

SHARP CONCENTRATION FOR SUMS OF MATRICES WITH MARKOVIAN DEPENDENCE THROUGH UNIVERSALITY

BY ALEXANDER VAN WERDE ^{1,2,a} , JARON SANDERS ^{1,b} 

¹*Eindhoven University of Technology, Department of Mathematics and Computer Science*

²*University of Münster, Institut für Mathematische Stochastik*

^a*a.van.werde@uni-muenster.de*; ^b*jaron.sanders@tue.nl*

We prove that a sum of random matrices generated by a ψ -mixing Markov chain has similar spectral properties to a Gaussian matrix with the same mean and covariance structure. This nonasymptotic universality principle enables sharp concentration inequalities when combined with recent advances in the Gaussian literature. We illustrate the theory with examples, showing how it enables polynomial dimensional improvements relative to previous Markovian matrix concentration results when applied to Wigner-type matrices, and how one can recover sharp limiting values for a model used to study spectral clustering techniques.

A key challenge in the proof is that techniques based only on classical cumulants, which can be used when summands are independent, are not sufficient on their own for efficient estimates in a Markovian setting. Our approach exploits Boolean cumulants and a change-of-measure argument.

1. Introduction. The study of random matrices is a rich topic in probability theory with numerous deep results and connections to other fields. Such connections are aided by the phenomenon of *universality* which states that the asymptotic properties of random matrices are remarkably robust to the entries' distribution. For instance, a classical result is that if \mathbf{W} is a *Wigner matrix*—a symmetric $d \times d$ matrix with independent and identically distributed entries of mean zero and unit variance—then the operator norm satisfies $\|\mathbf{W}\|/\sqrt{d} \rightarrow 2$ as $d \rightarrow \infty$ given only mild moment assumptions, no matter the specific law of the entries [8, 18].

Classical random matrix theory is however mostly limited to asymptotic results for homogeneous matrices, such as those with identically distributed entries. For settings that do not admit an asymptotic formulation or matrices with inhomogeneous structure, the more recent theory of *matrix concentration inequalities* has achieved notable success [50]. For example, suppose that we are given self-adjoint $d \times d$ matrices $\mathbf{X}_1, \dots, \mathbf{X}_n$ that are bounded $\|\mathbf{X}_i\| \leq R$ and centered $\mathbb{E}[\mathbf{X}_i] = 0$. If these matrices are independent, then the matrix Bernstein inequality of Tropp and Oliveira [35, 50] yields a constant $c > 0$ such that

$$\mathbb{E} \left[\left\| \sum_{i=1}^n \mathbf{X}_i \right\| \right] \leq c \sqrt{\ln(d) \left\| \sum_{i=1}^n \mathbb{E}[\mathbf{X}_i^2] \right\|} + c \ln(d) R. \quad (1.1)$$

Being applicable to any matrix that can be decomposed into independent bounded summands, this result is quite flexible. It covers classical models as a special case, but also permits settings where entries are neither independent nor identically distributed.

Moreover, results with dependent summands have also been developed, further enhancing the theory's flexibility. For example, Neeman, Shi, and Ward [32] recently achieved a variant of (1.1) with Markovian dependence. Assume that there exists a stationary Markov chain Z_1, \dots, Z_n with values in some state space \mathcal{Z} as well as matrix valued functions $\mathbf{F}_i : \mathcal{Z} \rightarrow$

MSC2020 subject classifications: 60B20, 60J05, 46L53.

Keywords and phrases: Matrix concentration, Markov chain, free probability, Boolean cumulant.

$\mathbb{C}^{d \times d}$ such that the self-adjoint matrices can be expressed as $\mathbf{X}_i = \mathbf{F}_i(Z_i)$ for every $i \geq 1$. Then, the Markovian matrix Bernstein inequality in [32] implies that there exists an absolute constant $c > 0$ with

$$\mathbb{E} \left[\left\| \sum_{i=1}^n \mathbf{X}_i \right\| \right] \leq c \sqrt{\frac{\ln(d)}{1-\lambda} \sum_{i=1}^n \left\| \mathbb{E}[\mathbf{X}_i^2] \right\|} + \frac{c}{1-\lambda} \ln(d)R. \quad (1.2)$$

Here, $1 - \lambda$ is the *absolute spectral gap* of the Markov chain Z and quantifies its decay of dependence; see [32, Definition 3.1].

Such results with dependent summands have significant practical relevance as these arise in numerous applications. For instance, sums of dependent random matrices generated by a stochastic process are crucial in the analysis of time series [11, 22], randomness-efficient sampling methods [19], random walk based graph embedding methods [38], convergence rates of Hessian matrices in stochastic gradient descent [32], and state space reduction methods in reinforcement learning [24, 41]. Matrix concentration inequalities allow analyzing such settings with only minimal model-specific computations, while the dependencies involved would otherwise complicate the analysis.

The flexibility however comes at a cost: most previous results are not sharp in typical applications. For example, the bound that results if one applies (1.1) to a Wigner matrix is loose by a logarithmic dimensional factor. Moreover, note that the variance proxy $\left\| \sum_i \mathbb{E}[\mathbf{X}_i^2] \right\|$ in (1.1) is replaced by $\frac{1}{1-\lambda} \sum_i \left\| \mathbb{E}[\mathbf{X}_i^2] \right\|$ in (1.2). Moving the sum outside the operator norm is insubstantial if summands are identically distributed, but it can be highly inefficient if summands encode different matrix structure as is necessary in some applications. For instance, if (1.2) is applied to a Wigner-type matrix, then the resulting bound is loose by a factor of order $\sqrt{\ln(d+1)}\sqrt{d}$; see also Section 3.1. Not only does this give a logarithmic dimensional factor, the suboptimality is here even polynomial in the dimensionality.

1.1. Universality-based concentration. With the goal to achieve sharp concentration estimates in dependent settings, we here pursue a relatively new approach. Our main result, Theorem 2.4, establishes that a sum of matrices generated by a Markov chain has similar spectral properties to a Gaussian matrix with matching covariance structure. By combining this universality principle with results from the Gaussian literature [9, 10, 29], we then achieve practical concentration estimates in Corollaries 2.5 and 2.6. Crucially, this is done in a flexible nonasymptotic setting allowing nonhomogeneous covariance and dependent summands.

A key advantage of a universality-based approach is that it enables a leading-order term that only depends on the covariance structure of the summed matrix. Direct dependence on the summands or on the specific structure of the Markov chain only occurs through lower-order terms coming from the approximation error in the universality statement. In particular, this yields better variance proxies than are used in [11, 19, 22, 32, 38]. Moreover, by relying on recent advances in the Gaussian literature [9, 10], we can achieve a sharp estimate of the form

$$\mathbb{E} \left[\left\| \sum_{i=1}^n \mathbf{X}_i \right\| \right] \leq C_{\text{free}} \sqrt{\left\| \mathbb{E} \left[\left(\sum_{i=1}^n \mathbf{X}_i \right)^2 \right] \right\|} + \varepsilon \ln(d). \quad (1.3)$$

Here, $1 \leq C_{\text{free}} \leq 2$ is a sharp free-probabilistic constant with explicit dependence on the covariance structure of the summed matrix, and the error term ε has explicit dependence on the model parameters and is small in typical applications; see Corollary 2.5.

Universality hence enables sharp results that are inaccessible though previous results with dependencies. We achieve an optimal constant on the main term that, in particular, removes the often-suboptimal logarithmic factor if ε is sufficiently small; note that $C_{\text{free}} \leq 2$. Moreover, even if one does not care about constants or logarithms, the main term in (1.3) has a

natural variance proxy that can yield significant improvements relative to results like (1.2) in at least two ways. First, by having the sum inside the operator norm, the variance proxy $\|\mathbb{E}[(\sum_i \mathbf{X}_i)^2]\|$ can enable polynomial dimensional improvements for applications where the summands have different matrix structure. Second, moving the sum inside the square optimally¹ incorporates the dependence between summands in the way that it naturally appears in the summed matrix. In particular, this is more efficient than worst-case variance proxies of the form $D \times \|\mathbb{E}[\sum_i \mathbf{X}_i^2]\|$ with D a dependence coefficient for the Markov chain, like one based on a spectral gap; see also the discussion after Lemma 3.4.

Examples illustrating the aforementioned features are given in Section 3. We briefly consider matrices that are filled entry-wise by a Markov chain in Section 3.1, demonstrating that our results recover the optimal asymptotic order for Wigner-type matrices. Next, Section 3.2 considers an application to *block Markov chains*. The latter are a model for Markov chains with a cluster structure in the state space, and the properties of associated random matrices are crucial in the analysis of spectral clustering algorithms that recover the clusters based on an observed sample path [24, 41, 42, 54]. We use our general-purpose nonasymptotic theory to achieve improvements of results that had been derived using model-specific asymptotic analysis in [43, 44]; see Theorem 3.5 and Proposition 3.6.

1.2. Related work. Universality in classical random matrix theory has attracted significant attention, including efforts to relax the independence and homogeneity assumptions [2, 16, 46]. However, most previous results adopt a semiclassical perspective where the departure from the classical regime remains manageable. For instance, the asymptotic results in [16] allow non-identically distributed and dependent entries but adopt mean-field conditions limiting the amount of nonhomogeneity in the variance profile of the matrix as well as assumptions that impose decaying correlations; see [16, Assumptions C to E].

That universality remains valid in the nonasymptotic and nonhomogeneous framework of matrix concentration results is not explained by such semiclassical results. This is a more recent development due to Brailovskaya and Van Handel [14] who considered a setting with independent summands, and established that universality arises in such models through an operator-theoretic mechanism. Their work served as an inspiration for our investigations. As discussed in Section 1.3, however, the extension from independent summands to a Markovian setting is nontrivial due to the dependence involved. Our technical contribution is to identify the appropriate proof techniques to handle the dependence.

In the previous literature on matrix concentration inequalities with dependence, one can distinguish the following models: Markovian [19, 32, 38, 39], β or τ -mixing stochastic processes that can be non-Markovian [11, 22], martingales [6, 35, 49], exchangeable pairs [30, 36], and negative dependence [1, 4, 25, 26]. None of these previous results can access sharp bounds with a sharp constant on the main term like (1.3). Indeed, sharp bounds in the flexible setting of matrix concentration inequalities are a recent development, even for Gaussian matrices [9, 10]. Specifically, free-probabilistic upper bounds in the Gaussian setting are due to Bandeira, Boedihardjo, and Van Handel [9], and lower bounds that establish the sharpness of [9] were recently proven by Bandeira, Cipolloni, Schröder, and Van Handel [10].

¹Jensen's inequality yields that $\mathbb{E}[\|\sum_i \mathbf{X}_i\|^2] \geq \|\mathbb{E}[(\sum_i \mathbf{X}_i)^2]\|$. This shows that the leading term in (1.3) is optimal up to the constant $C_{\text{free}} \leq 2$, at least if the L^1 norm is replaced by the L^2 norm. Moreover, Corollary 2.5 gives two-sided bounds on L^p norms that show that the constant C_{free} is also optimal when $p \approx \ln(d)$. One can make it rigorous that the L^1 norm may be replaced by an L^p norm conditional on a concentration-of-measure ingredient. We presently however only have that ingredient in special cases; see the discussion after (2.12).

To our knowledge, all previous results that are applicable in Markovian settings with general matrix-valued summands involve variance proxies of order $\geq D \times \sum_i \|\mathbb{E}[\mathbf{X}_i^2]\|$ on the leading term with D a dependence coefficient [11, 19, 22, 32, 38].² As explained earlier, having the norm inside the sum even results in polynomially suboptimal results in some applications like Wigner-type matrices. In the special case where $\mathbf{X}_i = f_i(Z_i)\mathbf{A}_i$ for scalar-valued functions f_i and deterministic matrices \mathbf{A}_i , a variance proxy $D \times \|\sum_i \mathbb{E}[\mathbf{X}_i^2]\|$ that moves the sum inside the norm follows from [39, Theorem 1.3]. That result yields bounds of the correct order of magnitude for Wigner-type matrices but does not cover general matrix-valued summands and does not give a sharp constant.

Moreover, note that all the aforementioned results involve a multiplicative dependence coefficient on their variance proxy and not the natural quantity $\|\mathbb{E}[(\sum_i \mathbf{X}_i)^2]\|$ present in the leading-order term in (1.3). That dependence coefficients occur somewhere in the results is necessary, of course, since one can not expect good concentration estimates if the summands may have arbitrary dependence. In our results, however, the dependence coefficients will only occur in the error term ε in (1.3) which is often small in applications. Thus, by passing through Gaussian theory, the leading-order term in our results can incorporate the dependence in the way that it actually appears in the summed matrix. This can lead to more efficient estimates, especially if the Markov chain exhibits slow decay of dependence.

Previous results that are specific to time-homogeneous Markov chains have quantified the dependence using quantities based on a spectral gap [19, 32, 38, 39], while results that extend to time-inhomogeneous or non-Markovian settings have used β - and τ -mixing dependence coefficients [11, 22]. We here consider a Markovian setting that allows time-inhomogeneity and use a ψ -dependence coefficient; see (2.4). This dependence coefficient is sufficient for the applications we have in mind, but note that it is a strong assumption relative to aforementioned notions of dependence. It would be interesting future work if one could establish extensions of our results with weaker dependence coefficients. In particular, extensions using a spectral gap would be interesting as these have applications in theoretical computer science such as randomness-efficient sampling methods [19].

1.3. Proof techniques. Our proof involves an interpolation $\{\mathbf{S}(t) : t \in [0, 1]\}$ from the summed matrix $\sum_i \mathbf{X}_i$ to an independent Gaussian matrix with matching mean and covariance structure. Then, to show that the summed matrix and the Gaussian have similar spectral properties, it suffices to control the rate of change along the interpolation. This approach is inspired by [14] who established universality in a setting with independent summands. In their setting, the rate of change along the interpolation could be controlled using an expansion in terms of classical cumulants. Such an expansion is efficient when one assumes independence due to an *independence–implies–vanishing property* of classical cumulants. This property allows [14] to neglect all terms involving independent summands.

We however lack independence—our summands have Markovian dependence—and hence risk a combinatorial explosion in the number of terms. To solve this, one could hope that classical cumulants with approximately independent random variables are small although not necessarily zero. Such a property indeed holds true to some extent, but we found that a proof using only classical cumulants does not yield efficient estimates. This can be explained in hindsight from the fact that Markov chains are time-ordered, which classical cumulants do not respect because they are permutation invariant.

²The results in [32] are the only ones with this exact shape for the variance proxy, but the proxies used in [11, 19, 22, 38] are at least as large. Specifically, [11, 22] replace $\sum_i \|\mathbb{E}[\mathbf{X}_i^2]\|$ by $\max\{(n/\#K) \sum_{i \in K} \|\mathbb{E}[\mathbf{X}_i^2]\| : K \subseteq \{1, \dots, n\}\}$ and the results in [19, 38] only use the boundedness of the summands' operator norm as an input, without an explicit variance proxy, effectively replacing $\mathbb{E}[\mathbf{X}_i^2]$ by its worst-case bound $\|\|\mathbf{X}_i\|^2\|_{L^\infty}$.

Our approach rather relies on *Boolean cumulants*. Boolean cumulants are not as well known as classical cumulants but enjoy a natural interaction with the underlying Markovian structure, thus solving the key issue mentioned above; see Proposition 4.2. This interaction was historically first exploited by Statulevičius [45, 48].³ Our setting is however more delicate than the one studied in [45, 48] because we are concerned with matrix-valued summands. To be specific, issues arise from noncommutativity preventing one from reordering a product of matrices in a time-ordered fashion. To circumvent this, we dualize the problem in terms of the transition densities of the Markov chain and subsequently rely on a change-of-measure argument to encode the decay of dependence in scalar-valued random variables; see Proposition 4.5. This allows us to leave the matrices in the original order and is one of the critical new ideas for our approach, as it is ultimately what enables us to circumvent the difficulties for a Markovian and noncommutative setting.

Combining this idea with a classical-to-Boolean relation due to Arizmendi, Hasebe, Lehner, and Vargas [5] allows us to establish an expansion for the rate of change along the interpolation which efficiently incorporates the decay of dependence in the Markovian sequence; see Proposition 4.6. The contribution of the random variables which we used to encode the decay of dependence is here not immediately obvious. The size of these random variables is namely linked to the nontrivial combinatorics associated with the classical-to-Boolean relation from [5]. We perform an analysis of the combinatorics involved in Section 4.2.

We finally note that Boolean cumulants are also studied in the free-probabilistic literature [47]. This may lead one to believe that the free-probabilistic constant in (1.3) arises in our proof in the step where we rely on Boolean cumulants. This is however not the case. As was sketched above, Boolean cumulants allow us to exploit the Markovianity whereas the free-probabilistic main term shows up when we use the Gaussian theory of [9, 10].

1.4. *Notation.* For a real-valued random variable W , we denote $\|W\|_{L^\infty}$ for the *essential supremum* of $|W|$. We equip \mathbb{C}^d with the Euclidean norm $\|v\| := \sqrt{\langle v, v \rangle}$. For a matrix $\mathbf{M} \in \mathbb{C}^{d \times d}$ we denote $\|\mathbf{M}\| := \sup_{\|v\|=\|w\|=1} |\langle v, \mathbf{M}w \rangle|$ for the *operator norm*. Denote $\text{tr } \mathbf{M} := \frac{1}{d} \text{Tr } \mathbf{M}$ for the *normalized trace* of \mathbf{M} and let it be understood that $\text{tr } \mathbf{M}^p = \text{tr}[\mathbf{M}^p]$. The collection of all self-adjoint $d \times d$ matrices is denoted $\mathbb{C}_{\text{sa}}^{d \times d}$.

1.5. *Structure of this paper.* We state our results in Section 2. Two illustrative applications of these results are discussed in Section 3. The proofs of the main universality result and its corollaries are given in Sections 4 and 5, respectively. Some supplemental details related to the applications of our results are deferred to the appendices.

2. Results.

2.1. *Matrix models and parameters.* Consider a sequence of random variables $Z := (Z_i)_{i=1}^n$ such that Z_i takes values in a standard Borel space \mathcal{Z}_i for any $i \geq 1$. This sequence is said to be *Markovian* if for each $i > 1$ and any measurable $E \subseteq \mathcal{Z}_i$,

$$\mathbb{P}(Z_i \in E \mid Z_1, \dots, Z_{i-1}) = \mathbb{P}(Z_i \in E \mid Z_{i-1}), \quad (2.1)$$

almost surely. A sequence of self-adjoint random matrices $(\mathbf{X}_i)_{i=0}^n$ is said to come from a *Markovian model associated to Z* if \mathbf{X}_0 is a deterministic matrix and there exist measurable

³In [45, 48], Boolean cumulants are called *centered moments*.

functions $\mathbf{F}_i : \mathcal{Z}_i \rightarrow \mathbb{C}_{\text{sa}}^{d \times d}$ such that $\mathbf{X}_i = \mathbf{F}_i(Z_i)$ with $\mathbb{E}[\mathbf{X}_i] = 0$ for every $i \geq 1$. We are concerned with the sum:

$$\mathbf{S} := \mathbf{X}_0 + \sum_{i=1}^n \mathbf{X}_i. \quad (2.2)$$

Note that this is more general than the setting in the introduction insofar that \mathbf{S} may have nonzero mean and the Markovian sequence may be time-inhomogeneous. Our main result compares the spectral properties of \mathbf{S} with those of a matrix with jointly Gaussian entries.

2.1.1. *Gaussian model.* Let $\text{Cov}(\mathbf{S})$ denote the $d^2 \times d^2$ covariance matrix of the entries of the self-adjoint matrix \mathbf{S} . That is, for any $i, j, k, l \in \{1, \dots, d\}$

$$\text{Cov}(\mathbf{S})_{ij,kl} := \mathbb{E}[(\mathbf{S} - \mathbb{E}[\mathbf{S}])_{i,j} \overline{(\mathbf{S} - \mathbb{E}[\mathbf{S}])_{k,l}}]. \quad (2.3)$$

Here, \bar{z} denotes the complex conjugate of a complex number $z \in \mathbb{C}$. A self-adjoint random matrix \mathbf{G} satisfying the following properties is called a *Gaussian model of \mathbf{S}* :

1. The $2d^2$ -dimensional real-valued vector consisting of the real and imaginary parts of the entries, $\{\text{Re } \mathbf{G}_{i,j}, \text{Im } \mathbf{G}_{i,j} : i, j \in \{1, \dots, d\}\}$, is Gaussian.
2. The mean and covariance match: $\mathbb{E}[\mathbf{G}] = \mathbb{E}[\mathbf{S}]$ and $\text{Cov}(\mathbf{G}) = \text{Cov}(\mathbf{S})$.
3. The random matrices \mathbf{G} and \mathbf{S} are independent.

We next introduce parameters that are used in our results to quantify to what extent the spectral properties of \mathbf{S} are matched by those of its Gaussian model.

2.1.2. *Dependence parameter.* Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and σ -algebras $\mathcal{A}, \mathcal{B} \subseteq \mathcal{F}$, the ψ -dependence coefficient of \mathcal{A} and \mathcal{B} is defined by

$$\psi(\mathcal{A}, \mathcal{B}) := \sup_{A \in \mathcal{A}: \mathbb{P}(A) > 0} \sup_{B \in \mathcal{B}: \mathbb{P}(B) > 0} \left| \frac{\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)}{\mathbb{P}(A)\mathbb{P}(B)} \right|. \quad (2.4)$$

For two random variables V, W defined on the same probability space we denote $\psi(V, W) := \psi(\sigma(V), \sigma(W))$ for the dependence coefficient between the associated σ -algebras.

The following parameter then allows us to bound the amount of dependence in the Markovian sequence of random variables Z :

$$\Psi(Z) := \min \left\{ j \geq 1 : \psi(Z_{i+j}, Z_i) \leq \frac{1}{4} \text{ for all } i \in \{1, \dots, n-j\} \right\}. \quad (2.5)$$

Let us remark that the occurrence of $1/4$ in the definition (2.5) of $\Psi(Z)$ is not significant. Similar results can be achieved if any other number between zero and one is used.

REMARK 2.1. Another quantity that is commonly used to quantify the dependence in a Markov chain is given by the *total variation mixing time* t_{mix} [28, Section 4.5]. To apply our results, it may be useful to know that $\Psi(Z)$ can be bounded using the latter if Z is a stationary ergodic Markov chain on a finite state space. Specifically, one then has $\Psi(Z) \leq (3 + \log_2(1/\min_i \pi_i))t_{\text{mix}}$ with π the stationary distribution; see Appendix A for a proof.

2.1.3. *Matrix parameters.* All our results assume that the summands \mathbf{X}_i are bounded in operator norm. We denote $R(\mathbf{X})$ for the corresponding parameter:

$$R(\mathbf{X}) := \left\| \max_{1 \leq i \leq n} \|\mathbf{X}_i\| \right\|_{L^\infty} < \infty. \quad (2.6)$$

We further introduce the following variance proxies:

$$\sigma(\mathbf{S})^2 = \|\mathbb{E}[(\mathbf{S} - \mathbb{E}[\mathbf{S}])^2]\| \quad \text{and} \quad \varsigma(\mathbf{X})^2 := \left\| \mathbb{E} \left[\sum_{i=1}^n \mathbf{X}_i^2 \right] \right\|. \quad (2.7)$$

Note that $\varsigma(\mathbf{X}) = \sigma(\mathbf{S})$ in the special case where $\mathbf{X}_1, \dots, \mathbf{X}_n$ are independent.

Theorem 2.4 establishes that \mathbf{S} can be controlled in terms of the Gaussian model \mathbf{G} . Thereafter, we rely on results from [9, 10] which control \mathbf{G} in terms of a *free-probabilistic model* \mathbf{S}_{free} . The quality of the bounds in [9, 10] is determined by the following parameter:

$$v(\mathbf{S})^2 := \|\text{Cov}(\mathbf{S})\|. \quad (2.8)$$

The only free-probabilistic quantity which is used in the statement of our results is the norm $\|\mathbf{S}_{\text{free}}\|$. For the sake of brevity, we hence forego the precise definition⁴ of the model itself and simply note that an identity by Lehner [27], [9, Lemma 2.4] allows one to compute the norm in terms of the mean and covariance structure of \mathbf{S} as

$$\|\mathbf{S}_{\text{free}}\| = \max_{\eta \in \{+1, -1\}} \inf_{\mathbf{W} \succ 0} \lambda_{\max} \left(\mathbf{W}^{-1} + \eta \mathbb{E}[\mathbf{S}] + \mathbb{E}[(\mathbf{S} - \mathbb{E}[\mathbf{S}]) \mathbf{W} (\mathbf{S} - \mathbb{E}[\mathbf{S}])] \right), \quad (2.9)$$

where the infimum runs over deterministic positive definite $d \times d$ matrices \mathbf{W} , and λ_{\max} refers to the greatest eigenvalue.

REMARK 2.2. An estimate by Pisier [37], [9, Lemma 2.5] provides a user-friendly bound on the free-probabilistic quantity:

$$\max\{\|\mathbb{E}[\mathbf{S}]\|, \sigma(\mathbf{S})\} \leq \|\mathbf{S}_{\text{free}}\| \leq \|\mathbb{E}[\mathbf{S}]\| + 2\sigma(\mathbf{S}). \quad (2.10)$$

In particular, it follows that $\sigma(\mathbf{S}) \leq \|\mathbf{S}_{\text{free}}\| \leq 2\sigma(\mathbf{S})$ if $\mathbb{E}[\mathbf{S}] = 0$. For centered random matrices, the role of the exact formula (2.9) is hence that it enables sharp constants on the leading-order term. If one is not concerned with constant factors, then one can simply use the bounds in terms of $\sigma(\mathbf{S})$ and avoid computing $\|\mathbf{S}_{\text{free}}\|$.

REMARK 2.3. For an additional simplification of the parameters, consider the following upper bounds proven in Section A.2:

$$\sigma(\mathbf{S})^2 \leq 3\Psi(Z)\varsigma(\mathbf{X})^2 \quad \text{and} \quad v(\mathbf{S})^2 \leq 3\Psi(Z) \left\| \sum_{i=1}^n \text{Cov}(\mathbf{X}_i) \right\|. \quad (2.11)$$

These bounds are tight up to the absolute constant if the summands \mathbf{X}_i are independent. If the Markov chain mixes slowly, however, then directly using the definition of $\sigma(\mathbf{S})$ and $v(\mathbf{S})$ can sometimes yield significant improvements relative to these bounds, depending on the structure of the summed matrix; see the discussion after Lemma 3.4.

2.2. *Results.* Our main result establishes universality for the tracial moments of \mathbf{S} , showing that these can be approximated by the tracial moments of the Gaussian model:

THEOREM 2.4. *There exists an absolute constant $c > 0$ such that for every integer $p \geq 1$,*

$$\left| \mathbb{E}[\text{tr} \mathbf{S}^{2p}]^{\frac{1}{2p}} - \mathbb{E}[\text{tr} \mathbf{G}^{2p}]^{\frac{1}{2p}} \right| \leq cR(\mathbf{X})^{\frac{1}{3}} \Psi(Z)^{\frac{2}{3}} \varsigma(\mathbf{X})^{\frac{2}{3}} p^{\frac{2}{3}} + cR(\mathbf{X})\Psi(Z)p.$$

⁴The free model of a Gaussian matrix is defined in [9, (2.1)] and we let $\mathbf{S}_{\text{free}} := \mathbf{G}_{\text{free}}$ with \mathbf{G} the Gaussian model of \mathbf{S} .

Most relevant for applications of this result, one can deduce estimates on the operator norm of a random matrix given an estimate on its tracial moments by using that $\|\mathbf{M}\|^{2p}/d \leq \text{tr} \mathbf{M}^{2p} \leq \|\mathbf{M}\|^{2p}$ for any self-adjoint $d \times d$ matrix \mathbf{M} . This implies that practical bounds on $\|\mathbf{S}\|$ can be deduced whenever such bounds are available for its Gaussian model; see Lemma 5.1. For example, by combining Theorem 2.4 with recent results from [9, 10], we find the following sharp two-sided bounds:

COROLLARY 2.5. *There exists an absolute constant $C > 0$ such that for every $p \geq 1$:*

$$d^{-\frac{1}{2p}} \|\mathbf{S}_{\text{free}}\| - C\mathcal{E}(p) \leq \mathbb{E}[\|\mathbf{S}\|^{2p}]^{\frac{1}{2p}} \leq d^{\frac{1}{2p}} \|\mathbf{S}_{\text{free}}\| + Cd^{\frac{1}{2p}} \mathcal{E}(p)$$

where the error term $\mathcal{E}(p)$ is defined by

$$\mathcal{E}(p) := R(\mathbf{X})^{\frac{1}{3}} \Psi(Z)^{\frac{2}{3}} \varsigma(\mathbf{X})^{\frac{2}{3}} p^{\frac{2}{3}} + R(\mathbf{X}) \Psi(Z) p + v(\mathbf{S})^{\frac{1}{2}} \sigma(\mathbf{S})^{\frac{1}{2}} \left(p^{\frac{1}{2}} + \ln(d+1)^{\frac{3}{4}} \right).$$

Taking p a sufficiently large multiple of $\ln(d)$ in Corollary 2.5 ensures that $d^{1/2p} \leq 1 + \delta$ for some arbitrarily small $\delta > 0$. If the parameters occurring in the error term \mathcal{E} are small, as is often the case in applications of the result, then it follows that both the lower and upper bounds in Corollary 2.5 are given by $\|\mathbf{S}_{\text{free}}\|$ up to a small error. In particular, this implies the upper bound summarized in (1.3) for centered matrices since Remark 2.2 then yields that $\|\mathbf{S}_{\text{free}}\| = c\sigma(\mathbf{S})$ for some $1 < c \leq 2$ and since $\mathbb{E}[\|\mathbf{S}\|] \leq \mathbb{E}[\|\mathbf{S}\|^{2p}]^{1/2p}$ by Jensen's inequality.

Further, combining Corollary 2.5 with Markov's inequality implies that for every $\delta > 0$ there exists a constant $c > 0$ such that the following upper tail bound holds for every $x > 0$:

$$\mathbb{P}(\|\mathbf{S}\| \geq (1 + \delta)\|\mathbf{S}_{\text{free}}\| + \mathcal{E}(x)) \leq (d+1) \exp(-cx). \quad (2.12)$$

The proof details are given in Section 5. We also expect a lower bound $\|\mathbf{S}\| \geq (1 - \delta)\|\mathbf{S}_{\text{free}}\|$ to hold with high probability, but this is not immediate from Corollary 2.5 as extracting lower bounds from moments requires a separate concentration-of-measure ingredient; see Lemma 3.2 and Remark 3.3 for further discussion in a special case with summands of the form $\mathbf{X}_i = f_i(Z_i)\mathbf{B}_i$ for scalar functions f_i and deterministic matrices $\mathbf{B}_i \in \mathbb{C}_{\text{sa}}^{d \times d}$.

The price for the sharpness of the bounds is that it necessitates estimating quite a few matrix parameters when applying the results. The universality principle can however also be combined with different results for Gaussian matrices, which may be more suitable if one is not concerned with sharp constants or logarithmic dimensional factors. For instance, using the matrix Khintchine inequality of Lust-Piquard [29], [51, Section 2.3] instead of the free-probabilistic results from [9, 10] yields the following estimate removing $v(\mathbf{S})$ at the cost of a polylogarithmic factor, but still with the natural variance proxy $\sigma(\mathbf{S})$ on the main term:

COROLLARY 2.6. *Additionally assume that $\mathbb{E}[\mathbf{S}] = 0$. Then, there exists an absolute constant $c > 0$ such that*

$$\mathbb{E}[\|\mathbf{S}\|] \leq c \ln(d+1)^{1/2} \sigma(\mathbf{S}) + c R(\mathbf{X})^{\frac{1}{3}} \Psi(Z)^{\frac{2}{3}} \varsigma(\mathbf{X})^{\frac{2}{3}} \ln(d+1)^{\frac{2}{3}} + c R(\mathbf{X}) \Psi(Z) \ln(d+1).$$

So, combining the universality result with appropriately chosen bounds from the Gaussian literature can enable practical concentration estimates. We illustrated this with Corollaries 2.5 and 2.6, and one could naturally also combine Theorem 2.4 with other results from the Gaussian literature. Another potential use-case for the universality of tracial moments from Theorem 2.4 is to establish limiting laws for empirical eigenvalue distributions or empirical singular value distributions. An example illustrating this is given in Section 3.2.2.

REMARK 2.7. One can extract concentration inequalities for extremal eigenvalues from Corollary 2.5. By replacing \mathbf{S} by $\mathbf{S}' := \mathbf{S} + t\mathbf{1}$ for $t > \|\mathbf{X}_0\| + nR(\mathbf{X})$ one can namely ensure that $\|\mathbf{S}'\| = \lambda_{\max}(\mathbf{S}) + t$. Similarly, one can consider $\mathbf{S} - t\mathbf{1}$ to establish a bound on $\lambda_{\min}(\mathbf{S})$.

This trick does not apply to nonextremal eigenvalues. In this context, let us note that [14, Theorem 2.4] establishes concentration for the entire spectrum with regard to the Hausdorff distance when the summands are independent. We believe that a similar result should hold true in a Markovian setting and that Boolean cumulants would be an important ingredient in the proof. For the sake of brevity, however, we do not pursue this extension here.

3. Examples.

3.1. *Markovian entries.* We start by briefly considering a symmetric matrix with entries defined by a Markov chain. The goal is to give some intuition on the matrix parameters in a simple setting. Further, it is possible to establish two-sided tail bounds in this case.

Consider scalar random variables of the form $f_t(Z_t)$ for Z_1, \dots, Z_n a Markovian sequence and f_t real-valued functions. Suppose that $n := d(d+1)/2$ and fix a bijective function $\varphi : \{1, \dots, n\} \rightarrow \mathcal{I}$ with \mathcal{I} the set of unordered pair $\{i, j\}$ satisfying $i, j \in \{1, \dots, d\}$. Then, we can define a symmetric $d \times d$ random matrix by

$$\mathbf{S} := \sum_{t \leq n} \mathbf{X}_t \quad \text{with} \quad \mathbf{X}_t := \begin{cases} f_t(Z_t)(e_i e_j^\top + e_j e_i^\top) & \text{if } \varphi(t) = \{i, j\} \text{ with } i \neq j, \\ f_t(Z_t)e_i e_i^\top & \text{if } \varphi(t) = \{i, i\}, \end{cases} \quad (3.1)$$

where $e_1, \dots, e_n \in \mathbb{R}^d$ is the standard basis. Assume that $\mathbb{E}[f_t(Z_t)] = 0$ for all t .

LEMMA 3.1. *For (3.1), it holds that*

$$R(\mathbf{X}) = \max_{i, j \leq d} \|\mathbf{S}_{i, j}\|_{L^\infty} \quad \text{and} \quad \varsigma(\mathbf{X})^2 = \max_{i \leq d} \sum_{j \leq d} \mathbb{E}[\mathbf{S}_{i, j}^2]. \quad (3.2)$$

Further, we have $\sigma(\mathbf{S})^2 \leq 3\Psi(Z)\varsigma(\mathbf{X})^2$ and $v(\mathbf{S})^2 \leq 6\Psi(Z) \max_{i, j \leq d} \mathbb{E}[\mathbf{S}_{i, j}^2]$.

PROOF. The bounds in (3.2) are immediate from the definitions (2.6) and (2.7), where we use that the operator norm of a diagonal matrix is equal to the maximum of its entries when computing $\varsigma(\mathbf{X})^2$. The bound on $\sigma(\mathbf{S})^2$ is in (2.11).

Finally, regarding the estimate on $v(\mathbf{S})$, recall (2.11) and note that $\sum_{t \leq n} \text{Cov}(\mathbf{X}_t)$ can be written in a block diagonal form with blocks of size ≤ 2 associated with the symmetric entries (i, j) and (j, i) . Those blocks have norm $\|\text{Cov}(\mathbf{X}_t)\| \leq 2\mathbb{E}[f_t(Z_t)^2]$ with equality if $\varphi(t) = \{i, j\}$ for $i \neq j$. The estimate then follows because the norm of a block diagonal matrix is the greatest norm of its blocks. \square

For comparison, the variance proxy in the Markovian Bernstein inequality (1.2) from [32] has the following expression for (3.1):

$$\sum_{t \leq n} \|\mathbb{E}[\mathbf{X}_t^2]\| = \sum_{i=1}^d \sum_{j \geq i} \mathbb{E}[f_{\varphi^{-1}(i, j)}(Z_{\varphi^{-1}(i, j)})^2] = \sum_{i=1}^d \sum_{j \geq i} \mathbb{E}[\mathbf{S}_{i, j}^2]. \quad (3.3)$$

So, the difference between the variance proxies is that the sum over i in (3.3) is replaced by a maximum in $\varsigma(\mathbf{X})$ and $\sigma(\mathbf{S})$. This allows for bounds with good dimensional dependence in applications like Wigner-type matrices, that are inaccessible with the weaker variance proxy.

For instance, suppose that the Z_t are generated by some fixed ψ -mixing Markov chain and that $\|\mathbf{S}_{i, j}\|_{L^\infty} \leq 1$ for all $i, j \leq d$. Then, using that $\|\mathbf{S}_{\text{free}}\| \leq 2\sigma(\mathbf{S})$ by (2.10), substituting the

estimates from Lemma 3.1 in the upper bound in Corollary 2.5 yields that $\mathbb{E}[\|\mathbf{S}\|] = O(\sqrt{d})$ as $d \rightarrow \infty$. Thus, we recover the correct asymptotic order in this example: recall that $\|\mathbf{S}\| \approx 2\sqrt{d}$ in the special case where the entries are independent and identically distributed. For comparison, bounds using $\sum_t \|\mathbb{E}[\mathbf{X}_t^2]\|$ as variance proxy like (1.2) would give a suboptimal asymptotic order $O(d\sqrt{\ln(d+1)})$ which is loose by a factor of order $\sqrt{d \ln(d)}$.

The appropriate order of magnitude in this specific model could also have been extracted from earlier results in the literature such as [39, Corollary 1.4], but the sharpness of our results allows one to go further. One could now also determine an explicit constant in the big- O notation based on the free-probabilistic quantity $\|\mathbf{S}_{\text{free}}\|$. The latter does not admit a simple exact expression at this level of generality, but this may be possible in some special cases depending on the covariance structure of the entries. For instance, a simplified expression in the case where \mathbf{S} has independent entries may be found in [9, Lemma 3.2]. Whatever it may be, the free-probabilistic bound is here asymptotically tight as we have two-sided tail bounds:

LEMMA 3.2. *For (3.1), assume that $\|\mathbf{S}_{i,j}\|_{L^\infty} \leq 1$ for all i, j . Then, for every $\delta > 0$, there exist constants $c, C > 0$ such that for every $x > 0$,*

$$\mathbb{P}\left(\min_{|\gamma| \leq \delta} \|\mathbf{S}\| - (1 + \gamma)\|\mathbf{S}_{\text{free}}\| > x + C\tilde{\mathcal{E}}\right) \leq \exp\left(-\frac{cx^2}{\Psi(Z)}\right). \quad (3.4)$$

Here, $\tilde{\mathcal{E}} := \Psi(Z)^{2/3} d^{1/3} \ln(d+1)^{2/3} + \Psi(Z) \ln(d+1)$.

The proof amounts to an application of a concentration-of-measure principle by Samson [40] for convex functions of (weakly) dependent scalar random variables. The latter yields that $\|\mathbf{S}\| - \mathbb{E}[\|\mathbf{S}\|]$ has sub-Gaussian deviations after which (3.4) follows from the expectation bounds in Corollary 2.5. We refer to Appendix B for the proof details.

REMARK 3.3. Note that the tail bound in (3.4) is sub-Gaussian and dimension-independent. In particular, this yields sharper bounds on the upper tails than the one in (2.12) that followed immediately from Corollary 2.5. This reflects a broader principle: the main content in matrix concentration results like ours lies in expectation bounds, not in the immediate tail bounds. Once a bound on $\mathbb{E}[\|\mathbf{S}\|]$ like (1.3) is known, scalar theory for controlling the deviations of $\|\mathbf{S}\| - \mathbb{E}[\|\mathbf{S}\|]$ can often be used to give sharper tail bounds. For instance, [40] can also be applied to a more general *matrix series model* with summands of the form $\mathbf{X}_t = f_t(Z_t)\mathbf{B}_t$ with \mathbf{B}_t deterministic matrices. Developing analogous results for the general case $\mathbf{X}_t = \mathbf{F}_t(Z_t)$ could be relevant future work.

3.2. *Block Markov chains.* We next consider a model that is used to study clustering algorithms for sequential data, and whose spectral properties were previously studied using asymptotic and model-specific methods. Our general-purpose results can here be used to painlessly establish nonasymptotic estimates and to sharpen previous asymptotic results.

Let $d \geq K \geq 1$ be positive integers, consider a partition $(\mathcal{V}_i)_{i=1}^K$ of $\{1, \dots, d\}$ into nonempty subsets, and let $\mathbf{p} \in [0, 1]^{K \times K}$ be the transition matrix of an ergodic Markov chain on $\{1, \dots, K\}$. Then, the *block Markov chain* [41] with cluster transition matrix \mathbf{p} and clusters $(\mathcal{V}_i)_{i=1}^K$ is the Markov chain $(Z_t)_{t=1}^n$ on $\{1, \dots, d\}$ whose transition probabilities only depend on the states' clusters:

$$\mathbb{P}(Z_t = j \mid Z_{t-1} = i) = \frac{\mathbf{p}_{k,m}^{k,m}}{\#\mathcal{V}_m} \text{ for all } i \in \mathcal{V}_k \text{ and } j \in \mathcal{V}_m. \quad (3.5)$$

A schematic depiction of a block Markov chain may be found in Figure 1.

The analysis of spectral clustering algorithms [24, 41, 42, 52, 54] which recover the clusters \mathcal{V}_i based on an observed sample path crucially require concentration estimates for the *sample frequency matrix* $\hat{\mathbf{N}}$ associated with the sample path defined by

$$\hat{\mathbf{N}} := (\hat{\mathbf{N}}_{i,j})_{i,j=1}^d \quad \text{where} \quad \hat{\mathbf{N}}_{i,j} := \sum_{t=1}^{n-1} \mathbb{1}\{Z_t = i, Z_{t+1} = j\}. \quad (3.6)$$

Using a model-specific analysis, the asymptotic order of magnitude of $\|\hat{\mathbf{N}} - \mathbb{E}[\hat{\mathbf{N}}]\|$ was established in [43] and the limiting distribution of all singular values was established in [44]. We establish refinements of these results in Theorem 3.5 and Proposition 3.6.

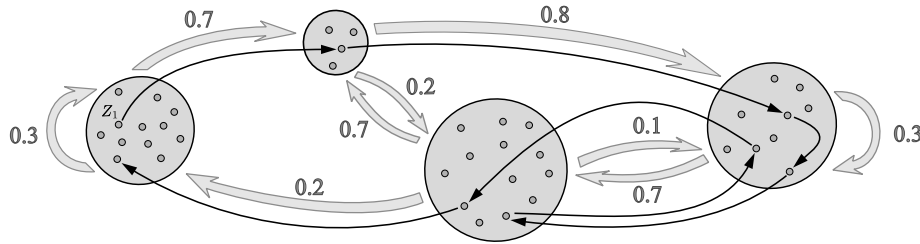


Fig 1: Visualization of a block Markov chain on $K = 4$ clusters with cluster transition matrix $\mathbf{p} = [[0.3, 0.7, 0, 0], [0, 0, 0.2, 0.8], [0.2, 0.7, 0, 0.1], [0, 0, 0.7, 0.3]]$ where we number the clusters from left to right. The thick arrows visualize the transition probabilities $\mathbf{p}_{i,j}$ and the thin arrows visualize the transitions (Z_t, Z_{t+1}) in the observed sample path $(Z_t)_{t=1}^n$.

3.2.1. *Concentration and sharp limiting value.* For simplicity, we assume that $(Z_t)_{t=1}^n$ starts in stationarity. This is to say that $\mathbb{P}(Z_1 = i) = \pi_k / \#\mathcal{V}_k$ for all $i \in \mathcal{V}_k$ and all $k \in \{1, \dots, K\}$ where $\pi \in [0, 1]^K$ is the stationary distribution associated with \mathbf{p} .

Let us denote $\mathbf{M} := \sqrt{d/n}(\hat{\mathbf{N}} - \mathbb{E}[\hat{\mathbf{N}}])$. Our goal is to estimate $\|\mathbf{M}\|$. Since the matrix \mathbf{M} is not self-adjoint with high probability, we need to consider a preliminary reduction before applying our results. We define a $2d \times 2d$ matrix, called the *self-adjoint dilation* of \mathbf{M} , by

$$\mathbf{S} := \begin{pmatrix} 0 & \mathbf{M} \\ \mathbf{M}^\top & 0 \end{pmatrix}. \quad (3.7)$$

Note that $\|\mathbf{S}\| = \|\mathbf{M}\|$. One can represent \mathbf{S} in terms of a Markovian model associated with the Markov chain of transitions $E := (E_t)_{t=1}^{n-1}$ defined by $E_t := (Z_t, Z_{t+1})$. Namely, note that with $e_i \in \mathbb{R}^d$ the i th standard basis vector we can write $\mathbf{S} = \sum_{t=1}^{n-1} \mathbf{X}_t$ with

$$\mathbf{X}_t := \sqrt{\frac{d}{n}} \sum_{i,j=1}^d (\mathbb{1}(E_t = (i, j)) - \mathbb{P}(E_t = (i, j))) \begin{pmatrix} 0 & e_i e_j^\top \\ e_j e_i^\top & 0 \end{pmatrix}. \quad (3.8)$$

Let us write $\hat{\alpha}_i := \#\mathcal{V}_i/d$ and $\hat{\alpha}_{\min} := \min_{i \leq K} \hat{\alpha}_i$. Further, let $\Psi(\mathbf{p})$ denote the ψ -mixing time for the Markov chain on $\{1, \dots, K\}$ associated with \mathbf{p} . The following lemma then provides an indication of the typical size of the matrix parameters.

LEMMA 3.4. *There exist constants $c_1, c_2, c_3 > 0$ depending only on $\hat{\alpha}_{\min}$ such that*

$$\begin{aligned} R(\mathbf{X}) &\leq c_1 \sqrt{d/n}, & \Psi(E) &\leq \Psi(\mathbf{p}) + 1, & \zeta(\mathbf{X})^2 &\leq \max\{\pi_i / \hat{\alpha}_i : i \leq K\}, \\ \sigma(\mathbf{S})^2 &\leq \max\{\pi_v / \hat{\alpha}_i : i \leq K\} + c_2 \Psi(\mathbf{p})/d, & v(\mathbf{S})^2 &\leq c_3 \Psi(\mathbf{p})/d. \end{aligned} \quad (3.9)$$

PROOF. This follows from Lemma C.5 which provides more precise estimates. \square

Note that the bounds on $\sigma(\mathbf{S})^2$ and $\varsigma(\mathbf{X})^2$ have the same leading-order contribution in (3.9) because the dependence on $\Psi(\mathbf{p})$ only occurs in a subleading term which is suppressed by a factor $1/d$. This illustrates another advantage of universality-based concentration results: the variance proxy $\sigma(\mathbf{S})^2$ incorporates how the dependence appears in the summed matrix which can be more efficient than worst-case bounds of the form $D \times \varsigma(\mathbf{X})^2$ with D a dependence coefficient such as $\Psi(\mathbf{p})$ or quantities based on a spectral gap.

Moreover, using Corollary 2.5, it follows that $\|\mathbf{S}\| \approx \|\mathbf{S}_{\text{free}}\|$. In the asymptotic regime $d \rightarrow \infty$, this allows us to refine one of the results in [42]. We let n and the clusters $(\mathcal{V}_k)_{k=1}^K$ depend on d but assume that the cluster transition matrix \mathbf{p} is kept fixed. Further, assume that there exist strictly positive numbers $\alpha_1, \dots, \alpha_k > 0$ such that $\lim_{d \rightarrow \infty} \#\mathcal{V}_k/d = \alpha_k$ for every k . Then, in the regime where $n \gg d \ln(d)^4$, the following theorem improves upon [42, Theorem 3] in the fact that we can determine the exact limiting value whereas [42] only proves a nonexplicit upper bound by characterizing the asymptotic order. Let us note however that [42] has a weaker assumption: it is there assumed that $n \gg d \ln(d)$.

THEOREM 3.5. *Assume that $\lim_{d \rightarrow \infty} d \ln(d)^4/n = 0$. Then, the random variable $\|\mathbf{M}\|$ converges in probability to the scalar $\mathfrak{m} > 0$ defined by*

$$\mathfrak{m} := \inf_{x \in \mathbb{R}_{>0}^{2K}} \max_{i=1, \dots, 2K} \left\{ \frac{1}{x_i} + \sum_{j=1}^{2K} c_{i,j} x_j \right\} \quad (3.10)$$

where the infimum runs over all vectors x with strictly positive coordinates and the coefficients $(c_{i,j})_{i,j=1}^{2K}$ are defined by

$$c_{i,j} := \begin{cases} 0 & \text{if } i \leq K \text{ and } j \leq K, \\ \alpha_i^{-1} \pi_i \mathbf{p}_{i,j-K} & \text{if } i \leq K \text{ and } j > K, \\ 0 & \text{if } i > K \text{ and } j > K, \\ \alpha_{i-K}^{-1} \pi_j \mathbf{p}_{j,i-K} & \text{if } i > K \text{ and } j \leq K. \end{cases}$$

The proof of this result is given in Section C.3 and relies on Corollary 2.5 to establish an upper bound on $\|\mathbf{M}\|$. Note, however, that we here not only provide an upper bound but rather the exact limiting value. The proof of this two-sidedness relies on our next result Proposition 3.6 which implies, in particular, that there is asymptotically an abundance of singular values which are close to the upper bound from Corollary 2.5. This is visualized by Figure 2 in Section 3.2.3 below.

3.2.2. Limiting singular value distribution. Universality of tracial moments can also be used to establish limiting laws for the singular value distribution. We start by introducing some terminology. The i th largest singular value of a square matrix $\mathbf{M} \in \mathbb{C}^{d \times d}$ can be defined in terms of the i th largest eigenvalue of $\mathbf{M}\mathbf{M}^*$ as $s_i(\mathbf{M}) := (\lambda_i(\mathbf{M}\mathbf{M}^*))^{1/2}$. The *singular value distribution* of \mathbf{M} is then the probability measure $\nu_{\mathbf{M}}$ defined by

$$\nu_{\mathbf{M}}([a, b]) := \frac{1}{d} \#\{i \in \{1, \dots, d\} : a \leq s_i(\mathbf{M}) \leq b\}. \quad (3.11)$$

A sequence of random probability measures ν_n is said to converge *weakly in probability* to a deterministic probability measure ν if $\int f(x) d\nu_n(x)$ converges in probability to $\int f(x) d\nu(x)$ for every continuous bounded function $f : \mathbb{R} \rightarrow \mathbb{R}$.

The *Stieltjes transform* of a probability measure ν on \mathbb{R} is the analytic function $s : \mathbb{C}^+ \rightarrow \mathbb{C}^-$ given by $s(z) := \int 1/(z-x)d\nu(x)$ where $\mathbb{C}^+ := \{z \in \mathbb{C} : \text{Im}(z) > 0\}$ is the upper half-plane and $\mathbb{C}^- := \{z \in \mathbb{C} : \text{Im}(z) < 0\}$ is the lower half-plane. The measure can be recovered from its Stieltjes transform using the Stieltjes inversion formula [7, Theorem B.8] which states that for any continuity points $a < b$ of ν ,

$$\nu([a, b]) = -\frac{1}{\pi} \lim_{\varepsilon \rightarrow 0^+} \int_a^b \text{Im}(s(x + \sqrt{-1}\varepsilon)). \quad (3.12)$$

Let us warn that [7] employs a definition for the Stieltjes transform which differs from the one used here by a minus sign. Finally, the *symmetrization* of a measure ν on $\mathbb{R}_{\geq 0}$ is the measure $\text{sym}(\nu)$ on \mathbb{R} given by $A \mapsto (\nu(A \cap \mathbb{R}_{\geq 0}) + \nu((-A) \cap \mathbb{R}_{\geq 0}))/2$ where A ranges over all measurable subsets of \mathbb{R} and $-A := \{-a : a \in A\}$.

The following results improve upon [44, Theorem 1.1]. In [44] it was namely assumed that $n \approx cd^2$ for some constant $c > 0$ whereas the following result only requires that $n \gg d$. This relaxed assumption allows short sample paths which give rise to a sparser sample frequency matrix. The assumption that $n \gg d$ is optimal, as the result has to fail if $n \approx cd$ for fixed $c > 0$. For instance, suppose that $K = 1$ so that the sequence Z_1, \dots, Z_n consists of independent and identically distributed random variables. Then, it follows from $n \approx cd$ that there exists some $u > 0$ such that the number of states in \mathcal{V} which are not visited by Z_t is at least ud with high probability. Then, \mathbf{M} has at least ud rows equal to zero which implies that $\nu_{\mathbf{M}}(0) \geq u$. On the other hand, when $K = 1$, the distribution $\text{sym}(\nu_{\infty})$ in Proposition 3.6 is the semicircular law and hence continuous. Considering $\int f(x)d\nu_{\mathbf{M}}(x)$ with $f : \mathbb{R} \rightarrow \mathbb{R}$ supported in a small neighborhood of 0 then shows that $\nu_{\mathbf{M}}$ does not converge weakly in probability to ν_{∞} .

PROPOSITION 3.6. *Assume that $\lim_{d \rightarrow \infty} d/n = 0$. Then, $\nu_{\mathbf{M}}$ converges weakly in probability to a compactly supported probability measure ν_{∞} on $\mathbb{R}_{\geq 0}$. Moreover, $\text{sym}(\nu_{\infty})$ has Stieltjes transform $s(z) = \sum_{i=1}^K \alpha_i (a_i(z) + a_{K+i}(z))/2$ where a_1, \dots, a_{2K} are the unique analytic functions from \mathbb{C}^+ to \mathbb{C}^- such that the following system of equations is satisfied*

$$\begin{aligned} a_i(z)^{-1} &= z - \sum_{j=1}^K \alpha_i^{-1} \pi_i \mathbf{p}_{i,j} a_{K+j}(z) \\ a_{i+K}(z)^{-1} &= z - \sum_{j=1}^K \alpha_i^{-1} \pi_j \mathbf{p}_{j,i} a_j(z) \end{aligned}$$

for $i = 1, \dots, K$.

COROLLARY 3.7. *Assume that $\lim_{d \rightarrow \infty} d/n = 0$ and let ν_{∞} be as in Proposition 3.6. Then, $\nu_{\sqrt{d/n}\hat{\mathbf{N}}}$ converges weakly in probability to ν_{∞} .*

The proof is given in Section C.2 and relies on the moment universality from Theorem 2.4. It is interesting to note that universality gives a simpler proof than in [44]. The main difficulty in [44] is namely to show that all joint moments of the entries of \mathbf{M} behave approximately as in the independent case. For the second moments, i.e., covariance, this is not too difficult [44, Proposition 4.8]. Estimating higher-order joint moments is however significantly more technical; see [44, Section 6.3.4]. A universality-based approach allows bypassing this technical step because the higher moments of a Gaussian are determined by its covariance.

REMARK 3.8. The quantitative bounds in Theorem 2.4 are only for tracial moments of even order. This is sufficient for singular value distributions, but other applications such as

eigenvalue distributions of self-adjoint random matrices also requires tracial moments of odd order. In this context, note that $\mathbb{E}[\text{tr}(\mathbf{S} + t\mathbf{1})^{2p}] = \sum_{k=0}^{2p} \binom{2p}{k} t^{2p-k} \mathbb{E}[\text{tr} \mathbf{S}^k]$. Since pointwise convergence of a polynomial implies convergence of coefficients, this also allows extracting asymptotic universality for odd tracial moments from Theorem 2.4. A similar trick applied to $(\mathbf{S} \otimes \mathbf{1} + t\mathbf{1} \otimes \mathbf{S})^{2p}$ yields universality for the variance of tracial moments; see the proof of Lemma C.7 in Section C for details.

3.2.3. Visualization of results. Figure 2 visualizes Theorem 3.5 and Proposition 3.6. Observe that the edge of the support of the empirical singular value distribution is visually indistinguishable from $\|\mathbf{S}_{\text{free}}\|$. This illustrates the sharpness of the leading-order term: recall that the greatest singular value of a matrix corresponds to its operator norm. Such sharp leading-order terms are inaccessible with previous general-purpose matrix concentration results with dependencies.

The experimental setup for this figure corresponds to the block Markov chain visualized in Figure 1. More precisely, we sampled a trajectory from a block Markov chain with $K = 4$, $\mathbf{p} = [[0.3, 0.7, 0, 0], [0, 0, 0.2, 0.8], [0.2, 0.7, 0, 0.1], [0, 0, 0.7, 0.3]]$, $(\alpha_1, \alpha_2, \alpha_3, \alpha_4) = (0.2, 0.1, 0.4, 0.3)$, $d = 10000$, and $n = 100d$. The system of equations in Proposition 3.6 was solved using the algorithm of [23] as implemented in `BMCToolkit` [53].

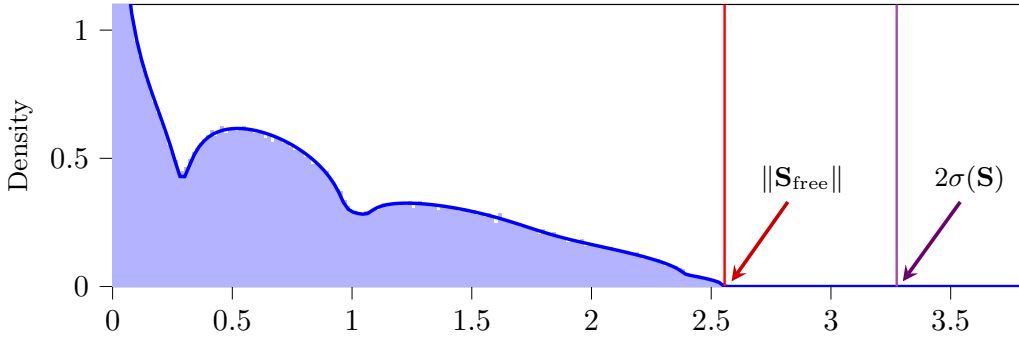


Fig 2: The bars display the empirical singular value distribution of the recentered and normalized sample frequency matrix \mathbf{M} for a sample path of a block Markov chain. The blue line displays the asymptotic theoretical prediction from Proposition 3.6. The red and purple vertical lines respectively display $\|\mathbf{S}_{\text{free}}\|$ and $2\sigma(\mathbf{S})$ with \mathbf{S} as in (3.7).

4. Proof of Theorem 2.4. Recall that Theorem 2.4 is a universality principle stating that the tracial moments of \mathbf{S} are well-approximated by those of \mathbf{G} . To establish this universality statement, we interpolate using

$$\mathbf{S}(t) := \mathbb{E}[\mathbf{S}] + \sqrt{t}(\mathbf{S} - \mathbb{E}[\mathbf{S}]) + \sqrt{1-t}(\mathbf{G} - \mathbb{E}[\mathbf{G}]) \quad (4.1)$$

for $t \in [0, 1]$. The reason for using weights \sqrt{t} and $\sqrt{1-t}$ in the interpolation is to ensure that the covariance of the entries of $\mathbf{S}(t)$ remains constant as t varies. In order to prove that $\mathbb{E}[\text{tr} \mathbf{S}^{2p}]^{1/2p} \approx \mathbb{E}[\text{tr} \mathbf{G}^{2p}]^{1/2p}$, it is now sufficient to show that $|\frac{d}{dt} \mathbb{E}[\text{tr} \mathbf{S}(t)^{2p}]|$ is small.

Section 4.1 establishes an exact expansion for $\frac{d}{dt} \mathbb{E}[g(\mathbf{S}(t))]$ where $g: \mathbb{C}^{d \times d} \rightarrow \mathbb{C}$ is a polynomial map. This involves combinatorics associated with classical-to-Boolean cumulant relations from [5] whose contribution we estimate in Section 4.2. Further, the individual terms in the expansion are defined in terms of directional derivatives of g . We estimate these individual terms using trace inequalities in Section 4.3. These ingredients are finally combined to bound $\frac{d}{dt} \mathbb{E}[\text{tr} \mathbf{S}(t)^{2p}]$ in Section 4.4, concluding the proof.

4.1. Expansion for the rate-of-change along the interpolation.

4.1.1. *Boolean joint cumulants.* The *Boolean joint cumulant* $b(Y_1, \dots, Y_k)$ of a sequence of bounded real random variables Y_1, \dots, Y_k is defined in terms of the joint moments as

$$b(Y_1, \dots, Y_k) := \sum_{m=0}^{k-1} \sum_{1 \leq j_1 < \dots < j_m \leq k-1} (-1)^m \mathbb{E}[Y_1 \cdots Y_{j_1}] \mathbb{E}[Y_{j_1+1} \cdots Y_{j_2}] \cdots \mathbb{E}[Y_{j_{m+1}} \cdots Y_k]. \quad (4.2)$$

Let us warn that Boolean cumulants are *not* permutation invariant. That is, it can occur that $b(Y_1, \dots, Y_k) \neq b(Y_{\rho(1)}, \dots, Y_{\rho(k)})$ for some permutation $\rho \in \mathcal{S}_k$. This warning is relevant in some of the subsequent proofs where we have to ensure that the random variables occur in the appropriate order to be able to exploit the decay of dependence in the underlying Markovian sequence (Z_1, \dots, Z_n) .

For any sequence of real values $i = (i_1, \dots, i_k)$ let us denote $\text{sort}(i)$ for the unique sequence of length k which is a nondecreasing permutation of i :

$$\min\{i_1, \dots, i_k\} = \text{sort}(i)_1 \leq \text{sort}(i)_2 \leq \dots \leq \text{sort}(i)_k = \max\{i_1, \dots, i_k\}. \quad (4.3)$$

For a nonempty subset $\mathcal{J} \subseteq \{1, \dots, k\}$ and $1 \leq \ell \leq \#\mathcal{J}$ we denote $\mathcal{J}(\ell) := \text{sort}((j)_{j \in \mathcal{J}})_\ell$ for the ℓ th smallest element in \mathcal{J} . We then denote $b(Y_j : j \in \mathcal{J})$ for the Boolean cumulant associated to \mathcal{J} with indices in increasing order:

$$b(Y_j : j \in \mathcal{J}) := b(Y_{\mathcal{J}(1)}, Y_{\mathcal{J}(2)}, \dots, Y_{\mathcal{J}(k)}). \quad (4.4)$$

Denote \mathcal{S}_k and \mathcal{P}_k for the sets consisting of all permutations or partitions of $\{1, \dots, k\}$, respectively. An (*increasing*) *run* in a permutation $\rho \in \mathcal{S}_k$ is an increasing subsegment of $(\rho(1), \dots, \rho(k))$ of maximal length. Here, a *subsegment* is a sequence of the form $(\rho(i), \rho(i+1), \rho(i+2), \dots, \rho(i+\ell))$. We denote $\pi_\rho \in \mathcal{P}_k$ for the partition of $\{1, \dots, k\}$ consisting of the runs. That is, for $i < j$ it holds that $i \sim j$ in the equivalence relation induced by π_ρ if and only if $i = \rho(r) < \rho(r+1) < \dots < \rho(r+\ell) = j$ for some r, ℓ . See Figure 3 for a visualization of a permutation ρ and the associated partition π_ρ .

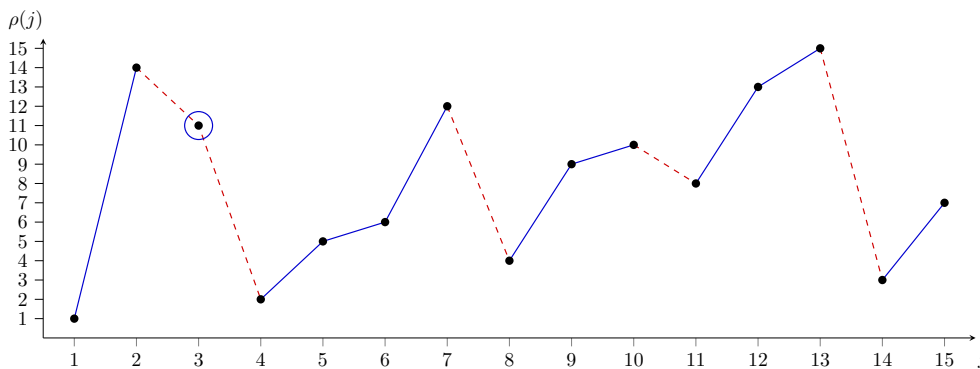


Fig 3: This figure displays a permutation $\rho \in \mathcal{S}_{15}$. Every part in the induced partition of runs π_ρ here corresponds to a connected component of the solid blue lines by way of the y -values in this connected component. The displayed permutation is given by $(\rho(1), \rho(2), \dots, \rho(15)) = (1, 14, 11, 2, 5, 6, 12, 4, 9, 10, 8, 13, 15, 3, 7)$. The induced partition of runs is given by $\pi_\rho = \{\{1, 14\}, \{11\}, \{2, 5, 6, 12\}, \{4, 9, 10\}, \{8, 13, 15\}, \{3, 7\}\}$.

PROPOSITION 4.1. *Let V_1, \dots, V_n be a sequence of, possibly dependent, centered and bounded \mathbb{R}^D -valued random vectors and consider a sequence of centered Gaussian random vectors satisfying that $\mathbb{E}[G_{i_1} G_{i_2}^\top] = \mathbb{E}[V_{i_1} V_{i_2}^\top]$ for any $i_1, i_2 \in \{1, \dots, n\}$. Assume that $\mathbf{V} := (V_1, \dots, V_n)^\top$ and $\mathbf{G} := (G_1, \dots, G_n)^\top$ are independent and let $\mathbf{V}(t) := \sqrt{t} \mathbf{V} + \sqrt{1-t} \mathbf{G}$. Then, for any polynomial $g : \mathbb{R}^{n \times D} \rightarrow \mathbb{C}$ and any $t \in [0, 1]$*

$$\begin{aligned} \frac{d}{dt} \mathbb{E}[g(\mathbf{V}(t))] &= \frac{1}{2} \sum_{k=3}^{\infty} \frac{t^{\frac{k}{2}-1}}{(k-1)!} \sum_{i \in \{1, \dots, n\}^k} \sum_{\rho \in \mathcal{S}_k : \rho(1)=1} (-1)^{\#\pi_\rho - 1} \sum_{\gamma \in \{1, \dots, D\}^k} \\ &\times \left\{ \prod_{\mathcal{J} \in \pi_\rho} b(V_{\text{sort}(i)_j, \gamma_j} : j \in \mathcal{J}) \right\} \mathbb{E} \left[\frac{\partial^k g}{\partial v_{\text{sort}(i)_1, \gamma_1} \cdots \partial v_{\text{sort}(i)_k, \gamma_k}}(\mathbf{V}(t)) \right]. \end{aligned}$$

PROOF. We will use an expansion in terms of classical cumulants from [14, Theorem 4.3] as our starting point and rephrase by using a classical-to-Boolean formula from [5, Corollary 1.6]. In this context, let us note that the classical cumulant of a sequence of bounded random variables Y_1, \dots, Y_k is defined by

$$\kappa(Y_1, \dots, Y_k) := \sum_{\pi \in \mathcal{P}_k} (-1)^{\#\pi - 1} (\#\pi - 1)! \prod_{\mathcal{J} \in \pi} \mathbb{E} \left[\prod_{j \in \mathcal{J}} Y_j \right]. \quad (4.5)$$

Note that it is immediate from (4.5) that classical cumulants are permutation invariant, meaning that $\kappa(Y_1, \dots, Y_k) = \kappa(Y_{\rho(1)}, \dots, Y_{\rho(k)})$ for all $\rho \in \mathcal{S}_k$.

We can also view $\mathbf{V}(t)$ as an nD -dimensional random vector whose entries are indexed by tuples $(i, \gamma) \in \{1, \dots, n\} \times \{1, \dots, D\}$. Hence, by the special case of the classical cumulant expansion [14, Theorem 4.3] for a *single* nD -dimensional random vector,

$$\begin{aligned} \frac{d}{dt} \mathbb{E}[g(\mathbf{V}(t))] & \\ &= \frac{1}{2} \sum_{k=3}^{\infty} \frac{t^{\frac{k}{2}-1}}{(k-1)!} \sum_{i \in \{1, \dots, n\}^k} \sum_{\gamma \in \{1, \dots, D\}^k} \kappa(V_{i_1, \gamma_1}, \dots, V_{i_k, \gamma_k}) \mathbb{E} \left[\frac{\partial^k g}{\partial v_{i_1, \gamma_1} \cdots \partial v_{i_k, \gamma_k}}(\mathbf{V}(t)) \right] \end{aligned} \quad (4.6)$$

for any $t \in [0, 1]$. Hence, since $\text{sort}(i)$ is a permutation of i and since classical cumulants are permutation invariant,

$$\begin{aligned} &\sum_{\gamma \in \{1, \dots, D\}^k} \kappa(V_{i_1, \gamma_1}, \dots, V_{i_k, \gamma_k}) \mathbb{E} \left[\frac{\partial^k g}{\partial v_{i_1, \gamma_1} \cdots \partial v_{i_k, \gamma_k}}(\mathbf{V}(t)) \right] \\ &= \sum_{\gamma \in \{1, \dots, D\}^k} \kappa(V_{\text{sort}(i)_1, \gamma_1}, \dots, V_{\text{sort}(i)_k, \gamma_k}) \mathbb{E} \left[\frac{\partial^k g}{\partial v_{\text{sort}(i)_1, \gamma_1} \cdots \partial v_{\text{sort}(i)_k, \gamma_k}}(\mathbf{V}(t)) \right]. \end{aligned} \quad (4.7)$$

We next apply the classical-to-Boolean formula from [5, Corollary 1.6].

This formula involves quantities $B_{\text{runs}(\rho)}$ and $d(\rho)$ where ρ is a permutation. In our notation, it holds that $\text{runs}(\rho) = \pi_\rho$ and $d(\rho) = \#\pi_\rho - 1$ due to [5, Section 7, page 79, item (1)]. Further, for any partition π , the quantity B_π is defined in terms of a product of Boolean joint cumulants in [5, Section 2, page 62]. Let us finally remark that our definitions (4.5) and (4.2) for classical and Boolean joint cumulants agree with the definitions used in [5]: take $\pi = \{\{1, \dots, k\}\}$ and $\varphi = \mathbb{E}$ in [5, (2.10) & (2.12)]. Hence, in our notation, [5, Corollary 1.6] states that

$$\kappa(V_{\text{sort}(i)_1, \gamma_1}, \dots, V_{\text{sort}(i)_k, \gamma_k}) = \sum_{\rho \in \mathcal{S}_k : \rho(1)=1} (-1)^{\#\pi_\rho - 1} \prod_{\mathcal{J} \in \pi_\rho} b(V_{\text{sort}(i)_j, \gamma_j} : j \in \mathcal{J}). \quad (4.8)$$

Combine (4.6)–(4.8) to find the desired result. \square

The reason why Boolean cumulants are useful in our Markovian setting is due to the following expression. This identity can also be found in [45, (1.62)].

PROPOSITION 4.2. *Consider a Markovian sequence of random variables W_1, \dots, W_k with values in standard Borel spaces $\mathcal{W}_1, \dots, \mathcal{W}_k$ as well as measurable functions $g_i : \mathcal{W}_i \rightarrow \mathbb{R}$. Then, denoting $Y_i = g_i(W_i)$ for any $i \in \{1, \dots, k\}$,*

$$b(Y_1, \dots, Y_k) = \int_{\mathcal{W}_1} \cdots \int_{\mathcal{W}_k} \left\{ \prod_{i=2}^k g_i(w_i) (\mathrm{d}\mathbb{P}_{W_i|W_{i-1}=w_{i-1}}(w_i) - \mathrm{d}\mathbb{P}_{W_i}(w_i)) \right\} g_1(w_1) \mathrm{d}\mathbb{P}_{W_1}(w_1). \quad (4.9)$$

PROOF. Expand the right-hand-side of (4.9) into a sum with 2^{k-1} terms and note that each of these terms corresponds to one of the terms in (4.2). For instance the term $(-1) \int_{\mathcal{W}_1} \int_{\mathcal{W}_2} \int_{\mathcal{W}_3} g_1(w_1) g_2(w_2) g_3(w_3) \mathrm{d}\mathbb{P}_{W_3}(w_3) \mathrm{d}\mathbb{P}_{W_2|W_1=w_1}(w_2) \mathrm{d}\mathbb{P}_{W_1}(w_1)$ in the expansion of (4.9) when $k = 3$ corresponds to the term $(-1) \mathbb{E}[Y_1 Y_2] \mathbb{E}[Y_3]$ in (4.2). \square

4.1.2. Properties of the ψ -dependence coefficient. Note the factors with $i \geq 2$ in the Boolean cumulant expression (4.9) provide a suppression when $\mathbb{P}_{W_i|W_{i-1}=w_{i-1}} \approx \mathbb{P}_{W_i}$. That is, the Boolean cumulant tends to be small if the Markovian sequence is almost a sequence of independent random variables. We will exploit this fact together with our assumption that the ψ -dependence in the Markov chain is decaying by using the following properties:

PROPOSITION 4.3. *Assume that V, W are random variables taking values in standard Borel spaces such that $\psi(V, W) < \infty$. Then, $\mathbb{P}_{V, W}$ is absolutely continuous with respect to $\mathbb{P}_V \otimes \mathbb{P}_W$, and for $(\mathbb{P}_V \otimes \mathbb{P}_W)$ -almost every (v, w) ,*

$$1 - \psi(V, W) \leq \frac{\mathrm{d}\mathbb{P}_{V, W}}{\mathrm{d}(\mathbb{P}_V \otimes \mathbb{P}_W)}(v, w) \leq 1 + \psi(V, W).$$

PROPOSITION 4.4. *Let Z be a Markovian sequence of random variables. Then, for any $i \in \{1, \dots, n\}$ and $j \geq \Psi(Z)$,*

$$\psi(Z_{i+j}, Z_i) \leq \left(\frac{1}{4}\right)^{\lfloor j/\Psi(Z) \rfloor}.$$

Both properties follow from the definition (2.4) by direct computations using the Radon-Nikodym theorem and the Markov property. This is classical, dating back to the 1963 work of Blum, Hanson, and Koopmans [12, Pages 8–10], so we omit the details.⁵

4.1.3. Encoding suppression in random variables. If we were interested in results for real-valued random variables, then we could now exploit our assumptions regarding decay of dependence by applying Hölder's inequality to (4.9); see e.g., [45, Chapter 4] and [17, Section 10] for such arguments. Our primary interest is however in random matrices, and we only pass by the real-valued random variables as an intermediate stage. This makes it so that we have to postpone the application of estimates. To this end, we will encode the decay of dependence in scalar-valued random variable which allows us to maintain exact equalities, leading to a practical expansion for the rate of change along the interpolation in Proposition 4.6.

⁵Detailed computations can also be found in the first arXiv version of this paper; see arXiv:2307.11632v1.

PROPOSITION 4.5. *Let W_1, \dots, W_k be a Markovian sequence of random variables taking values in standard Borel spaces $\mathcal{W}_1, \dots, \mathcal{W}_k$, respectively. Then, there exist random variables W'_1, \dots, W'_k and Δ' with the following three properties:*

- (1) **Marginal distribution:** *For any fixed $j \leq k$, it holds that W'_j has the same law as W_j .*
(2) **Suppression:** *The random variable Δ' takes values in \mathbb{R} and satisfies that almost surely*

$$|\Delta'| \leq 2^{k-1} \prod_{j=2}^k \min\{1, \psi(W_j, W_{j-1})\}.$$

When $k = 1$, this bound should be understood as the statement that $|\Delta'| \leq 1$.

- (3) **Expression for Boolean cumulants:** *For any sequence of real random variables Y_1, \dots, Y_k associated to the Markovian sequence, i.e., with $Y_i = g_i(W_i)$ for certain $g_i : \mathcal{W}_i \rightarrow \mathbb{R}$,*

$$b(Y_1, \dots, Y_k) = \mathbb{E}[\Delta' Y'_1 \cdots Y'_k],$$

where $Y'_i := g_i(W'_i)$.

PROOF. Let us partition the index set as $\{1, \dots, k\} = \{1\} \cup \mathcal{I}_1 \cup \mathcal{I}_2$ with

$$\mathcal{I}_1 := \{i \geq 2 : \psi(W_i, W_{i-1}) < 1\} \text{ and } \mathcal{I}_2 := \{i \geq 2 : \psi(W_i, W_{i-1}) \geq 1\}. \quad (4.10)$$

The general idea of the subsequent argument is to execute a change-of-measure on all factors with $i \in \mathcal{I}_1$ in (4.9) and to expand all remaining factors in a summation. The random variable Δ' will then arise from the Radon–Nikodym derivative in the change-of-measure.

We start by introducing some notation for bookkeeping purposes. For any $i \in \{1, \dots, k\}$ and $\alpha_i \in \{0, 1\}$, let $\mathbb{Q}_{W_i|W_{i-1}}^{(\alpha_i)} : \mathcal{B}(\mathcal{W}_i) \times \mathcal{W}_{i-1} \rightarrow \mathbb{R}$ denote the regular conditional probability measure defined for every measurable $E \subseteq \mathcal{W}_i$ and $w_{i-1} \in \mathcal{W}_{i-1}$ by

$$\mathbb{Q}_{W_i|W_{i-1}=w_{i-1}}^{(\alpha_i)}(E) := \begin{cases} \mathbb{P}_{W_i}(E) & \text{if } i \in \mathcal{I}_1 \cup \{1\}, \\ \mathbb{P}_{W_i|W_{i-1}=w_{i-1}}(E) & \text{if } \alpha_i = 0 \text{ and } i \in \mathcal{I}_2, \\ \mathbb{P}_{W_i}(E) & \text{if } \alpha_i = 1 \text{ and } i \in \mathcal{I}_2. \end{cases} \quad (4.11)$$

Then, expanding the factors with $i \in \mathcal{I}_2$ in the expression (4.9) for the Boolean cumulant into a sum yields that

$$b(Y_1, \dots, Y_k) = \sum_{\alpha_i \in \{0,1\}: i \in \{1\} \cup \mathcal{I}_2} \int_{\mathcal{W}_1} \cdots \int_{\mathcal{W}_k} \frac{(-1)^{\#\{i \in \mathcal{I}_2: \alpha_i = 1\}}}{2} \times \prod_{i \in \{1\} \cup \mathcal{I}_2} g_i(w_i) d\mathbb{Q}_{W_i|W_{i-1}=w_{i-1}}^{(\alpha_i)}(w_i) \prod_{i \in \mathcal{I}_1} g_i(w_i) (d\mathbb{P}_{W_i|W_{i-1}=w_{i-1}}(w_i) - d\mathbb{P}_{W_i}(w_i)). \quad (4.12)$$

The factor $1/2$ here accounts for double counting: recall that the factor with $i = 1$ in (4.9) does not involve a difference of conditional and unconditional probability measures.

We next apply a change-of-measure to encode the decay of dependence. For any $i \in \mathcal{I}_1$ let us define a measurable function $\delta_i : \mathcal{W}_i \times \mathcal{W}_{i-1} \rightarrow \mathbb{R}$ by

$$\delta_i(w_i, w_{i-1}) := \frac{d\mathbb{P}_{W_i, W_{i-1}}}{d(\mathbb{P}_{W_i} \otimes \mathbb{P}_{W_{i-1}})}(w_i, w_{i-1}) - 1. \quad (4.13)$$

Note that (4.13) is well-defined due to Proposition 4.3 and that for $(\mathbb{P}_{W_i} \otimes \mathbb{P}_{W_{i-1}})$ -almost every w_i, w_{i-1} ,

$$-\psi(W_i, W_{i-1}) \leq \delta_i(w_i, w_{i-1}) \leq \psi(W_i, W_{i-1}). \quad (4.14)$$

Further, by the definition (4.13) of δ_i in terms of a Radon–Nikodym derivative, we have that $d\mathbb{P}_{W_i|W_{i-1}=w_{i-1}}(w_i) - d\mathbb{P}_{W_i}(w_i) = \delta_i(w_i, w_{i-1})d\mathbb{P}_{W_i}(w_i)$. Hence, using this on the factors with $i \in \mathcal{I}_1$ in (4.12) and subsequently using that $\mathbb{P}_{W_i} = \mathbb{Q}_{W_i|W_{i-1}}^{(\alpha_i)}$ for any $i \in \mathcal{I}_1$,

$$b(Y_1, \dots, Y_k) = \sum_{\alpha_1, \dots, \alpha_k \in \{0,1\}} \int_{\mathcal{W}_1} \cdots \int_{\mathcal{W}_k} \delta(w_1, \dots, w_k) \prod_{i=1}^k g_i(w_i) d\mathbb{Q}_{W_i|W_{i-1}=w_{i-1}}^{(\alpha_i)}(w_i), \quad (4.15)$$

where $\delta : \mathcal{W}_1 \times \dots \times \mathcal{W}_k \rightarrow \mathbb{R}$ is defined by

$$\delta(w_1, \dots, w_k) := \frac{(-1)^{\#\{i \in \mathcal{I}_2 : \alpha_i = 1\}}}{2^{\#\mathcal{I}_1 + 1}} \prod_{i \in \mathcal{I}_1} \delta_i(w_i, w_{i-1}). \quad (4.16)$$

The additional factor $(1/2)^{\#\mathcal{I}_1}$ relative to (4.12) accounts for double counting: note that the sum in (4.12) runs only over α_i with $i \in \{1\} \cup \mathcal{I}_2$, while (4.15) also allows $i \in \mathcal{I}_1$.

For any $\alpha = (\alpha_1, \dots, \alpha_k)$ as in (4.15), let us define $(W_1^{(\alpha)}, \dots, W_k^{(\alpha)})$ to be random variables with joint law given by $\prod_{i=1}^k d\mathbb{Q}_{W_i|W_{i-1}}^{(\alpha_i)}$. Further, define $\Delta^{(\alpha)} := \delta(W_1^{(\alpha)}, \dots, W_k^{(\alpha)})$. Recall from statement (3) of Proposition 4.5 that $Y_i = g_i(W_i)$. Hence, (4.15) yields that

$$b(Y_1, \dots, Y_k) = \sum_{\alpha \in \{0,1\}^k} \mathbb{E}[\Delta^{(\alpha)} g_1(W_1^{(\alpha)}) \cdots g_k(W_k^{(\alpha)})]. \quad (4.17)$$

To remove the dependence on α from the notation, we can encode the summation into a mixture of measures. Let $A \sim \text{Unif}\{0, 1\}^k$ be uniformly distributed, independent from the preceding data. Then, it holds with $W'_i := W_i^{(A)}$ and $\Delta' := 2^k \Delta^{(A)}$ that

$$b(Y_1, \dots, Y_k) = \mathbb{E}[\Delta' g_1(W'_1) \cdots g_k(W'_k)]. \quad (4.18)$$

It remains to verify that the properties claimed in Proposition 4.5 are satisfied.

First, it follows from (4.11) that $\prod_i \mathbb{Q}_{W_i|W_{i-1}}^{(\alpha_i)}$ has marginal distribution \mathbb{P}_{W_i} at the i th coordinate; to verify this use induction on i together with the Markov property and Bayes' theorem. Consequently, W'_i has the same law as W_i for every α , and it follows that the same holds for the mixture W'_i . This yields the distributional property in item (1). Second, note that $\|\Delta'\|_{L^\infty} \leq 2^k \max_\alpha \|\Delta^{(\alpha)}\|_{L^\infty}$. Item (2) hence follows from (4.14) and (4.16) since $\#\mathcal{I}_1 \geq 0$. Finally, the expansion in item (3) is explicit in (4.18). \square

We next combine Propositions 4.1 and 4.5. First, however, let us set up some notation. Given a smooth function $g : \mathbb{C}^{d \times d} \rightarrow \mathbb{C}$, let it be understood that $\partial_{\mathbf{B}} g : \mathbb{C}^{d \times d} \rightarrow \mathbb{C}$ denotes the directional derivative of g in the direction of a matrix $\mathbf{B} \in \mathbb{C}^{d \times d}$:

$$(\partial_{\mathbf{B}} g)(\mathbf{M}) := \lim_{\varepsilon \rightarrow 0} \frac{g(\mathbf{M} + \varepsilon \mathbf{B}) - g(\mathbf{M})}{\varepsilon}. \quad (4.19)$$

Recall that the expansion in Proposition 4.1 involves a sum over permutations $\rho \in \mathcal{S}_k$ with $\rho(1) = 1$. To bookkeep these permutations as well as indices $i_1, \dots, i_k \in \{1, \dots, n\}$ for the Markovian sequence, we define an index set for any $k, n \geq 1$ by

$$\mathcal{I}_{k,n} := \{\rho \in \mathcal{S}_k : \rho(1) = 1\} \times \{1, \dots, n\}^k. \quad (4.20)$$

Proposition 4.6 below establishes an expansion for the rate-of-change along the interpolation $\mathbf{S}(t)$ from (4.1) using new random variables whose law depends on an index $(\rho, i) \in \mathcal{I}_{k,n}$. The exact joint distribution is however not required in the subsequent arguments, so we only explicitly state those properties that we need and refer to the proof for the construction.

PROPOSITION 4.6. *Adopt the assumptions and notation from Section 2.1. That is, let Z_1, \dots, Z_n be a Markovian sequence and consider the associated centered random matrices $\mathbf{X}_i = \mathbf{F}_i(Z_i)$ for $i \geq 1$ as well as $\mathbf{S} = \mathbf{X}_0 + \sum_{i=1}^n \mathbf{X}_i$ and a Gaussian model \mathbf{G} .*

Then, there exist random variables $(Z_{i,1}^{(\rho)}, \dots, Z_{i,k}^{(\rho)}, \Delta_i^{(\rho)})_{(\rho,i) \in \mathcal{I}_{k,n}}$ for every $k \geq 3$ that satisfy the following four properties for any $(\rho, i) \in \mathcal{I}_{k,n}$ and $j \in \{1, \dots, k\}$:

- (1) **Marginal distribution:** *It holds that $Z_{i,j}^{(\rho)}$ has the same law as $Z_{i,j}$.*
- (2) **Independence:** *The tuple of random variables $(Z_{i,1}^{(\rho)}, \dots, Z_{i,k}^{(\rho)}, \Delta_i^{(\rho)})$ is independent of the Markovian sequence Z_1, \dots, Z_n and the Gaussian model \mathbf{G} .*
- (3) **Suppression:** *The random variable $\Delta_i^{(\rho)}$ takes values in \mathbb{R} and satisfies that almost surely*

$$|\Delta_i^{(\rho)}| \leq 2^{k-\#\pi_\rho} \prod_{\mathcal{J} \in \pi_\rho} \prod_{j=2}^{\#\mathcal{J}} \min \left\{ 1, \psi \left(Z_{\text{sort}(i)_{\mathcal{J}(j)}}, Z_{\text{sort}(i)_{\mathcal{J}(j-1)}} \right) \right\}$$

where we recall that $\mathcal{J}(j)$ is the j th smallest element in \mathcal{J} , and we use the convention that an empty product yields unity when $\#\mathcal{J} = 1$.

- (4) **Expansion for Gaussian interpolations:** *Define random matrices by $\mathbf{X}_{i,j}^{(\rho)} := \mathbf{F}_{i,j}(Z_{i,j}^{(\rho)})$.*

Then, for every polynomial $g : \mathbb{C}^{d \times d} \rightarrow \mathbb{C}$ and $t \in [0, 1]$,

$$\frac{d}{dt} \mathbb{E}[g(\mathbf{S}(t))] = \frac{1}{2} \sum_{k=3}^{\infty} \frac{t^{\frac{k}{2}-1}}{(k-1)!} \mathbb{E} \left[\sum_{(\rho,i) \in \mathcal{I}_{k,n}} \Delta_i^{(\rho)} \partial_{\mathbf{X}_{i,1}^{(\rho)}} \cdots \partial_{\mathbf{X}_{i,k}^{(\rho)}} g(\mathbf{S}(t)) \right].$$

PROOF. We may assume without loss of generality that $\mathbf{G} = \mathbf{X}_0 + \sum_{i=1}^n \mathbf{G}_i$ where $(\mathbf{G}_1, \dots, \mathbf{G}_n)$ is a Gaussian model for the rectangular matrix $(\mathbf{X}_1, \dots, \mathbf{X}_n)$.⁶ Then, viewing $\mathbf{S}(t)$ as a function of these rectangular matrices and identifying complex matrices with vectors in \mathbb{R}^{2d^2} by taking the entries' real and imaginary parts, Proposition 4.1 yields that

$$\begin{aligned} \frac{d}{dt} \mathbb{E}[g(\mathbf{S}(t))] &= \frac{1}{2} \sum_{k=3}^{\infty} \frac{t^{\frac{k}{2}-1}}{(k-1)!} \sum_{i \in \{1, \dots, n\}^k} \sum_{\rho \in \mathcal{S}_k : \rho(1)=1} \sum_{\Gamma_1, \dots, \Gamma_k \in \{1, \dots, d\}^2 \times \{\text{Re}, \text{Im}\}} (-1)^{\#\pi_\rho - 1} \\ &\quad \times \left\{ \prod_{\mathcal{J} \in \pi_\rho} b(\mathbf{X}_{\text{sort}(i)_{\mathcal{J}}, \Gamma_j} : j \in \mathcal{J}) \right\} \mathbb{E} \left[\frac{\partial^k g}{\partial \mathbf{M}_{\Gamma_1} \cdots \partial \mathbf{M}_{\Gamma_k}}(\mathbf{S}(t)) \right] \end{aligned} \quad (4.21)$$

where the Γ -subscripts refer to the entries' real and imaginary parts. For instance, given $\Gamma = (v, w, \text{Re})$ with $v, w \in \{1, \dots, d\}$, we use $\mathbf{X}_{i,j,\Gamma}$ to refer to the real part of the vw th entry of the matrix $\mathbf{X}_{i,j}$. Similarly, the partial derivatives in (4.21) are taken in the direction of the real/imaginary part of the corresponding argument of the function $g : \mathbb{C}^{d \times d} \rightarrow \mathbb{C}$.

To rewrite this expansion, we next apply Proposition 4.5 to the Boolean cumulants and then exploit that expectation factorizes over products of independent random variables. To make this precise, define random variables $(Z_{i,1}^{(\rho)}, \dots, Z_{i,k}^{(\rho)}, \{\Delta_{i,\mathcal{J}}^{(\rho)} : \mathcal{J} \in \pi_\rho\})_{(\rho,i) \in \mathcal{I}_{k,n}, k \geq 3}$ that are independent of Z_1, \dots, Z_n and \mathbf{G} with joint law specified by the following properties:

1. Given $(\rho, i) \in \mathcal{I}_{k,n}$ and some part $\mathcal{J} \in \pi_\rho$ in the partition induced by ρ , set

$$(W_1, \dots, W_{\#\mathcal{J}}) := (Z_{\text{sort}(i)_{\mathcal{J}(1)}}, \dots, Z_{\text{sort}(i)_{\mathcal{J}(\#\mathcal{J})}}). \quad (4.22)$$

Then, we define the joint law of $\Delta_{i,\mathcal{J}}^{(\rho)}$ and the random variables $Z_{i,j}^{(\rho)}$ with $j \in \mathcal{J}$ to be given by the random variables resulting from Proposition 4.5 applied to $W_1, \dots, W_{\#\mathcal{J}}$, with the

⁶For instance, because the independence from all the other random variables in Proposition 4.6 implies that the statement can only depend on the law of \mathbf{G} . Alternatively, if one insists on maintaining the original matrix, one can use the joint distribution of $(\mathbf{G}_1, \dots, \mathbf{G}_n)$ and $\mathbf{X}_0 + \sum_i \mathbf{G}_i$ to define the new matrices with a coupling to \mathbf{G} .

new variables ordered in such a fashion to remove the $\text{sort}(\cdot)$ -operation from (4.22). More precisely, if $\mu_i \in \mathcal{S}_k$ is a permutation such that $i_{\mu_i(j)} = \text{sort}(i)_j$ for every $j \leq k$, then

$$(Z_{i, \mu_i(\mathcal{J}(1))}^{(\rho)}, \dots, Z_{i, \mu_i(\mathcal{J}(\#\mathcal{J}))}^{(\rho)}, \Delta_{i, \mathcal{J}}^{(\rho)}) \sim (W'_1, \dots, W'_{\#\mathcal{J}}, \Delta'). \quad (4.23)$$

2. The random variables $Z_{i,j}^{(\rho)}$ and $\Delta_{i,\mathcal{J}}^{(\rho)}$ associated with different $(\rho, i) \in \mathcal{I}_{k,n}$ or different parts $\mathcal{J} \in \pi_\rho$ are independent. More precisely, if $T_{i,\rho,\mathcal{J}}$ is the tuple on the left-hand side of (4.23), then the variables $(T_{i,\rho,\mathcal{J}} : k \geq 3, (i, \rho) \in \mathcal{I}_{k,n}, \mathcal{J} \in \pi_\rho)$ are jointly independent.

Recall that $\mathbf{X}_{i_j} = \mathbf{F}_{i_j}(Z_{i_j})$. Hence, for any fixed $(\rho, i) \in \mathcal{I}_{k,n}$, the reference to Proposition 4.5 in the definitions (4.22)–(4.23) ensures that for any $\Gamma_1, \dots, \Gamma_k$ and any $\mathcal{J} \in \pi_\rho$,

$$b(\mathbf{X}_{\text{sort}(i)_j, \Gamma_j} : j \in \mathcal{J}) = \mathbb{E} \left[\Delta_{i, \mathcal{J}}^{(\rho)} \prod_{j \in \mathcal{J}} \mathbf{X}_{i, \mu_i(j), \Gamma_j}^{(\rho)} \right] \quad (4.24)$$

where we recall from item (4) in Proposition 4.6 that $\mathbf{X}_{i_j}^{(\rho)} = \mathbf{F}_{i_j}(Z_{i_j}^{(\rho)})$. Now, using (4.24) and the fact that the variables were defined to be independent of Z_1, \dots, Z_n and \mathbf{G} as well as the joint independence in the second property of the definition,

$$\begin{aligned} & \left(\prod_{\mathcal{J} \in \pi_\rho} b(\mathbf{X}_{\text{sort}(i)_j, \Gamma_j} : j \in \mathcal{J}) \right) \mathbb{E} \left[\frac{\partial^k g}{\partial \mathbf{M}_{\Gamma_1} \cdots \partial \mathbf{M}_{\Gamma_k}}(\mathbf{S}(t)) \right] \\ &= \mathbb{E} \left[\left(\prod_{\mathcal{J} \in \pi_\rho} \Delta_{i, \mathcal{J}}^{(\rho)} \prod_{j \in \mathcal{J}} \mathbf{X}_{i, \mu_i(j), \Gamma_j}^{(\rho)} \right) \frac{\partial^k g}{\partial \mathbf{M}_{\Gamma_1} \cdots \partial \mathbf{M}_{\Gamma_k}}(\mathbf{S}(t)) \right]. \end{aligned} \quad (4.25)$$

Hence, if we define $\Delta_i^{(\rho)} := (-1)^{\#\pi_\rho - 1} \prod_{\mathcal{J} \in \pi_\rho} \Delta_{i, \mathcal{J}}^{(\rho)}$, then

$$\begin{aligned} & (-1)^{\#\pi_\rho - 1} \sum_{\Gamma_1, \dots, \Gamma_k \in \{1, \dots, d\}^2 \times \{\text{Re}, \text{Im}\}} \left(\prod_{\mathcal{J} \in \pi_\rho} b(\mathbf{X}_{\text{sort}(i)_j, \Gamma_j} : j \in \mathcal{J}) \right) \mathbb{E} \left[\frac{\partial^k g}{\partial \mathbf{M}_{\Gamma_1} \cdots \partial \mathbf{M}_{\Gamma_k}}(\mathbf{S}(t)) \right] \\ &= \sum_{\tilde{\Gamma}_1, \dots, \tilde{\Gamma}_k \in \{1, \dots, d\}^2 \times \{\text{Re}, \text{Im}\}} \mathbb{E} \left[\Delta_i^{(\rho)} \prod_{j=1}^k \mathbf{X}_{i, j, \tilde{\Gamma}_j}^{(\rho)} \frac{\partial^k g}{\partial \mathbf{M}_{\tilde{\Gamma}_1} \cdots \partial \mathbf{M}_{\tilde{\Gamma}_k}}(\mathbf{S}(t)) \right] \\ &= \mathbb{E} \left[\Delta_i^{(\rho)} \partial_{\mathbf{X}_{i,1}^{(\rho)}} \cdots \partial_{\mathbf{X}_{i,k}^{(\rho)}} g(\mathbf{S}(t)) \right] \end{aligned} \quad (4.26)$$

where the first equality renumbered the indices by taking $\tilde{\Gamma}_j := \Gamma_{\mu_i^{-1}(j)}$, and the second equality follows by recognizing the standard entry-wise decomposition for a directional derivative. Substitution of (4.26) in (4.21) yields the expansion in item (4) of Proposition 4.6.

It remains to verify items (1) to (3). The marginal distribution in item (1) follows from corresponding property in Proposition 4.5 after chasing the indices through (4.22) and (4.23). The independence in item (2) is explicit in the definitions; see the paragraph preceding (4.22). Finally, by (4.23) and the suppression property in Proposition 4.5, for any (i, ρ) and $\mathcal{J} \in \pi_\rho$,

$$|\Delta_{i, \mathcal{J}}^{(\rho)}| \leq 2^{\#\mathcal{J} - 1} \prod_{j=2}^k \min \{1, \psi(Z_{\text{sort}(i)_{\mathcal{J}(j)}}, Z_{\text{sort}(i)_{\mathcal{J}(j-1)}})\}. \quad (4.27)$$

The desired suppression property in item (3) hence follows using that $|\Delta_i^{(\rho)}| = \prod_{\mathcal{J} \in \pi_\rho} |\Delta_{i, \mathcal{J}}^{(\rho)}|$ and that $\sum_{\mathcal{J} \in \pi_\rho} (\#\mathcal{J} - 1) = k - \#\pi_\rho$ since π_ρ is a partition of $\{1, \dots, k\}$. \square

4.2. *Combining classical-to-Boolean combinatorics and the decay of dependence.* Passing through Boolean cumulants in the proofs leading up to Proposition 4.6 allowed us to efficiently exploit the Markovian structure. This however comes at a cost: we now have to understand the combinatorics associated with the classical-to-Boolean relations.

Specifically, our goal in this section is to estimate the total contribution of the scalar-valued random variables $\Delta_i^{(\rho)}$ that encode the suppression due to the decay of dependence. Regarding the individual random variables, we have the following upper bound:

LEMMA 4.7. *With notation as in Proposition 4.6, it holds for every $(\rho, i) \in \mathcal{I}_{k,n}$ that*

$$\|\Delta_i^{(\rho)}\|_{L^\infty} \leq 2^{k-1} \exp\left(-\sum_{j=2}^k \mathbb{1}\{\text{sort}(i)_{\rho(j)} > \text{sort}(i)_{\rho(j-1)}\} \left\lfloor \frac{\text{sort}(i)_{\rho(j)} - \text{sort}(i)_{\rho(j-1)}}{\Psi(Z)} \right\rfloor\right).$$

PROOF. Due to Proposition 4.4 it holds that for all $l \in \{1, \dots, n\}$ and $t \in \{0, \dots, n-l\}$,

$$\min\{1, \psi(Z_{l+t}, Z_l)\} \leq \left(\frac{1}{4}\right)^{\lfloor t/\Psi(Z) \rfloor} \leq \exp\left(-\left\lfloor \frac{t}{\Psi(Z)} \right\rfloor\right). \quad (4.28)$$

Hence, using item (3) in Proposition 4.6 and that $\#\pi_\rho \geq 1$,

$$\|\Delta_i^{(\rho)}\|_{L^\infty} \leq 2^{k-1} \prod_{\mathcal{J} \in \pi_\rho} \prod_{j=2}^{\#\mathcal{J}} \exp\left(-\left\lfloor \frac{\text{sort}(i)_{\mathcal{J}(j)} - \text{sort}(i)_{\mathcal{J}(j-1)}}{\Psi(Z)} \right\rfloor\right). \quad (4.29)$$

Recall from Section 4.1.1 that π_ρ denotes the partition of increasing runs associated with the permutation $\rho \in \mathcal{S}_k$. So, for instance, the part $\mathcal{J}_1 \in \pi_\rho$ that contains $\rho(1)$ is of the form $\{\rho(1), \rho(2), \dots, \rho(\ell_1)\}$ where $\ell_1 \geq 1$ is the least value with $\rho(\ell_1 + 1) < \rho(\ell_1)$, or where $\ell_1 = k$ if no such value exists. Hence, bringing the product inside the exponential,

$$\prod_{j=2}^{\#\mathcal{J}_1} \exp\left(-\left\lfloor \frac{\text{sort}(i)_{\mathcal{J}_1(j)} - \text{sort}(i)_{\mathcal{J}_1(j-1)}}{\Psi(Z)} \right\rfloor\right) = \exp\left(-\sum_{j=2}^{\ell_1} \left\lfloor \frac{\text{sort}(i)_{\rho(j)} - \text{sort}(i)_{\rho(j-1)}}{\Psi(Z)} \right\rfloor\right),$$

where it is to be understood that the right-hand side is unity when $\ell_1 = 1$. Continuing sequentially, the part that contains $\rho(\ell_1 + 1)$ is $\mathcal{J}_2 := \{\rho(\ell_1 + 1), \dots, \rho(\ell_2)\}$ where $\ell_2 \geq \ell_1 + 1$ is the least value with $\rho(\ell_2 + 1) < \rho(\ell_2)$, or where $\ell_2 = k$ if no such value exists. Hence,

$$\begin{aligned} & \prod_{m=1}^2 \prod_{j=2}^{\#\mathcal{J}_m} \exp\left(-\left\lfloor \frac{\text{sort}(i)_{\mathcal{J}_m(j)} - \text{sort}(i)_{\mathcal{J}_m(j-1)}}{\Psi(Z)} \right\rfloor\right) \\ &= \exp\left(-\sum_{j=2}^{\ell_1} \left\lfloor \frac{\text{sort}(i)_{\rho(j)} - \text{sort}(i)_{\rho(j-1)}}{\Psi(Z)} \right\rfloor + \sum_{j=\ell_1+2}^{\ell_2} \left\lfloor \frac{\text{sort}(i)_{\rho(j)} - \text{sort}(i)_{\rho(j-1)}}{\Psi(Z)} \right\rfloor\right) \\ &= \exp\left(-\sum_{j=2}^{\ell_2} \mathbb{1}\{\rho(j) > \rho(j-1)\} \left\lfloor \frac{\text{sort}(i)_{\rho(j)} - \text{sort}(i)_{\rho(j-1)}}{\Psi(Z)} \right\rfloor\right), \end{aligned} \quad (4.30)$$

where the final equality used the definition of ℓ_1 and ℓ_2 . Continue in this fashion until ℓ_m reaches k . Then, using that $\mathbb{1}\{\rho(j) > \rho(j-1)\} \mathbb{1}\{\text{sort}(i)_{\rho(j)} \neq \text{sort}(i)_{\rho(j-1)}\} = \mathbb{1}\{\text{sort}(i)_{\rho(j)} > \text{sort}(i)_{\rho(j-1)}\}$ because $\text{sort}(i)$ is nondecreasing concludes the proof. \square

We next sum over ρ and i to quantify the total contribution of the random variables $\Delta_i^{(\rho)}$. Specifically, we study a restricted sum wherein one of the coordinates of $i \in \{1, \dots, n\}^k$ is kept fixed. (This restriction is used in the proof of Lemma 4.17.) We start with initial simplifications that only exploit the general structure of the right-hand side of Lemma 4.7.

4.2.1. *General-purpose simplifications.* For any sequence $i \in \mathbb{R}^k$ and permutation $\rho \in \mathcal{S}_k$, let us abbreviate i_ρ for the permuted sequence $(i_{\rho(j)})_{j=1}^k$.

LEMMA 4.8. *Fix some $I \in \{1, \dots, n\}$ and $J \in \{1, \dots, k\}$. Then, for every nonnegative function $E : \{1, \dots, n\}^k \rightarrow \mathbb{R}_{\geq 0}$,*

$$\sum_{\substack{\rho \in \mathcal{S}_k \\ \rho(1)=1}} \sum_{\substack{i \in \{1, \dots, n\}^k \\ i_J = I}} E(\text{sort}(i)_\rho) \leq k! \sum_{T=1}^k \sum_{\substack{i \in \{1, \dots, n\}^k \\ i_T = I, i_1 = \min(i_j : j \leq k)}} E(i). \quad (4.31)$$

PROOF. Fix some arbitrary sequence $t \in \{1, \dots, n\}^k$. We verify that (4.31) holds for the corresponding basis function $E(\cdot) := \mathbb{1}\{\cdot = t\}$. The case with arbitrary nonnegative E then follows since both sides of the desired inequality are linear in E .

In particular, sorting i and thereafter applying an arbitrary permutation ρ with $\rho(1) = 1$ brings the minimal element to the front. Consequently, the left-hand side of (4.31) can only be nonzero if $t_1 = \min(t_j : j \leq k)$ and $t_T = I$ for some $T \leq k$. The desired bound (4.31) is trivial if the left-hand side is zero, so we may assume that t satisfies the latter conditions. Then, there is at least one nonzero term in the right-hand side of (4.31) and it suffices to show that the left-hand side is $\leq k!$.

To this end, note that it follows from the assumption that $E(\cdot) = \mathbb{1}\{\cdot = t\}$ that

$$\sum_{\substack{\rho \in \mathcal{S}_k \\ \rho(1)=1}} \sum_{\substack{i \in \{1, \dots, n\}^k \\ i_J = I}} E(\text{sort}(i)_\rho) = \#\{(i, \rho) \in \{1, \dots, n\}^k \times \mathcal{S}_k : \text{sort}(i)_\rho = t, i_J = I, \rho(1) = 1\} \\ \leq \#\{i \in \{1, \dots, n\}^k : \text{sort}(i) = \text{sort}(t)\} \times \#\{\rho \in \mathcal{S}_k : \text{sort}(t)_\rho = t\} \quad (4.32)$$

where the inequality follows by neglecting final two constraints in the first line.

Now, consider the (right) group action of \mathcal{S}_k on $\{1, \dots, k\}^n$ defined by $i * \rho := i_\rho$. Then, the first set on the right-hand side of (4.32) is simply the orbit of t under this group action:

$$\{i \in \{1, \dots, n\}^k : \text{sort}(i) = \text{sort}(t)\} = \{t_\rho : \rho \in \mathcal{S}_k\}. \quad (4.33)$$

Moreover, the cardinality of the second set agrees with that of the stabilizer of t ,

$$\#\{\rho \in \mathcal{S}_k : \text{sort}(t)_\rho = t\} = \#\{\theta \in \mathcal{S}_k : t_\theta = t\}, \quad (4.34)$$

since a bijection may be defined by setting $\theta := \mu \circ \rho$ with $\mu \in \mathcal{S}_k$ a permutation satisfying $t_\mu = \text{sort}(t)$. It hence follows from the orbit-stabilizer theorem that

$$\#\{t_\rho : \rho \in \mathcal{S}_k\} \times \#\{\theta \in \mathcal{S}_k : t_\theta = t\} = \#\mathcal{S}_k = k!. \quad (4.35)$$

Combine (4.32)–(4.35) to conclude that the left-hand side of (4.31) is $\leq k!$, as desired. \square

LEMMA 4.9. *Fix some integer $I \in \mathbb{Z}$ and $T \in \{1, \dots, k\}$ as well as a nonnegative function $E : \mathbb{Z}^k \rightarrow \mathbb{R}_{\geq 0}$. Assume that E is translation invariant in the sense that $E(i_1, \dots, i_k) = E(i_1 + x, \dots, i_k + x)$ for every $i_1, \dots, i_k \in \mathbb{Z}$ and $x \in \mathbb{Z}$. Then,*

$$\sum_{\substack{i \in \mathbb{Z}^k, i_T = I \\ i_1 = \min(i_j : j \leq k)}} E(i) = \sum_{\substack{i \in \mathbb{Z}_{\geq 0}^k, i_1 = 0}} E(i) \quad (4.36)$$

PROOF. Note that i runs over integer sequences in the left-hand side of (4.36) that may also include negative values. We subdivide the sum by the minimal value i_{\min} attained by the sequence i and subsequently use translation invariance with $x = -i_{\min}$

$$\sum_{\substack{i \in \mathbb{Z}^k, i_T = I \\ i_1 = \min(i_j : j \leq k)}} E(i) = \sum_{i_{\min} = -\infty}^I \sum_{\substack{i \in \mathbb{Z}^k, i_T = I \\ i_1 = \min(i_j : j \leq k) = i_{\min}}} E(i) = \sum_{i_{\min} = -\infty}^I \sum_{\substack{i \in \mathbb{Z}^k, i_T = I - i_{\min} \\ i_1 = \min(i_j : j \leq k) = 0}} E(i). \quad (4.37)$$

Here, making a change of variables $\mathfrak{J} := I - i_{\min}$ and rewriting the constraints in the second sum into an equivalent formulation,

$$\sum_{i_{\min}=-\infty}^I \sum_{\substack{i \in \mathbb{Z}^k, i_T=I-i_{\min} \\ i_1=\min(i_j:j \leq k)=0}} E(i) = \sum_{\mathfrak{J} \geq 0} \sum_{\substack{i \in \mathbb{Z}_{\geq 0}^k \\ i_T=\mathfrak{J}, i_1=0}} E(i) = \sum_{i \in \mathbb{Z}_{\geq 0}^k, i_1=0} E(i). \quad (4.38)$$

The combination of (4.37) and (4.38) yields (4.36), as desired. \square

COROLLARY 4.10. *For any fixed $I \in \{1, \dots, n\}$ and $J \in \{1, \dots, k\}$,*

$$\sum_{(\rho, \alpha, i) \in \mathcal{I}_{k,n}: i_J=I} \|\Delta_i^{(\rho)}\|_{L^\infty} \leq 4^{k-1} k! \sum_{i \in \{0\} \times \mathbb{Z}_{\geq 0}^{k-1}} \exp\left(-\sum_{j=2}^k \mathbb{1}\{i_j > i_{j-1}\} \left\lfloor \frac{i_j - i_{j-1}}{\Psi(Z)} \right\rfloor\right). \quad (4.39)$$

PROOF. Let us define a function $E : \mathbb{Z}^k \rightarrow \mathbb{R}_{\geq 0}$ by

$$E(i) := 2^{k-1} \exp\left(-\sum_{j=2}^k \mathbb{1}\{i_j > i_{j-1}\} \left\lfloor \frac{i_j - i_{j-1}}{\Psi(Z)} \right\rfloor\right). \quad (4.40)$$

Then, Lemma 4.7 yields that $\|\Delta_i^{(\rho)}\|_{L^\infty} \leq E(\text{sort}(i)_\rho)$. Hence, recalling the definition of $\mathcal{I}_{k,n}$ from (4.20) and using Lemma 4.8,

$$\sum_{(\rho, i) \in \mathcal{I}_{k,n}: i_J=I} \|\Delta_i^{(\rho)}\|_{L^\infty} = \sum_{\substack{\rho \in \mathcal{S}_k \\ \rho(1)=1}} \sum_{\substack{i \in \{1, \dots, n\}^k \\ i_J=I}} E(\text{sort}(i)_\rho) \leq k! \sum_{T=1}^k \sum_{\substack{i \in \{1, \dots, n\}^k \\ i_T=I, i_1=\min(i_j:j \leq k)}} E(i). \quad (4.41)$$

Note that $E(i)$ is translation invariant. Hence, by enlarging the sum and using Lemma 4.9,

$$\sum_{\substack{i \in \{1, \dots, n\}^k \\ i_T=I, i_1=\min(i_j:j \leq k)}} E(i) \leq \sum_{\substack{i \in \mathbb{Z}^k \\ i_T=I, i_1=\min(i_j:j \leq k)}} E(i) = \sum_{i \in \{0\} \times \mathbb{Z}_{\geq 0}^{k-1}} E(i). \quad (4.42)$$

The desired estimate (4.39) now follows by combining (4.41) and (4.42), where we use that $k \leq 2^{k-1}$ to bound the factor arising from the sum over T in (4.41). \square

The sum in (4.39) can be interpreted as the normalization constant of a $\mathbb{Z}_{\geq 0}$ -valued stochastic process for which going up by more than $\Psi(Z)$ in a single step is penalized by an exponential factor, and with steps downward or up by less than $\Psi(Z)$ being cost-free. Our goal in the subsequent arguments is to establish an upper bound on this normalization constant.

REMARK 4.11. The aforementioned interpretation suggests that a significant contribution to the sum should come from the case where the path starts with a number of penalized steps upwards and subsequently only takes unpenalized steps, as this creates the most possible combinations given a fixed amount of penalization. An earlier version of this work utilized this structure to bound the normalization constant; the interested reader is referred to arXiv:2307.11632v2. A direct computation however turns out to be shorter, which we present below. We thank an anonymous referee for suggesting the more efficient argument.

4.2.2. *Bounding the “normalization constant”.* We start with preparatory estimates:

LEMMA 4.12. *Fix integers $i, P \geq 1$ and $q \geq 0$. Then,*

$$\sum_{\ell=0}^{i-1} \ell^q \leq \frac{i^{q+1}}{q+1} \quad \text{and} \quad \sum_{\ell=0}^{\infty} \ell^q \exp\left(-\left\lfloor \frac{\ell}{P} \right\rfloor\right) \leq e 2^q P^{q+1} q!. \quad (4.43)$$

Here, it should be understood that $\ell^q = 1$ if $\ell = 0 = q$.

PROOF. The first estimate in (4.43) is an equality if $q = 0$. Now let $q > 0$. Because $x \mapsto x^q$ is nondecreasing for $x \geq 0$ and $q > 0$, we have $\sum_{\ell=0}^{i-1} \ell^q \leq \int_1^i x^q dx$. The first estimate in (4.43) then follows from $\int_1^i x^q < \int_0^i x^q dx = i^{q+1}/(q+1)$.

For the second estimate, we start by subdividing the sum based on $\lfloor \ell/P \rfloor$:

$$\sum_{\ell=0}^{\infty} \ell^q \exp\left(-\left\lfloor \frac{\ell}{P} \right\rfloor\right) = \sum_{j=0}^{\infty} \sum_{\ell=jP}^{(j+1)P-1} \ell^q \exp(-j) \leq P^{q+1} \sum_{j=0}^{\infty} (j+1)^q \exp(-j). \quad (4.44)$$

The inequality here used that $\ell^q \leq P^q (j+1)^q$ for all $\ell \in \{jP, \dots, (j+1)P-1\}$. Note that $x+1 \geq j+1$ and $\exp(-(x-1)) \geq \exp(-j)$ for all $x \in [j, j+1)$. Consequently,

$$\sum_{j=0}^{\infty} (j+1)^q \exp(-j) \leq \int_0^{\infty} (x+1)^q \exp(-(x-1)) dx = e \int_0^{\infty} (x+1)^q \exp(-x) dx. \quad (4.45)$$

Conclude by noting that $(x+1)^q \leq 2^q x^q$ and $\int_0^{\infty} x^q \exp(-x) dx = \Gamma(q+1) = q!$. \square

LEMMA 4.13. *For integers $P \geq 1$ and $k \geq 2$,*

$$\sum_{i \in \{0\} \times \mathbb{Z}_{\geq 0}^{k-1}} \exp\left(-\sum_{j=2}^k \mathbb{1}\{i_j > i_{j-1}\} \left\lfloor \frac{i_j - i_{j-1}}{P} \right\rfloor\right) \leq 16^{k-1} P^{k-1}. \quad (4.46)$$

PROOF. We consider a more general quantity. For an integer $q \geq 0$, define

$$S(k, q) := \sum_{(i_1, \dots, i_k) \in \{0\} \times \mathbb{Z}_{\geq 0}^{k-1}} i_k^q \exp\left(-\sum_{j=2}^k \mathbb{1}\{i_j > i_{j-1}\} \left\lfloor \frac{i_j - i_{j-1}}{P} \right\rfloor\right). \quad (4.47)$$

Here, it should be understood that $i_k^q = 1$ if $i_k = 0 = q$. Observe that $S(k, 0)$ equals the left-hand side of (4.46). It hence suffices to prove that

$$S(k, q) \leq 16^{k-1+q/2} P^{k-1+q} q!. \quad (4.48)$$

We proceed by induction on $k \geq 2$.

First, note that $S(2, q) = \sum_{i_2=0}^{\infty} i_2^q \exp(-\lfloor i_2/P \rfloor)$. The second estimate in Lemma 4.12 then yields that $S(2, q) \leq e 2^q P^{q+1} q!$. This shows that (4.48) holds for $k = 2$.

Next, suppose that (4.48) holds for a given $k \geq 2$ and all $q \geq 0$. We will now show that the bound then remains valid for $k+1$. For any fixed $i_k \geq 0$,

$$\begin{aligned} \sum_{i_{k+1}=0}^{\infty} i_k^q \exp\left(-\mathbb{1}\{i_{k+1} > i_k\} \left\lfloor \frac{i_{k+1} - i_k}{P} \right\rfloor\right) &= \sum_{i_{k+1}=0}^{i_k-1} i_k^q + \sum_{\ell=0}^{\infty} (\ell + i_k)^q \exp\left(-\left\lfloor \frac{\ell}{P} \right\rfloor\right) \\ &= \sum_{i_{k+1}=0}^{i_k-1} i_k^q + \sum_{j=0}^q \binom{q}{j} i_k^{q-j} \sum_{\ell=0}^{\infty} \ell^j \exp\left(-\left\lfloor \frac{\ell}{P} \right\rfloor\right) \leq \frac{i_k^{q+1}}{q+1} + e \sum_{j=0}^q \binom{q}{j} i_k^{q-j} 2^j P^{j+1} j!. \end{aligned} \quad (4.49)$$

Here, we used the binomial theorem for the second equality and Lemma 4.12 for the inequality. Substituting (4.49) in (4.47) and using the induction hypothesis (4.48) yields

$$\begin{aligned} S(k+1, q) &= \frac{1}{q+1} S(k, q+1) + e \sum_{j=0}^q \binom{q}{j} 2^j P^{j+1} j! S(k, q-j) \\ &\leq \left(16^{-1/2} + e \sum_{j=0}^q 2^j 16^{-j/2-1} \right) 16^{k+q/2} P^{k+q} q!. \end{aligned} \quad (4.50)$$

Note that the term within brackets in (4.50) is no greater than 1 to conclude that (4.48) also holds when k is replaced by $k+1$. This concludes the proof. \square

COROLLARY 4.14. *For all fixed $k \geq 3$, $I \in \{1, \dots, n\}$, $J \in \{1, \dots, k\}$,*

$$\sum_{(\rho, \alpha, i) \in \mathcal{I}_{k, n}: i_J = I} \|\Delta_i^{(\rho)}\|_{L^\infty} \leq 64^{k-1} k! \Psi(Z)^{k-1}. \quad (4.51)$$

PROOF. This follows by combining Corollary 4.10 with Lemma 4.13. \square

4.3. *Derivative of tracial moments.* Recall that our goal is to control the rate of change of tracial moments along the interpolation $\mathbf{S}(t)$. To control the individual terms in the expansion of Proposition 4.6, we must then understand the directional derivatives of $g(\mathbf{M}) := \text{tr}[\mathbf{M}^p]$. The following lemma, which also appears in [14, Lemma 6.3], provides an explicit expression:

LEMMA 4.15. *For any positive integers $1 \leq k \leq p$ and any matrices $\mathbf{B}_1, \dots, \mathbf{B}_k \in \mathbb{C}^{d \times d}$ the polynomial function from $\mathbb{C}^{d \times d}$ to \mathbb{C} defined by $\mathbf{M} \mapsto \text{tr}[\mathbf{M}^p]$ satisfies*

$$\partial_{\mathbf{B}_1} \cdots \partial_{\mathbf{B}_k} \text{tr}[\mathbf{M}^p] = \sum_{\theta \in \mathcal{S}_k} \sum_{\substack{r_1, \dots, r_{k+1} \geq 0 \\ r_1 + \dots + r_{k+1} = p-k}} \text{tr}[\mathbf{M}^{r_1} \mathbf{B}_{\theta(1)} \mathbf{M}^{r_2} \mathbf{B}_{\theta(2)} \cdots \mathbf{M}^{r_k} \mathbf{B}_{\theta(k)} \mathbf{M}^{r_{k+1}}].$$

PROOF. This follows by entry-wise expansion for the trace and the product rule. \square

We next establish a trace inequality that will be used to control the terms that arise on the right-hand side of Lemma 4.15. For a random matrix \mathbf{M} and a scalar $1 \leq p \leq \infty$, we define

$$\|\mathbf{M}\|_p := \begin{cases} \mathbb{E}[\text{tr}[\|\mathbf{M}\|^p]]^{\frac{1}{p}} & \text{if } p < \infty, \\ \|\|\mathbf{M}\|\|_{L^\infty} & \text{if } p = \infty. \end{cases} \quad (4.52)$$

It is known that $\|\cdot\|_p$ defines a norm on bounded random matrices [33, (26)]. The following result is related to [14, Proposition 5.1], as is its proof.

LEMMA 4.16. *Fix a positive integer $k \geq 2$ and consider a finite set \mathcal{I} . Let $\mathbf{X} := (\mathbf{X}_{i,j})_{i \in \mathcal{I}, j \in \{1, \dots, k\}}$ and $\mathbf{M} := (\mathbf{M}_j)_{j=1}^k$ be bounded self-adjoint random matrices and consider bounded real-valued random variables $\Delta := (\Delta_i)_{i \in \mathcal{I}}$. Suppose that \mathbf{M} is independent of (\mathbf{X}, Δ) . Then, for any scalars $1 \leq p_1, \dots, p_k \leq \infty$ with $\sum_{j=1}^k 1/p_j = 1$*

$$\left| \mathbb{E} \left[\sum_{i \in \mathcal{I}} \Delta_i \text{tr}[\mathbf{X}_{i,1} \mathbf{M}_1 \mathbf{X}_{i,2} \mathbf{M}_2 \cdots \mathbf{X}_{i,k} \mathbf{M}_k] \right] \right| \leq R_{\mathcal{I}}(\mathbf{X})^{k-2} \varsigma_{\Delta}(\mathbf{X})^2 \prod_{j=1}^k \|\mathbf{M}_j\|_{p_j} \quad (4.53)$$

where

$$R_{\mathcal{I}}(\mathbf{X}) = \max_{j=1, \dots, k} \max_{i \in \mathcal{I}} \|\|\mathbf{X}_{i,j}\|\|_{L^\infty} \quad \text{and} \quad \varsigma_{\Delta}(\mathbf{X})^2 := \max_{j=1, \dots, k} \left\| \mathbb{E} \left[\sum_{i \in \mathcal{I}} |\Delta_i| \mathbf{X}_{i,j}^2 \right] \right\|.$$

PROOF. As in [14, Proof of Proposition 5.1, Step 1] one can employ a convexity result, namely [14, Lemma 5.2], and the cyclic property of the trace to reduce to the case where $p_k = 1$. Note that it then necessarily holds that $p_j = \infty$ for any $j \neq k$. By rescaling both sides of (4.53) it may further be assumed that $\|\mathbf{M}_k\|_1 = 1$ and $\|\mathbf{M}_j\|_\infty = 1$ for any $j < k$.

For the sake of notational simplicity, let I be a random index which is uniformly distributed in \mathcal{I} and independent of \mathbf{X}, Δ and \mathbf{M} . Then, also using the tower property,

$$\begin{aligned} \left| \mathbb{E} \left[\sum_{i \in \mathcal{I}} \Delta_i \operatorname{tr}[\mathbf{X}_{i,1} \mathbf{M}_1 \cdots \mathbf{M}_{k-1} \mathbf{X}_{i,k} \mathbf{M}_k] \right] \right| &= \#\mathcal{I} \left| \mathbb{E}[\operatorname{tr}[\Delta_I \mathbf{X}_{I,1} \mathbf{M}_1 \cdots \mathbf{M}_{k-1} \mathbf{X}_{I,k} \mathbf{M}_k]] \right| \\ &\leq \#\mathcal{I} \left| \mathbb{E}[\operatorname{tr} \mathbb{E}[\Delta_I \mathbf{X}_{I,1} \mathbf{M}_1 \cdots \mathbf{M}_{k-1} \mathbf{X}_{I,k} \mid \mathbf{M}] \mathbf{M}_k] \right|. \end{aligned} \quad (4.54)$$

The norm (4.52) admits a Hölder-type inequality that implies that $|\mathbb{E}[\operatorname{tr} \mathbf{A} \mathbf{B}]| \leq \|\mathbf{A}\|_p \|\mathbf{B}\|_q$ for any random matrices \mathbf{A}, \mathbf{B} and $1/p + 1/q = 1$; see e.g., [14, Lemma 5.3] for a proof. Using this with $\mathbf{A} = \mathbb{E}[\Delta_I \mathbf{X}_{I,1} \mathbf{M}_1 \cdots \mathbf{M}_{k-1} \mathbf{X}_{I,k} \mid \mathbf{M}]$ and $\mathbf{B} = \mathbf{M}_k$ and $p = \infty$ and $q = 1$,

$$\mathbb{E}[\operatorname{tr} \mathbb{E}[\Delta_I \mathbf{X}_{I,1} \mathbf{M}_1 \cdots \mathbf{M}_{k-1} \mathbf{X}_{I,k} \mid \mathbf{M}] \mathbf{M}_k] \leq \|\mathbb{E}[\Delta_I \mathbf{X}_{I,1} \mathbf{M}_1 \cdots \mathbf{M}_{k-1} \mathbf{X}_{I,k} \mid \mathbf{M}]\|_{L^\infty} \|\mathbf{M}_k\|_1 \quad (4.55)$$

where we used that $\mathbb{E}[\operatorname{tr} \mathbf{M}_k] = \|\mathbf{M}_k\|_1 = 1$ by the preliminary reductions.

Denote $\mathbf{Z} = \mathbf{M}_1 \mathbf{X}_{I,2} \mathbf{M}_3 \mathbf{X}_{I,3} \cdots \mathbf{M}_{k-1}$ and note that, almost surely, $\mathbf{Z} \mathbf{Z}^* \leq R_{\mathcal{I}}(\mathbf{X})^{2k-4} \mathbf{1}$ with respect to the positive semidefinite order where $\mathbf{1}$ is the identity matrix. By definition of the operator norm, the Cauchy–Schwarz inequality for random vectors, and the assumption that (\mathbf{X}, Δ) and \mathbf{M} are independent, it follows that almost surely

$$\begin{aligned} &\|\mathbb{E}[\Delta_I \mathbf{X}_{I,1} \mathbf{M}_1 \cdots \mathbf{M}_{k-1} \mathbf{X}_{I,k} \mid \mathbf{M}]\| && (4.56) \\ &= \sup_{v, w \in S^{d-1}} \left| \mathbb{E}[(v^* \operatorname{sign}(\Delta_I) |\Delta_I|^{\frac{1}{2}} \mathbf{X}_{I,1} \mathbf{Z}) (|\Delta_I|^{\frac{1}{2}} \mathbf{X}_{I,k} w) \mid \mathbf{M}] \right| \\ &\leq \sup_{v, w \in S^{d-1}} \mathbb{E}[v^* |\Delta_I| \mathbf{X}_{I,1} \mathbf{Z} \mathbf{Z}^* \mathbf{X}_{I,1} v \mid \mathbf{M}]^{\frac{1}{2}} \mathbb{E}[w^* |\Delta_I| \mathbf{X}_{I,k} \mathbf{X}_{I,k} w \mid \mathbf{M}]^{\frac{1}{2}} \\ &\leq R_{\mathcal{I}}^{k-2}(\mathbf{X}) \sup_{v, w \in S^{d-1}} \mathbb{E}[v^* |\Delta_I| \mathbf{X}_{I,1}^2 v]^{\frac{1}{2}} \mathbb{E}[w^* |\Delta_I| \mathbf{X}_{I,k}^2 w]^{\frac{1}{2}} \end{aligned}$$

with $S^{d-1} \subseteq \mathbb{C}^d$ the unit sphere. Here, for any $j \in \{1, \dots, k\}$ and $v \in S^{d-1}$

$$\mathbb{E}[v^* |\Delta_I| \mathbf{X}_{I,j}^2 v] = (\#\mathcal{I})^{-1} v^* \mathbb{E} \left[\sum_{i \in \mathcal{I}} |\Delta_i| \mathbf{X}_{i,j}^2 \right] v \leq (\#\mathcal{I})^{-1} \zeta_{\Delta}(\mathbf{X})^2. \quad (4.57)$$

Combine (4.54)–(4.57) to conclude the proof. \square

4.4. *Proof of Theorem 2.4.* We finally combine the preceding ingredients with a direct calculation to control the rate of change of $\mathbb{E}[\operatorname{tr} \mathbf{S}(t)^{2p}]$ along the interpolation from (4.1) and use this to prove the desired universality principle for the tracial moments of even order.

LEMMA 4.17. *For any integer $p \geq 2$ and any $t \in [0, 1]$*

$$\begin{aligned} &\left| \frac{d}{dt} \mathbb{E}[\operatorname{tr} \mathbf{S}(t)^{2p}] \right| && (4.58) \\ &\leq (400p)^3 R(\mathbf{X}) \Psi(Z)^2 \zeta(\mathbf{X})^2 \max \left\{ \mathbb{E}[\operatorname{tr} \mathbf{S}(t)^{2p}]^{1-\frac{3}{2p}}, (400p R(\mathbf{X}) \Psi(Z))^{2p-3} \right\}. \end{aligned}$$

PROOF. The combination of Proposition 4.6 and Lemma 4.15 yields that

$$\frac{d}{dt} \mathbb{E}[\operatorname{tr} \mathbf{S}(t)^{2p}] = \frac{1}{2} \sum_{k=3}^{2p} \frac{t^{\frac{k}{2}-1}}{(k-1)!} \sum_{\theta \in \mathcal{S}_k} \sum_{\substack{r_1, \dots, r_{k+1} \geq 0 \\ r_1 + \dots + r_{k+1} = 2p-k}} \quad (4.59)$$

$$\times \mathbb{E} \left[\sum_{(\rho, i) \in \mathcal{I}_{k, n}} \operatorname{tr} [\Delta_i^{(\rho)} \mathbf{S}(t)^{r_1} \mathbf{X}_{i, \theta(1)}^{(\rho)} \cdots \mathbf{S}(t)^{r_k} \mathbf{X}_{i, \theta(k)}^{(\rho)} \mathbf{S}(t)^{r_{k+1}}] \right].$$

We here used that $\mathbf{M} \rightarrow \operatorname{tr} \mathbf{M}^{2p}$ is a polynomial of degree $2p$ to neglect all terms in Proposition 4.6 with $k > 2p$. Any $\theta \in \mathcal{S}_k$ defines a bijection from $\mathcal{I}_{k, n}$ into itself as $(\rho, i) \mapsto (\rho, i_\theta)$ where we recall that $(i_\theta)_j = i_{\theta(j)}$. Hence, the sum over θ in (4.59) can be eliminated in exchange for a factor $\#\mathcal{S}_k = k!$:

$$\frac{d}{dt} \mathbb{E} [\operatorname{tr} \mathbf{S}(t)^{2p}] \leq \frac{1}{2} \sum_{k=3}^{2p} \sum_{\substack{r_1, \dots, r_{k+1} \geq 0 \\ r_1 + \dots + r_{k+1} = 2p - k}} t^{\frac{k}{2} - 1} k \times \quad (4.60)$$

$$\left| \mathbb{E} \left[\sum_{(\rho, i) \in \mathcal{I}_{k, n}} \operatorname{tr} [\Delta_i^{(\rho)} \mathbf{S}(t)^{r_1} \mathbf{X}_{i, 1}^{(\rho)} \cdots \mathbf{S}(t)^{r_k} \mathbf{X}_{i, k}^{(\rho)} \mathbf{S}(t)^{r_{k+1}}] \right] \right|.$$

We next apply Lemma 4.16 with $\mathbf{M}_j = \mathbf{S}(t)^{r_{j+1}}$ and $p_j = (2p - k)/r_{j+1}$ for $j < k$ and $\mathbf{M}_k = \mathbf{S}(t)^{r_{k+1} + r_1}$ and $p_k = (2p - k)/(r_{k+1} + r_1)$ for $j = k$. Note that the independence condition in Lemma 4.16 is then satisfied because $\mathbf{S}(t)$ only depends on \mathbf{G} and Z_1, \dots, Z_n which are independent of $\mathbf{X}_{i_j}^{(\rho)} = \mathbf{F}_{i_j}(Z_{i_j}^{(\rho)})$ and $\Delta_i^{(\rho)}$ due item (2) in Proposition 4.6. Thus,

$$\begin{aligned} & \left| \mathbb{E} \left[\sum_{(\rho, i) \in \mathcal{I}_{k, n}} \operatorname{tr} [\Delta_i^{(\rho)} \mathbf{S}(t)^{r_1} \mathbf{X}_{i, 1}^{(\rho)} \cdots \mathbf{S}(t)^{r_k} \mathbf{X}_{i, k}^{(\rho)} \mathbf{S}(t)^{r_{k+1}}] \right] \right| \quad (4.61) \\ & \leq R_{\mathcal{I}_{k, n}}(\mathbf{X}^{(\rho)})^{k-2} \varsigma_{\Delta}(\mathbf{X}^{(\rho)})^2 \mathbb{E} [\operatorname{tr} |\mathbf{S}(t)|^{2p-k}] \end{aligned}$$

where

$$R_{\mathcal{I}_{k, n}}(\mathbf{X}^{(\rho)}) := \max_{j=1, \dots, k} \max_{(\rho, i) \in \mathcal{I}_{k, n}} \|\mathbf{X}_{i, j}^{(\rho)}\|_{L^\infty}, \quad (4.62)$$

$$\varsigma_{\Delta}(\mathbf{X}^{(\rho)})^2 := \max_{j=1, \dots, k} \left\| \mathbb{E} \left[\sum_{(\rho, i) \in \mathcal{I}_{k, n}} |\Delta_i^{(\rho)}| (\mathbf{X}_{i, j}^{(\rho)})^2 \right] \right\|. \quad (4.63)$$

It follows from item (1) in Proposition 4.6 that $\mathbf{X}_{i, j}^{(\rho)}$ has the same marginal distribution as $\mathbf{X}_{i, j}$. Hence, $R_{\mathcal{I}_{k, n}}(\mathbf{X}^{(\rho)}) = R(\mathbf{X})$ and $\mathbb{E}[(\mathbf{X}_{i, j}^{(\rho)})^2] = \mathbb{E}[\mathbf{X}_{i, j}^2]$ with $R(\mathbf{X})$ as in (2.6). Thus, the following inequality holds with respect to the positive semidefinite order for any fixed $j \leq k$:

$$\begin{aligned} & \mathbb{E} \left[\sum_{(\rho, i) \in \mathcal{I}_{k, n}} |\Delta_i^{(\rho)}| (\mathbf{X}_{i, j}^{(\rho)})^2 \right] \preceq \sum_{(\rho, i) \in \mathcal{I}_{k, n}} \|\Delta_i^{(\rho)}\|_{L^\infty} \mathbb{E} \left[(\mathbf{X}_{i, j}^{(\rho)})^2 \right] \quad (4.64) \\ & = \sum_{I=1}^n \left(\sum_{(\rho, i) \in \mathcal{I}_{k, n}: i_j = I} \|\Delta_i^{(\rho)}\|_{L^\infty} \right) \mathbb{E} \left[\mathbf{X}_I^2 \right]. \end{aligned}$$

Here, we have $\sum_{(\rho, i) \in \mathcal{I}_{k, n}: i_j = I} \|\Delta_i^{(\rho)}\|_{L^\infty} \leq 64^{k-1} k! \Psi(Z)^{k-1}$ by Corollary 4.14. Recall the definitions of $\varsigma(\mathbf{X})^2$ and $\varsigma_{\Delta}(\mathbf{X}^{(\rho)})^2$ from (2.7) and (4.63), respectively. We conclude that

$$\varsigma_{\Delta}(\mathbf{X}^{(\rho)})^2 \leq 64^{k-1} k! \Psi(Z)^{k-1} \varsigma(\mathbf{X})^2. \quad (4.65)$$

We next combine (4.60)–(4.65). Note that the number of $(k+1)$ -tuples (r_1, \dots, r_{k+1}) satisfying $r_j \geq 0$ and $\sum_{j=1}^{k+1} r_j = 2p - k$ is equal to $\binom{2p}{k} \leq (2p)^k / k!$. Hence,

$$\begin{aligned} & \left| \frac{d}{dt} \mathbb{E} [\operatorname{tr} \mathbf{S}(t)^{2p}] \right| \leq \frac{1}{2} \sum_{k=3}^{2p} t^{\frac{k}{2} - 1} k 64^{k-1} (2p)^k R(\mathbf{X})^{k-2} \Psi(Z)^{k-1} \varsigma(\mathbf{X})^2 \mathbb{E} [\operatorname{tr} |\mathbf{S}(t)|^{2p-k}] \\ & \leq \frac{1}{2} \sum_{k=3}^{2p} (200p)^k R(\mathbf{X})^{k-2} \Psi(Z)^{k-1} \varsigma(\mathbf{X})^2 \mathbb{E} [\operatorname{tr} \mathbf{S}(t)^{2p}]^{1 - \frac{k}{2p}} \quad (4.66) \end{aligned}$$

where the second inequality used that $\mathbb{E}[\text{tr}|\mathbf{S}(t)|^{2p-k}] \leq \mathbb{E}[\text{tr}\mathbf{S}(t)^{2p}]^{1-\frac{k}{2p}}$ by Jensen's inequality as well as the fact that $t^{\frac{k}{2}-1} \leq 1$ and $k(64)^{k-1} \leq 100^k$ for all $k \geq 3$. Here, we can further simplify by using Hölder's inequality with the fact that $\sum_{k=3}^{2p} (1/2)^k \leq \sum_{k=0}^{2p} (1/2)^k = 2$,

$$\begin{aligned} & \frac{1}{2} \sum_{k=3}^{2p} (200p)^k R(\mathbf{X})^{k-2} \Psi(Z)^{k-1} \varsigma(\mathbf{X})^2 \mathbb{E}[\text{tr}\mathbf{S}(t)^{2p}]^{1-\frac{k}{2p}} \\ &= \frac{1}{2} \sum_{k=3}^{2p} 2^{-k} (400p)^k R(\mathbf{X})^{k-2} \Psi(Z)^{k-1} \varsigma(\mathbf{X})^2 \mathbb{E}[\text{tr}\mathbf{S}(t)^{2p}]^{1-\frac{k}{2p}} \\ &\leq \max_{3 \leq k \leq 2p} (400p)^k R(\mathbf{X})^{k-2} \Psi(Z)^{k-1} \varsigma(\mathbf{X})^2 \mathbb{E}[\text{tr}\mathbf{S}(t)^{2p}]^{1-\frac{k}{2p}}. \end{aligned} \quad (4.67)$$

Finally, note that the function $k \mapsto x^k$ is convex for any fixed $x \geq 0$. It follows that the maximum on the right-hand side of (4.67) is achieved at $k = 3$ or at $k = 2p$. Combine (4.66) and (4.67) to complete the proof. \square

Solving the differential inequality in Lemma 4.17 now yields the desired result:

PROOF OF THEOREM 2.4. If $p = 1$, then the result is true because $\mathbb{E}[\mathbf{G}^2] = \mathbb{E}[\mathbf{S}^2]$ by definition of a Gaussian model. Now assume that $p \geq 2$.

Every differentiable function $f : [0, 1] \rightarrow \mathbb{R}_{\geq 0}$ with $|\frac{d}{dt}f(t)| \leq C \max\{f(t)^{1-\alpha}, K^{1-\alpha}\}$ for constants $C, K \geq 0$ and $\alpha \in [0, 1]$ satisfies that $|f(1)^\alpha - f(0)^\alpha| \leq C\alpha + K^\alpha$; see [14, Lemma 6.6] for a proof. Applying this with $\alpha = 3/(2p)$ to the conclusion of Lemma 4.17

$$|\mathbb{E}[\text{tr}\mathbf{S}^{2p}]^{\frac{3}{2p}} - \mathbb{E}[\text{tr}\mathbf{G}^{2p}]^{\frac{3}{2p}}| \leq \frac{3}{2p} (400p)^3 R(\mathbf{X}) \Psi(Z)^2 \varsigma(\mathbf{X})^2 + (400p R(\mathbf{X}) \Psi(Z))^3. \quad (4.68)$$

Using that $v \mapsto (w+v)^{1/3} - v^{1/3}$ is a nonincreasing function in $v \geq 0$ for any fixed $w \geq 0$, it can be deduced that $|x^{1/3} - y^{1/3}| \leq |x - y|^{1/3}$ and $(x+y)^{1/3} \leq x^{1/3} + y^{1/3}$ for any $x, y \geq 0$. The desired result hence follows with absolute constant $c = 500$ by raising both sides of (4.68) to the power $1/3$ and using that $(3/2)^{1/3} 400 \leq 500$. \square

5. Proof of Corollaries 2.5 and 2.6 as well as Equation (2.12). We now demonstrate how the universality of tracial moments proved in Theorem 2.4 can be combined with Gaussian theory to give bounds on the operator norm. We use yet another variance proxy:

$$\sigma_*(\mathbf{S})^2 := \sup_{\|v\|=\|w\|=1} \mathbb{E}[|\langle v, (\mathbf{S} - \mathbb{E}[\mathbf{S}])w \rangle|^2]. \quad (5.1)$$

This quantity is used to state the following general-purpose estimate in its strongest form, but it will suffice in our application that $\sigma_*(\mathbf{S}) \leq v(\mathbf{S})$ and $\sigma_*(\mathbf{S}) \leq \sigma(\mathbf{S})$ [9, Section 2.1].

LEMMA 5.1. *There exists an absolute constant $C > 0$ such that for every integer $p \geq 1$,*

$$d^{-\frac{1}{2p}} \mathbb{E}[\|\mathbf{G}\|] - C\mathcal{E}_*(p) \leq \mathbb{E}[\|\mathbf{S}\|^{2p}]^{1/2p} \leq d^{\frac{1}{2p}} \mathbb{E}[\|\mathbf{G}\|] + Cd^{\frac{1}{2p}} \mathcal{E}_*(p) \quad (5.2)$$

where $\mathcal{E}_*(p) := R(\mathbf{X})^{1/3} \Psi(Z)^{2/3} \varsigma(\mathbf{X})^{2/3} p^{2/3} + R(\mathbf{X}) \Psi(Z) p + \sigma_*(\mathbf{S}) \sqrt{p}$.

PROOF. Using that $d^{-1} \|\mathbf{M}\|^{2p} \leq \text{tr}[\mathbf{M}^{2p}] \leq \|\mathbf{M}\|^{2p}$ for any $\mathbf{M} \in \mathbb{C}_{\text{sa}}^{d \times d}$, it follows from Theorem 2.4 that there exists an absolute constant $c > 0$ such that

$$d^{-\frac{1}{2p}} \mathbb{E}[\|\mathbf{G}\|^{2p}]^{\frac{1}{2p}} - cE(p) \leq \mathbb{E}[\|\mathbf{S}^{2p}\|]^{1/2p} \leq d^{\frac{1}{2p}} \mathbb{E}[\|\mathbf{G}\|^{2p}]^{\frac{1}{2p}} + cd^{\frac{1}{2p}} E(p) \quad (5.3)$$

where $E(p) := R(\mathbf{X})^{1/3}\Psi(Z)^{2/3}\zeta(\mathbf{X})^{2/3}p^{2/3} + R(\mathbf{X})\Psi(Z)p$.

Viewing the operator norm $\|\mathbf{G}\| = \sup_{\|v\|=\|w\|=1} \operatorname{Re}(\langle v, \mathbf{G}w \rangle)$ as a Gaussian process, it follows from [13, Theorem 5.8] that $\mathbb{P}(\|\mathbf{G}\| - \mathbb{E}\|\mathbf{G}\| \geq x) \leq 2 \exp(-x^2/2\sigma_*^2(\mathbf{G}))$. Hence, using that sub-Gaussianity is equivalent to moment bounds [13, Theorem 2.1], we have that

$$\mathbb{E}[\|\mathbf{G}\| - \mathbb{E}\|\mathbf{G}\|]^{2p} \leq 2\sigma_*(\mathbf{G})\sqrt{p}. \quad (5.4)$$

Note that $\sigma_*(\mathbf{G}) = \sigma_*(\mathbf{S})$ since this quantity only depends on the covariance structure. Combining (5.3) and (5.4) using the triangle inequality for the L^p -norm hence yields (5.2). \square

PROOF OF COROLLARY 2.5. It is shown in [9, Theorem 2.3] that $|\mathbb{E}\|\mathbf{G}\| - \|\mathbf{G}_{\text{free}}\|| \leq Cv(\mathbf{S})^{1/2}\sigma(\mathbf{S})^{1/2}(\log d)^{3/4}$. The bounds in Corollary 2.5 are then immediate from Lemma 5.1 since $\|\mathbf{G}_{\text{free}}\| = \|\mathbf{S}_{\text{free}}\|$ and $\sigma_*(\mathbf{S}) \leq \min\{v(\mathbf{S}), \sigma(\mathbf{S})\} \leq v(\mathbf{S})^{1/2}\sigma(\mathbf{S})^{1/2}$. \square

PROOF OF COROLLARY 2.6. The matrix Khintchine inequality of Lust–Piquard [29], [51, Corollary 2.4] implies that $\mathbb{E}\|\mathbf{G}\| \leq C\sqrt{\ln(d+1)}\sigma(\mathbf{S})$ for some absolute constant $C > 0$. Substituting this in Lemma 5.1 with $p = \lceil \ln(d+1) \rceil$ and using that $\sigma_*(\mathbf{S}) \leq \sigma(\mathbf{S})$ yields the bound in Corollary 2.6 since $\mathbb{E}\|\mathbf{S}\| \leq \mathbb{E}\|\mathbf{S}\|^{2p}^{1/2p}$ by Jensen’s inequality. \square

Finally, the following result implies the tail bound that was claimed in (2.12).

PROPOSITION 5.2. *There exists an absolute constant $c > 0$ such that for every $0 < \delta \leq 1$ and $x > 0$ it holds with \mathcal{E} as in Corollary 2.5 that*

$$\mathbb{P}(\|\mathbf{S}\| \geq (1 + \delta)\|\mathbf{S}_{\text{free}}\| + c\mathcal{E}(x)) \leq (d + 1)(1 + \delta)^{-x}. \quad (5.5)$$

PROOF. Using the upper bound in Corollary 2.5 and Markov’s inequality, there exists an absolute constant $C > 0$ such that for every $y > 0$ and every integer $p \geq 1$,

$$\mathbb{P}(\|\mathbf{S}\| \geq y(\|\mathbf{S}_{\text{free}}\| + C\mathcal{E}(p))) \leq \mathbb{P}(\|\mathbf{S}\| \geq yd^{-1/2p}\mathbb{E}\|\mathbf{S}\|^{2p}) \leq dy^{-2p}. \quad (5.6)$$

Let $y := 1 + \delta$ and $p := \lceil x/2 \rceil$. We may assume without loss of generality that $x \geq \log_{1+\delta}(d+1)$, since (5.5) is vacuous otherwise. Using this in (5.6) as well as the fact that $x/2 \leq p \leq x$,

$$\mathbb{P}(\|\mathbf{S}\| \geq (1 + \delta)\|\mathbf{S}_{\text{free}}\| + (1 + \delta)C\mathcal{E}(x)) \leq (d + 1)(1 + \delta)^{-x}. \quad (5.7)$$

This yields (5.5) with $c := 2C$ since $1 + \delta \leq 2$ by the assumption that $\delta \leq 1$. \square

Acknowledgments. This work is part of the project Clustering and Spectral Concentration in Markov Chains with project number OCENW.KLEIN.324 of the research programme Open Competition Domain Science – M which is partly financed by the Dutch Research Council (NWO).

We thank Ramon van Handel for discussions at the 50th Probability Summer School of Saint–Flour which led to a simplified proof for Lemma 4.16. This manuscript further benefitted significantly from feedback from an anonymous referee, who we thank sincerely for an exceptionally helpful report.

Declarations. The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

- [1] ADAMCZAK, R. and KAVVADIAS, I. (2025). Matrix concentration inequalities for dependent binary random variables. *arXiv preprint arXiv:2504.08138*. <http://doi.org/10.48550/arXiv.2504.08138>.
- [2] AJANKI, O. H., ERDŐS, L. and KRÜGER, T. (2019). Stability of the matrix Dyson equation and random matrices with correlations. *Probability Theory and Related Fields*. <http://doi.org/10.1007/s00440-018-0835-z>.
- [3] ANDERSON, G. W., GUIONNET, A. and ZEITOUNI, O. (2010). *An introduction to random matrices*. Cambridge University Press <http://doi.org/10.1017/CBO9780511801334>.
- [4] AOUN, R., BANNA, M. and YOUSSEF, P. (2020). Matrix Poincaré inequalities and concentration. *Advances in Mathematics*. <http://doi.org/10.1016/j.aim.2020.107251>.
- [5] ARIZMENDI, O., HASEBE, T., LEHNER, F. and VARGAS, C. (2015). Relations between cumulants in non-commutative probability. *Advances in Mathematics*. <http://doi.org/10.1016/j.aim.2015.03.029>.
- [6] BACRY, E., GAÏFFAS, S. and MUZY, J. F. (2018). Concentration inequalities for matrix martingales in continuous time. *Probability Theory and Related Fields*. <http://doi.org/10.1007/s00440-017-0786-9>.
- [7] BAI, Z. and SILVERSTEIN, J. W. (2010). *Spectral analysis of large dimensional random matrices*. Springer <http://doi.org/10.1007/978-1-4419-0661-8>.
- [8] BAI, Z. D. and YIN, Y. Q. (1988). Necessary and sufficient conditions for almost sure convergence of the largest eigenvalue of a Wigner matrix. *The Annals of Probability*. <http://doi.org/10.1214/aop/1176991594>.
- [9] BANDEIRA, A. S., BOEDIHARDJO, M. T. and VAN HANDEL, R. (2023). Matrix concentration inequalities and free probability. *Inventiones Mathematicae*. <http://doi.org/10.1007/s00222-023-01204-6>.
- [10] BANDEIRA, A. S., CIPOLLONI, G., SCHRÖDER, D. and VAN HANDEL, R. (2024). Matrix Concentration Inequalities and Free Probability II. Two-sided Bounds and Applications. *arXiv preprint arXiv:2406.11453*. <http://doi.org/10.48550/arXiv.2406.11453>.
- [11] BANNA, M., MERLEVÈDE, F. and YOUSSEF, P. (2016). Bernstein-type inequality for a class of dependent random matrices. *Random Matrices: Theory and Applications*. <http://doi.org/10.1142/S2010326316500064>.
- [12] BLUM, J. R., HANSON, D. L. and KOOPMANS, L. H. (1963). On the Strong Law of Large numbers for a Class of Stochastic Processes. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*. <http://doi.org/10.1007/BF00535293>.
- [13] BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press <http://doi.org/10.1093/acprof:oso/9780199535255.001.0001>.
- [14] BRAILOVSKAYA, T. and VAN HANDEL, R. (2024). Universality and sharp matrix concentration inequalities. *Geometric and Functional Analysis*. <http://doi.org/10.1007/s00039-024-00692-9>.
- [15] EATON, M. L. (1983). *Multivariate statistics: a vector space approach*. Institute of Mathematical Statistics <http://doi.org/10.1214/lnms/1196285102>.
- [16] ERDŐS, L., KRÜGER, T. and SCHRÖDER, D. (2019). Random matrices with slow correlation decay. In *Forum of Mathematics, Sigma*. Cambridge University Press <http://doi.org/10.1017/fms.2019.2>.
- [17] FÉRAY, V. (2018). Weighted dependency graphs. *Electronic Journal of Probability*. <http://doi.org/10.1214/18-EJP222>.
- [18] FÜREDI, Z. and KOMLÓS, J. (1981). The eigenvalues of random symmetric matrices. *Combinatorica*. <http://doi.org/10.1007/BF02579329>.
- [19] GARG, A., LEE, Y. T., SONG, Z. and SRIVASTAVA, N. (2018). A matrix expander Chernoff bound. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing (STOC'18)*. <http://doi.org/10.1145/3188745.3188890>.
- [20] GOLUB, G. H. and VAN LOAN, C. F. (2013). *Matrix computations*, fourth ed. JHU press.
- [21] HAAGERUP, U. and THORBJØRNSSEN, S. (2005). A new application of random matrices: $\text{Ext}(C_{\text{red}}^*(F_2))$ is not a group. *Annals of Mathematics* 711–775. <https://www.jstor.org/stable/20159928>.
- [22] HAN, F. and LI, Y. (2020). Moment bounds for large autocovariance matrices under dependence. *Journal of Theoretical Probability*. <http://doi.org/10.1007/s10959-019-00922-z>.
- [23] HELTON, J. W., FAR, R. R. and SPEICHER, R. (2007). Operator-valued Semicircular Elements: Solving A Quadratic Matrix Equation with Positivity Constraints. *International Mathematics Research Notices*. <http://doi.org/10.1093/imrn/rnm086>.

- [24] JEDRA, Y., LEE, J., PROUTIERE, A. and YUN, S. Y. (2023). Nearly Optimal Latent State Decoding in Block MDPs. In *International Conference on Artificial Intelligence and Statistics*. <https://doi.org/10.48550/arXiv.2208.08480>.
- [25] KAUFMAN, T., KYNG, R. and SOLDÁ, F. (2022). Scalar and matrix Chernoff bounds from ℓ_∞ -independence. In *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)* 3732–3753. SIAM <http://doi.org/10.1137/1.9781611977073.147>.
- [26] KYNG, R. and SONG, Z. (2018). A matrix Chernoff bound for strongly Rayleigh distributions and spectral sparsifiers from a few random spanning trees. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)* 373–384. IEEE <http://doi.org/10.1109/FOCS.2018.00043>.
- [27] LEHNER, F. (1999). Computing norms of free operators with matrix coefficients. *American Journal of Mathematics*. <http://doi.org/10.1353/ajm.1999.0022>.
- [28] LEVIN, D. A. and PERES, Y. (2017). *Markov chains and mixing times*, Second ed. American Mathematical Society.
- [29] LUST-PIQUARD, F. (1986). Inégalités de Khintchine dans C_p ($1 < p < \infty$). *Comptes Rendus de l'Académie des Sciences Paris*.
- [30] MACKEY, L., JORDAN, M. I., CHEN, R. Y., FARRELL, B. and TROPP, J. A. (2014). Matrix concentration inequalities via the method of exchangeable pairs. *The Annals of Probability*. <http://doi.org/10.1214/13-AOP892>.
- [31] MINGO, J. A. and SPEICHER, R. (2017). *Free probability and random matrices*. Springer <http://doi.org/10.1007/978-1-4939-6942-5>.
- [32] NEEMAN, J., SHI, B. and WARD, R. (2024). Concentration inequalities for sums of Markov dependent random matrices. *Information and Inference: a Journal of the IMA*. <http://doi.org/10.1093/imaiai/iaae032>.
- [33] NELSON, E. (1974). Notes on non-commutative integration. *Journal of Functional Analysis*. [http://doi.org/10.1016/0022-1236\(74\)90014-7](http://doi.org/10.1016/0022-1236(74)90014-7).
- [34] NICA, A. and SPEICHER, R. (2006). *Lectures on the Combinatorics of Free Probability*. Cambridge University Press <http://doi.org/10.1017/CBO9780511735127>.
- [35] OLIVEIRA, R. I. (2009). Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges. *arXiv preprint arXiv:0911.0600*. <http://doi.org/10.48550/arXiv.0911.0600>.
- [36] PAULIN, D., MACKEY, L. and TROPP, J. A. (2016). Efron–Stein inequalities for random matrices. *The Annals of Probability*. <http://doi.org/10.1214/15-AOP1054>.
- [37] PISIER, G. (2003). *Introduction to Operator Space Theory*. Cambridge University Press <http://doi.org/10.1017/CBO9781107360235>.
- [38] QIU, J., WANG, C., LIAO, B., PENG, R. and TANG, J. (2020). A matrix Chernoff bound for Markov chains and its application to co-occurrence matrices. *Advances in Neural Information Processing Systems*. <http://doi.org/10.48550/arXiv.2008.02464>.
- [39] RAO, S. (2018). A Hoeffding inequality for Markov chains. *Electronic Communications in Probability*. <http://doi.org/10.1214/19-ECP219>.
- [40] SAMSON, P. M. (2000). Concentration of measure inequalities for Markov chains and Φ -mixing processes. *The Annals of Probability*. <http://doi.org/10.1214/aop/1019160125>.
- [41] SANDERS, J., PROUTIERE, A. and YUN, S. Y. (2020). Clustering in block Markov chains. *The Annals of Statistics*. <http://doi.org/10.1214/19-AOS1939>.
- [42] SANDERS, J. and SENEN-CERDA, A. (2023). Spectral norm bounds for block Markov chain random matrices. *Stochastic Processes and their Applications*. <http://doi.org/10.1016/j.spa.2022.12.004>.
- [43] SANDERS, J. and SENEN-CERDA, A. (2023). Spectral norm bounds for block Markov chain random matrices. *Stochastic Processes and their Applications*. <http://doi.org/10.1016/j.spa.2022.12.004>.
- [44] SANDERS, J. and VAN WERDE, A. (2023). Singular value distribution of dense random matrices with block Markovian dependence. *Stochastic Processes and their Applications*. <http://doi.org/10.1016/j.spa.2023.01.001>.
- [45] SAULIS, L. and STATULEVIČIUS, V. A. (1991). *Limit theorems for large deviations*. Springer Science & Business Media <http://doi.org/10.1007/978-94-011-3530-6>.
- [46] SODIN, S. (2010). The spectral edge of some random band matrices. *Annals of Mathematics*. <https://www.jstor.org/stable/29764668>.
- [47] SPEICHER, R. and WOROUDI, R. (1997). Boolean convolutions. *Fields Institute Communications*. <http://doi.org/10.1090/fic/012/13>.
- [48] STATULEVIČIUS, V. (1969, 1970). Limit theorems for the sums of random variables related to a Markov chain. I, II, III. *Lithuanian Mathematical Journal*.

- [49] TROPP, J. A. (2011). Freedman's inequality for matrix martingales. *Electronic Communications in Probability*. <http://doi.org/10.1214/ECP.v16-1624>.
- [50] TROPP, J. A. (2015). An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*. <http://doi.org/10.1561/22000000048>.
- [51] TROPP, J. A. (2018). Second-order matrix concentration inequalities. *Applied and Computational Harmonic Analysis*. <http://doi.org/10.1016/j.acha.2016.07.005>.
- [52] VAN VUREN, T., CRONK, T. and SANDERS, J. (2024). Estimating the number of clusters of a block Markov chain. *arXiv preprint arXiv:2407.18287*. <https://doi.org/10.48550/arXiv.2407.18287>.
- [53] VAN WERDE, A., SENEN CERDA, A., KOSMELLA, G. and SANDERS, J. (2022). Detection and Evaluation of Clusters within Sequential Data. *arXiv preprint arXiv:2210.01679*. <http://doi.org/10.48550/arXiv.2210.01679>.
- [54] ZHANG, A. and WANG, M. (2020). Spectral state compression of Markov processes. *IEEE Transactions on Information Theory*. <http://doi.org/10.1109/TIT.2019.2956737>.

APPENDIX A: PROPERTIES OF THE ψ -DEPENDENCE COEFFICIENT

In this section, we prove some properties of the ψ -dependence coefficient that may be helpful in the estimation of the parameters when one needs to apply our results. First, we prove a bound on $\Psi(Z)$ in terms of the total variation mixing time of the Markov chain in Section A.1. Second, we prove bounds on $\sigma(\mathbf{S})$ and $v(\mathbf{S})$ in Section A.2.

A.1. Bound on $\Psi(Z)$ in terms of total variation mixing time. Another common way to quantify the decay of dependence in a Markov chain is the *total variation mixing time* [28, Section 4.5]. As was announced in Remark 2.3, one can bound $\Psi(Z)$ in terms of the mixing time whenever the state space is finite. We prove this claim in Proposition A.2.

LEMMA A.1. *Let $Z = (Z_i)_{i=1}^n$ be an ergodic stationary Markov chain on a finite state space \mathcal{Z} . Denote $\mathbf{P} \in [0, 1]^{\mathcal{Z} \times \mathcal{Z}}$ and $\pi \in [0, 1]^{\mathcal{Z}}$ for the transition matrix and stationary distribution of Z . Then,*

$$\Psi(Z) = \min \left\{ n, \min \left\{ t \geq 1 : \max_{i,j \in \mathcal{Z}} \left| \frac{(\mathbf{P}^t)_{i,j} - \pi_j}{\pi_j} \right| \leq \frac{1}{4} \right\} \right\}. \quad (\text{A.1})$$

PROOF. We claim that if X and Y are random variables with values supported on a finite set \mathcal{Z} , then the suprema in (2.4) are realized by singleton sets:

$$\psi(X, Y) = \max_{x \in \mathcal{Z}, y \in \mathcal{Z}} \left| \frac{\mathbb{P}(X = x, Y = y) - \mathbb{P}(X = x)\mathbb{P}(Y = y)}{\mathbb{P}(X = x)\mathbb{P}(Y = y)} \right|. \quad (\text{A.2})$$

To see this, note that for any absolutely continuous measures $\mu \ll \nu$,

$$\sup_{E: \nu(E) > 0} \left| \frac{\mu(E)}{\nu(E)} - 1 \right| = \sup_{E: \nu(E) > 0} \frac{1}{\nu(E)} \left| \int \mathbb{1}_E \left(\frac{d\mu}{d\nu} - 1 \right) d\nu \right| \leq \left\| \frac{d\mu}{d\nu} - 1 \right\|_{L^\infty}. \quad (\text{A.3})$$

Applying this with $\mu := \mathbb{P}_{X,Y}$ and $\nu := \mathbb{P}_X \otimes \mathbb{P}_Y$ shows that the left-hand side of (A.2) is no greater than the right-hand side. The other inequality follows by using singleton sets in (2.4). Using (A.2) in the definition (2.4) of $\Psi(Z)$ now yields (A.1). \square

The *total variation distance* between two probability measures μ, ν on the same space \mathcal{Z} is

$$d_{\text{TV}}(\mu, \nu) := \max_{E \subseteq \mathcal{Z}} |\mu(E) - \nu(E)|. \quad (\text{A.4})$$

For any $\varepsilon > 0$ the ε -*mixing time* of an ergodic Markov chain Z on a finite state space is defined as $t_{\text{mix}}(\varepsilon) := \min\{t \geq 1 : d(t) \leq \varepsilon\}$ where $d(t) := \sup_{i \in \mathcal{Z}} d_{\text{TV}}(\mathbb{P}(Z_t = \cdot | Z_0 = i), \pi)$. We refer to $t_{\text{mix}} := t_{\text{mix}}(1/4)$ as the *total variation mixing time* of Z .

PROPOSITION A.2. *Let Z and $\pi \in [0, 1]^{\mathcal{Z}}$ be as in Lemma A.1 and denote $\pi_{\min} := \min\{\pi_x : x \in \mathcal{Z}\}$. Then, $\Psi(Z) \leq (\log_2(1/\pi_{\min}) + 3)t_{\text{mix}}$*

PROOF. By taking $E = \{j\}$ in (A.4) it follows that for every $t \geq 1$, $\max_{i,j \in \mathcal{Z}} |\mathbf{P}_{i,j}^t - \pi_j| \leq d(t)$. Denote $\ell := \lceil \log_2(1/\pi_{\min}) + 2 \rceil$ and note that $\ell \geq \log_2(1/\pi_{\min}) + 2$. It consequently follows from [28, (4.35)] that $d(\ell t_{\text{mix}}) \leq 2^{-\ell} \leq 4^{-1} \pi_{\min}$. Hence,

$$\max_{i,j \in \mathcal{Z}} |(\mathbf{P}_{i,j}^{\ell t_{\text{mix}}} - \pi_j)/\pi_j| \leq 1/4. \quad (\text{A.5})$$

The desired result now follows from Lemma A.1 and the fact that $\ell \leq \log_2(1/\pi_{\min}) + 3$. \square

A.2. Bounds on $\sigma(\mathbf{S})$ and $v(\mathbf{S})$. We now prove Remark 2.3 in Lemmas A.3 and A.4.

LEMMA A.3. *It holds that $\sigma(\mathbf{S})^2 \leq 3\Psi(Z)\zeta(\mathbf{X})^2$.*

PROOF. By expanding the summation in the definition of \mathbf{S} and regrouping terms,

$$\mathbb{E}[(\mathbf{S} - \mathbb{E}[\mathbf{S}])^2] = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}[\mathbf{X}_i \mathbf{X}_j + \mathbf{X}_j \mathbf{X}_i]. \quad (\text{A.6})$$

For any i, j with $\psi(Z_i, Z_j) < \infty$ using the definition that $\mathbf{X}_i = \mathbf{F}_i(Z_i)$,

$$\begin{aligned} \mathbb{E}[\mathbf{X}_i \mathbf{X}_j + \mathbf{X}_j \mathbf{X}_i] &= \int_{\mathcal{Z}_i \times \mathcal{Z}_j} (\mathbf{F}_i(z_i) \mathbf{F}_j(z_j) + \mathbf{F}_j(z_j) \mathbf{F}_i(z_i)) d\mathbb{P}_{Z_i, Z_j}(z_i, z_j) \\ &= \int_{\mathcal{Z}_i \times \mathcal{Z}_j} (\mathbf{F}_i(z_i) \mathbf{F}_j(z_j) + \mathbf{F}_j(z_j) \mathbf{F}_i(z_i)) \frac{d\mathbb{P}_{Z_i, Z_j}}{d\mathbb{P}_{Z_i} \otimes \mathbb{P}_{Z_j}}(z_i, z_j) d(\mathbb{P}_{Z_i} \otimes \mathbb{P}_{Z_j})(z_i, z_j). \end{aligned} \quad (\text{A.7})$$

Let \tilde{Z}_i and \tilde{Z}_j be independent random variables with the same marginal distribution as Z_i and Z_j , respectively. Then, with $\Delta_{i,j} := \frac{d\mathbb{P}_{Z_i, Z_j}}{d\mathbb{P}_{Z_i} \otimes \mathbb{P}_{Z_j}}(\tilde{Z}_i, \tilde{Z}_j) - 1$ and $\tilde{\mathbf{X}}_i = \mathbf{F}_i(\tilde{Z}_i)$, it follows from (A.7) and the assumption that \mathbf{X}_i is centered that

$$\mathbb{E}[\mathbf{X}_i \mathbf{X}_j + \mathbf{X}_j \mathbf{X}_i] = \mathbb{E}[(1 + \Delta_{i,j})(\tilde{\mathbf{X}}_i \tilde{\mathbf{X}}_j + \tilde{\mathbf{X}}_j \tilde{\mathbf{X}}_i)] = \mathbb{E}[\Delta_{i,j}(\tilde{\mathbf{X}}_i \tilde{\mathbf{X}}_j + \tilde{\mathbf{X}}_j \tilde{\mathbf{X}}_i)]. \quad (\text{A.8})$$

In general, for any $\mathbf{A}, \mathbf{B} \in \mathbb{C}_{\text{sa}}^{d \times d}$ and scalar $\delta \in \mathbb{R}$, we have $\delta(\mathbf{A}\mathbf{B} + \mathbf{B}\mathbf{A}) \preceq |\delta|(\mathbf{A}^2 + \mathbf{B}^2)$ with respect to the positive semidefinite order.⁷ Hence, using Proposition 4.3 and (A.8) if and only if $\psi(Z_i, Z_j) < 1$, and applying the inequality directly otherwise,

$$\mathbb{E}[\mathbf{X}_i \mathbf{X}_j + \mathbf{X}_j \mathbf{X}_i] \preceq \min\{1, \psi(Z_i, Z_j)\} (\mathbb{E}[\mathbf{X}_i^2] + \mathbb{E}[\mathbf{X}_j^2]) \quad (\text{A.9})$$

for all $i, j \in \{1, \dots, n\}$. Substitution in (A.6) and reorganizing the terms again yields that

$$\mathbb{E}[(\mathbf{S} - \mathbb{E}[\mathbf{S}])^2] \preceq \sum_{i=1}^n \sum_{j=1}^n \min\{1, \psi(Z_i, Z_j)\} \mathbb{E}[\mathbf{X}_i^2]. \quad (\text{A.10})$$

Finally, using that $\min\{1, \psi(Z_i, Z_j)\} \leq (1/4)^{\lfloor |i-j|/\Psi(Z) \rfloor}$ by Proposition 4.4, we here have that $\sum_{j=1}^n \min\{1, \psi(Z_i, Z_j)\} \leq 2\Psi(Z) \sum_{v=0}^{\infty} (1/4)^v \leq 3\Psi(Z)$. The claimed estimate hence follows from the fact that the operator norm respects the positive semidefinite order when restricted to the set of positive semidefinite matrices. \square

LEMMA A.4. *It holds that $v(\mathbf{S})^2 \leq 3\Psi(Z) \|\sum_{i=1}^n \text{Cov}(\mathbf{X}_i)\|$.*

PROOF. For any random vector V , it holds that $\|\text{Cov}(V)\| = \sup_{\|W\| \leq 1} \mathbb{E}[|\langle V, W \rangle|^2]$. In particular, $\|\text{Cov}(\mathbf{S})\| = \sup_{\text{Tr}[\mathbf{M}]^2 \leq 1} \mathbb{E}[|\text{Tr}[\mathbf{S}\mathbf{M}]|^2]$ with $\text{Tr}[\mathbf{M}] = \sum_i \mathbf{M}_{i,i}$ the unnormalized trace. Here, proceeding similarly to (A.6)–(A.10),

$$\mathbb{E}[|\text{Tr}[\mathbf{S}\mathbf{M}]|^2] \leq \sum_{i=1}^n \sum_{j=1}^n \min\{1, \psi(Z_i, Z_j)\} \mathbb{E}[|\text{Tr}[\mathbf{X}_i \mathbf{M}]|^2]. \quad (\text{A.11})$$

Thus, taking the supremum over \mathbf{M} and bounding $\sum_{j=1}^n \min\{1, \psi(Z_i, Z_j)\} \leq 3\Psi(Z)$, we have $\|\text{Cov}(\mathbf{S})\| \leq 3\Psi(Z) \|\sum_i \text{Cov}(\mathbf{X}_i)\|$. This yields the desired result. \square

⁷This follows by using that $\delta(\mathbf{A} - \mathbf{B})^2 \succeq 0$ if $\delta \geq 0$ and by using that $\delta(\mathbf{A} + \mathbf{B})^2 \preceq 0$ if $\delta \leq 0$.

APPENDIX B: PROOF OF LEMMA 3.2

The following is a special case of a result by Samson [40].

LEMMA B.1. *Consider scalar random variables of the form $Y_i = f_i(Z_i)$ for functions $f_i : \mathcal{Z}_i \rightarrow [0, 1]$, and consider deterministic matrices $\mathbf{B}_1, \dots, \mathbf{B}_n \in \mathbb{R}^{d \times d}$. Then, there exists an absolute constant $C > 0$ such that for every $x \geq 0$,*

$$\mathbb{P}\left(\left\|\sum_{i=1}^n Y_i \mathbf{B}_i\right\| - \mathbb{E}\left\|\sum_{i=1}^n Y_i \mathbf{B}_i\right\| \geq x\right) \leq \exp\left(-C \frac{x^2}{\varsigma_*^2(\mathbf{B})^2 \Psi(Z)}\right) \quad (\text{B.1})$$

where $\varsigma_*^2(\mathbf{B}) := \sup_{\|v\|=\|w\|=1} \sum_{i=1}^n \langle v, \mathbf{B}_i w \rangle^2$.

PROOF. We use [40, (2.23)], which provides a concentration-of-measure principle for random sums of deterministic vectors b_1, \dots, b_n in an arbitrary Banach space $(B, \|\cdot\|)$:

$$\mathbb{P}\left(\left\|\sum_{i=1}^n Y_i b_i\right\| - \mathbb{E}\left\|\sum_{i=1}^n Y_i b_i\right\| \geq x\right) \leq \exp\left(-\frac{x^2}{8\varsigma_*(b)^2 \|\Gamma\|^2}\right) \quad (\text{B.2})$$

where $\|\Gamma\|$ is a dependence quantity [40, Section 2] and $\varsigma_*(b)^2 := \sup_{\xi \in B^* : \|\xi\| \leq 1} \sum_{i=1}^n \xi(b_i)^2$ with supremum running over linear functionals $\xi : B \rightarrow \mathbb{R}$ of norm ≤ 1 .

Suppose that the Banach space is $\mathbb{R}^{d \times d}$ with the operator norm. Then, the linear functionals of norm ≤ 1 are convex combinations of those of the form $\xi(\mathbf{M}) = \langle v, \mathbf{M}w \rangle$ for fixed vectors $v, w \in \mathbb{R}^d$ with $\|v\| = \|w\| = 1$. Hence, taking $b_i = \mathbf{B}_i$ and using that $\sum_{i=1}^n \xi(b_i)^2$ depends convexly on ξ yields the variance proxy specified in Lemma B.1.

Regarding the dependence quantity $\|\Gamma\|$, it suffices for our purposes that it can be bounded whenever the dependence in the Markov chain decays at an exponential rate. Specifically, since the total variation distance between the laws of Z_i and Z_j is at most $\min\{2, \psi(Z_i, Z_j)\}$ it follows from Proposition 4.4 with the same argumentation as in [40, pages 421 to 422] that

$$\|\Gamma\| \leq \sum_{k=0}^{n-1} \sqrt{\min\{2, (1/4)^{k/\Psi(Z)}\}}. \quad (\text{B.3})$$

Thus, $\|\Gamma\| \leq c\Psi(Z)$ for some absolute constant $c > 0$. Use this in (B.2) to conclude. \square

COROLLARY B.2. *With notation as in Lemma B.1, there exists an absolute constant $c > 0$ such that for any integer $p \geq 1$,*

$$\left| \mathbb{E}\left[\left\|\sum_{i=1}^n Y_i \mathbf{B}_i\right\|^{2p}\right] - \mathbb{E}\left[\left\|\sum_{i=1}^n Y_i \mathbf{B}_i\right\|^{2p}\right]^{1/2p} \right| \leq c\sqrt{p\Psi(Z)\varsigma_*(\mathbf{B})}. \quad (\text{B.4})$$

PROOF. This is immediate from Lemma B.1 since sub-Gaussianity is equivalent to moment bounds [13, Theorem 2.1]. \square

PROOF OF LEMMA 3.2. Note that the model (3.1) is of the form described in Lemma B.1, with $\mathbf{B}_t = e_i e_j^\top + e_j e_i^\top$ or $\mathbf{B}_t = e_i e_i^\top$ depending on $\varphi(t) = \{i, j\}$. Here,

$$\varsigma_*^2(\mathbf{B}) = \sup_{\|v\|=\|w\|=1} \sum_{i \leq j} \langle v, \mathbf{B}_{\varphi^{-1}(i,j)} w \rangle^2 = \sup_{\|v\|=\|w\|=1} \sum_{i=1}^n \sum_{j=1}^n v_i^2 w_j^2 = 1 \quad (\text{B.5})$$

where we used that $\|v_i\| = \|w_j\| = 1$. Thus, using Lemma B.1 and Corollary B.2 with the triangle inequality, there exist $c, C > 0$ such that for every $p \geq 1$,

$$\mathbb{P}(\|\mathbf{S}\| - \mathbb{E}\|\mathbf{S}\|^{2p}]^{1/2p} \geq c\sqrt{p\Psi(Z)} + x) \leq \exp(-Cx^2/\Psi(Z)). \quad (\text{B.6})$$

On the other hand, taking p a sufficiently large multiple of $\ln(d+1)$ depending on the desired multiplicative error δ , and using the two-sided bounds in Corollary 2.5 with the parameter estimates from Lemma 3.1 and the assumption that $\|\mathbf{S}_{i,j}\|_{L^\infty} \leq 1$,

$$\begin{aligned} & \max_{-\delta \leq \gamma \leq \delta} \left| \mathbb{E}[\|\mathbf{S}\|^{2p}]^{1/2p} - (1+\gamma)\|\mathbf{S}_{\text{free}}\| \right| \\ & \leq c\Psi(Z)^{\frac{2}{3}}d^{\frac{1}{3}}\ln(d+1)^{\frac{2}{3}} + c\Psi(Z)\ln(d+1) + c\Psi(Z)^{\frac{1}{2}}d^{\frac{1}{4}}\ln(d+1)^{\frac{3}{4}} \end{aligned} \quad (\text{B.7})$$

for some $c > 0$. Combine (B.6)–(B.7) and simplify using that $\sqrt{\ln(d+1)\Psi(Z)} \leq \ln(d+1)\Psi(Z)$ and $\Psi(Z)^{1/2}d^{1/4}\ln(d+1)^{3/4} \leq c'\Psi(Z)^{2/3}d^{1/3}\ln(d+1)^{2/3}$ for some $c' > 0$. \square

APPENDIX C: PROOFS CONCERNING BLOCK MARKOV CHAINS

This section concerns the proofs for Section 3.2. We adopt the notation that was used there. In particular, $\mathbf{M} = \sqrt{d/n}(\hat{\mathbf{N}} - \mathbb{E}[\hat{\mathbf{N}}])$, and \mathbf{S} is the self-adjoint dilation defined in (3.7).

This section is split in the following main parts. In Section C.1, we estimate the entries of $\text{Cov}(\mathbf{M})$ and use these to provide precise estimates on the parameters of \mathbf{S} . We prove Proposition 3.6 in Section C.2. Finally, the proof of Theorem 3.5 is given in Section C.3 where we also establish a nonasymptotic concentration inequality in Proposition C.13.

C.1. Estimates on the parameters of block Markovian random matrices. For any fixed $d, n \geq 1$, we introduce the following abbreviations

$$\mathbf{c}_1 := d \max \left\{ \frac{\pi_v}{\#\mathcal{V}_v} : v \in \{1, \dots, K\} \right\}, \quad (\text{C.1})$$

$$\mathbf{c}_2 := d^2 \max \left\{ \frac{\pi_v}{\#\mathcal{V}_v} \frac{\mathbf{P}_{v,w}}{\#\mathcal{V}_w} : v, w \in \{1, \dots, K\} \right\}, \quad (\text{C.2})$$

$$\mathbf{c}_3 := d^3 \max \left\{ \frac{\pi_u}{\#\mathcal{V}_u} \frac{\mathbf{P}_{u,v}}{\#\mathcal{V}_v} \frac{\mathbf{P}_{v,w}}{\#\mathcal{V}_w} : u, v, w \in \{1, \dots, K\} \right\}, \quad (\text{C.3})$$

$$\mathfrak{d} := d^2 \max \left\{ \left| \frac{1}{n} \sum_{t=1}^{n-3} (n-2-t) \frac{(\mathbf{P}^t)_{u,v} - \pi_v}{\#\mathcal{V}_v} \middle| \frac{\mathbf{P}_{v,w}}{\#\mathcal{V}_w} : u, v, w \in \{1, \dots, K\} \right\}. \quad (\text{C.4})$$

Here, we let $\mathfrak{d} = 0$ if $n-3 < 1$. Simple upper bounds on the \mathbf{c}_i and \mathfrak{d} are given in Lemma C.4.

LEMMA C.1. *For any $i, j, k, m \in \{1, \dots, d\}$ it holds that*

$$\begin{aligned} & |\text{Cov}(\mathbf{M})_{ij,km}| \\ & \leq \begin{cases} \frac{1}{d}(\mathbf{c}_2 + \frac{2}{d^2}\mathbf{c}_2^2 + \frac{2}{d}\mathbf{c}_3 + \frac{2}{d^2}\mathbf{c}_2\mathfrak{d}) & \text{if } (i, j) = (k, m), \\ \frac{1}{d^2}(\frac{3}{d}\mathbf{c}_2^2 + 2\mathbf{c}_3 + \frac{2}{d}\mathbf{c}_2\mathfrak{d}) & \text{if } (i, j) \neq (k, m) \text{ and } (i = m \text{ or } j = k), \\ \frac{1}{d^3}(3\mathbf{c}_2^2 + 2\mathbf{c}_2\mathfrak{d}) & \text{else.} \end{cases} \end{aligned} \quad (\text{C.5})$$

Furthermore, if $i \in \mathcal{V}_a$ and $j \in \mathcal{V}_b$ for $a, b \in \{1, \dots, K\}$, then

$$\left| \text{Cov}(\mathbf{M})_{ij,ij} - d \frac{\pi_a \mathbf{P}_{a,b}}{\#\mathcal{V}_a \#\mathcal{V}_b} \right| \leq \frac{1}{d^2} \left(\frac{d}{n} \mathbf{c}_2 + \frac{3}{d} \mathbf{c}_2^2 + 2\mathbf{c}_3 + \frac{2}{d} \mathbf{c}_2 \mathfrak{d} \right). \quad (\text{C.6})$$

PROOF. Recall that the Markov chain of transitions $E = (E_t)_{t=1}^{n-1}$ is defined by $E_t = (Z_t, Z_{t+1})$. For any $i, j, k, m \in \{1, \dots, n\}$ we can write

$$\text{Cov}(\mathbf{M})_{ij,km} = \frac{d}{n} \mathbb{E} \left[(\hat{\mathbf{N}} - \mathbb{E}[\hat{\mathbf{N}}])_{i,j} (\hat{\mathbf{N}} - \mathbb{E}[\hat{\mathbf{N}}])_{k,m} \right] \quad (\text{C.7})$$

$$\begin{aligned}
&= \frac{d}{n} \sum_{t_1, t_2=1}^{n-1} \mathbb{E} \left[\mathbb{1}\{E_{t_1} = (i, j)\} \mathbb{1}\{E_{t_2} = (k, m)\} - \mathbb{P}(E_{t_1} = (i, j)) \mathbb{P}(E_{t_2} = (k, m)) \right] \\
&=: \frac{d}{n} \sum_{t_1, t_2=1}^{n-1} \mathcal{E}_{t_1, t_2}(i, j, k, m).
\end{aligned}$$

Let us separately consider the case where $t_1 = t_2$, $|t_1 - t_2| = 1$, and $|t_1 - t_2| > 1$. If $t_1 = t_2$, then since the block Markov chain starts from stationarity,

$$\frac{d}{n} \sum_{t_1=1}^{n-1} \mathcal{E}_{t_1, t_1}(i, j, k, m) = \begin{cases} -\frac{d(n-1)}{n} \mathbb{P}(E_1 = (i, j)) \mathbb{P}(E_1 = (k, m)) & \text{if } (i, j) \neq (k, m), \\ \frac{d(n-1)}{n} (\mathbb{P}(E_1 = (i, j)) - \mathbb{P}(E_1 = (i, j))^2) & \text{if } (i, j) = (k, m), \end{cases} \quad (\text{C.8})$$

It here holds for $i \in \mathcal{V}_a$ and $j \in \mathcal{V}_b$ that $\mathbb{P}(E_1 = (i, j)) = \pi_a \mathbf{P}_{a,b} / \#\mathcal{V}_a \#\mathcal{V}_b \leq \mathbf{c}_2 / d^2$. Hence, using that $\mathbb{P}(E_1 = (i, j)) - \mathbb{P}(E_1 = (i, j))^2 \leq \mathbb{P}(E_1 = (i, j))$ and $(n-1)/n \leq 1$,

$$\left| \frac{d}{n} \sum_{t_1=1}^{n-1} \mathcal{E}_{t_1, t_1}(i, j, k, m) \right| \leq \begin{cases} \frac{1}{d^3} \mathbf{c}_2^2, & \text{if } (i, j) \neq (k, m), \\ \frac{1}{d} \mathbf{c}_2 & \text{if } (i, j) = (k, m). \end{cases} \quad (\text{C.9})$$

Furthermore, for $i \in \mathcal{V}_a$ and $j \in \mathcal{V}_b$,

$$\begin{aligned}
&\left| \frac{d}{n} \sum_{t_1=1}^{n-1} \mathcal{E}_{t_1, t_1}(i, j, i, j) - d \frac{\pi_a \mathbf{P}_{a,b}}{\#\mathcal{V}_a \#\mathcal{V}_b} \right| & (\text{C.10}) \\
&\leq \left(\frac{dn}{n} - \frac{d(n-1)}{n} \right) \mathbb{P}(E_1 = (i, j)) + \frac{d(n-1)}{n} \mathbb{P}(E_1 = (i, j))^2 \leq \frac{\mathbf{c}_2}{dn} + \frac{\mathbf{c}_2^2}{d^3}.
\end{aligned}$$

Now suppose that $|t_1 - t_2| = 1$. By symmetry, it suffices to consider $t_2 = t_1 + 1$. Then,

$$\begin{aligned}
&\left| \frac{d}{n} \sum_{t_1=1}^{n-2} \mathcal{E}_{t_1, t_1+1}(i, j, k, m) \right| & (\text{C.11}) \\
&= \begin{cases} \frac{d(n-2)}{n} \mathbb{P}(E_1 = (i, j)) \mathbb{P}(E_2 = (k, m)) & \text{if } j \neq k, \\ \frac{d(n-2)}{n} |\mathbb{P}(E_1 = (i, j), E_2 = (j, m)) - \mathbb{P}(E_1 = (i, j)) \mathbb{P}(E_2 = (j, m))| & \text{if } j = k. \end{cases}
\end{aligned}$$

A similar estimate applies when $t_1 = t_2 + 1$. The only difference is that the case distinction then depends on whether $i = m$ or $i \neq m$ because $\mathcal{E}_{t_1, t_2}(i, j, k, m) = \mathcal{E}_{t_2, t_1}(k, m, i, j)$. Hence,

$$\left| \frac{d}{n} \sum_{|t_1 - t_2|=1} \mathcal{E}_{t_1, t_2}(i, j, k, m) \right| \leq \begin{cases} \frac{2}{d^3} \mathbf{c}_2^2, & \text{if } j \neq k \text{ and } i \neq m, \\ \frac{2}{d^2} (\mathbf{c}_3 + \frac{1}{d} \mathbf{c}_2^2) & \text{if } j = k \text{ or } i = m. \end{cases} \quad (\text{C.12})$$

Finally, consider the case where $|t_1 - t_2| > 1$. Then, if $t_2 > t_1 + 1$,

$$\begin{aligned}
&\left| \frac{d}{n} \sum_{t_1=1}^{n-3} \sum_{t_2 > t_1+1} \mathcal{E}_{t_1, t_2}(i, j, k, m) \right| = \left| \frac{d}{n} \sum_{s=2}^{n-2} (n-1-s) \mathcal{E}_{1, 1+s}(i, j, k, m) \right| & (\text{C.13}) \\
&= d \mathbb{P}(E_1 = (i, j)) \mathbb{P}(Z_{2+s} = m \mid Z_{1+s} = k) \\
&\quad \times \left| \sum_{s=2}^{n-2} \frac{n-1-s}{n} (\mathbb{P}(Z_{1+s} = k \mid Z_2 = j) - \mathbb{P}(Z_{1+s} = k)) \right|.
\end{aligned}$$

Recall (C.4) and note that $\mathbb{P}(Z_{1+s} = k \mid Z_2 = j) = (\mathbf{p}^{s-1})_{a,b}/\#\mathcal{V}_b$ for any $j \in \mathcal{V}_a$ and $k \in \mathcal{V}_b$. A substitution $t = s - 1$ hence yields that $\left| \frac{d}{n} \sum_{t_1=1}^{n-3} \sum_{t_2>t_1+1} \mathcal{E}_{t_1,t_2}(i, j, k, m) \right| \leq \frac{1}{d^3} \mathbf{c}_2 \mathfrak{d}$. By using a similar estimate for the case $t_1 > t_2 + 1$ we conclude that

$$\left| \frac{d}{n} \sum_{|t_1-t_2|>1} \mathcal{E}_{t_1,t_2}(i, j, k, m) \right| \leq \frac{2}{d^3} \mathbf{c}_2 \mathfrak{d}. \quad (\text{C.14})$$

We finally combine the foregoing estimates. Due to (C.7) it holds that

$$\begin{aligned} |\text{Cov}(\mathbf{M})_{ij,km}| &\leq \left| \frac{d}{n} \sum_{t_1=1}^{n-1} \mathcal{E}_{t_1,t_1}(i, j, k, m) \right| + \left| \frac{d}{n} \sum_{|t_1-t_2|=1} \mathcal{E}_{t_1,t_2}(i, j, k, m) \right| \\ &\quad + \left| \frac{d}{n} \sum_{|t_1-t_2|>1} \mathcal{E}_{t_1,t_1+1}(i, j, k, m) \right|. \end{aligned} \quad (\text{C.15})$$

Estimate the terms in the right-hand side of (C.15) using (C.9), (C.12), and (C.14) to complete the proof of (C.5). Similarly, the proof of (C.6) can be completed by combining (C.10), (C.12), and (C.14) together with the fact that

$$\begin{aligned} \left| \text{Cov}(\mathbf{M})_{ij,ij} - d \frac{\pi_a \mathbf{P}_{a,b}}{\#\mathcal{V}_a \#\mathcal{V}_b} \right| &\leq \left| \frac{d}{n} \sum_{t_1=1}^{n-1} \mathcal{E}_{t_1,t_1}(i, j, i, j) - d \frac{\pi_a \mathbf{P}_{a,b}}{\#\mathcal{V}_a \#\mathcal{V}_b} \right| \\ &\quad + \left| \frac{d}{n} \sum_{|t_1-t_2|=1} \mathcal{E}_{t_1,t_2}(i, j, i, j) \right| + \left| \frac{d}{n} \sum_{|t_1-t_2|>1} \mathcal{E}_{t_1,t_1+1}(i, j, i, j) \right|. \end{aligned} \quad (\text{C.16})$$

This concludes the proof. \square

We next show that one can estimate $\Psi(Z)$ and $\Psi(E)$ in terms of $\Psi(\mathbf{p})$.

LEMMA C.2. *It holds that $\Psi(Z) = \min\{n, \Psi(\mathbf{p})\}$.*

PROOF. For any $i \in \mathcal{V}_a$ and $j \in \mathcal{V}_b$ it holds that $\mathbb{P}(Z_{1+t} = j \mid Z_1 = i) = (\mathbf{p}^t)_{a,b}/\#\mathcal{V}_b$. The desired estimate is hence immediate from Lemma A.1. \square

LEMMA C.3. *It holds that $\Psi(E) \leq \Psi(Z) + 1$.*

PROOF. Since Z starts in stationarity, the same holds for E . Further, for any $t > 1$ and $i, j, k, m \in \{1, \dots, d\}$ using the definition that $E_t = (Z_t, Z_{t+1})$,

$$\begin{aligned} &\max_{i,j,k,m \in \{1, \dots, d\}} \left| \frac{\mathbb{P}(E_{1+t} = (i, j) \mid E_1 = (k, m)) - \mathbb{P}(E_{1+t} = (i, j))}{\mathbb{P}(E_{1+t} = (i, j))} \right| \leq \frac{1}{4} \} \\ &= \max_{i,m \in \{1, \dots, d\}} \left| \frac{\mathbb{P}(Z_{1+t-1} = i \mid Z_1 = m) - \mathbb{P}(Z_{1+t-1} = i)}{\mathbb{P}(Z_{t+1} = i)} \right| \leq \frac{1}{4} \}. \end{aligned} \quad (\text{C.17})$$

The desired result now follows from Lemma A.1. \square

We bound the parameters (C.1)–(C.4). Recall that $\hat{\alpha}_{\min} = \min\{\#\mathcal{V}_k/d : k \in \{1, \dots, K\}\}$.

LEMMA C.4. *For any $i \in \{1, 2, 3\}$ it holds that $\mathbf{c}_i \leq \hat{\alpha}_{\min}^{-i}$ and $\mathfrak{d} \leq \frac{4}{3} \Psi(\mathbf{p}) \hat{\alpha}_{\min}^{-2}$.*

PROOF. The estimate $\mathbf{c}_i \leq \hat{\alpha}_{\min}^{-i}$ is immediate from the definitions (C.1)–(C.3) and the fact that $\pi_v \leq 1$ and $\mathbf{p}_{v,w} \leq 1$ for all $v, w \in \{1, \dots, K\}$. Further, for any $u, v, w \in \{1, \dots, K\}$ the triangle inequality together the estimates $\#\mathcal{V}_v/d \geq \hat{\alpha}_{\min}$ and $\mathbf{p}_{v,w} \leq 1$ yields that

$$\begin{aligned} \left| \frac{1}{n} \sum_{t=1}^{n-2} (n-2-t) \frac{(\mathbf{p}^t)_{u,v} - \pi_v}{\#\mathcal{V}_v} \middle| \frac{\mathbf{p}_{v,w}}{\#\mathcal{V}_w} \right| &\leq \sum_{t=1}^{n-2} \left| \frac{(\mathbf{p}^t)_{u,v} - \pi_v}{\#\mathcal{V}_v} \middle| \frac{\mathbf{p}_{v,w}}{\#\mathcal{V}_w} \right| \\ &\leq d^{-2} \hat{\alpha}_{\min}^{-2} \left((\Psi(\mathbf{p}) - 1) + \sum_{t=\Psi(\mathbf{p})}^{\infty} |(\mathbf{p}^t)_{u,v} - \pi_v| \right). \end{aligned} \quad (\text{C.18})$$

Since $\pi_v \leq 1$, one can relate the right-hand side of (C.18) to the ψ -dependence coefficient:

$$\max \left\{ |(\mathbf{p}^t)_{u,v} - \pi_v| : u, v \in \{1, \dots, K\} \right\} \leq \max \left\{ \frac{|(\mathbf{p}^t)_{u,v} - \pi_v|}{\pi_v} : u, v \in \{1, \dots, K\} \right\}.$$

Due to Proposition 4.4 it hence follows that

$$\mathfrak{d} \leq \hat{\alpha}_{\min}^{-2} \left((\Psi(\mathbf{p}) - 1) + \sum_{t=\Psi(\mathbf{p})}^{\infty} \left(\frac{1}{4} \right)^{\lfloor t/\Psi(\mathbf{p}) \rfloor} \right) = \hat{\alpha}_{\min}^{-2} \left((\Psi(\mathbf{p}) - 1) + \Psi(\mathbf{p}) \sum_{s=1}^{\infty} \left(\frac{1}{4} \right)^s \right). \quad (\text{C.19})$$

Use that $\sum_{t=1}^{\infty} (1/4)^t = 1/3$ to conclude the proof. \square

Recall that $\Psi(\mathbf{p})$ refers to the ψ -mixing time of the Markov chain on $\{1, \dots, K\}$ with transition matrix \mathbf{p} . By Lemma A.1, one can express $\Psi(\mathbf{p})$ more explicitly as

$$\Psi(\mathbf{p}) := \min \left\{ t \geq 1 : \max_{i,j \in \{1, \dots, K\}} \left| \frac{(\mathbf{p}^t)_{i,j} - \pi_j}{\pi_j} \right| \leq \frac{1}{4} \right\}. \quad (\text{C.20})$$

LEMMA C.5. *With \mathbf{S} , \mathbf{X} , and E as in (3.7)–(3.8) it holds that*

$$R(\mathbf{X}) \leq 2\sqrt{d/n}, \quad \Psi(E) \leq \Psi(\mathbf{p}) + 1, \quad \varsigma(\mathbf{X})^2 \leq \mathbf{c}_1, \quad \sigma(\mathbf{S})^2 \leq \mathfrak{g}, \quad v(\mathbf{S})^2 \leq d^{-1} \mathfrak{v},$$

where \mathfrak{g} and \mathfrak{v} are explicit and satisfy $\mathfrak{g} \leq \mathbf{c}_1 + Cd^{-1} \hat{\alpha}_{\min}^{-4} \Psi(\mathbf{p})$ and $\mathfrak{v} \leq C' \hat{\alpha}_{\min}^{-4} \Psi(\mathbf{p})$ for certain absolute constants $C, C' > 0$.

PROOF. The estimate $R(\mathbf{X}) \leq 2\sqrt{d/n}$ is immediate from the definition (3.8) of the matrices \mathbf{X}_t . Similarly, the estimate on $\Psi(E)$ is immediate from Lemmas C.2 and C.3.

We next consider $\varsigma(\mathbf{X})$. The \mathbf{X}_t are identically distributed since the block Markov chain is assumed to start in stationarity. Hence, using that for any self-adjoint random matrix \mathbf{A} one has $\mathbb{E}[(\mathbf{A} - \mathbb{E}[\mathbf{A}])^2] = \mathbb{E}[\mathbf{A}^2] - \mathbb{E}[\mathbf{A}]^2 \preceq \mathbb{E}[\mathbf{A}^2]$ with the positive semidefinite order,

$$\begin{aligned} \sum_{t=1}^{n-1} \mathbb{E}[\mathbf{X}_t^2] &= (n-1) \mathbb{E}[\mathbf{X}_1^2] \preceq d \sum_{i,j=1}^d \mathbb{P}(E_1 = (i,j)) \begin{pmatrix} 0 & e_i e_j^\top \\ e_j e_i^\top & 0 \end{pmatrix}^2 \\ &= d \sum_{i=1}^d \mathbb{P}(Z_1 = i) \begin{pmatrix} e_i e_i^\top & 0 \\ 0 & 0 \end{pmatrix} + d \sum_{j=1}^d \mathbb{P}(Z_2 = j) \begin{pmatrix} 0 & 0 \\ 0 & e_j e_j^\top \end{pmatrix}. \end{aligned}$$

Since the operator norm of a positive diagonal matrix is its maximal element, it follows that $\varsigma(\mathbf{X})^2 \leq \max\{d\pi_k/\#\mathcal{V}_k : k = 1, \dots, K\} = \mathbf{c}_1$, as desired.

We next consider $v(\mathbf{S})^2$. Using [9, Lemma 4.10] and that $\text{Cov}(\mathbf{M})$ being symmetric implies that its operator norm is bounded by the absolute row sums by [20, Corollary 2.3.2],

$$v(\mathbf{S})^2 \leq 2 \|\text{Cov}(\mathbf{M})\| \leq \max_{i,j=1, \dots, d} \sum_{k,m=1}^d |\text{Cov}(\mathbf{M})_{ij,km}|. \quad (\text{C.21})$$

For any fixed i, j there is precisely one term in (C.21) with $(k, m) = (i, j)$; at most $2d$ terms with $(k, m) \neq (i, j)$ but $i = m$ or $j = k$; and at most d^2 remaining terms. The combination of (C.21) and (C.5) hence yields that $v(\mathbf{S})^2 \leq d^{-1}\mathfrak{v}$ with

$$\mathfrak{v} := 2\left(\left(\mathfrak{c}_2 + \frac{2}{d^2}\mathfrak{c}_2^2 + \frac{2}{d}\mathfrak{c}_3 + \frac{2}{d^2}\mathfrak{c}_2\mathfrak{d}\right) + 2\left(\frac{3}{d}\mathfrak{c}_2^2 + 2\mathfrak{c}_3 + \frac{2}{d}\mathfrak{c}_2\mathfrak{d}\right) + \left(3\mathfrak{c}_2^2 + 2\mathfrak{c}_2\mathfrak{d}\right)\right). \quad (\text{C.22})$$

The claimed upper bound on \mathfrak{v} now follows from Lemma C.4.

We next consider the estimate on $\sigma(\mathbf{S})^2$. A direct computation shows that \mathbf{S}^2 is a block diagonal matrix with diagonal blocks $\mathbf{M}\mathbf{M}^\top$ and $\mathbf{M}^\top\mathbf{M}$. Consequently, taking expectations and the operator norm on both sides, $\sigma(\mathbf{S})^2 = \max\{\|\mathbf{A}\|, \|\mathbf{B}\|\}$ where \mathbf{A} and \mathbf{B} are the self-adjoint $d \times d$ matrices whose entries are given by

$$\mathbf{A}_{i,j} = \sum_{\ell=1}^d \text{Cov}(\mathbf{M})_{i\ell,j\ell}, \quad \mathbf{B}_{i,j} = \sum_{\ell=1}^d \text{Cov}(\mathbf{M})_{\ell i,\ell j}. \quad (\text{C.23})$$

We consider the diagonal and the off-diagonal terms of these matrices separately. First, consider the case where $i \neq j$. Then, $(i, \ell) \neq (j, \ell)$ and $(\ell, i) \neq (\ell, j)$ for all $\ell \in \{1, \dots, d\}$. Hence, using (C.5) with separate consideration of the case where $i = \ell$ or $j = \ell$

$$|\mathbf{A}_{i,j}| \leq \frac{2}{d^2}\left(\frac{3}{d}\mathfrak{c}_2^2 + 2\mathfrak{c}_3 + \frac{2}{d}\mathfrak{c}_2\mathfrak{d}\right) + \frac{d}{d^3}(3\mathfrak{c}_2^2 + 2\mathfrak{c}_2\mathfrak{d}), \quad (\text{C.24})$$

$$|\mathbf{B}_{i,j}| \leq \frac{2}{d^2}\left(\frac{3}{d}\mathfrak{c}_2^2 + 2\mathfrak{c}_3 + \frac{2}{d}\mathfrak{c}_2\mathfrak{d}\right) + \frac{d}{d^3}(3\mathfrak{c}_2^2 + 2\mathfrak{c}_2\mathfrak{d}). \quad (\text{C.25})$$

For the case with $i = j$ we get better leading-order term if we replace (C.9) by the following:

$$\begin{aligned} \left|\frac{d}{n} \sum_{\ell=1}^d \sum_{t_1=1}^{n-1} \mathcal{E}_{t_1,t_1}(i, \ell, i, \ell)\right| &= \frac{d(n-1)}{n} \sum_{\ell=1}^d (\mathbb{P}(E_1 = (i, \ell)) - \mathbb{P}(E_1 = (i, \ell))^2) \\ &\leq d\mathbb{P}(Z_1 = i) \leq \mathfrak{c}_1. \end{aligned} \quad (\text{C.26})$$

It was here used that $\mathbb{P}(E_1 = (i, \ell)) \geq \mathbb{P}(E_1 = (i, \ell))^2$. Now observe that by (C.7),

$$\begin{aligned} \left|\sum_{\ell=1}^d \text{Cov}(\mathbf{M})_{i\ell,i\ell}\right| &\leq \left|\frac{d}{n} \sum_{\ell=1}^d \sum_{t_1=1}^{n-1} \mathcal{E}_{t_1,t_1}(i, \ell, i, \ell)\right| + \left|\frac{d}{n} \sum_{\ell=1}^d \sum_{|t_1-t_2|=1} \mathcal{E}_{t_1,t_2}(i, \ell, i, \ell)\right| \\ &\quad + \left|\frac{d}{n} \sum_{\ell=1}^d \sum_{|t_1-t_2|>1} \mathcal{E}_{t_1,t_1+1}(i, \ell, i, \ell)\right|. \end{aligned} \quad (\text{C.27})$$

Hence, by using (C.12), (C.14), and (C.26) in (C.27), $|\mathbf{A}_{i,i}| \leq \mathfrak{c}_1 + d^{-1}(2d^{-1}\mathfrak{c}_2^2 + 2\mathfrak{c}_3 + 2d^{-1}\mathfrak{c}_2\mathfrak{d})$. Exactly the same estimate applies to $|\mathbf{B}_{i,i}|$. Now using [20, Corollary 2.3.2] as in (C.21), we conclude that $\sigma(\mathbf{S})^2 \leq \mathfrak{g}$ with

$$\mathfrak{g} := \mathfrak{c}_1 + \frac{1}{d}\left(\frac{2}{d}\mathfrak{c}_2^2 + 2\mathfrak{c}_3 + \frac{2}{d}\mathfrak{c}_2\mathfrak{d}\right) + \frac{1}{d}\left(2\left(\frac{3}{d}\mathfrak{c}_2^2 + 2\mathfrak{c}_3 + \frac{2}{d}\mathfrak{c}_2\mathfrak{d}\right) + (3\mathfrak{c}_2^2 + 2\mathfrak{c}_2\mathfrak{d})\right). \quad (\text{C.28})$$

The claimed upper bound on \mathfrak{g} now follows from Lemma C.4. \square

C.2. Convergence of singular value distributions. Recall the definition of the empirical singular value (3.11) and that \mathbf{S} was defined by a self-adjoint dilation in (3.7). It follows that $\text{sym}(\nu_{\mathbf{M}}) = \mu_{\mathbf{S}}$ where $\mu_{\mathbf{S}}$ is the *empirical eigenvalue distribution* of \mathbf{S} , defined by

$$\mu_{\mathbf{S}}([a, b]) := \frac{1}{2d} \#\{i \in \{1, \dots, 2d\} : a \leq \lambda_i(\mathbf{S}) \leq b\} \quad (\text{C.29})$$

where the $\lambda_i(\mathbf{S})$ are the eigenvalues of \mathbf{S} . We rely on the well known moment method to establish the limiting law of $\mu_{\mathbf{S}}$ and state it as a lemma for the sake of clarity.

LEMMA C.6. Consider a sequence of self-adjoint random matrices $(\mathbf{S}_i)_{i=1}^\infty$ and let μ be a deterministic compactly supported probability measure on \mathbb{R} . If for every integer $p \geq 1$,

$$\lim_{i \rightarrow \infty} \mathbb{E}[\operatorname{tr} \mathbf{S}_i^p] = \int x^p d\mu(x) \quad \text{and} \quad \lim_{i \rightarrow \infty} \operatorname{Var}[\operatorname{tr} \mathbf{S}_i^p] = 0,$$

then $\mu_{\mathbf{S}_i}$ converges weakly in probability to μ as i tends to infinity.

PROOF. This is standard, for instance implicit in the proof of the Wigner semicircular law in [3, Section 2.1.2]. \square

LEMMA C.7. If $\lim_{d \rightarrow \infty} d/n = 0$, then for every positive integer $p \geq 1$

$$\lim_{d \rightarrow \infty} |\mathbb{E}[\operatorname{tr} \mathbf{S}^p] - \mathbb{E}[\operatorname{tr} \mathbf{G}^p]| = 0 \quad \text{and} \quad \lim_{d \rightarrow \infty} |\operatorname{Var}[\operatorname{tr} \mathbf{S}^p] - \operatorname{Var}[\operatorname{tr} \mathbf{G}^p]| = 0.$$

PROOF. Using that $\liminf_{d \rightarrow \infty} \min\{\#\mathcal{V}_k/d : k = 1, \dots, K\} > 0$ in the considered limiting regime and that $\lim_{d \rightarrow \infty} d/n = 0$ it follows from Lemmas C.4 and C.5 that

$$\lim_{d \rightarrow \infty} R(\mathbf{X}) = 0, \quad \limsup_{d \rightarrow \infty} \Psi(E) < \infty, \quad \limsup_{d \rightarrow \infty} \varsigma(\mathbf{X})^2 < \infty. \quad (\text{C.30})$$

Hence, it follows from Theorem 2.4 that $\lim_{d \rightarrow \infty} |\mathbb{E}[\operatorname{tr} \mathbf{S}^{2p}] - \mathbb{E}[\operatorname{tr} \mathbf{G}^{2p}]| = 0$. Further, it holds for every odd p that $\operatorname{tr} \mathbf{S}^p = 0 = \operatorname{tr} \mathbf{G}^p$ by definition of a self-adjoint dilation (3.7).

It now remains to show that the variance of even tracial moments is universal, for which purpose it suffices to show that $\lim_{d \rightarrow \infty} |\mathbb{E}[(\operatorname{tr} \mathbf{S}^{2p})^2] - \mathbb{E}[(\operatorname{tr} \mathbf{G}^{2p})^2]| = 0$. For this purpose, we use the trick outlined in Remark 3.8: note that for every fixed $t \in \mathbb{R}$,

$$\mathbb{E}[\operatorname{tr}(\mathbf{S} \otimes \mathbf{1} + t\mathbf{1} \otimes \mathbf{S})^{2p}] = \sum_{j=0}^{2p} \binom{2p}{j} t^j \mathbb{E}[\operatorname{tr}[\mathbf{S}^{2p-j}] \operatorname{tr}[\mathbf{S}^j]]. \quad (\text{C.31})$$

Hence, since pointwise convergence of polynomials implies convergence of coefficients, it suffices to prove universality for $\mathbb{E}[\operatorname{tr}(\mathbf{S} \otimes \mathbf{1} + t\mathbf{1} \otimes \mathbf{S})^{2p}]$. (Consider $j = p$.)

To this end, note that the matrix $\mathbf{S} \otimes \mathbf{1} + t\mathbf{1} \otimes \mathbf{S}$ can be represented as a Markovian model. Moreover, direct computation shows that the parameters are again of the same asymptotic order. Indeed, $\|\mathbf{X}_i \otimes \mathbf{1} + t\mathbf{1} \otimes \mathbf{X}_i\| \leq (1+t)R(\mathbf{X})$ and $\mathbb{E}[(\mathbf{X}_i \otimes \mathbf{1} + t\mathbf{1} \otimes \mathbf{X}_i)^2] = \mathbb{E}[\mathbf{X}_i^2] \otimes \mathbf{1} + t\mathbf{1} \otimes \mathbb{E}[\mathbf{X}_i^2]$ since $\mathbb{E}[\mathbf{X}_i] = 0$ so that the variance quantity is bounded by $(1+t)\varsigma(\mathbf{X})$. Hence, using the parameter estimates from (C.30) together with Theorem 2.4 gives pointwise convergence for the polynomial (C.31), concluding the proof. \square

In order to prove Proposition 3.6 it is now sufficient to consider the empirical eigenvalue distribution of the Gaussian model. We will do this by using [44, Theorem 4.2].

The proofs in [44] are not directly applicable to \mathbf{S} in the sparse regime $n \ll d^2$. And outside the sparse regime, [44] relies on quite some computation to prove that [44, Theorem 4.2] is applicable to \mathbf{S} , by directly determining the joint moments of the entries of the matrix. Gaussian universality allows one to bypass this, since the higher joint moments of a Gaussian vector are determined by its covariance structure.

REMARK C.8. Given Gaussian universality, there are also other arguments that can recover the following results. For instance, the Gaussian comparison results in [14, Section 8.1] can show that the spectral distribution and norm of \mathbf{G} are well-approximated by those of a self-adjoint Gaussian matrix $\tilde{\mathbf{G}}$ with *independent* entries. Approximating $\tilde{\mathbf{G}}$ next by a free probabilistic object using [9], and then using the matrix Dyson equation [21, Equation (1.5)] for the spectral law of the latter object, would do the trick.

For any positive semidefinite $D \times D$ matrix $\Sigma \succcurlyeq 0$ we define the following parameter measuring the size of the off-diagonal terms:

$$\epsilon(\Sigma) := \max\{|\Sigma_{i,j}| : i \neq j, i, j \in \{1, \dots, D\}\}. \quad (\text{C.32})$$

The following lemma is used to verify a condition in [44].

LEMMA C.9. *Fix a positive integer $D \geq 1$ and positive scalars $\ell, u > 0$ and define a set of positive semidefinite matrices by*

$$\mathfrak{S}_D(\ell, u) := \left\{ \Sigma \in \mathbb{R}^{D \times D} : \Sigma \succcurlyeq 0 \text{ and } \ell \leq \Sigma_{i,i} \leq u \text{ for all } i \in \{1, \dots, D\} \right\}.$$

Then, for any nonnegative integers $0 \leq r \leq D$ and $p_1, \dots, p_D \geq 0$ with $p_i = 1$ for $i \in \{1, \dots, r\}$ there exists $C > 0$ such that for any centered Gaussian vector g with covariance matrix $\Sigma \in \mathfrak{S}_D(\ell, u)$,

$$\left| \mathbb{E}[g_1^{p_1} g_2^{p_2} \cdots g_D^{p_D}] \right| \leq C \epsilon(\Sigma)^{\frac{r}{2}} \quad \text{and} \quad \left| \mathbb{E}[g_1^2 \cdots g_D^2] - \mathbb{E}[g_1^2] \cdots \mathbb{E}[g_D^2] \right| \leq C \epsilon(\Sigma)^2. \quad (\text{C.33})$$

PROOF. This follows readily from direct calculations with the properties of the Gaussian distribution. For instance, one can proceed by induction on D . The base case $D = 1$ is immediate from the assumption that $\mathbb{E}[g_1] = 0$, and in the inductive step one can exploit that the conditional distribution of g_1 given g_2, \dots, g_D is explicit [15, Proposition 3.13]. This computation is straightforward but tedious, so we omit the details.⁸ \square

PROOF OF PROPOSITION 3.6. For any $d \geq 1$ let $\mathbf{A}_d := \sqrt{d}\mathbf{G}$ denote a rescaled version of the corresponding $2d \times 2d$ Gaussian model \mathbf{G} for \mathbf{S} . We claim that the sequence of random matrices $(\mathbf{A}_d)_{d=K}^\infty$ is *approximately uncorrelated with variance profile* as defined in [44, Definition 4.1]. This means that we have to show that for any fixed non-negative integers $0 \leq r \leq D$ and $p_1, \dots, p_D \geq 0$ with $p_i = 1$ for $i = 1, \dots, r$,

$$\limsup_{d \rightarrow \infty} \max_{\forall k \neq l: \{i_k, j_k\} \neq \{i_l, j_l\}} d^{\frac{r}{2}} \left| \mathbb{E}[\mathbf{A}_{d, i_1 j_1}^{p_1} \mathbf{A}_{d, i_2 j_2}^{p_2} \cdots \mathbf{A}_{d, i_D j_D}^{p_D}] \right| < \infty \quad (\text{C.34})$$

and

$$\limsup_{d \rightarrow \infty} \max_{\forall k \neq l: \{i_k, j_k\} \neq \{i_l, j_l\}} \left| \mathbb{E}[\mathbf{A}_{d, i_1 j_1}^2 \cdots \mathbf{A}_{d, i_D j_D}^2] - \mathbb{E}[\mathbf{A}_{d, i_1 j_1}^2] \cdots \mathbb{E}[\mathbf{A}_{d, i_D j_D}^2] \right| = 0 \quad (\text{C.35})$$

with the maxima running over all values of $(i_1, j_1), \dots, (i_D, j_D) \in \{1, \dots, 2d\}^2$ with $\{i_k, j_k\} \neq \{i_l, j_l\}$ for all $k \neq l$.

If $\mathbf{A}_{d, i_k, j_k} = 0$ almost surely, then there is nothing to prove so assume that $\text{Var}[\mathbf{A}_{i_k, j_k}] \neq 0$. Recall that \mathbf{G} is a Gaussian model of \mathbf{S} which is a self-adjoint dilation of \mathbf{M} . The covariance of the entries of \mathbf{A} is hence a function of the covariance of the entries of \mathbf{M} . Hence, applying Lemma C.9 to the vector $g := (\mathbf{A}_{d, i_1 j_1}, \dots, \mathbf{A}_{d, i_D j_D})$ yields (C.34) and (C.35) if there exist constants $\ell, u, c > 0$ such that $\ell \leq d \text{Cov}(\mathbf{M})_{ij, ij} \leq u$ and $|d \text{Cov}(\mathbf{M})_{ij, km}| \leq c/d$ for all $(i, j) \neq (k, m)$. Indeed, $\ell \leq d \text{Cov}(\mathbf{M})_{ij, ij} \leq u$ shows that the assumption of Lemma C.9 is satisfied while $|d \text{Cov}(\mathbf{M})_{ij, km}| \leq c/d$ yields that $\epsilon(\text{Cov}(g)) \leq c/d$ with ϵ as in (C.32) which ensures that Lemma C.9 provides a sufficiently strong conclusion.

The required upper bounds on $|d \text{Cov}(\mathbf{M})_{ij, km}|$ and $d \text{Cov}(\mathbf{M})_{ij, ij}$ can be found in (C.5). For the lower bound on $\text{Cov}(\mathbf{M})_{ij, ij}$, note that (C.6) implies that for all $a, b \in \{1, \dots, K\}$,

$$\lim_{d \rightarrow \infty} \max_{i \in \mathcal{V}_a, j \in \mathcal{V}_b} \left| d \text{Cov}(\mathbf{M})_{ij, ij} - \frac{\pi_a \mathbf{P}_{a,b}}{\alpha_a \alpha_b} \right| = 0. \quad (\text{C.36})$$

⁸If required, these details can be found in the first arXiv version of this paper; see arXiv:2307.11632v1.

In particular, since $\frac{\pi_a \mathbf{P}_{a,b}}{\alpha_a \alpha_b} \neq 0$ due to the preliminary reduction to the case where $\text{Var}[A_{d,i_k,j_k}] \neq 0$, there exists a constant $\ell > 0$ such that $\ell \leq d \text{Cov}(\mathbf{M})_{ij,ij}$ for all d .

We may note that (C.36) is exactly the same limiting variance profile as occurs in the dense regime with $n = d^2$ in [44, Equation (23)]. Consequently, since the universal limiting eigenvalue distribution in [44, Theorem 4.2] only depends on the limiting variance profile, the empirical eigenvalue distribution of $\mathbf{G} = \mathbf{A}_d/\sqrt{d}$ has the same universal limit as is predicted by [44, Theorem 1.1]. More precisely, it follows from [44, Lemma 6.5 and Lemma 6.6] that

$$\lim_{d \rightarrow \infty} \mathbb{E}[\text{tr} \mathbf{G}^p] = \int x^p d\text{sym}(\nu_\infty)(x) \quad \text{and} \quad \lim_{i \rightarrow \infty} \text{Var}[\text{tr} \mathbf{G}^p] = 0 \quad (\text{C.37})$$

with ν_∞ as in Proposition 3.6. The result now follows from Lemmas C.6 and C.7, and the fact that $\mu_{\mathbf{S}} = \text{sym}(\nu_{\mathbf{M}})$. \square

PROOF OF COROLLARY 3.7. Note that $\text{rank} \mathbb{E}[\hat{\mathbf{N}}] \leq K$. The desired result hence follows from Proposition 3.6 since low-rank perturbations do not affect limiting singular value distributions. More precisely, one can combine [7, Theorem A.43] with a self-adjoint dilation. \square

PROPOSITION C.10 (Abundance of singular values near \mathfrak{m}). *Adopt the assumptions of Proposition 3.6. Then, for any $\varepsilon > 0$ there exists some constant $c > 0$ such that*

$$\lim_{d \rightarrow \infty} \mathbb{P}(\#\{i \in \{1, \dots, d\} : s_i(\mathbf{M}) \in (\mathfrak{m} - \varepsilon, \mathfrak{m} + \varepsilon)\} < cd) = 0. \quad (\text{C.38})$$

In particular, since $\|\mathbf{M}\|$ is the greatest singular value, $\lim_{d \rightarrow \infty} \mathbb{P}(\|\mathbf{M}\| < \mathfrak{m} - \varepsilon) = 0$.

PROOF. We can interpret the measure $\text{sym}(\nu_\infty)$ in Proposition 3.6 as the spectral law of a $2K \times 2K$ free-probabilistic object. Specifically, if \mathbf{H} is the symmetric $2K \times 2K$ Gaussian matrix with independent entries satisfying that for every $i, j \leq K$,

$$\text{Var}[\mathbf{H}_{i,j}] = 0 \quad \text{and} \quad \text{Var}[\mathbf{H}_{i,j+K}] = \alpha_i^{-1} \pi_i \mathbf{p}_{i,j} \quad (\text{C.39})$$

then, it is classical (see e.g., [31, Chapter 9] or [21, Equation (1.5)]) that the system of equations in Proposition 3.6 corresponds to the Stieltjes transform of the spectral law of the free-probabilistic object \mathbf{H}_{free} defined in [9, Section 2.1].

To be more specific, it is part of the definition of the latter object that it is an element of a free probability space of the form $\mathcal{A}^{2K \times 2K}$ where \mathcal{A} is a C^* algebra equipped with a faithful state $\tau : \mathcal{A} \rightarrow \mathbb{C}$, and the statement that $\text{sym}(\nu_\infty)$ corresponds to the spectral law means that $(\text{tr} \otimes \tau)(f(\mathbf{H}_{\text{free}})) = \int f(x) d\text{sym}(\nu_\infty)$ for every continuous function $f : \mathbb{R} \rightarrow \mathbb{R}$ where $f(\mathbf{H}_{\text{free}})$ is defined by functional calculus. Thus, it follows from [34, (3.26)] that $\|\mathbf{H}_{\text{free}}\|$ is the upper edge of the support of this measure $\text{sym}(\nu_\infty)$. In particular, the measure assigns nonzero mass to any open interval containing $\|\mathbf{H}_{\text{free}}\|$. Now, (C.38) follows from Proposition 3.6 since $\|\mathbf{H}_{\text{free}}\| = \mathfrak{m}$ with \mathfrak{m} as in Theorem 3.5 by [9, Lemma 3.2]. \square

C.3. Sharp upper bounds on $\|\mathbf{M}\|$. The main technical result in this section concerns a nonasymptotic upper bound on $\|\mathbf{S}_{\text{free}}\|$; see Lemma C.12. We further prove Theorem 3.5, and establish a nonasymptotic concentration inequality in Proposition C.13.

Recall that $\hat{\alpha}_i := \#\mathcal{V}_i/d$ and define a scalar $\hat{\mathfrak{m}}$ analogously to Theorem 3.5 by

$$\hat{\mathfrak{m}} := \inf_{x \in \mathbb{R}_{>0}^{2K}} \max_{i=1, \dots, 2K} \left\{ \frac{1}{x_i} + \sum_{j=1}^{2K} \hat{c}_{i,j} x_j \right\} \quad (\text{C.40})$$

where the infimum runs over all vectors x with strictly positive coordinates and the coefficients $(\hat{c}_{i,j})_{i,j=1}^{2K}$ are defined by

$$\hat{c}_{i,j} = \begin{cases} 0 & \text{if } i \leq K \text{ and } j \leq K, \\ \hat{\alpha}_i^{-1} \pi_i \mathbf{p}_{i,j-K} & \text{if } i \leq K \text{ and } j > K, \\ 0 & \text{if } i > K \text{ and } j > K, \\ \hat{\alpha}_{i-K}^{-1} \pi_j \mathbf{p}_{j,i-K} & \text{if } i > K \text{ and } j \leq K. \end{cases} \quad (\text{C.41})$$

Let us further introduce the following parameter:

$$\mathbf{u} := \max_{i \in \{1, \dots, 2K\}} \min \left\{ \frac{2\sqrt{c_1}}{\hat{c}_{i,j}} : j \in \{1, \dots, 2K\} \right\}. \quad (\text{C.42})$$

LEMMA C.11. *There exists a vector $x^* \in \mathbb{R}_{>0}^{2K}$ which realizes the infimum in (C.40) and this vector satisfies that $\frac{1}{2c_1} \leq x_i^* \leq \mathbf{u}$ for every $i \in \{1, \dots, 2K\}$.*

PROOF. As in the proof of Proposition C.10, it follows from [9, Lemma 3.2] that $\hat{\mathbf{m}} = \|\hat{\mathbf{H}}_{\text{free}}\|$ where $\hat{\mathbf{H}}$ is the symmetric $2K \times 2K$ Gaussian matrix with $\text{Var}[\hat{\mathbf{H}}_{i,j}] = 0$ and $\text{Var}[\hat{\mathbf{H}}_{i,j+K}] = \alpha_i^{-1} \pi_i \mathbf{p}_{i,j}$ for every $i, j \leq K$. Due to (2.10) it consequently holds that

$$\hat{\mathbf{m}} \leq 2\sigma(\hat{\mathbf{H}}) = 2\|\mathbb{E}[\hat{\mathbf{H}}^2]\|^{1/2} = 2\sqrt{\max\left\{\frac{\pi_a}{\hat{\alpha}_a} : a \in \{1, \dots, K\}\right\}} = 2\sqrt{c_1} \quad (\text{C.43})$$

where we used that the operator norm of the diagonal matrix $\mathbb{E}[\hat{\mathbf{H}}^2]$ is its maximal element.

Let us now define a subset of $\mathbb{R}_{>0}^{2K}$ by

$$\mathfrak{X} := \left\{ x \in \mathbb{R}_{>0}^{2K} : \frac{1}{2\sqrt{c_1}} \leq x_i \leq \mathbf{u} \text{ for all } i \in \{1, \dots, 2K\} \right\}. \quad (\text{C.44})$$

Then, due to (C.43), the infimum in (C.40) may be restricted to \mathfrak{X} . Further, since \mathbf{p} is assumed to define an ergodic Markov chain, it holds for every j that there is some i with $\hat{c}_{i,j} \neq 0$. Thus, \mathbf{u} is finite and the set \mathfrak{X} is compact. Consequently, there exists $x^* \in \mathfrak{X}$ with

$$\hat{\mathbf{m}} = \inf_{x \in \mathfrak{X}} \max_{i=1, \dots, 2K} \left\{ \frac{1}{x_i} + \sum_{j=1}^{2K} \hat{c}_{i,j} x_j \right\} = \max_{i=1, \dots, 2K} \left\{ \frac{1}{x_i^*} + \sum_{j=1}^{2K} \hat{c}_{i,j} x_j^* \right\}. \quad (\text{C.45})$$

This concludes the proof. \square

LEMMA C.12. *With \mathbf{S} as in (3.7), there exists an explicit $\mathfrak{C} > 0$ with*

$$\|\mathbf{S}_{\text{free}}\| \leq \hat{\mathbf{m}} + \frac{1}{d} \mathfrak{C}.$$

Moreover, it holds that $\mathfrak{C} \leq n^{-1} \text{duc}_2 + C \mathbf{u} \Psi(\mathbf{p}) \hat{\alpha}_{\min}^{-4}$ for some absolute constant $C > 0$.

PROOF. Let $x^* \in \mathbb{R}_{>0}^{2K}$ be the vector provided by Lemma C.11 and let \mathbf{W} be the $2d \times 2d$ diagonal matrix with $\mathbf{W}_{i,i} = x_a^*$ and $\mathbf{W}_{i+d,i+d} = x_{a+K}^*$ for any $i \in \mathcal{V}_a$ and $a \in \{1, \dots, K\}$. Then, due to (2.9) and the fact that eigenvalues are dominated by the operator norm

$$\|\mathbf{S}_{\text{free}}\| \leq \lambda_{\max}(\mathbf{W}^{-1} + \mathbb{E}[\mathbf{S}\mathbf{W}\mathbf{S}]) \leq \|\mathbf{W}^{-1} + \mathbb{E}[\mathbf{S}\mathbf{W}\mathbf{S}]\|. \quad (\text{C.46})$$

For any $i, j \in \{1, \dots, 2d\}$ the (i, j) th entry of $\mathbb{E}[\mathbf{SWS}]$ is given by

$$\begin{aligned} \mathbb{E}[\mathbf{SWS}]_{i,j} &= \sum_{k=1}^{2d} \mathbf{W}_k \mathbb{E}[\mathbf{S}_{i,k} \mathbf{S}_{k,j}] \\ &= \begin{cases} 0 & \text{if } (i \leq d \text{ and } j > d) \text{ or } (i > d \text{ and } j \leq d), \\ \sum_{k=1}^d \mathbf{W}_{k+d, k+d} \text{Cov}(\mathbf{M})_{ik, jk} & \text{if } i \leq d \text{ and } j \leq d, \\ \sum_{k=1}^d \mathbf{W}_{k, k} \text{Cov}(\mathbf{M})_{k(i-d), k(j-d)} & \text{if } i > d \text{ and } j > d. \end{cases} \end{aligned} \quad (\text{C.47})$$

We next study the diagonal and off-diagonal entries separately, starting with the diagonal.

Recall the estimate (C.6) on $\text{Cov}(\mathbf{M})_{ik, ik}$ and note that, since $\hat{\alpha}_a = \#\mathcal{V}_a/d$,

$$\left| \text{Cov}(\mathbf{M})_{ik, ik} - \frac{1}{d} \frac{\pi_a \mathbf{P}_{a,b}}{\hat{\alpha}_a \hat{\alpha}_b} \right| \leq \frac{1}{d^2} \left(\frac{d}{n} \mathbf{c}_2 + \frac{3}{d} \mathbf{c}_2^2 + 2\mathbf{c}_3 + \frac{2}{d} \mathbf{c}_2 \mathfrak{d} \right) \quad (\text{C.48})$$

for every $i \in \mathcal{V}_a$ and $k \in \mathcal{V}_b$. Recall from (C.41) and that $\hat{c}_{a,b} = 0$ if $\max\{a, b\} \leq K$ or if $\min\{a, b\} > K$. Further, note that $\max_{i=1, \dots, d} \mathbf{W}_{i,i} = \max_{b=1, \dots, 2K} x_b^* \leq \mathbf{u}$ by Lemma C.11. Hence, grouping terms along the clusters in (C.47) and substituting (C.48) yields that

$$\left| \mathbb{E}[\mathbf{SWS}]_{i,i} - \sum_{b=1}^{2K} \hat{c}_{a,b} x_b^* \right| \leq \frac{\mathbf{u}}{d} \left(\frac{d}{n} \mathbf{c}_2 + \frac{3}{d} \mathbf{c}_2^2 + 2\mathbf{c}_3 + \frac{2}{d} \mathbf{c}_2 \mathfrak{d} \right) \quad (\text{C.49})$$

for every i, a with $i \in \mathcal{V}_a$ and $a \leq K$ or $i \in \mathcal{V}_{a-K}$ and $a > K$.

Now suppose that $i \neq j$. Then, using the estimate (C.5) in (C.47) with the fact that there are precisely 2 values of k with $k \in \{i, j\}$ and $d-2 \leq d$ remaining values,

$$\begin{aligned} |\mathbb{E}[\mathbf{SWS}]_{i,j}| &\leq \frac{2\mathbf{u}}{d^2} \left(\frac{3}{d} \mathbf{c}_2^2 + 2\mathbf{c}_3 + \frac{2}{d} \mathbf{c}_2 \mathfrak{d} \right) + \frac{\mathbf{u}}{d^2} (3\mathbf{c}_2^2 + 2\mathbf{c}_2 \mathfrak{d}) \\ &= \frac{\mathbf{u}}{d^2} \left((3\mathbf{c}_2 + 4\mathbf{c}_3 + 2\mathbf{c}_2^2 \mathfrak{d}) + \frac{1}{d} (6\mathbf{c}_2^2 + 4\mathbf{c}_2 \mathfrak{d}) \right). \end{aligned} \quad (\text{C.50})$$

Let us now split $\mathbf{W}^{-1} + \mathbb{E}[\mathbf{SWS}] = \mathbf{D} + \mathbf{R}$ where \mathbf{D} is the matrix containing all diagonal entries and \mathbf{R} is the matrix containing all off-diagonal entries. Then, since the operator norm of a diagonal matrix is the greatest absolute value of its elements it follows from (C.49) that

$$\|\mathbf{D}\| \leq \hat{\mathbf{m}} + \frac{\mathbf{u}}{d} \left(\frac{d}{n} \mathbf{c}_2 + \frac{3}{d} \mathbf{c}_2^2 + 2\mathbf{c}_3 + \frac{2}{d} \mathbf{c}_2 \mathfrak{d} \right). \quad (\text{C.51})$$

Further, using that the operator norm of a block diagonal matrix is the maximum of the operator norms of the blocks as well as the fact that the operator norm is dominated by the Frobenius norm, it follows from (C.50) that

$$\|\mathbf{R}\| = \left\| \begin{pmatrix} \mathbf{R}_1 & 0 \\ 0 & \mathbf{R}_2 \end{pmatrix} \right\| \leq \frac{\mathbf{u}}{d} (3\mathbf{c}_2^2 + 4\mathbf{c}_3 + 2\mathbf{c}_2 \mathfrak{d}) + \frac{\mathbf{u}}{d^2} (6\mathbf{c}_2^2 + 4\mathbf{c}_2 \mathfrak{d}). \quad (\text{C.52})$$

Hence, since $\|\mathbf{W}^{-1} + \mathbb{E}[\mathbf{SWS}]\| \leq \|\mathbf{D}\| + \|\mathbf{R}\|$, the combination of (C.46), (C.51), and (C.52) yields the desired result with

$$\mathfrak{E} := \mathbf{u} \left(\frac{d}{n} \mathbf{c}_2 + 3\mathbf{c}_2^2 + 6\mathbf{c}_3 + 2\mathbf{c}_2 \mathfrak{d} \right) + \frac{\mathbf{u}}{d} (9\mathbf{c}_2^2 + 6\mathbf{c}_2 \mathfrak{d}). \quad (\text{C.53})$$

The claimed upper bound on \mathfrak{E} further follows from Lemma C.4. \square

The combination of Lemmas C.5 and C.12 provides close-to-optimal estimates on the parameters of \mathbf{S} . This yields the following concentration inequality:

PROPOSITION C.13. *There exists an absolute constant $c > 0$ such that, for every $0 < \delta \leq 1$ and $x > 0$, the matrix $\mathbf{M} = \sqrt{d/n}(\hat{\mathbf{N}} - \mathbb{E}[\hat{\mathbf{N}}])$ satisfies*

$$\mathbb{P}(\|\mathbf{M}\| \geq (1 + \delta)(\hat{m} + d^{-1}\mathfrak{E}) + c\mathcal{E}(x)) \leq (d + 1)(1 + \delta)^{-x}.$$

where $\mathcal{E}(x) := (dn^{-1}\Psi(E)^4\mathfrak{c}_1^2)^{1/6}x^{2/3} + d^{1/2}n^{-1/2}\Psi(E)x + (d^{-1}\mathfrak{v}\mathfrak{g})^{1/4}(x^{1/2} + \ln(d + 1)^{3/4})$ with \hat{m} , \mathfrak{E} , \mathfrak{g} , \mathfrak{v} , and \mathfrak{c}_1 as in (C.40), (C.53), (C.28), (C.22), and (C.1), respectively.

PROOF. Apply Proposition 5.2 with Lemmas C.5 and C.12 and use that $\|\mathbf{M}\| = \|\mathbf{S}\|$. \square

LEMMA C.14. *Adopt the notation and assumptions of Theorem 3.5. Then, for any $\varepsilon > 0$*

$$\lim_{d \rightarrow \infty} \mathbb{P}(\|\mathbf{M}\| > m + \varepsilon) = 0. \quad (\text{C.54})$$

PROOF. A comparison of (3.10) and (C.40) shows that $\lim_{d \rightarrow \infty} \hat{m} = m$; recall that we assume that $\lim_{d \rightarrow \infty} \#\mathcal{V}_a/d = \alpha_a$ and that \mathfrak{p} is kept fixed. Further, the parameter \mathfrak{E} defined in (C.53) remains bounded as d tends to infinity, so $\mathfrak{E}/d \rightarrow 0$. Now, using Proposition C.13 with δ fixed at a sufficiently small value and subsequently taking x a large multiple of $\ln(d)$, a direct calculation using the assumption $\lim_{d \rightarrow \infty} d \ln(d)^4/n = 0$ yields the result. \square

PROOF OF THEOREM 3.5. Combine Proposition C.10 and Lemma C.14. \square