \eta_{i-k_i} \leq 2\eta_i$. Set $k_i := \tau_{mix}\left(\gamma\eta_i^2\right), \gamma \in (0,1]$. Let $G \in \mathbb{R}^{d\times d}$ be a constant positive semi-definite matrix, and $P_t := \mathrm{Tr}\left(B_{t-1}B_{t-1}^T G (A_t - \Sigma)\right)$, then,*

$$\mathbb{E}\left[P_t\right] \leq \eta_{t-k_t}\left(\frac{2\mathcal{V}\left|\lambda_2\left(P\right)\right|}{1-\left|\lambda_2\left(P\right)\right|} + \eta_{t-k_t}\mathcal{M}\left(2\gamma\left(1+8\epsilon\right) + \left(2+\left(1+\epsilon\right)^2\right) k_t^2 \left(\mathcal{M}+\lambda_1\right)^2\right)\right) \|G\|_2 \, \mathbb{E}\left[\mathrm{Tr}\left(B_{t-k_t}B_{t-k_t}^T\right)\right]$$

where B_t is defined in 2.

Proof. Let $B_t = \left(I+R\right) B_{t-k_t}$ with $\|R\|_2 \leq r$. Then,

$$\begin{aligned}
\mathbb{E}\left[P_t\right] = \mathbb{E}\Bigg[\underbrace{\mathrm{Tr}\left(B_{t-k_t}B_{t-k_t}^T G(A_t-\Sigma)\right)}_{P_{t,1}}\Bigg] + \mathbb{E}\Bigg[\underbrace{\mathrm{Tr}\left(B_{t-k_t}B_{t-k_t}^T R^T G(A_t-\Sigma)\right)}_{P_{t,2}}\Bigg] \\
+ \mathbb{E}\Bigg[\underbrace{\mathrm{Tr}\left(B_{t-k_t}B_{t-k_t}^T G(A_t-\Sigma)R\right)}_{P_{t,3}}\Bigg] + \mathbb{E}\Bigg[\underbrace{\mathrm{Tr}\left(B_{t-k_t}B_{t-k_t}^T R^T G(A_t-\Sigma)R\right)}_{P_{t,4}}\Bigg]
\end{aligned}$$

Let's consider each of the terms above. Using Von-Neumann's trace inequality and 34, we have,

$$\begin{aligned}
\mathbb{E}\left[P_{t,1}\right] &= \mathbb{E}\left[\mathrm{Tr}\left(B_{t-k_t}B_{t-k_t}^T \mathbb{E}\left[G(A_t-\Sigma)|s_1, s_2, \ldots, s_{t-k_t}\right]\right)\right] \\
&\leq \mathbb{E}\left[\mathrm{Tr}\left(B_{t-k_t}B_{t-k_t}^T G \mathbb{E}\left[\left(A_t-\Sigma\right)|s_{t-k_t}\right]\right)\right] \\
&\leq \left\|G\mathbb{E}\left[\left(A_t-\Sigma\right)|s_{t-k_t}\right]\right\|_2 \mathbb{E}\left[\mathrm{Tr}\left(B_{t-k_t}B_{t-k_t}^T\right)\right] \\
&\leq 2\mathcal{M} d_{\text{mix}}\left(k_t\right) \|G\|_2 \, \mathbb{E}\left[\mathrm{Tr}\left(B_{t-k_t}B_{t-k_t}^T\right)\right] \text{ using 31} \\
&\leq 2\gamma\eta_t^2\mathcal{M} \|G\|_2 \, \mathbb{E}\left[\mathrm{Tr}\left(B_{t-k_t}B_{t-k_t}^T\right)\right]
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}\left[P_{t,2}\right] &= \mathbb{E}\left[\mathrm{Tr}\left(B_{t-k_t}B_{t-k_t}^T, \mathbb{E}\left[R^T G(A_t-\Sigma)U|s_1, s_2, \ldots, s_{t-k_t}\right]\right)\right] \\
&\leq \left\|\mathbb{E}\left[R^T G(A_t-\Sigma)|s_1, s_2, \ldots, s_{t-k_t}\right]\right\|_2 \mathbb{E}\left[\mathrm{Tr}\left(B_{t-k_t}B_{t-k_t}^T\right)\right] \\
&= \left\|\mathbb{E}\left[R^T G(A_t-\Sigma)|s_{t-k_t}\right]\right\|_2 \mathbb{E}\left[\mathrm{Tr}\left(B_{t-k_t}B_{t-k_t}^T\right)\right] \\
&\leq \eta_{t-k_t} \|G\|_2 \left(\frac{\mathcal{V}\left|\lambda_2\left(P\right)\right|}{1-\left|\lambda_2\left(P\right)\right|} + 8\gamma\eta_{t-k_t}^2 k_t \mathcal{M}\left(\mathcal{M}+\lambda_1\right) + \eta_{t-k_t} k_t^2 \mathcal{M}\left(\mathcal{M}+\lambda_1\right)^2\right) \mathbb{E}\left[\mathrm{Tr}\left(B_{t-k_t}B_{t-k_t}^T\right)\right] \text{ using 34}
\end{aligned}$$

$$\mathbb{E}\left[P_{t,3}\right] \leq \eta_{t-k_t}\left\|G\right\|_2\left(\frac{\mathcal{V}\left|\lambda_2\left(P\right)\right|}{1-\left|\lambda_2\left(P\right)\right|}+8\gamma\eta_{t-k_t}^2k_t\mathcal{M}\left(\mathcal{M}+\lambda_1\right)+\eta_{t-k_t}k_t^2\mathcal{M}\left(\mathcal{M}+\lambda_1\right)^2\right)\mathbb{E}\left[\mathrm{Tr}\left(B_{t-k_t}B_{t-k_t}^T\right)\right]$$

using similar steps as $\mathbb{E}\left[P_{t,2}\right]$

$$\begin{aligned}
\mathbb{E}\left[P_{t,4}\right] &= \mathbb{E}\left[\mathrm{Tr}\left(B_{t-k_t}B_{t-k_t}^TR^TG(A_t-\Sigma)R\right)\right]\\
&\leq r^2\mathcal{M}\left\|G\right\|_2\mathbb{E}\left[\mathrm{Tr}\left(B_{t-k_t}B_{t-k_t}^T\right)\right]\\
&\leq \left(1+\epsilon\right)^2\eta_{t-k_t+1}^2k_t^2\mathcal{M}\left(\mathcal{M}+\lambda_1\right)^2\left\|G\right\|_2\mathbb{E}\left[\mathrm{Tr}\left(B_{t-k_t}B_{t-k_t}^T\right)\right] \text{ using Lemma 2}\\
&\leq \left(1+\epsilon\right)^2\eta_{t-k_t}^2k_t^2\mathcal{M}\left(\mathcal{M}+\lambda_1\right)^2\left\|G\right\|_2\mathbb{E}\left[\mathrm{Tr}\left(B_{t-k_t}B_{t-k_t}^T\right)\right]
\end{aligned}$$

Therefore we have,

$$\begin{aligned}
&\mathbb{E}\left[P_t\right]\\
&\leq \eta_{t-k_t}\left(\frac{2\mathcal{V}\left|\lambda_2\left(P\right)\right|}{1-\left|\lambda_2\left(P\right)\right|}+\mathcal{M}\left(2\gamma\eta_t+16\gamma\eta_{t-k_t}^2k_t\left(\mathcal{M}+\lambda_1\right)+\left(2+\left(1+\epsilon\right)^2\right)\eta_{t-k_t}k_t^2\left(\mathcal{M}+\lambda_1\right)^2\right)\right)\left\|G\right\|_2\\
&\qquad\times\mathbb{E}\left[\mathrm{Tr}\left(B_{t-k_t}B_{t-k_t}^T\right)\right]\\
&\overset{(i)}{\leq} \eta_{t-k_t}\left(\frac{2\mathcal{V}\left|\lambda_2\left(P\right)\right|}{1-\left|\lambda_2\left(P\right)\right|}+\eta_{t-k_t}\mathcal{M}\left(2\gamma+16\gamma\eta_tk_t\left(\mathcal{M}+\lambda_1\right)+\left(2+\left(1+\epsilon\right)^2\right)k_t^2\left(\mathcal{M}+\lambda_1\right)^2\right)\right)\left\|G\right\|_2\\
&\qquad\times\mathbb{E}\left[\mathrm{Tr}\left(B_{t-k_t}B_{t-k_t}^T\right)\right]\\
&\overset{(ii)}{\leq} \eta_{t-k_t}\left(\frac{2\mathcal{V}\left|\lambda_2\left(P\right)\right|}{1-\left|\lambda_2\left(P\right)\right|}+\eta_{t-k_t}\mathcal{M}\left(2\gamma\left(1+8\epsilon\right)+\left(2+\left(1+\epsilon\right)^2\right)k_t^2\left(\mathcal{M}+\lambda_1\right)^2\right)\right)\left\|G\right\|_2\mathbb{E}\left[\mathrm{Tr}\left(B_{t-k_t}B_{t-k_t}^T\right)\right]
\end{aligned}$$

where in (i) we used $2\eta_{t-k_t}\leq\eta_t\leq\eta_{t-k_t}$ along with $\eta_tk_t\left(\mathcal{M}+\lambda_1\right)\leq\epsilon$ in (ii). Hence proved. □

Lemma S.11. *Let $\forall i\in[n],\eta_ik_i\left(\mathcal{M}+\lambda_1\right)\leq\epsilon,\epsilon\in(0,1)$ and η_i forms a non-increasing sequence. Set $k_i:=\tau_{mix}\left(\gamma\eta_i^2\right),\gamma\in(0,1]$. Let $U\in\mathbb{R}^{d\times d}$ be a constant matrix and $Q_t:=\mathrm{Tr}\left(B_{t-1}B_{t-1}^T(A_t-\Sigma)U(A_t-\Sigma)\right)$. Further, let the decay of the step-sizes be slow such that $\forall i,\ \eta_i\leq\eta_{i-k_i}\leq2\eta_i$. Then*

$$\mathbb{E}\left[Q_t\right]\leq\left(\mathcal{V}+\eta_{t-k_t+1}\mathcal{M}^2\left(2\gamma\eta_t+2\left(1+\epsilon\right)\left(1+\epsilon\left(1+\epsilon\right)\right)k_t\left(\mathcal{M}+\lambda_1\right)\right)\right)\left\|U\right\|_2\mathbb{E}\left[\mathrm{Tr}\left(B_{t-k_t}B_{t-k_t}^T\right)\right]$$

where B_t is defined in 2.

Proof. Let $B_t=\left(I+R\right)B_{t-k_t}$ with $\left\|R\right\|_2\leq r$. Then,

$$\begin{aligned}
\mathbb{E}\left[Q_t\right] = \mathbb{E}\Bigg[\underbrace{\mathrm{Tr}\left(B_{t-k_t}B_{t-k_t}^T(A_t-\Sigma)U(A_t-\Sigma)\right)}_{Q_{t,1}}\Bigg]+\mathbb{E}\Bigg[\underbrace{\mathrm{Tr}\left(B_{t-k_t}B_{t-k_t}^TR^T(A_t-\Sigma)U(A_t-\Sigma)\right)}_{Q_{t,2}}\Bigg]\\
+\mathbb{E}\Bigg[\underbrace{\mathrm{Tr}\left(RB_{t-k_t}B_{t-k_t}^T(A_t-\Sigma)U(A_t-\Sigma)\right)}_{Q_{t,3}}\Bigg]+\mathbb{E}\Bigg[\underbrace{\mathrm{Tr}\left(RB_{t-k_t}B_{t-k_t}^TR^T(A_t-\Sigma)U(A_t-\Sigma)\right)}_{Q_{t,4}}\Bigg]
\end{aligned}$$

Let's consider each of the terms above. Using Von-Neumann's trace inequality and noting that $\left\|\mathbb{E}_\pi\left[\left(A_t-\Sigma\right)^2\right]\right\|_2\leq\mathcal{V}$, we have

$$\begin{aligned}
\mathbb{E}\left[Q_{t,1}\right] &= \mathbb{E}\left[\mathrm{Tr}\left(B_{t-k_t}B_{t-k_t}^T\mathbb{E}\left[(A_t-\Sigma)U(A_t-\Sigma)|s_1,s_2,\ldots,s_{t-k_t}\right]\right)\right]\\
&= \mathbb{E}\left[\mathrm{Tr}\left(B_{t-k_t}B_{t-k_t}^T\mathbb{E}\left[(A_t-\Sigma)U(A_t-\Sigma)|s_{t-k_t}\right]\right)\right]\\
&\leq \left\|\mathbb{E}\left[(A_t-\Sigma)U(A_t-\Sigma)|s_{t-k_t}\right]\right\|_2\mathbb{E}\left[\mathrm{Tr}\left(B_{t-k_t}B_{t-k_t}^T\right)\right]\\
&\leq \left\|U\right\|_2\left\|\mathbb{E}\left[\left(A_t-\Sigma\right)^2|s_{t-k_t}\right]\right\|_2\mathbb{E}\left[\mathrm{Tr}\left(B_{t-k_t}B_{t-k_t}^T\right)\right] \text{ using 37}\\
&\leq \left\|U\right\|_2\left(\mathcal{V}+2d_{\text{mix}}\left(k_t\right)\mathcal{M}^2\right)\mathbb{E}\left[\mathrm{Tr}\left(B_{t-k_t}B_{t-k_t}^T\right)\right]\\
&\leq \left\|U\right\|_2\left(\mathcal{V}+2\gamma\eta_t^2\mathcal{M}^2\right)\mathbb{E}\left[\mathrm{Tr}\left(B_{t-k_t}B_{t-k_t}^T\right)\right]
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}\left[Q_{t,2}\right] &= \mathbb{E}\left[\mathrm{Tr}\left(B_{t-k_t}B_{t-k_t}^T\mathbb{E}\left[R^T(A_t-\Sigma)U(A_t-\Sigma)|s_1,s_2,\ldots,s_{t-k_t}\right]\right)\right] \\
&\leq \left\|\mathbb{E}\left[R^T(A_t-\Sigma)U(A_t-\Sigma)|s_{t-k_t}\right]\right\|_2 \mathbb{E}\left[\mathrm{Tr}\left(B_{t-k_t}B_{t-k_t}^T\right)\right] \\
&\leq (1+\epsilon)\,\eta_{t-k_t+1}k_t\mathcal{M}^2\left(\mathcal{M}+\lambda_1\right)\|U\|_2\,\mathbb{E}\left[\mathrm{Tr}\left(B_{t-k_t}B_{t-k_t}^T\right)\right] \quad \text{using Lemma 2}
\end{aligned}$$

$$\mathbb{E}\left[Q_{t,3}\right] \leq (1+\epsilon)\,\eta_{t-k_t+1}k_t\mathcal{M}^2\left(\mathcal{M}+\lambda_1\right)\|U\|_2\,\mathbb{E}\left[\mathrm{Tr}\left(B_{t-k_t}B_{t-k_t}^T\right)\right] \text{ using a similar argument as } Q_{t,2}$$

$$\begin{aligned}
\mathbb{E}\left[Q_{t,4}\right] &= \mathbb{E}\left[\mathrm{Tr}\left(RB_{t-k_t}B_{t-k_t}^TR^T(A_t-\Sigma)U(A_t-\Sigma)\right)\right] \\
&= \mathbb{E}\left[\mathrm{Tr}\left(B_{t-k_t}B_{t-k_t}^TR^T(A_t-\Sigma)U(A_t-\Sigma)R\right)\right] \\
&\leq r^2\|U\|_2\,\mathcal{M}^2\mathbb{E}\left[\mathrm{Tr}\left(B_{t-k_t}B_{t-k_t}^T\right)\right] \\
&\leq (1+\epsilon)^2\,\eta_{t-k_t+1}^2k_t^2\mathcal{M}^2\left(\mathcal{M}+\lambda_1\right)^2\|U\|_2\,\mathbb{E}\left[\mathrm{Tr}\left(B_{t-k_t}B_{t-k_t}^T\right)\right] \text{ using Lemma 2}
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&\mathbb{E}\left[Q_t\right] \\
&\leq \left(\mathcal{V}+\eta_{t-k_t+1}\left(2\gamma\eta_t\mathcal{M}^2+2\left(1+\epsilon\right)k_t\mathcal{M}^2\left(\mathcal{M}+\lambda_1\right)+\left(1+\epsilon\right)^2\eta_{t-k_t+1}k_t^2\mathcal{M}^2\left(\mathcal{M}+\lambda_1\right)^2\right)\right)\|U\|_2\,\mathbb{E}\left[\mathrm{Tr}\left(B_{t-k_t}B_{t-k_t}^T\right)\right. \\
&\overset{(i)}{\leq} \left(\mathcal{V}+\eta_{t-k_t+1}\mathcal{M}^2\left(2\gamma\eta_t+2\left(1+\epsilon\right)k_t\left(\mathcal{M}+\lambda_1\right)+2\epsilon\left(1+\epsilon\right)^2k_t\left(\mathcal{M}+\lambda_1\right)\right)\right)\|U\|_2\,\mathbb{E}\left[\mathrm{Tr}\left(B_{t-k_t}B_{t-k_t}^T\right)\right] \\
&= \left(\mathcal{V}+\eta_{t-k_t+1}\mathcal{M}^2\left(2\gamma\eta_t+2\left(1+\epsilon\right)\left(1+\epsilon\left(1+\epsilon\right)\right)k_t\left(\mathcal{M}+\lambda_1\right)\right)\right)\|U\|_2\,\mathbb{E}\left[\mathrm{Tr}\left(B_{t-k_t}B_{t-k_t}^T\right)\right]
\end{aligned}$$

In (i), we used the slow-decay assumption on η_i mentioned in the lemma statement along with $\eta_i k_i\left(\mathcal{M}+\lambda_1\right)\leq\epsilon$. Hence proved. □

Lemma S.12. *(Learning Rate Schedule) Fix any $\delta\in(0,1)$. Set $k_i := \tau_{mix}\left(\eta_i^2\right)$. Suppose the step sizes are set such that*

$$\eta_i = \frac{\alpha}{\left(\lambda_1-\lambda_2\right)\left(\beta+i\right)}$$

Define the linear function

$$\forall i\in[n],\ f\left(i\right) := \frac{1}{\eta_i} = \frac{\left(\lambda_1-\lambda_2\right)\left(\beta+i\right)}{\alpha},$$

With $\epsilon := \frac{1}{100}$ and $\xi_{k,t},\zeta_{k,t},\mathcal{V}',\overline{\mathcal{V}_{k,t}}$ defined in 58, set $\alpha>2,\ f\left(0\right)\geq e,\ m:=200$ and

$$\beta := 600\max\left\{\frac{\tau_{mix}\ln\left(f\left(0\right)\right)\left(\mathcal{M}+\lambda_1\right)\alpha}{\lambda_1-\lambda_2},\ \frac{5\tau_{mix}\ln\left(f\left(0\right)\right)\left(\mathcal{M}+\lambda_1\right)^2\alpha^2}{3\left(\lambda_1-\lambda_2\right)^2\ln\left(1+\frac{\delta}{m}\right)},\ \frac{\left(\mathcal{V}'+5\lambda_1^2\right)\alpha^2}{300\left(\lambda_1-\lambda_2\right)^2\ln\left(1+\frac{\delta}{m}\right)}\right\}$$

then we have

1. $\eta_i k_i\left(\mathcal{M}+\lambda_1\right)\leq\epsilon$
2. $\forall i,\ \eta_i\leq\eta_{i-k_i}\leq\left(1+2\epsilon\right)\eta_i\leq 2\eta_i$ *(slow-decay)*
3. $\sum_{i=1}^{n}\left(\overline{\mathcal{V}_{k,i}}+\zeta_{k,i}+4\lambda_1^2\right)\eta_i^2\leq\ln\left(1+\frac{\delta}{m}\right)$
4. $\sum_{i=1}^{n}\left(\mathcal{V}'+\xi_{k,i}\right)\eta_{i-k_i}^2\exp\left(-\sum_{j=i+1}^{n}2\eta_j\left(\lambda_1-\lambda_2\right)\right)\leq$

$$\left(\frac{2\left(1+10\epsilon\right)\alpha^2}{2\alpha-1}\right)\frac{\mathcal{V}'}{\left(\lambda_1-\lambda_2\right)^2}\frac{1}{n}+\left(\frac{24\left(1+10\epsilon\right)\alpha^3}{\left(\alpha-1\right)}\right)\frac{\mathcal{M}\left(\mathcal{M}+\lambda_1\right)^2}{\left(\lambda_1-\lambda_2\right)^3}\frac{k_n^2}{n^2}$$

Proof. We use the following inequalities -

$$\sum_{j=i}^{t} \eta_j^2 \leq \frac{\alpha^2}{(\lambda_1 - \lambda_2)^2 (\beta + i - 1)} \quad \left(\text{Using } \frac{1}{x+1} \leq \sum_{i=1}^{\infty} \frac{1}{(x+i)^2} \leq \frac{1}{x} \right) \tag{39}$$

$$\sum_{j=i}^{t} \eta_j \geq \frac{\alpha}{(\lambda_1 - \lambda_2)} \log \left(\frac{t + \beta + 1}{i + \beta} \right) \tag{40}$$

$$\sum_{j=i}^{t} \eta_j \leq \frac{\alpha}{(\lambda_1 - \lambda_2)} \log \left(\frac{t + \beta}{i + \beta - 1} \right) \tag{41}$$

$$\sum_{j=i}^{t} (j + \beta)^{\ell} \leq \frac{(t + \beta + 1)^{\ell+1} - (i + \beta)^{\ell+1}}{\ell + 1} \leq \frac{(t + \beta + 1)^{\ell+1}}{\ell + 1} \; \forall \; \ell > 0 \tag{42}$$

For the first result, we observe that $f(x) = \frac{\ln(x)}{x}$ is a decreasing function of x for $x \geq e$. Using 5, note that

$$k_i := \tau_{\text{mix}} \left(\eta_i^2 \right) \leq \frac{2\tau_{\text{mix}}}{\ln(2)} \ln \left(\frac{1}{\eta_i^2} \right) = \frac{4\tau_{\text{mix}}}{\ln(2)} \ln \left(\frac{(\beta + i)(\lambda_1 - \lambda_2)}{\alpha} \right) = \frac{4\tau_{\text{mix}}}{\ln(2)} \ln \left(f(i) \right) \tag{43}$$

for $\eta_i < 1$. For $i \geq 0$

$$f(i) \geq f(0) = \frac{\beta (\lambda_1 - \lambda_2)}{\alpha} \geq e$$

Therefore,

$$\begin{aligned} \eta_i k_i (\mathcal{M} + \lambda_1) &\leq \frac{4\tau_{\text{mix}} (\mathcal{M} + \lambda_1)}{\ln(2)} \frac{\alpha}{(\beta + i)(\lambda_1 - \lambda_2)} \ln \left(\frac{(\beta + i)(\lambda_1 - \lambda_2)}{\alpha} \right) \\ &= \frac{4\tau_{\text{mix}} (\mathcal{M} + \lambda_1)}{\ln(2)} \frac{\ln(f(i))}{f(i)} \\ &\leq \frac{4\tau_{\text{mix}} (\mathcal{M} + \lambda_1)}{\ln(2)} \frac{\ln(f(0))}{f(0)} \end{aligned}$$

From the assumptions mentioned in the Lemma statement, we have

$$\frac{\ln(f(0))}{f(0)} < \frac{\epsilon \ln(2)}{4\tau_{\text{mix}} (\mathcal{M} + \lambda_1)} = \frac{\ln(2)}{400\tau_{\text{mix}} (\mathcal{M} + \lambda_1)} \tag{44}$$

Therefore,

$$\forall \; i, \eta_i k_i (\mathcal{M} + \lambda_1) \leq \epsilon \tag{45}$$

For the second result, we note that $\forall i \in [n]$,

$$\begin{aligned} \frac{\eta_{i-k_i}}{\eta_i} &= \frac{\beta + i}{\beta + i - k_i} \\ &= 1 + \frac{k_i}{\beta + i - k_i} \\ &= 1 + \frac{1}{\frac{\beta + i}{k_i} - 1} \end{aligned}$$

Consider the fraction $\frac{\beta+i}{k_i}$. We can simplify it as :

$$\begin{aligned}\frac{\beta+i}{k_i} &\geq \frac{\ln(2)}{4\tau_{\text{mix}}} \frac{\beta+i}{\ln\left(\frac{(\beta+i)(\lambda_1-\lambda_2)}{\alpha}\right)} \\ &= \frac{\alpha\ln(2)}{4\tau_{\text{mix}}(\lambda_1-\lambda_2)} \frac{f(i)}{\ln(f(i))} \\ &\geq \frac{\alpha\ln(2)}{4\tau_{\text{mix}}(\lambda_1-\lambda_2)} \frac{f(0)}{\ln(f(0))} \\ &\geq \frac{1}{\epsilon} \text{ from } 44\end{aligned}$$

where we used the fact that $\frac{x}{\ln(x)}$ is an increasing function for $x \geq e$. Therefore, we have that

$$\begin{aligned}\frac{\eta_{i-k_i}}{\eta_i} &\leq 1 + \frac{1}{\frac{1}{\epsilon}-1} \\ &= \frac{1}{1-\epsilon} \\ &\leq 1+2\epsilon \text{ for } \epsilon \in (0, 0.1)\end{aligned}$$

For the third result, we note that

$$\begin{aligned}&\zeta_{k,t} := 40k_{t+1}\left(\mathcal{M}+\lambda_1\right)^2, \\ &\xi_{k,t} := 2\eta_t\mathcal{M}\left[3+9k_{t+1}^2\left(\mathcal{M}+\lambda_1\right)^2\right] \\ &\quad \leq 24\eta_t\mathcal{M}\left[k_{t+1}^2\left(\mathcal{M}+\lambda_1\right)^2\right] \text{ since } (\mathcal{M}+\lambda_1) \geq 1 \text{ WLOG} \\ &\quad \leq 24\epsilon(1+\epsilon)k_{t+1}\left(\mathcal{M}+\lambda_1\right)^2 \text{since } \eta_t \leq (1+2\epsilon)\eta_{t+1} \text{ and } \eta_{t+1}k_{t+1}\left(\mathcal{M}+\lambda_1\right) \leq \epsilon\end{aligned}$$

Therefore,

$$\begin{aligned}\sum_{i=1}^{n}\left(\overline{\mathcal{V}_{k,i}}+\zeta_{k,i}\right)\eta_i^2 &= \left(\mathcal{V}'+5\lambda_1^2\right)\sum_{i=1}^{n}\eta_i^2 + 41\left(\mathcal{M}+\lambda_1\right)^2\sum_{i=1}^{n}\eta_i^2k_{i+1} \\ &\overset{(i)}{\leq} \underbrace{\left(\mathcal{V}'+5\lambda_1^2\right)\sum_{i=1}^{n}\eta_i^2}_{T_1} + 45\left(\mathcal{M}+\lambda_1\right)^2\underbrace{\sum_{i=1}^{n}\eta_{i+1}^2k_{i+1}}_{T_2}\end{aligned} \tag{46}$$

where (i) follows from the slow decay property of η_i.

For T_1, using 39 we have,

$$T_1 \leq \frac{\alpha^2}{\left(\lambda_1-\lambda_2\right)^2\beta} \tag{47}$$

For T_2, substituting the value of k_i from 43 for $\eta_i < 1$ we have,

$$T_2 := \sum_{i=1}^{n}\eta_{i+1}^2k_{i+1} \leq \frac{4\tau_{\text{mix}}}{\ln(2)}\sum_{i=1}^{n}\left(\frac{\alpha}{(\lambda_1-\lambda_2)(\beta+i+1)}\right)^2\ln\left(\frac{(\lambda_1-\lambda_2)(\beta+i+1)}{\alpha}\right) \tag{48}$$

$$= \frac{4\tau_{\text{mix}}}{\ln(2)}\sum_{i=1}^{n}\frac{\ln(f(i+1))}{f(i+1)^2} \tag{49}$$

Note that $f(i)$ is a linear function of i and $\forall i\ f(i+1)-f(i) = \frac{\lambda_1-\lambda_2}{\alpha}$. We observe that $g(x) = \frac{\ln(x)}{x^2}$ is a decreasing function of x for $x \geq e^{\frac{1}{2}} \sim 1.65$. Therefore,

$$\left(\frac{\lambda_1-\lambda_2}{\alpha}\right)\sum_{i=1}^{n}\frac{\ln(f(i+1))}{f(i+1)^2} \leq \int_{f(1)}^{f(n+1)}\frac{\ln(x)}{x^2}\,dx$$

Substituting in 49 we have,

$$\begin{aligned}
T_2 &\leq \frac{4\tau_{\text{mix}}}{\ln(2)}\left(\frac{\alpha}{\lambda_1-\lambda_2}\right)\int_{f(1)}^{f(n+1)}\frac{\ln(x)}{x^2}\,dx \\
&= \frac{4\tau_{\text{mix}}}{\ln(2)}\left(\frac{\alpha}{\lambda_1-\lambda_2}\right)\left(-\left(\frac{\ln(x)}{x}+\frac{1}{x}\right)\Bigg|_{f(1)}^{f(n)}\right) \\
&\leq \frac{4\tau_{\text{mix}}}{\ln(2)}\left(\frac{\alpha}{\lambda_1-\lambda_2}\right)\left(\frac{\ln(f(1))}{f(1)}+\frac{1}{f(1)}\right) \\
&\leq \frac{8\tau_{\text{mix}}}{\ln(2)}\left(\frac{\alpha}{\lambda_1-\lambda_2}\right)\left(\frac{\ln(f(1))}{f(1)}\right) \\
&\leq \frac{8\tau_{\text{mix}}}{\ln(2)}\left(\frac{\alpha}{\lambda_1-\lambda_2}\right)\left(\frac{\ln(f(0))}{f(0)}\right) \text{ since } \frac{\ln(x)}{x} \text{ is a decreasing function of } x \text{ for } x \geq e
\end{aligned}$$

Putting everything together in 46 and using the bounds on $\beta, f(0)$ mentioned in the lemma statement, we have,

$$\begin{aligned}
\sum_{i=1}^{n}\left(\overline{\mathcal{V}_{k,i}}+\zeta_{k,i}\right)\eta_i^2 &\leq 460\left(\mathcal{M}+\lambda_1\right)^2\tau_{\text{mix}}\left(\frac{\alpha}{\lambda_1-\lambda_2}\right)\frac{\ln(f(0))}{f(0)}+\frac{\alpha^2}{\left(\lambda_1-\lambda_2\right)^2\beta}\left(\mathcal{V}'+5\lambda_1^2\right) \\
&= 460\tau_{\text{mix}}\ln(f(0))\frac{\alpha^2}{\left(\lambda_1-\lambda_2\right)^2\beta}\left(\mathcal{M}+\lambda_1\right)^2+\frac{\alpha^2}{\left(\lambda_1-\lambda_2\right)^2\beta}\left(\mathcal{V}'+5\lambda_1^2\right) \\
&\leq \ln\left(1+\frac{\delta}{m}\right)
\end{aligned}$$

Finally, for the last result we first note that

$$\begin{aligned}
\xi_{k,t} &:= 2\eta_t\mathcal{M}\left[3+9k_{t+1}^2\left(\mathcal{M}+\lambda_1\right)^2\right] \\
&\leq 24\eta_t\mathcal{M}\left[k_{t+1}^2\left(\mathcal{M}+\lambda_1\right)^2\right] \text{ since } \left(\mathcal{M}+\lambda_1\right)\geq 1 \text{ WLOG}
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\sum_{i=1}^{n}\left(\mathcal{V}'+\xi_{k,i}\right)\eta_{i-k_i}^2\exp\left(-\sum_{j=i+1}^{n}2\eta_j\left(\lambda_1-\lambda_2\right)\right) \\
&\quad\leq (1+2\epsilon)^2\sum_{i=1}^{n}\left(\mathcal{V}'+\xi_{k,i}\right)\eta_i^2\exp\left(-\sum_{j=i+1}^{n}2\eta_j\left(\lambda_1-\lambda_2\right)\right) \\
&\quad\leq (1+5\epsilon)\sum_{i=1}^{n}\left(\mathcal{V}'+\xi_{k,i}\right)\eta_i^2\exp\left(-\sum_{j=i+1}^{n}2\eta_j\left(\lambda_1-\lambda_2\right)\right) \text{ since } \epsilon\in(0,0.1) \\
&\quad= (1+5\epsilon)\left[\sum_{i=1}^{n}\mathcal{V}'\eta_i^2\exp\left(-\sum_{j=i+1}^{n}2\eta_j\left(\lambda_1-\lambda_2\right)\right)+\sum_{i=1}^{n}\xi_{k,i}\eta_i^2\exp\left(-\sum_{j=i+1}^{n}2\eta_j\left(\lambda_1-\lambda_2\right)\right)\right]
\end{aligned} \tag{50}$$

Let's define

$$g(i) := \exp\left(-\sum_{j=i+1}^{n}2\eta_j\left(\lambda_1-\lambda_2\right)\right),\quad T_3 := \sum_{i=1}^{n}\eta_i^2 g(i),\quad T_4 := \sum_{i=1}^{n}\eta_i^3 g(i),\quad T_5 := \sum_{i=1}^{n}\eta_i^3 k_i^2 g(i),$$

Note that since $k_n \geq k_i$,

$$T_5 = \sum_{i=1}^{n} \eta_i^3 k_i^2 g(i) \leq k_n^2 \sum_{i=1}^{n} \eta_i^3 g(i) = k_n^2 T_4$$

Then,

$$\begin{aligned}\sum_{i=1}^{n} (\mathcal{V}' + \xi_{k,i}) \eta_{i-k_i}^2 \exp\left(-\sum_{j=i+1}^{n} 2\eta_j (\lambda_1 - \lambda_2)\right) &\leq (1+5\epsilon)\left[\mathcal{V}' T_3 + 24\mathcal{M}(\mathcal{M}+\lambda_1)^2 T_5\right] \\ &\leq (1+5\epsilon)\left[\mathcal{V}' T_3 + 24\mathcal{M}(\mathcal{M}+\lambda_1)^2 k_n^2 T_4\right]\end{aligned} \tag{51}$$

Using 40, $g(i) \leq \left(\frac{i+\beta+1}{n+\beta+1}\right)^{2\alpha}$. Noting that $\left(\frac{\beta+1}{\beta}\right)^2 \leq \left(\frac{\beta+1}{\beta}\right)^3 \leq 2$, we have

$$\begin{aligned}T_3 &:= \sum_{i=1}^{n} \eta_i^2 \exp\left(-2\sum_{j=i+1}^{n} \eta_j (\lambda_1 - \lambda_2)\right) \\ &= \left(\frac{\alpha}{\lambda_1 - \lambda_2}\right)^2 \sum_{i=1}^{n} \frac{1}{(\beta+i)^2} \left(\frac{i+\beta+1}{n+\beta+1}\right)^{2\alpha} \\ &\leq \left(\frac{\alpha}{\lambda_1 - \lambda_2}\right)^2 \left(\frac{\beta+1}{\beta}\right)^2 \sum_{i=1}^{n} \frac{1}{(\beta+i+1)^2} \left(\frac{i+\beta+1}{n+\beta+1}\right)^{2\alpha} \\ &= \left(\frac{\alpha}{\lambda_1 - \lambda_2}\right)^2 \left(\frac{\beta+1}{\beta}\right)^2 \sum_{i=1}^{n} \frac{1}{(\beta+i+1)^2} \left(\frac{i+\beta+1}{n+\beta+1}\right)^{2\alpha} \\ &\leq 2\left(\frac{\alpha}{\lambda_1 - \lambda_2}\right)^2 \frac{1}{(n+\beta+1)^{2\alpha}} \sum_{i=1}^{n} (i+\beta+1)^{2\alpha-2} \\ &\leq \frac{2}{2\alpha-1}\left(\frac{\alpha}{\lambda_1 - \lambda_2}\right)^2 \frac{1}{(n+\beta+2)} \left(\frac{n+\beta+2}{n+\beta+1}\right)^{2\alpha} \text{ using 42} \\ &= \frac{2}{2\alpha-1}\left(\frac{\alpha}{\lambda_1 - \lambda_2}\right)^2 \frac{1}{(n+\beta+2)} \left(1 + \frac{1}{n+\beta+1}\right)^{2\alpha}\end{aligned} \tag{52}$$

and similarly,

$$\begin{aligned}T_4 &:= \sum_{i=1}^{n} \eta_i^3 \exp\left(-2\sum_{j=i+1}^{n} \eta_j (\lambda_1 - \lambda_2)\right) \\ &= \left(\frac{\alpha}{\lambda_1 - \lambda_2}\right)^3 \sum_{i=1}^{n} \frac{1}{(\beta+i)^3} \left(\frac{i+\beta+1}{n+\beta+1}\right)^{2\alpha} \\ &\leq \left(\frac{\alpha}{\lambda_1 - \lambda_2}\right)^3 \left(\frac{\beta+1}{\beta}\right)^3 \sum_{i=1}^{n} \frac{1}{(\beta+i+1)^3} \left(\frac{i+\beta+1}{n+\beta+1}\right)^{2\alpha} \\ &= \left(\frac{\alpha}{\lambda_1 - \lambda_2}\right)^3 \left(\frac{\beta+1}{\beta}\right)^3 \sum_{i=1}^{n} \frac{1}{(\beta+i+1)^2} \left(\frac{i+\beta+1}{n+\beta+1}\right)^{2\alpha} \\ &\leq 2\left(\frac{\alpha}{\lambda_1 - \lambda_2}\right)^3 \frac{1}{(n+\beta+1)^{2\alpha}} \sum_{i=1}^{n} (i+\beta+1)^{2\alpha-3}\end{aligned}$$

$$\leq \frac{1}{\alpha-1}\left(\frac{\alpha}{\lambda_1-\lambda_2}\right)^3 \frac{1}{(n+\beta+2)^2}\left(\frac{n+\beta+2}{n+\beta+1}\right)^{2\alpha} \text{ using 42}$$
$$= \frac{1}{\alpha-1}\left(\frac{\alpha}{\lambda_1-\lambda_2}\right)^3 \frac{1}{(n+\beta+2)^2}\left(1+\frac{1}{n+\beta+1}\right)^{2\alpha} \tag{53}$$

Using 45, we have

$$\frac{\alpha}{n+\beta+1} = \eta_n(\lambda_1-\lambda_2) \leq \eta_n\lambda_1 \leq \eta_n k_n \lambda_1 \leq \epsilon \leq 0.1 \tag{54}$$

Therefore, using [17]

$$\left(1+\frac{1}{n+\beta+1}\right)^{2\alpha} \overset{(i)}{\leq} \frac{1}{1-\frac{2\alpha}{n+\beta+1}} \overset{(ii)}{\leq} 1+\frac{4\alpha}{n+\beta+1} \leq 1+4\epsilon \tag{55}$$

where (i) follows since $\frac{2\alpha}{n+\beta+1} < 1$ by 54 and (ii) follows since $\frac{1}{1-x} \leq 1+2x$ for $x \in [0, \frac{1}{2}]$.

Using 55 with 52, we have

$$T_3 \leq \frac{2}{2\alpha-1}\left(\frac{\alpha}{\lambda_1-\lambda_2}\right)^2 \frac{1}{(n+\beta+2)}\left(1+\frac{4\alpha}{n+\beta+1}\right)$$
$$\leq \frac{2(1+4\epsilon)}{2\alpha-1}\left(\frac{\alpha}{\lambda_1-\lambda_2}\right)^2 \frac{1}{(n+\beta+2)} \tag{56}$$

Using 55 with 53, we have

$$T_4 \leq \frac{1+4\epsilon}{\alpha-1}\left(\frac{\alpha}{\lambda_1-\lambda_2}\right)^3 \frac{1}{(n+\beta+2)^2} \tag{57}$$

Let

$$C_1 := \frac{2(1+10\epsilon)\alpha^2}{2\alpha-1}, C_2 := \frac{24(1+10\epsilon)\alpha^3}{(\alpha-1)},$$

Putting together 56, 57 in 51 and using the definition of k_i in 43 we have

$$(1+5\epsilon)\mathcal{V}'T_3 \leq \frac{2(1+5\epsilon)(1+4\epsilon)}{2\alpha-1}\left(\frac{\alpha}{\lambda_1-\lambda_2}\right)^2 \frac{\mathcal{V}'}{(n+\beta+2)}$$
$$\leq \frac{2(1+10\epsilon)\alpha^2}{2\alpha-1}\frac{\mathcal{V}'}{(\lambda_1-\lambda_2)^2}\frac{1}{n} \text{ since } \epsilon \leq 0.05$$

and similarly,

$$24(1+5\epsilon)\mathcal{M}(\mathcal{M}+\lambda_1)^2 k_n^2 T_4 \leq \frac{24(1+5\epsilon)(1+4\epsilon)\alpha^3}{\alpha-1}\frac{\mathcal{M}(\mathcal{M}+\lambda_1)^2}{(\lambda_1-\lambda_2)^3}\frac{k_n^2}{n^2}$$

Therefore from 51, we have

$$\sum_{i=1}^{n}(\mathcal{V}'+\xi_{k,i})\eta_{i-k_i}^2 \exp\left(-\sum_{j=i+1}^{n} 2\eta_j(\lambda_1-\lambda_2)\right) \leq C_1\frac{\mathcal{V}'}{(\lambda_1-\lambda_2)^2}\frac{1}{n} + C_2\frac{\mathcal{M}(\mathcal{M}+\lambda_1)^2}{(\lambda_1-\lambda_2)^3}\frac{k_n^2}{n^2}$$

Hence proved. □

S.4 Proofs : Convergence Analysis of Oja's Algorithm for Markovian Data

In this section, we present proofs of Theorems 2, 3, 4 and 5. We state versions of these theorems that are valid under more general conditions on the step sizes. Specifically, for the following, we only require a sequence of non-increasing step-sizes which satisfy, for $\epsilon := \frac{1}{100}, \forall i \in [n]$ -

1. $\eta_i k_i \left(\mathcal{M} + \lambda_1\right) \leq \epsilon$
2. $\eta_i \leq \eta_{i-k_i} \leq (1+2\epsilon)\,\eta_i \leq 2\eta_i$ (slow-decay)

The version of these theorems stated in the manuscript are obtained by plugging in the step-sizes as $\eta_i := \frac{\alpha}{(\lambda_1-\lambda_2)(\beta+i)}$ for the values of α, β provided in Lemma S.12. Before starting with the proofs, we define the following scalar variables -

$$\begin{aligned} r &:= 2\left(1+\epsilon\right) k_n \eta_n \left(\mathcal{M}+\lambda_1\right), \qquad \zeta_{k,t} := 40 k_{t+1} \left(\mathcal{M}+\lambda_1\right)^2 \\ \psi_{k,t} &:= 6\mathcal{M}\left[1 + 3k_{t+1}^2 \left(\mathcal{M}+\lambda_1\right)^2\right], \qquad \xi_{k,t} := \eta_{t-k_t}\psi_{k,t} \\ \mathcal{V}' &:= \frac{1+(3+4\epsilon)\left|\lambda_2\left(P\right)\right|}{1-\left|\lambda_2\left(P\right)\right|}\mathcal{V}, \qquad \overline{\mathcal{V}_{k,t}} := \mathcal{V}' + \lambda_1^2 + \xi_{k,t} \end{aligned} \tag{58}$$

Theorem 2. *(General Version)*

$$\mathbb{E}\left[v_1^T B_{n,1} B_{n,1}^T v_1\right] \leq \left(1+r\right)^2 \exp\left(\sum_{t=1}^{n-k_n} \left(2\eta_t\lambda_1 + \eta_t^2\left(\mathcal{V}' + \lambda_1^2 + \xi_{k,t}\right)\right)\right)$$

where $B_{j,i}$ is defined in 12.

Proof. Define $\alpha_{n,t} := \mathbb{E}\left[\mathrm{Tr}\left(v_1^T B_{n,t} B_{n,t}^T v_1\right)\right] = \mathbb{E}\left[v_1^T B_{n,t} B_{n,t}^T v_1\right], i \leq t \leq n$. Then, we have

$$\begin{aligned} v_1^T B_{n,t} B_{n,t}^T v_1 &= v_1^T B_{n,t+1}(I+\eta_t\Sigma)^2 B_{n,t+1}^T v_1 \;+\; 2\eta_t \underbrace{\left(v_1^T B_{n,t+1}(I+\eta_t\Sigma)(A_t-\Sigma)B_{n,t+1}^T v_1\right)}_{P_{n,t}} \\ &+\; \eta_t^2 \underbrace{\left(v_1^T B_{n,t+1}(A_t-\Sigma)^2 B_{n,t+1}^T v_1\right)}_{Q_{n,t}} \\ &\leq v_1^T B_{j,t+1} B_{j,t+1}^T v_1 ((1+\eta_t\lambda_1)^2) + \eta_t^2 Q_{n,t} + 2\eta_t P_{n,t} \end{aligned} \tag{59}$$

Using Lemma S.8 with $U = v_1, G = \left(I+\eta_t\Sigma\right), \gamma = 1$ and noting that $\mathbb{E}_\pi\left[A_t - \Sigma\right] = 0$, along with observing that $\alpha_{n,t+k_{t+1}} \leq \alpha_{n,t+k_t}$ from Lemma S.3, we have

$$\left|\mathbb{E}\left[P_{n,t}\right]\right| \leq \eta_{t+1}\left(1+\eta_t\lambda_1\right)\left(\frac{2\mathcal{V}\left|\lambda_2\left(P\right)\right|}{1-\left|\lambda_2\left(P\right)\right|} + \eta_{t+1}\mathcal{M}\left(2 + 16\epsilon + \left(2+\left(1+\epsilon\right)^2\right)k_{t+1}^2\left(\mathcal{M}+\lambda_1\right)^2\right)\right)\alpha_{n,t+k_t}$$

We note that $\forall i,\ k_i \geq 1$, therefore, using the assumption in 58, $1+\eta_t\lambda_1 \leq 1+\eta_t k_t\left(\mathcal{M}+\lambda_1\right) \leq 1+\epsilon$.

Next, using Lemma S.9 with $U = v_1, G = I, \gamma = 1$ and noting that $\left\|\mathbb{E}_\pi\left[\left(A_t-\Sigma\right)^2\right]\right\|_2 \leq \mathcal{V}$ along with observing that $\alpha_{n,t+k_{t+1}} \leq \alpha_{n,t+k_t}$ using Lemma S.3, we have

$$\begin{aligned} \left|\mathbb{E}\left[Q_{n,t}\right]\right| &\leq \left(\mathcal{V} + \eta_{t+1}\mathcal{M}^2\left(2\eta_{t+1} + \left(1+\epsilon\right)\left(2+\epsilon\left(1+\epsilon\right)\right)k_{t+1}\left(\mathcal{M}+\lambda_1\right)\right)\right)\alpha_{n,t+k_t} \\ &\leq \left(\mathcal{V} + 2\epsilon\eta_{t+1}\mathcal{M} + \eta_{t+1}\mathcal{M}^2\left(\left(1+\epsilon\right)\left(2+\epsilon\left(1+\epsilon\right)\right)k_{t+1}\left(\mathcal{M}+\lambda_1\right)\right)\right)\alpha_{n,t+k_t} \end{aligned}$$

where in the last line, we used $\eta_{t+1}\mathcal{M} \leq \eta_{t+1}\left(\mathcal{M}+\lambda_1\right) \leq \eta_{t+1}k_{t+1}\left(\mathcal{M}+\lambda_1\right) \leq \epsilon$.

Then from 59 for $n - k_t \geq t \geq 1$,

$$\alpha_{n,t} \leq \left(1+\eta_t\lambda_1\right)^2\alpha_{n,t+1} + \left(\frac{1+(3+4\epsilon)\left|\lambda_2\left(P\right)\right|}{1-\left|\lambda_2\left(P\right)\right|}\right)\mathcal{V}\eta_t^2\alpha_{n,t+k_t} + C_{k,t}\eta_t^3\alpha_{n,t+k_t} \tag{60}$$

where $C_{k,t}$ is defined as

$$\begin{aligned}
C_{k,t} &:= \mathcal{M}\left[4\left(1+\epsilon\right)\left(1+8\epsilon\right)+2\epsilon+k_{t+1}\left(\mathcal{M}+\lambda_1\right)\left(\left(1+\epsilon\right)\left(2+\epsilon\left(1+\epsilon\right)\right)\mathcal{M}+2\left(2+\left(1+\epsilon\right)^2\right)k_{t+1}\left(\mathcal{M}+\lambda_1\right)\right)\right] \\
&\overset{(i)}{\leq} \mathcal{M}\left[4\left(1+\epsilon\right)\left(1+8\epsilon\right)+2\epsilon+\left(\left(1+\epsilon\right)\left(2+\epsilon\left(1+\epsilon\right)\right)+2\left(2+\left(1+\epsilon\right)^2\right)\right)k_{t+1}^2\left(\mathcal{M}+\lambda_1\right)^2\right] \\
&= \mathcal{M}\left[4+38\epsilon+32\epsilon^2+\left(6+2\epsilon+\left(1+\epsilon\right)^2\left(1+2\epsilon\right)\right)k_{t+1}^2\left(\mathcal{M}+\lambda_1\right)^2\right]
\end{aligned}$$

where in (i) we used $\mathcal{M} \leq k_{t+1}\left(\mathcal{M}+\lambda_1\right)$.

Then recalling the definition of $\xi_{k,t}$ in 58, and noting that $\alpha_{n,t+k_t} \leq \alpha_{n,t+1}$ using Lemma S.3 we have from 60,

$$\begin{aligned}
\alpha_{n,t} &\leq \left(1+\eta_t\lambda_1\right)^2\alpha_{n,t+1}+\left(\left(\frac{1+\left(3+4\epsilon\right)\left|\lambda_2\left(P\right)\right|}{1-\left|\lambda_2\left(P\right)\right|}\right)\mathcal{V}+\xi_{k,t}\right)\eta_t^2\alpha_{n,t+k_t} \\
&= \left(1+2\eta_t\lambda_1+\eta_t^2\left(\left(\frac{1+\left(3+4\epsilon\right)\left|\lambda_2\left(P\right)\right|}{1-\left|\lambda_2\left(P\right)\right|}\right)\mathcal{V}+\lambda_1^2+\xi_{k,t}\right)\right)\alpha_{n,t+1}
\end{aligned}$$

Therefore using this recursion, we have,

$$\alpha_{n,1} \leq \alpha_{n,n-k_n+1}\exp\left(2\lambda_1\sum_{t=1}^{n-k_n}\eta_t+\sum_{t=1}^{n-k_n}\eta_t^2\left(\left(\frac{1+\left(3+4\epsilon\right)\left|\lambda_2\left(P\right)\right|}{1-\left|\lambda_2\left(P\right)\right|}\right)\mathcal{V}+\lambda_1^2+\xi_{k,t}\right)\right)$$

Let $B_{n,n-k_n+1} = I + R'$, where $\|R'\| \leq r$ a.s.

$$\begin{aligned}
\alpha_{n,n-k_n+1} &= \mathbb{E}\left[v_1^T B_{n,n-k_n+1}B_{n,n-k_n+1}^T v_1\right] \\
&= \mathbb{E}\left[v_1^T v_1\right]+\mathbb{E}\left[v_1^T\left(R'+R'^T\right)v_1\right]+\mathbb{E}\left[v_1^T R'R'^T v_1\right] \\
&\leq 1+2r+r^2
\end{aligned}$$

Using Lemma 2 we have

$$\begin{aligned}
r &\leq \left(1+\epsilon\right)k_n\eta_{n-k_n+1}\left(\mathcal{M}+\lambda_1\right) \\
&\leq \left(1+\epsilon\right)k_n\eta_{n-k_n}\left(\mathcal{M}+\lambda_1\right) \\
&\leq 2\left(1+\epsilon\right)k_n\eta_n\left(\mathcal{M}+\lambda_1\right) \quad \text{since } \eta_{n-k_n} \leq 2\eta_n
\end{aligned}$$

Therefore,

$$\alpha_{n,1} \leq \left(1+2r+r^2\right)\exp\left(2\lambda_1\sum_{t=1}^{n-k_n}\eta_t+\sum_{t=1}^{n-k_n}\eta_t^2\left(\left(\frac{1+\left(3+4\epsilon\right)\left|\lambda_2\left(P\right)\right|}{1-\left|\lambda_2\left(P\right)\right|}\right)\mathcal{V}+\lambda_1^2+\xi_{k,t}\right)\right)$$

Hence proved. □

Theorem 3. *(General Version) Let* $u := \min\left\{i : i \in [n], i-k_i \geq 0\right\}$. *Then,*

$$\begin{aligned}
\mathbb{E}\left[\mathrm{Tr}\left(V_\perp^T B_n B_n^T V_\perp\right)\right] \leq \left(1+5\epsilon\right)\exp&\left(\sum_{i=u+1}^{n}2\eta_i\lambda_2+\left(\mathcal{V}'+\lambda_1^2+\xi_{k,i}\right)\eta_{i-k_i}^2\right) \\
&\times\left(d+\sum_{i=u+1}^{n}\left(\mathcal{V}'+\xi_{k,i}\right)C'_{k,i}\eta_{i-k_i}^2\exp\left(\sum_{j=u+1}^{i}2\eta_j\left(\lambda_1-\lambda_2\right)\right)\right)
\end{aligned}$$

where $C'_{k,t} := \exp\left(2\lambda_1\sum_{j=1}^{u}\left(\eta_j-\eta_{t-u+j}\right)+\sum_{j=1}^{t-u}\eta_j^2\left(\overline{\mathcal{V}_{k,j}}-\overline{\mathcal{V}_{k,j+u}}\right)\right)$ *and* B_t *is defined in 2.*

Proof. For $t \leq n$, let

$$\alpha_t := \alpha_{t,1} = \mathbb{E}\left[v_1^T B_t B_t^T v_1\right] = \mathbb{E}\left[\mathrm{Tr}\left(v_1^T B_t B_t^T v_1\right)\right], \text{ as defined in Theorem 2}$$
$$\beta_t := \mathbb{E}\left[\mathrm{Tr}\left(V_\perp^T B_t B_t^T V_\perp\right)\right]$$

Note that $\alpha_t + \beta_t = \mathrm{Tr}\left(B_t B_t^T\right)$ by definition. Then,

$$\begin{aligned}
\mathrm{Tr}\left(B_t B_t^T V_\perp V_\perp^T\right) &= \mathrm{Tr}\left(B_{t-1} B_{t-1}^T (I+\eta_t\Sigma) V_\perp V_\perp^T (I+\eta_t\Sigma)\right) + \eta_t \,\mathrm{Tr}\left(B_{t-1}^T (I+\eta_t\Sigma) V_\perp V_\perp^T (A_t-\Sigma) B_{t-1}\right) \\
&\quad + \eta_t \,\mathrm{Tr}\left(B_{t-1}^T (A_t-\Sigma) V_\perp V_\perp^T (I+\eta_t\Sigma) B_{t-1}\right) + \eta_t^2 \,\mathrm{Tr}\left(B_{t-1} B_{t-1}^T (A_t-\Sigma) V_\perp V_\perp^T (A_t-\Sigma)\right) \\
&\leq (1+\eta_t\lambda_2)^2 \,\mathrm{Tr}\left(B_{t-1} B_{t-1}^T V_\perp V_\perp^T\right) + 2\eta_t \underbrace{\mathrm{Tr}\left(B_{t-1} B_{t-1}^T (I+\eta_t\Sigma) V_\perp V_\perp^T (A_t-\Sigma)\right)}_{P_t} \\
&\quad + \eta_t^2 \underbrace{\mathrm{Tr}\left(B_{t-1} B_{t-1}^T (A_t-\Sigma) V_\perp V_\perp^T (A_t-\Sigma)\right)}_{Q_t}
\end{aligned}$$

Let $B_{t-1} = (I+R) B_{t-k_t}$ with $\|R\|_2 \leq r$. Using Lemma S.10 with $G = (I+\eta_t\Sigma) V_\perp V_\perp^T = V_\perp (I+\eta_t\Lambda_\perp) V_\perp^T, \gamma = 1$, where $\Lambda_\perp$ is a $d-1 \times d-1$ diagonal matrix of eigenvalues $\lambda_2, \ldots, \lambda_d$ of Σ, and noting that $\left\|V_\perp V_\perp^T\right\|_2 = 1$,

$$\begin{aligned}
\mathbb{E}\left[P_t\right] &\leq (1+\eta_t\lambda_1)\,\eta_{t-k_t}\left(\frac{2\mathcal{V}\left|\lambda_2\left(P\right)\right|}{1-\left|\lambda_2\left(P\right)\right|} + \eta_{t-k_t}\mathcal{M}\left(2\left(1+8\epsilon\right) + \left(2+\left(1+\epsilon\right)^2\right)k_t^2\left(\mathcal{M}+\lambda_1\right)^2\right)\right)\left(\alpha_{t-k_t}+\beta_{t-k_t}\right) \\
&\leq (1+\epsilon)\,\eta_{t-k_t}\left(\frac{2\mathcal{V}\left|\lambda_2\left(P\right)\right|}{1-\left|\lambda_2\left(P\right)\right|} + \eta_{t-k_t}\mathcal{M}\left(2\left(1+8\epsilon\right) + \left(2+\left(1+\epsilon\right)^2\right)k_t^2\left(\mathcal{M}+\lambda_1\right)^2\right)\right)\left(\alpha_{t-k_t}+\beta_{t-k_t}\right)
\end{aligned}$$

where in the last line, we used $\eta_t\lambda_1 \leq \eta_t k_t\left(\mathcal{M}+\lambda_1\right) \leq \epsilon$.

Using Lemma S.11 with $U = V_\perp V_\perp^T, \gamma = 1$,

$$\begin{aligned}
\mathbb{E}\left[Q_t\right] &\leq \left(\mathcal{V} + \eta_{t-k_t+1}\mathcal{M}^2\left(2\eta_t + 2\left(1+\epsilon\right)\left(1+\epsilon\left(1+\epsilon\right)\right)k_t\left(\mathcal{M}+\lambda_1\right)\right)\right)\left(\alpha_{t-k_t}+\beta_{t-k_t}\right) \\
&\overset{(i)}{\leq} \left(\mathcal{V} + 2\epsilon\eta_t\mathcal{M} + 2\eta_{t-k_t+1}\mathcal{M}^2\left(\left(1+\epsilon\right)\left(1+\epsilon\left(1+\epsilon\right)\right)k_t\left(\mathcal{M}+\lambda_1\right)\right)\right)\left(\alpha_{t-k_t}+\beta_{t-k_t}\right) \\
&\overset{(ii)}{\leq} \left(\mathcal{V} + 2\epsilon\eta_t\mathcal{M} + 2\eta_{t-k_t+1}\mathcal{M}\left(\left(1+\epsilon\right)\left(1+\epsilon\left(1+\epsilon\right)\right)k_t^2\left(\mathcal{M}+\lambda_1\right)^2\right)\right)\left(\alpha_{t-k_t}+\beta_{t-k_t}\right)
\end{aligned}$$

where in (i) we used $\forall i,\ \eta_i\mathcal{M} \leq \eta_i k_i\left(\mathcal{M}+\lambda_1\right) \leq \epsilon$ and in (ii) we used $\mathcal{M} \leq k_t\left(\mathcal{M}+\lambda_1\right)$. Putting everything together, we have,

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{Tr}\left(B_t B_t^T V_\perp V_\perp^T\right)\right] \\
&\leq \left(1+\eta_t\lambda_2\right)^2\beta_{t-1} \\
&\quad + 2\left(1+\epsilon\right)\eta_t\eta_{t-k_t}\left(\frac{2\mathcal{V}\left|\lambda_2\left(P\right)\right|}{1-\left|\lambda_2\left(P\right)\right|} + \eta_{t-k_t}\mathcal{M}\left(2\left(1+8\epsilon\right) + \left(2+\left(1+\epsilon\right)^2\right)k_t^2\left(\mathcal{M}+\lambda_1\right)^2\right)\right)\left(\alpha_{t-k_t}+\beta_{t-k_t}\right) \\
&\quad + \eta_t^2\left(\mathcal{V} + 2\epsilon\eta_t\mathcal{M} + 2\eta_{t-k_t+1}\mathcal{M}\left(\left(1+\epsilon\right)\left(1+\epsilon\left(1+\epsilon\right)\right)k_t^2\left(\mathcal{M}+\lambda_1\right)^2\right)\right)\left(\alpha_{t-k_t}+\beta_{t-k_t}\right) \\
&\leq \left(1+\eta_t\lambda_2\right)^2\beta_{t-1} \\
&\quad + 2\left(1+\epsilon\right)\eta_{t-k_t}^2\left(\frac{2\mathcal{V}\left|\lambda_2\left(P\right)\right|}{1-\left|\lambda_2\left(P\right)\right|} + \eta_{t-k_t}\mathcal{M}\left(2\left(1+8\epsilon\right) + \left(2+\left(1+\epsilon\right)^2\right)k_t^2\left(\mathcal{M}+\lambda_1\right)^2\right)\right)\left(\alpha_{t-k_t}+\beta_{t-k_t}\right) \\
&\quad + \eta_{t-k_t}^2\left(\mathcal{V} + 2\epsilon\eta_t\mathcal{M} + 2\eta_{t-k_t+1}\mathcal{M}\left(\left(1+\epsilon\right)\left(1+\epsilon\left(1+\epsilon\right)\right)k_t^2\left(\mathcal{M}+\lambda_1\right)^2\right)\right)\left(\alpha_{t-k_t}+\beta_{t-k_t}\right) \\
&\leq \left(1+\eta_t\lambda_2\right)^2\beta_{t-1} + \eta_{t-k_t}^2\left(\left(\frac{1+\left(3+4\epsilon\right)\left|\lambda_2\left(P\right)\right|}{1-\left|\lambda_2\left(P\right)\right|}\right)\mathcal{V} + \xi_{k,t}\right)\left(\alpha_{t-k_t}+\beta_{t-k_t}\right)
\end{aligned}$$

where $\xi_{k,t}$ is as defined in 58. Therefore using Lemma S.4,

$$\mathbb{E}\left[\mathrm{Tr}\left(B_t B_t^T V_\perp V_\perp^T\right)\right] \leq \left(1+2\eta_t\lambda_2+\eta_{t-k_t}^2\left(\left(\frac{1+(3+4\epsilon)\left|\lambda_2\left(P\right)\right|}{1-\left|\lambda_2\left(P\right)\right|}\right)\mathcal{V}+\lambda_2^2+\xi_{k,t}\right)\right)\beta_{t-1} + \eta_{t-k_t}^2\left(\left(\frac{1+(3+4\epsilon)\left|\lambda_2\left(P\right)\right|}{1-\left|\lambda_2\left(P\right)\right|}\right)\mathcal{V}+\xi_{k,t}\right)\alpha_{t-1} \tag{61}$$

Let $\chi_\epsilon := 1+4\epsilon\left(1+\epsilon\right)\left(1+\epsilon+\epsilon^2\right) \leq 1.05$. From Theorem 2 denoting

$$r_{k,t} := 1+4\left(1+\epsilon\right)\eta_{t-1}k_{t-1}\left(\mathcal{M}+\lambda_1\right)+4\left(1+c\right)^2\eta_{t-1}^2k_{t-1}^2\left(\mathcal{M}+\lambda_1\right)^2 \leq 1+4\epsilon\left(1+\epsilon\right)\left(1+\epsilon+\epsilon^2\right)=\chi_\epsilon, \tag{62}$$

we have,

$$\alpha_{t-1} \leq r_{k,t}\exp\left(2\lambda_1\sum_{i=1}^{t-k_t-1}\eta_t+\sum_{i=1}^{t-k_t-1}\eta_i^2\left(\left(\frac{1+(3+4\epsilon)\left|\lambda_2\left(P\right)\right|}{1-\left|\lambda_2\left(P\right)\right|}\right)\mathcal{V}+\lambda_1^2+\xi_{k,i}\right)\right)$$

Now, we note the definition of $\overline{\mathcal{V}_{k,t}}$ and $\mathcal{V}'$ as mentioned in 58 -

$$\begin{aligned}\overline{\mathcal{V}_{k,t}} &:= \left(\frac{1+(3+4\epsilon)\left|\lambda_2\left(P\right)\right|}{1-\left|\lambda_2\left(P\right)\right|}\right)\mathcal{V}+\lambda_1^2+\xi_{k,t}\\ &= \mathcal{V}'+\lambda_1^2+\xi_{k,t}\end{aligned}$$

Therefore using 61,

$$\beta_t \leq \left(1+2\eta_t\lambda_2+\eta_{t-k_t}^2\overline{\mathcal{V}_{k,t}}\right)\beta_{t-1}+\eta_{t-k_t}^2 r_{k,t}\left(\mathcal{V}'+\xi_{k,t}\right)\exp\left(2\lambda_1\sum_{i=1}^{t-k_t-1}\eta_i+\sum_{i=1}^{t-k_t-1}\eta_i^2\overline{\mathcal{V}_{k,i}}\right)$$

Recursing on the above inequality for $u < t \leq n$ where $u = \min\left\{i : i \in [n], i-k_i \geq 0\right\}$, we have,

$$\begin{aligned}\beta_n &\leq \beta_u\exp\left(2\sum_{i=u+1}^{n}\eta_i\lambda_2+\sum_{i=u+1}^{n}\overline{\mathcal{V}_{k,i}}\eta_{i-k_i}^2\right)\\ &\quad+\sum_{i=u+1}^{n}r_{k,i}\left(\mathcal{V}'+\xi_{k,i}\right)\eta_{i-k_i}^2\exp\left(\sum_{j=i+1}^{n}\left(2\eta_j\lambda_2+\overline{\mathcal{V}_{k,j}}\eta_{j-k_j}^2\right)\right)\exp\left(\sum_{j=1}^{i-k_i}2\eta_j\lambda_1+\overline{\mathcal{V}_{k,j}}\eta_j^2\right)\\ &\leq \exp\left(\sum_{i=u+1}^{n}2\eta_i\lambda_2+\overline{\mathcal{V}_{k,i}}\eta_{i-k_i}^2\right)\\ &\quad\times\left(\beta_u+\sum_{i=u+1}^{n}r_{k,i}\left(\mathcal{V}'+\xi_{k,i}\right)\eta_{i-k_i}^2\exp\left(\sum_{j=1}^{i-k_i}(2\eta_j\lambda_1+\overline{\mathcal{V}_{k,j}}\eta_j^2)-\sum_{j=u+1}^{i}\left(2\eta_j\lambda_2+\overline{\mathcal{V}_{k,j}}\eta_{j-k_j}^2\right)\right)\right)\end{aligned}$$

Now, since $k_i, k_j \geq k_u = u$, therefore, we have

$$\begin{aligned}\beta_n &\leq \exp\left(\sum_{i=u+1}^{n}2\eta_i\lambda_2+\overline{\mathcal{V}_{k,i}}\eta_{i-k_i}^2\right)\times\\ &\quad\left(\beta_u+\sum_{i=u+1}^{n}r_{k,i}\left(\mathcal{V}'+\xi_{k,i}\right)\eta_{i-k_i}^2\exp\left(\sum_{j=1}^{i-u}(2\eta_j\lambda_1+\overline{\mathcal{V}_{k,j}}\eta_j^2)-\sum_{j=u+1}^{i}\left(2\eta_j\lambda_2+\overline{\mathcal{V}_{k,j}}\eta_{j-u}^2\right)\right)\right)\end{aligned}$$

Recall that $C'_{k,i} := \exp\left(2\lambda_1\sum_{j=1}^{u}\left(\eta_j-\eta_{i-u+j}\right)+\sum_{j=1}^{i-u}\eta_j^2\left(\overline{\mathcal{V}_{k,j}}-\overline{\mathcal{V}_{k,j+u}}\right)\right)$ as defined in 58.

Therefore,

$$\beta_n \leq \exp\left(\sum_{i=u+1}^{n} 2\eta_i\lambda_2 + \overline{\mathcal{V}_{k,i}}\eta_{i-k_i}^2\right) \times \left(\beta_u + \sum_{i=u+1}^{n} r_{k,i}\left(\mathcal{V}' + \xi_{k,i}\right) C'_{k,i}\eta_{i-k_i}^2 \exp\left(\sum_{j=u+1}^{i} 2\eta_j\left(\lambda_1 - \lambda_2\right)\right)\right)$$

Let $B_u = I + R'$ with $\|R'\| \leq r'$ a.s. Using Lemma 2 we have

$$\begin{aligned} r' &\leq (1+\epsilon)\, k_u\eta_1\left(\mathcal{M} + \lambda_1\right) \\ &\leq (1+\epsilon)\, k_u\eta_0\left(\mathcal{M} + \lambda_1\right) \\ &\leq 2\,(1+\epsilon)\, k_u\eta_u\left(\mathcal{M} + \lambda_1\right) \quad \text{since } \eta_0 = \eta_{u-k_u} \leq 2\eta_u \\ &< 2\epsilon\,(1+\epsilon) \end{aligned}$$

Therefore,

$$\begin{aligned} \beta_u &= \mathbb{E}\left[\mathrm{Tr}\left(V_\perp^T B_u B_u^T V_\perp\right)\right] \\ &= \mathbb{E}\left[\mathrm{Tr}\left(V_\perp^T V_\perp\right)\right] + \mathbb{E}\left[\mathrm{Tr}\left(V_\perp^T (R' + R'^T) V_\perp\right)\right] + \mathbb{E}\left[\mathrm{Tr}\left(V_\perp^T R' R'^T V_\perp\right)\right] \\ &\leq d\left(1 + 2r' + r'^2\right) \\ &\leq d\left(1 + 4\epsilon\,(1+\epsilon) + 4\epsilon^2\,(1+\epsilon)^2\right) \\ &= d\left(1 + 4\epsilon\,(1+\epsilon)\left(1 + \epsilon + \epsilon^2\right)\right) \\ &= \chi_\epsilon d \end{aligned}$$

The proof follows by noting that $r_{k,t} \leq \chi_\epsilon$ as shown in 62. □

Theorem 4. *(General Version)*

$$\mathbb{E}\left[v_1^T B_{n,1} B_{n,1}^T v_1\right] \geq (1-t)\exp\left(\sum_{i=1}^{n-k_n} 2\eta_i\lambda_1 - \sum_{i=1}^{n-k_n} 4\eta_i^2\lambda_1^2\right)$$

where $t := 2r + s, s := 3\,(1+r)^2 \exp\left(2\lambda_1^2 \sum_{i=1}^{n} \eta_i^2\right) \sum_{t=1}^{n-k_n} W_{k,t}\eta_t^2 \exp\left(\sum_{i=t+1}^{n-k_n} \eta_i^2\right), W_{k,t} := \mathcal{V}' + \xi_{k,t}$ *and* $B_{j,i}$ *has been defined in 12.*

Proof. We will start will expanding the quantity of interest using Eq 59.

$$\alpha_{n,t} = \mathbb{E}\left[v_1^T B_{n,t} B_{n,t}^T v_1\right] \geq \mathbb{E}\left[v_1^T B_{n,t+1}\left(I + \eta_t\Sigma\right)^2 B_{n,t+1}^T v_1 + 2\eta_t P_{n,t}\right] \tag{63}$$

where $P_{n,t}$ has been defined in Theorem 2. Let's define

$$\begin{aligned} &\mathcal{S}_t := \prod_{i=t}^{1}\left(I + \eta_i\Sigma\right)\prod_{i=1}^{t}\left(I + \eta_i\Sigma\right),\ \mathcal{S}_0 = I \ \text{ and} \\ &\delta_{n,t} := \mathbb{E}\left[v_1^T B_{n,t+1}\mathcal{S}_t B_{n,t+1}^T v_1\right] \end{aligned}$$

Note that $\delta_{n,0} = \alpha_{n,1}$. First we bound $\delta_{n,n-k_n}$. Let $B_{n,n-k_n} = I + R'$. By Lemma 2 along with the slow-decay assumption on the step-sizes, we know that $\|R'\|_2 \leq r := 2\,(1+\epsilon)\,\eta_n k_n\left(\mathcal{M} + \lambda_1\right)$ a.s. Then,

$$\delta_{n,n-k_n} - \prod_{i=1}^{n-k_n}\left(1 + \eta_i\lambda_1\right)^2 \geq -2\left|E[v_1^T R'\mathcal{S}_{n-k_n} v_1]\right| \geq -2r\prod_{i=1}^{n-k_n}\left(1 + \eta_i\lambda_1\right)^2$$

Therefore,

$$\begin{aligned}\delta_{n,n-k_n} &\geq \prod_{i=1}^{n-k_n}\left(1+\eta_i\lambda_1\right)^2(1-2r)\\ &= (1-2r)\left\|\mathcal{S}_{n-k_n}\right\|_2 \end{aligned} \tag{64}$$

Now using 63, we have

$$\delta_{n,t-1} \geq \delta_{n,t} + 2\eta_t\mathbb{E}\left[\underbrace{v_1^T B_{n,t+1}\left(I+\eta_t\Sigma\right)\mathcal{S}_{t-1}\left(A_t-\Sigma\right)B_{n,t+1}^T v_1}_{U_t}\right]$$

First, observe that $\mathcal{S}_{t-1} = U\Lambda U^T$, where U denotes a matrix of eigenvectors of Σ, and Λ is a PSD diagonal matrix. Since $I+\eta_t\Sigma = U\Lambda' U^T$ for some other PSD diagonal matrix Λ', the product will also be PSD.

By using Lemma S.8 with $U = v_1, G = \left(I+\eta_t\Sigma\right)\mathcal{S}_{t-1}, \gamma = 1$ and noting that $\mathbb{E}_\pi\left[A_t-\Sigma\right] = 0$, we have

$$\begin{aligned}\left|\mathbb{E}\left[U_t\right]\right| &\leq \left(1+\eta_t\lambda_1\right)\eta_{t+1}\left\|\mathcal{S}_{t-1}\right\|_2\left(\frac{2\mathcal{V}\left|\lambda_2\left(P\right)\right|}{1-\left|\lambda_2\left(P\right)\right|} + \eta_{t+1}\mathcal{M}\left(2\left(1+8\epsilon\right)+\left(2+\left(1+\epsilon\right)^2\right)k_{t+1}^2\left(\mathcal{M}+\lambda_1\right)^2\right)\right)\alpha_{n,t+k_{t+1}}\\ &\leq \left(1+\epsilon\right)\eta_{t+1}\left\|\mathcal{S}_{t-1}\right\|_2 W_{k,t}\alpha_{n,t+1}\end{aligned}$$

where $W_{k,t} = \mathcal{V}' + \xi_{k,t}$. Therefore,

$$\delta_{n,t-1} \geq \delta_{n,t} - 2\left(1+\epsilon\right)W_{k,t}\eta_t^2\alpha_{n,t+1}\left\|\mathcal{S}_{t-1}\right\|_2 \text{ for } t \leq n-k_n$$

Let

$$\mathcal{V}' := \left(\frac{1+\left(3+4\epsilon\right)\left|\lambda_2\left(P\right)\right|}{1-\left|\lambda_2\left(P\right)\right|}\right)\mathcal{V}$$

as defined in 58. Unwinding the recursion for $t \leq n-k_n$, we have,

$$\begin{aligned}\delta_{n,0} &\geq \delta_{n,n-k_n} - 2\left(1+\epsilon\right)\sum_{t=1}^{n-k_n} W_{k,t}\eta_t^2\alpha_{n,t+1}\left\|\mathcal{S}_{t-1}\right\|_2\\ &\geq \left(1-2r\right)\left\|\mathcal{S}_{n-k_n}\right\|_2 - 2\left(1+\epsilon\right)\left(1+r\right)^2\sum_{t=1}^{n-k_n} W_{k,t}\eta_t^2\exp\left(2\lambda_1\sum_{i=t+1}^{n-k_n}\eta_i + \sum_{i=t+1}^{n-k_n}\eta_i^2\left(\mathcal{V}'+\lambda_1^2+C_{k,i}\right)\right)\left\|\mathcal{S}_{t-1}\right\|_2\end{aligned}$$

where second step followed from Theorem 2 and 64.

Using the inequalities $\forall x \in \mathbb{R},\ 1+x \leq e^x$ and $\forall x \in \mathbb{R},\ x \geq 0, 1+x \geq e^{x-x^2}, \quad \forall t$ we have,

$$\begin{aligned}\left\|\mathcal{S}_t\right\|_2 &= \prod_{i=1}^{t}(1+\eta_i\lambda_1)^2 \leq \exp\left(2\lambda_1\sum_{i=1}^{t}\eta_i\right), \text{ and}\\ \left\|\mathcal{S}_t\right\|_2 &= \prod_{i=1}^{t}(1+\eta_i\lambda_1)^2 \geq \exp\left(2\lambda_1\sum_{i=1}^{t}\eta_i - 4\lambda_1^2\sum_{i=1}^{t}\eta_i^2\right)\end{aligned}$$

Therefore denoting $\theta_\epsilon := 2\left(1+\epsilon\right)\exp\left(2\lambda_1^2\sum_{i=1}^n \eta_i^2\right)$, we have

$$\begin{aligned}
\delta_{n,0} &\geq \exp\left(2\lambda_1\sum_{i=1}^{n-k_n}\eta_i - 4\lambda_1^2\sum_{i=1}^{n-k_n}\eta_i^2\right)\left[\left(1-2r\right) - \theta_\epsilon\left(1+r\right)^2\sum_{t=1}^{n-k_n} W_{k,t}\eta_t^2\exp\left(\sum_{i=t+1}^{n-k_n}\eta_i^2\left(\mathcal{V}' + \lambda_1^2 + C_{k,i}\right)\right)\right] \\
&\geq \exp\left(2\lambda_1\sum_{i=1}^{n-k_n}\eta_i - 4\lambda_1^2\sum_{i=1}^{n-k_n}\eta_i^2\right)\left[\left(1-2r\right) - \theta_\epsilon\left(1+r\right)^2\sum_{t=1}^{n-k_n} W_{k,t}\eta_t^2\exp\left(\sum_{i=t+1}^{n-k_n}\eta_i^2\left(\mathcal{V}' + \lambda_1^2 + C_{k,i}\right)\right)\right] \\
&\geq \exp\left(2\lambda_1\sum_{i=1}^{n-k_n}\eta_i - 4\lambda_1^2\sum_{i=1}^{n-k_n}\eta_i^2\right)\left[1 - \left(2r + \theta_\epsilon\left(1+r\right)^2\sum_{t=1}^{n-k_n} W_{k,t}\eta_t^2\exp\left(\sum_{i=t+1}^{n-k_n}\eta_i^2\overline{\mathcal{V}_{k,i}}\right)\right)\right]
\end{aligned}$$

where $\overline{\mathcal{V}_{k,i}}$ is defined in 58. Hence proved. □

Theorem 5. *(General Version)*

$$\mathbb{E}\left[\left(v_1^T B_{n,1} B_{n,1}^T v_1\right)^2\right] \leq \left(1+r\right)^4\exp\left(\sum_{i=1}^{n-k_n} 4\eta_i\lambda_1 + \sum_{i=1}^{n-k_n}\eta_i^2\zeta_{k,i}\right)$$

where $B_{j,i}$ has been defined in 12.

Proof. Define $Q_{n,t} := v_1^T B_{n,t+1}(A_t-\Sigma)^2 B_{n,t+1}^T v_1$, and $P_{n,t} := v_1^T B_{n,t+1}(I+\eta_t\Sigma)(A_t-\Sigma)B_{n,t+1}^T v_1$. Using 59, we have, for $n \geq t \geq 1$,

$$\begin{aligned}
0 \leq v_1^T B_{n,t} B_{n,t}^T v_1 &= v_1^T B_{n,t+1}(I+\eta_t\Sigma)^2 B_{n,t+1}^T v_1 + \eta_t^2 Q_{n,t} + 2\eta_t P_{n,t} \\
&\leq v_1^T B_{j,t+1} B_{j,t+1}^T v_1 (1+\eta_t\lambda_1)^2 + \eta_t^2\mathcal{M}^2\left(v_1^T B_{n,t+1}B_{n,t+1}^T v_1\right) + 2\eta_t P_{n,t} \\
&\leq v_1^T B_{j,t+1} B_{j,t+1}^T v_1 \underbrace{\left((1+\eta_t\lambda_1)^2 + \eta_t^2\mathcal{M}^2\right)}_{c_t} + 2\eta_t P_{n,t}
\end{aligned}$$

Thus, we have -

$$\begin{aligned}
\kappa_{n,t} := \mathbb{E}\left[(v_1^T B_{n,t}B_{n,t}^T v_1)^2\right] &\leq \mathbb{E}\left[\left(c_t v_1^T B_{n,t+1}B_{n,t+1}^T v_1 + 2\eta_t P_{n,t}\right)^2\right] \\
&\leq c_t^2\kappa_{n,t+1} + 4\eta_t^2\mathbb{E}\left[P_{n,t}^2\right] + 4c_t\eta_t\mathbb{E}\left[\left(v_1^T B_{n,t+1}B_{n,t+1}^T v_1\right)P_{n,t}\right]
\end{aligned} \tag{65}$$

Note that,

$$\begin{aligned}
\mathbb{E}\left[P_{n,t}^2\right] &\leq \mathbb{E}\left[\left(v_1^T B_{n,t+1}(I+\eta_t\Sigma)(A_t-\Sigma)B_{n,t+1}^T v_1\right)^2\right] \\
&\leq (1+\eta_t\lambda_1)^2\mathcal{M}^2\mathbb{E}\left[\left(v_1^T B_{n,t+1}B_{n,t+1}^T v_1\right)^2\right] \\
&= (1+\eta_t\lambda_1)^2\mathcal{M}^2\kappa_{n,t+1}
\end{aligned}$$

Now we work on the cross-term. For the convenience of notation, let's denote $k := k_{t+1}$ unless otherwise specified. Let $B_{n,t+1} = B_{n,t+k}\left(I+R\right)$ with,

$$\|R\|_2 \leq (1+c)\eta_{t+1}k(\mathcal{M}+\lambda_1) =: r_t \leq \epsilon\left(1+\epsilon\right)$$

Using Lemma 2, we have

$$\begin{aligned}
\underbrace{|v_1^T B_{n,t+1}B_{n,t+1}^T v_1 - v_1^T B_{n,t+k}B_{n,t+k}^T v_1|}_{Y_1} &= |v_1^T B_{n,t+k}(R+R^T+RR^T)B_{n,t+k}^T v_1| \\
&\leq |v_1^T B_{n,t+k}B_{n,t+k}^T v_1|\left(2r_t + r_t^2\right)
\end{aligned} \tag{66}$$

We will also bound

$$
\begin{aligned}
&\underbrace{|v_1^T B_{n,t+1}(I+\eta_t\Sigma)(A_t-\Sigma)B_{n,t+1}^T v_1 - v_1^T B_{n,t+k}(I+\eta_t\Sigma)(A_t-\Sigma)B_{n,t+k}^T v_1|}_{Y_2} \\
&\quad = |v_1^T B_{n,t+k} R(I+\eta_t\Sigma)(A_t-\Sigma)(I+R^T)B_{n,t+k}^T v_1 + v_1^T B_{n,t+k}(I+\eta_t\Sigma)(A_t-\Sigma)R^T B_{n,t+k}^T v_1| \\
&\quad \le (2r_t + r_t^2)\left(1+\eta_t\lambda_1\right)\mathcal{M}|v_1^T B_{n,t+k}B_{n,t+k}^T v_1| \qquad (67)
\end{aligned}
$$

So, now we have:

$$
\begin{aligned}
&\mathbb{E}\left[(v_1^T B_{n,t+1}B_{n,t+1}^T v_1 P_{n,t})\right] \\
&= \mathbb{E}\left[(v_1^T B_{n,t+1}B_{n,t+1}^T v_1)(v_1^T B_{n,t+1}(I+\eta_t\Sigma)(A_t-\Sigma)B_{n,t+1}^T v_1)\right] \\
&= \mathbb{E}\left[(Y_1 + v_1^T B_{n,t+k}B_{n,t+k}^T v_1)(Y_2 + v_1^T B_{n,t+k}(I+\eta_t\Sigma)(A_t-\Sigma)B_{n,t+k}^T v_1)\right] \\
&= \underbrace{\mathbb{E}\left[Y_1 Y_2\right]}_{T_1} + \underbrace{\mathbb{E}\left[Y_1 v_1^T B_{n,t+k}(I+\eta_t\Sigma)(A_t-\Sigma)B_{n,t+k}^T v_1\right]}_{T_2} + \underbrace{\mathbb{E}\left[Y_2 v_1^T B_{n,t+k}B_{n,t+k}^T v_1\right]}_{T_3} \\
&\quad + \underbrace{\mathbb{E}\left[(v_1^T B_{n,t+k}B_{n,t+k}^T v_1)(v_1^T B_{n,t+k}(I+\eta_t\Sigma)(A_t-\Sigma)B_{n,t+k}^T v_1)\right]}_{T_4}
\end{aligned}
$$

Lets start with the last term, T_4. Using Lemma S.3 we have,

$$
\begin{aligned}
|T_4| &\le \left|\mathbb{E}\left[(v_1^T B_{n,t+k}B_{n,t+k}^T v_1)(v_1^T B_{n,t+k}(I+\eta_t\Sigma)\mathbb{E}\left[(A_t-\Sigma)|s_{t+k}\right]B_{n,t+k}^T v_1)\right]\right| \\
&\le 2(1+\eta_t\lambda_1)\mathcal{M}d_{\text{mix}}\left(k\right)\kappa_{n,t+k} \\
&\le 2\eta_{t+1}^2(1+\eta_t\lambda_1)\mathcal{M}\kappa_{n,t+k} \\
&\le 2\eta_{t+1}^2(1+\eta_t\lambda_1)\mathcal{M}\kappa_{n,t+1}
\end{aligned}
$$

Using Eqs 66 and 67 the first three terms can be bounded as:

$$
\begin{aligned}
|T_1| \le \mathbb{E}\left[|Y_1 Y_2|\right] &\le \left(2r_t + r_t^2\right)^2\left(1+\eta_t\lambda_1\right)\mathcal{M}\kappa_{n,t+k} \\
&\le \left(2r_t + r_t^2\right)^2\left(1+\eta_t\lambda_1\right)\mathcal{M}\kappa_{n,t+1} \text{ using Lemma } S.3 \\
&= \left(2+r_t\right)^2 r_t^2\left(1+\eta_t\lambda_1\right)\mathcal{M}\kappa_{n,t+1} \\
&\le \left(1+\epsilon\right)^2\left(2+\epsilon\left(1+\epsilon\right)\right)^2\left(1+\eta_t\lambda_1\right)\eta_{t+1}^2 k_{t+1}^2\mathcal{M}\left(\mathcal{M}+\lambda_1\right)^2\kappa_{n,t+1} \\
&\le \left(1+\epsilon\right)^3\left(2+\epsilon+\epsilon^2\right)^2\eta_{t+1}^2 k_{t+1}^2\mathcal{M}\left(\mathcal{M}+\lambda_1\right)^2\kappa_{n,t+1} \text{ since } \eta_t\lambda_1 \le \epsilon
\end{aligned}
$$

$$
\begin{aligned}
|T_2| \le \mathbb{E}\left[|Y_1 v_1^T B_{n,t+k}(I+\eta_t\Sigma)(A_t-\Sigma)B_{n,t+k}^T v_1|\right] &\le \left(2+r_t\right) r_t\left(1+\eta_t\lambda_1\right)\mathcal{M}\kappa_{n,t+k} \\
&\le \left(2+r_t\right) r_t\left(1+\eta_t\lambda_1\right)\mathcal{M}\kappa_{n,t+1} \text{ using Lemma } S.3 \\
&\le \left(2+\epsilon+\epsilon^2\right)\left(1+\epsilon\right)\left(1+\eta_t\lambda_1\right)\eta_{t+1}k_{t+1}\left(\mathcal{M}+\lambda_1\right)\mathcal{M}\kappa_{n,t+1} \\
&\le \left(1+\epsilon\right)^2\left(2+\epsilon+\epsilon^2\right)\eta_{t+1}k_{t+1}\mathcal{M}\left(\mathcal{M}+\lambda_1\right)\kappa_{n,t+1}
\end{aligned}
$$

and similarly,

$$
\begin{aligned}
|T_3| \le \mathbb{E}\left[Y_2 v_1^T B_{n,t+k}B_{n,t+k}^T v_1\right] &\le r_t\left(2+r_t\right)\left(1+\eta_t\lambda_1\right)\mathcal{M}\kappa_{n,t+k} \\
&\le \left(1+\epsilon\right)\left(2+\epsilon+\epsilon^2\right)\left(1+\eta_t\lambda_1\right)\eta_{t+1}k_{t+1}\mathcal{M}\left(\mathcal{M}+\lambda_1\right)\kappa_{n,t+k} \\
&\le \left(1+\epsilon\right)^2\left(2+\epsilon+\epsilon^2\right)\eta_{t+1}k_{t+1}\mathcal{M}\left(\mathcal{M}+\lambda_1\right)\kappa_{n,t+k} \\
&\le \left(1+\epsilon\right)^2\left(2+\epsilon+\epsilon^2\right)\eta_{t+1}k_{t+1}\mathcal{M}\left(\mathcal{M}+\lambda_1\right)\kappa_{n,t+1} \text{ using Lemma } S.3
\end{aligned}
$$

Note that

$$
\begin{aligned}
c_t &:= (1+\eta_t\lambda_1)^2 + \eta_t^2\mathcal{M}^2 \leq 1 + 2\epsilon + 2\epsilon^2, \text{ and} \\
c_t^2 &= \left(1 + 2\eta_t\lambda_1 + \eta_t^2\left(\mathcal{M}^2 + \lambda_1^2\right)\right)^2 \\
&= 1 + 4\eta_t^2\lambda_1^2 + \eta_t^4\left(\mathcal{M}^2+\lambda_1^2\right)^2 + 4\eta_t\lambda_1 + 4\eta_t^3\lambda_1\left(\mathcal{M}^2+\lambda_1^2\right) + 2\eta_t^2\left(\mathcal{M}^2+\lambda_1^2\right) \\
&\leq 1 + 4\eta_t\lambda_1 + \eta_t^2\left(2\mathcal{M}^2 + 6\lambda_1^2 + \epsilon\mathcal{M} + \epsilon\lambda_1\right) + 4\eta_t^3\lambda_1\left(\mathcal{M}^2+\lambda_1^2\right) \\
&\leq 1 + 4\eta_t\lambda_1 + 6\eta_t^2\left(\mathcal{M}+\lambda_1\right)^2 + 4\eta_t^3\lambda_1\left(\mathcal{M}+\lambda_1\right)^2
\end{aligned}
$$

Define

$$
\begin{aligned}
\phi_\epsilon &:= (1+\epsilon)\left(2+\epsilon+\epsilon^2\right) \\
\omega_\epsilon &:= 1 + 2\epsilon + 2\epsilon^2 \\
\zeta_{k,t} &:= \left(10 + 8\left(1+\epsilon\right) + 4\left(1+2\epsilon\right)\phi_\epsilon\right)\phi_\epsilon c_t k_{t+1}\left(\mathcal{M}+\lambda_1\right)^2
\end{aligned}
$$

Putting everything together in Eq 65, for $t \leq n - k_{t+1}$ we have,

$$
\begin{aligned}
\frac{\kappa_{n,t}}{\kappa_{n,t+1}} &\leq c_t^2 + 4\eta_t^2\left(1+\eta_t\lambda_1\right)^2\mathcal{M}^2 + 4\left(1+\epsilon\right)c_t\eta_t\mathcal{M}\left(2\phi_\epsilon\eta_{t+1}k_{t+1}\left(\mathcal{M}+\lambda_1\right) + \left(2+\phi_\epsilon^2 k_{t+1}^2\left(\mathcal{M}+\lambda_1\right)^2\right)\eta_{t+1}^2\right) \\
&\leq c_t^2 + 4\eta_t^2\left(1+\eta_t\lambda_1\right)^2\mathcal{M}^2 + 4\left(1+\epsilon\right)c_t\eta_t\mathcal{M}\left(2\phi_\epsilon\eta_t k_{t+1}\left(\mathcal{M}+\lambda_1\right) + \left(2+\phi_\epsilon^2 k_{t+1}^2\left(\mathcal{M}+\lambda_1\right)^2\right)\eta_t^2\right) \\
&= c_t^2 + 4\eta_t^2\left[\mathcal{M}^2 + 2\phi_\epsilon\left(1+\epsilon\right)c_t\mathcal{M}\left(\mathcal{M}+\lambda_1\right)k_{t+1}\right] + 4\eta_t^3\left[\left(1+2\epsilon\right)\lambda_1 + \left(1+\epsilon\right)c_t\mathcal{M}\left(2+\phi_\epsilon^2 k_{t+1}^2\left(\mathcal{M}+\lambda_1\right)^2\right)\right] \\
&\leq c_t^2 + 4\eta_t^2\left[2 + 2\phi_\epsilon\left(1+\epsilon\right)c_t k_{t+1}\right]\mathcal{M}\left(\mathcal{M}+\lambda_1\right) + 4\left(1+2\epsilon\right)\eta_t^3\left[\lambda_1 + c_t\mathcal{M}\left(2+\phi_\epsilon^2 k_{t+1}^2\left(\mathcal{M}+\lambda_1\right)^2\right)\right] \\
&\leq 1 + 4\eta_t\lambda_1 + \eta_t^2\left[10 + 8\phi_\epsilon\left(2+\epsilon\right)c_t k_{t+1}\right]\left(\mathcal{M}+\lambda_1\right)^2 + 4\left(1+2\epsilon\right)\eta_t^3\left[\lambda_1 + 2c_t\mathcal{M} + c_t\phi_\epsilon^2 k_{t+1}^2\left(\mathcal{M}+\lambda_1\right)^3\right]
\end{aligned}
$$

$$
\begin{aligned}
&\leq \exp\left(4\eta_t\lambda_1 + \eta_t^2\left(10 + 8\phi_\epsilon\left(1+\epsilon\right)c_t k_{t+1}\right)\left(\mathcal{M}+\lambda_1\right)^2 + 4\left(1+2\epsilon\right)\eta_t^3\left(\lambda_1 + c_t\mathcal{M} + 2c_t\phi_\epsilon^2 k_{t+1}^2\left(\mathcal{M}+\lambda_1\right)^3\right)\right) \\
&\leq \exp\left(4\eta_t\lambda_1 + \eta_t^2\left(10 + 8\phi_\epsilon\left(1+\epsilon\right)c_t k_{t+1}\right)\left(\mathcal{M}+\lambda_1\right)^2 + 4\epsilon\left(1+2\epsilon\right)\eta_t^2\left(2c_t + c_t\phi_\epsilon^2 k_{t+1}\left(\mathcal{M}+\lambda_1\right)^2\right)\right) \\
&\leq \exp\left(4\eta_t\lambda_1 + \eta_t^2\left(8\epsilon\left(1+2\epsilon\right)\omega_\epsilon + \left(10 + \left(8\left(1+\epsilon\right) + 4\epsilon\left(1+2\epsilon\right)\phi_\epsilon\right)\phi_\epsilon\omega_\epsilon k_{t+1}\right)\left(\mathcal{M}+\lambda_1\right)^2\right)\right) \\
&\leq \exp\left(4\eta_t\lambda_1 + \eta_t^2\left(1 + \left(10 + 20k_{t+1}\right)\left(\mathcal{M}+\lambda_1\right)^2\right)\right) \\
&\leq \exp\left(4\eta_t\lambda_1 + \eta_t^2\left(1 + \left(10 + 20k_{t+1}\right)\left(\mathcal{M}+\lambda_1\right)^2\right)\right) \\
&\leq \exp\left(4\eta_t\lambda_1 + 40\eta_t^2 k_{t+1}\left(\mathcal{M}+\lambda_1\right)^2\right) \text{ since } \left(\mathcal{M}+\lambda_1\right), k_{t+1} \geq 1
\end{aligned}
$$

Recall our definition of $k := k_{t+1}$. We can use the above recursion for $1 \leq t \leq n - k_{t+1}$. We note that $t = n - k_n$ satisfies the conditions. Therefore,

$$
\kappa_{n,1} \leq \exp\left(\sum_{i=1}^{n-k_n} 4\eta_i\lambda_1 + \sum_{i=1}^{n-k_n}\eta_i^2\zeta_{k,i}\right)\kappa_{n,n-k_n+1}
$$

Let $B_{n,n-k_n+1} = I + R'$, with $\|R'\|_2 \leq r$ a.s.

$$
\begin{aligned}
\kappa_{n,n-k_n+1} &= \mathbb{E}\left[\left(v_1^T B_{n,n-k_n+1} B_{n,n-k_n+1}^T v_1\right)^2\right] \\
&= \mathbb{E}\left[\left(v_1^T v_1 + v_1^T(R' + R'^T)v_1 + v_1^T R' R'^T v_1\right)^2\right] \\
&\leq \left(1 + 2r + r^2\right)^2 \mathbb{E}\left[\left(v_1^T v_1\right)^2\right]
\end{aligned}
$$

Using Lemma 2, we have

$$\begin{aligned} r &\leq (1+\epsilon)\, k_{n+1}\eta_{n-k_{n+1}} (\mathcal{M}+\lambda_1) \\ &\leq (1+\epsilon)\, k_n\eta_{n-k_n} (\mathcal{M}+\lambda_1) \\ &\leq 2\,(1+\epsilon)\, k_n\eta_n (\mathcal{M}+\lambda_1) \quad \text{since } \eta_{n-k_n} \leq 2\eta_n \end{aligned}$$

which completes our proof. □

S.5 Main Results : Details and Proofs

S.5.1 Proof of Theorem 1

Lemma S.13. *This lemma proves conditions required later in the proof. Let the step-sizes be set according to Lemma S.12 and $m := 200$. Define*

$$\begin{aligned} r &:= 2\,(1+\epsilon)\,\eta_n k_n (\mathcal{M}+\lambda_1)\,, \\ s &:= 3(1+r)^2 \sum_{t=1}^{n-k_n-1} W_{k,t}\eta_t^2 \exp\left(\sum_{i=t+1}^{n-k_n-1} \overline{\mathcal{V}_{k,i}}\eta_i^2\right) \end{aligned}$$

where $W_{k,t}$ is defined in Theorem 4, $\overline{\mathcal{V}_{k,i}}$ is defined in 58 and $\alpha, \beta, f(.), \delta$ are defined in Lemma S.12. Then for sufficiently large number of samples n, such that

$$\frac{n}{\ln(f(n))} > \frac{\beta}{\ln(f(0))}$$

we have

1. $2r + s \leq \frac{1}{2}$ *(72)*
2. $r = 2\,(1+\epsilon)\,\eta_n k_n (\mathcal{M}+\lambda_1) < \frac{1}{50}\frac{\delta/m}{1+\delta/m}$ *(75)*

Proof. For (1), using Lemma S.12-(3), we note that

$$\begin{aligned} s &\leq 3(1+r)^2 \sum_{t=1}^{n-k_n-1} W_{k,t}\eta_t^2 \exp\left(\sum_{i=t+1}^{n-k_n-1} \overline{\mathcal{V}_{k,i}}\eta_i^2\right) \\ &\leq 3(1+r)^2 \sum_{t=1}^{n-k_n-1} W_{k,t}\eta_t^2 \left(1+\frac{\delta}{m}\right) \\ &\leq \frac{3(1+r)^2}{100}\left(1+\frac{\delta}{m}\right)\ln\left(1+\frac{\delta}{m}\right) \qquad (68) \\ &\leq \frac{3(1+r)^2\ln(2)}{50} \text{ since } \frac{\delta}{m} < 1 \end{aligned}$$

Therefore,

$$\begin{aligned} 2r + s &\leq 2r + \frac{3\,(1+r)^2}{25} \\ &= \frac{3}{25} + \frac{56}{25}r + \frac{3}{25}r^2 \qquad (69) \end{aligned}$$

Setting $\frac{3}{25} + \frac{56}{25}r + \frac{3}{25}r^2 \leq \frac{1}{2}$, we have,

$$\begin{aligned} &\frac{3}{25} + \frac{56}{25}r + \frac{3}{25}r^2 \leq \frac{1}{2} \\ \implies\ &6r^2 + 112r - 19 \leq 0 \end{aligned}$$

which holds for $r \in \left[0, \frac{1}{10}\right]$.

For (2), using Lemma S.12 and substituting the value of $k_i := \tau_{\text{mix}}\left(\eta_i^2\right) \leq \frac{2\tau_{\text{mix}}}{\ln(2)} \ln\left(\frac{1}{\eta_i^2}\right)$ for $\eta_i < 1$, we note that

$$\begin{aligned} r &\leq \frac{8\left(1+\epsilon\right)\tau_{\text{mix}}\left(\mathcal{M}+\lambda_1\right)}{\ln\left(2\right)} \frac{\alpha}{\left(\lambda_1-\lambda_2\right)\left(\beta+n\right)} \ln\left(\frac{\left(\lambda_1-\lambda_2\right)\left(\beta+n\right)}{\alpha}\right) \\ &= \frac{8\left(1+\epsilon\right)\tau_{\text{mix}}\left(\mathcal{M}+\lambda_1\right)}{\ln\left(2\right)} \frac{\ln\left(\frac{\left(\lambda_1-\lambda_2\right)\left(\beta+n\right)}{\alpha}\right)}{\frac{\left(\lambda_1-\lambda_2\right)\left(\beta+n\right)}{\alpha}} \\ &= \frac{8\left(1+\epsilon\right)\tau_{\text{mix}}\left(\mathcal{M}+\lambda_1\right)}{\ln\left(2\right)} \frac{\ln\left(f\left(n\right)\right)}{f\left(n\right)} \end{aligned}$$

Therefore (2) holds for sufficiently large n, i.e,

$$\frac{f\left(n\right)}{\ln\left(f\left(n\right)\right)} \geq \frac{400\left(1+\frac{\delta}{m}\right)\left(1+\epsilon\right)\tau_{\text{mix}}\left(\mathcal{M}+\lambda_1\right)}{\ln\left(2\right)\frac{\delta}{m}}$$

This is satisfied if

$$\frac{n}{\ln\left(f\left(n\right)\right)} \geq \frac{400\tau_{\text{mix}}\left(1+\frac{\delta}{m}\right)\left(1+\epsilon\right)}{\ln\left(2\right)} \frac{\left(\mathcal{M}+\lambda_1\right)\alpha}{\left(\lambda_1-\lambda_2\right)\frac{\delta}{m}} \tag{70}$$

From Lemma S.12, we have

$$\frac{\beta}{\ln\left(f\left(0\right)\right)} \geq \frac{600\tau_{\text{mix}}\left(1+2\epsilon\right)^2\left(\mathcal{M}+\lambda_1\right)^2\alpha^2}{\left(\lambda_1-\lambda_2\right)^2\ln\left(1+\frac{\delta}{m}\right)} \overset{(i)}{\geq} \frac{400\tau_{\text{mix}}\left(1+\frac{\delta}{m}\right)\left(1+\epsilon\right)}{\ln\left(2\right)} \frac{\left(\mathcal{M}+\lambda_1\right)\alpha}{\left(\lambda_1-\lambda_2\right)\frac{\delta}{m}}$$

where (i) follows since $\frac{\mathcal{M}+\lambda_1}{\lambda_1-\lambda_2} > 1, \alpha > 2$ and $\ln\left(1+x\right) \leq x \ \forall x$. Therefore, $\frac{n}{\ln(f(n))} > \frac{\beta}{\ln(f(0))}$ suffices. Further, we note that (2) implies (1) for m = 200, $\delta \leq 1$. Therefore, the condition on n is sufficient for both results. Hence proved. □

Lemma S.14. *Let*

$$u := \min\left\{i : i \in [n], i - k_i \geq 0\right\}$$

where k_i is defined in Lemma S.12. Then,

$$u \leq \lfloor\beta\rfloor \leq \beta$$

Proof. Using the definition of k_i mentioned in Lemma S.12, we have

$$\begin{aligned} k_i := \tau_{\text{mix}}\left(\eta_i^2\right) &\leq \frac{2\tau_{\text{mix}}}{\ln\left(2\right)} \ln\left(\frac{1}{\eta_i^2}\right) \\ &= \frac{4\tau_{\text{mix}}}{\ln\left(2\right)} \ln\left(\frac{\left(\lambda_1-\lambda_2\right)\left(\beta+i\right)}{\alpha}\right) \end{aligned}$$

Therefore,

$$\begin{aligned} \lfloor\beta\rfloor - k_{\lfloor\beta\rfloor} &\geq \lfloor\beta\rfloor - \frac{4\tau_{\text{mix}}}{\ln\left(2\right)} \ln\left(\frac{\beta+\lfloor\beta\rfloor}{\frac{\alpha}{\lambda_1-\lambda_2}}\right) \\ &\geq \frac{\beta}{2} - \frac{4\tau_{\text{mix}}}{\ln\left(2\right)} \ln\left(\frac{2\beta}{\frac{\alpha}{\lambda_1-\lambda_2}}\right) \text{ since } \beta > 1 \\ &= \beta\left[\frac{1}{2} - \frac{4\tau_{\text{mix}}}{\ln\left(2\right)} \frac{\ln\left(2f\left(0\right)\right)}{\beta}\right], \text{ where } f\left(.\right) \text{ is defined in Lemma } S.12 \end{aligned}$$

Now, from Lemma S.12, we know that $f(0) > e$. Therefore, $\ln(2f(0)) \leq 2\ln(f(0))$. Then,

$$\lfloor\beta\rfloor - k_{\lfloor\beta\rfloor} \geq \beta\left[\frac{1}{2} - \frac{8\tau_{\text{mix}}}{\ln(2)}\frac{\ln(f(0))}{\beta}\right]$$

Again, from the conditions in Lemma S.12, we know that

$$\frac{\ln(f(0))}{\beta} \leq \frac{\epsilon}{6\tau_{\text{mix}}}\frac{\lambda_1-\lambda_2}{(\mathcal{M}+\lambda_1)\alpha} \leq \frac{1}{120\tau_{\text{mix}}} \text{ since } \alpha > 2, \frac{\lambda_1-\lambda_2}{\mathcal{M}+\lambda_1} \leq 1, \epsilon \leq \frac{1}{100}$$

Therefore,

$$\lfloor\beta\rfloor - k_{\lfloor\beta\rfloor} \geq \beta\left(\frac{1}{2} - \frac{8}{120\ln(2)}\right) \geq 0$$

Hence proved. □

S.5.1.1 Numerator

Using Theorem 3 and Markov's Inequality, we have with probability atleast $(1-\delta)$

$$\begin{aligned}&\text{Tr}\left(V_\perp^T B_n B_n^T V_\perp\right) \leq \\ &1.05\frac{\exp\left(\sum_{i=u+1}^{n} 2\eta_i\lambda_2 + \overline{\mathcal{V}_{k,i}}\eta_{i-k_i}^2\right)}{\delta}\left(d + \sum_{i=u+1}^{n}\left(\mathcal{V}' + \xi_{k,i}\right)C'_{k,i}\eta_{i-k_i}^2\exp\left(\sum_{j=u+1}^{i} 2\eta_j\left(\lambda_1-\lambda_2\right)\right)\right)\end{aligned}$$

S.5.1.2 Denominator

Using Chebyshev's Inequality we have, with probability atleast $(1-\delta)$

$$v_1^T B_n B_n^T v_1 \geq \mathbb{E}\left[v_1^T B_n B_n^T v_1\right]\left(1 - \sqrt{\frac{1}{\delta}}\sqrt{\frac{\mathbb{E}\left[\left(v_1^T B_n B_n^T v_1\right)^2\right]}{\mathbb{E}\left[v_1^T B_n B_n^T v_1\right]^2} - 1}\right) \tag{71}$$

Let $r := 2(1+\epsilon)\eta_n k_n(\mathcal{M}+\lambda_1) \leq \frac{1}{10}$. Using Theorem 3, we have

$$\mathbb{E}\left[\left(v_1^T B_n B_n^T v_1\right)^2\right] \leq (1+r)^4\exp\left(\sum_{i=1}^{n-k_n} 4\eta_i\lambda_1 + \sum_{i=1}^{n-k_n}\eta_i^2\zeta_{k,t}\right)$$

Using Theorem 4, we have

$$\mathbb{E}\left[v_1^T B_{n,1} B_{n,1}^T v_1\right] \geq \exp\left(2\lambda_1\sum_{i=1}^{n-k_n}\eta_i - 4\lambda_1^2\sum_{i=1}^{n-k_n}\eta_i^2\right)\left[1 - \left(2r + 3(1+r)^2\sum_{t=1}^{n-k_n} W_{k,t}\eta_t^2\exp\left(\sum_{i=t+1}^{n-k_n}\eta_i^2\overline{\mathcal{V}_{k,i}}\right)\right)\right]$$

Let

$$s := 3(1+r)^2\sum_{t=1}^{n-k_n} W_{k,t}\eta_t^2\exp\left(\sum_{i=t+1}^{n-k_n}\eta_i^2\overline{\mathcal{V}_{k,i}}\right)$$

Then,

$$\frac{\mathbb{E}\left[\left(v_1^T B_n B_n^T v_1\right)^2\right]}{\mathbb{E}\left[v_1^T B_n B_n^T v_1\right]^2} \leq \frac{(1+r)^4}{(1-2r-s)^2}\exp\left(\sum_{i=1}^{n-k_n}\eta_i^2\left(\zeta_{k,i} + 4\lambda_1^2\right)\right)$$

By Lemma S.13, we have that

$$2r + s \leq \frac{1}{2}. \tag{72}$$

Then, using

$$\frac{1}{(1-x)^2} \leq 1 + 6x \text{ for } x \in \left[0, \frac{1}{2}\right] \text{ and, } (1+x)^4 \leq 1 + 5x \text{ for } x \in \left[0, \frac{1}{10}\right]$$

we have,

$$\begin{aligned}
\frac{\mathbb{E}\left[\left(v_1^T B_n B_n^T v_1\right)^2\right]}{\mathbb{E}\left[v_1^T B_n B_n^T v_1\right]^2} &\leq (1+5r)(1+12r+6s)\exp\left(\sum_{i=1}^{n-k_n} \eta_i^2\left(\zeta_{k,i} + 4\lambda_1^2\right)\right) \\
&\leq \left(1 + 17r + 6s + 60r^2 + 30rs\right)\exp\left(\sum_{i=1}^{n-k_n} \eta_i^2\left(\zeta_{k,i} + 4\lambda_1^2\right)\right) \\
&\leq (1 + 22r + 12s)\exp\left(\sum_{i=1}^{n-k_n} \eta_i^2\left(\zeta_{k,i} + 4\lambda_1^2\right)\right) \text{ since } r \leq \frac{1}{10}
\end{aligned}$$

By Lemma S.12-(3), we have that

$$\exp\left(\sum_{i=1}^{n-k_n} \eta_i^2\left(\zeta_{k,i} + 4\lambda_1^2\right)\right) \leq 1 + \frac{\delta}{m} \tag{73}$$

By 68, we have that

$$\begin{aligned}
12s &\leq \frac{48(1+r)^2}{100}\left(1 + \frac{\delta}{m}\right)^2 \ln\left(1 + \frac{\delta}{m}\right) \\
&\leq \frac{3}{5}\left(1 + \frac{\delta}{m}\right)^2 \ln\left(1 + \frac{\delta}{m}\right) \text{ since } r \leq \frac{1}{10}
\end{aligned} \tag{74}$$

By Lemma S.13, we have that

$$r = 2(1+\epsilon)\eta_n k_n(\mathcal{M} + \lambda_1) < \frac{1}{50}\frac{\delta/m}{1+\delta/m} \tag{75}$$

Then,

$$\begin{aligned}
\frac{\mathbb{E}\left[\left(v_1^T B_n B_n^T v_1\right)^2\right]}{\mathbb{E}\left[v_1^T B_n B_n^T v_1\right]^2} &\leq (1 + 22r + 12s)\left(1 + \frac{\delta}{m}\right) \\
&= 1 + \frac{\delta}{m} + 22r\left(1 + \frac{\delta}{m}\right) + 12s\left(1 + \frac{\delta}{m}\right) \\
&\leq 1 + \frac{\delta}{m} + \frac{22}{50}\frac{\delta}{m} + \frac{3}{5}\left(1 + \frac{\delta}{m}\right)^3 \ln\left(1 + \frac{\delta}{m}\right) \\
&\leq 1 + \frac{\delta}{m} + \frac{22}{50}\frac{\delta}{m} + \frac{7}{10}\ln\left(1 + \frac{\delta}{m}\right) \text{ since } \delta \leq 1, m = 200 \\
&\leq 1 + \frac{\delta}{m} + \frac{22}{50}\frac{\delta}{m} + \frac{7}{10}\frac{\delta}{m} \text{ since } \forall x,\ \ln(1+x) \leq x \\
&\leq 1 + 3\frac{\delta}{m}
\end{aligned}$$

Then setting $m = 200$, from 71 we have

$$\begin{aligned}
v_1^T B_n B_n^T v_1 &\geq \exp\left(\sum_{i=1}^{n-k_n} 2\eta_i\lambda_1 - 4\eta_i^2\lambda_1^2\right)(1-2r-s)\left(1-\sqrt{\frac{1}{\delta}}\sqrt{\frac{3\delta}{m}}\right) \\
&\geq \exp\left(\sum_{i=1}^{n-k_n} 2\eta_i\lambda_1 - 4\eta_i^2\lambda_1^2\right)\left(1-\frac{1}{25}\frac{\delta/m}{1+\delta/m} - \frac{1}{20}\left(1+\frac{\delta}{m}\right)^2 \ln\left(1+\frac{\delta}{m}\right)\right)\left(1-\sqrt{\frac{3}{m}}\right) \text{ using } 74, 75 \\
&\geq \frac{5}{6}\exp\left(\sum_{i=1}^{n-k_n} 2\eta_i\lambda_1 - 4\eta_i^2\lambda_1^2\right) \text{ since } \delta \leq 1 \text{ and } m = 200
\end{aligned}$$

S.5.1.3 Fraction

Now that we have established this result let's calculate the fraction. Let the step-sizes be set according to Lemma S.12. Define

$$\begin{aligned}
\mathcal{S} &:= \exp\left(\sum_{i=u+1}^{n} \overline{\mathcal{V}_{k,i}}\eta_{i-k_i}^2 + \sum_{i=1}^{n-k_n} 4\lambda_1^2\eta_i^2\right) \\
Q_u &:= \exp\left(2\lambda_1\left(\sum_{j=1}^{u}\eta_j - \sum_{j=n-k_n+1}^{n}\eta_j\right)\right) \\
\mathcal{R}_{k,t} &:= \frac{\exp\left(\sum_{j=1}^{t-u}\eta_j^2\left(\overline{\mathcal{V}_{k,j}} - \overline{\mathcal{V}_{k,j+u}}\right)\right)\exp\left(2\lambda_1\sum_{j=n-k_n+1}^{n}\eta_j\right)}{\exp\left(2\lambda_1\sum_{j=1}^{u}\eta_{t-u+j}\right)}
\end{aligned}$$

Then, recall that

$$\begin{aligned}
u &:= \min\{i : i \in [n], i - k_i \geq 0\} \\
\xi_{k,t} &:= 6\eta_{t-k_t}\mathcal{M}\left[1 + 3k_{t+1}^2\left(\mathcal{M}+\lambda_1\right)^2\right] \\
\mathcal{V}' &:= \left(\frac{1+(3+4\epsilon)\left|\lambda_2\left(P\right)\right|}{1-\left|\lambda_2\left(P\right)\right|}\right)\mathcal{V} \\
\overline{\mathcal{V}_{k,t}} &:= \mathcal{V}' + \lambda_1^2 + \xi_{k,t} \\
C'_{k,t} &:= \exp\left(2\lambda_1\sum_{j=1}^{u}\left(\eta_j - \eta_{t-u+j}\right) + \sum_{j=1}^{t-u}\eta_j^2\left(\overline{\mathcal{V}_{k,j}} - \overline{\mathcal{V}_{k,j+u}}\right)\right) = Q_u\mathcal{R}_{k,t}
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\frac{\mathrm{Tr}\left(V_\perp^T B_n B_n^T V_\perp\right)}{v_1^T B_n B_n^T v_1} \\
&\leq \frac{1.3}{\delta}\frac{\exp\left(\sum_{i=u+1}^{n} 2\eta_i\lambda_2 + \overline{\mathcal{V}_{k,i}}\eta_{i-k_i}^2\right)}{\exp\left(\sum_{i=1}^{n-k_n} 2\eta_i\lambda_1 - 4\eta_i^2\lambda_1^2\right)}\left(d + \sum_{i=u+1}^{n}\left(\mathcal{V}'+\xi_{k,i}\right)C'_{k,i}\eta_{i-k_i}^2\exp\left(\sum_{j=u+1}^{i} 2\eta_j\left(\lambda_1-\lambda_2\right)\right)\right) \\
&\leq \frac{1.3}{\delta}\frac{\mathcal{S}}{Q_u}\exp\left(\sum_{i=u+1}^{n} 2\eta_i\left(\lambda_2-\lambda_1\right)\right)\left(d + \sum_{i=u+1}^{n}\left(\mathcal{V}'+\xi_{k,i}\right)C'_{k,i}\eta_{i-k_i}^2\exp\left(\sum_{j=u+1}^{i} 2\eta_j\left(\lambda_1-\lambda_2\right)\right)\right) \\
&\leq \frac{1.3}{\delta}\mathcal{S}\left(\underbrace{\frac{d\exp\left(\sum_{i=u+1}^{n} 2\eta_i\left(\lambda_2-\lambda_1\right)\right)}{Q_u}}_{X_1} + \underbrace{\sum_{i=u+1}^{n}\left(\mathcal{V}'+\xi_{k,i}\right)\mathcal{R}_{k,i}\eta_{i-k_i}^2\exp\left(-\sum_{j=i+1}^{n} 2\eta_j\left(\lambda_1-\lambda_2\right)\right)}_{X_2}\right)
\end{aligned} \tag{76}$$

For X_1, we have

$$\begin{aligned}
X_1 &\leq \frac{d\exp\left(\sum_{i=u+1}^{n} 2\eta_i\left(\lambda_2-\lambda_1\right)\right)}{Q_u} \\
&= \frac{d\exp\left(\sum_{i=u+1}^{n} 2\eta_i\left(\lambda_2-\lambda_1\right)\right)}{\exp\left(2\lambda_1\left(\sum_{j=1}^{u}\eta_j-\sum_{j=n-k_n+1}^{n}\eta_j\right)\right)} \\
&\leq \frac{d\exp\left(\sum_{i=u+1}^{n} 2\eta_i\left(\lambda_2-\lambda_1\right)\right)}{\exp\left(-2\lambda_1\left(\sum_{j=n-k_n+1}^{n}\eta_j\right)\right)} \\
&\leq d\exp\left(\sum_{i=u+1}^{n} 2\eta_i\left(\lambda_2-\lambda_1\right)\right)\exp\left(2\lambda_1\left(\sum_{j=n-k_n+1}^{n}\eta_j\right)\right)
\end{aligned}$$

Note that

$$\begin{aligned}
\exp\left(2\lambda_1\sum_{j=n-k_n+1}^{n}\eta_j\right) &\leq \exp\left(2\left(1+2\epsilon\right)\lambda_1 k_n\eta_{n-k_n+1}\right) \text{ using monotonicity of } \eta_i \\
&\leq \exp\left(4\left(1+2\epsilon\right)\lambda_1 k_n\eta_n\right) \text{ using slow-decay of } \eta_i \\
&\leq 1+2\frac{\delta}{m} \text{ using Lemma } S.13 \text{ along with } e^x\leq 1+x+x^2 \text{ for } x\in(0,1)
\end{aligned}$$

Therefore, using 40

$$X_1 \leq d\left(1+\frac{2\delta}{m}\right)\left(\frac{\beta+u}{n}\right)^{2\alpha}$$

Next, for X_2, we first have

$$\begin{aligned}
\mathcal{R}_{k,t} &:= \frac{\exp\left(\sum_{j=1}^{t-u}\eta_j^2\left(\overline{\mathcal{V}_{k,j}}-\overline{\mathcal{V}_{k,j+u}}\right)\right)\exp\left(2\lambda_1\sum_{j=n-k_n+1}^{n}\eta_j\right)}{\exp\left(2\lambda_1\sum_{j=1}^{u}\eta_{t-u+j}\right)} \\
&\leq \exp\left(\sum_{j=1}^{t-u}\eta_j^2\overline{\mathcal{V}_{k,j}}\right)\exp\left(2\lambda_1\sum_{j=n-k_n+1}^{n}\eta_j\right) \\
&\leq \left(1+\frac{2\delta}{m}\right)^2 \text{ using Lemmas } S.12-(3), S.13 \text{ and } e^x\leq 1+x+x^2 \text{ for } x\in(0,1)
\end{aligned}$$

Now, using S.12-(4) we have,

$$\sum_{i=1}^{n}\overline{\mathcal{V}_{k,i}}\eta_{i-k_i}^2\exp\left(-\sum_{j=i+1}^{n}2\eta_j\left(\lambda_1-\lambda_2\right)\right) \leq$$

$$\left(\frac{2\left(1+10\epsilon\right)\alpha^2}{2\alpha-1}\right)\frac{\mathcal{V}'}{\left(\lambda_1-\lambda_2\right)^2}\frac{1}{n}+\left(\frac{800\left(1+10\epsilon\right)\alpha^3}{\left(\alpha-1\right)}\right)\frac{\mathcal{M}\left(\mathcal{M}+\lambda_1\right)^2\tau_{\text{mix}}^2}{\left(\lambda_1-\lambda_2\right)^3}\frac{\ln^2\left(\frac{(\beta+n)(\lambda_1-\lambda_2)}{\alpha}\right)}{n^2}$$

Then,

$$X_2 \leq \left(1+\frac{2\delta}{m}\right)^2\left[\underbrace{\left(\frac{2\left(1+10\epsilon\right)\alpha^2}{2\alpha-1}\right)}_{C_1}\frac{\mathcal{V}'}{\left(\lambda_1-\lambda_2\right)^2}\frac{1}{n}+\underbrace{\left(\frac{24\left(1+10\epsilon\right)\alpha^3}{\left(\alpha-1\right)}\right)}_{C_2}\frac{\mathcal{M}\left(\mathcal{M}+\lambda_1\right)^2}{\left(\lambda_1-\lambda_2\right)^3}\frac{k_n^2}{n^2}\right]$$

Therefore substituting in 76,

$$\frac{\mathrm{Tr}\left(V_{\perp}^{T}B_{n}B_{n}^{T}V_{\perp}\right)}{v_{1}^{T}B_{n}B_{n}^{T}v_{1}} \leq \frac{1.3\mathcal{S}}{\delta}\left(1+\frac{2\delta}{m}\right)^{2}\left[d\left(\frac{\beta+u}{n}\right)^{2\alpha}+\frac{C_{1}\mathcal{V}'}{\left(\lambda_{1}-\lambda_{2}\right)^{2}}\frac{1}{n}+\frac{C_{2}\mathcal{M}\left(\mathcal{M}+\lambda_{1}\right)^{2}}{\left(\lambda_{1}-\lambda_{2}\right)^{3}}\frac{k_{n}^{2}}{n^{2}}\right] \tag{77}$$

Proof of Theorem 1. To complete our proof, we bound $\mathcal{S}$ to simplify 77. We note that under the learning rate schedule presented in Lemma S.12-(3),

$$\mathcal{S} \leq \left(1+\frac{\delta}{m}\right)$$

Therefore,

$$\begin{aligned}\frac{\mathrm{Tr}\left(V_{\perp}^{T}B_{n}B_{n}^{T}V_{\perp}\right)}{v_{1}^{T}B_{n}B_{n}^{T}v_{1}} &\leq \frac{1.3}{\delta}\left(1+\frac{2\delta}{m}\right)^{3}\left[d\left(\frac{\beta+u}{n}\right)^{2\alpha}+\frac{C_{1}\mathcal{V}'}{\left(\lambda_{1}-\lambda_{2}\right)^{2}}\frac{1}{n}+\frac{C_{2}\mathcal{M}\left(\mathcal{M}+\lambda_{1}\right)^{2}}{\left(\lambda_{1}-\lambda_{2}\right)^{3}}\frac{k_{n}^{2}}{n^{2}}\right] \\ &\leq \frac{1.4}{\delta}\left[d\left(\frac{\beta+u}{n}\right)^{2\alpha}+\frac{C_{1}\mathcal{V}'}{\left(\lambda_{1}-\lambda_{2}\right)^{2}}\frac{1}{n}+\frac{C_{2}\mathcal{M}\left(\mathcal{M}+\lambda_{1}\right)^{2}}{\left(\lambda_{1}-\lambda_{2}\right)^{3}}\frac{k_{n}^{2}}{n^{2}}\right]\end{aligned}$$

Using lemma S.14, we have that $u \leq \beta$. Then, using Lemma 3.1 from [14] completes our proof. □

S.5.2 Proof of Corollary 1

Proof of Corollary 1. We note that the downsampled data stream can be considered to be drawn from a Markov chain with transition kernel $P^{k}\left(.,.\right)$ since each data-point is k steps away from the previous one. For sufficiently large k, this implies that the mixing time of this chain is $\Theta\left(1\right)$.

Therefore, as noted in the theorem statement, we plug in the modified parameters in the bound proved for Theorem 1.

Next, we note that for the transition kernel $P^{k}\left(.,.\right)$, the second-largest absolute eigenvalue is given as $\left|\lambda_{2}\left(P\right)\right|^{k}$. Therefore, for $k := \tau_{\text{mix}}\left(\gamma\eta_{n}^{2}\right) \leq \frac{2\tau_{\text{mix}}}{\ln(2)}\ln\left(\frac{1}{\gamma\eta_{n}^{2}}\right) \leq C\frac{1}{1-\left|\lambda_{2}(P)\right|}\ln\left(n\right), C>1$, using standard bounds on mixing times for reversible Markov chains (see for example, [19]). Consider the function $f\left(x\right) := x^{\frac{1}{1-x}}$ for $x \in (0,1)$. Then,

$$f'\left(x\right) = f\left(x\right)\left(\frac{1-x-x\ln\left(x\right)}{x\left(1-x\right)^{2}}\right) > 0$$

Therefore, $f\left(x\right) < \lim_{x\to 1} f\left(x\right) = \frac{1}{e} < 1$. which implies $\left|\lambda_{2}\left(P\right)\right|^{k} \overset{(i)}{\leq} \left(\frac{1}{e}\right)^{C\ln(n)} < \frac{1}{e}$. Here (i) follows if $C>1, n>3$, which is true. Therefore,

$$\mathcal{V}' := \left(\frac{1+\left(3+4\epsilon\right)\left|\lambda_{2}\left(P\right)\right|^{k}}{1-\left|\lambda_{2}\left(P\right)\right|^{k}}\right)\mathcal{V} \leq 5\mathcal{V}$$

This also implies that mixing time for the new Markov chain for sub-sampled data is $\Theta\left(1\right)$. The bound then follows by substituting n to be $\frac{n}{k} = n_{k} = \Theta\left(\frac{n}{C\tau_{\text{mix}}\ln(n)}\right)$ and setting the τ_{mix} in the original expression of Theorem 1 to a constant. □