# Concentration Inequalities for Sums of Markov Dependent Random Matrices

JOE NEEMAN, BOBBY SHI*, AND RACHEL WARD
*The University of Texas at Austin*
*Corresponding author: bhshi@utexas.edu

We give Hoeffding and Bernstein-type concentration inequalities for the largest eigenvalue of sums of random matrices arising from a Markov chain. We consider time-dependent matrix-valued functions on a general state space, generalizing previous results that had only considered Hoeffding-type inequalities, and only for time-independent functions on a finite state space. In particular, we study a kind of noncommutative moment generating function, provide tight bounds on this object, and use a method of Garg et al. to turn this into tail bounds. Our proof proceeds spectrally, bounding the norm of a certain perturbed operator. In the process we make an interesting connection to dynamical systems and Banach space theory to prove a crucial result on the limiting behavior of our moment generating function that may be of independent interest.

## 1. Introduction

The study of concentration inequalities has now become textbook material, with a variety of applications [6]. Two of the most widely used and studied inequalities for scalar-valued random variables in the independent setting are Hoeffding's inequality [28], which operates under a boundedness assumption, and Bernstein's inequality [5], which operates under an additional bounded variance assumption.

With the wide usage of these inequalities, various generalizations have been made. One line of work has sought to relax the independence assumption, deriving concentration inequalities for sums of Markov-dependent random variables. Starting with the result of Gillman [19], improvements and refinements have been made to address the general setting of time-independent functions of a nonreversible Markov chain on a general state space [15, 26, 33, 41, 42, 43, 54], resulting in Hoeffding and Bernstein-type concentration inequalities that generalize nicely the independent setting; these results are largely spectral in nature.

Another line of work has sought to generalize the scalar setting to concentration inequalities of sums of independent random matrices, beginning with the work of Ahlswede and Winter [2]. Several authors have followed the Ahlswede-Winter approach to develop analogs of Hoeffding's and Bernstein's inequalities for sums of random matrices; some works include [8, 24, 52, 53, 57]. However, results using this framework often have variance proxies that are suboptimal. To this end, [63, 64] developed techniques to circumvent this barrier, allowing for concentration inequalities that are much tighter in many cases. These results are powerful and easy to use [65, 66], with various applications [13, 45, 47].

We combine the above two lines of study in our work to develop concentration inequalities – Hoeffding and Bernstein-type – for the largest eigenvalue of sums of Markov dependent Hermitian random matrices, by combining the Ahlswede-Winter style argument for random matrices with the spectral techniques used to study Markov dependent scalar-valued random variables. The starting point is a deep result of Garg et al. [16], which provides a multi-matrix Golden-Thompson inequality that supports a spectral approach. Our results are broad and general: we provide inequalities for *time-dependent* matrix-valued functions of a *nonreversible* Markov chain on *general (continuous) state spaces*. In this way, our Hoeffding-type bounds vastly generalize previous work and provide sharper constants, and our Bernstein-type bounds are new in the literature (Section 6). Along the way we prove a novel Perron-Frobenius type limit lemma for the matrix setting (Lemmas 5.4, 5.5), necessitated by

our assumptions of nonversibility of the chain on continuous state spaces, that may be of independent interest.

Thus, this work can be seen as a first systematic study of concentration inequalities in the Markov-dependent matrix setting. Each of our results involve two bounds: one bounding a noncommutative moment generating function of the form

$$\mathbb{E}\left[\left\|\prod_{j=1}^{n}\exp\left(\frac{\theta}{2}X_j\right)\right\|_F^2\right]$$

and one bounding the tail probabilities

$$\Pr\left(\lambda_{\max}\left(\sum_{j=1}^{n}X_j\right)\geq t\right)$$

for Hermitian random matrices $X_1,\ldots,X_n$ arising as functions of a Markov chain. We expect our results on the former to be optimal, directly generalizing the scalar setting.

These results are readily applicable to a wide variety of settings. For example, the best known guarantees for offline principal component analysis (PCA) are derived from a combination of the standard matrix Bernstein's inequality along with Wedin's perturbation theorem [32, 38, 69]. Now, if the sample matrices are not independent, but instead arise from a Markov chain, we again have a bound for leading eigenvector estimation using our bounds. Our results are also directly applicable to sums of matrix-valued random variables sampled from a random walk on an expander; Theorem 2.5 improves on the best known results in this setting [16, 70]. There is also a line of work in machine learning that realizes stochastic gradient descent (SGD) with constant step size as a Markov chain with a stationary distribution [4, 11, 48]. Additionally, Hessian information has been used widely to give improved algorithms and better understand model performance [18, 44]. Our results are able to give precise bounds on the convergence of the largest eigenvalue of the empirical mean of the Hessian matrices along the path of SGD to the largest eigenvalue of the expected Hessian with respect to the stationary measure. This has many interesting and relevant applications in statistical learning and generalization. More broadly, our bounds are useful in the study of online algorithms, where matrix-valued data is received and processed in a streaming fashion from an underlying Markov chain. In Section 7 we more closely focus on an application of our main results to offline PCA for samples arising as a function of an underlying Markov chain; we present a bound that is a direct generalization of the best known bound for offline PCA in the i.i.d. setting. This result was first stated in [38], where to our knowledge it was the first in the literature to give this sort of bound for offline Markov PCA ; it directly uses our Theorem 2.6.

### 1.1. *Related work*

In terms of results, the literature on concentration inequalities for sums of Markov-dependent random matrices is fairly small, in contrast to the scalar setting. Recent progress in this area builds on the work of [16, 60], which develop a powerful Golden-Thompson inequality [21, 62] that allows for a spectral analysis of a useful moment generating function; the latter provides a corresponding Hoeffding's inequality for when the stationary distribution is uniform and the chain is stationary. The work of [56] extends this to arbitrary stationary distributions and initial distributions; these works are valid only for finite state spaces and reversible chains. Our work generalizes these results to general state spaces

and nonreversible chains using significantly different techniques, which allows us to also improve the constants in the bounds. Furthermore, our Hoeffding-type result is in terms of both the largest and smallest eigenvalues, matching the type given in the scalar setting; and we give a Bernstein's inequality, which is new in the literature.

In terms of techniques, most related to ours are the works of [15, 26, 33, 41, 42, 49, 54], which give Hoeffding's and Bernstein's inequalities for sums of Markov-dependent *scalar* random variables. Broadly speaking, the idea is to bound the largest norm of a perturbed operator. We prove the corresponding spectral properties of a Markov operator that is the tensored form of the operator appearing in these works in Section 5; in particular, we generalize to continuous state spaces using related functional analytic ideas. However, previous techniques are not sufficient for us to prove our main results; our proof of the result on the limit of the moment generating function (Lemma 5.4) takes a detour through dynamical systems theory and Banach space theory.

In contrast to the Markov-dependent setting, the sums of independent random matrices have been studied in depth, and is the model for which we give our bounds. In particular, our results mirror the type given in [2, 8, 24] in terms of variance proxy, and the analysis proceeds from the same starting point: bounding a matrix moment generating function; in contrast, later works of [52, 63] sharpen the variance proxy. Our analysis of the moment generating function arising from the matrix-valued Golden-Thompson inequality result in bounds resembling more of the former; we believe our analysis is likely sharp given the form of the moment generating function, so an improvement in terms of the variance proxy would likely require a different approach and is an interesting open problem.

Besides the spectral approach, there have been a myriad of ways to derive concentration inequalities relaxing the independence assumption, both in the scalar and matrix settings. The papers of [1, 20] exploit regeneration-type minorization conditions to derive exponential concentration for ergodic sums; though these works do not assume a nonzero spectral gap they often have less explicit constants. The work of [54] uses Marton couplings to obtain concentration inequalities for a scalar-valued function of a single random sample. The papers of [46, 55] leverage Efron-Stein inequalities and exchangeability to develop concentration inequalities for random matrices, also relaxing the independence assumption. Similarly, recent works of [3, 30, 36] demonstrate that matrix functional inequalities, i.e., Poincaré, directly translate to matrix concentration inequalities; [36] develops a Bernstein's inequality for strongly Rayleigh distributions; [39] gives a Chernoff bound for a similar setting. These largely address single sample concentration of (Lipschitz) functionals; however, some methods can be generalized to product measures [31], recovering some of the Efron-Stein results. And the work of [16] sketches a method to reduce studying concentration of random variables sampled from a Markov chain to concentration of sums of martingale random variables [9, 63]; though the resulting bounds are suboptimal, they show that qualitatively concentration for Markov chains is more generic than just the scalar or matrix settings (see [40]). These techniques may be very useful in improving the variance proxy, better matching the independent setting, where a direct spectral analysis may prove insufficient; in particular, functional inequalities have emerged as a powerful general tool to derive concentration. It may be interesting to see if this can be demonstrated in the scalar setting as well.

## 2. Main Results

In this section we give our main results. Each will hold under a combination of the following assumptions.

**Assumption 2.1** *$P$ is a discrete-time Markov chain on a continuous state space $\mathscr{X}$ with stationary distribution $\mu$ and absolute spectral gap $\lambda$ (Definition 3.1). $s_1,\dots,s_n$ is a sequence of states driven by $P$ with initial distribution $\mu$.*

**Assumption 2.2** *$F_j : \mathscr{X} \to \mathbb{R}^{d\times d}$, $j = 1,\dots,n$, is a sequence of functions each mapping to real symmetric $d \times d$ matrices. Each $F_j$ is $\ell_2(\mu \otimes \mathbf{1})$ measurable (see Section 3).*

**Assumption 2.3** *For all $j$, $\mathbb{E}_\mu[F_j(x)] = 0$ and $a_j I \preceq F_j(x) \preceq b_j I$ for all $x \in \mathscr{X}$.*

**Assumption 2.4** *For all $j$, $\mathbb{E}_\mu[F_j(x)] = 0$ and there exists a constant $\mathscr{V}_j$ such that $\left\|\mathbb{E}_\mu\left[F_j(x)^2\right]\right\| \le \mathscr{V}_j$ for all $x \in \mathscr{X}$. Moreover, there exists an absolute constant $\mathscr{M}$ such that $\left\|F_j(x)\right\| \le \mathscr{M}$ for all $j, x \in \mathscr{X}$.*

The following two theorems each consist of two statements. The first bounds a certain type of moment generating function arising from the expected Frobenius norm of a product of matrix exponentials. The second statement turns this into a tail bound. We only give upper tail bounds; the corresponding statement for lower tail bounds are clear and the proof proceeds in almost the exact same way.

**Theorem 2.5** (Markov Matrix Hoeffding, Real) *Instate Assumptions 2.1, 2.2, 2.3. Then for any $\theta > 0$,*

$$\mathbb{E}_\mu\left[\left\|\prod_{j=1}^{n}\exp\left(\frac{\theta}{2}F_j(s_j)\right)\right\|_F^2\right] \le d\exp\left(\frac{\theta^2}{2}\cdot\alpha(\lambda)\cdot\frac{\sum_{j=1}^n(b_j-a_j)^2}{4}\right)$$

*and for any $t > 0$,*

$$\Pr_\mu\left(\lambda_{max}\left(\sum_{j=1}^{n}F_j(s_j)\right)\ge t\right) \le d^{2-\pi/4}\exp\left(\frac{-t^2/(8/\pi^2)}{\alpha(\lambda)\cdot\sum_{j=1}^n(b_j-a_j)^2}\right).$$

*Here $\alpha(\lambda) = (1+\lambda)/(1-\lambda)$, where $\lambda$ is the absolute spectral gap (Definition 3.1).*

The first statement in the above result reveals the sub-Gaussian nature of the moment generating function; in the scalar case, it exactly reduces to the standard moment generating function. Using our methods, this bound matches the scalar case given in [15]. The second statement provides a large deviation tail bound; the bound on the lower eigenvalue follows analogously. Our results offer four main improvements to that of [56]: first, we give strictly better constants, both in terms of $d$ and the absolute constants in the exponent (the $\alpha(\lambda)$ differs slightly from their equivalent term for the mixing of the Markov chain, but is strictly better and is more classical); second, we allow for time-dependent functions $F_j$, which is a strict generalization; third, our inequality is in terms of both a lower and upper eigenvalue bound (the $a_j, b_j$ parameters), giving a result that is directly comparable to the classic Hoeffding's lemma; and fourth, our bounds apply to general state spaces.

**Theorem 2.6** (Markov Matrix Bernstein, Real) *Instate Assumptions 2.1, 2.2, 2.4. Let* $\sigma^2 = \sum_{j=1}^n \mathscr{V}_j$. *Then for any* $0 < \theta < \log(1-\lambda)/\mathscr{M}$,

$$\mathbb{E}_\mu \left[ \left\| \prod_{j=1}^n \exp\left(\frac{\theta}{2} F_j(s_j)\right) \right\|_F^2 \right] \leq d \exp\left( \frac{\sigma^2}{\mathscr{M}^2} \left( e^{\mathscr{M}\theta} - \mathscr{M}\theta - 1 \right) + \frac{\sigma^2}{\mathscr{M}^2} \cdot \frac{\lambda (e^{\mathscr{M}\theta} - 1)^2}{1 - \lambda e^{\mathscr{M}\theta}} \right).$$

*Furthermore, if* $\theta < (1-\lambda)/(8\mathscr{M}/\pi)$, *then for any* $t > 0$,

$$\Pr_\mu \left( \lambda_{max} \left( \sum_{j=1}^n F_j(s_j) \right) \geq t \right) \leq d^{2-\pi/4} \exp\left( \frac{-t^2/(32/\pi^2)}{\alpha(\lambda) \cdot \sigma^2 + \beta(\lambda) \cdot \mathscr{M} t} \right).$$

*Here*

$$\alpha(\lambda) = \frac{1+\lambda}{1-\lambda}, \quad \beta(\lambda) = \begin{cases} \frac{4}{3\pi}, & \lambda = 0, \\ \frac{8/\pi}{1-\lambda}, & 0 < \lambda < 1 \end{cases},$$

*where* $\lambda$ *is the absolute spectral gap (Definition 3.1).*

This is to our knowledge the first result giving a Bernstein-type inequality for sums of random matrices arising from markov chains. The first statement in the above result is made up of two terms; the $(e^{\mathscr{M}\theta} - \mathscr{M}\theta - 1)\sigma^2/\mathscr{M}^2$ term coincides with the convex function that makes up the classic Bernstein's and Bennett's inequality. The second term in the exponent reflects the influence of the Markov chain; compared to previous scalar results [33, 54], our techniques are able to recover a slightly improved version of this second term in terms of absolute constants. A linear algebraic perspective of the first statement above is that the first term bounds how much the product of matrix exponentials increases the magnitude of vectors in the direction of $\mathbf{1}$, which is the leading eigenvector of the operator $P$. The second term is a bound on how much the product of matrix exponentials increases the magnitude of orthogonal directions, hence involving $\lambda$.

We now give two corollaries of the above results, generalizing the assumptions above; the proofs are in the appendix and are classical. Our first corollary generalizes to complex Hermitian matrices via a complexification technique [12]:

**Corollary 2.7** (Extension to Complex Matrices) *Instate Assumption 2.1. Now assume that* $F_j : \mathscr{X} \to \mathbb{C}^{d\times d}$, $j = 1, \ldots, n$, *is a sequence of functions each mapping to complex Hermitian* $d \times d$ *matrices.*

*Under Assumption 2.3 (resp. Assumption 2.4), the conclusion of Theorem 2.5 (resp. Theorem 2.6) holds with an extra multiplicative factor of* 2 *on the right-hand side.*

Our second corollary generalizes to the case that the Markov chain starts at a distribution $\nu$:

**Corollary 2.8** (Extension to Nonstationary Chains) *Let* $P$ *be a Markov chain on general state space* $\mathscr{X}$ *with stationary distribution* $\mu$ *and absolute spectral gap* $\lambda$. *Let* $s_1, \ldots, s_n$ *be a sequence of states driven by* $P$ *with initial distribution* $\nu$, *where* $\nu \ll \mu$. *Instate Assumption 2.2. Let* $\operatorname{ess\,sup} \frac{d\nu}{d\mu}$ *be the essential supremum of the Radon-Nikodym derivative* $\frac{d\nu}{d\mu}$.

*Under Assumption 2.3 (resp. Assumption 2.4), the conclusion of Theorem 2.5 (resp. Theorem 2.6) holds with an extra multiplicative factor of* $\operatorname{ess\,sup} \frac{d\nu}{d\mu}$ *on the right-hand side.*

**Remark 2.9** *In many applications these bounds are used to determine how many samples are needed (i.e., how long the chain has to run) in order for the tail probability to be less than some fixed value; in these cases the number of samples will depend only logarithmically on* $\operatorname{ess\,sup} \frac{d\nu}{d\mu}$. *Alternatively, the probability of an event under the Markov chain driven by P initialized at* $\nu$ *can be bounded by the probability of an event under the Markov chain driven by P initialized at* $\mu$ *up to an extra multiplicative factor of* $1+\chi^2(\nu \,\|\, \mu)$*; see Lemma 29 of [37].*

We now give some remarks on the above theorems. First, for the first statements regarding the moment generating function in both Theorems 2.5 and 2.6, the variance proxies are asymptotically optimal for a class of Markov chains. Let $\mu$ be a distribution on $\mathscr{X}$ and let $P$ be a transition kernel such that $P(x,y) = \lambda \mathbb{1}_{x=y} + (1-\lambda)\mu(y)$, so that $\mu$ is the stationary distribution. Let $f : \mathscr{X} \to \mathbb{R}$ be any scalar-valued function, and let $F : \mathscr{X} \to \mathbb{R}^{d\times d}$ be defined as $F(x) = f(x)I$, so that these are all real symmetric. It is straightforward that

$$\mathbb{E}_\mu \left[ \left\| \prod_{j=1}^{n} \exp\left(\frac{\theta}{2} F(s_j)\right) \right\|_F^2 \right] = d\mathbb{E}_\mu \left[ \exp\left( \theta \sum_{j=1}^{n} f(s_j) \right) \right] \tag{2.1}$$

since all $F(x)$ commute.

Under this setup, start with our Bernstein assumptions. Let $f$ such that $\mathbb{E}_\mu[f] = 0$, $\mathbb{E}_\mu[f^2] = \sigma^2$, and $|f| \le \mathscr{M}$. A central limit theorem of [17] states that for a two state Markov chain driven by a $P$ of our above form that

$$\mathscr{V}_{\text{asy}} := \lim_{n\to\infty} \operatorname{Var}\left( \frac{1}{\sqrt{n}} \sum_{j=1}^{n} f(s_j) \right) = \frac{1+\lambda}{1-\lambda} \cdot \sigma^2, \tag{2.2}$$

which is called the asymptotic variance. Now suppose that $\mathscr{V}$ is such that $\mathbb{E}_\mu[f^2] \le \mathscr{V}$; it is classical that any such variance proxy $\mathscr{V}$ that satisfies $g(\theta) = n\mathscr{V}\theta^2/2 + o(\theta^2)$ for a function $g$ that upper bounds the scalar moment generating function via

$$\mathbb{E}_\mu \left[ \exp\left( \theta \sum_{j=1}^{n} f(s_j) \right) \right] \le e^{g(\theta)}$$

upper bounds $\operatorname{Var}\left( \frac{1}{\sqrt{n}} \sum_{j=1}^{n} f(s_j) \right)$ [33]. This lower bound is indeed obtained by our Markov kernel $P$ and choice of function $f$, and therefore the same holds for our matrix-valued function $F$.

Now assume $f$ are Rademacher, i.e., the probability under $\mu$ that $f(x) = 1$ is $1/2$ and $f(x) = -1$ is $1/2$. Define $F$ the same way as above, and so we see that $\mathbb{E}_\mu[F] = 0$ and $-I \preceq F(x) \preceq I$ for all $x \in \mathscr{X}$. Via Equation 2.1 we again reduce to the scalar case since $F$ commute, and through 2.2 we have that $\mathscr{V}_{\text{asy}} = (1+\lambda)/(1-\lambda)$, since the variance of $f$ is 1. Now in this case we have

$$\mathbb{E}_\mu \left[ \exp\left( \theta \sum_{j=1}^{n} f(s_j) \right) \right] \le \exp\left( \frac{\theta^2}{2} \cdot \alpha(\lambda) \cdot n \right),$$

as can be seen from our results with $d = 1$. We see that $n\alpha(\lambda)$ is the variance proxy for a sub-Gaussian random variable and so naturally upper bounds the asymptotic variance. This lower bound is again attained for $P$ and $f$, and therefore the same holds for the matrix-valued function $F$.

2.1. *Roadmap*

We begin with the quantity

$$\mathbb{E}_\mu \left[ \left\| \prod_{j=1}^{n} \exp\left( \frac{\theta e^{i\phi}}{2} F_j(s_j) \right) \right\|_F^2 \right]$$

for $\phi \in [-\pi/2, \pi/2]$ and $\theta > 0$, developed first in [16, 60]. This will play the role of our moment generating function. Using properties of the Kronecker product, in Section 4 we show that this moment generating function is equal to

$$\left\langle \mathbf{1} \otimes \mathrm{vec}(I_d), E_1^{\theta/2} \left( \prod_{j=1}^{n-1} E_j^{\theta/2} \tilde{P} E_{j+1}^{\theta/2} \right) E_n^{\theta/2} (\mathbf{1} \otimes \mathrm{vec}(I_d)) \right\rangle_\mu ,$$

where $\tilde{P}$ is the lifted Markov operator and $E_j^{\theta/2}$ is a certain multiplication operator (Definition 3.4). This quantity is a matrix version of a classical quantity that appears in the analysis of sums of Markov-dependent scalar random variables – indeed, when $d = 1$ these are equivalent. Thus, we can apply a spectral analysis to bound this quantity.

In Section 5 we proceed with this study; this section addresses the main challenges of extending to continuous state spaces. Using Cauchy-Schwartz we bound the above by

$$\left\| E_1^{\theta/2^*} (\mathbf{1} \otimes \mathrm{vec}(I_d)) \right\|_\mu \left\| E_n^{\theta/2} (\mathbf{1} \otimes \mathrm{vec}(I_d)) \right\|_\mu \prod_{j=1}^{n-1} \left\| E_j^{\theta/2} \tilde{P} E_{j+1}^{\theta/2} \right\|_\mu .$$

Lemma 5.1 is a symmetrizing result, allowing us to shift our attention to the lifted Leon-Perron operator $\hat{P}$ by giving the bound

$$\left\| E_1^{\theta/2^*} (\mathbf{1} \otimes \mathrm{vec}(I_d)) \right\|_\mu \left\| E_n^{\theta/2} (\mathbf{1} \otimes \mathrm{vec}(I_d)) \right\|_\mu \prod_{j=1}^{n-1} \left\| E_j^{\theta/2} \tilde{P} E_{j+1}^{\theta/2} \right\|_\mu \leq d \prod_{j=1}^{n} \left\| E_j^{\theta/2} \hat{P} E_j^{\theta/2^*} \right\|_\mu .$$

This lets us replace the operator $\tilde{P}$, which represents a chain that is not necessarily reversible, by $\hat{P}$, which represents a very simple reversible chain with many of the same important spectral properties as $\tilde{P}$. In doing so, we obtain operators $E_j^{\theta/2} \hat{P} E_j^{\theta/2^*}$ that are now self-adjoint (and even positive semidefinite) on $\ell_2(\mu \otimes \mathbf{1})$. We focus on bounding the leading eigenvalue of each of these self-adjoint matrices.

To simplify the results, we then notice that the eigenvalues of $E_j^{\theta/2} \hat{P} E_j^{\theta/2^*}$ are equal to the eigenvalues of $E_{T_j}^{\theta/2} \hat{P} E_{T_j}^{\theta/2}$, for some other mulitiplication operator $E_{T_j}^{\theta/2}$ that is in fact *real*, since the underlying matrix-valued functions $F_j$ are real-valued. Our next main result states that the leading eigenvalue of $E_{T_j}^{\theta/2} \hat{P} E_{T_j}^{\theta/2}$ is a sort of limit for a corresponding moment generating function. More precisely, we show that

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}_\mu \left[ \left\| \prod_{k=1}^{n} \exp\left( \frac{\theta \cos(\phi)}{2} F_j(s_k) \right) \right\|_F^2 \right] = \log \left\| E_{T_j}^{\theta/2} \hat{P} E_{T_j}^{\theta/2} \right\|_\mu ,$$

where $s_1, \ldots, s_n$ is driven by the Markov chain represented by the Leon-Perron operator $\hat{P}$; this result appears in Lemma 5.4 and 5.5. The realness assumption on the $F_j$ is important for the proof of this

result – we invoke an interesting generalization of the Perron-Frobenius theorem that applies to linear transformations leaving invariant a cone. This result is our main technical innovation and is crucial in extending existing results to the continuous state space setting; conveniently it also allows us to sharpen known bounds in discrete state spaces.

Section 6 gives our final bounds. We first show how to transfer bounds on our moment generating function to tail bounds; the technique is straightforward and uses the multi-matrix Golden-Thompson inequality from [16]. We then give these tight bounds on the moment generating function under both Hoeffding and Bernstein-type assumptions. For Hoeffding-type assumptions we use a coupling technique to exhibit a two-state chain that acts as the "limit" of our chain. The leading eigenvalue of the corresponding operator can be solved exactly, giving optimal bounds. For Bernstein-type assumptions, we use a robust linear algebraic approach to bound the operator norm of a related matrix directly; this approach will also give optimal results.

## 3. Preliminaries

### 3.1. *Notation*

Lower case, unbolded $a$ denotes scalars or scalar-valued functions, bold $\mathbf{a}$ denotes vectors or vector-valued functions, and upper case $A$ denotes matrices or matrix-valued functions.

The operator $\otimes$ will denote the Kronecker product, where for matrices $A, B$ of size $a \times b$, $c \times d$, respectively,

$$A \otimes B = \begin{bmatrix} A_{1,1}B & \dots & A_{1,b}B \\ \vdots & & \vdots \\ A_{a,1}B & \dots & A_{a,b}B \end{bmatrix}$$

of size $ac \times bd$. This operation has an identification with the tensor product. An important fact we use is $(A \otimes C)(B \otimes D) = AB \otimes CD$ for matrices of the appropriate sizes. The Kronecker product has an interesting relationship with the vectorization operator vec, where for a matrix $X$ of size $a \times b$, $\mathrm{vec}(X)$ is the flattened vector of size $ab$. Importantly, $(B^\top \otimes A)\,\mathrm{vec}(X) = \mathrm{vec}(AXB)$, and so

$$\mathrm{tr}[AB] = \langle \mathrm{vec}(I_d), \mathrm{vec}(AB) \rangle = \left\langle \mathrm{vec}(I_d), (B^\top \otimes A)\,\mathrm{vec}(I_d) \right\rangle.$$

### 3.2. *Markov chains*

Let $\mathscr{X}$ be a general state space with $\sigma$-algebra $\mathscr{B} := \mathscr{B}(\mathscr{X})$. Let $P$ be a Markov kernel on $\mathscr{X}$ defined, for a sequence of random variables $X_1, \dots, X_n$, in the standard way as

$$P(x, B) = \Pr\left(X_k \in B \mid X_{k-1} = x\right), \quad \forall B \in \mathscr{B}$$

with stationary measure $\mu$ so that

$$\mu(B) = \int P(x, B)\mu(dx), \quad \forall B \in \mathscr{B}.$$

Define $\ell_2(\mu)$ as

$$\ell_2(\mu) = \{h : \mathscr{X} \to \mathbb{C} \mid \mathbb{E}_\mu[h(x)^2] < \infty\}$$

for $h : \mathscr{X} \to \mathbb{C}$ measurable. This is a a Hilbert space equipped with the following inner product:

$$\langle f, g \rangle_\mu = \int f(x)g(x)\,d\mu(x), \quad \forall f, g \in \ell_2(\mu)$$

and corresponding norm $\|f\|_\mu = \sqrt{\langle f, f\rangle_\mu}$. The norm of a linear operator $T$ is defined in the usual way:

$$\|T\|_\mu = \sup_{\|h\|_\mu = 1} \|Th\|_\mu .$$

A transition kernel $P$ acts as an operator on $\ell_2(\mu)$:

$$(Ph)(x) = \int h(y)P(x,dy), \quad \forall x \in \mathscr{X}, h \in \ell_2(\mu).$$

The projection operator $\Pi$ corresponding to the distribution $\mu$ is defined as $(\Pi h)(x) = \mathbb{E}_\mu[h]\mathbf{1}$ ($\mathbf{1}$ is the function such that $\mathbf{1}(x) = 1$ for all $x$); this is a rank-1 operator by definition, and is indeed a projection onto $\mathbf{1}$ as $\mathbb{E}_\mu[h] = \langle \mathbf{1}, h\rangle_\mu$. If $\mu$ is stationary for the transition kernel $P$ then $P\Pi = \Pi P = \Pi$.

An important subset of elements of $\ell_2(\mu)$ will be the class of "mean-zero functions," denoted

$$\ell_2^0(\mu) = \{h \in \ell_2(\mu) \mid \Pi h = 0\}.$$

Then we have the following definition:

**Definition 3.1** (Absolute spectral gap) *A Markov kernel $P$ with stationary measure $\mu$ admits an absolute spectral gap $1 - \lambda(P)$ if*

$$\lambda(P) := \sup_{h \in \ell_2^0(\mu), h \neq 0} \frac{\|Ph\|_\mu}{\|h\|_\mu} = \|P - \Pi\|_\mu < 1.$$

Note that $\lambda(P) \leq 1$ always, as $\|Ph\|_\mu \leq \|h\|_\mu$ by Jensen's inequality, with equality for $h = \mathbf{1}$. When it is clear from context we will just use $\lambda = \lambda(P)$. The absolute spectral gap characterizes the convergence of a Markov chain to its invariant measure [59]. For reversible finite-state chains, the value $\lambda(P)$ corresponds to the second largest eigenvalue and the existence of the gap corresponds to ergodicity. A Markov chain driven by $P$ is reversible if and only if $P$ is a self-adjoint operator on $\ell_2(\mu)$.

We define what is known as a Leon-Perron operator [15, 33, 41] that will be important in the sequel:

**Definition 3.2** (Leon-Perron operator) *Let $P$ be a Markov kernel with stationary measure $\mu$. For a constant $c \in [0,1]$, define $\hat{P}_c$ as*

$$\hat{P}_c := cI + (1-c)\Pi.$$

*If $c = \lambda(P)$ we drop the subscript and call $\hat{P}$ the Leon-Perron version of $P$.*

We can interpret $\hat{P}_c$ as a transition kernel such that at a state, it stays at that state with probability $c$ or samples a new state independently from $\mu$ with probability $1 - c$. Note that if $P$ admits an absolute spectral gap then $\hat{P}$ does so as well with the same absolute spectral gap.

### 3.3. *Behavior in tensored space*

As we work with matrices, the above will have to be lifted to a tensored space as necessary. We first consider the product space $\mathbb{C}^d \otimes \mathbb{C}^d \simeq \mathbb{C}^{d^2}$ with inner product induced from the standard Euclidean product, i.e., $\langle \mathbf{a} \otimes \mathbf{c}, \mathbf{b} \otimes \mathbf{d}\rangle = \langle \mathbf{a}, \mathbf{b}\rangle \langle \mathbf{c}, \mathbf{d}\rangle$, and extend by linearity. Oftentimes we will simply identify

an element as $\mathbf{y} \in \mathbb{C}^{d^2}$; it is straightforward that the standard Euclidean inner product on this larger space is equivalent to the tensored inner product. So it is no loss to move between one representation to the other.

Now define $\ell_2(\mu \otimes \mathbf{1})$ as the following "lift" of $\ell_2(\mu)$: formally, define

$$\ell_2(\mu \otimes \mathbf{1}) = \{\mathbf{h} : \mathscr{X} \to \mathbb{C}^{d^2} \mid \mathbb{E}_\mu[\|\mathbf{h}(x)\|^2] < \infty\}$$

for $\mathbf{h} : \mathscr{X} \to \mathbb{C}^{d^2}$ measurable as a vector-valued function. This is equipped with the inner product

$$\langle \mathbf{f}, \mathbf{g} \rangle_\mu = \int \langle \mathbf{f}(x), \mathbf{g}(x) \rangle \, d\mu(x)$$

and corresponding norm $\|\mathbf{f}\|_\mu = \sqrt{\langle \mathbf{f}, \mathbf{f} \rangle_\mu}$. This gives the understanding of $\ell_2(\mu \otimes \mathbf{1})$ as a direct integral of Hilbert spaces indexed by the points of $\mathscr{X}$; decomposable elements in this space are understood to be of the form $f \otimes \mathbf{v}$ for $f \in \ell_2(\mu), \mathbf{v} \in \mathbb{C}^{d^2}$ such that $(f \otimes \mathbf{v})(x) = f(x)\mathbf{v}$. Norms of linear operators are thus defined with respect to this norm in the usual way.

For a Markov operator $P$ on $\ell_2(\mu)$ we can define the operator $\tilde{P} := P \otimes I_{d^2}$ as an operator on $\ell_2(\mu \otimes \mathbf{1})$. This operator acts as

$$(\tilde{P}\mathbf{h})(x) = \int \mathbf{h}(x) P(x, dy), \quad \forall x \in \mathscr{X}, \mathbf{h} \in \ell_2(\mu \otimes \mathbf{1});$$

note that $(\tilde{P}\mathbf{h})(x)$ is a vector. Whereas for the operator $P$ the leading eigenfunction was $\mathbf{1}$ and spans the entire eigenspace for the leading eigenvalue of 1 (assuming nonzero spectral gap), for $\tilde{P}$ the leading eigenspace is $d^2$-dimensional, spanned by $\{\mathbf{1} \otimes \mathbf{e}_1, \ldots, \mathbf{1} \otimes \mathbf{e}_{d^2}\}$. This eigenspace is $\mathbf{1} \otimes \mathbb{C}^{d^2}$.

For a decomposable element $f \otimes \mathbf{v}$, where $f \in \ell_2(\mu), \mathbf{v} \in \mathbb{C}^{d^2}$, projection onto $\mathbf{1} \otimes \mathbb{C}^{d^2}$ is $\langle \mathbf{1}, f \rangle_\mu (\mathbf{1} \otimes \mathbf{v}) = \mathbb{E}_\mu[f](\mathbf{1} \otimes \mathbf{v})$. For the projection operator $\Pi$ we can define the lifted version as $\tilde{\Pi} := \Pi \otimes I_{d^2}$, again satisfying $\tilde{\Pi}\tilde{P} = \tilde{P}\tilde{\Pi} = \tilde{\Pi}$, that performs exactly this operation, i.e., $\tilde{\Pi}(f \otimes \mathbf{v}) = \langle \mathbf{1}, f \rangle_\mu (\mathbf{1} \otimes \mathbf{v})$. This can all be extended by linearity and convergence in $\ell_2$. Then there is a corresponding lift of "mean-zero functions" as

$$\ell_2^0(\mu \otimes \mathbf{1}) = \{\mathbf{h} \in \ell_2(\mu \otimes \mathbf{1}) \mid \tilde{\Pi}\mathbf{h} = 0\},$$

and can also be identified as $\ell_2^0(\mu \otimes \mathbf{1}) = \ell_2^0(\mu) \otimes \mathbb{C}^{d^2}$. Similarly, define

$$\lambda(\tilde{P}) = \sup_{\mathbf{h} \in \ell_2^0(\mu \otimes \mathbf{1}), \mathbf{h} \neq 0} \frac{\left\|\tilde{P}\mathbf{h}\right\|_\mu}{\|\mathbf{h}\|_\mu} = \left\|\tilde{P} - \tilde{\Pi}\right\|_\mu.$$

The following lemma, Lemma 3 from [56], will imply that if $P$ admits an absolute spectral gap, then so does $\tilde{P}$ with the same absolute spectral gap. Thus, many of the essential spectral properties of $P$ and $\tilde{P}$ are the same. The proof is deferred to the appendix.

**Lemma 3.3** $\lambda(P) = \lambda(\tilde{P})$.

Lastly, we can define Leon-Perron operators of $\tilde{P}$ as the corresponding lifts of Leon-Perron operators of $P$, which is tensorization by $I_{d^2}$. With some abuse of notation, we will often interchange the notation

$\hat{P}$ to refer to both Leon-Perron operators of $P$ and Leon-Perron operators of $\tilde{P}$, but the usage will be clear from context.

Recall for a scalar-valued function $g$, the multiplication operator $M_g$ is defined as $(M_g h)(x) = g(x)h(x)$. For a matrix-valued function $G$ we have the following definition:

**Definition 3.4** (Multiplication operator) *Let $G$ be a matrix-valued function mapping $\mathscr{X}$ to $\mathbb{C}^{d^2\times d^2}$ matrices. Define $M_G$ be the operator on $\ell_2(\mu\otimes\mathbf{1})$ such that for any $\mathbf{h}\in\ell_2(\mu\otimes\mathbf{1})$, $(M_G\mathbf{h})(x) = G(x)\mathbf{h}(x)$. We also call these block diagonal operators.*

In other words, $M_G$ is multiplication by a function $G$ in the direct integral of Hilbert spaces [50]. We will often use the multiplication operator $E_H^\theta$ defined as $(E^\theta\mathbf{h})(x) = \exp(\theta H(x))\mathbf{h}(x)$, sometimes dropping the subscript $H$ when it is clear from context.

## 4. Starting point

Our starting point is the following matrix moment generating function:

$$\mathbb{E}_\mu\left[\left\|\prod_{j=1}^n \exp\left(\frac{\theta e^{i\phi}}{2}F_j(s_j)\right)\right\|_F^2\right], \tag{4.1}$$

where $\phi\in[-\pi/2,\pi/2]$, $\theta>0$, and $s_1,\ldots,s_n$ is a sequence of states driven by the Markov chain with transition matrix $P$.

To simplify this moment generating function, we first use the property that $\mathrm{tr}[AB^\top] = \mathrm{vec}(I_d)^\top(A\otimes B)\,\mathrm{vec}(I_d)$, where $\otimes$ is the Kronecker product and $I_d$ is the $d\times d$ identity matrix to write

$$\begin{aligned}
&\left\|\prod_{j=1}^n \exp\left(\frac{\theta e^{i\phi}}{2}F_j(s_j)\right)\right\|_F^2 \\
&= \mathrm{tr}\left[\prod_{j=1}^n \exp\left(\frac{\theta e^{i\phi}}{2}F_j(s_j)\right)\prod_{j=n}^1 \exp\left(\frac{\theta e^{-i\phi}}{2}F_j(s_j)\right)\right] \\
&= \mathrm{vec}(I_d)^\top\left(\prod_{j=1}^n \exp\left(\frac{\theta e^{i\phi}}{2}F_j(s_j)\right)\otimes\prod_{j=1}^n \exp\left(\frac{\theta e^{-i\phi}}{2}F_j(s_j)\right)\right)\mathrm{vec}(I_d)
\end{aligned} \tag{4.2}$$

and then use successive applications of the property $(AC)\otimes(BD) = (A\otimes B)(C\otimes D)$ to write the above as

$$\mathrm{vec}(I_d)^\top\left(\prod_{j=1}^n \exp\left(\frac{\theta e^{i\phi}}{2}F_j(s_j)\right)\otimes\exp\left(\frac{\theta e^{-i\phi}}{2}F_j(s_j)\right)\right)\mathrm{vec}(I_d).$$

Lastly, we use the fact that $\exp(A)\otimes\exp(B) = \exp(A\otimes I_d + I_d\otimes B)$ to ultimately write

$$\mathrm{tr}\left[\prod_{j=1}^n \exp\left(\frac{\theta e^{i\phi}}{2}F_j(s_j)\right)\prod_{j=n}^1 \exp\left(\frac{\theta e^{-i\phi}}{2}F_j(s_j)\right)\right] = \mathrm{vec}(I_d)^*\left(\prod_{j=1}^n \exp(\theta H_j(s_j))\right)\mathrm{vec}(I_d) \tag{4.3}$$

where $H_1, \dots, H_n$ is a sequence of matrix-valued function on $\mathscr{X}$, implicitly with respect to $\phi$, defined as

$$H_j(x) = \frac{e^{i\phi}}{2} F_j(x) \otimes I_d + \frac{e^{-i\phi}}{2} I_d \otimes F_j(x) \in \mathbb{C}^{d^2 \times d^2} \tag{4.4}$$

for $x \in \mathscr{X}$. Note that unfortunately $H_j(x)$ is not in general Hermitian for any $x$, though its real and imaginary parts are both symmetric. We emphasize the identity

$$\exp(A) \otimes \exp(B) = \exp(A \otimes I_d + I_d \otimes B)$$

as we will switch back and forth from these two forms as needed.

The above allows us to focus on the matrices $H_j$, which are measurable and satisfy many of the same probabilistic properties as $F_j$ – a more precise statement is given in the next section. We then write

$$\begin{aligned} &\mathbb{E}_\mu \left[ \prod_{j=1}^n \exp(\theta H_j(s_j)) \right] \\ &= \int P(s_1, ds_2) \dots P(s_{n-1}, ds_n) \prod_{j=1}^n \exp(\theta H_j(s_j))\, d\mu(s_1) \\ &= \int d\mu(s_1) \int P(s_1, ds_2) \exp(\theta H_1(s_1)) \cdots \int P(s_{n-1}, ds_n) \exp(\theta H_{n-1}(s_{n-1})) \exp(\theta H_n(s_n)). \end{aligned} \tag{4.5}$$

Recall that $E_j^\theta$ the multiplication operator defined as $(E_j^\theta \mathbf{h})(x) = \exp(\theta H_j(x))\mathbf{h}(x)$ for any vector-valued function $\mathbf{h}$ – see Definition 3.4, and recall that $\tilde{P} = P \otimes I_{d^2}$. $E_j^\theta$ and $\tilde{P}$ act via

$$(\tilde{P} E_j^\theta \mathbf{h})(x) = \int \exp(\theta H_j(y)) \mathbf{h}(y) P(x, dy)$$

and

$$(E_j^{\theta/2} \tilde{P} E_j^{\theta/2} \mathbf{h})(x) = \int \exp\left(\frac{\theta}{2} H_j(x)\right) \exp\left(\frac{\theta}{2} H_j(y)\right) \mathbf{h}(y) P(x, dy)$$

as operators. Then simplifying Equation 4.5, and using the definition of the inner product on $\ell_2(\mu \otimes \mathbf{1})$, we have

$$\mathbb{E}_\mu \left[ \prod_{j=1}^n \exp(\theta H_j(s_j)) \right] = \left\langle \mathbf{1} \otimes \mathrm{vec}(I_d), E_1^{\theta/2} \left( \prod_{j=1}^{n-1} E_j^{\theta/2} \tilde{P} E_{j+1}^{\theta/2} \right) E_n^{\theta/2} (\mathbf{1} \otimes \mathrm{vec}(I_d)) \right\rangle_\mu . \tag{4.6}$$

## 5. Bounding the operator norm

Having derived the above, in this section we give a series of bounds that shift our focus to bounding the norm of a perturbed operator; we address the main technical challenges of studying general continuous state spaces.

From Equation 4.6, we apply Cauchy-Schwartz and submultiplicativity of the spectral norm to obtain

$$\begin{aligned}
&\left\langle \mathbf{1}\otimes \mathrm{vec}(I_d), E_1^{\theta/2}\left(\prod_{j=1}^{n-1} E_j^{\theta/2}\tilde{P}E_{j+1}^{\theta/2}\right)E_n^{\theta/2}(\mathbf{1}\otimes \mathrm{vec}(I_d))\right\rangle_\mu \\
&= \left\langle E_1^{\theta/2^*}(\mathbf{1}\otimes \mathrm{vec}(I_d)), \left(\prod_{j=1}^{n-1} E_j^{\theta/2}\tilde{P}E_{j+1}^{\theta/2}\right)E_n^{\theta/2}(\mathbf{1}\otimes \mathrm{vec}(I_d))\right\rangle_\mu \qquad (5.1)\\
&\leq \left\|E_1^{\theta/2^*}(\mathbf{1}\otimes \mathrm{vec}(I_d))\right\|_\mu \left\|E_n^{\theta/2}(\mathbf{1}\otimes \mathrm{vec}(I_d))\right\|_\mu \prod_{j=1}^{n-1}\left\|E_j^{\theta/2}\tilde{P}E_{j+1}^{\theta/2}\right\|_\mu .
\end{aligned}$$

Now $\tilde{P}$ possesses many of the same spectral properties as $P$, most notably the result from Lemma 3.3. Recall that $\hat{P}$ is the "lifted" Leon-Perron operator for $\tilde{P}$, defined as $\hat{P} = (\lambda I + (1-\lambda)\Pi)\otimes I_{d^2} = \lambda I + (1-\lambda)\tilde{\Pi}$ with $\tilde{\Pi} = \Pi \otimes I_{d^2}$. The following lemma allows us to replace $\tilde{P}$ with its Leon-Perron version $\hat{P}$.

**Lemma 5.1** *The operator $\tilde{P}$ and its Leon-Perron version $\hat{P}$ satisfy the following:*

*(1) For any $\mathbf{g},\mathbf{h}\in\ell_2(\mu\otimes\mathbf{1})$, $\left|\left\langle \mathbf{g},\tilde{P}\mathbf{h}\right\rangle_\mu\right| \leq \left\langle \mathbf{g},\hat{P}\mathbf{g}\right\rangle_\mu^{1/2}\left\langle \mathbf{h},\hat{P}\mathbf{h}\right\rangle_\mu^{1/2}$.*

*(2) For operators $S_1,S_2$ on $\ell_2(\mu\otimes\mathbf{1})$, $\left\|S_1\tilde{P}S_2\right\|_\mu \leq \left\|S_1\hat{P}S_1^*\right\|_\mu^{1/2}\left\|S_2\hat{P}S_2^*\right\|_\mu^{1/2}$.*

*(3) For any multiplication operator $M_G$ with respect to a measurable matrix valued function $G$, and for any vector $\mathbf{v}$, $\|M_G(\mathbf{1}\otimes\mathbf{v})\|_\mu \leq \|\mathbf{v}\|\left\|M_G\hat{P}M_G^*\right\|_\mu^{1/2}$.*

*Proof* For (1), we have

$$\begin{aligned}
\left|\left\langle \mathbf{g},\tilde{P}\mathbf{h}\right\rangle_\mu\right| &= \left|\left\langle \mathbf{g},(\tilde{P}-\tilde{\Pi}+\tilde{\Pi})\mathbf{h}\right\rangle\right| \\
&= \left|\left\langle \mathbf{g},\tilde{\Pi}\mathbf{h}\right\rangle + \left\langle \mathbf{g},(\tilde{P}-\tilde{\Pi})\mathbf{h}\right\rangle\right| \\
&\leq \left|\left\langle \mathbf{g},\tilde{\Pi}\mathbf{h}\right\rangle_\mu\right| + \left|\left\langle \mathbf{g}-\tilde{\Pi}\mathbf{g},(\tilde{P}-\tilde{\Pi})(\mathbf{h}-\tilde{\Pi}\mathbf{h})\right\rangle_\mu\right| \\
&\leq \left|\left\langle \mathbf{g},\tilde{\Pi}\mathbf{h}\right\rangle_\mu\right| + \lambda(\tilde{P})\left\|(I-\tilde{\Pi})\mathbf{g}\right\|_\mu\left\|(I-\tilde{\Pi})\mathbf{h}\right\|_\mu .
\end{aligned}$$

where the first inequality follows because $\tilde{\Pi}\tilde{P} = \tilde{P}\tilde{\Pi} = \tilde{\Pi}$ and because $\tilde{\Pi}$ is a projection and is thus self-adjoint on $\ell_2(\mu\otimes\mathbf{1})$. Because of this we also have $\left\langle \mathbf{g},\tilde{\Pi}\mathbf{h}\right\rangle_\mu = \left\langle \tilde{\Pi}\mathbf{g},\tilde{\Pi}\mathbf{h}\right\rangle_\mu$. Therefore, by Cauchy-Schwartz, $\left|\left\langle \mathbf{g},\tilde{\Pi}\mathbf{h}\right\rangle_\mu\right| \leq \left\|\tilde{\Pi}\mathbf{g}\right\|_\mu\left\|\tilde{\Pi}\mathbf{h}\right\|_\mu$. Then

$$\begin{aligned}
\left|\left\langle \mathbf{g},\tilde{P}\mathbf{h}\right\rangle_\mu\right| &\leq \left|\left\langle \mathbf{g},\tilde{\Pi}\mathbf{h}\right\rangle_\mu\right| + \lambda(\tilde{P})\left\|(I-\tilde{\Pi})\mathbf{g}\right\|_\mu\left\|(I-\tilde{\Pi})\mathbf{h}\right\|_\mu \\
&\leq \left\|\tilde{\Pi}\mathbf{g}\right\|_\mu\left\|\tilde{\Pi}\mathbf{h}\right\|_\mu + \lambda(\tilde{P})\left\|(I-\tilde{\Pi})\mathbf{g}\right\|_\mu\left\|(I-\tilde{\Pi})\mathbf{h}\right\|_\mu \\
&\leq \sqrt{\lambda\left\|(I-\tilde{\Pi})\mathbf{g}\right\|_\mu^2 + \left\|\tilde{\Pi}\mathbf{g}\right\|_\mu^2}\cdot\sqrt{\lambda\left\|(I-\tilde{\Pi})\mathbf{h}\right\|_\mu^2 + \left\|\tilde{\Pi}\mathbf{h}\right\|_\mu^2} \\
&= \left\langle \mathbf{g},\hat{P}\mathbf{g}\right\rangle_\mu^{1/2}\left\langle \mathbf{h},\hat{P}\mathbf{h}\right\rangle_\mu^{1/2} .
\end{aligned}$$

For (2), we have

$$\begin{aligned}
\left\|S_1\tilde{P}S_2\right\|_\mu &= \sup_{\|\mathbf{g}\|_\mu=\|\mathbf{h}\|_\mu=1} \left|\left\langle \mathbf{g}, S_1\tilde{P}S_2\mathbf{h}\right\rangle_\mu\right| \\
&= \sup_{\|\mathbf{g}\|_\mu=\|\mathbf{h}\|_\mu=1} \left|\left\langle S_1^*\mathbf{g}, \tilde{P}S_2\mathbf{h}\right\rangle_\mu\right| \\
&\leq \sup_{\|\mathbf{g}\|_\mu=\|\mathbf{h}\|_\mu=1} \left\langle S_1^*\mathbf{g}, \hat{P}S_1^*\mathbf{g}\right\rangle_\mu^{1/2} \left\langle S_2\mathbf{h}, \hat{P}S_2\mathbf{h}\right\rangle_\mu^{1/2} \\
&= \left\|S_1\hat{P}S_1^*\right\|_\mu^{1/2} \left\|S_2^*\hat{P}S_2\right\|_\mu^{1/2}.
\end{aligned}$$

The result follows by noting that $\left\|S_1^*\hat{P}S_1\right\|_\mu = \left\|S_1\hat{P}S_1^*\right\|_\mu$.

For (3), the case $G = 0$ is clear. Otherwise,

$$\begin{aligned}
\left\|M_G\hat{P}M_G^*\right\|_\mu &\geq \frac{\left\langle M_G(\mathbf{1}\otimes\mathbf{v}), M_G\hat{P}M_G^*M_G(\mathbf{1}\otimes\mathbf{v})\right\rangle_\mu}{\|M_G(\mathbf{1}\otimes\mathbf{v})\|_\mu^2} \\
&= \frac{\left\langle M_G^*M_G(\mathbf{1}\otimes\mathbf{v}), \tilde{\Pi}M_G^*M_G(\mathbf{1}\otimes\mathbf{v})\right\rangle_\mu}{\|M_G(\mathbf{1}\otimes\mathbf{v})\|_\mu^2} + \frac{\lambda\left\langle M_G^*M_G(\mathbf{1}\otimes\mathbf{v}), (I-\tilde{\Pi})M_G^*M_G(\mathbf{1}\otimes\mathbf{v})\right\rangle_\mu}{\|M_G(\mathbf{1}\otimes\mathbf{v})\|_\mu^2} \\
&= \frac{\left\langle M_G^*M_G(\mathbf{1}\otimes\mathbf{v}), \tilde{\Pi}M_G^*M_G(\mathbf{1}\otimes\mathbf{v})\right\rangle_\mu}{\|M_G(\mathbf{1}\otimes\mathbf{v})\|_\mu^2} + \frac{\lambda\left\|(I-\tilde{\Pi})M_G^*M_G(\mathbf{1}\otimes\mathbf{v})\right\|_\mu^*}{\|M_G(\mathbf{1}\otimes\mathbf{v})\|_\mu^2} \\
&\geq \frac{\left\langle M_G^*M_G(\mathbf{1}\otimes\mathbf{v}), \tilde{\Pi}M_G^*M_G(\mathbf{1}\otimes\mathbf{v})\right\rangle_\mu}{\|M_G(\mathbf{1}\otimes\mathbf{v})\|_\mu^2},
\end{aligned}$$

where the second to last line follows because $I-\tilde{\Pi}$ is a projection. We can express these terms more explicitly. First, for the denominator, we have $(M_G(\mathbf{1}\otimes\mathbf{v}))(x) = G(x)\mathbf{v}$. Therefore, we have

$$\|M_G(\mathbf{1}\otimes\mathbf{v})\|_\mu^2 = \left\langle \mathbf{v}, \mathbb{E}_\mu[G(x)^*G(x)]\mathbf{v}\right\rangle.$$

Similarly, we have $\tilde{\Pi}M_G^*M_G(\mathbf{1}\otimes\mathbf{v})(x) = \mathbb{E}_\mu[G(y)^*G(y)]\mathbf{v}$ for any $x \in \mathscr{X}$. Then

$$\left\langle M_G^*M_G(\mathbf{1}\otimes\mathbf{v}), \tilde{\Pi}M_G^*M_G(\mathbf{1}\otimes\mathbf{v})\right\rangle_\mu = \left\langle \mathbb{E}_\mu[G(x)^*G(x)]\mathbf{v}, \mathbb{E}_\mu[G(x)^*G(x)]\mathbf{v}\right\rangle.$$

Now for any Hermitian matrix $A$ and any vector $\mathbf{v} \neq 0$ it holds that $\langle A\mathbf{v}, A\mathbf{v}\rangle \geq \langle \mathbf{v}, A\mathbf{v}\rangle^2 / \|\mathbf{v}\|^2$ as a simple consequence of Cauchy-Schwartz. Therefore, if $A$ is positive semidefinite, then $\langle A\mathbf{v}, A\mathbf{v}\rangle / \langle \mathbf{v}, A\mathbf{v}\rangle \geq$

$\langle \mathbf{v}, A\mathbf{v}\rangle / \|\mathbf{v}\|^2$. As $\mathbb{E}_\mu[G(x)^* G(x)]$ is indeed positive semidefinite,

$$\begin{aligned}
\left\|M_G \hat{P} M_G^*\right\|_\mu &\geq \frac{\left\langle M_G^* M_G(\mathbf{1}\otimes\mathbf{v}), \tilde{\Pi} M_G^* M_G(\mathbf{1}\otimes\mathbf{v})\right\rangle_\mu}{\|M_G(\mathbf{1}\otimes\mathbf{v})\|_\mu^2} \\
&= \frac{\left\langle \mathbb{E}_\mu[G(x)^*G(x)]\mathbf{v}, \mathbb{E}_\mu[G(x)^*G(x)]\mathbf{v}\right\rangle}{\left\langle \mathbf{v}, \mathbb{E}_\mu[G(x)^*G(x)]\mathbf{v}\right\rangle} \\
&\geq \frac{\left\langle \mathbf{v}, \mathbb{E}_\mu[G(x)^*G(x)]\mathbf{v}\right\rangle}{\|\mathbf{v}\|^2} \\
&= \frac{\|M_G(\mathbf{1}\otimes\mathbf{v})\|_\mu^2}{\|\mathbf{v}\|^2}.
\end{aligned}$$

Rearranging finishes the proof. □

Going back to Equation 5.1, applying Lemma 5.1 parts (2) and (3), and noticing that $\|\mathrm{vec}(I_d)\| = \sqrt{d}$, we establish

$$\begin{aligned}
&\left\langle \mathbf{1}\otimes\mathrm{vec}(I_d), E_1^{\theta/2}\left(\prod_{j=1}^{n-1} E_j^{\theta/2}\tilde{P}E_{j+1}^{\theta/2}\right)E_n^{\theta/2}(\mathbf{1}\otimes\mathrm{vec}(I_d))\right\rangle_\mu \\
&\leq \left\|E_1^{\theta/2^*}(\mathbf{1}\otimes\mathrm{vec}(I_d))\right\|_\mu \left\|E_n^{\theta/2}(\mathbf{1}\otimes\mathrm{vec}(I_d))\right\|_\mu \prod_{j=1}^{n-1}\left\|E_j^{\theta/2}\tilde{P}E_{j+1}^{\theta/2}\right\|_\mu \qquad (5.2) \\
&\leq d\prod_{j=1}^{n}\left\|E_j^{\theta/2}\hat{P}E_j^{\theta/2^*}\right\|_\mu.
\end{aligned}$$

Thus, we have transferred the study of the problem from $\tilde{P}$, which represents a general, nonreversible Markov chain, to $\hat{P}$, which represents a reversible chain and thus is much simpler to analyze. Therefore, the operators $E_j^{\theta/2}\hat{P}E_j^{\theta/2^*}$ are self-adjoint on $\ell_2(\mu\otimes\mathbf{1})$. Now we focus on bounding the leading eigenvalues of these operators; we make one more simplification:

**Proposition 5.2** *Let $T_j(x) = \frac{\cos(\phi)}{2}(F_j(x)\otimes I_d + I_d\otimes F(x)))$. Let $E_{T_j}^\theta$ be the operator defined as $(E_{T_j}^\theta \mathbf{h})(x) = \exp(\theta T_j(x))\mathbf{h}(x)$. Then the norms of $E_j^{\theta/2}\hat{P}E_j^{\theta/2^*}$ and $E_{T_j}^{\theta/2}\hat{P}E_{T_j}^{\theta/2}$ are the same.*

*Proof* The operator $E_j^{\theta/2}\hat{P}E_j^{\theta/2^*}$ is similar to the operator $\hat{P}E_j^{\theta/2^*}E_j^{\theta/2}$. The operator $E_j^{\theta/2^*}E_j^{\theta/2}$ acts via

$$\begin{aligned}(E_j^{\theta/2^*}E_j^{\theta/2}\mathbf{h})(x) &= \exp\left(\frac{\theta}{2}(H_j^*(x)+H_j(x))\right)\mathbf{h}(x)\\ &= \exp\left(\frac{\theta}{2}\left(\frac{e^{-i\phi}+e^{i\phi}}{2}F_j(x)\otimes I_d + \frac{e^{i\phi}+e^{-i\phi}}{2}I_d\otimes F_j(x)\right)\right)\mathbf{h}(x)\\ &= \exp\left(\frac{\theta\cos(\phi)}{2}(F_j(x)\otimes I_d + I_d\otimes F_j(x))\right)\mathbf{h}(x)\\ &= (E_{T_j}^{\theta}\mathbf{h})(x),\end{aligned}$$

where the first equality holds because $H_j(x), H_j^*(x)$ commute. As $\hat{P}E_{T_j}^{\theta}$ is similar to $E_{T_j}^{\theta/2}\hat{P}E_{T_j}^{\theta/2}$, we see that $E_j^{\theta/2}\hat{P}E_j^{\theta/2^*}$ and $E_{T_j}^{\theta/2}\hat{P}E_{T_j}^{\theta/2}$ have the same spectrum; as they are both self-adjoint, the spectral radius equals the operator norm. □

At this point, we have reduced our problem to a considerably simpler form. We can now show that the operators $T_j$ satisfy many of the same probabilistic properties as $F_j$; the proof is in the appendix.

**Proposition 5.3** *Let $T$ be the operator defined as $T(x) = \frac{\cos(\phi)}{2}(F(x)\otimes I_d + I_d\otimes F(x))$ with $\phi \in [-\pi/2, \pi/2]$ so that $\cos(\phi)$ is always nonnegative. Then*

*(1) If $\mathbb{E}[F(x)] = 0$, then $\mathbb{E}[T(x)] = 0$,*
*(2) If $aI \preceq F(x) \preceq bI$ for all $x \in \mathscr{X}$, then $a\cos(\phi)I \preceq T(x) \preceq b\cos(\phi)I$ for all $x \in \mathscr{X}$,*
*(3) If $\left\|\mathbb{E}[F(x)^2]\right\| \leq \mathscr{V}$, then $\left\|\mathbb{E}[T(x)^2]\right\| \leq \cos^2(\phi)\mathscr{V}$,*

*where all expectations are taken with respect to a measure on $\mathscr{X}$, and $F$ and $T$ are measurable.*

### 5.1. *The leading eigenvalue as the limit*

In this subsection we will prove a limit lemma, justifying the use of the leading eigenvalue as a characterization of the moment generating function, mirroring the scalar case as laid out in [10] – essentially, what we will prove is that the log limit of the moment generating function is the logarithm of the leading eigenvalue. This will allow us to shift between time-dependent and time-independent functions as needed. The main difficulty in extending this theory from the scalar case is proving the limit statement in part (3) of Lemma 5.5, and in particular, a specific lower bound in the proof. This lower bound is not difficult to prove in the scalar setting; however, the matrix setting is more subtle and requires some more care.

Our main lemma is the following:

**Lemma 5.4** *Let $F$ be a map from $\mathscr{X}$ to real symmetric $d\times d$ matrices with $aI \preceq F(x) \preceq bI$ for all $x \in \mathscr{X}$. Let $T(x) = \frac{1}{2}(F(x)\otimes I_d + I_d\otimes F(x))$, $x \in \mathscr{X}$, and let $E_T^{\theta/2}$ be the operator defined as $(E_T^{\theta}\mathbf{h})(x) = \exp(\theta T(x))\mathbf{h}(x)$. Let $\mu$ be a distribution, let $s_1,\ldots,s_n$ be a sequence of states driven by the stationary Markov chain with transition matrix $\lambda I + (1-\lambda)\mathbf{1}\mu^\top$, and let $\hat{P}$ be this transition matrix*

*tensored with $I_{d^2}$. Then*

$$\lim_{n\to\infty}\frac{1}{n}\log\mathbb{E}_\mu\left[\left\|\prod_{j=1}^{n}\exp\left(\frac{\theta}{2}F(s_j)\right)\right\|_F^2\right]=\log\left\|E_T^{\theta/2}\hat{P}E_T^{\theta/2}\right\|_\mu.$$

In other words, we take $s_1,\dots,s_n$ to be driven by a Leon-Perron operator, where at a given state the chain stays at that state with probability $\lambda$ and samples a state from $\mu$ with probability $1-\lambda$. We first address the case of a simple matrix-valued function $F$.

**Lemma 5.5** *Let $\hat{P}=(\lambda I+(1-\lambda)\Pi)\otimes I_{d^2}$ be a Leon-Perron operator. Let $F$ be a simple function $\mu$-almost everywhere, that is, there exist a finite set of real symmetric matrices $\{B_1,\dots,B_m\}$ such that $F^{-1}(B_j)=\{x\in\mathscr{X}\mid F(x)=B_j\}$ satisfies*

$$\mu(F^{-1}(B_j))>0,\forall j\in[m],\quad \sum_{j\in[m]}\mu(F^{-1}(B_j))=1.$$

*Assume that for all $B_j$, $aI\preceq B_j\preceq bI$, and let these be tight. Let $T$ be the operator defined as $T(x)=\frac{1}{2}(F(x)\otimes I_d+I_d\otimes F(x))$, with exponential multiplication operator $E_T^\theta$. Define the matrix-valued function*

$$\mathscr{F}(r)=\mathbb{E}_\mu\left[(1-\lambda)\exp\left(\frac{\theta}{2}T(x)\right)(rI-\lambda\exp(\theta T(x)))^{-1}\exp\left(\frac{\theta}{2}T(x)\right)\right].$$

*Then the following hold:*

*(1) Let $r^*$ be such that $\mathscr{F}(r^*)$ has an eigenvalue of 1, and let $\mathbf{v}$ be an eigenvector corresponding to the eigenvalue 1 of $\mathscr{F}(r^*)$. Then $r^*$ is an eigenvalue of $E_T^{\theta/2}\hat{P}E_T^{\theta/2}$ with corresponding eigenfunction $\mathbf{h}$ such that*

$$\mathbf{h}(x)=(1-\lambda)(r^*I-\lambda\exp(\theta T(x)))^{-1}\exp\left(\frac{\theta}{2}T(x)\right)\mathbf{v}.$$

*Conversely, every eigenvalue of $E_T^{\theta/2}\hat{P}E_T^{\theta/2}$ is defined this way.*

*(2) Let $\rho$ be the largest eigenvalue of $E_T^{\theta/2}\hat{P}E_T^{\theta/2}$. Then $\rho>\lambda e^{\theta b}$.*

*(3) $\left\|E_T^{\theta/2}\hat{P}E_T^{\theta/2}\right\|_\mu=\rho$ and*

$$\lim_{n\to\infty}\frac{1}{n}\log\left\langle\mathbf{1}\otimes\mathrm{vec}(I_d),E_T^{\theta/2}(E_T^{\theta/2}\hat{P}E_T^{\theta/2})^{n-1}E_T^{\theta/2}(\mathbf{1}\otimes\mathrm{vec}(I_d))\right\rangle_\mu=\log\rho=\log\left\|E_T^{\theta/2}\hat{P}E_T^{\theta/2}\right\|_\mu.$$

*Proof* For (1), if $r^*$ is an eigenvalue of $E_T^{\theta/2}\hat{P}E_T^{\theta/2}$ with eigenfunction $\mathbf{h}$, then

$$\begin{aligned}
&E_T^{\theta/2}\hat{P}E_T^{\theta/2}\mathbf{h}-r^*\mathbf{h}=0\\
\iff&\lambda E_T^{\theta}\mathbf{h}+(1-\lambda)E_T^{\theta/2}\tilde{\Pi}E_T^{\theta/2}\mathbf{h}-r^*\mathbf{h}=0\\
\iff&\lambda\exp(\theta T(x))\mathbf{h}(x)+(1-\lambda)\exp\left(\frac{\theta}{2}T(x)\right)\mathbb{E}_\mu\left[\exp\left(\frac{\theta}{2}T(y)\right)\mathbf{h}(y)\right]-r^*\mathbf{h}(x)=0.
\end{aligned}$$

Let $\mathbf{v}$ as defined in the statement of part (1). Plugging in $\mathbf{h}$, the first term on the left-hand side is equal to

$$\lambda(1-\lambda)\exp(\theta T(x))(r^*I-\lambda\exp(\theta T(x)))^{-1}\exp\left(\frac{\theta}{2}T(x)\right)\mathbf{v}.$$

The second term is equal to

$$\begin{aligned}
&(1-\lambda)\exp\left(\frac{\theta}{2}T(x)\right)\mathbb{E}_\mu\left[(1-\lambda)\exp\left(\frac{\theta}{2}T(y)\right)(r^*I-\lambda\exp(\theta T(y)))^{-1}\exp\left(\frac{\theta}{2}T(y)\right)\mathbf{v}\right]\\
&=(1-\lambda)\exp\left(\frac{\theta}{2}T(x)\right)\mathscr{F}(r^*)\mathbf{v}\\
&=(1-\lambda)\exp\left(\frac{\theta}{2}T(x)\right)\mathbf{v}\\
&=(1-\lambda)(r^*I-\lambda\exp(\theta T(x)))(r^*I-\lambda\exp(\theta T(x)))^{-1}\exp\left(\frac{\theta}{2}T(x)\right)\mathbf{v}\\
&=r^*(1-\lambda)(r^*I-\lambda\exp(\theta T(x)))^{-1}\exp\left(\frac{\theta}{2}T(x)\right)\mathbf{v}\\
&\quad-\lambda(1-\lambda)\exp(\theta T(x))(r^*I-\lambda\exp(\theta T(x)))^{-1}\exp\left(\frac{\theta}{2}T(x)\right)\mathbf{v}
\end{aligned}$$

The third term is equal to

$$r^*(1-\lambda)(r^*I-\lambda\exp(\theta T(x)))^{-1}\exp\left(\frac{\theta}{2}T(x)\right)\mathbf{v}.$$

Adding the first two and subtracting the third gives zero, and shows that $r^*$ is indeed an eigenvalue with eigenfunction $\mathbf{h}$.

Conversely, if $(r^*,\mathbf{h})$ is an eigenpair of $E_T^{\theta/2}\hat{P}E_T^{\theta/2}$, then solving for $\mathbf{h}$ in the equation

$$\lambda\exp(\theta T(x))\mathbf{h}(x)+(1-\lambda)\exp\left(\frac{\theta}{2}T(x)\right)\mathbb{E}_\mu\left[\exp\left(\frac{\theta}{2}T(y)\right)\mathbf{h}(y)\right]-r^*\mathbf{h}(x)=0$$

gives

$$\mathbf{h}(x)=(1-\lambda)(r^*I-\lambda\exp(\theta T(x)))^{-1}\exp\left(\frac{\theta}{2}T(x)\right)\mathbb{E}_\mu\left[\exp\left(\frac{\theta}{2}T(x)\right)\mathbf{h}(x)\right].$$

Multiplying by $\exp\left(\frac{\theta}{2}T(x)\right)$ on both sides and taking expectations, we see that

$$\mathbb{E}_\mu\left[\exp\left(\frac{\theta}{2}T(x)\right)\mathbf{h}(x)\right]=\mathscr{F}(r^*)\mathbb{E}_\mu\left[\exp\left(\frac{\theta}{2}T(x)\right)\mathbf{h}(x)\right]$$

so indeed $r^*$ is such that $\mathscr{F}(r^*)$ has an eigenvalue of 1 with $\mathbf{v}=\mathbb{E}_\mu\left[\exp\left(\frac{\theta}{2}T(x)\right)\mathbf{h}(x)\right]$.

For (2), the function $\mathscr{F}(r)$ is continuous in $r$ when $r$ is large enough (so that the inverses exist); indeed, since $F$ is simple the expectation decomposes as a finite sum, so $\mathscr{F}$ is continuous if all of the summands are continuous. In particular, since the eigenvalues of all $B_j$ are bounded above by $b$, when

$r > \lambda e^{\theta b}$ $\mathscr{F}(r)$ is continuous in $r$. Now when $r \to \infty$ the entries of $\mathscr{F}(r)$ approach zero, implying that all eigenvalues of $\mathscr{F}(r)$ also approach zero as the eigenvalues are continuous in the entries. On the other hand, $\mathscr{F}(r)$ decomposes as a finite sum. Consider the summand corresponding to some $B_j$ such that $b$ is its largest eigenvalue. As $r \to \lambda e^{\theta b}$ from above it is clear that the largest eigenvalue of this summand goes to $\infty$; then it is clear that the largest eigenvalue of this finite sum must also go to $\infty$; implying that at least one eigenvalue of $\mathscr{F}(r)$ goes to $\infty$. Thus, as $\mathscr{F}(r)$ is entrywise continuous in $r$, and as the eigenvalues are continuous in the entries, for some $r^* \in (\lambda e^{\theta b}, \infty)$ $\mathscr{F}(r^*)$ has an eigenvalue of 1. By part (1). this implies that $r^*$ is an eigenvalue of $E_T^{\theta/2} \hat{P} E_T^{\theta/2}$. Now let $\rho$ be the largest of all such eigenvalues, and so we have shown that $\rho > \lambda e^{\theta b}$.

For (3), let $\sigma(E_T^{\theta/2} \hat{P} E_T^{\theta/2})$ be the spectrum of $E_T^{\theta/2} \hat{P} E_T^{\theta/2}$. As $E_T^{\theta/2} \hat{P} E_T^{\theta/2}$ is self-adjoint on $\ell_2(\mu \otimes \mathbf{1})$, the spectrum is real and decomposes as the disjoint union of the essential spectrum $\sigma_{\text{ess}}(E_T^{\theta/2} \hat{P} E_T^{\theta/2})$ and the discrete spectrum $\sigma_{\text{d}}(E_T^{\theta/2} \hat{P} E_T^{\theta/2})$, the latter consisting of all eigenvalues with finite multiplicity. Define the essential spectral radius as the largest element of the essential spectrum and the spectral radius as the largest element of the spectrum, so the latter is always at least the former. Now $E_T^{\theta/2} \hat{P} E_T^{\theta/2} = \lambda E_T^{\theta} + (1-\lambda) E^{\theta/2} \tilde{\Pi} E^{\theta/2}$ and thus is a compact perturbation of the operator $\lambda E_T^{\theta}$, as $(1-\lambda) E^{\theta/2} \tilde{\Pi} E^{\theta/2}$ is finite rank (rank at most $d^2$). Weyl's perturbation theorem [67] ensures that $\sigma_{\text{ess}}(E_T^{\theta/2} \hat{P} E_T^{\theta/2}) = \sigma_{\text{ess}}(\lambda E_T^{\theta/2})$. As $\lambda E_T^{\theta/2}$ is a block diagonal operator (Definition 3.4), its spectrum is the union of the spectra of $\lambda \exp(\theta T(x)), x \in \mathscr{X}$ [50], and so the essential spectral radius is bounded above by $\lambda e^{\theta b}$. Therefore, by part (2). the largest eigenvalue $\rho$ is equal to the spectral radius of $E_T^{\theta/2} \hat{P} E_T^{\theta/2}$ and thus $\left\| E_T^{\theta/2} \hat{P} E_T^{\theta/2} \right\|_{\mu} = \rho$.

To prove the second statement, from Lemma 5.1 we have

$$\left\langle \mathbf{1} \otimes \text{vec}(I_d), E_T^{\theta/2} \left( E_T^{\theta/2} \hat{P} E_T^{\theta/2} \right)^{n-1} E_T^{\theta/2} (\mathbf{1} \otimes \text{vec}(I_d)) \right\rangle_{\mu} \leq d \left\| E_T^{\theta/2} \hat{P} E_T^{\theta/2} \right\|_{\mu}^{n} = d \rho^n.$$

This implies

$$\limsup_{n \to \infty} \frac{1}{n} \log \left\langle \mathbf{1} \otimes \text{vec}(I_d), E_T^{\theta/2} \left( E_T^{\theta/2} \hat{P} E_T^{\theta/2} \right)^{n-1} E_T^{\theta/2} (\mathbf{1} \otimes \text{vec}(I_d)) \right\rangle_{\mu} \leq \log \rho. \tag{5.3}$$

Let $\mathbf{h}$ be a unit norm eigenfunction corresponding to $\rho$, and let $\tilde{\mathbf{h}}$ be the projection of $E_T^{\theta/2}(\mathbf{1} \otimes \text{vec}(I_d))$ onto $\mathbf{h}$; in other words,

$$\tilde{\mathbf{h}} = \left\langle E_T^{\theta/2} (\mathbf{1} \otimes \text{vec}(I_d)), \mathbf{h} \right\rangle_{\mu} \mathbf{h}. \tag{5.4}$$

Then

$$\left\langle E_T^{\theta/2} (\mathbf{1} \otimes \text{vec}(I_d)) - \tilde{\mathbf{h}}, (E_T^{\theta/2} \hat{P} E_T^{\theta/2})^{n-1} \tilde{\mathbf{h}} \right\rangle_{\mu} = 0$$

by definition of $\tilde{\mathbf{h}}$ being an eigenfunction and it being a projection of $E_T^{\theta/2}(\mathbf{1} \otimes \text{vec}(I_d))$. Next, we have

$$\left\langle E_T^{\theta/2} (\mathbf{1} \otimes \text{vec}(I_d)) - \tilde{\mathbf{h}}, (E_T^{\theta/2} \hat{P} E_T^{\theta/2})^{n-1} (E_T^{\theta/2} (\mathbf{1} \otimes \text{vec}(I_d)) - \tilde{\mathbf{h}}) \right\rangle_{\mu} \geq 0$$

as $E_T^{\theta/2}\hat{P}E_T^{\theta/2}$ is positive semidefinite on $\ell_2(\mu\otimes\mathbf{1})$ – this follows from $E_T^{\theta/2}$ being self-adjoint and $\hat{P}$ being positive semidefinite on $\ell_2(\mu\otimes\mathbf{1})$. Then

$$\begin{aligned}
&\left\langle E_T^{\theta/2}(\mathbf{1}\otimes\mathrm{vec}(I_d)), (E_T^{\theta/2}\hat{P}E_T^{\theta/2})^{n-1}E_T^{\theta/2}(\mathbf{1}\otimes\mathrm{vec}(I_d))\right\rangle_\mu \\
&= \left\langle \tilde{\mathbf{h}}, (E_T^{\theta/2}\hat{P}E_T^{\theta/2})^{n-1}\tilde{\mathbf{h}}\right\rangle_\mu \\
&\quad + 2\left\langle E_T^{\theta/2}(\mathbf{1}\otimes\mathrm{vec}(I_d)) - \tilde{\mathbf{h}}, (E_T^{\theta/2}\hat{P}E_T^{\theta/2})^{n-1}\tilde{\mathbf{h}}\right\rangle_\mu \\
&\quad + \left\langle E_T^{\theta/2}(\mathbf{1}\otimes\mathrm{vec}(I_d)) - \tilde{\mathbf{h}}, (E_T^{\theta/2}\hat{P}E_T^{\theta/2})^{n-1}(E_T^{\theta/2}(\mathbf{1}\otimes\mathrm{vec}(I_d)) - \tilde{\mathbf{h}})\right\rangle_\mu \\
&\geq \left\langle \tilde{\mathbf{h}}, (E_T^{\theta/2}\hat{P}E_T^{\theta/2})^{n-1}\tilde{\mathbf{h}}\right\rangle_\mu \\
&= \left\langle E_T^{\theta/2}(\mathbf{1}\otimes\mathrm{vec}(I_d)), \mathbf{h}\right\rangle_\mu^2 \rho^{n-1}.
\end{aligned}$$

Assume $\left\langle E_T^{\theta/2}(\mathbf{1}\otimes\mathrm{vec}(I_d)), \mathbf{h}\right\rangle_\mu \neq 0$ for now. Then the above shows that

$$\liminf_{n\to\infty}\frac{1}{n}\log\left\langle \mathbf{1}\otimes\mathrm{vec}(I_d), E_T^{\theta/2}\left(E_T^{\theta/2}\hat{P}E_T^{\theta/2}\right)^{n-1}E_T^{\theta/2}(\mathbf{1}\otimes\mathrm{vec}(I_d))\right\rangle_\mu \geq \log\rho. \tag{5.5}$$

Putting Equations 5.3 and 5.5 together proves the statement.

We now show that $\left\langle E_T^{\theta/2}(\mathbf{1}\otimes\mathrm{vec}(I_d)), \mathbf{h}\right\rangle_\mu \neq 0$. To do this we make a connection to dynamical systems and the theory of positive operators in Banach spaces, and which uses our realness assumption on $F$. For a cone $K$ in a Banach space, an operator $A$ is *positive* with respect to $K$ if $AK\subseteq K$. In our case, let the Banach space under consideration be the space of $d\times d$ symmetric matrices equipped with the Frobenius norm, or equivalently, the space of length $d^2$ vectors whose reshaping to a $d\times d$ matrix is symmetric equipped with the Euclidean norm; we will show that it is no loss to restrict to this space. Let $\mathbf{S}_+^d$ be the cone of real positive semidefinite $d\times d$ matrices; it is a closed, convex cone. It is also *reproducing* in the space of real symmetric $d\times d$ matrices, in that every symmetric matrix $X$ can be written as $X = X_1 - X_2$ for $X_1, X_2$ positive semidefinite [27]. We can also identify $\mathbf{S}_+^d$ as the space of length $d^2$ vectors whose reshaping to a $d\times d$ matrix is positive semidefinite; we now make this identification. Now let $K\subset\ell_2(\mu\otimes\mathbf{1})$ be the infinite direct product of $\mathbf{S}_+^d$ indexed by elements of $\mathscr{X}$; in other words, for $\mathbf{h}\in\ell_2(\mu\otimes\mathbf{1})$, $\mathbf{h}\in K$ if and only if $\mathbf{h}(x)\in\mathbf{S}_+^d$ for all $x\in\mathscr{X}$. This cone is is closed, convex, and reproducing in the space of all $\mathbf{h}\in\ell_2(\mu\otimes\mathbf{1})$ such that $\mathbf{h}(x)$ can be reshaped to a symmetric $d\times d$ matrix for all $x\in\mathscr{X}$.

Next, we show that we can indeed restrict ourselves to the Banach space of symmetric matrices equipped with the Frobenius norm, in the sense that any eigenfunction $\mathbf{h}$ of $E_T^{\theta/2}\hat{P}E_T^{\theta/2}$ can be such that $\mathbf{h}(x)$ can be reshaped into a $d\times d$ symmetric matrix. As an operator, we see that

$$(E_T^{\theta/2}\hat{P}E_T^{\theta/2}\mathbf{h})(x) = \int \exp\left(\frac{\theta}{2}T(x)\right)\exp\left(\frac{\theta}{2}T(y)\right)\mathbf{h}(y)P(x,dy).$$

From the definition of $T$ we see that $\exp\left(\frac{\theta}{2}T(x)\right) = \exp\left(\frac{\theta}{4}F(x)\right)\otimes\exp\left(\frac{\theta}{4}F(x)\right)$, and so $\exp\left(\frac{\theta}{2}T(x)\right)\exp\left(\frac{\theta}{2}T(y)\right) = \exp\left(\frac{\theta}{4}F(x)\right)\exp\left(\frac{\theta}{4}F(y)\right)\otimes\exp\left(\frac{\theta}{4}F(x)\right)\exp\left(\frac{\theta}{4}F(y)\right)$. If $\mathbf{h}(y)$ has a

decomposition $\sum_j \mathbf{a}_j(y) \otimes \mathbf{b}_j(y)$ then

$$\begin{aligned}
&\exp\left(\frac{\theta}{2}T(x)\right)\exp\left(\frac{\theta}{2}T(y)\right)\mathbf{h}(y) \\
&=\sum_j \exp\left(\frac{\theta}{4}F(x)\right)\exp\left(\frac{\theta}{4}F(y)\right)\mathbf{a}_j(y)\otimes\exp\left(\frac{\theta}{4}F(x)\right)\exp\left(\frac{\theta}{4}F(y)\right)\mathbf{b}_j(y).
\end{aligned}$$

Then it is not hard to see that the function $\mathbf{h}'$ such that $\mathbf{h}'(y) = \sum_j \mathbf{b}_j(y) \otimes \mathbf{a}_j(y)$ is also an eigenfunction of $E_T^{\theta/2}\hat{P}E_T^{\theta/2}$. So take $(\mathbf{h}+\mathbf{h}')/2$ as the eigenfunction we desire. Next, we see that $E_T^{\theta/2}\hat{P}E_T^{\theta/2}$ is positive with respect to $K$. For $\mathbf{h} \in \ell_2(\mu \otimes \mathbf{1})$ denote $\hat{\mathbf{h}}(x)$ as the $d \times d$ matrix reshaped from the length $d^2$ vector $\mathbf{h}(x)$. From the action of $E_T^{\theta/2}\hat{P}E_T^{\theta/2}$ and properties of the Kronecker product we have

$$\exp\left(\frac{\theta}{2}T(x)\right)\exp\left(\frac{\theta}{2}T(y)\right)\mathbf{h}(y) = \exp\left(\frac{\theta}{4}F(x)\right)\exp\left(\frac{\theta}{4}F(y)\right)\hat{\mathbf{h}}(y)\exp\left(\frac{\theta}{4}F(y)\right)\exp\left(\frac{\theta}{4}F(x)\right).$$

It is clear that since $\hat{\mathbf{h}}(y)$ is positive semidefinite so is the right-hand side above. Then by linearity we see that the integral of this with respect to $P(x,dy)$ is also positive semidefinite. Therefore, $E_T^{\theta/2}\hat{P}E_T^{\theta/2}\mathbf{h} \in K$, since $x$ was arbitrary.

Now we use Corollary 2.2 of [51], which ensures that if a cone is closed, convex, and reproducing, and if $A$ is a positive linear operator with respect to the cone with spectral radius strictly larger than the essential spectral radius, then the spectral radius is an eigenvalue and there exists a corresponding nonzero eigenfunction *that lies in the cone*. In our case, we indeed have all the above are satisfied: our $K$ is closed, convex, and reproducing, $E_T^{\theta/2}\hat{P}E_T^{\theta/2}$ is positive with respect to $K$, and the spectral radius $\rho$ is strictly larger than the essential spectral radius, which is at most $\lambda e^{\theta b}$. Then there is a nonzero eigenfunction $\mathbf{h}$ in $K \subseteq \ell_2(\mu \otimes \mathbf{1})$.

This allows us to assert that in the derivation of Equation 5.5 we could have chosen $\mathbf{h} \in K$. With this choice of $\mathbf{h}$ it is straightforward in showing $\left\langle E_T^{\theta/2}(\mathbf{1}\otimes \mathrm{vec}(I_d)), \mathbf{h}\right\rangle_\mu \neq 0$. Indeed, we have

$$\begin{aligned}
\left\langle E_T^{\theta/2}(\mathbf{1}\otimes \mathrm{vec}(I_d)), \mathbf{h}\right\rangle_\mu &= \left\langle \mathbf{1}\otimes \mathrm{vec}(I_d), E_T^{\theta/2}\mathbf{h}\right\rangle_\mu \\
&= \left\langle \mathrm{vec}(I_d), \mathbb{E}_\mu\left[\exp\left(\frac{\theta}{2}T(x)\right)\mathbf{h}(x)\right]\right\rangle \\
&= \mathrm{tr}\,\mathbb{E}_\mu\left[\exp\left(\frac{\theta}{4}F(x)\right)\hat{\mathbf{h}}(x)\exp\left(\frac{\theta}{4}F(x)\right)\right] \\
&= \mathbb{E}_\mu\left[\mathrm{tr}\left[\exp\left(\frac{\theta}{4}F(x)\right)\hat{\mathbf{h}}(x)\exp\left(\frac{\theta}{4}F(x)\right)\right]\right].
\end{aligned}$$

Each of these matrices $\exp\left(\frac{\theta}{4}F(x)\right)\hat{\mathbf{h}}(x)\exp\left(\frac{\theta}{4}F(x)\right)$ are positive semidefinite; moreover, as $\mathbf{h}$ is nonzero in the $\ell_2$ space $\hat{\mathbf{h}}(x)$ is nonzero on a set of positive measure (i.e., $\hat{\mathbf{h}}(x)$ cannot be nonzero only on a set of measure zero). Therefore, since $\exp\left(\frac{\theta}{4}F(x)\right)$ is full rank, by Sylvester's inertia theorem [61] we have that $\exp\left(\frac{\theta}{4}F(x)\right)\hat{\mathbf{h}}(x)\exp\left(\frac{\theta}{4}F(x)\right)$ has at least one strictly positive eigenvalue for all $x$ in

a set of positive measure, meaning that the trace is strictly positive. Then

$$\left\langle E_T^{\theta/2}(\mathbf{1}\otimes \mathrm{vec}(I_d)), \mathbf{h}\right\rangle_\mu > 0.$$

□

Using this lemma, we are now able to give the proof of Lemma 5.4.

*Proof of Lemma 5.4* We first reduce to the simple case. The range of $F$ is compact, as $F$ can be seen as a continuous map of the compact set $O(d)\times[a,b]$. Then for $\varepsilon>0$ let $\mathscr{N}_\varepsilon$ be an $\varepsilon$-net of the range of $F$ with respect to the metric induced by the operator norm. Define $F_\varepsilon$ as $F_\varepsilon(x) = \arg\min_{B\in\mathscr{N}_\varepsilon}\|F(x)-B\|$. Then it is clear that for all $x\in\mathscr{X}$, $\|F_\varepsilon(x)-F(x)\|\le\varepsilon$. Define $T_\varepsilon(x) = \frac{1}{2}(F_\varepsilon(x)\otimes I_d + I_d\otimes F_\varepsilon(x))$ and $E_{T_\varepsilon}^{\theta/2}$ the corresponding block diagonal operator as usual. From this we see that $E_{T_\varepsilon}^{\theta}$ converges to $E_T^\theta$ in norm, and thus $E_{T_\varepsilon}^{\theta/2}\hat{P}E_{T_\varepsilon}^{\theta/2}$ converges to $E_T^{\theta/2}\hat{P}E_T^{\theta/2}$ in norm via any sequence of $\varepsilon$-nets with $\varepsilon\to 0$. Then $\lim_{\varepsilon\to 0}\log\left\|E_{T_\varepsilon}^{\theta/2}\hat{P}E_{T_\varepsilon}^{\theta/2}\right\|_\mu = \log\left\|E_T^{\theta/2}\hat{P}E_T^{\theta/2}\right\|_\mu$.

From Lemma 5.5 we have

$$\lim_{n\to\infty}\frac{1}{n}\log\left\langle \mathbf{1}\otimes\mathrm{vec}(I_d), E_{T_\varepsilon}^{\theta/2}(E_{T_\varepsilon}^{\theta/2}\hat{P}E_{T_\varepsilon}^{\theta/2})^{n-1}E_{T_\varepsilon}^{\theta/2}(\mathbf{1}\otimes\mathrm{vec}(I_d))\right\rangle_\mu = \log\left\|E_{T_\varepsilon}^{\theta/2}\hat{P}E_{T_\varepsilon}^{\theta/2}\right\|_\mu,$$

so that, denoting $a_{\varepsilon,n} = \frac{1}{n}\log\left\langle \mathbf{1}\otimes\mathrm{vec}(I_d), E_{T_\varepsilon}^{\theta/2}(E_{T_\varepsilon}^{\theta/2}\hat{P}E_{T_\varepsilon}^{\theta/2})^{n-1}E_{T_\varepsilon}^{\theta/2}(\mathbf{1}\otimes\mathrm{vec}(I_d))\right\rangle_\mu$,

$$\lim_{\varepsilon\to 0}\lim_{n\to\infty} a_{\varepsilon,n} = \lim_{\varepsilon\to 0}\log\left\|E_{T_\varepsilon}^{\theta/2}\hat{P}E_{T_\varepsilon}^{\theta/2}\right\|_\mu = \log\left\|E_T^{\theta/2}\hat{P}E_T^{\theta/2}\right\|_\mu.$$

In fact, we can give a more refined proof of part (3). of Lemma 5.5 to show that $\lim_{n\to\infty} a_{\varepsilon,n}$ uniformly converges in $\varepsilon$. Let $\rho_\varepsilon$ be the leading eigenvalue of $E_{T_\varepsilon}^{\theta/2}\hat{P}E_{T_\varepsilon}^{\theta/2}$. We first have by Lemma 5.1 that

$$\left\langle \mathbf{1}\otimes\mathrm{vec}(I_d), E_{T_\varepsilon}^{\theta/2}(E_{T_\varepsilon}^{\theta/2}\hat{P}E_{T_\varepsilon}^{\theta/2})^{n-1}E_{T_\varepsilon}^{\theta/2}(\mathbf{1}\otimes\mathrm{vec}(I_d))\right\rangle_\mu \le d\left\|E_{T_\varepsilon}^{\theta/2}\hat{P}E_{T_\varepsilon}^{\theta/2}\right\|_\mu^n = d\rho_\varepsilon^n,$$

so $a_{\varepsilon,n}-\log\rho_\varepsilon\le(\log d)/n$. Next, we showed for any unit norm eigenfunction $\mathbf{h}_\varepsilon$ corresponding to the leading eigenvalue $\rho_\varepsilon$ that

$$\left\langle \mathbf{1}\otimes\mathrm{vec}(I_d), E_{T_\varepsilon}^{\theta/2}(E_{T_\varepsilon}^{\theta/2}\hat{P}E_{T_\varepsilon}^{\theta/2})^{n-1}E_{T_\varepsilon}^{\theta/2}(\mathbf{1}\otimes\mathrm{vec}(I_d))\right\rangle_\mu \ge \left\langle E_{T_\varepsilon}^{\theta/2}(\mathbf{1}\otimes\mathrm{vec}(I_d)), \mathbf{h}_\varepsilon\right\rangle_\mu^2\rho_\varepsilon^{n-1},$$

so $a_{\varepsilon,m}-\log\rho_\varepsilon \ge \left(\log\left\langle E_{T_\varepsilon}^{\theta/2}(\mathbf{1}\otimes\mathrm{vec}(I_d)), \mathbf{h}\right\rangle_\mu^2\right)/n - (\log\rho_\varepsilon)/n$.

In the proof of part (3). of Lemma 5.5 we showed that $\mathbf{h}_\varepsilon$ can in fact be chosen such that for all $x\in\mathscr{X}$, $\mathbf{h}_\varepsilon(x)$ can be reshaped to a positive semidefinite $d\times d$ matrix. With such structure we lower bounded $\left\langle E_{T_\varepsilon}^{\theta/2}(\mathbf{1}\otimes\mathrm{vec}(I_d)), \mathbf{h}_\varepsilon\right\rangle_\mu > 0$. We can actually give a more quantitative lower bound of this quantity that is independent of $\varepsilon$; in addition, we can give an upper bound $\rho_\varepsilon$ also independent of $\varepsilon$.

First note that under an $\varepsilon$-net of the range of $F$, since $aI \preceq F(x) \preceq bI$, that $aI \preceq F_\varepsilon(x) \preceq bI$ also. Then the upper bound is simple; we have $\rho_\varepsilon = \left\| E_{T_\varepsilon}^{\theta/2} \hat{P} E_{T_\varepsilon}^{\theta/2} \right\|_\mu \leq e^{\theta b}$. For the lower bound we first write

$$\begin{aligned}\left\langle E_{T_\varepsilon}^{\theta/2}(\mathbf{1}\otimes \mathrm{vec}(I_d)), \mathbf{h}_\varepsilon \right\rangle_\mu &= \left\langle \mathrm{vec}(I_d), \mathbb{E}_\mu \left[ \exp\left(\frac{\theta}{2} T_\varepsilon(x)\right) \mathbf{h}_\varepsilon(x) \right] \right\rangle \\ &= \mathrm{tr}\, \mathbb{E}_\mu \left[ \exp\left(\frac{\theta}{4} F_\varepsilon(x)\right) \hat{\mathbf{h}}_\varepsilon(x) \exp\left(\frac{\theta}{4} F_\varepsilon(x)\right) \right].\end{aligned}$$

Since we have established $\exp\left(\frac{\theta}{2}F_\varepsilon(x)\right) \hat{\mathbf{h}}_\varepsilon(x) \exp\left(\frac{\theta}{2}F_\varepsilon(x)\right)$ is positive semidefinite for all $x \in \mathscr{X}$, the trace is equal to the nuclear norm, which upper bounds the Frobenius norm. Therefore,

$$\begin{aligned}\mathrm{tr}\, \mathbb{E}_\mu \left[ \exp\left(\frac{\theta}{4} F_\varepsilon(x)\right) \hat{\mathbf{h}}_\varepsilon(x) \exp\left(\frac{\theta}{4} F_\varepsilon(x)\right) \right] &= \mathbb{E}_\mu \left[ \left\| \exp\left(\frac{\theta}{4} F_\varepsilon(x)\right) \hat{\mathbf{h}}_\varepsilon(x) \exp\left(\frac{\theta}{4} F_\varepsilon(x)\right) \right\|_* \right] \\ &\geq \mathbb{E}_\mu \left[ \left\| \exp\left(\frac{\theta}{4} F_\varepsilon(x)\right) \hat{\mathbf{h}}_\varepsilon(x) \exp\left(\frac{\theta}{4} F_\varepsilon(x)\right) \right\|_F \right] \\ &\geq \mathbb{E}_\mu \left[ e^{\theta a/2} \left\| \hat{\mathbf{h}}_\varepsilon(x) \right\|_F \right] \\ &= e^{\theta a/2} \mathbb{E}_\mu \left[ \|\mathbf{h}_\varepsilon(x)\| \right].\end{aligned}$$

Now since $F_\varepsilon$ takes only finite number of values, we have that $\mathbf{h}_\varepsilon$ can also only take a finite number of values (see part (1) of Lemma 5.5). Letting $A_j = F^{-1}(B_j)$ and denoting $\mathbf{h}_j = \mathbf{h}_\varepsilon(x)$ for $x \in A_j$, we have

$$\mathbb{E}_\mu \left[ \|\mathbf{h}_\varepsilon(x)\| \right] = \sum_{j\in[m]} \mu(A_j) \left\| \mathbf{h}_j \right\|.$$

Now as $\|\mathbf{h}_\varepsilon\|_\mu = 1$ we must have $\left\|\mathbf{h}_j\right\| \leq 1$, and so then $\left\|\mathbf{h}_j\right\| \geq \left\|\mathbf{h}_j\right\|^2$. But $\sum_{j\in[m]} \mu(A_j) \left\|\mathbf{h}_j\right\|^2 = \|\mathbf{h}_\varepsilon\|_\mu^2$, and thus $\mathbb{E}_\mu[\|\mathbf{h}_\varepsilon(x)\|] \geq 1$.

Therefore, $\left\langle E_{T_\varepsilon}^{\theta/2}(\mathbf{1}\otimes \mathrm{vec}(I_d)), \mathbf{h} \right\rangle_\mu^2 \geq e^{\theta a}$, and we thus have $a_{\varepsilon,n} - \log \rho_\varepsilon \geq (\theta(a-b))/n$. Putting this all together, we have shown

$$|a_{\varepsilon,n} - \log \rho_\varepsilon| \leq \max\left\{ \frac{\log d}{n}, \frac{\theta(b-a)}{n} \right\},$$

which implies uniform convergence, since $d, \theta, a$, and $b$ are all constants independent of $\varepsilon$. We can then swap limits:

$$\lim_{\varepsilon\to 0} \lim_{n\to\infty} a_{\varepsilon,n} = \lim_{n\to\infty} \lim_{\varepsilon\to 0} a_{\varepsilon,n}.$$

But as $E_{T_\varepsilon}^\theta$ converges to $E_T^{\theta/2}$ in norm, so does $E_{T_\varepsilon}^{\theta/2}(E_{T_\varepsilon}^{\theta/2}\hat{P}E_{T_\varepsilon}^{\theta/2})^{n-1}E_{T_\varepsilon}^{\theta/2}$ to $E_T^{\theta/2}(E_T^{\theta/2}\hat{P}E_T^{\theta/2})^{n-1}E_T^{\theta/2}$, which implies weak convergence, that is,

$$\begin{aligned}&\lim_{\varepsilon\to 0} \left\langle \mathbf{1}\otimes \mathrm{vec}(I_d), E_{T_\varepsilon}^{\theta/2}(E_{T_\varepsilon}^{\theta/2}\hat{P}E_{T_\varepsilon}^{\theta/2})^{n-1}E_{T_\varepsilon}^{\theta/2}(\mathbf{1}\otimes \mathrm{vec}(I_d)) \right\rangle_\mu \\ &= \left\langle \mathbf{1}\otimes \mathrm{vec}(I_d), E_T^{\theta/2}(E_T^{\theta/2}\hat{P}E_T^{\theta/2})^{n-1}E_T^{\theta/2}(\mathbf{1}\otimes \mathrm{vec}(I_d)) \right\rangle_\mu.\end{aligned}$$

Then we have the chain of equalities

$$\begin{aligned}
&\lim_{n\to\infty}\frac{1}{n}\log\mathbb{E}_\mu\left[\left\|\prod_{j=1}^{n}\exp\left(\frac{\theta}{2}F(s_j)\right)\right\|_F^2\right]\\
&=\lim_{n\to\infty}\frac{1}{n}\log\left\langle \mathbf{1}\otimes\mathrm{vec}(I_d), E_T^{\theta/2}(E_T^{\theta/2}\hat{P}E_T^{\theta/2})^{n-1}E_T^{\theta/2}(\mathbf{1}\otimes\mathrm{vec}(I_d))\right\rangle_\mu\\
&=\lim_{n\to\infty}\lim_{\varepsilon\to 0}\frac{1}{n}\log\left\langle \mathbf{1}\otimes\mathrm{vec}(I_d), E_{T_\varepsilon}^{\theta/2}(E_{T_\varepsilon}^{\theta/2}\hat{P}E_{T_\varepsilon}^{\theta/2})^{n-1}E_{T_\varepsilon}^{\theta/2}(\mathbf{1}\otimes\mathrm{vec}(I_d))\right\rangle_\mu\\
&=\lim_{\varepsilon\to 0}\lim_{n\to\infty}\frac{1}{n}\log\left\langle \mathbf{1}\otimes\mathrm{vec}(I_d), E_{T_\varepsilon}^{\theta/2}(E_{T_\varepsilon}^{\theta/2}\hat{P}E_{T_\varepsilon}^{\theta/2})^{n-1}E_{T_\varepsilon}^{\theta/2}(\mathbf{1}\otimes\mathrm{vec}(I_d))\right\rangle_\mu\\
&=\lim_{\varepsilon\to 0}\log\left\|E_{T_\varepsilon}^{\theta/2}\hat{P}E_{T_\varepsilon}^{\theta/2}\right\|_\mu\\
&=\log\left\|E_T^{\theta/2}\hat{P}E_T^{\theta/2}\right\|_\mu.
\end{aligned}$$

□

## 6. Final bounds

In this section we prove our main theorems regarding the moment generating function and give concentration inequalities. We first show how to go from bounds on the moment generating function of Equation 4.1 to tail bounds.

### 6.1. *From the MGF to tail bounds*

The reason we have studied the moment generating function of Equation 4.1 is due to the following result:

**Theorem 6.1** (Multi-matrix Golden-Thompson inequality, [16]) *Let $H_1,\dots,H_k\in\mathbb{C}^{d\times d}$ be Hermitian matrices. Then*

$$\log\left(\mathrm{tr}\left[\exp\left(\sum_{j=1}^{k}H_j\right)\right]\right)\le\frac{4}{\pi}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\log\left(\mathrm{tr}\left[\prod_{j=1}^{k}\exp\left(\frac{e^{i\phi}}{2}H_j\right)\prod_{j=k}^{1}\exp\left(\frac{e^{-i\phi}}{2}H_j\right)\right]\right)d\xi(\phi),$$

*where $\xi$ is some probability distribution on $\left[-\frac{\pi}{2},\frac{\pi}{2}\right]$.*

From this, we can obtain tail bounds. Recall that $F_j : \mathscr{X} \to \mathbb{R}^{d\times d}$ are a sequence of measurable functions from $\mathscr{X}$ to $d \times d$ real symmetric matrices, and let $\mathbb{E}[F_j] = 0$. First,

$$\begin{aligned}\Pr_\mu\left(\lambda_{\max}\left(\sum_j F_j(s_j)\right) \geq t\right) &\leq e^{-\theta t}\mathbb{E}_\mu\left[\exp\left(\theta\lambda_{\max}\left(\sum_j F_j(s_j)\right)\right)\right] \\ &= e^{-\theta t}\mathbb{E}_\mu\left[\lambda_{\max}\left(\exp\left(\theta\sum_j F_j(s_j)\right)\right)\right] \\ &\leq e^{-\theta t}\mathbb{E}_\mu\left[\operatorname{tr}\exp\left(\theta\sum_j F_j(s_j)\right)\right].\end{aligned} \tag{6.1}$$

We would like to apply Theorem 6.1; we follow the standard outline given in [16]. An immediate application of Jensen's inequality on the right-hand side of Theorem 6.1 furnishes an upper bound, and then taking exponents, we have

$$\operatorname{tr}\left[\exp\left(\sum_{j=1}^n F_j(s_j)\right)\right] \leq \left(\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \operatorname{tr}\left[\prod_{j=1}^n \exp\left(\frac{e^{i\phi}}{2}F_j(s_j)\right)\prod_{j=n}^1 \exp\left(\frac{e^{-i\phi}}{2}F_j(s_j)\right)\right] d\xi(\phi)\right)^{4/\pi}.$$

Using $\|x\|_p \leq d^{1/p-1}\|x\|_1$ for $p \in (0,1)$ and choosing $p = \pi/4$ we have

$$\operatorname{tr}\left[\exp\left(\frac{\pi}{4}\sum_{j=1}^n F_j(s_j)\right)\right]^{4/\pi} \leq d^{4/\pi-1}\operatorname{tr}\left[\exp\left(\sum_{j=1}^n F_j(s_j)\right)\right].$$

Combining the above two lines, adding in $\theta$, and taking expectation, we have

$$\begin{aligned}&\mathbb{E}\left[\operatorname{tr}\exp\left(\frac{\pi}{4}\theta\sum_{j=1}^n F_j(s_j)\right)\right] \\ \leq& d^{1-\pi/4}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \mathbb{E}\left[\operatorname{tr}\left[\prod_{j=1}^n \exp\left(\frac{\theta e^{i\phi}}{2}F_j(s_j)\right)\prod_{j=n}^1 \exp\left(\frac{\theta e^{-i\phi}}{2}F_j(s_j)\right)\right]\right] d\xi(\phi) \\ =& d^{1-\pi/4}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \mathbb{E}\left[\left\|\prod_{j=1}^n \exp\left(\frac{\theta e^{i\phi}}{2}\right)F_j(s_j)\right\|_F^2\right] d\xi(\phi).\end{aligned} \tag{6.2}$$

Now let $M_F(\theta)$ be such that

$$M_F(\theta) \geq \mathbb{E}\left[\left\|\prod_{j=1}^n \exp\left(\frac{\theta e^{i\phi}}{2}\right)F_j(s_j)\right\|_F^2\right] \tag{6.3}$$

that is independent of $\phi$, i.e., an upper bound on the moment generating function in Equation 4.1 valid for any $\phi \in [-\pi/2, \pi/2]$. Then

$$\begin{aligned}\Pr_\mu\left(\lambda_{\max}\left(\sum_j F_j(s_j)\right) \geq t\right) &\leq e^{-\theta t}\mathbb{E}_\mu\left[\operatorname{tr}\exp\left(\theta\sum_j F_j(s_j)\right)\right] \\ &\leq d^{1-\pi/4}e^{-\theta t}\int_{\frac{-\pi}{2}}^{\frac{\pi}{2}} M_F\left(\frac{4}{\pi}\theta\right)d\xi(\phi) \\ &= d^{1-\pi/4}\exp\left(-\theta t + \log M_F\left(\frac{4}{\pi}\theta\right)\right),\end{aligned} \tag{6.4}$$

where the last line holds because $\xi$ is a probability distribution and $M_F$ does not depend on $\phi$.

We now give our final bounds on the moment generating function

$$\mathbb{E}_\mu\left[\left\|\prod_{j=1}^n \exp\left(\frac{\theta e^{i\phi}}{2}\right)F_j(s_j)\right\|_F^2\right]$$

under both Hoeffding and Bernstein-type assumptions. In Sections 4 and 5 we have reduced the problem to estimating the leading eigenvalues of the operators

$$E_{T_j}^{\theta/2}\hat{P}E_{T_j}^{\theta/2} \tag{6.5}$$

for $j = 1,\ldots,n$ and $\hat{P}$ the lifted Leon-Perron version of $P$. Lemma 5.4 has shown that each eigenvalue is a limit a corresponding moment generating function. Thus, the general idea will be to bound

$$\mathbb{E}_\mu\left[\left\|\prod_{k=1}^n \exp\left(\frac{\theta\cos(\phi)}{2}F_j(s_k)\right)\right\|_F^2\right]$$

for each $j$, and where $s_1,\ldots,s_n$ is a sequence of states driven by the Markov chain corresponding to $\hat{P}$. In these bounds we will proceed by way of defining a function $M_F$ to bound these operators.

### 6.2. *Hoeffding*

In this section, fix a matrix-valued function $F$, and let $\mathbb{E}_\mu[F(x)] = 0$ and $aI \preceq F(x) \preceq bI$ for all $x \in \mathscr{X}$. For a scalar convex functions $\Psi$, we have the following proposition:

**Proposition 6.2** *Let $H$ be a Hermitian $d \times d$ matrix such that $aI \preceq H \preceq bI$. Then for any scalar convex function $\Psi$,*

$$\Psi(H) \preceq \frac{bI - H}{b-a}\Psi(a) + \frac{H - aI}{b-a}\Psi(b).$$

*Here $\Psi$ acts spectrally; if $H = U\Lambda U^*$, then $\Psi(H) = U\Psi(\Lambda)U^*$, where $\Psi$ acts on the diagonal matrix $\Lambda$ diagonally, i.e., $\Psi(\Lambda)$ is the diagonal matrix such that $\Psi(\Lambda)_{ii} = \Psi(\Lambda_{ii})$.*

*Proof* We simply need to show that for any unit vector $\mathbf{u}$ that

$$\mathbf{u}^*\Psi(H)\mathbf{u} \leq \frac{b-\mathbf{u}^*H\mathbf{u}}{b-a}\Psi(a) + \frac{\mathbf{u}^*H\mathbf{u}-aI}{b-a}\Psi(b).$$

It suffices to show this for eigenvectors $\mathbf{v}$ of $H$. If $\lambda$ is the corresponding eigenvalue, then $\mathbf{v}^*\Psi(H)\mathbf{v} = \Psi(\lambda)$. The right-hand side is

$$\frac{b-\lambda}{b-a}\Psi(a) + \frac{\lambda-a}{b-a}\Psi(b).$$

Since $\Psi$ is a scalar convex function and $a \leq \lambda \leq b$,

$$\Psi(\lambda) \leq \frac{b-\lambda}{b-a}\Psi(a) + \frac{\lambda-a}{b-a}\Psi(b).$$

For general $\mathbf{u}$, let $\mathbf{u} = \sum_{k=1}^d c_k\mathbf{v}_k$ be the representation of $\mathbf{u}$ in the basis of eigenvectors of $H$. Then

$$\begin{aligned}
\mathbf{u}^*\Psi(H)\mathbf{u} &= \sum_{k=1}^d |c_k|^2\Psi(\lambda_k) \\
&\leq \sum_{k=1}^d |c_k|^2\left(\frac{b-\lambda_k}{b-a}\Psi(a) + \frac{\lambda_k-a}{b-a}\Psi(b)\right) \\
&= \frac{b-\mathbf{u}^*H\mathbf{u}}{b-a}\Psi(a) + \frac{\mathbf{u}^*H\mathbf{u}-a}{b-a}\Psi(b).
\end{aligned}$$

□

If in the above $H$ is a random matrix such that $\mathbb{E}[H] = 0$, then taking expectations we have

$$\mathbb{E}[H] \preceq \frac{b}{b-a}\Psi(aI) + \frac{-a}{b-a}\Psi(bI),$$

the right-hand side of which defines an implicit two-point distribution on matrices. Introduce some notation; let $\mu = [b/(b-a), -a/(b-a)]^\top$ and $Q = \lambda I + (1-\lambda)\mathbf{1}\mu^\top$. Then $Q$ is a transition matrix on two states $\{0,1\}$ with stationary distribution $\mu$. Let $G$ be a function such that $G(0) = aI_d$ and $G(1) = bI_d$.

This allows us to make the following argument about a convex ordering of distributions, with proof in the appendix:

**Lemma 6.3** *Let $F$ be a function from $\mathscr{X}$ to Hermitian $d \times d$ matrices, and assume $\mathbb{E}_\mu[F] = 0$ and $aI \preceq F(x) \preceq bI$ for all $x \in \mathscr{X}$. Let $s_1, \ldots, s_n$ be driven by the Leon-Perron operator $\hat{P}$. Let $Q, \mu, G$ be as above. Let $y_1, \ldots, y_n$ be driven by $Q$. Then for any scalar nonnegative convex function $\Psi$,*

$$\mathbb{E}_\mu\left[\left\|\prod_{j=1}^n \Psi(F(s_j))\right\|_F^2\right] \leq \mathbb{E}_\mu\left[\left\|\prod_{j=1}^n \Psi(G(y_j))\right\|_F^2\right].$$

**Remark 6.4** *The condition on $\Psi$ above does not seem to be very strict. Indeed, it applies to many natural matrix-valued functions, such as even matrix powers and the matrix exponential.*

Applying this to the scalar convex function $x \mapsto \exp\left(\frac{\theta\cos(\phi)}{2}x\right)$ yields

$$\mathbb{E}_\mu\left[\left\|\prod_{j=1}^{n}\exp\left(\frac{\theta\cos(\phi)}{2}F(s_j)\right)\right\|_F^2\right] \le \mathbb{E}_\mu\left[\left\|\prod_{j=1}^{n}\exp\left(\frac{\theta\cos(\phi)}{2}G(y_j)\right)\right\|_F^2\right]. \tag{6.6}$$

Now let $M^\theta \in \mathbb{R}^{2d^2\times 2d^2}$ be the operator on the two-state state space $\{0,1\}$ such that $(M^\theta \mathbf{h})(0) = \exp(\theta\cos(\phi)a)I_{d^2}$ and $(M^\theta\mathbf{h})(1) = \exp(\theta\cos(\phi)b)I_{d^2}$, i.e.,

$$M^\theta = \begin{bmatrix}\exp(\theta\cos(\phi)a)I_{d^2} & 0\\ 0 & \exp(\theta\cos(\phi)b)I_{d^2}\end{bmatrix}.$$

Then Lemma 5.4 indicates that

$$\lim_{n\to\infty}\frac{1}{n}\log\mathbb{E}_\mu\left[\left\|\prod_{j=1}^{n}\exp\left(\frac{\theta\cos(\phi)}{2}G(y_j)\right)\right\|_F^2\right] = \log\eta, \tag{6.7}$$

where $\eta$ is the largest eigenvalue of $M^{\theta/2}\hat{Q}M^{\theta/2}$, $\hat{Q} = Q\otimes I_{d^2}$. Now note that

$$\begin{aligned} &M^{\theta/2}\hat{Q}M^{\theta/2}\\ &= \begin{bmatrix}\exp\left(\frac{\theta\cos(\phi)}{2}a\right) & 0\\ 0 & \exp\left(\frac{\theta\cos(\phi)}{2}b\right)\end{bmatrix} Q \begin{bmatrix}\exp\left(\frac{\theta\cos(\phi)}{2}a\right) & 0\\ 0 & \exp\left(\frac{\theta\cos(\phi)}{2}b\right)\end{bmatrix}\otimes I_{d^2} \qquad (6.8)\\ &=: K^\theta\otimes I_{d^2}. \end{aligned}$$

Then by properties of the Kronecker product, the eigenvalues of the matrix on the left-hand side are equal to the eigenvalues of $K^\theta$, a $2\times 2$ matrix, the largest of which can be solved for explicitly.

Putting this all together,

$$\begin{aligned} \log\rho &= \lim_{n\to\infty}\frac{1}{n}\log\mathbb{E}_\mu\left[\left\|\prod_{j=1}^{n}\exp\left(\frac{\theta\cos(\phi)}{2}F(s_j)\right)\right\|_F^2\right]\\ &\le \lim_{n\to\infty}\frac{1}{n}\log\mathbb{E}_\mu\left[\left\|\prod_{j=1}^{n}\exp\left(\frac{\theta\cos(\phi)}{2}G(y_j)\right)\right\|_F^2\right] \qquad (6.9)\\ &= \log\eta\\ &= \log\lambda_{\max}(K^\theta). \end{aligned}$$

With the above we can establish the following result, with proof in the appendix.

**Lemma 6.5** *Let $K^\theta$ be the $2\times 2$ matrix defined in Equation 6.8, and let $\eta$ be its largest eigenvalue. Then*

$$\eta \le \tilde{\eta}(\theta) := \exp\left(\alpha(\lambda)\cdot\frac{\theta^2\cos^2(\phi)(b-a)^2}{8}\right)$$

*where $\alpha : \lambda \mapsto (1+\lambda)/(1-\lambda)$.*

Putting our work above together, we can now prove the two statements of Theorem 2.5.

*Proof of Theorem 2.5* The discussion in Lemma 5.1 (in particular Equation 5.2) and Proposition 5.2 reveal

$$\mathbb{E}_\mu\left[\left\|\prod_{j=1}^{n}\exp\left(\frac{\theta e^{i\phi}}{2}F_j(s_j)\right)\right\|_F^2\right] \leq d\prod_{j=1}^{n}\left\|E_{T_j}^{\theta/2}\hat{P}E_{T_j}^{\theta/2}\right\|_\mu, \tag{6.10}$$

where recall $T_j(x) = \frac{\cos(\phi)}{2}(F_j(x)\otimes I_d + I_d \otimes F_j(x))$, $E_{T_j}^{\theta/2}$ is the multiplication operator with respect to the matrix-valued function $T_j$ and $\hat{P} = (\lambda I + (1-\lambda)\Pi)\otimes I_{d^2}$. Then Lemma 5.4 and Lemma 6.3 allow us to bound each of these norms with the norms of corresponding matrices arising from a two state chain, each of which can be bounded using Lemma 6.5. In other words, for each $j$ we have, verifying that the operators $T_j$ satisfy the assumptions of the lemmas using Proposition 5.3,

$$\left\|E_{T_j}^{\theta/2}\hat{P}E_{T_j}^{\theta/2}\right\|_\mu \leq \exp\left(\alpha(\lambda)\cdot\frac{\theta^2\cos(\phi)^2(b_j-a_j)^2}{8}\right). \tag{6.11}$$

Then

$$\mathbb{E}_\mu\left[\left\|\prod_{j=1}^{n}\exp\left(\frac{\theta e^{i\phi}}{2}F_j(s_j)\right)\right\|_F^2\right] \leq d\exp\left(\theta^2\cdot\alpha(\lambda)\cdot\frac{\cos(\phi)^2\sum_{j=1}^{n}(b_j-a_j)^2}{8}\right), \tag{6.12}$$

and setting $\phi = 0$ we obtain the first statement of the theorem.

To turn this into a tail bound, since $\cos(\phi)^2 \leq 1$, we can let $M_F(\theta)$, as described in Equation 6.3, be

$$M_F(\theta) := d\exp\left(\theta^2\cdot\alpha(\lambda)\cdot\frac{\sum_{j=1}^{n}(b_j-a_j)^2}{8}\right). \tag{6.13}$$

Using Equation 6.4, we then have

$$\begin{aligned}\Pr_\mu\left(\lambda_{\max}\left(\sum_{j=1}^{n}F_j(s_j)\right)\geq t\right) &\leq \inf_{\theta>0} d^{1-\pi/4}\exp\left(-\theta\cdot t + \log M_F\left(\frac{4}{\pi}\theta\right)\right)\\ &= \inf_{\theta>0} d^{2-\pi/4}\exp\left(-\theta\cdot t + \theta^2\cdot\alpha(\lambda)\cdot\frac{2\sum_{j=1}^{n}(b_j-a_j)^2}{\pi^2}\right).\end{aligned} \tag{6.14}$$

Optimizing this quadratic over $\theta$ we have

$$\Pr_\mu\left(\lambda_{\max}\left(\sum_{j=1}^{n}F_j(s_j)\right)\geq t\right) \leq d^{2-\pi/4}\exp\left(\frac{-t^2/(8/\pi^2)}{\alpha(\lambda)\cdot\sum_{j=1}^{n}(b_j-a_j)^2}\right), \tag{6.15}$$

as desired. □

6.3. *Bernstein*

In this section, fix a matrix-valued function $F$, and let $\mathbb{E}_\mu[F(x)] = 0$, $\left\|\mathbb{E}_\mu[F(x)^2]\right\| \le \mathscr{V}$, and $\|F(x)\| \le \mathscr{M}$ for all $x \in \mathscr{X}$.

We take a slightly different approach to bounding the norms of the operators $E_{T_j}^{\theta/2}\hat{P}E_{T_j}^{\theta/2}$. In particular, consider again Equation 4.1 for a time-independent function $T$ and the Leon-Perron operator $\hat{P}$. Then

$$\begin{aligned}
&\mathbb{E}_\mu\left[\left\|\prod_{j=1}^{n}\exp\left(\frac{\theta\cos(\phi)}{2}\right)F(s_j)\right\|_F^2\right] \\
&= \left\langle \mathbf{1}\otimes\mathrm{vec}(I_d), E_T^{\theta/2}(E_T^{\theta/2}\hat{P}E_T^{\theta/2})^{n-1}E_T^{\theta/2}(\mathbf{1}\otimes\mathrm{vec}(I_d))\right\rangle_\mu \\
&= \left\langle \tilde{\Pi}(\mathbf{1}\otimes\mathrm{vec}(I_d)), E_T^{\theta/2}(E_T^{\theta/2}\hat{P}E_T^{\theta/2})^{n-1}E_T^{\theta/2}\tilde{\Pi}(\mathbf{1}\otimes\mathrm{vec}(I_d))\right\rangle_\mu \\
&= \left\langle \mathbf{1}\otimes\mathrm{vec}(I_d), \tilde{\Pi}E_T^{\theta/2}(E_T^{\theta/2}\hat{P}E_T^{\theta/2})^{n-1}E_T^{\theta/2}\tilde{\Pi}(\mathbf{1}\otimes\mathrm{vec}(I_d))\right\rangle_\mu,
\end{aligned} \tag{6.16}$$

where recall $\tilde{\Pi} = \Pi\otimes I_{d^2}$ and $T(x) = \frac{\cos(\phi)}{2}(F(x)\otimes I_{d^2} + I_{d^2}\otimes F(x))$. The third line holds as $\tilde{\Pi}(\mathbf{1}\otimes\mathrm{vec}(I_d)) = \mathbf{1}\otimes\mathrm{vec}(I_d)$ and the last line holds as $\tilde{\Pi}$ is self-adjoint with respect to the inner product. Next,

$$\begin{aligned}
&\left\langle \mathbf{1}\otimes\mathrm{vec}(I_d), \tilde{\Pi}E_T^{\theta/2}(E_T^{\theta/2}\hat{P}E_T^{\theta/2})^{n-1}E_T^{\theta/2}\tilde{\Pi}(\mathbf{1}\otimes\mathrm{vec}(I_d))\right\rangle_\mu \\
&\le \left\|\tilde{\Pi}E_T^{\theta/2}(E_T^{\theta/2}\hat{P}E_T^{\theta/2})^{n-1}E_T^{\theta/2}\tilde{\Pi}\right\|_\mu \|\mathbf{1}\otimes\mathrm{vec}(I_d)\|_\mu^2 \\
&= d\left\|\tilde{\Pi}E_T^{\theta/2}(E_T^{\theta/2}\hat{P}E_T^{\theta/2})^{n-1}E_T^{\theta/2}\tilde{\Pi}\right\|_\mu \\
&= d\left\|\tilde{\Pi}(\hat{P}E_T^{\theta})^{n}\tilde{\Pi}\right\|_\mu,
\end{aligned} \tag{6.17}$$

with the last line following from $\tilde{\Pi}\hat{P} = \tilde{\Pi}$ so we can introduce another $\hat{P}$. Now from Lemma 5.4 and the above we have

$$\begin{aligned}
\log\left\|E_T^{\theta/2}\hat{P}E_T^{\theta/2}\right\|_\mu &= \lim_{n\to\infty}\frac{1}{n}\log\left\langle \mathbf{1}\otimes\mathrm{vec}(I_d), E_T^{\theta/2}(E_T^{\theta/2}\hat{P}E_T^{\theta/2})^{n-1}E_T^{\theta/2}(\mathbf{1}\otimes\mathrm{vec}(I_d))\right\rangle_\mu \\
&\le \lim_{n\to\infty}\frac{1}{n}\log d\left\|\tilde{\Pi}(\hat{P}E_T^{\theta})^{n}\tilde{\Pi}\right\|_\mu \\
&= \lim_{n\to\infty}\frac{1}{n}\log\left\|\tilde{\Pi}(\hat{P}E_T^{\theta})^{n}\tilde{\Pi}\right\|_\mu.
\end{aligned} \tag{6.18}$$

Although the above may seem a bit contrived, it allows for a general recursive method used in [16, 26, 56, 68]; more importantly, we can extend the final result to time-dependent functions via our limit lemma. In particular, for an initial function $\mathbf{z}_0$, we can track the updates $\mathbf{z}_k := (\hat{P}E_T^{\theta})^k\mathbf{z}_0$ recursively, and then provide a bound. This method is strong enough to give the sorts of inequalities we are after.

Now we introduce some notation. For an element $\mathbf{z}\in\ell_2(\mu\otimes\mathbf{1})$ let $\mathbf{z}^{\|} = \tilde{\Pi}\mathbf{z}$ and $\mathbf{z}^{\perp} = (I-\tilde{\Pi})\mathbf{z}$. In other words, $\mathbf{z}$ decomposes as $\mathbf{z} = \left\|\mathbf{z}^{\|}\right\|_\mu(\mathbf{1}\otimes\mathbf{v}) + \left\|\mathbf{z}^{\perp}\right\|_\mu\rho$ for some $\mathbf{v}$ and some $\rho$ orthogonal to $\mathbf{1}\otimes\mathbb{C}^{d^2}$.

As per our discussion above, we will bound the evolution of $\tilde{P}E_T^{\theta}\mathbf{z}$ for any vector $\mathbf{z}$. We first have the following proposition:

**Proposition 6.6** *For any* $\mathbf{z} \in \ell_2(\mu \otimes \mathbf{1})$,

*(1)* $\hat{P}\mathbf{z}^{\|} = \mathbf{z}^{\|}$,
*(2)* $(\hat{P}\mathbf{z})^{\perp} = \tilde{P}\mathbf{z}^{\perp}$.

*Proof* This follows immediately from the definitions. □

We next have a crucial lemma, the proof deferred to the appendix. For ease of exposition, we will assume that $\phi = 0$ throughout, since the factor of $\cos(\phi)$ is just a scalar and can be pulled through.

**Lemma 6.7** *For any* $\mathbf{z} \in \ell_2(\mu \otimes \mathbf{1})$,

*(1)* $\left\|(E_T^{\theta}\mathbf{z}^{\|})^{\|}\right\|_{\mu} \leq \alpha_1$,
*(2)* $\left\|(E_T^{\theta}\mathbf{z}^{\perp})^{\|}\right\|_{\mu} \leq \alpha_2$,
*(3)* $\left\|(E_T^{\theta}\mathbf{z}^{\|})^{\perp}\right\|_{\mu} \leq \alpha_2$,
*(4)* $\left\|(E_T^{\theta}\mathbf{z}^{\perp})^{\perp}\right\|_{\mu} \leq \alpha_3$,

*where*

$$\alpha_1 = 1 + \frac{\mathscr{V}(e^{\mathscr{M}\theta} - \mathscr{M} - \theta - 1))}{\mathscr{M}^2}, \alpha_2 = \frac{\sqrt{V}(e^{\mathscr{M}\theta} - 1)}{\mathscr{M}}, \alpha_3 = e^{\mathscr{M}\theta}.$$

With this lemma we have the following two claims, with proofs in the appendix. These are recursive bounds.

**Claim 6.8** *For any k, if* $\theta < \log(1/\lambda)/\mathscr{M}$, *then*

$$\left\|\mathbf{z}_k^{\perp}\right\|_{\mu} \leq \frac{\lambda\alpha_2}{1-\lambda\alpha_3} \max_{0\leq j\leq k} \left\|\mathbf{z}_j^{\|}\right\|_{\mu}.$$

**Claim 6.9** *For any k, if* $\theta < \log(1/\lambda)/\mathscr{M}$, *then*

$$\left\|\mathbf{z}_k^{\|}\right\|_{\mu} \leq \left(\alpha_1 + \frac{\lambda\alpha_2^2}{1-\lambda\alpha_3}\right) \max_{0\leq j\leq k} \left\|\mathbf{z}_j^{\|}\right\|_{\mu}.$$

We are now able to prove Theorem 2.6.

*Proof of Theorem 2.6* As in the proof of Theorem 2.5, Lemma 5.1 and Proposition 5.2 give

$$\mathbb{E}_\mu \left[ \left\| \prod_{j=1}^{n} \exp\left( \frac{\theta e^{i\phi}}{2} F_j(s_j) \right) \right\|_F^2 \right] \le d \prod_{j=1}^{n} \left\| E_{T_j}^{\theta/2} \hat{P} E_{T_j}^{\theta/2} \right\|_\mu . \tag{6.19}$$

Assume throughout that $\phi = 0$ for simplicity, so we can directly use Lemma 6.7 without modification. Equation 6.18 indicates that

$$\log \left\| E_{T_j}^{\theta/2} \hat{P} E_{T_j}^{\theta/2} \right\|_\mu \le \lim_{n\to\infty} \frac{1}{n} \log \left\| \tilde{\Pi} (\hat{P} E_{T_j}^{\theta})^n \tilde{\Pi} \right\|_\mu .$$

Claim 6.9 exactly gives a bound on this term; letting $\mathbf{z}_0 \in \ell_2(\mu \otimes \mathbf{1})$, the claim indicates that

$$\left\| \tilde{\Pi} (\hat{P} E_{T_j})^n \tilde{\Pi} \mathbf{z}_0 \right\|_\mu \le \left( \alpha_{j,1} + \frac{\lambda \alpha_{j,2}^2}{1 - \lambda \alpha_{j,3}} \right)^n \left\| \tilde{\Pi} \mathbf{z}_0 \right\|_\mu \le \left( \alpha_{j,1} + \frac{\lambda \alpha_{j,2}^2}{1 - \lambda \alpha_{j,3}} \right)^n \left\| \mathbf{z}_0 \right\|_\mu , \tag{6.20}$$

with $\alpha_{j,1}, \alpha_{j,2}, \alpha_{j,3}$ as in Lemma 6.7 and using $\mathscr{V}_j$ in the definitions of the $\alpha_j$ as we are dealing with time-dependent functions. Then $\left\| \tilde{\Pi} (\hat{P} E_{T_j}^{\theta})^n \tilde{\Pi} \right\|_\mu \le \left( \alpha_{j,1} + \frac{\lambda \alpha_{j,2}^2}{1-\lambda \alpha_{j,3}} \right)^n$. Now

$$\begin{aligned} \left( \alpha_{j,1} + \frac{\lambda \alpha_{j,2}^2}{1 - \lambda \alpha_{j,3}} \right)^n &= \left( 1 + \frac{1}{\mathscr{M}^2} \left( \mathscr{V}_j (e^{\mathscr{M}\theta} - \mathscr{M}\theta - 1) + \frac{\lambda \mathscr{V}_j (e^{\mathscr{M}\theta} - 1)^2}{1 - \lambda e^{\mathscr{M}\theta}} \right) \right)^n \\ &\le \exp\left( \frac{n \mathscr{V}_j}{\mathscr{M}^2} \left( e^{\mathscr{M}\theta} - \mathscr{M}\theta - 1 + \frac{\lambda (e^{\mathscr{M}\theta} - 1)^2}{1 - \lambda e^{\mathscr{M}\theta}} \right) \right), \end{aligned} \tag{6.21}$$

implying

$$\begin{aligned} \log \left\| E_{T_j}^{\theta/2} \hat{P} E_{T_j}^{\theta/2} \right\|_\mu &\le \lim_{n\to\infty} \frac{1}{n} \log \left\| \tilde{\Pi} (\hat{P} E_{T_j}^{\theta})^n \tilde{\Pi} \right\|_\mu \\ &\le \lim_{n\to\infty} \frac{1}{n} \log \exp\left( \frac{n \mathscr{V}_j}{\mathscr{M}^2} \left( e^{\mathscr{M}\theta} - \mathscr{M}\theta - 1 + \frac{\lambda (e^{\mathscr{M}\theta} - 1)^2}{1 - \lambda e^{\mathscr{M}\theta}} \right) \right) \\ &= \frac{\mathscr{V}_j}{\mathscr{M}^2} \left( e^{\mathscr{M}\theta} - \mathscr{M}\theta - 1 + \frac{\lambda (e^{\mathscr{M}\theta} - 1)^2}{1 - \lambda e^{\mathscr{M}\theta}} \right), \end{aligned} \tag{6.22}$$

and thus

$$\begin{aligned} \mathbb{E}_\mu \left[ \left\| \prod_{j=1}^{n} \exp\left( \frac{\theta}{2} F_j(s_j) \right) \right\|_F^2 \right] &\le d \exp\left( \frac{1}{\mathscr{M}^2} \sum_{j=1}^{n} \mathscr{V}_j \left( e^{\mathscr{M}\theta} - \mathscr{M}\theta - 1 + \frac{\lambda (e^{\mathscr{M}\theta} - 1)^2}{1 - \lambda e^{\mathscr{M}\theta}} \right) \right) \\ &= d \exp\left( \frac{\sigma^2}{\mathscr{M}^2} \left( e^{\mathscr{M}\theta} - \mathscr{M}\theta - 1 + \frac{\lambda (e^{\mathscr{M}\theta} - 1)^2}{1 - \lambda e^{\mathscr{M}\theta}} \right) \right) \end{aligned} \tag{6.23}$$

if $\theta < \log(1/\lambda)/\mathscr{M}$, as desired.

Let

$$M_F(\theta) := d\exp\left(\frac{\sigma^2}{\mathscr{M}^2}\left(e^{\mathscr{M}\theta} - \mathscr{M}\theta - 1 + \frac{\lambda(e^{\mathscr{M}\theta}-1)^2}{1-\lambda e^{\mathscr{M}\theta}}\right)\right). \tag{6.24}$$

To turn this into a tail bound, we will first assume $\mathscr{M} = 1$, i.e., $\|F_j\| \leq 1$ for all $j$, and then drop this assumption later. Using Equation 6.4 we have

$$\Pr_\mu\left(\lambda_{\max}\left(\sum_{j=1}^n F_j(s_j)\right) \geq t\right) \leq \inf_{\theta>0} d^{1-\pi/4}\exp\left(-\theta\cdot t + \log M_F\left(\frac{4}{\pi}\theta\right)\right). \tag{6.25}$$

Then

$$\begin{aligned} M_F\left(\frac{4}{\pi}\theta\right) &= d\exp\left(\sigma^2\left(e^{4\theta/\pi} - 4\theta/\pi - 1 + \frac{\lambda(e^{4\theta/\pi}-1)^2}{1-\lambda e^{4\theta/\pi}}\right)\right) \\ &= d\exp\left(\sigma^2\left(e^{4\theta/\pi} - 4\theta/\pi - 1 + \frac{\lambda(e^{4\theta/\pi}-1)^2}{1-\lambda e^{4\theta/\pi}}\right)\right) \\ &= d\exp\left(\frac{16}{\pi^2}\left(\frac{\sigma^2}{16/\pi^2}\left(e^{4\theta/\pi} - 4\theta/\pi - 1\right) + \frac{\sigma^2}{16/\pi^2}\cdot\frac{\lambda(e^{4\theta/\pi}-1)^2}{1-\lambda e^{4\theta/\pi}}\right)\right). \end{aligned} \tag{6.26}$$

Write $L = 4/\pi$. Using the inequality $e^x - 1 \leq xe^x$, the above becomes

$$\begin{aligned} &d\exp\left(L^2\left(\frac{\sigma^2}{L^2}\left(e^{L\theta} - L\theta - 1\right) + \frac{\sigma^2}{L^2}\cdot\frac{\lambda(e^{L\theta}-1)^2}{1-\lambda e^{L\theta}}\right)\right) \\ \leq &d\exp\left(L^2\left(\frac{\sigma^2}{L^2}\left(e^{L\theta} - L\theta - 1\right) + \frac{\sigma^2}{L^2}\cdot\frac{\lambda\theta^2 e^{2L\theta}}{1-\lambda e^{L\theta}}\right)\right) \\ = &d\exp\left(\sigma^2\left(e^{L\theta} - L\theta - 1\right) + \frac{\sigma^2\lambda\theta^2 e^{2L\theta}}{1-\lambda e^{L\theta}}\right). \end{aligned} \tag{6.27}$$

Now we argue that

$$\frac{\sigma^2\lambda\theta^2 e^{2L\theta}}{1-\lambda e^{L\theta}} \leq \frac{\sigma^2\lambda\theta^2}{1-\lambda-2L\theta}$$

when $0 \leq \theta < (1-\lambda)/2L$, which is less than $\log(1/\lambda)$ when $0 < \lambda < 1$. Indeed, comparing both sides, it suffices to show

$$\frac{e^{2L\theta}}{1-\lambda e^{L\theta}} \leq \frac{1}{1-\lambda-2L\theta}$$

for this range of $\theta$, which is equivalent to showing $e^{-2L\theta} - \lambda e^{-L\theta} \geq 1-\lambda-2L\theta$. At $\theta = 0$ they are equivalent, so we compare derivatives. The derivative of the left-hand side is $\lambda L e^{-L\theta} - 2Le^{-2L\theta} = Le^{-L\theta}(\lambda - 2e^{-L\theta})$, while the derivative of the right-hand side is just $-2L$. Then

$$\begin{aligned} &Le^{-L\theta}(\lambda - 2e^{-L\theta}) \geq -2L \\ \iff &\frac{e^{-L\theta}(2e^{-L\theta} - \lambda)}{2} \leq 1, \end{aligned}$$

which is true by seeing that the left-hand side is a decreasing function in $\theta$, so for $\theta \geq 0$ attains its maximum at $\theta = 0$. At this value of $\theta$, the left-hand side is $1 - \lambda/2 < 1$.

Then

$$\begin{aligned} &d\exp\left(\sigma^2\left(e^{L\theta}-L\theta-1\right)+\frac{\sigma^2\lambda\theta^2e^{2L\theta}}{1-\lambda e^{L\theta}}\right)\\ \le\, &d\exp\left(\sigma^2\left(e^{L\theta}-L\theta-1\right)+\frac{\sigma^2\lambda\theta^2}{1-\lambda-2L\theta}\right). \end{aligned} \tag{6.28}$$

To recap, at this point we have bounded

$$M_F(L\theta)\le d\exp\left(\sigma^2\left(e^{L\theta}-L\theta-1\right)+\frac{\sigma^2\lambda\theta^2}{1-\lambda-2L\theta}\right).$$

Then

$$\begin{aligned} \Pr_\mu\left(\lambda_{\max}\left(\sum_{j=1}^n F_j(s_j)\right)\ge t\right) &\le \inf_{\theta>0} d^{1-\pi/4}\exp\left(-\theta t+\log M_F\left(\frac{4}{\pi}\theta\right)\right)\\ &\le d^{2-\pi/4}\inf_{\theta>0}\exp\left(-\theta t+g(\theta)\right) \end{aligned}$$

where

$$g(\theta):=\sigma^2\left(e^{L\theta}-L\theta-1\right)+\frac{\sigma^2\lambda L^2\theta^2}{1-\lambda-2L\theta}.$$

It is standard that the conjuguate $g^*(t)=\sup_\theta\{\theta t-g(\theta)\}$ bounds the tail probability via

$$\Pr\left(\lambda_{\max}\left(\sum_{j=1}^n F_j(s_j)\right)\ge t\right)\le d\exp(g^*(t)) \tag{6.29}$$

for a convex function $g$ [22]. Using Lemma D.2 with $L=4/\pi$, the second statement in Theorem 2.6 is immediate in the case $\mathscr{M}=1$.

More generally, we have

$$\Pr\left(\lambda_{\max}\left(\sum_{j=1}^n F_j(s_j)\right)\ge t\right)=\Pr\left(\lambda_{\max}\left(\sum_{j=1}^n\frac{F_j(s_j)}{\mathscr{M}}\right)\ge\frac{t}{\mathscr{M}}\right).$$

Now $\mathbb{E}_\mu[F_j(x)/\mathscr{M}]=0$, $\|F_j\|\le 1$, and $\left\|\mathbb{E}_\mu[F_j(x)^2/\mathscr{M}^2]\right\|\le\mathscr{V}_j/\mathscr{M}^2$. Plugging this in and rearranging delivers the second statement of the theorem. □

## 7. An Application to Covariance Estimation and Markov PCA

Principal Component Analysis (PCA) is one of the most fundamental problems in data science, used to reduce the dimensionality of multidimensional datasets while retaining as much of the variance as possible and implemented in a wide variety of downstream applications [35]. The task is to estimate the largest eigenvector of a population covariance matrix from samples. Classically, vector-valued samples $\mathbf{x}_1,\dots,\mathbf{x}_n$ are received i.i.d., and the leading eigenvector of the sample covariance matrix is used to estimate the leading eigenvector of the population covariance matrix.

More precisely, we fix a distribution $\mu$ and let $\Sigma=\mathbb{E}_\mu\left[\mathbf{x}\mathbf{x}^\top\right]\in\mathbb{R}^{d\times d}$. Let $\chi_1\ge\cdots\ge\chi_d$ be the eigenvalues of $\Sigma$ in descending order, and let $\mathbf{v}_1$ be the leading eigenvector of $\Sigma$. The following is a standard consequence of the classical matrix Bernstein inequality [64] and Wedin's theorem [69]:

**Theorem 7.1** (Offline PCA, [32]) *Fix $\delta \in (0,1)$. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be drawn i.i.d. from $\mu$. Assume that $\left\| \mathbb{E}_\mu \left[ \left( \mathbf{x_j} \mathbf{x_j}^\top - \Sigma \right)^2 \right] \right\| \leq \mathscr{V}_j$ and $\left\| \mathbf{x}_j \mathbf{x}_j^\top - \Sigma \right\| \leq \mathscr{M}$ almost surely; let $\sigma^2 = \sum_{j=1}^n \mathscr{V}_j$. Let $\hat{\mathbf{v}}$ be the leading eigenvector of $\frac{1}{n} \sum_{j=1}^n \mathbf{x}_j \mathbf{x}_j^\top$. Then with probability $1-\delta$*

$$1 - \langle \hat{\mathbf{v}}, \mathbf{v}_1 \rangle^2 \leq C_1 \frac{\frac{\sigma^2}{n} \ln\left(\frac{d}{\delta}\right)}{(\mu_1 - \mu_2)^2} \cdot \frac{1}{n} + C_2 \left( \frac{\mathscr{M} \ln\left(\frac{d}{\delta}\right)}{\chi_1 - \chi_2} \right)^2 \cdot \frac{1}{n^2} \tag{7.1}$$

*where $C_1, C_2$ are absolute constants.*

In many data tasks the samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are not drawn i.i.d. but have instead some dependent structure; for example, this can occur with time series data [29, 34]. Specifically, the dependent structure can be Markovian, where the vectors $\mathbf{x}_j := \mathbf{x}_j(s_j)$ are vector-valued functions of an underlying Markov chain. An example of this is in the context of token algorithms for Federated PCA, where multiple machines are connected via a graph [14, 23, 25]: each machine contains some fraction of the total dataset and the goal is to obtain the principal component of the entire dataset with respect to some target stationary distribution; work has also been done to adapt this to the streaming setting [38] (whereas our following bound is applicable to the offline setting). The following bound is the adaptation of Theorem 7.1 to the Markov setting, which is a consequence of Theorem 2.6 and Wedin's theorem; it was first found in [38] and is derived as an immediate application of our results. Thus, our theorems are able to directly give the first known bounds for offline Markov PCA in the literature.

**Theorem 7.2** (Offline Markov PCA, [38]) *Fix $\delta \in (0,1)$. Let $P$ be a discrete-time Markov chain on general state space with stationary distribution $\mu$ and absolute spectral gap $\lambda$. Let $s_1, \ldots, s_n$ be a sequence of states driven by $P$ with initial distribution $\mu$. Consider a sequence of vector valued functions $\mathbf{x}_j := \mathbf{x}_j(s_j), j = 1, \ldots, n$. Assume that $\left\| \mathbb{E}_\mu \left[ \left( \mathbf{x_j} \mathbf{x_j}^\top - \Sigma \right)^2 \right] \right\| \leq \mathscr{V}_j$ and $\left\| \mathbf{x}_j \mathbf{x}_j^\top - \Sigma \right\| \leq \mathscr{M}$ almost surely; let $\sigma^2 = \sum_{j=1}^n \mathscr{V}_j$. Let $\hat{\mathbf{v}}$ be the leading eigenvector of $\frac{1}{n} \sum_{j=1}^n \mathbf{x}_j \mathbf{x}_j^\top$. Then with probability $1-\delta$*

$$1 - \langle \hat{\mathbf{v}}, \mathbf{v}_1 \rangle^2 \leq C_1' \frac{\frac{\sigma^2}{n} \ln\left(\frac{d^{2-\pi/4}}{\delta}\right)}{(\chi_1 - \chi_2)^2} \left( \frac{1+\lambda}{1-\lambda} \right) \cdot \frac{1}{n} + C_2' \left( \frac{\mathscr{M} \ln\left(\frac{d^{2-\pi/4}}{\delta}\right)}{(\chi_1 - \chi_2)(1-\lambda)} \right)^2 \cdot \frac{1}{n^2} \tag{7.2}$$

*where $C_1', C_2'$ are absolute constants.*

Observe that the bound from Theorem 7.2 is a natural generalization of the bound from Theorem 7.1 up to constants, the dependence on the absolute spectral gap $\lambda$, and the extra $2 - \pi/4$ factor in the dimension (this can be removed with a slightly worse probability of success); when $\lambda = 0$, which recovers the independent setting, the bound essentially matches that of Theorem 7.1. Note that there are now two spectral gaps: the absolute spectral gap $\lambda$ corresponding to the underlying Markov chain, which governs mixing and thus the spectral norm bound between the sample and population covariances, and the difference $\chi_1 - \chi_2$ corresponding to the population covariance matrix $\Sigma$ that governs the quality of the estimated eigenvector from Wedin's theorem.

More generally, our bounds can be useful for analyzing algorithms for covariance matrix estimation for a stationary distribution based on dependent samples from a Markov chain [37].

## 8. Discussion

In this work we gave a systematic study of concentration inequalities for sums of Markov-dependent random matrices, namely, sums of time-dependent matrix-valued functions of a nonreversible Markov chain on continuousgi state spaces. Our techniques broadly follow the classic literature on spectral methods for sums of Markov-dependent random variables, as we bound the largest eigenvalue of certain perturbed operators. To address time-dependent Markov chains on continuous state spaces we first gave a crucial limit lemma justifying the analysis of the largest eigenvalue of a certain operator, analogous to the scalar case; in the process, we make an interesting connection to the Krein-Rutman theorem, the theory of operators leaving invariant a cone, to tackle the spectral analysis.

In addition to this, we gave a tighter bound on the moment generating function in the Hoeffding setting, exposing the sub-Gaussian nature of the sum, and used this to give improved constants in the final tail bound. Our technique here was to construct a natural coupling with a two-state chain via a convex majorization argument, similar to the proof of the classical Hoeffding's lemma. We also gave the first Bernstein-type inequality for sums of Markov-dependent random matrices via a recursive, linear algebraic, argument – our bound on the moment generating function nicely reveals the Markov dependence, and provides corresponding tail bounds. Both the Hoeffding and Bernstein inequalities thus generalize exactly, and in the Bernstein case even improves, the scalar setting. In addition, we expect our general spectral framework to readily extend to other concentration inequalities, such as Bennett's inequality and Bernstein's inequality with unbounded summands (when there is control over the moments). We hope that our work opens up avenues for application and contributes to a growing body of work on concentration inequalities of sums of random matrices beyond independence.

One interesting direction is a possible improvement of Corollary 2.8. In general, finite moments of the Radon-Nikodym derivative $\frac{d\nu}{d\mu}$ can be much smaller than the essential supremum; specifically, it is possible in the scalar setting to tradeoff an improved multiplicative factor involving the $p^{\text{th}}$ moment of the Radon-Nikodym derivative with worse constants in the exponent – the proof is just Hölder's inequality. However, for matrices this argument would force us to consider not the Frobenius norm squared in the moment generating function but the Frobenius norm raised to some power depending on $p$. It would be interesting if such a tradeoff could be made in our setting as well.

Another interesting direction is with regards to the spectral gap of the chain. Our "absolute spectral gap" is another name for multiplicative reversibilization, commonly used to quantify the mixing of nonreversible chains and is used to reduce analysis to a reversible chain. Though widespread in the literature, it can be a pessimistic estimate of mixing [7]; it would be nice to have a bound that depends on a more "nonreversible" quantity. In addition, compared to the independent setting, our variance proxy is different and potentially slightly suboptimal; it is in term of a "sum of norms" rather than a "norm of a sum." However, we doubt this can be recovered using spectral methods; we believe an improved variance proxy would require new techniques – such as an application of functional inequalities – that are not spectral in nature or at least do not rely on the Golden-Thompson inequality utilized here, and thus imply an entirely new approach for ergodic sums of scalar random variables as well.

## Acknowledgments

JN is supported by NSF CAREER Award 2145800 and NSF DMS-2204449, with partial support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC-2047/1 – 390685813. BS acknowledges support through NSF IFML grant 2019844. RW is supported by AFOSR MURI FA9550-19-1-0005, NSF DMS-1952735, NSF DMS-2109155, and

NSF IFML grant 2019844. The authors thank Sanjay Shakkottai for useful references and Purnamrita Sarkar for helpful discussions.

## REFERENCES


1. Radosław Adamczak and Witold Bednorz. Exponential concentration inequalities for additive functionals of markov chains. *ESAIM: PS*, 19:440–481, 2015.
2. Rudolf Ahlswede and Andreas Winter. Strong converse for identification via quantum channels. *IEEE Transactions on Information Theory*, 48(3):569–579, 2002.
3. Richard Aoun, Marwa Banna, and Pierre Youssef. Matrix poincaré inequalities and concentration. *Advances in Mathematics*, 371:107251, 09 2020.
4. Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate o(1/n). In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'13, page 773–781, Red Hook, NY, USA, 2013. Curran Associates Inc.
5. Sergei Bernstein. Theory of probability (russian). 1927.
6. Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 02 2013.
7. Sourav Chatterjee. Spectral gap of nonreversible markov chains, 2023.
8. Demetres Christofides and Klas Markström. Expansion properties of random cayley graphs and vertex transitive graphs via matrix martingales. *Random Structures & Algorithms*, 32(1):88–100, 2008.
9. Fan Chung and Linyuan Lu. Concentration inequalities and martingale inequalities: a survey. *Internet Mathematics*, 3(1):79 – 127, 2006.
10. Amir Dembo and Ofer Zeitouni. *Large deviations techniques and applications*, volume 38. Springer Science & Business Media, 2009.
11. Aymeric Dieuleveut, Alain Durmus, and Francis Bach. Bridging the gap between constant step size stochastic gradient descent and Markov chains. *The Annals of Statistics*, 48(3):1348 – 1382, 2020.
12. J. J. Dongarra, J. R. Gabriel, D. D. Koelling, and J. H. Wilkinson. The eigenvalue problem for hermitian matrices with time reversal symmetry. *Linear Algebra and its Applications*, 60:27–42, 1984.
13. Petros Drineas and Anastasios Zouzias. A note on element-wise matrix sparsification via a matrix-valued bernstein inequality. *Information Processing Letters*, 111(8):385–389, 2011.
14. Mathieu Even. Stochastic gradient descent under markovian sampling schemes, 2023.
15. Jianqing Fan, Bai Jiang, and Qiang Sun. Hoeffding's inequality for general markov chains and its applications to statistical learning. *Journal of Machine Learning Research*, 22(139):1–35, 2021.
16. Ankit Garg, Yin Tat Lee, Zhao Song, and Nikhil Srivastava. A matrix expander chernoff bound. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2018, page 1102–1114, New York, NY, USA, 2018. Association for Computing Machinery.
17. Charles J. Geyer. Practical Markov Chain Monte Carlo. *Statistical Science*, 7(4):473 – 483, 1992.
18. Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via hessian eigenvalue density. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2232–2241. PMLR, 09–15 Jun 2019.
19. David Wallace Gillman. *Hidden Markov chains: convergence rates and the complexity of inference*. PhD thesis, Massachusetts Institute of Technology, 1993.
20. Peter W. Glynn and Dirk Ormoneit. Hoeffding's inequality for uniformly ergodic markov chains. *Statistics & Probability Letters*, 56(2):143–146, 2002.
21. Sidney Golden. Lower bounds for the helmholtz function. *Phys. Rev.*, 137:B1127–B1128, Feb 1965.
22. Gianluca Gorni. Conjugation and second-order properties of convex functions. *Journal of Mathematical Analysis and Applications*, 158(2):293–315, 1991.

23. Andreas Grammenos, Rodrigo Mendoza-Smith, Jon Crowcroft, and Cecilia Mascolo. Federated principal component analysis, 2020.
24. David Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.
25. Anne Hartebrodt, Reza Nasirigerdeh, David B. Blumenthal, and Richard Röttger. Federated principal component analysis for genome-wide association studies. In *2021 IEEE International Conference on Data Mining (ICDM)*, pages 1090–1095, 2021.
26. Alexander D. Healy. Randomness-efficient sampling within nc1. *Comput. Complex.*, 17(1):3–37, apr 2008.
27. Richard D Hill and Steven R Waters. On the cone of positive semidefinite matrices. *Linear Algebra and its Applications*, 90:81–88, 1987.
28. Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.
29. Siegfried Hörmann, Łukasz Kidziński, and Marc Hallin. Dynamic functional principal components. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 77(2):319–348, 2015.
30. De Huang and Joel A. Tropp. From Poincaré inequalities to nonlinear matrix concentration. *Bernoulli*, 27(3):1724 – 1744, 2021.
31. De Huang and Joel A. Tropp. Nonlinear matrix concentration via semigroup methods. *Electronic Journal of Probability*, 26(none):1 – 31, 2021.
32. Prateek Jain, Sham M. Kakade, Rahul Kidambi, Praneeth Netrapalli, Venkata Krishna Pillutla, and Aaron Sidford. A markov chain theory approach to characterizing the minimax optimality of stochastic gradient descent (for least squares). 2018.
33. Bai Jiang, Qiang Sun, and Jianqing Fan. Bernstein's inequality for general markov chains, 2018.
34. Ian T. Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.
35. I.T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer, 2002.
36. Tarun Kathuria. A matrix bernstein inequality for strong rayleigh distributions, 2020.
37. Yunbum Kook and Matthew S. Zhang. Covariance estimation using markov chain monte carlo, 2024.
38. Syamantak Kumar and Purnamrita Sarkar. Streaming pca for markovian data. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2024. Curran Associates Inc.
39. Rasmus Kyng and Zhao Song. A matrix chernoff bound for strongly rayleigh distributions and spectral sparsifiers from a few random spanning trees, 2018.
40. M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. A Series of Modern Surveys in Mathematics Series. Springer, 1991.
41. Carlos A. León and François Perron. Optimal Hoeffding bounds for discrete reversible Markov chains. *The Annals of Applied Probability*, 14(2):958 – 970, 2004.
42. Pascal Lezaud. Chernoff-type bound for finite Markov chains. *The Annals of Applied Probability*, 8(3):849 – 867, 1998.
43. Pascal Lezaud. Chernoff and berry–esséen inequalities for markov processes. *ESAIM: Probability and Statistics*, 5:183–201, 2001.
44. Xinyan Li, Qilong Gu, Yingxue Zhou, Tiancong Chen, and Arindam Banerjee. *Hessian based analysis of SGD for Deep Nets: Dynamics and Generalization*, pages 190–198.
45. David Lopez-Paz, Suvrit Sra, Alex Smola, Zoubin Ghahramani, and Bernhard Schoelkopf. Randomized nonlinear component analysis. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1359–1367, Bejing, China, 2014. PMLR.
46. Lester Mackey, Michael I. Jordan, Richard Y. Chen, Brendan Farrell, and Joel A. Tropp. Matrix concentration inequalities via the method of exchangeable pairs. *The Annals of Probability*, 42(3):906–945, 05 2014.

47. Avner Magen and Anastasios Zouzias. Low rank matrix-valued chernoff bounds and approximate matrix multiplication. In *Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '11, page 1422–1436, USA, 2011. Society for Industrial and Applied Mathematics.
48. Stephan Mandt, Matthew D. Hoffman, and David M. Blei. Stochastic gradient descent as approximate bayesian inference. *J. Mach. Learn. Res.*, 18(1):4873–4907, jan 2017.
49. Błażej Miasojedow. Hoeffding's inequalities for geometrically ergodic markov chains on general state space. *Statistics & Probability Letters*, 87:115–120, 2014.
50. Mark Aronovich Nauimark. *Normed rings*. P. Noordhoff, 1964.
51. Roger D. Nussbaum. Eigenvectors of nonlinear positive operators and the linear krein-rutman theorem. In Edward Fadell and Gilles Fournier, editors, *Fixed Point Theory*, pages 309–330, Berlin, Heidelberg, 1981. Springer Berlin Heidelberg.
52. Roberto Oliveira. Sums of random Hermitian matrices and an inequality by Rudelson. *Electronic Communications in Probability*, 15(none):203 – 212, 2010.
53. Roberto Imbuzeiro Oliveira. Concentration of the adjacency matrix and of the laplacian in random graphs with independent edges. *arXiv preprint arXiv:0911.0600*, 2009.
54. Daniel Paulin. Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electronic Journal of Probability*, 20(none):1 – 32, 2015.
55. Daniel Paulin, Lester Mackey, and Joel A. Tropp. Efron–Stein inequalities for random matrices. *The Annals of Probability*, 44(5):3431 – 3473, 2016.
56. Jiezhong Qiu, Chi Wang, Ben Liao, Richard Peng, and Jie Tang. A matrix chernoff bound for markov chains and its application to co-occurrence matrices. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc.
57. Benjamin Recht. A simpler approach to matrix completion. *J. Mach. Learn. Res.*, 12(null):3413–3430, dec 2011.
58. R.T. Rockafellar. *Convex Analysis*. Convex Analysis. Princeton University Press, 1997.
59. Daniel Rudolf. Explicit error bounds for markov chain monte carlo. *Dissertationes Mathematicae*, 485:1–93, 2012.
60. David Sutter, Mario Berta, and Marco Tomamichel. Multivariate trace inequalities. *Communications in Mathematical Physics*, 352(1):37–58, May 2017.
61. J.J. Sylvester. Xix. a demonstration of the theorem that every homogeneous quadratic polynomial is reducible by real orthogonal substitutions to the form of a sum of positive and negative squares. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 4(23):138–142, 1852.
62. Colin J. Thompson. Inequality with applications in statistical mechanics. *Journal of Mathematical Physics*, 6(11):1812–1813, November 1965.
63. Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, aug 2011.
64. Joel A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.
65. Joel A. Tropp. The expected norm of a sum of independent random matrices: An elementary approach. In Christian Houdré, David M. Mason, Patricia Reynaud-Bouret, and Jan Rosiński, editors, *High Dimensional Probability VII*, pages 173–202, Cham, 2016. Springer International Publishing.
66. Joel A. Tropp. Second-order matrix concentration inequalities. *Applied and Computational Harmonic Analysis*, 44(3):700–736, 2018.
67. Hermann Von Weyl. Über beschränkte quadratische formen, deren differenz vollstetig ist. *Rendiconti del Circolo Matematico di Palermo (1884-1940)*, 27(1):373–392, 1909.
68. Roy Wagner. Tail estimates for sums of variables sampled from a random walk, 2006.
69. Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12:99–111, 1972.
70. A. Wigderson and D. Xiao. A randomness-efficient sampler for matrix-valued functions and applications. In *46th Annual IEEE Symposium on Foundations of Computer Science (FOCS'05)*, pages 397–406, 2005.

# Appendix

## A. Proofs for Section 2 (Main Results)

*Proof of Corollary 2.7* If $Z = X + iY$ is a complex Hermitian $d \times d$ matrix, then $X$ is real symmetric and $Y$ is skew-symmetric. Then $Z$ is Hermitian if and only if

$$Z' = \begin{bmatrix} X & Y \\ -Y & X \end{bmatrix}$$

is a real symmetric $2d \times 2d$ matrix. If $Z$ has eigenvalues $\lambda_1, \ldots, \lambda_d$ then $Z'$ has eigenvalues $\lambda_1, \lambda_1, \ldots, \lambda_d, \lambda_d$, i.e., has the same eigenvalues each with twice the multiplicity. Then call Theorems 2.5 and 2.6 with the matrices $Z'$. □

*Proof of Corollary 2.8* Let $H(s_1, \ldots, s_n)$ be a measurable function. Then

$$\begin{aligned} \mathbb{E}_\nu[H(s_1, \ldots, s_n)] &= \mathbb{E}_\mu \left[ \frac{d\nu}{d\mu} \cdot H(s_1, \ldots, s_n) \right] \\ &\leq \operatorname{ess\,sup} \frac{d\nu}{d\mu} \cdot \mathbb{E}_\mu[H(s_1, \ldots, s_n)]. \end{aligned}$$

This gives the bounds on the moment generating function and the tail bounds then follow. □

## B. Proofs for Section 3 (Preliminaries)

*Proof of Lemma 3.3* The proof follows in a straightforward manner using properties of the operator norm of a tensor product. We have

$$\begin{aligned} \lambda(\tilde{P}) &= \left\| \tilde{P} - \tilde{\Pi} \right\|_\mu \\ &= \left\| P \otimes I_{d^2} - \Pi \otimes I_{d^2} \right\|_\mu \\ &= \left\| (P - \Pi) \otimes I_{d^2} \right\|_\mu \\ &= \left\| P - \Pi \right\|_\mu \left\| I_{d^2} \right\| \\ &= \lambda(P). \end{aligned}$$

□

## C. Proofs for Section 5 (Bounding the operator norm)

*Proof of Proposition 5.3* For (1), we have

$$\begin{aligned} \mathbb{E}[T(x)] &= \frac{\cos(\phi)}{2} (\mathbb{E}[F(x) \otimes I_d] + \mathbb{E}[I_d \otimes F(x)]) \\ &= \frac{\cos(\phi)}{2} (\mathbb{E}[F(x)] \otimes I_d + I_d \otimes \mathbb{E}[F(x)]) \\ &= 0, \end{aligned}$$

since the Kronecker product is a linear operation.

Now we prove (2). For some $x \in \mathscr{X}$, let $\rho_1(x) \geq \cdots \geq \rho_d(x)$ be the eigenvalues of $F(x)$ (the $\rho$ are implicitly scalar-valued functions). We have that the eigenvalues of $F(x) \otimes I_d$ are exactly these eigenvalues but each with multiplicity $d$ – this also holds for $I_d \otimes F(x)$. Now as $F(x) \otimes I_d$ and $I_d \otimes F(x)$ are Hermitian, a crude estimate is that the maximum eigenvalue of the sum is bounded above by $2\rho_1(x) \leq 2b$ and the minimum eigenvalue of the sum is bounded below by $2\rho_d(x) \geq 2a$. The statement follows.

For (3), we have

$$T(x)^2 = \frac{\cos^2(\phi)}{4}(F(x)^2 \otimes I_d + I_d \otimes F(x)^2 + 2 \cdot F(x) \otimes F(x)).$$

Then

$$\left\|\mathbb{E}[F(x)^2 \otimes I_d]\right\| = \left\|\mathbb{E}[F(x)^2] \otimes I_d\right\| \leq \mathscr{V},$$

and the same holds for $\left\|\mathbb{E}[I_d \otimes F(x)^2]\right\|$. For $F(x) \otimes F(x)$, we consider $(\mathbf{u} \otimes \mathbf{u})^*(F(x) \otimes F(x))(\mathbf{u} \otimes \mathbf{u})$ for some unit vector $\mathbf{u}$. This is equal to $(\mathbf{u}^* F(x)\mathbf{u})^2$. Now by Cauchy-Schwartz,

$$\begin{aligned}(\mathbf{u}^* F(x)\mathbf{u})^2 &\leq (\|F(x)\mathbf{u}\|)^2 \\ &= \mathbf{u}^* F(x)^2 \mathbf{u}.\end{aligned}$$

Then $(\mathbf{u} \otimes \mathbf{u})^* \mathbb{E}[F(x) \otimes F(x)](\mathbf{u} \otimes \mathbf{u}) \leq \mathbf{u}^* \mathbb{E}[F(x)^2]\mathbf{u}$ for any unit vector $\mathbf{u}$. It then follows that $\|\mathbb{E}[F(x) \otimes F(x)]\| \leq \mathscr{V}$. The statement follows by putting the above together. □

## D. Proofs for Section 6 (Final bounds)

### D.1. *Hoeffding*

*Proof of Lemma 6.3* We first note that for any nonnegative convex function $\Psi$, the function $\Psi^p$ is convex for any $p \geq 1$. We first describe a generating process for $s_j$. Draw random variables $I_j$ with $I_1 = 1$ and $I_j \sim \mathrm{Ber}(1-\lambda)$ for $j > 1$ and variables $Z_j \sim \mu$ uniformly and independent from each other and let

$$s_k = \sum_{j=1}^{k} \left( \prod_{\ell=j+1}^{k} (1 - I_\ell) \right) I_j Z_j \tag{D.1}$$

so that

$$\Psi(F(s_k)) = \sum_{j=1}^{k} \left( \prod_{\ell=j+1}^{k} (1 - I_\ell) \right) I_j \Psi(F(Z_j)). \tag{D.2}$$

We can verify that this is a valid construction for a Markov chain driven by $\hat{P}$. We now couple $(y_j, s_j)$ by drawing random variables $B_j \sim \mu$ independent from each other and let

$$y_k = \sum_{j=1}^{k} \left( \prod_{\ell=j+1}^{k} (1 - I_\ell) \right) I_j B_j \tag{D.3}$$

so that

$$\Psi(G(y_k)) = \sum_{j=1}^{k} \left( \prod_{\ell=j+1}^{k} (1 - I_\ell) \right) I_j \Psi(G(B_j)). \tag{D.4}$$

We can again verify that this is a valid construction for a Markov chain driven by $\hat{Q}$ and is also a valid coupling. Then

$$\prod_{k=1}^{n} \Psi(F(s_j)) = \prod_{k=1}^{n} \left( \sum_{k=1}^{n} \left( \prod_{\ell=j+1}^{k} (1 - I_\ell) \right) I_j \Psi(F(Z_j)) \right). \tag{D.5}$$

We can expand this out as a matrix polynomial. Each monomial is of degree $n$ and can be determined by a tuple $\alpha \in \{0,1\}^n$ with $\alpha_1 = 1$ always, each entry corresponding to $I_j$ and $\Psi(F(Z_j))$. For each $\alpha$ construct a coefficient $c_\alpha$ and a partition $a$ of $n$ as follows: if $\alpha_j = 0$ then $(1 - I_j)$ appears in $c_\alpha$ and $a_j = 0$. If $\alpha_j = 1$ then $I_j$ appears in $c_\alpha$ and $a_j$ is equal to the one plus the number of zeros appearing sequentially after $\alpha_j$ until the next one. Then we have $c_\alpha \prod_{j=1}^{n} \Psi(F(Z_j))^{a_j}$ as our matrix monomial corresponding to $\alpha$. There are clearly $2^{n-1}$ monomials in this matrix polynomial.

For example, let $n = 6$. The monomial corresponding to $\alpha = (1,0,1,0,1,0)$ is

$$(1 - I_6) I_5 (1 - I_4) I_3 (1 - I_2) I_1 \Psi(F(Z_5))^2 \Psi(F(Z_3))^2 \Psi(F(Z_1))^2,$$

the monomial corresponding to $\alpha = (1,0,1,0,0,1)$ is

$$I_6 (1 - I_5)(1 - I_4) I_3 (1 - I_2) I_1 \Psi(F(Z_6)) \Psi(F(Z_3))^3 \Psi(F(Z_1))^2,$$

the monomial corresponding to $\alpha = (1,1,1,1,1,1)$ is

$$I_6 I_5 I_4 I_3 I_2 I_1 \Psi(F(Z_6)) \Psi(F(Z_5)) \Psi(F(Z_4)) \Psi(F(Z_3)) \Psi(F(Z_2)) \Psi(F(Z_1)),$$

and the monomial corresponding to $\alpha = (1,0,0,0,0,0)$ is

$$(1 - I_6)(1 - I_5)(1 - I_4)(1 - I_3)(1 - I_2)(1 - I_1) \Psi(F(Z_1))^6.$$

Moreover, it is not hard to see that the monomials in this matrix polynomial are pairwise orthogonal (with respect to the trace inner product) using the property that $(1 - I_j) I_j = 0$. Therefore

$$\begin{aligned} \mathbb{E}_\mu \left[ \left\| \prod_{j=1}^{n} \Psi(F(s_j)) \right\|_F^2 \right] &= \mathbb{E}_\mu \left[ \left\| \sum_\alpha c_\alpha \prod_{j=1}^{n} \Psi(F(Z_j))^{a_j} \right\|_F^2 \right] \\ &= \mathbb{E}_\mu \left[ \sum_\alpha c_\alpha \left\| \prod_{j=1}^{n} \Psi(F(Z_j))^{a_j} \right\|_F^2 \right] \quad , \\ &= \sum_\alpha \mathbb{E}[c_\alpha] \mathbb{E}_\mu \left[ \left\| \prod_{j=1}^{n} \Psi(F(Z_j))^{a_j} \right\|_F^2 \right] \end{aligned} \tag{D.6}$$

where the last line follows from independence of the $I_j$ and $Z_j$. We now apply Lemma D.1 to see that

$$\begin{aligned}\mathbb{E}_\mu\left[\left\|\prod_{j=1}^{n}\Psi(F(s_j))\right\|_F^2\right] &= \sum_\alpha \mathbb{E}[c_\alpha]\mathbb{E}_\mu\left[\left\|\prod_{j=1}^{n}\Psi(F(Z_j))^{a_j}\right\|_F^2\right] \\ &\leq \sum_\alpha \mathbb{E}[c_\alpha]\mathbb{E}_\mu\left[\left\|\prod_{j=1}^{n}\Psi(G(B_j))^{a_j}\right\|_F^2\right] \\ &= \mathbb{E}_\mu\left[\left\|\prod_{j=1}^{n}\Psi(G(y_j))\right\|_F^2\right],\end{aligned} \tag{D.7}$$

where we use the coupling between $y_j$ and $s_j$. □

**Lemma D.1** *For $Z_j$ drawn i.i.d. from $\mu$ and $B_j$ drawn i.i.d. from $\mu$, and if $\mathbb{E}_\mu[F] = \mathbb{E}_\mu[G] = 0$, and for a nonnegative scalar convex function $\Psi$, we have*

$$\mathbb{E}_\mu\left[\left\|\prod_{j=1}^{n}\Psi(F(Z_j))^{a_j}\right\|_F^2\right] \leq \mathbb{E}_\mu\left[\left\|\prod_{j=1}^{n}\Psi(G(B_j))^{a_j}\right\|_F^2\right].$$

*Proof* We have

$$\begin{aligned}\mathbb{E}_\mu\left[\left\|\prod_{j=1}^{n}\Psi(F(Z_j))^{a_j}\right\|_F^2\right] &= \mathbb{E}_\mu\left[\left\langle\prod_{j=1}^{n}\Psi(F(Z_j))^{a_j}, \prod_{j=1}^{n}\Psi(F(Z_j))^{a_j}\right\rangle\right] \\ &= \mathbb{E}_\mu\left[\left\langle\mathbb{E}_\mu\left[\Psi(F(Z_n))^{2a_n}\right], \prod_{j=1}^{n-1}\Psi(F(Z_j))^{a_j}\prod_{j=n-1}^{1}\Psi(F(Z_j))^{a_j}\right\rangle\right]\end{aligned} \tag{D.8}$$

using the independence of the $Z_j$. Now as $\Psi$ is nonnegative and convex, $\Psi^{a_j}$ is convex. Therefore, by Proposition 6.2 and the fact that $F$ is mean zero, we have

$$\mathbb{E}_\mu\left[\Psi(F(Z_n))^{2a_n}\right] \preceq \mathbb{E}_\mu\left[\Psi(G(B_n))^{2a_n}\right] \tag{D.9}$$

so that

$$\begin{aligned}&\mathbb{E}_\mu\left[\left\langle\mathbb{E}_\mu\left[\Psi(F(Z_n))^{2a_n}\right], \prod_{j=1}^{n-1}\Psi(F(Z_j))^{a_j}\prod_{j=n-1}^{1}\Psi(F(Z_j))^{a_j}\right\rangle\right] \\ \leq &\mathbb{E}_\mu\left[\left\langle\mathbb{E}_\mu\left[\Psi(G(B_n))^{2a_n}\right], \prod_{j=1}^{n-1}\Psi(F(Z_j))^{a_j}\prod_{j=n-1}^{1}\Psi(F(Z_j))^{a_j}\right\rangle\right].\end{aligned} \tag{D.10}$$

Now since $G$ maps to matrices that are a constant times the identity, the matrices $\mathbb{E}_\mu\left[\Psi(G(B_j))^{2a_j}\right]$ commute with all other matrices. This allows us to write

$$\begin{aligned}&\mathbb{E}_\mu\left[\left\langle \mathbb{E}_\mu\left[\Psi(G(B_n))^{2a_n}\right], \prod_{j=1}^{n-1}\Psi(F(Z_j))^{a_j}\prod_{j=n-1}^{1}\Psi(F(Z_j))^{a_j}\right\rangle\right]\\ =&\mathbb{E}_\mu\left[\left\langle \mathbb{E}_\mu\left[\Psi(F(Z_{n-1}))^{2a_{n-1}}\right], \prod_{j=1}^{n-2}\Psi(F(Z_j))^{a_j}\mathbb{E}_\mu\left[\Psi(G(B_n))^{2a_n}\right]\prod_{j=n-2}^{1}\Psi(F(Z_j))^{a_j}\right\rangle\right].\end{aligned} \tag{D.11}$$

We can iterate this process, and using independence of the $B_j$ and rearranging (recall all these resulting matrices commute) we obtain the statement of the lemma. □

*Proof of Lemma 6.5, [15]* With some abuse of notation, reuse $a := \cos(\phi)a$ and $b := \cos(\phi)b$. Since $\phi \in [-\pi/2, \pi/2]$, $\phi$ is always nonnegative, so the ordering remains the same.

Let $q = -a/(b-a)$. The eigenvalues of $K^\theta$ are the roots of the quadratic

$$\begin{aligned}0 &= \det(\tilde{\eta} I - K^\theta)\\ &= \tilde{\eta}^2 - (1+\lambda)[(1-p)e^{\theta a} + pe^{\theta b}]\tilde{\eta} + \lambda e^{\theta(a+b)},\end{aligned} \tag{D.12}$$

where

$$p = \frac{\lambda + (1-\lambda)q}{1+\lambda}. \tag{D.13}$$

Then $\eta$ is simply the larger of these two. Instead of solving directly for $\eta$, we can choose any value that upper bounds $\eta$ that has a considerably simpler form. In particular, if a quadratic is of the form $x^2 - bx + c$, we can choose any $x$ such that $x^2 - bx + c \geq 0$ and $x^2 \geq c$; it can be verified that such an $x$ upper bounds both roots of the quadratic. Thus, with our choice of $\tilde{\eta}(\theta) = \exp(\alpha(\lambda)\cdot\theta^2(b-a)^2/8)$, we show

$$\begin{aligned}&\tilde{\eta}(\theta)^2 - (1+\lambda)[(1-p)e^{\theta a} + pe^{\theta b}]\tilde{\eta}(\theta) + \lambda e^{\theta(a+b)} \geq 0,\\ &\tilde{\eta}(\theta)^2 \geq \lambda e^{\theta(a+b)}.\end{aligned} \tag{D.14}$$

The first point in Equation D.14 holds if and only if

$$\frac{\tilde{\eta}(\theta) + \lambda e^{\theta(a+b)}\tilde{\eta}(\theta)^{-1}}{1+\lambda} \geq (1-p)e^{\theta a} + pe^{\theta b}. \tag{D.15}$$

Then for the left-hand side of Equation D.15 we have

$$\begin{aligned}
&\frac{\tilde{\eta}(\theta)+\lambda e^{\theta(a+b)}\tilde{\eta}(\theta)^{-1}}{1+\lambda} \\
&= \frac{\exp(\alpha(\lambda)\theta^2(b-a)^2/8)+\lambda\exp(\theta(a+b)-\alpha(\lambda)\theta^2(a-b)^2/8)}{1+\lambda} \\
&\geq \exp\left(\frac{\alpha(\lambda)\theta^2(b-a)^2/8+\lambda\theta(a+b)-\lambda\alpha(\lambda)\theta^2(b-a)^2/8}{1+\lambda}\right) \\
&= \exp\left(\theta\cdot\frac{\lambda(a+b)}{1+\lambda}+\frac{(b-a)^2\theta^2}{8}\cdot\frac{(1-\lambda)\alpha(\lambda)}{1+\lambda}\right) \\
&= \exp\left(\theta\cdot[(1-p)a+pb]+\theta^2\cdot\frac{(b-a)^2}{8}\right),
\end{aligned}$$

where the lower bound is by convexity of the exponential map. Now the right-hand side of Equation D.15 is the moment generating function of the random variable taking value $a$ with probability $1-p$ and value $b$ with probability $p$, and thus is upper bounded by Hoeffding's lemma as

$$(1-p)e^{\theta a}+pe^{\theta b} \leq \exp\left(\theta\cdot[(1-p)a+pb]+\theta^2\cdot\frac{(b-a)^2}{8}\right).$$

Putting the two together proves the first line of Equation D.14.

The second line of Equation D.14 can be developed as

$$\begin{aligned}
\log\left(\tilde{\eta}(\theta)^2 e^{-\theta(a+b)}\right) &= \log\left(\exp\left(\frac{\alpha(\lambda)\theta^2(b-a)^2}{4}-\theta(a+b)\right)\right) \\
&= \frac{\alpha(\lambda)\theta^2(b-a)^2}{4}-\theta(a+b).
\end{aligned}$$

We show that this is at least $-1/\alpha(\lambda)$. A sufficient condition for this is

$$\alpha(\lambda)(a+b)\theta-\frac{\alpha(\lambda)^2(b-a)^2}{4}\theta^2 \leq 1.$$

The above is maximized at $\theta = 2(a+b)/(\alpha(\lambda)(b-a)^2)$, and plugging this in, we have

$$\alpha(\lambda)(a+b)\theta-\frac{\alpha(\lambda)^2(b-a)^2}{4}\theta^2 = \frac{(a+b)^2}{(b-a)^2}$$

which is always at most 1 since $a \leq 0 \leq b$. Then

$$\begin{aligned}
\log\left(\tilde{\eta}(\theta)^2 e^{-\theta(a+b)}\right) &= \frac{\alpha(\lambda)\theta^2(b-a)^2}{4}-\theta(a+b) \\
&\geq -\frac{1}{\alpha(\lambda)} \\
&= -\frac{1-\lambda}{1+\lambda} \\
&\geq \log\lambda
\end{aligned}$$

when $0 \leq \lambda \leq 1$, proving the second line of Equation D.14.

As $\tilde{\eta}(\theta)$ has been established to be a suitable upper bound, we then replace $a$ and $b$ in its definition by $\cos(\phi)a$, $\cos(\phi)b$, respectively to obtain the statement of the lemma. □

D.2. *Bernstein*

*Proof of Lemma 6.7* First we use an explicit decomposition of $\mathbf{z}$. Let $\mathbf{z} = \left\|\mathbf{z}^{\|}\right\| (\mathbf{1} \otimes \mathbf{v}_1) + \left\|\mathbf{z}^{\perp}\right\|_{\mu} \rho$, and $E_T^{\theta}(\mathbf{1} \otimes \mathbf{v}_1) = a(\mathbf{1} \otimes \mathbf{v}_2) + c\sigma$, $E_T^{\theta}\rho = b(\mathbf{1} \otimes \mathbf{v}_3) + d\tau$, with $\rho, \sigma, \tau$ norm 1 and orthogonal to $\mathbf{1} \otimes \mathbb{C}^{d^2}$ and $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ all norm 1. Let $D$ be the multiplication operator that acts diagonally by $T$, so that $E_T^{\theta} = \exp(\theta D)$.

(1) We have $\left\|(E_T^{\theta}\mathbf{z}^{\|})^{\|}\right\|_{\mu} = |a| \left\|\mathbf{z}^{\|}\right\|_{\mu}$. It suffices to bound $|a|$. First we see that $a = \left\langle \mathbf{1} \otimes \mathbf{v}_2, E_T^{\theta}(\mathbf{1} \otimes \mathbf{v}_1) \right\rangle_{\mu}$. Then

$$\left\langle \mathbf{1} \otimes \mathbf{v}_2, E_T^{\theta}(\mathbf{1} \otimes \mathbf{v}_1) \right\rangle_{\mu} = \left\langle \mathbf{v}_2, \mathbb{E}_{\mu}[\exp(\theta T(x))]\mathbf{v}_1 \right\rangle. \tag{D.16}$$

We see this by definition of the multiplication operator $E_T^{\theta}$ and the definition of the inner product of $\ell_2(\mu \otimes \mathbf{1})$. Then

$$\begin{aligned} \left\langle \mathbf{v}_2, \mathbb{E}_{\mu}[\exp(\theta T(x))]\mathbf{v}_1 \right\rangle &= \mathbb{E}_{\mu}\left[\langle \mathbf{v}_2, \exp(\theta T(x))\mathbf{v}_1 \rangle\right] \\ &= \mathbb{E}_{\mu}\left[\left\langle \mathbf{v}_2, I + \sum_{k=2}^{\infty} \frac{\theta^k}{k!} T(x)^k \mathbf{v}_1 \right\rangle\right]. \end{aligned} \tag{D.17}$$

We used the fact that $\mathbb{E}_{\mu}[T(x)] = 0$ to remove the $k = 1$ term. Next,

$$\mathbb{E}_{\mu}\left[\left\langle \mathbf{v}_2, I + \sum_{k=2}^{\infty} \frac{\theta^k}{k!} T(x)^k \mathbf{v}_1 \right\rangle\right] = \langle \mathbf{v}_2, \mathbf{v}_1 \rangle + \sum_{k=2}^{\infty} \frac{\theta^k}{k!} \left\langle \mathbf{v}_2, \mathbb{E}_{\mu}[T(x)^k]\mathbf{v}_1 \right\rangle. \tag{D.18}$$

But the inner products in the sum are bounded by the spectral norm of $\mathbb{E}_{\mu}[T(x)^k]$, so we apply Lemma 6.6.2 from [64] and Proposition 5.3. Thus

$$\begin{aligned} \left|\left\langle \mathbf{v}_2, \mathbb{E}_{\mu}[\exp(\theta T(x))]\mathbf{v}_1 \right\rangle\right| &\leq 1 + \mathscr{V} \sum_{k=2}^{\infty} \frac{\theta^k}{k!} \mathscr{M}^{k-2} \\ &= 1 + \frac{\mathscr{V}(e^{\mathscr{M}\theta} - \mathscr{M}\theta - 1)}{\mathscr{M}^2}. \end{aligned} \tag{D.19}$$

(2) We have $\left\|(E_T^{\theta}\mathbf{z}^{\perp})^{\|}\right\|_{\mu} = |b| \left\|\mathbf{z}^{\perp}\right\|_{\mu}$. It suffices to bound $|b|$. As before we see $b = \left\langle \mathbf{1} \otimes \mathbf{v}_3, E_T^{\theta}\rho \right\rangle_{\mu}$. From our definition of $D$, we have

$$\begin{aligned} \left\langle \mathbf{1} \otimes \mathbf{v}_3, E_T^{\theta}\rho \right\rangle_{\mu} &= \left\langle \mathbf{1} \otimes \mathbf{v}_3, I + \sum_{k=1}^{\infty} \frac{\theta^k}{k!} D^k \rho \right\rangle_{\mu} \\ &= \langle \mathbf{1} \otimes \mathbf{v}_3, \rho \rangle_{\mu} + \sum_{k=1}^{\infty} \frac{\theta^k}{k!} \left\langle \mathbf{1} \otimes \mathbf{v}_3, D^k \rho \right\rangle_{\mu} \\ &= \sum_{k=1}^{\infty} \frac{\theta^k}{k!} \left\langle \mathbf{1} \otimes \mathbf{v}_3, D^k \rho \right\rangle_{\mu}, \end{aligned} \tag{D.20}$$

since $\mathbf{1}\otimes\mathbf{v}_3$ and $\rho$ are orthogonal. Now since $\|T\|\le\mathscr{M}$ we also have $\|D\|_\mu\le\mathscr{M}$, and so

$$-\mathscr{M}^{k-1}\langle\mathbf{1}\otimes\mathbf{v}_3,D\rho\rangle_\mu\le\left\langle\mathbf{1}\otimes\mathbf{v}_3,D^k\rho\right\rangle_\mu\le\mathscr{M}^{k-1}\langle\mathbf{1}\otimes\mathbf{v}_3,D\rho\rangle_\mu.$$

This is in turn upper bounded by $\|D(\mathbf{1}\otimes\mathbf{v}_3)\|_\mu$ and lower bounded by its negation by Cauchy-Schwartz and seeing that $\|\rho\|_\mu=1$. To bound this norm, we have

$$\begin{aligned}\|D(\mathbf{1}\otimes\mathbf{v}_3)\|_\mu^2&=\langle D(\mathbf{1}\otimes\mathbf{v}_3),D(\mathbf{1}\otimes\mathbf{v}_3)\rangle_\mu\\&=\left\langle\mathbf{1}\otimes\mathbf{v}_3,\mathbb{E}_\mu[T(x)^2](\mathbf{1}\otimes\mathbf{v}_3)\right\rangle\\&\le\mathscr{V}.\end{aligned}\tag{D.21}$$

It then follows that

$$\begin{aligned}\left|\left\langle\mathbf{1}\otimes\mathbf{v}_3,E_T^\theta\rho\right\rangle_\mu\right|&\le\sqrt{\mathscr{V}}\sum_{k=1}^\infty\frac{\theta^k}{k!}\mathscr{M}^{k-1}\\&=\frac{\sqrt{\mathscr{V}}(e^{\mathscr{M}\theta}-1)}{\mathscr{M}}.\end{aligned}\tag{D.22}$$

(3) $\left\|(E_T^\theta\mathbf{z}^\parallel)^\perp\right\|=|c|\left\|\mathbf{z}^\parallel\right\|_\mu$. Bounding $|c|$, we see $c=\left\langle\sigma,E^\theta(\mathbf{1}\otimes\mathbf{v}_1)\right\rangle_\mu$. Then the result follows as in the previous part.

(4) We simply use the uniform upper bound of $\|\exp(\theta T(x))\|\le\exp(\mathscr{M}\theta)$ for all $x\in\mathscr{X}$.

□

*Proof of Claim 6.8* We have

$$\begin{aligned}\left\|\mathbf{z}_k^\perp\right\|_\mu&=\left\|\left(\hat{P}E_T^\theta\mathbf{z}_{k-1}\right)^\perp\right\|_\mu\\&\le\left\|\left(\hat{P}E_T^\theta\mathbf{z}_{k-1}^\parallel\right)^\perp\right\|_\mu+\left\|\left(\hat{P}E_T^\theta\mathbf{z}_{k-1}^\perp\right)^\perp\right\|_\mu\\&=\left\|\hat{P}\left(E_T^\theta\mathbf{z}_{k-1}^\parallel\right)^\perp\right\|_\mu+\left\|\hat{P}\left(E_T^\theta\mathbf{z}_{k-1}^\perp\right)^\perp\right\|_\mu\\&\le\lambda\left\|\left(E_T^\theta\mathbf{z}_{k-1}^\parallel\right)^\perp\right\|_\mu+\lambda\left\|\left(E_T^\theta\mathbf{z}_{k-1}^\perp\right)^\perp\right\|_\mu\\&\le\lambda\alpha_2\left\|\mathbf{z}_{k-1}^\parallel\right\|_\mu+\lambda\alpha_3\left\|\mathbf{z}_{k-1}^\perp\right\|_\mu\\&\le\left(\lambda\alpha_2+(\lambda\alpha_2)(\lambda\alpha_3)+\cdots+(\lambda\alpha_2)(\lambda\alpha_3)^{k-1}\right)\max_{0\le j\le k}\left\|\mathbf{z}_j^\parallel\right\|_\mu\\&\le\frac{\lambda\alpha_2}{1-\lambda\alpha_3}\max_{0\le j\le k}\left\|\mathbf{z}_j^\parallel\right\|_\mu,\end{aligned}\tag{D.23}$$

where the last step holds if $\lambda\alpha_3=\lambda e^{\mathscr{M}\theta}<1$, implied if $\theta<\log(1/\lambda)/\mathscr{M}$. □

*Proof of Claim 6.9* We have

$$\begin{aligned}
\left\|\mathbf{z}_k^{\parallel}\right\|_\mu &= \left\|\left(\hat{P}E_T^\theta \mathbf{z}_{k-1}\right)^{\parallel}\right\|_\mu \\
&\le \left\|\left(\hat{P}E_T^\theta \mathbf{z}_{k-1}^{\parallel}\right)^{\parallel}\right\|_\mu + \left\|\left(\hat{P}E_T^\theta \mathbf{z}_{k-1}^{\perp}\right)^{\parallel}\right\|_\mu \\
&\le \left\|\hat{P}\left(E_T^\theta \mathbf{z}_{k-1}^{\parallel}\right)^{\parallel}\right\|_\mu + \left\|\tilde{P}\left(E_T^\theta \mathbf{z}_{k-1}^{\perp}\right)^{\parallel}\right\|_\mu \\
&= \left\|\left(E_T^\theta \mathbf{z}_{k-1}^{\parallel}\right)^{\parallel}\right\|_\mu + \left\|\left(E_T^\theta \mathbf{z}_{k-1}^{\perp}\right)^{\parallel}\right\|_\mu \\
&\le \alpha_1 \left\|\mathbf{z}_{k-1}^{\parallel}\right\|_\mu + \alpha_2 \left\|\mathbf{z}_{k-1}^{\perp}\right\|_\mu \\
&\le \alpha_1 \left\|\mathbf{z}_{k-1}^{\parallel}\right\|_\mu + \frac{\lambda\alpha_2^2}{1-\lambda\alpha_3} \max_{0\le j\le k-1} \left\|\mathbf{z}_j^{\parallel}\right\|_\mu \\
&\le \left(\alpha_1 + \frac{\lambda\alpha_2^2}{1-\lambda\alpha_3}\right) \max_{0\le j\le k-1} \left\|\mathbf{z}_j^{\parallel}\right\|_\mu,
\end{aligned} \tag{D.24}$$

where we apply Claim 6.8. □

The following lemma bounds the conjugate function in Equation 6.29. The proof is adapted from [33] to our setting and is standard.

**Lemma D.2** *Define*

$$g_1(t) = \sigma^2\left(e^{L\theta} - L\theta - 1\right), \quad g_2(t) = \frac{\sigma^2\lambda L^2\theta^2}{1-\lambda-2L\theta}$$

*for any* $0 \le \theta < (1-\lambda)/2L$. *If* $0 < \lambda < 1$ *then*

$$(g_1+g_2)^*(t) \ge \frac{t^2}{2L^2}\left(\frac{1+\lambda}{1-\lambda}\cdot\sigma^2 + \frac{2Lt}{1-\lambda}\right)^{-1}.$$

*If* $\lambda = 0$ *then*

$$(g_1+g_2)^*(t) \ge \frac{t^2}{2L^2}\left(\sigma^2 + \frac{Lt}{3}\right)^{-1}.$$

*Proof* Extend $g_1$ and $g_2$ to all of $\mathbb{R}$:

$$g_1(\theta) = \begin{cases} 0, & \theta < 0 \\ \sigma^2(e^{L\theta} - L\theta - 1), & \theta \ge 0 \end{cases},$$

$$g_2(\theta) = \begin{cases} 0, & \theta < 0 \\ \frac{\sigma^2\lambda L^2\theta^2}{1-\lambda-2L\theta}, & 0 \le \theta < \frac{1-\lambda}{2L} \\ \infty, & \theta \ge \frac{1-\lambda}{2L} \end{cases}.$$

These are close and convex and so admit convex conjugates:

$$g_1^*(t) = \begin{cases} \sigma^2 h_1\left(\frac{t}{L\sigma^2}\right), & t \geq 0, \\ \infty, & t < 0 \end{cases}, \tag{D.25}$$

where $h_1(x) = (1+x)\log(1+x) - x$ and

$$g_2^*(t) = \begin{cases} \frac{(1-\lambda)t^2}{2\lambda\sigma^2 L^2} h_2\left(\frac{2t}{\lambda\sigma^2 L}\right), & t \geq 0, \\ \infty, & t < 0 \end{cases}, \tag{D.26}$$

where $h_2(x) = (\sqrt{1+x} + x/2 + 1)^{-1}$.

Now $(g_1+g_2)^*$ is consistent with the above definitions, and so $(g_1+g_2)^*(t) = \sup\{\theta t - g_1(\theta) - g_2(\theta)\}$. If $\lambda > 0$, the Moreau-Rockafellar formula [58] states

$$\begin{aligned}(g_1+g_2)^*(t) &= \inf\{g_1^*(t_1) + g_2^*(t_2) \mid t_1 + t_2 = t\} \\ &= \inf\left\{\sigma^2 h_1\left(\frac{t_1}{L\sigma^2}\right) + \frac{(1-\lambda)t_2^2}{2\lambda\sigma^2 L^2} h_2\left(\frac{2t_2}{\lambda\sigma^2 L}\right) \mid t_1 + t_2 = t, t_1, t_2 \geq 0\right\}\end{aligned}$$

Using $h_1(x) \geq x^2/(2(1+x/3))$ and $h_2(x) \geq (2+x)^{-1}$ delivers

$$\begin{aligned}&\inf\left\{\sigma^2 h_1\left(\frac{t_1}{L\sigma^2}\right) + \frac{(1-\lambda)t_2^2}{2\lambda\sigma^2 L^2} h_2\left(\frac{2t_2}{\lambda\sigma^2 L}\right) \mid t_1 + t_2 = t, t_1, t_2 \geq 0\right\} \\ &\geq \inf\left\{\frac{t_1^2/(2L^2)}{\sigma^2 + \frac{Lt_1}{3}} + \frac{t_2^2/(2L^2)}{\frac{2\lambda\sigma^2}{1-\lambda} + \frac{2Lt_2}{1-\lambda}} \mid t_1 + t_2 = t, t_1, t_2 \geq 0\right\} \\ &\geq \inf\left\{\frac{(t_1+t_2)^2/(2L^2)}{\sigma^2 + \frac{Lt_1}{3} + \frac{2\lambda\sigma^2}{1-\lambda} + \frac{2Lt_2}{1-\lambda}} \mid t_1 + t_2 = t, t_1, t_2 \geq 0\right\} \\ &= \frac{t^2/(2L^2)}{\frac{1+\lambda}{1-\lambda}\cdot\sigma^2 + \frac{2Lt}{1-\lambda}},\end{aligned}$$

where the second to last line uses $t_1^2/a + t_2^2/b \geq (t_1+t_2)^2/(a+b)$ for nonnegative $t_1, t_2$ and positive $a, b$.

If $\lambda = 0$ then $(g_1+g_2)^*(t) = g_1^*(t)$ and follows from the above analogously. □