Universality-based concentration for sums of dependent random matrices

Jaron Sanders and Alexander Van Werde

Department of Mathematics and Computer Science, Eindhoven University of Technology

TU/e

Matrix concentration inequalities

The following topics have a common feature: they all benefit from matrix concentration inequalities!

- Community detection
- Time series analysis
- Quantum information theory
- Randomized linear algebra
- Spectral graph theory
- Compressed sensing

An example of such a matrix concentration inequality is the matrix Bernstein inequality [Oli09, Tro12]. Let $\mathbf{X}_1, \dots, \mathbf{X}_n$ be independent self-adjoint random matrices satisfying $\mathbb{E}[\mathbf{X}_i] = 0$ and $\|\mathbf{X}_i\| \le R$. Then, $\mathbf{S} := \sum_i \mathbf{X}_i$ satisfies

$$\mathbb{E}[\|\mathbf{S}\|] \le c\sigma\sqrt{\ln(2d)} + R\ln(2d)$$

where $\sigma^2 := \|\mathbb{E}[\mathbf{S}^2]\|$.

This result is extremely flexible and underlies many of the applications above. It is however not always a sharp bound. In particular, the dimensional dependence by $\sqrt{\ln(2d)}$ on the first term is often found to be suboptimal. For instance, for a Wigner matrix one has $\mathbb{E}[\|\mathbf{S}\|] \approx 2\sigma$. In such a case, matrix Bernstein does not provide the correct asymptotic order. A further limitation is the independence assumption which can be restrictive in applications such as the analysis of time series.

Question 1. When can the dimensional dependence by $\sqrt{\ln(2d)}$ be removed?

Question 2. What if the summands $\mathbf{X}_i$ are dependent?

We were inspired to work on this problem by [BBvH21] which made fundamental progress on Question 1 in the case where $\mathbf{S}$ is Gaussian and by [BvH22] which concerns the case with independent non-Gaussian summands.

Markovian model

Let $Z_1, \dots, Z_n$ be a Markovian sequence of random variables. Suppose that the random matrices are instantaneous functions of this Markov chain:

$$\mathbf{X}_i = f_i(\mathbf{Z}_i).$$

Dependence parameter: For two random variables V, W we define

$$\psi(V \mid W) := \sup_{A,B}\left|\frac{\mathbb{P}(V\in A, W\in B) - \mathbb{P}(V\in A)\mathbb{P}(W\in B)}{\mathbb{P}(V\in A)\mathbb{P}(W\in B)}\right|$$

and quantify the dependence in the Markovian sequence as

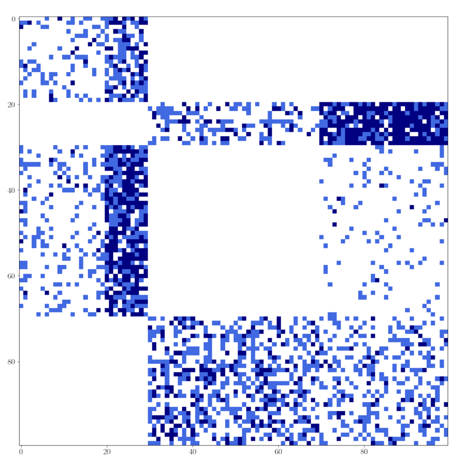
$$\Psi := \min\{j \ge 1 : \psi(Z_{i+j} \mid Z_i) \le 1/4 \text{ for all } i \in \{1, \dots, n-j\}\}$$

Free-probabilistic quantity: We define a quantity $\|\mathbf{S}_{\text{free}}\|$ satisfying $\sigma \le \|\mathbf{S}_{\text{free}}\| \le 2\sigma$ by $\|\mathbf{S}_{\text{free}}\| = \inf_{\mathbf{W}\succ 0} \lambda_{\max}(\mathbf{W}^{-1} + \mathbb{E}[\mathbf{SWS}])$.

Matrix parameters: Recall that $\sigma^2 = \|\mathbb{E}[\mathbf{S}^2]\|$ and $\|\mathbf{X}_i\| \le R$. We further also use $\varsigma^2 := \|\sum_{i=1}^n \mathbb{E}[\mathbf{X}_i^2]\|$ and $v^2 = \|\operatorname{Cov}(\mathbf{S})\|$.

Application: Block Markov chains

Fix a positive integer $K \ge 1$ and let $\mathcal{V}_1, \dots, \mathcal{V}_K$ be a partition of $\{1, \dots, d\}$ into nonempty sets. Then, a *block Markov chain* with cluster transition matrix $\mathbf{p} \in [0,1]^{K\times K}$ is a Markov chain with transition probabilities $\mathbb{P}(Z_{t+1} = j \mid Z_t = i) = \mathbf{p}_{a,b}/\#\mathcal{V}_b$ for $i \in \mathcal{V}_a$ and $j \in \mathcal{V}_b$.



The associated *empirical frequency matrix* is the $d \times d$ matrix $\hat{\mathbf{N}}$ with entries given by $\hat{\mathbf{N}}_{i,j} := \sum_{t=1}^{n-1} \mathbb{1}\{(Z_t, Z_{t+1}) = (i,j)\}$. Concentration properties of this matrix are essential in the study of spectral clustering algorithms [SPY20].

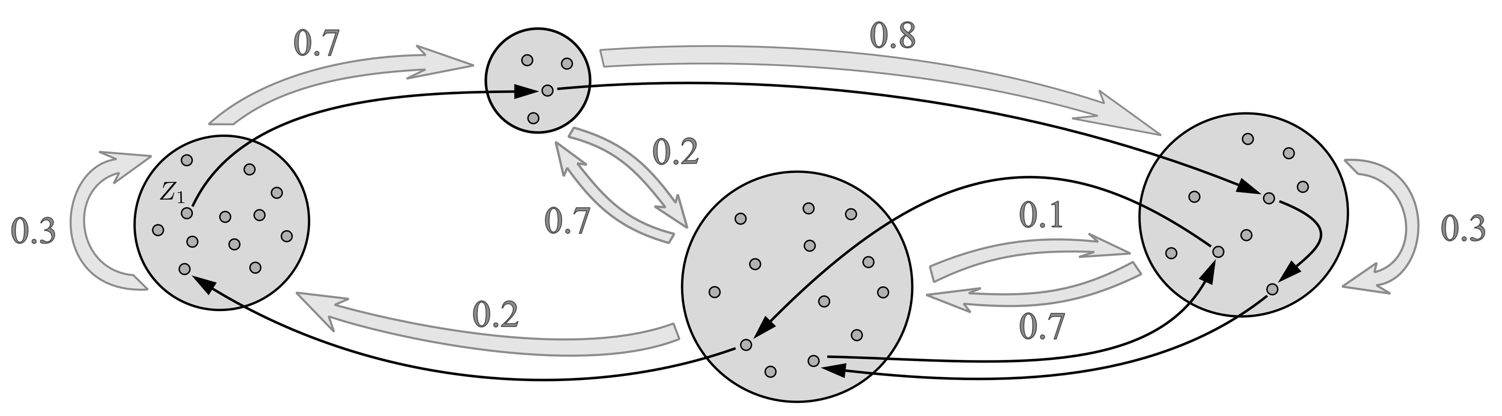


Figure 1. Visualization of a block Markov chain on $K = 4$ clusters.

Lemma. Suppose $\lim_{d\to\infty} \#\mathcal{V}_i/d > 0$ for every $i \in \{1, \dots, K\}$. One can then represent $\sqrt{n/d}(\hat{\mathbf{N}} - \mathbb{E}[\hat{\mathbf{N}}])$ in terms of the Markovian model with

$$\sigma^2 = O(1), \quad \varsigma^2 = O(1), \quad \Psi = O(1), \quad R = O(\sqrt{d/n}), \quad v^2 = O(1/d).$$

In particular, when $n = \omega(d\ln(2d)^4)$, our main result yields that $\mathbb{E}[\|\mathbf{S}\|] \le \|\mathbf{S}_{\text{free}}\| + o(1)$.

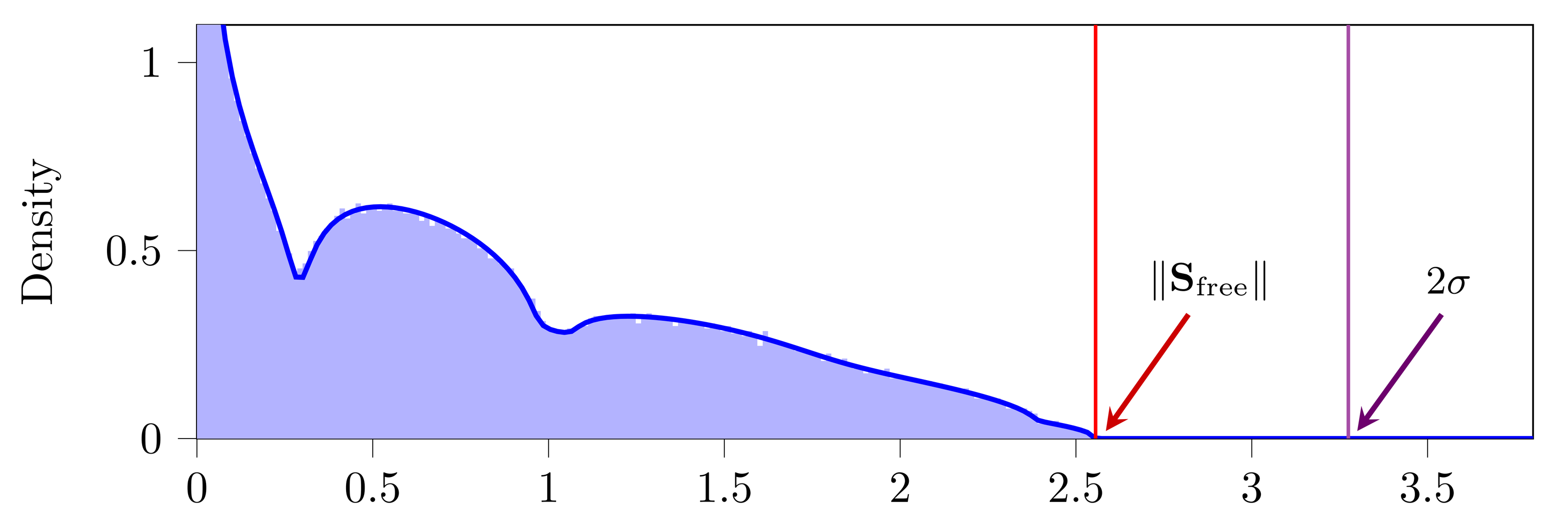


Figure 2. Singular value distribution of $\sqrt{d/n}(\hat{\mathbf{N}} - \mathbb{E}[\hat{\mathbf{N}}])$ as well as our leading-order term $\|\mathbf{S}_{\text{free}}\|$. We observe that the inequality $\|\mathbf{S}_{\text{free}}\| \le 2\sigma$ is strict. Not only have we reduced the logarithmic dimensional dependence, we also get a sharp leading-order term!

Result: Sharp concentration inequality

Theorem. In the Markovian model, there exists an absolute constant $c > 0$ such that for any $0 \le \delta \le 1$ and $x \ge 0$

$$\mathbb{P}\big(\|\mathbf{S}\| \ge (1+\delta)\|\mathbf{S}_{\text{free}}\| + c\varepsilon(x)\big) \le 2d(1+\delta)^{-x}$$

where

$$\varepsilon(x) = v^{1/2}\sigma^{1/2}x^{3/4} + R^{1/3}\Psi^{2/3}\varsigma^{2/3}x^{2/3} + R\Psi x.$$

In particular,

$$\mathbb{E}[\|\mathbf{S}\|] \le (1+\delta)\|\mathbf{S}_{\text{free}}\| + c\varepsilon(\ln(2d)).$$

Proof sketch

Recall that the Gaussian case is well-understood due to [BBvH21]. Our general strategy involves a universality argument: we show that $\mathbb{E}[\operatorname{tr}\mathbf{S}^{2p}] \approx \mathbb{E}[\operatorname{tr}\mathbf{G}^{2p}]$ for all $p \ge 1$. The desired result then follows using the Markov inequality.

Let $\mathbf{G}$ be a *Gaussian model* of $\mathbf{S}$. For $t \in [0,1]$ we show that $\frac{d}{dt}\mathbb{E}[\operatorname{tr}\mathbf{S}(t)^{2p}]$ is small where

$$\mathbf{S}(t) := \sqrt{t}\mathbf{S} + \sqrt{1-t}\mathbf{G}.$$

In a setting with independent summands, a similar strategy was employed by [BvH22]. Their argument relies on an expansion in terms of classical cumulants. The efficiency of this approach is due to an independence-implies-vanishing property of classical cumulants. This property does not apply to us.

Our new idea is to instead use Boolean cumulants. These are not as well-known as classical cumulants but have the advantage that they enjoy a nice interaction with the Markovian structure. By also using an identity from [AHLV15] and a change-of-measure we show that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[\operatorname{tr}\mathbf{S}(t)^{2p}] = \frac{1}{2}\sum_{k=3}^{\infty}\frac{t^{k/2}}{(k-1)!}\mathbb{E}[\sum_{(\rho,\alpha,i)} \Delta^{(\rho,\alpha,i)}\partial_{\mathbf{X}_{i,1}^{(\rho,\alpha,i)}}\cdots\partial_{\mathbf{X}_{i,k}^{(\rho,\alpha,i)}}\operatorname{tr}\mathbf{S}(t)^{2p}]$$

where $\mathbf{X}_{i,j}^{(\rho,\alpha,i)}$ has the same marginal distribution as $\mathbf{X}_{i_j}$ and $\Delta^{(\rho,\alpha,i)}$ are scalar random variables measuring the decay of dependence. Inspecting the combinatorics associated with the identity of [AHLV15], we show that

$$\sum_{(\rho,\alpha,i)} \|\Delta^{(\rho,\alpha,i)}\|_{\mathcal{L}^\infty} \le 2^{k-1}k! \sum_{i\in\{0\}\times\mathbb{Z}_{\ge 0}^{k-1}} \exp\Big(-\ln(4)\sum_{j=1}^{k}\mathbb{1}\{i_j > i_{j-1}\}\Big\lfloor\frac{i_j - i_{j-1}}{\Psi}\Big\rfloor\Big) \le k!c^k\Psi^k.$$

The directional derivative of tracial moments is explicit:

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

Combining this with a trace inequality and the fundamental theorem of calculus it can be shown that $\mathbb{E}[\operatorname{tr}\mathbf{S}^{2p}] \approx \mathbb{E}[\operatorname{tr}\mathbf{G}^{2p}]$ as desired.

Further results

Using the same method of proof, we establish the following:

1. **Universality for tracial moments:** For any positive integer $p \ge 1$ one has that $\mathbb{E}[\operatorname{tr}\mathbf{S}^p] \approx \mathbb{E}[\operatorname{tr}\mathbf{G}^p]$ and $\operatorname{Var}[\operatorname{tr}\mathbf{S}^p] \approx \operatorname{Var}[\operatorname{tr}\mathbf{G}^p]$. This can be used to establish universality of empirical eigenvalue or singular value distributions improving upon [SVW23].
2. **Similar results in the matrix series model:** In the matrix series model we have $\mathbf{X}_i = Y_i\mathbf{A}_i$ for deterministic $\mathbf{A}_i$ and random real-valued Y_i. We here quantify dependence and heavy-tailedness using a cumulant-based parameter $K_k := \max_{i_1}\sum_{i_2,\dots,i_k}|\kappa(Y_{i_1}, Y_{i_2}\dots, Y_{i_k})|$.
3. **Applications of the matrix series model:** We show concentration for matrices with independent sub-Weibull entries. Further, we establish convergence of the empirical eigenvalue distribution of the adjacency matrix in the $G(n,m)$ random graph model.

Acknowledgements

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).

References

[AHLV15] O. Arizmendi, T. Hasebe, F. Lehner, and C. Vargas. Relations between cumulants in noncommutative probability. *Advances in Mathematics*, 2015.

[BBvH21] A.S. Bandeira, M.T. Boedihardjo, and R. van Handel. Matrix Concentration Inequalities and Free Probability. *arXiv preprint arXiv:2108.06312v2*, 2021.

[BvH22] T. Brailovskaya and R. van Handel. Universality and sharp matrix concentration inequalities. *arXiv preprint arXiv:2201.05142*, 2022.

[Oli09] R.I. Oliveira.Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges. *arXiv preprint arXiv:0911.0600*, 2009.

[SPY20] J. Sanders, A. Proutière, and S.-Y. Yun. Clustering in block Markov chains. *The Annals of Statistics*, 2020.

[SVW23] J. Sanders and A. Van Werde. Singular value distribution of dense random matrices with block Markovian dependence. *Stochastic Processes and their Applications*, 2023.

[Tro12] J.A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 2012.