Intrinsic-dimension empirical Bernstein inequalities for bounded self-adjoint operators

Diego Martinez-Taboada
Department of Statistics & Data Science
Carnegie Mellon University
Pittsburgh, PA 15213
diegomar@andrew.cmu.edu

Aaditya Ramdas
Department of Statistics
Stanford University
Stanford, CA 94305
aramdas@stanford.edu

Abstract

Operator-valued concentration inequalities are foundational to the analysis of modern high-dimensional statistics and randomized algorithms. However, standard oracle bounds are frequently limited in practice: they require explicit a priori knowledge of the true variance, and often explicitly scale with the ambient dimension, rendering them vacuous for infinite-dimensional or heavily structured operators. Motivated by these challenges, we establish the first empirical Bennett and Bernstein inequalities for sums of independent, bounded, self-adjoint Hilbert-Schmidt operators. Our fully data-driven bounds replace the unknown variance with an empirical estimate and rely strictly on the intrinsic dimension rather than the ambient dimension. This structural shift yields computable, dimension-free guarantees with a sharper first-order asymptotic radius for non-isotropic random matrices and seamlessly extends to infinite-dimensional Hilbert spaces. We demonstrate that our empirical bounds achieve asymptotic sharpness with the best known oracle rates. Finally, as an independent byproduct, we derive novel empirical concentration guarantees for the intrinsic dimension itself.

1 Introduction

Matrices and linear operators are ubiquitous across modern high-dimensional statistics and machine learning [42, 41]. In particular, symmetric matrices and, more generally, compact self-adjoint operators serve as the mathematical foundation for representing complex data structures and transformations. This specific class of operators is of fundamental interest because, unlike general asymmetric matrices or non-compact operators, they possess a rigorously well-behaved, real, and discrete spectrum [33, 7]. This spectral structure is critical: it ensures that quantities like eigenvalues and the trace remain well-defined even in infinite-dimensional Hilbert spaces, allowing finite-dimensional intuition to seamlessly extend to functional spaces. Consequently, compact self-adjoint operators naturally represent empirical covariance structures [15, 17], Hessians in optimization [5, 8], graph adjacency matrices for undirected networks [6, 35], and physical observables in quantum mechanics [29, 12].

Because these systems are frequently governed by the accumulation of random noise or data, a core challenge in the field is understanding the spectral behavior of sums of independent random compact self-adjoint operators (X_i), which we denote as

$$S_n := \sum_{i=1}^{n} X_i.$$

Establishing rigorous bounds on the fluctuations of S_n is essential for proving the reliability and convergence rates of a vast array of learning algorithms. To provide probabilistic tail bounds for the spectral properties of S_n, operator-valued concentration inequalities have become fundamental

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

tools [1, 37, 18]. Traditionally, these bounds depend explicitly on the ambient dimension d of the underlying space. However, a significant line of work has shifted focus toward bounds that depend instead on the *intrinsic dimension* of the variance structure [14, 27, 40]. This shift enables the natural extension of concentration phenomena to infinite-dimensional operators and yields strictly sharper, dimension-free guarantees for non-isotropic random matrices, where the spectrum decays rapidly, making the ambient dimension a severe overestimate of the true complexity of the problem.

Despite their theoretical elegance, standard oracle inequalities, such as the operator-valued Bennett's and Bernstein's inequalities, are limited in practice. Throughout this work, we are specifically interested in the regime where these random operators are uniformly bounded in operator norm, as this boundedness is a fundamental prerequisite for deriving Bennett- and Bernstein-type tail bounds. However, even under this boundedness assumption, these oracle inequalities remain restricted in their application because they additionally require *a priori* knowledge of the exact variance of the random operators. This limitation motivates the study of empirical Bernstein inequalities [3, 25, 43]. By substituting the true variance with an empirical estimate, these bounds theoretically guarantee the existence of fully data-driven algorithms and computable confidence sets that adapt to the observed variability of the data, circumventing the need for prior knowledge of the variance.

The central ML application is operator estimation: learning procedures replace a population covariance, kernel, curvature, or spectral operator by a sample average. An observable operator-norm radius transfers through perturbation bounds to uncertainty in eigenvalues, learned subspaces, and regularized transformations. For kernel and functional PCA [24, 34], this certifies the stability of the learned representation even in infinite-dimensional feature spaces. More broadly, it can assess sample adequacy for a desired operator accuracy and calibrate regularization against estimation error. Intrinsic dimension is especially useful when the feature dimension is huge or infinite but the variance spectrum is non-isotropic; related settings include random graphs [30] and quantum state estimation [9, 10].

Motivated by these challenges, we revisit the derivation of empirical Bennett and Bernstein inequalities in the operator-valued setting. Our contributions are summarized as follows:

- We derive the first known empirical Bennett and Bernstein inequalities applicable to bounded self-adjoint Hilbert-Schmidt operators, including infinite-dimensional settings.
- Our inequalities improve the first-order asymptotic radius for non-isotropic random matrices, because they scale with respect to the intrinsic dimension $\mathrm{tr}(\Sigma)/\|\Sigma\| \leq d$, as opposed to the ambient dimension d. In particular, our inequality is *sharp* (asymptotically equivalent) to the oracle inequality from Martinez-Taboada and Ramdas [23].
- As a byproduct of our theoretical analysis, we provide empirical concentration guarantees for the intrinsic dimension $\mathrm{tr}(\Sigma)/\|\Sigma\|$ itself.

2 Related work

Operator-valued dimension-free concentration inequalities. Our theoretical framework shares deep connections with the operator-valued intrinsic-dimension inequalities established by Minsker [27] and Tropp et al. [40], recently unified by Martinez-Taboada and Ramdas [23]. These works notably refined the earlier bounds from Hsu et al. [14]. Intrinsic-dimension guarantees have also been explored through the lens of PAC-Bayes theory by Zhivotovskiy [45]. It is also worth noting other dimension-free approaches, such as those by Oliveira [31] and Magen and Zouzias [19], though these bounds depend explicitly on the rank of the operators rather than their intrinsic dimension.

Operator-valued dimension-dependent concentration inequalities. Historically, the prevailing operator-valued concentration bounds have relied explicitly on the ambient dimension. A comprehensive collection of such results can be found in Tropp [37] and Tropp et al. [40]. Related contributions include Tropp [36], Oliveira [30], Mackey et al. [18]. Within this regime, Wang and Ramdas [43] extended some of these results to obtain an empirical Bernstein inequality for matrices.

Empirical Bernstein inequalities. The core focus of this paper is the development of novel empirical Bernstein inequalities for operators. In the scalar setting, empirical Bernstein bounds were first introduced by Audibert et al. [3] and Maurer and Pontil [25], spurring a robust line of subsequent research [28, 26, 13, 32, 44]. Moving beyond the one-dimensional case, empirical Bernstein-type guarantees have recently been extended to estimators of the variance, martingales in Banach spaces,

self-normalized processes, and random matrices [21, 22, 20, 43]. The closest antecedent, Wang and Ramdas [43], provides a plug-in bound for independent matrices and a self-normalized, time-uniform bound under martingale dependence. Their two-sided first-order radius uses $\log(2d/\delta)$, whereas ours uses $\log(2d_{\text{eff}}/\delta)$, with $d_{\text{eff}} = \text{tr}(\Sigma)/\|\Sigma\|$. Theorem 5.6 targets the intrinsic oracle rather than the ambient Tropp oracle. The gain is largest when $d_{\text{eff}} \ll d$, and disappears asymptotically when $d_{\text{eff}} = d$. We do not claim uniform finite-sample dominance: the additional trace and fourth-moment estimation can make the self-normalized bound tighter at small n.

3 Background

3.1 Operator theory

Let $\mathcal{B}(\mathbb{H})$ denote bounded operators on a separable Hilbert space $\mathbb{H}$, and $\mathcal{B}^{\text{s}}(\mathbb{H})$, $\mathcal{B}^{++}(\mathbb{H})$, and $\mathcal{B}^{\text{cs}}(\mathbb{H})$ its self-adjoint, strictly positive, and compact self-adjoint classes. Write $\mathcal{S}_2^{\text{s}}(\mathbb{H})$ for self-adjoint Hilbert-Schmidt operators: $\|A\|_{\text{HS}}^2 = \text{tr}(A^2) < \infty$. These are compact but need not be trace-class. We use $\text{tr}(A)$ only for trace-class A, the operator norm $\|A\| = \max\{|\inf \text{spec}(A)|, |\sup \text{spec}(A)|\}$, and the Loewner order $\preceq$. Scalar functions extend by spectral calculus.

To control operator moment generating functions (MGFs), we use the following finite-dimensional trace results [11, 16, 2].

Theorem 3.1 (Trace Jensen inequality)**.** *On self-adjoint matrices, $X \mapsto tr f(X)$ is convex for every continuous convex $f : \mathbb{R} \to \mathbb{R}_{\geq 0}$.*

Theorem 3.2 (Lieb's concavity theorem)**.** *For a fixed self-adjoint matrix A, $X \mapsto tr \exp(\log X + A)$ is concave on positive definite matrices.*

The trace Jensen inequality also yields the submartingale property needed to apply Doob's maximal inequality.

Lemma 3.3. *Let X_i be independent mean-zero self-adjoint matrices, $S_k = \sum_{i \leq k} X_i$, and f be continuous, convex, and nonnegative. If $\mathbb{E} tr f(\theta S_k) < \infty$ for every $k \leq n$, then $(tr f(\theta S_k))_{k=0}^n$ is a nonnegative submartingale.*

To apply these tools in infinite-dimensional spaces, we use finite-rank compressions, as in Minsker [27]. The limiting argument in Appendix E retains centered expressions: $\varphi(A) = \cosh(A) - I$ is trace-class for $A \in \mathcal{S}_2^{\text{s}}(\mathbb{H})$, as is $e^{\psi V} - I$ for positive trace-class V and finite $\psi \geq 0$. This keeps the traces finite while extending the MGF bound and submartingale property to the full space.

3.2 Operator-norm Bernstein and Bennett inequalities

The following theorem [23] is a strict generalization of the one-dimensional Bennett inequality to operators. Compared to its one-dimensional counterpart, it pays an extra pre-factor of $\text{tr}(V_n)/\sigma^2$.

Theorem 3.4 (Operator-valued Bennett's inequality)**.** *Assume independent $X_i \in \mathcal{S}_2^s(\mathbb{H})$ with $\mathbb{E}\|X_i\|_{\text{HS}}^2 < \infty$, such that $\mathbb{E}X_i = 0$ and $\|X_i\| \leq c$ almost surely for all i. Define $V_n := \sum_{i \leq n} \mathbb{E}X_i^2$ and $h(u) := (1+u)\log(1+u) - u$. If $0 < \sigma^2 \in [\|V_n/n\|, tr(V_n/n)]$, then*

$$\mathbb{P}\left(\sup_{i \leq n} \|S_i\| \geq r\right) \leq 2\frac{tr\,(V_n/n)}{\sigma^2} \exp\left(-\frac{n\sigma^2}{c^2} h\left(\frac{cr}{n\sigma^2}\right)\right). \tag{1}$$

Theorem 3.4 unified previous results, including those from Tropp [37] and Minsker [27]. In particular, (1) depends on the *intrinsic dimension* $d_{\text{eff}} = \text{tr}(V_n)/\|V_n\|$ if $\sigma^2 = \|V_n\|$. Importantly, it holds that $1 \leq \text{tr}(V_n)/\sigma^2 \leq \text{rank}(V_n) \leq d$. In the isotropic case $V_n = I$, it follows that $\text{tr}(V_n)/\sigma^2 = d$. However, the inequality $\text{tr}(V_n)/\sigma^2 \leq d$ is strict if V_n is not isotropic; for instance, if all but one of the eigenvalues are zero, then $\text{tr}(V_n)/\sigma^2 = 1$ even if d is infinite.

We recall that Bennett's inequality is sharper than Bernstein's inequality, and so the latter can be recovered from the former. Applying a common lower bound on h (Lemma C.2), we can obtain an operator-valued Bernstein's inequality from Theorem 3.4. In particular, for $\mathbb{E}X_i = \mu$, $\|X_i\| \leq c$, and $\Sigma = \mathbb{E}(X_i - \mu)^2 = V_n/n$, taking $\sigma^2 = \|\Sigma\| > 0$, Theorem 3.4 and Lemma C.2 yield

$$\|S_n/n - \mu\| \leq \sqrt{\frac{2\|\Sigma\|}{n} \log\left(\frac{\text{tr}\,(\Sigma)}{\sigma^2}\frac{2}{\delta}\right)} + \frac{2c}{3n}\log\left(\frac{\text{tr}\,(\Sigma)}{\sigma^2}\frac{2}{\delta}\right). \tag{2}$$

3.3 Scalar empirical Bennett and Bernstein inequalities

The practical utility of Bernstein and Bennett inequalities is limited by the requirement of knowing the variance, a quantity that is generally unknown. *Empirical* Bernstein inequalities adapt to the observed variability of the data, typically yielding sharper bounds than Hoeffding's inequality, while remaining fully data-driven and requiring no additional assumptions beyond boundedness.

In particular, Audibert et al. [3], Maurer and Pontil [25] presented the first scalar empirical Bernstein inequalities. These rely on invoking an oracle Bernstein (or Bennett) inequality in conjunction with another concentration inequality for the variance, combining the two via a union bound. In particular, Maurer and Pontil [25, Theorem 11] proved the following scalar empirical Bernstein bound.

Theorem 3.5. *Let $Z_1, \dots, Z_n$ be independent random variables in $[0, B]$, with $n \geq 2$. It holds that*

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}Z_i - \frac{1}{n}\sum_{i=1}^{n}Z_i \leq \varsigma_n(\mathbf{Z})\sqrt{\frac{2\log(2/\delta)}{n}} + \frac{7B\log(2/\delta)}{3(n-1)},$$

with probability at least $1-\delta$, where

$$\varsigma_n^2(\mathbf{Z}) = \frac{1}{n(n-1)}\sum_{1\leq i<j\leq n}(Z_i - Z_j)^2. \tag{3}$$

We anchor our methodology to this specific empirical Bernstein formulation due to its clean and highly interpretable structure. However, we emphasize that other one-dimensional empirical Bernstein inequalities exist [28, 26, 13, 32, 44].

Methodologically, our approach mirrors the analytical framework established by Maurer and Pontil [25] for the proof of Theorem 3.5. Specifically, we rely on upper bounds for our variance terms, and combine them with an oracle inequality via union bounding. While this structural strategy was recently adapted by Wang and Ramdas [43] to develop a matrix-valued empirical Bernstein inequality, their resulting guarantees remain fundamentally bottlenecked by an explicit dependence on the ambient dimension. Building upon this foundation, the remainder of this work is dedicated to establishing an empirical inequality that adapts to intrinsic dimension.

4 Oracle concentration inequalities

The main approach of this contribution is to build empirical inequalities via union bounding the inequalities (1) and (2) with upper bounds (in high probability) for the parameters $\mathrm{tr}(V_n)$ and σ. This approach is in close spirit to that of Maurer and Pontil [25], but in the operator-valued regime.

A critical challenge in empirically applying (1) and (2) is that their right-hand sides are not monotonically increasing with respect to σ^2. For the Bernstein bound (2), this is evident since $\log\big(2\mathrm{tr}(V_n)/(\sigma^2\delta)\big) \to \infty$ as $\sigma^2 \to 0$, while the non-monotonicity of the Bennett bound (1) is formally established in Lemma C.3. Naively, this structural property suggests that one must construct both high-probability upper and lower bounds for the variance parameter σ^2. Fortunately, we can refine the foundational master theorems underlying (1) and (2) to circumvent this issue, showing that only a high-probability upper bound on σ^2 is strictly necessary.

We formalize such a generalized result in the following theorem, deferring its proof to Appendix B.1. Throughout, we let $\psi : [0, \theta_{\max}) \to [0, \infty)$, with $\psi(0) = 0$, be a scalar cumulant-generating-function (CGF). The deterministic operators $\Delta V_i \succeq 0$ are trace-class pseudo-variance increments, with $V_n = \sum_{i\leq n}\Delta V_i$. Operator MGF expectations below are assumed to exist.

Theorem 4.1. *Let $X_1, \dots, X_n \in \mathcal{S}_2^s(\mathbb{H})$ be independent random operators such that $\mathbb{E}X_i = 0$ and*

$$\log \mathbb{E}\exp(\theta X_i) \preceq \psi(|\theta|)\Delta V_i, \tag{4}$$

for all $\theta \in (-\theta_{\max}, \theta_{\max})$ and all $i \in \{1, \dots n\}$. If finite $\tau_u(\delta_2) \geq 0$ and $\sigma_u^2(\delta_3) > 0$ satisfy

$$\mathbb{P}\left(tr(V_n) \geq n\tau_u(\delta_2)\right) \leq \delta_2, \quad \mathbb{P}\left(\|V_n\| \geq n\sigma_u^2(\delta_3)\right) \leq \delta_3, \tag{5}$$

where $V_n := \sum_{i\leq n}\Delta V_i$, then

$$\mathbb{P}\left(\sup_{i\leq n}\|S_i\| \geq \Pi^{-1}_{\psi,\delta_2,\delta_3}(\delta_1)\right) \leq \delta_1 + \delta_2 + \delta_3,$$

$$\Pi_{\psi,\delta_2,\delta_3}(r) := 2\left(\frac{\tau_u(\delta_2)}{\sigma_u^2(\delta_3)} \vee 1\right)\left[\inf_{\theta\in[0,\theta_{\max})} \exp\left(n\psi(\theta)\sigma_u^2(\delta_3) - \theta r\right)\right]. \tag{6}$$

On $E = \{\mathrm{tr}(V_n) \le n\tau_u\} \cap \{\|V_n\| \le n\sigma_u^2\}$, the MGF master bound is at most the observable $\Pi_{\psi,\delta_2,\delta_3}(r)$ in (6). Thus upper confidence bounds alone suffice, despite the non-monotonicity of (1). Comparing the empirical threshold with a deterministic valid threshold on E, and applying a union bound to E^c, gives the additional failure probability $\delta_2 + \delta_3$.

With the generalized master theorem established, we can now specialize it to derive an oracle operator-valued Bennett's inequality; this result is central to our approach. Crucially, the structural separation achieved in Theorem 4.1 allows us to safely substitute the unknown, true variance parameters with their respective high-probability upper bounds. We formalize this result below, deferring the rigorous algebraic derivation to Appendix B.2.

Corollary 4.2 (Operator-valued Bennett's inequality)**.** *Assume independent $X_i \in \mathcal{S}_2^s(\mathbb{H})$ with $\mathbb{E}\|X_i\|_{\mathrm{HS}}^2 < \infty$, such that $\mathbb{E}X_i = 0$ and $\|X_i\| \le c$ almost surely for all i. Define $V_n := \sum_{i\le n}\mathbb{E}X_i^2$ and $h(u) := (1+u)\log(1+u) - u$. If τ_u and σ_u are functions such that* (5) *holds, then*

$$\mathbb{P}\left(\sup_{i\le n}\|S_i\| \ge \Pi^{-1}_{c,\delta_2,\delta_3}(\delta_1)\right) \le \delta_1 + \delta_2 + \delta_3,$$

where

$$\Pi_{c,\delta_2,\delta_3}(r) := 2\left(\frac{\tau_u(\delta_2)}{\sigma_u^2(\delta_3)} \vee 1\right)\exp\left(-\frac{n\sigma_u^2(\delta_3)}{c^2}h\left(\frac{cr}{n\sigma_u^2(\delta_3)}\right)\right). \tag{7}$$

5 Empirical concentration inequalities

Having established the requisite oracle bounds, we now transition to developing fully empirical concentration inequalities, our primary objective. Across the literature, the derivation of empirical bounds (whether in scalar, vector, or matrix settings) fundamentally relies on structural boundedness assumptions. In our infinite-dimensional operator framework, this necessitates controlling both the operator norm of the sequence and the trace of its squared elements. To this end, we impose the following conditions throughout this section.

Assumption 5.1. *Let $X_1, \dots, X_n \in \mathcal{S}_2^s(\mathbb{H})$ be independent self-adjoint Hilbert-Schmidt operators such that*

$$\|X_i\| \le c, \quad tr\left(X_i^2\right) \le B, \quad a.s., \tag{boundedness}$$

as well as

$$\mathbb{E}X_i = \mu, \quad \mathbb{V}X_i = \Sigma, \quad \mathbb{V}\left[(X_i-\mu)^2 - \Sigma\right] = \Sigma'. \tag{constant moments}$$

Here $c \ge \operatorname{ess\,sup}_i \|X_i\|$ and $B \ge \operatorname{ess\,sup}_i \|X_i\|_{\mathrm{HS}}^2$ are known envelopes, not tuned hyperparameters. "Fully empirical" refers to variance and intrinsic dimension. Rank-r updates admit $B = rc^2$; for $X_i = x_i x_i^*$ with $\|x_i\| \le R$, take $c = R^2$, $B = R^4$. Kernel updates with $k(x,x) \le \kappa^2$ admit $c = \kappa^2$, $B = \kappa^4$ (both equal to 1 for our normalized kernel). These structural choices can be much tighter than $B = dc^2$. The observations need not be trace-class, but $(X_i - \mu)^2$, Σ, and the paired estimators and their second moments are positive trace-class. Larger c enlarges the linear and second-level estimation terms; larger B enlarges trace bounds and can inflate the empirical dimension factor. These effects matter at moderate n but are lower order for fixed valid c, B in Theorem 5.6. Unbounded observations require truncation and bias control, outside our scope.

The remainder of this section is dedicated to constructing empirical Bennett and Bernstein inequalities under Assumption 5.1. We start by defining second and fourth moment random variables (X_i') and $(X_i'i)$ in Section 5.1, which serve as foundational tools for our subsequent analysis. Next, in Section 5.2, we introduce a secondary master theorem (Theorem 5.2) that yields operator-valued empirical inequalities when combined with Theorem 4.1. We leverage this machinery to derive our core empirical Bennett inequality in Section 5.3. In Section 5.4, we present a cleaner, albeit slightly looser, empirical Bernstein inequality, rigorously establishing its asymptotic sharpness relative to the oracle bound (2) in Theorem 5.6. We defer to Appendix A an extension of the aforementioned results to the concentration phenomena of the intrinsic dimension itself.

5.1 Auxiliary random variables

To streamline our analysis, we assume throughout that the sample size is a multiple of four, i.e., $n \equiv 0 \pmod 4$, with $n \geq 8$ so both scalar variance estimates are defined. We begin by constructing a sequence of auxiliary random variables that serve as unbiased, independent estimators for the true variance Σ. Under Assumption 5.1, we take

$$X_i' := \frac{1}{2}\left(X_{2i-1} - X_{2i}\right)^2, \quad i \in \left[\frac{n}{2}\right]; \quad \mathbb{E}X_i' = \frac{1}{2}\mathbb{E}\left[(X_{2i-1} - \mu) - (X_{2i} - \mu)\right]^2 = \Sigma.$$

Thus, we can generate $n/2$ independent random operators X_i' that are unbiased for Σ. Crucially, these new operators inherit the boundedness properties of the original sequence. Under Assumption 5.1, the trace and operator norm of X_i' are uniformly bounded as

$$0 \leq \mathrm{tr}\left(X_i'\right) = \frac{1}{2}\mathrm{tr}\left((X_{2i-1} - X_{2i})^2\right) \leq \frac{1}{2}\mathrm{tr}\left(X_{2i-1}^2 + X_{2i}^2\right) + \sqrt{\mathrm{tr}\left(X_{2i-1}^2\right)\mathrm{tr}\left(X_{2i}^2\right)} \leq 2B, \quad (8)$$

$$0 \leq \|X_i'\| = \frac{1}{2}\left\|(X_{2i-1} - X_{2i})^2\right\| \leq \frac{1}{2}\left[\|X_{2i-1}\|^2 + \|X_{2i}\|^2 + 2\|X_{2i-1}\|\|X_{2i}\|\right] \leq 2c^2. \quad (9)$$

Building upon this first auxiliary sequence, we apply the exact same differencing technique to construct a second sequence, designed to estimate the variance of the paired estimator. Write $\widetilde{\Sigma}_i' = \mathbb{E}[(X_i' - \Sigma)^2]$ and $\widetilde{\Sigma}' = (2/n)\sum_{i \leq n/2}\widetilde{\Sigma}_i'$. In the i.i.d. case, $\widetilde{\Sigma}' = \mathbb{E}[(X_i' - \Sigma)^2]$; it is distinct from Σ' in Assumption 5.1. We define

$$X_i'' := \frac{1}{2}\left(X_{2i-1}' - X_{2i}'\right)^2, \quad i \in \left[\frac{n}{4}\right]; \quad \mathbb{E}X_i'' = \tfrac{1}{2}(\widetilde{\Sigma}_{2i-1}' + \widetilde{\Sigma}_{2i}').$$

Thus $(4/n)\sum_{i \leq n/4} X_i''$ is unbiased for $\widetilde{\Sigma}'$. Furthermore, the trace of X_i'' is uniformly upper bounded. Combining the positive semi-definiteness of X_i' with (8) and (9), we obtain

$$\begin{aligned} 0 \leq \mathrm{tr}\left(X_i''\right) &= \frac{1}{2}\mathrm{tr}\left(\left(X_{2i-1}'\right)^2 + \left(X_{2i}'\right)^2 - 2X_{2i-1}'X_{2i}'\right) \leq \frac{1}{2}\mathrm{tr}\left(\left(X_{2i-1}'\right)^2 + \left(X_{2i}'\right)^2\right) \\ &\leq \frac{1}{2}\left[\mathrm{tr}\left(X_{2i-1}'\right)\|X_{2i-1}'\| + \mathrm{tr}\left(X_{2i}'\right)\|X_{2i}'\|\right] \leq \frac{1}{2}\left(4Bc^2 + 4Bc^2\right) = 4Bc^2. \end{aligned}$$

5.2 A master theorem

To derive a fully empirical operator-valued inequality inspired by the scalar framework of Maurer and Pontil [25], we must construct a high-probability upper bound for the variance parameter $\|\Sigma\|$. A naive strategy would be to apply our oracle bound directly to the auxiliary sequence (X_i'). However, because $\mathbb{V}X_i' = \widetilde{\Sigma}_i'$, this would yield a concentration inequality for $\|\Sigma\|$ that inherently depends on the fourth-moment parameter $\|\widetilde{\Sigma}'\|$. Repeating this process to bound $\|\widetilde{\Sigma}'\|$ would simply shift the dependence to the eighth moment, trapping the analysis in an infinite recursion of higher-order variance terms. Therefore, we must establish a self-contained concentration inequality for $\|\Sigma\|$ that successfully halts this recursive chain.

To achieve this, we derive a secondary, structurally distinct master theorem. While this new master theorem yields a fundamentally looser bound than Theorem 4.1, it naturally closes the estimation loop, as it can be invoked without requiring recursive bounds on higher-order variances. This mirrors the underlying philosophy of Maurer and Pontil [25], who utilized a looser variance bound to avoid recursion in the scalar setting. However, their techniques and our techniques are completely different, with them relying on self-bounding random variables and our tools fundamentally resorting to trace-operator norm inequalities. We formalize this result below, deferring the proof to Appendix B.3.

Theorem 5.2. *Let $X_1', \ldots, X_n' \in \mathcal{S}_2^s(\mathbb{H})$ be independent random operators, with ψ and trace-class $\Delta V_i' \succeq 0$ as in Theorem 4.1, such that $\mathbb{E}X_i' = 0$ and*

$$\log \mathbb{E}\exp(\theta X_i') \preceq \psi(|\theta|)\Delta V_i', \quad (10)$$

for all $\theta \in (-\theta_{\max}, \theta_{\max})$ and all $i \in \{1, \ldots n\}$. Defining $S_n' = \sum_{i \leq n} X_i'$ and $V_n' = \sum_{i \leq n}\Delta V_i'$, it holds that

$$\mathbb{P}\left(\sup_{i \leq n}\|S_i'\| \geq r\right) \leq 2\inf_{\theta \in (0, \theta_{\max})}\exp\left(\psi(\theta)tr(V_n') - \theta r\right). \quad (11)$$

Crucially, the bound established in Theorem 5.2 depends exclusively on the trace of the pseudo-covariance operators, completely bypassing both the operator norm of the variance and the ambient dimension of the space. When instantiated for bounded random operators, Theorem 5.2 yields dimension-free Bennett and Bernstein bounds for the variance as long as we know an upper bound on $\mathrm{tr}(\widetilde{\Sigma}')$. We formulate the resulting trace Bernstein inequality in the following corollary, and we defer its proof to Appendix B.4.

Corollary 5.3 (Trace Bernstein's inequality)**.** *For independent $X_i' \in \mathcal{S}_2^s(\mathbb{H})$, assume $\mathbb{E}X_i' = \Sigma$ and $\|X_i' - \Sigma\| \le 2c^2$ almost surely for all i. Let $S'_{n/2} = \sum_{i \le n/2} X_i'$ and $\widetilde{\Sigma}' = (2/n)\sum_{i \le n/2} \mathbb{E}[(X_i' - \Sigma)^2]$. Define*

$$\bar{\sigma}_u^2(\tau;\delta) := \|2S'_{n/2}/n\| + \Gamma_{c,\delta}(\tau), \quad \Gamma_{c,\delta}(\tau) := \sqrt{\frac{2\tau}{n/2}\log(2/\delta)} + \frac{2c^2}{3(n/2)}\log(2/\delta). \tag{12}$$

If $\tau \ge \mathrm{tr}(\widetilde{\Sigma}')$, then

$$\mathbb{P}\left(\|\Sigma\| \ge \bar{\sigma}_u^2(\tau;\delta)\right) \le \delta.$$

5.3 The empirical Bennett inequality

The core strategy for deriving our empirical Bennett inequality is to couple the oracle bound from Theorem 3.4 with high-probability estimators for the variance terms, drawing inspiration from the scalar framework of Maurer and Pontil [25]. Specifically, Corollary 4.2 guarantees a valid empirical bound, provided we can construct observable functions τ_u and σ_u that satisfy the probability conditions in (5). The remainder of this subsection is dedicated to constructing these functions.

We begin by defining the empirical estimators for the variance and the variance of the variance, relying on the independent blocks constructed in Section 5.1, as

$$\Sigma_n := \frac{2}{n}\left(X_1' + \ldots + X'_{\frac{n}{2}}\right) = 2S'_{n/2}/n, \quad \Sigma_n' := \frac{4}{n}\left(X_1'' + \ldots + X''_{\frac{n}{4}}\right).$$

Because the scalar random variables $Z_i = \mathrm{tr}(X_i')$ and $Z_i' = \mathrm{tr}(X_i'')$ are strictly bounded (as established in Section 5.1), we can directly apply the scalar empirical Bernstein bound (Theorem 3.5) to control their empirical means. We thus define the high-probability upper bound for the trace of the variance as

$$\tau_u(\delta) := \mathrm{tr}(\Sigma_n) + \varsigma_{n/2}(\mathbf{Z})\sqrt{\frac{2\log(2/\delta)}{n/2}} + \frac{7(2B)\log(2/\delta)}{3(n/2-1)}, \quad Z_i = \mathrm{tr}(X_i'). \tag{13}$$

Following the exact same procedure for the fourth moments yields

$$\tau_u'(\delta) := \mathrm{tr}(\Sigma_n') + \varsigma_{n/4}(\mathbf{Z}')\sqrt{\frac{2\log(2/\delta)}{n/4}} + \frac{7(4Bc^2)\log(2/\delta)}{3(n/4-1)}, \quad Z_i' = \mathrm{tr}(X_i''). \tag{14}$$

To bound the operator norm of the variance, we plug our empirical trace bound τ_u' into the trace Bernstein inequality derived in Corollary 5.3. Applying a standard union bound over the two error probabilities yields the following surrogate for σ^2:

$$\sigma_u^2(\delta_3) := \bar{\sigma}_u^2\left(\tau_u'\left(\frac{\delta_3}{2}\right); \frac{\delta_3}{2}\right). \tag{15}$$

The construction has two levels: pairing X_i gives Σ_n, and scalar Bernstein on $Z_i = \mathrm{tr}(X_i')$ gives τ_u in (13). Pairing again gives Σ_n' and the trace bound τ_u' in (14). Corollary 5.3 then converts $\tau = \tau_u'(\delta_3/2)$ into the norm bound (15), with the other $\delta_3/2$ allocated to that corollary. Hence $\mathbb{P}\{\|\Sigma\| > \sigma_u^2(\delta_3)\} \le \delta_3$. Since $V_n = n\Sigma$, equations (13) and (15) are exactly the two inputs to Corollary 4.2, yielding the theorem below (proof in Appendix B.5).

Theorem 5.4 (Operator-valued empirical Bennett inequality)**.** *Let $X_1, \ldots, X_n \in \mathcal{S}_2^s(\mathbb{H})$ satisfy Assumption 5.1. If* (7) *is defined with τ_u and σ_u as in* (13) *and* (15)*, then*

$$\mathbb{P}\left(\sup_{i \le n}\|S_i - i\mu\| \ge \Pi^{-1}_{2c,\delta_2,\delta_3}(\delta_1)\right) \le \delta_1 + \delta_2 + \delta_3.$$

The intrinsic-dimension improvement already applies at the Bennett level: with $V_n = n\Sigma$ and $\sigma^2 = \|\Sigma\| > 0$, the oracle bound (1) has prefactor $2d_{\text{eff}}$, where $d_{\text{eff}} = \text{tr}(\Sigma)/\|\Sigma\|$, instead of the ambient prefactor $2d$, with the same exponential rate. Thus $d_{\text{eff}} \ll d$ yields a smaller implicit confidence radius even before the Bernstein relaxation. We focus the comparison and sharpness discussion on Bernstein bounds because the closest prior empirical matrix results [43] are Bernstein-type.

While Theorem 5.4 provides a tight Bennett-style guarantee, its implicit rate function makes computing explicit confidence intervals analytically cumbersome. To remedy this, the following section relaxes this result into an empirical Bernstein inequality. This provides a much cleaner, closed-form confidence bound, whose theoretical sharpness relative to the oracle setting is more readily analyzed.

5.4 The empirical Bernstein inequality

We proposed a novel empirical Bennett inequality in Section 5.3. We shall remind the reader that Bennett's inequality is sharper than Bernstein's inequality for bounded random variables; however, the Bennett function h cannot be inverted in closed-form to yield a confidence interval. Hence, there is a widespread interest in obtaining Bernstein-type closed-form confidence intervals that are readily actionable, which we present here.

To construct a $(1-\delta)$-confidence set, we seek a radius r such that our Bennett bound achieves $\Pi_{2c,\delta_2,\delta_3}(r) \leq \delta - \delta_2 - \delta_3$. By employing the standard quadratic lower bound for h detailed in Lemma C.2, we can analytically invert the rate function. This inversion immediately yields the following closed-form empirical Bernstein inequality, whose proof is detailed in Appendix B.6.

Theorem 5.5 (Operator-valued empirical Bernstein inequality)**.** *Let $X_1, \ldots, X_n \in \mathcal{S}_2^s(\mathbb{H})$ attain Assumption 5.1. Define*

$$R_n(\delta_1, \delta_2, \delta_3) := \sigma_u(\delta_3)\sqrt{\frac{2}{n}\log\left(\frac{2}{\delta_1}\left[\frac{\tau_u(\delta_2)}{\sigma_u^2(\delta_3)} \vee 1\right]\right)} + \frac{2c}{3n}\log\left(\frac{2}{\delta_1}\left[\frac{\tau_u(\delta_2)}{\sigma_u^2(\delta_3)} \vee 1\right]\right), \tag{16}$$

where τ_u, τ_u', and σ_u are respectively defined in (13)*,* (14) *and* (15)*. If $\delta_1 + \delta_2 + \delta_3 \leq \delta$, then*

$$\{m \in \mathcal{S}_2^s(\mathbb{H}) : \|S_n/n - m\| \leq R_n(\delta_1, \delta_2, \delta_3)\}$$

is a $(1-\delta)$-confidence set for μ.

In the literature, an empirical Bernstein inequality with a random radius R_n is considered *asymptotically sharp* with respect to a deterministic oracle radius r_n if the scaled empirical radius converges almost surely to the oracle limit, i.e., $\sqrt{n}R_n \to \sqrt{n}r_n$ almost surely. For instance, in the matrix setting, Wang and Ramdas [43] established an empirical inequality that is sharp relative to the dimension-dependent oracle Bernstein bound of Tropp [37].

Instead, Theorem 5.5 can achieve sharpness with respect to the strictly tighter, intrinsic dimension-dependent oracle bound given in (2). Because our oracle bound is superior to that of Tropp [37] in the non-isotropic regime, our empirical counterpart seamlessly inherits this strict improvement. Specifically, by carefully scaling the error probabilities δ_1, δ_2, and δ_3 as functions of the sample size n, we guarantee that our empirical radius recovers the exact asymptotic behavior of the oracle intrinsic bound. This sharpness property is formalized below, with the proof deferred to Appendix B.7.

Theorem 5.6 (Sharpness)**.** *Let $X_1, \ldots, X_n \in \mathcal{S}_2^s(\mathbb{H})$ attain Assumption 5.1 with $\|\Sigma\| > 0$, and let R_n be defined as in* (16)*. If $\delta_{1,n} := (n-2)\delta/n$, $\delta_{2,n} := \delta/n$, and $\delta_{3,n} := \delta/n$, then*

$$\sqrt{n}R_n(\delta_{1,n}, \delta_{2,n}, \delta_{3,n}) \overset{a.s.}{\to} \sqrt{2\|\Sigma\|\log\left(\frac{2}{\delta}\frac{tr(\Sigma)}{\|\Sigma\|}\right)}.$$

6 Applications

Operator estimation connects our bounds to ML: a radius for $\|\widehat{A} - A\|$ transfers to any learning procedure stable under operator-norm perturbations, including spectral representations and regularized transformations. We illustrate this with covariance estimation and kernel PCA; Appendix D provides experimental details and two additional tests of intrinsic-dimension adaptation.

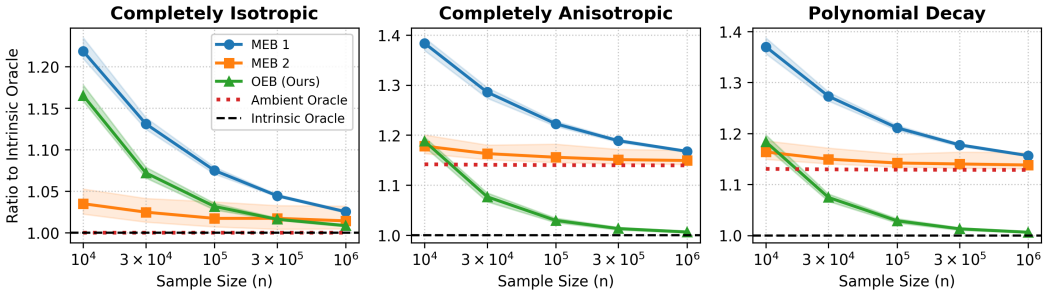


Figure 1: Ratio of empirical confidence radii to the intrinsic-dimension oracle (Equation (2)) across three spectral profiles ($d = 3, \delta = 0.05$). Solid lines represent the mean ratio over 50 independent trials, while the shaded regions indicate central 95% empirical intervals across trials. The dashed line represents the tight intrinsic oracle, and the dotted red line represents the standard ambient oracle. For non-isotropic spectra (anisotropic and polynomial decay), our operator empirical Bernstein (OEB) bound systematically outperforms both matrix empirical Bernstein (MEB) for sufficiently large sample sizes, successfully breaking the ambient dimension barrier to converge to the intrinsic oracle.

Covariance matrices. To illustrate the practical advantages of adapting to the effective rank, we first evaluate our operator-valued empirical Bernstein (OEB) bound in the context of covariance matrix estimation, comparing it directly against the dimension-dependent matrix empirical Bernstein inequalities (MEB 1 and MEB 2) derived by Wang and Ramdas [43]. We adopt their experimental setting, simulating sums of commuting (3×3)-dimensional random matrices where the eigenvalues are drawn from uniform distributions $\mathcal{U}[0, a_i]$ for $i \in [3]$. By modulating the decay of the spectrum a_i, we evaluate the bounds under three distinct geometric regimes: completely isotropic, completely anisotropic, and polynomial decay. Figure 1 displays the ratio of the empirical radii to the theoretical intrinsic-dimension oracle across increasing sample sizes n. Crucially, the simulations demonstrate that our bound successfully exploits the effective rank of the variance for non-isotropic data. As the sample size becomes sufficiently large, our bound strictly outperforms MEB 1 and MEB 2, crossing the ambient oracle barrier and converging to the intrinsic oracle.

Kernel principal component analysis (PCA). Our bounds allow for operators in infinite-dimensional Hilbert spaces, a setting where standard ambient bounds [43] break down. Consider the foundational problem of kernel PCA, where data is implicitly mapped to a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ via a continuous, bounded positive definite kernel $k(\cdot, \cdot)$. Assuming a shift-invariant kernel where $\sup_x k(x, x) \leq 1$, the rank-one empirical covariance updates $Y_i = \phi(X_i) \otimes \phi(X_i)$ are strictly bounded in both operator norm and trace ($\|Y_i\|_{\text{op}} = \text{tr}(Y_i) = 1$), and define $C = \mathbb{E}Y_i$ as well as $\widehat{C} = n^{-1} \sum_i Y_i$. Theorem 5.5 gives an observable R_n with $\|\widehat{C} - C\| \leq R_n$ with high probability. Importantly, in this setting, the variance operators are trace-class, and the intrinsic dimension is finite and governed purely by the spectral decay of the kernel. In Figure 2, we empirically track the ratio of our confidence radius against the intrinsic-dimension oracle within the infinite-dimensional RKHS induced by the Gaussian RBF kernel. Across all distributions, our bound yields a finite geometric radius that converges toward the optimal oracle limit.

Crucially, via standard spectral perturbation results such as Weyl's inequality and the Davis-Kahan $\sin\Theta$ theorem, bounding the previous operator norm directly yields finite-sample, computable confidence sets for the eigenvalues and eigenfunctions extracted during kernel PCA. The stability of the learned principal subspace depends on the population eigengap $g_k = \lambda_k(C) - \lambda_{k+1}(C)$. Specifically, when $g_k > 0$, the Davis–Kahan theorem yields an error of order R_n/g_k, so smaller gaps require larger sample sizes for an informative guarantee. The eigengap itself also admits a fully data-driven lower confidence bound: on the event $\|\widehat{C} - C\| \leq R_n$, Weyl's inequality gives $|\widehat{\lambda}_j - \lambda_j(C)| \leq R_n$, where $\widehat{\lambda}_j = \lambda_j(\widehat{C})$, and consequently $g_k \geq \widehat{\lambda}_k - \widehat{\lambda}_{k+1} - 2R_n$. Thus, an empirical gap exceeding $2R_n$ establishes a positive population gap and provides an observable lower bound that can be substituted into the Davis–Kahan guarantee. In the absence of an eigengap, individual directions within a repeated eigenspace are not identifiable. Nevertheless, the eigenvalue confidence intervals remain valid, and Davis–Kahan still controls the invariant subspace associated with a cluster separated from the remaining spectrum, without requiring separation within the cluster.

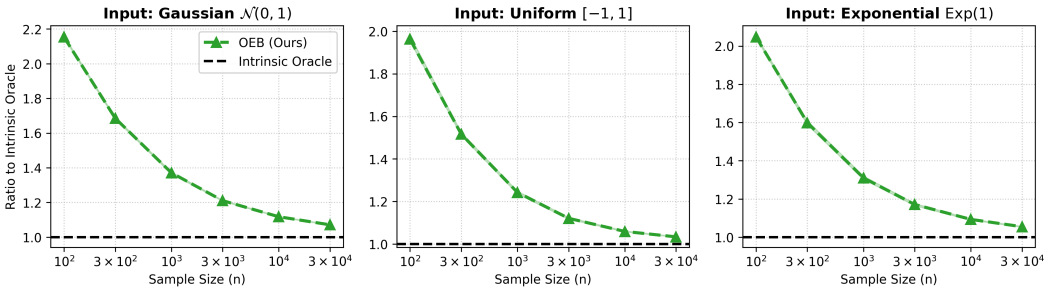


Figure 2: Ratio of our fully empirical operator empirical Bernstein (OEB) radius to a Monte Carlo estimate of the intrinsic oracle from Equation (2) in the infinite-dimensional RKHS induced by the Gaussian RBF kernel. Existing ambient bounds explicitly evaluate $\log(d) = \log(\infty)$ and are therefore omitted as they yield vacuous (infinite) radii. Across all distributions, our intrinsic radius remains strictly finite, computable, and approaches the estimated oracle threshold despite the infinite dimension of the feature space.

Moreover, kernel PCA represents each feature vector by its orthogonal projection onto a k-dimensional subspace. The corresponding rank-k projector retains the component in that subspace and discards the orthogonal remainder, whose expected squared norm is the reconstruction error. Writing $\widehat{P}_k$ for the projector onto the leading k empirical eigenfunctions and P_k^* for a projector onto a population subspace minimizing this error, our bound gives $\mathrm{tr}(P_k^* C) - \mathrm{tr}(\widehat{P}_k C) \leq 2kR_n$ on the same event. Since these traces measure the expected squared norm retained by the respective projections, their difference is precisely the excess reconstruction error of the learned representation. Thus, kernel PCA achieves reconstruction error within $2kR_n$ of the best k-dimensional representation, even when the principal directions themselves are not identifiable.

Additional adaptation tests. For noncommuting rank-one updates in $d = 1000$, OEB is smaller than both ambient MEB baselines at effective ranks $3, 5, 10, 20$; at rank 10, the reductions are 27.6% and 17.5%. For finite-rank truncations of a bounded Rademacher-feature RKHS, OEB changes by only 0.6% from rank 32 to 256, while MEB 1 and MEB 2 increase by 15.0% and 13.9%. MEB 2 is tighter at ranks 32 and 64, illustrating the absence of uniform finite-sample dominance. Tables 1 and 2 report means and Monte Carlo standard errors.

7 Conclusion

In this paper, we introduced the first empirical Bennett and Bernstein inequalities designed for self-adjoint Hilbert-Schmidt operators, including infinite-dimensional settings. By moving away from the ambient dimension and relying strictly on the intrinsic dimension, our bounds yield computable guarantees with a sharper first-order radius for non-isotropic random matrices that asymptotically match the tightest known oracle rates. As an added byproduct, we also developed novel empirical concentration bounds for the intrinsic dimension itself, offering fully data-driven guarantees for high-dimensional and functional settings.

This work is part of a larger push to replace ambient dimension dependencies with intrinsic ones across modern probability. A compelling avenue for future research is investigating whether the techniques developed here can be generalized to capture the underlying degree of non-commutativity within random matrices [38, 39, 4]. Additionally, relaxing independence to accommodate dependent structures, such as operator-valued martingales, represents a natural next step.

References

[1] Ahlswede, R. and Winter, A. (2002). Strong converse for identification via quantum channels. *IEEE Transactions on Information Theory*, 48(3):569–579.

[2] Araki, H. (1975). Relative entropy of states of von Neumann algebras. *Publications of the Research Institute for Mathematical Sciences*, 11(3):809–833.

[3] Audibert, J.-Y., Munos, R., and Szepesvári, C. (2009). Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902.

[4] Bandeira, A. S., Boedihardjo, M. T., and Van Handel, R. (2023). Matrix concentration inequalities and free probability. *Inventiones mathematicae*, 234(1):419–487.

[5] Bottou, L., Curtis, F. E., and Nocedal, J. (2018). Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311.

[6] Chung, F. R. (1997). *Spectral Graph Theory*, volume 92. American Mathematical Society.

[7] Conway, J. B. (2019). *A Course in Functional Analysis*, volume 96. Springer.

[8] Dauphin, Y. N., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., and Bengio, Y. (2014). Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. *Advances in neural information processing systems*, 27.

[9] Flammia, S. T., Gross, D., Liu, Y.-K., and Eisert, J. (2012). Quantum tomography via compressed sensing: error bounds, sample complexity and efficient estimators. *New Journal of Physics*, 14:095022.

[10] Guta, M., Kahn, J., Kueng, R., and Tropp, J. A. (2020). Fast state tomography with optimal error bounds. *Journal of Physics A: Mathematical and Theoretical*, 53:204001.

[11] Hansen, F. and Pedersen, G. K. (2003). Jensen's operator inequality. *Bulletin of the London Mathematical Society*, 35(4):553–564.

[12] Holevo, A. S. (2011). *Probabilistic and Statistical Aspects of Quantum Theory*, volume 1. Springer Science & Business Media.

[13] Howard, S. R., Ramdas, A., McAuliffe, J., and Sekhon, J. (2021). Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080.

[14] Hsu, D., Kakadey, S. M., and Zhang, T. (2012). Tail inequalities for sums of random matrices that depend on the intrinsic dimension. *Electronic Communications in Probability*, 17:1–13.

[15] Koltchinskii, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*, volume 2033. Springer Science & Business Media.

[16] Lieb, E. H. (1973). Convex trace functions and the Wigner-Yanase-Dyson conjecture. *Les rencontres physiciens-mathématiciens de Strasbourg-RCP25*, 19:0–35.

[17] Lounici, K. (2014). High-dimensional covariance matrix estimation with missing observations. *Bernoulli*, 20(3):1029–1058.

[18] 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*, 42(3):906–945.

[19] Magen, A. and Zouzias, A. (2011). Low rank matrix-valued Chernoff bounds and approximate matrix multiplication. In *Proceedings of the twenty-second annual ACM-SIAM symposium on Discrete Algorithms*, pages 1422–1436. SIAM.

[20] Martinez-Taboada, D., Gonzalez, T., and Ramdas, A. (2026). Vector-valued self-normalized concentration inequalities beyond sub-Gaussianity. In *Algorithmic Learning Theory*. PMLR.

[21] Martinez-Taboada, D. and Ramdas, A. (2025). Sharp empirical Bernstein bounds for the variance of bounded random variables. *arXiv preprint arXiv:2505.01987*.

[22] Martinez-Taboada, D. and Ramdas, A. (2026a). Empirical Bernstein in smooth Banach spaces. *Annals of Applied Probability*.

[23] Martinez-Taboada, D. and Ramdas, A. (2026b). Intrinsic dimension concentration inequalities for self-adjoint operators. *arXiv preprint arXiv:2602.13465*.

[24] Mas, A. and Ruymgaart, F. (2014). High-dimensional principal projections. *Complex Analysis and Operator Theory*, 9:35–63.

[25] Maurer, A. and Pontil, M. (2009). Empirical Bernstein bounds and sample variance penalization. *Conference on Learning Theory*.

[26] Mhammedi, Z., Grünwald, P., and Guedj, B. (2019). PAC-Bayes un-expected Bernstein inequality. *Advances in Neural Information Processing Systems*, 32.

[27] Minsker, S. (2017). On some extensions of Bernstein's inequality for self-adjoint operators. *Statistics & Probability Letters*, 127:111–119.

[28] Mnih, V., Szepesvári, C., and Audibert, J.-Y. (2008). Empirical Bernstein stopping. In *Proceedings of the 25th international conference on Machine learning*, pages 672–679.

[29] Nielsen, M. A. and Chuang, I. L. (2010). *Quantum Computation and Quantum Information*. Cambridge University Press.

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

[31] Oliveira, R. I. (2010). Sums of random Hermitian matrices and an inequality by Rudelson. *Electronic Communications in Probability*, 15:203–212.

[32] Orabona, F. and Jun, K.-S. (2023). Tight concentrations and confidence sequences from the regret of universal portfolio. *IEEE Transactions on Information Theory*, 70(1):436–455.

[33] Reed, M. and Simon, B. (1980). *Methods of Modern Mathematical Physics I: Functional Analysis*. Academic Press.

[34] Reiß, M. and Wahl, M. (2020). Nonasymptotic upper bounds for the reconstruction error of PCA. *The Annals of Statistics*, 48.

[35] Spielman, D. A. (2012). Spectral graph theory and its applications. *Foundations of computer science (FOCS)*.

[36] Tropp, J. A. (2011). Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16:262–270.

[37] Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434.

[38] Tropp, J. A. (2016). The expected norm of a sum of independent random matrices: An elementary approach. In *High Dimensional Probability VII: The Cargese Volume*, pages 173–202. Springer.

[39] Tropp, J. A. (2018). Second-order matrix concentration inequalities. *Applied and Computational Harmonic Analysis*, 44(3):700–736.

[40] Tropp, J. A. et al. (2015). An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230.

[41] Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge University Press.

[42] Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48. Cambridge University Press.

[43] Wang, H. and Ramdas, A. (2024). Sharp matrix empirical Bernstein inequalities. *arXiv preprint arXiv:2411.09516*.

[44] Waudby-Smith, I. and Ramdas, A. (2024). Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27.

[45] Zhivotovskiy, N. (2024). Dimension-free bounds for sums of independent matrices and simple tensors via the variational principle. *Electronic Journal of Probability*, 29:1–28.

A Concentration for the intrinsic dimension

Of independent interest is the study of the intrinsic dimension $\mathrm{tr}(\Sigma)/\|\Sigma\|$. This quantity fundamentally characterizes the complexity of random operators, acting as a continuous, stable analogue to the discrete algebraic rank. The theoretical tools developed in this contribution naturally yield concentration guarantees for the intrinsic dimension as a byproduct. In previous sections, we considered the ratio τ_u/σ_u^2, where τ_u and σ_u^2 are high-probability upper bounds for $\mathrm{tr}(\Sigma)$ and $\|\Sigma\|$. However, this cross-ratio establishes neither a valid upper nor a valid lower bound for the intrinsic dimension.

Instead, we must couple opposing bounds, considering $\tau_l(\delta_2)/\sigma_u^2(\delta_3)$ and $\tau_u(\delta_2)/\sigma_l^2(\delta_3)$, where τ_l and σ_l satisfy $\mathbb{P}\left(\mathrm{tr}(V_n) \le n\tau_l(\delta_2)\right) \le \delta_2$ and $\mathbb{P}\left(\|V_n\| \le n\sigma_l^2(\delta_3)\right) \le \delta_3$. These bounds can be immediately obtained from the frameworks we have already derived. In particular, flipping the empirical Bernstein bound in Theorem 3.5 yields the trace lower bound

$$\tau_l(\delta) := \mathrm{tr}(\Sigma_n) - \varsigma_{n/2}(\mathbf{Z})\sqrt{\frac{2\log(2/\delta)}{n/2}} - \frac{7(2B)\log(2/\delta)}{3(n/2-1)}, \quad Z_i = \mathrm{tr}(X_i'), \tag{17}$$

and modifying Corollary 5.3 produces the lower bound for the operator norm:

$$\sigma_l^2(\delta_3) := \sigma_l^2\left(\tau_u'\left(\frac{\delta_3}{2}\right); \frac{\delta_3}{2}\right), \quad \sigma_l^2(\tau;\delta) := \max\left(0, \|S'_{n/2}/(n/2)\| - \Gamma_{c,\delta}(\tau)\right). \tag{18}$$

Combining these guarantees, we naturally obtain fully empirical confidence intervals for the intrinsic dimension. We formalize this in the following theorem, deferring its proof to Appendix B.8.

Theorem A.1. *Let $X_1, \dots, X_n \in \mathcal{S}_2^s(\mathbb{H})$ attain Assumption 5.1. Assume $\|\Sigma\| > 0$ and interpret a zero lower-norm denominator as an infinite upper bound. If τ_l, τ_u, σ_l, and σ_u are defined as in* (17)*,* (13)*,* (18)*, and* (15)*, then*

$$\mathbb{P}\left(\frac{tr(V_n)}{\|V_n\|} \ge \frac{\tau_u(\delta_2)}{\sigma_l^2(\delta_3)}\right) \le \delta_2 + \delta_3, \quad \mathbb{P}\left(\frac{tr(V_n)}{\|V_n\|} \le \frac{\tau_l(\delta_2)}{\sigma_u^2(\delta_3)}\right) \le \delta_2 + \delta_3. \tag{19}$$

Moreover, the interval $[\tau_l(\delta_2)/\sigma_u^2(\delta_3),\ \tau_u(\delta_2)/\sigma_l^2(\delta_3)]$ contains $tr(\Sigma)/\|\Sigma\|$ with probability at least $1 - 2\delta_2 - \delta_3$. Thus $2\delta_2 + \delta_3 \le \delta$ gives a $(1-\delta)$-confidence interval.

While Theorem A.1 establishes rigorous, finite-sample confidence sets, it is instructive to examine their asymptotic behavior. Specifically, we demonstrate that the width of these confidence sets around the empirical intrinsic dimension estimator $\mathrm{tr}(\Sigma_n)/\|\Sigma_n\|$ shrinks precisely at the parametric rate $1/\sqrt{n}$. The following theorem, proved in Appendix B.9, establishes this almost sure convergence.

Theorem A.2 (Asymptotic width)**.** *Let $X_1, \dots, X_n \in \mathcal{S}_2^s(\mathbb{H})$ attain Assumption 5.1. If the observations are i.i.d., $\|\Sigma\| > 0$, and τ_l, τ_u, σ_l, and σ_u are defined as in* (17)*,* (13)*,* (18)*, and* (15)*, then there exists $K \ge 0$ such that*

$$\sqrt{n}\left(\frac{\tau_u(\delta_2)}{\sigma_l^2(\delta_3)} - \frac{tr(\Sigma_n)}{\|\Sigma_n\|}\right) \xrightarrow{a.s.} K, \quad \sqrt{n}\left(\frac{tr(\Sigma_n)}{\|\Sigma_n\|} - \frac{\tau_l(\delta_2)}{\sigma_u^2(\delta_3)}\right) \xrightarrow{a.s.} K. \tag{20}$$

The above result means that (19) yields fully empirical confidence intervals for the intrinsic dimension of width $O(1/\sqrt{n})$, which are the first of their kind to the best of our knowledge.

B Proofs

B.1 Proof of Theorem 4.1

The proof combines a bound on $\mathbb{E}\mathrm{tr}\varphi(\theta S_n)$, where $\varphi(u) = \cosh(u) - 1$, with Doob's maximal inequality, before substituting the empirical variance bounds. The trace MGF bound and submartingale property in the infinite-dimensional setting are established in Appendix E.

The trace estimate (40) and the scalar spectral bound give, for any $n\sigma^2 \ge \|V_n\|$ with $\sigma^2 > 0$,

$$\begin{aligned}\mathbb{E}\mathrm{tr}\varphi(\theta S_n) &\le \mathrm{tr}\left[e^{\psi(|\theta|)V_n} - I\right] \\ &\le \left[e^{n\psi(|\theta|)\sigma^2} - 1\right]\frac{\mathrm{tr}(V_n)}{n\sigma^2} =: g(\sigma^2).\end{aligned} \tag{21}$$

The function g is increasing on σ^2 (this can be easily observed by considering the Taylor expansion of the $\exp$ function).

Applying Doob's maximal inequality as in Martinez-Taboada and Ramdas [23], we obtain, for $\theta \in [0, \theta_{\max})$,

$$\mathbb{P}\left(\sup_{i\le n} \|S_i\| \ge r\right) \le 2\frac{\mathbb{E}\mathrm{tr}(\varphi(\theta S_n)) + d'}{\exp(\theta r)}, \tag{22}$$

where d' is any number greater or equal than 1. Combining (21) and (22), we obtain

$$\mathbb{P}\left(\sup_{i\le n} \|S_i\| \ge r\right) \le \inf_{\sigma^2 \ge \|V_n\|/n, d' \ge 1} h(\sigma^2, d'), \quad h(\sigma^2, d') := 2 \inf_{\theta\in[0,\theta_{\max})} \frac{g(\sigma^2) + d'}{\exp(\theta r)}.$$

Fix $\sigma^2 = \|V_n\|/n > 0$ (the zero case is handled below). We use the shorthand $\tau_u \equiv \tau_u(\delta_2)$ and $\sigma_u \equiv \sigma_u(\delta_3)$. The function h is increasing both on σ^2 (as g is increasing) and d'. So, under the event $\left\{n\sigma_u^2 \ge \|V_n\|\right\}$,

$$\mathbb{P}\left(\sup_{i\le n} \|S_i\| \ge r\right) \le h\left(\sigma^2, 1\right) \le h\left(\sigma^2, \frac{\tau_u}{\sigma_u^2} \vee 1\right) \le h\left(\sigma_u^2, \frac{\tau_u}{\sigma_u^2} \vee 1\right), \tag{23}$$

where the second inequality holds deterministically as $(\tau_u/\sigma_u^2) \vee 1 \ge 1$. It remains to observe that, under the event $\{n\tau_u \ge \mathrm{tr}(V_n)\}$,

$$\begin{aligned}
h\left(\sigma_u^2, \frac{\tau_u}{\sigma_u^2} \vee 1\right) &= 2 \inf_{\theta\in[0,\theta_{\max})} \frac{g(\sigma_u^2) + \frac{\tau_u}{\sigma_u^2} \vee 1}{\exp(\theta r)} \\
&= 2 \inf_{\theta\in[0,\theta_{\max})} \frac{\left[\exp\left(n\psi(\theta)\sigma_u^2\right) - 1\right] \frac{\mathrm{tr}(V_n)}{n\sigma_u^2} + \frac{\tau_u}{\sigma_u^2} \vee 1}{\exp(\theta r)} \\
&\le 2 \inf_{\theta\in[0,\theta_{\max})} \frac{\left[\exp\left(n\psi(\theta)\sigma_u^2\right) - 1\right] \frac{\tau_u}{\sigma_u^2} + \frac{\tau_u}{\sigma_u^2} \vee 1}{\exp(\theta r)} \\
&\le 2 \inf_{\theta\in[0,\theta_{\max})} \frac{\left[\exp\left(n\psi(\theta)\sigma_u^2\right) - 1\right] \left(\frac{\tau_u}{\sigma_u^2} \vee 1\right) + \frac{\tau_u}{\sigma_u^2} \vee 1}{\exp(\theta r)} \\
&\le 2\left(\frac{\tau_u}{\sigma_u^2} \vee 1\right) \inf_{\theta\in[0,\theta_{\max})} \frac{\exp\left(n\psi(\theta)\sigma_u^2\right)}{\exp(\theta r)}.
\end{aligned} \tag{24}$$

Combining (23) and (24), we obtain that, under the event $\{\sigma_u \ge \sigma\} \cap \{n\tau_u \ge \mathrm{tr}(V_n)\}$,

$$\mathbb{P}\left(\sup_{i\le n} \|S_i\| \ge r\right) \le 2\left(\frac{\tau_u}{\sigma_u^2} \vee 1\right) \inf_{\theta\in[0,\theta_{\max})} \frac{\exp\left(n\psi(\theta)\sigma_u^2\right)}{\exp(\theta r)}. \tag{25}$$

To complete the proof, write $h(\sigma^2, 1; r)$ to make its dependence on r explicit. Inequalities (23) and (24) show that this deterministic tail bound is at most $\Pi_{\psi,\delta_2,\delta_3}(r)$ on $E = \{\|V_n\| \le n\sigma_u^2\} \cap \{\mathrm{tr}(V_n) \le n\tau_u\}$. Thus the empirical inverse threshold is at least the deterministic δ_1 threshold of $h(\sigma^2, 1; \cdot)$ on E. Exceedance of the former is contained in exceedance of the latter or E^c. A union bound with (5) proves (6). If $V_n = 0$, the MGF hypothesis implies $X_i = 0$ almost surely and the conclusion is immediate for positive thresholds.

□

B.2 Proof of Corollary 4.2

Define

$$\psi(\theta) = \psi_{P,c}(\theta) := \frac{e^{\theta c} - \theta c - 1}{c^2}.$$

For $\theta \in \mathbb{R}$,

$$\mathbb{E}\exp(\theta X_i) = \mathbb{E}\left[\sum_{k\ge 0} \frac{(\theta X_i)^k}{k!}\right] = I + \mathbb{E}\left[\sum_{k\ge 2} \frac{(\theta X_i)^k}{k!}\right] = I + \sum_{k\ge 2} \mathbb{E}\left[(\theta X_i)^2 \frac{(\theta X_i)^{k-2}}{k!}\right]$$

$$\preceq I + \sum_{k\geq 2} \mathbb{E}\left[(\theta X_i)^2 \frac{(|\theta|\|X_i\|)^{k-2}}{k!}\right] \preceq I + \sum_{k\geq 2} \mathbb{E}\left[(\theta X_i)^2 \frac{(|\theta|c)^{k-2}}{k!}\right]$$
$$= I + \frac{\mathbb{E}\left(X_i^2\right)}{c^2} \sum_{k\geq 2} \frac{(|\theta|c)^k}{k!} = I + \mathbb{E}\left(X_i^2\right) \psi_{P,c}(|\theta|) \preceq \exp\left(\mathbb{E}\left(X_i^2\right) \psi_{P,c}(|\theta|)\right),$$

where the last inequality follows from $1 + x \leq \exp(x)$ for all $x \in \mathbb{R}$. Given the monotonicity of the operator $\log$ function, we obtain that $\log \mathbb{E}\exp(\theta X_i) \preceq \mathbb{E}\left(X_i^2\right) \psi_{P,c}(|\theta|)$. It now suffices to apply Theorem 4.1 with $\psi(\theta) = \psi_{P,c}(\theta)$ and $\theta_{\max} = \infty$, and take the minimizer

$$\theta^* = \frac{1}{c} \log\left(1 + \frac{cr}{n\sigma_u(\delta_3)^2}\right).$$

□

B.3 Proof of Theorem 5.2

We follow the same trace MGF and maximal-inequality argument as in Appendix B.1, but bound the centered exponential using only $\text{tr}(V_n')$. The finite-rank reduction in Appendix E ensures that the centered traces below are well-defined and integrable, and that the submartingale property extends to the full space.

Define

$$\varphi(u) := \cosh(u) - 1 = \frac{\exp(u) + \exp(-u)}{2} - 1$$

for $u \in \mathbb{R}$. For $\theta > 0$,

$$\begin{aligned}
\mathbb{P}\left(\sup_{i\leq n} \|S_i'\| \geq r\right) &\overset{(i)}{=} \mathbb{P}\left(\sup_{i\leq n} \|\theta S_i'\| \geq \theta r\right) \overset{(ii)}{=} \mathbb{P}\left(\sup_{i\leq n} \varphi(\|\theta S_i'\|) \geq \varphi(\theta r)\right) \\
&\overset{(iii)}{=} \mathbb{P}\left(\sup_{i\leq n} \|\varphi(\theta S_i')\| \geq \varphi(\theta r)\right) \overset{(iv)}{\leq} \mathbb{P}\left(\sup_{i\leq n} \text{tr}(\varphi(\theta S_i')) \geq \varphi(\theta r)\right) \\
&= \mathbb{P}\left(\sup_{i\leq n} \text{tr}(\varphi(\theta S_i')) + 1 \geq \varphi(\theta r) + 1\right) \overset{(v)}{\leq} \frac{\mathbb{E}\text{tr}(\varphi(\theta S_n')) + 1}{\varphi(\theta r) + 1}, \qquad (26)
\end{aligned}$$

where (i) follows from $\theta > 0$, (ii) follows from φ being strictly increasing on $\mathbb{R}_{>0}$, (iii) follows from φ being symmetric and increasing on $\mathbb{R}_{>0}$, (iv) follows from the fact that φ is nonnegative, and (v) follows from Lemma 3.3.

We can further upper bound $\mathbb{E}\text{tr}\varphi(\theta S_n')$, for $\theta \in (-\theta_{\max}, \theta_{\max})$, as

$$\begin{aligned}
\mathbb{E}\text{tr}\varphi(\theta S_n') &= \text{tr}\mathbb{E}\varphi(\theta S_n') \\
&= \mathbb{E}\text{tr}\left(\frac{\exp\left(\sum_{i=1}^n \log\exp(\theta X_i')\right) + \exp\left(\sum_{i=1}^n \log\exp(-\theta X_i')\right)}{2} - I\right) \\
&\leq \text{tr}\left(\frac{\exp\left(\sum_{i=1}^n \log\mathbb{E}\exp(\theta X_i')\right) + \exp\left(\sum_{i=1}^n \log\mathbb{E}\exp(-\theta X_i')\right)}{2} - I\right) \\
&\leq \text{tr}\left(\exp\left(\sum_{i=1}^n \psi(|\theta|)\Delta V_i'\right) - I\right) = \text{tr}\left(\exp\left(\psi(|\theta|)V_n'\right) - I\right), \qquad (27)
\end{aligned}$$

where the first equality follows from $\varphi \geq 0$ and Tonelli's theorem. For the first inequality, we apply the finite-rank argument of step 3 in Appendix E to the primed sequence. The second inequality then applies (10) to the original operators.

Lastly, given that $\|V_n'\| \leq \text{tr}(V_n')$ as $V_n' \succeq 0$, we observe that

$$\text{tr}\left[\exp\left(\psi(|\theta|)V_n'\right) - I\right] = \text{tr}\left[\sum_{k\geq 1} \frac{\left(\psi(|\theta|)V_n'\right)^k}{k!}\right]$$

$$
\begin{aligned}
&= \mathrm{tr}\left[\psi(|\theta|)\left(V_n'\right)^{\frac{1}{2}}\left\{\sum_{k\geq 0}\frac{\left(\psi(|\theta|)V_n'\right)^k}{(k+1)!}\right\}\left(V_n'\right)^{\frac{1}{2}}\right] \\
&\leq \mathrm{tr}\left[\psi(|\theta|)\left\|\sum_{k\geq 0}\frac{\left(\psi(|\theta|)V_n'\right)^k}{(k+1)!}\right\| V_n'\right] \\
&\leq \psi(|\theta|)\left[\sum_{k\geq 0}\frac{\left(\psi(|\theta|)\left\|V_n'\right\|\right)^k}{(k+1)!}\right]\mathrm{tr}\left(V_n'\right) \\
&\leq \psi(|\theta|)\left[\sum_{k\geq 0}\frac{\left(\psi(|\theta|)\mathrm{tr}\left(V_n'\right)\right)^k}{(k+1)!}\right]\mathrm{tr}\left(V_n'\right) \\
&= \sum_{k\geq 1}\frac{\left(\psi(|\theta|)\mathrm{tr}\left(V_n'\right)\right)^k}{k!} = \exp\left(\psi(|\theta|)\mathrm{tr}\left(V_n'\right)\right) - 1. \qquad (28)
\end{aligned}
$$

Combining (26), (27), and (28), we obtain

$$
\mathbb{P}\left(\sup_{i\leq n}\|S_i'\| \geq r\right) \leq \frac{\exp\left(\psi(|\theta|)\mathrm{tr}\left(V_n'\right)\right) - 1 + 1}{\varphi(\theta r)+1} = \frac{\exp\left(\psi(\theta)\mathrm{tr}\left(V_n'\right)\right)}{\varphi(\theta r)+1}, \quad \theta \in (0, \theta_{\max}).
$$

Based on $\varphi(\theta r) + 1 = \cosh(\theta r) \geq \frac{\exp(\theta r)}{2}$, we conclude

$$
\mathbb{P}\left(\sup_{i\leq n}\|S_i'\| \geq r\right) \leq \frac{\exp\left(\psi(\theta)\mathrm{tr}\left(V_n'\right)\right)}{\varphi(\theta r)+1} \leq 2\frac{\exp\left(\psi(\theta)\mathrm{tr}\left(V_n'\right)\right)}{\exp(\theta r)}, \quad \theta \in (0, \theta_{\max})
$$

Inequality (11) also holds trivially for $\theta = 0$.

□

B.4 Proof of Corollary 5.3

Put $Y_i = X_i' - \Sigma$, $N = n/2$, and $V_N' = \sum_{i\leq N}\mathbb{E}Y_i^2 = N\widetilde{\Sigma}'$. These independent mean-zero increments satisfy $\|Y_i\| \leq 2c^2$. The argument in Appendix B.2 gives the CGF envelope $\psi_{P,2c^2}$ with pseudo-variance increments $\mathbb{E}Y_i^2$. Theorem 5.2 therefore yields

$$
\mathbb{P}\left(\sup_{j\leq N}\|S_j' - j\Sigma\| \geq r\right) \leq 2\exp\left(-\frac{\mathrm{tr}(V_N')}{(2c^2)^2}h\left(\frac{2c^2 r}{\mathrm{tr}(V_N')}\right)\right), \qquad (29)
$$

by optimizing at $\theta^* = (2c^2)^{-1}\log(1 + 2c^2 r/\mathrm{tr}(V_N'))$ (the zero-variance case is immediate).

Applying the Bennett inversion from Lemma C.2 to (29) with $V_N' = N\widetilde{\Sigma}'$ leads to

$$
\|2S_{n/2}'/n - \Sigma\| \leq \Gamma_{c,\delta}(\mathrm{tr}(\widetilde{\Sigma}'))
$$

with probability $1-\delta$. Given that $\Gamma_{c,\delta}$ is increasing, for $\tau \geq \mathrm{tr}(\widetilde{\Sigma}')$,

$$
\|\Sigma\| - \|2S_{n/2}'/n\| \leq \|2S_{n/2}'/n - \Sigma\| \leq \Gamma_{c,\delta}(\mathrm{tr}(\widetilde{\Sigma}')) \leq \Gamma_{c,\delta}(\tau),
$$

with probability $1-\delta$.

□

B.5 Proof of Theorem 5.4

By Corollary 4.2,

$$
\mathbb{P}\left(\sup_{i\leq n}\|S_i - i\mu\| \geq \Pi^{-1}_{2c,\delta_2,\delta_3}(\delta_1)\right) \leq \delta_1 + \delta_2 + \delta_3
$$

if τ_u and σ_u are functions such that (5) holds. Note that we consider $\Pi_{2c,\delta_2,\delta_3}$ with the parameter $2c$ in view of $\|X_i - \mu\| \le \|X_i\| + \|\mu\| \le 2c$. By (13) (definition of τ_u) and Theorem 3.5,

$$\mathbb{P}\left(\mathrm{tr}(V_n) \ge n\tau_u(\delta_2)\right) \le \delta_2.$$

If $\tau \ge \mathrm{tr}(\widetilde{\Sigma}')$, by Corollary 5.3,

$$\mathbb{P}(\|V_n\| \ge n\bar{\sigma}_u^2(\tau; \delta_3/2)) \le \frac{\delta_3}{2}. \tag{30}$$

Furthermore, in view of (14) (definition of τ_u') and Theorem 3.5,

$$\mathbb{P}\left(\mathrm{tr}(\widetilde{\Sigma}') > \tau_u'(\delta_3/2)\right) \le \frac{\delta_3}{2}. \tag{31}$$

On $\{\tau_u'(\delta_3/2) \ge \mathrm{tr}(\widetilde{\Sigma}')\}$, monotonicity of $\Gamma_{c,\delta}$ permits substitution of the random trace upper bound into the deterministic error radius. Union bounding the guarantees (30) and (31), we obtain

$$\mathbb{P}(\|V_n\| \ge n\bar{\sigma}_u^2(\tau_u'(\delta_3/2); \delta_3/2)) = \mathbb{P}(\|V_n\| \ge n\sigma_u^2(\delta_3)) \le \delta_3.$$

Thus, we have proven that both τ_u and σ_u attain (5), concluding the result.

B.6 Proof of Theorem 5.5

We invoke Theorem 5.4 and solve for r such that

$$\Pi_{2c,\delta_2,\delta_3}(r) \le \delta_1.$$

In particular, the former inequality is attained if

$$2\left(\frac{\tau_u(\delta_2)}{\sigma_u^2(\delta_3)} \vee 1\right) \exp\left(-\frac{n\sigma_u^2(\delta_3)}{(2c)^2} h\left(\frac{2cr}{n\sigma_u^2(\delta_3)}\right)\right) \overset{(i)}{\le} \delta_1 \le \delta - \delta_2 - \delta_3.$$

Inverting (i) via Lemma C.2 concludes the result.

□

B.7 Proof of Theorem 5.6

The variables $Z_i = \mathrm{tr}(X_i') \in [0, 2B]$ are independent with common mean $\mathrm{tr}(\Sigma)$. The bounded-variable strong law gives $\mathrm{tr}(\Sigma_n) \to \mathrm{tr}(\Sigma)$ almost surely. Their empirical standard deviation is uniformly bounded, so the correction in (13) is $O(\sqrt{\log n/n} + \log n/n)$ and $\tau_u(\delta/n) \to \mathrm{tr}(\Sigma)$ almost surely.

Similarly, $Z_i' = \mathrm{tr}(X_i'') \in [0, 4Bc^2]$ implies $\tau_u'(\delta/(2n)) = O(1)$. No convergence to a common fourth moment is needed here. Since the independent $X_i' - \Sigma$ have uniformly bounded Hilbert-Schmidt norms, the Hilbert-space strong law gives $\|\Sigma_n - \Sigma\|_{\mathrm{HS}} \to 0$, hence $\|\Sigma_n\| \to \|\Sigma\|$. Consequently,

$$\begin{aligned}\sigma_u^2(\delta/n) &= \bar{\sigma}_u^2\left(\tau_u'\left(\frac{\delta}{2n}\right); \frac{\delta}{2n}\right) \\ &= \|\Sigma_n\| + \sqrt{\frac{4\tau_u'(\delta/(2n))\log(4n/\delta)}{n}} + \frac{4c^2\log(4n/\delta)}{3n} \xrightarrow{a.s.} \|\Sigma\|.\end{aligned}$$

It now remains to observe that

$$\begin{aligned}\sqrt{n}R_n &\equiv \sqrt{n}R_n(\delta_{1,n}, \delta_{2,n}, \delta_{3,n}) \\ &= \sqrt{2\sigma_u^2(\delta_{3,n})\log\left(\frac{2}{\delta_{1,n}}\left[\frac{\tau_u(\delta_{2,n})}{\sigma_u^2(\delta_{3,n})} \vee 1\right]\right)} + \frac{2c}{3\sqrt{n}}\log\left(\frac{2}{\delta_{1,n}}\left[\frac{\tau_u(\delta_{2,n})}{\sigma_u^2(\delta_{3,n})} \vee 1\right]\right) \\ &\overset{a.s.}{\to} \sqrt{2\|\Sigma\|\log\left(\frac{2}{\delta}\frac{\mathrm{tr}(\Sigma)}{\|\Sigma\|}\right)} + 0\end{aligned}$$

in order to conclude the result.

□

B.8 Proof of Theorem A.1

To prove this theorem, we rely on a standard union bound argument applied to the high-probability tail bounds of the trace and the operator norm established earlier. By definition of our upper and lower bound functions, we observe that

$$\mathbb{P}\left(\mathrm{tr}(V_n) \geq n\tau_u(\delta_2)\right) \leq \delta_2, \quad \mathbb{P}\left(\|V_n\| \leq n\sigma_l^2(\delta_3)\right) \leq \delta_3,$$
$$\mathbb{P}\left(\mathrm{tr}(V_n) \leq n\tau_l(\delta_2)\right) \leq \delta_2, \quad \mathbb{P}\left(\|V_n\| \geq n\sigma_u^2(\delta_3)\right) \leq \delta_3.$$

We begin by establishing the upper bound on the intrinsic dimension ratio. Consider the events $\{\mathrm{tr}(V_n) \geq n\tau_u(\delta_2)\}$ and $\{\|V_n\| \leq n\sigma_l^2(\delta_3)\}$. By the union bound, the probability that either event occurs is bounded by $\delta_2 + \delta_3$. Under the complement event, both $\mathrm{tr}(V_n) < n\tau_u(\delta_2)$ and $\|V_n\| > n\sigma_l^2(\delta_3)$ must hold simultaneously. Dividing these inequalities yields

$$\frac{\mathrm{tr}(V_n)}{\|V_n\|} < \frac{\tau_u(\delta_2)}{\sigma_l^2(\delta_3)}.$$

Consequently, the event where this ratio is violated implies that the union of the tail events must have occurred. Therefore, we conclude that the first inequality in (19) holds. Analogously, the second inequality in (19) also holds.

For the simultaneous interval, the proof of Corollary 5.3, combined with the trace bound $\tau_u'(\delta_3/2)$ as in Appendix B.5, gives $\|\Sigma_n - \bar{\Sigma}\| \leq \Gamma_{c,\delta_3/2}(\tau_u'(\delta_3/2))$ with probability at least $1 - \delta_3$. On this single event, $\sigma_l^2(\delta_3) \leq \|\Sigma\| \leq \sigma_u^2(\delta_3)$. Intersecting it with the two trace confidence events, each with failure probability at most δ_2, proves the interval coverage by a union bound.

□

B.9 Proof of Theorem A.2

We begin by isolating the deterministic error margins from our high-probability empirical bounds. Expressing the upper bound for the trace and the lower bound for the norm strictly in terms of the empirical matrix Σ_n, we have $\tau_u(\delta_2) = \mathrm{tr}(\Sigma_n) + E_{\tau,n}$ and $\sigma_l^2(\delta_3) = \max(0, \|\Sigma_n\| - E_{\sigma,n})$, where the respective error sequences are given by

$$E_{\tau,n} := \varsigma_{n/2}(\mathbf{Z})\sqrt{\frac{4\log(2/\delta_2)}{n}} + \frac{7(2B)\log(2/\delta_2)}{3(n/2-1)},$$
$$E_{\sigma,n} := \sqrt{\frac{4\tau_u'(\delta_3/2)}{n}\log(4/\delta_3)} + \frac{2(2c^2)}{3n}\log(4/\delta_3).$$

By the strong law of large numbers and the almost sure convergence of U-statistics, we obtain

$$\mathrm{tr}(\Sigma_n) \overset{a.s.}{\to} \mathrm{tr}(\Sigma), \quad \|\Sigma_n\| \overset{a.s.}{\to} \|\Sigma\|, \quad \varsigma_{n/2}^2(\mathbf{Z}) \overset{a.s.}{\to} \mathbb{V}(\mathrm{tr}(X_1')), \quad \tau_u'(\delta_3/2) \overset{a.s.}{\to} \mathrm{tr}(\widetilde{\Sigma}').$$

Scaling the error terms by $\sqrt{n}$ and taking the limit as $n \to \infty$ isolates the dominant $1/\sqrt{n}$ fluctuations

$$\lim_{n\to\infty} \sqrt{n}E_{\tau,n} = 2\sqrt{\mathbb{V}(\mathrm{tr}(X_1'))\log(2/\delta_2)} \quad \text{a.s.},$$
$$\lim_{n\to\infty} \sqrt{n}E_{\sigma,n} = 2\sqrt{\mathrm{tr}(\widetilde{\Sigma}')\log(4/\delta_3)} \quad \text{a.s.} \tag{32}$$

Almost surely, $\|\Sigma_n\| \to \|\Sigma\| > 0$ and $E_{\sigma,n} \to 0$, so $\|\Sigma_n\| > E_{\sigma,n}$ for all sufficiently large n. Hence $\sigma_l^2(\delta_3) = \|\Sigma_n\| - E_{\sigma,n} > 0$ for all such n, to which the following algebra is restricted. Any earlier undefined ratios may be assigned arbitrary values without affecting the limits.

We now analyze the asymptotic width of the upper bound relative to the unified empirical intrinsic dimension estimator $\mathrm{tr}(\Sigma_n)/\|\Sigma_n\|$; algebraic rearrangement yields

$$\frac{\tau_u(\delta_2)}{\sigma_l^2(\delta_3)} - \frac{\mathrm{tr}(\Sigma_n)}{\|\Sigma_n\|} = \frac{\mathrm{tr}(\Sigma_n) + E_{\tau,n}}{\|\Sigma_n\| - E_{\sigma,n}} - \frac{\mathrm{tr}(\Sigma_n)}{\|\Sigma_n\|} = \frac{E_{\tau,n}\|\Sigma_n\| + \mathrm{tr}(\Sigma_n)E_{\sigma,n}}{\|\Sigma_n\|(\|\Sigma_n\| - E_{\sigma,n})},$$

and multiplying both sides by $\sqrt{n}$ gives

$$\sqrt{n}\left(\frac{\tau_u(\delta_2)}{\sigma_l^2(\delta_3)} - \frac{\text{tr}(\Sigma_n)}{\|\Sigma_n\|}\right) = \frac{(\sqrt{n}E_{\tau,n})\|\Sigma_n\| + \text{tr}(\Sigma_n)(\sqrt{n}E_{\sigma,n})}{\|\Sigma_n\|(\|\Sigma_n\| - E_{\sigma,n})}.$$

As $n \to \infty$, the unscaled error term $E_{\sigma,n} \to 0$. Applying the almost sure limits of the empirical quantities alongside the scaled error limits from (32), the denominator converges to $\|\Sigma\|^2$, and the numerator converges to $2\|\Sigma\|\sqrt{\mathbb{V}(\text{tr}(X_1'))\log(2/\delta_2)} + 2\text{tr}(\Sigma)\sqrt{\text{tr}(\widetilde{\Sigma}')\log(4/\delta_3)}$. Distributing the denominator establishes the limit for the upper bound.

The derivation for the lower bound follows via a symmetric argument. Observing that $\tau_l(\delta_2) = \text{tr}(\Sigma_n) - E_{\tau,n} + O(1/n)$ and $\sigma_u^2(\delta_3) = \|\Sigma_n\| + E_{\sigma,n} + O(1/n)$, the exact same algebraic limits apply. Consequently, both bounds converge to the identical limit, allowing us to explicitly define the constant K as

$$K := 2\left(\frac{\sqrt{\mathbb{V}(\text{tr}(X_1'))\log(2/\delta_2)}}{\|\Sigma\|} + \frac{\text{tr}(\Sigma)\sqrt{\text{tr}(\widetilde{\Sigma}')\log(4/\delta_3)}}{\|\Sigma\|^2}\right). \tag{33}$$

□

C Auxiliary results

Lemma C.1. *For all $y \geq 0$, the inverse of the Bennett rate function $h(u) = (1+u)\log(1+u) - u$ satisfies*

$$h^{-1}(y) \leq \sqrt{2y} + \frac{y}{3}$$

Proof. Because $h(u)$ is strictly monotonically increasing for $u \geq 0$, applying h to both sides preserves the inequality. Therefore, proving the proposition is equivalent to proving that, for all $y \geq 0$,

$$h\left(\sqrt{2y} + \frac{y}{3}\right) \geq y.$$

To simplify the differentiation, let $x = \sqrt{2y}$. Since $y \geq 0$. Thus we need to show that, for all $x \geq 0$,

$$h\left(x + \frac{x^2}{6}\right) \geq \frac{x^2}{2}.$$

Let us define the difference function

$$f(x) = h\left(x + \frac{x^2}{6}\right) - \frac{x^2}{2}.$$

We aim to show $f(x) \geq 0$ for all $x \geq 0$. We know $f(0) = h(0) - 0 = 0$, so it suffices to show that $f'(x) \geq 0$ on $[0, \infty)$. Recalling that $h'(u) = \log(1+u)$,

$$f'(x) = h'\left(x + \frac{x^2}{6}\right)\left(1 + \frac{x}{3}\right) - x = \log\left(1 + x + \frac{x^2}{6}\right)\left(1 + \frac{x}{3}\right) - x,$$

and so

$$f'(x) \geq 0 \iff \log\left(1 + x + \frac{x^2}{6}\right) \geq \frac{x}{1 + x/3} = \frac{3x}{x+3}$$

Letting $g(x) = \log\left(1 + x + \frac{x^2}{6}\right) - \frac{3x}{x+3}$, we obtain that $f'(x) \geq 0 \iff g(x) \geq 0$. In view of $g(0) = \log(1) - 0 = 0$, $g \geq 0$ if $g' \geq 0$ on $[0, \infty)$. We observe that

$$g'(x) = \frac{1 + x/3}{1 + x + x^2/6} - \frac{3(x+3) - 3x}{(x+3)^2} = \frac{x+3}{3 + 3x + x^2/2} - \frac{9}{(x+3)^2}.$$

To determine the sign of $g'(x)$, we place the terms over a common denominator. Because the denominators are strictly positive for $x \geq 0$, the sign of $g'(x)$ depends entirely on the cross-multiplied numerators:

$$(x+3)^3 \quad \text{vs.} \quad 9\left(3+3x+\frac{x^2}{2}\right)$$

Expanding the left side yields

$$(x+3)^3 = x^3 + 9x^2 + 27x + 27,$$

and expanding the right side yields

$$9\left(3+3x+\frac{x^2}{2}\right) = 4.5x^2 + 27x + 27.$$

Subtracting the right side from the left side gives

$$(x^3 + 9x^2 + 27x + 27) - (4.5x^2 + 27x + 27) = x^3 + 4.5x^2,$$

which is nonnegative for $x \geq 0$.

□

Lemma C.2 (Bennett's Inequality Inversion)**.** *Let Z be a random variable such that*

$$\mathbb{P}(Z \geq \epsilon) \leq \exp\left(-\frac{\sigma^2}{r^2} h\left(\frac{r\epsilon}{\sigma^2}\right)\right),$$

where $h(u) = (1+u)\log(1+u) - u$. Then, with probability at least $1-\delta$,

$$Z \leq \sqrt{2\sigma^2 \log(1/\delta)} + \frac{r\log(1/\delta)}{3}.$$

Proof. Let $z = \log(1/\delta)$. To find the high-probability upper bound for Z, we solve for the smallest ϵ such that the exponent ensures the tail probability is bounded by δ; that is,

$$\frac{\sigma^2}{r^2} h\left(\frac{r\epsilon}{\sigma^2}\right) \geq z.$$

This is equivalent to

$$h\left(\frac{r\epsilon}{\sigma^2}\right) \geq \frac{r^2 z}{\sigma^2}.$$

Since $h(u)$ is strictly increasing for $u \geq 0$, we can invert the function to isolate ϵ as

$$\frac{r\epsilon}{\sigma^2} \geq h^{-1}\left(\frac{r^2 z}{\sigma^2}\right).$$

Hence, it suffices to find ϵ such that

$$\frac{r\epsilon}{\sigma^2} \geq \sqrt{2\left(\frac{r^2 z}{\sigma^2}\right)} + \frac{1}{3}\left(\frac{r^2 z}{\sigma^2}\right) \geq h^{-1}\left(\frac{r^2 z}{\sigma^2}\right),$$

where we bounded the inverse rate function (second inequality) via Lemma C.1. The first inequality is equivalent to

$$\epsilon \geq \sqrt{2\sigma^2 z} + \frac{rz}{3}.$$

Substituting $z = \log(1/\delta)$ completes the proof. □

Lemma C.3. *Let V_n be a positive semi-definite matrix such that $\operatorname{tr}(V_n) > 0$. Define the function $f : (0,\infty) \to (0,\infty)$ by*

$$f(\sigma^2) = \frac{\operatorname{tr}(V_n)}{\sigma^2} \exp\left(-\frac{\sigma^2}{c^2} h\left(\frac{cr}{\sigma^2}\right)\right),$$

where $h(u) = (1+u)\log(1+u) - u$. There exist $c, r > 0$ such that $f(\sigma^2)$ is strictly increasing or strictly decreasing on some interval $(0, U)$, $U > 0$.

Proof. Let $x = \sigma^2 > 0$ and $T = \operatorname{tr}(V_n) > 0$. Since $f(x) > 0$ for all $x > 0$, $f(x)$ is strictly increasing if and only if $\log f(x)$ is strictly increasing. We therefore examine the derivative of $\log f(x)$ with respect to x; that is,

$$\log f(x) = \log T - \log x - \frac{x}{c^2} h\left(\frac{cr}{x}\right).$$

Using the definition $h(u) = (1+u)\log(1+u) - u$, we expand the last term to obtain

$$\frac{x}{c^2} h\left(\frac{cr}{x}\right) = \frac{x}{c^2}\left[\left(1+\frac{cr}{x}\right)\log\left(1+\frac{cr}{x}\right) - \frac{cr}{x}\right] = \frac{x+cr}{c^2}\log\left(1+\frac{cr}{x}\right) - \frac{r}{c}.$$

Taking the derivative with respect to x yields

$$\begin{aligned}\frac{d}{dx}\left[\frac{x}{c^2} h\left(\frac{cr}{x}\right)\right] &= \frac{1}{c^2}\log\left(1+\frac{cr}{x}\right) + \frac{x+cr}{c^2}\cdot\frac{1}{1+\frac{cr}{x}}\left(-\frac{cr}{x^2}\right) \\ &= \frac{1}{c^2}\log\left(1+\frac{cr}{x}\right) - \frac{r}{cx}.\end{aligned}$$

Substituting this back into the derivative of $\log f(x)$, we obtain

$$\frac{d}{dx}\log f(x) = -\frac{1}{x} - \frac{1}{c^2}\log\left(1+\frac{cr}{x}\right) + \frac{r}{cx} = \frac{1}{x}\left(\frac{r}{c} - 1 - \frac{x}{c^2}\log\left(1+\frac{cr}{x}\right)\right).$$

To analyze the term in the parentheses, let $u = \frac{cr}{x}$. Noting that $\frac{x}{c^2} = \frac{r}{cu}$, we can rewrite the derivative as

$$\frac{d}{dx}\log f(x) = \frac{1}{x}\left[\frac{r}{c}\left(1 - \frac{\log(1+u)}{u}\right) - 1\right].$$

Consider the limit of the bracketed expression as $x \to 0^+$, which corresponds to $u \to \infty$. By an application of L'Hôpital's rule, $\lim_{u\to\infty} \frac{\log(1+u)}{u} = 0$. Consequently,

$$\lim_{x\to 0^+}\left[\frac{r}{c}\left(1 - \frac{\log(1+u)}{u}\right) - 1\right] = \frac{r}{c} - 1.$$

If $r > c$, there exists some threshold $U > 0$ such that for all $x \in (0, U)$, $\log f(x)$ is strictly increasing on the interval $(0, U)$. Analogously, if $r < c$, $\log f(x)$ is strictly decreasing on some interval $(0, U)$. □

D Experiments description

We describe the simulations in Section 6. The accompanying `empirical_bernstein_matrix` repository provides code and execution instructions for all experiments. The code can also be found at https://github.com/DMartinezT/empirical_bernstein_matrix. Its `rank_one/` directory includes pinned dependencies, seeds, and trial-level results for the additional experiments in Appendices D.3 and D.4.

D.1 Covariance matrix experiment

Data Generation. We simulate independent, commuting random operators $X_1, \ldots, X_n \in \mathbb{H}$. Because the matrices commute, they share a common eigenbasis, allowing us to simulate their d-dimensional eigenvalues directly to drastically reduce computational overhead. Each observation is generated as a diagonal matrix $X_k = \operatorname{diag}(x_{k,1}, \ldots, x_{k,d})$, where the coordinates are drawn independently from a uniform distribution: $x_{k,i} \sim \mathcal{U}[0, a_i]$. Hence, we can take c instead of $2c$ in the definition of R_n (Equation (16)), because $0 \preceq X_i \preceq cI$, instead of only $\|X_i\| \le c$.

We fix the ambient dimension to $d = 3$ and explore sample sizes $n \in \{10^4, 3\times10^4, 10^5, 3\times10^5, 10^6\}$. We evaluate three spectral profiles by varying the sequence $\{a_i\}_{i=1}^d$:

1. **Completely Isotropic:** $a_i = 1$ for all $i \in [d]$.
2. **Completely Anisotropic:** $a_1 = 1$, and $a_i = 0$ for $i > 1$.
3. **Polynomial Decay:** $a_i = i^{-2}$ for $i \in [d]$.

Algorithmic Parameters. For all bounds, the failure probability was set to $\delta = 0.05$. The global uniform upper bound on the operators was conservatively set to $c = 1.0$, and the upper bound on the squared Hilbert-Schmidt norm used in our framework was set to the worst-case ambient limit $B = dc^2 = 3.0$. The true variance $V = \mathbb{E}[X_k^2] - (\mathbb{E}[X_k])^2$ is purely diagonal with $V_{i,i} = a_i^2/12$. The intrinsic dimension used to compute the oracle baseline is thus exactly $\text{tr}(V)/\|V\|$. We use the probability partition $\delta_1 = \delta(n-2)/n$, $\delta_2 = \delta_3 = \delta/n$.

Implementation of Baselines. We compare our intrinsic bound (OEB) against MEB 1 (the Maurer-Pontil union bound approach) and MEB 2 (the time-uniform supermartingale approach) from Wang and Ramdas [43]. Because their original results are formulated for bounding the maximum eigenvalue $\lambda_{\max}$, we apply a standard union bound over $\lambda_{\max}$ and $\lambda_{\min}$ to obtain guarantees for the operator norm. Consequently, the ambient dimension penalty in the logarithmic terms for MEB 1, MEB 2, and the Ambient Oracle (Tropp) is evaluated at $2d$ rather than d.

Monte Carlo Evaluation. For each sample size n and scenario, we perform 50 independent Monte Carlo trials. The plotted lines in Figure 1 represent the empirical mean of the ratio between each generated bound and the intrinsic oracle. The shaded error bands represent the central 95% empirical intervals, calculated via the 2.5th and 97.5th percentiles of the ratios observed across the 50 trials.

D.2 Kernel PCA experiment

We evaluate our empirical Bernstein inequality in an infinite-dimensional RKHS without relying on finite-dimensional feature approximations. Let $K \in \mathbb{R}^{n\times n}$ be the kernel matrix evaluated on the data $X_1, \dots, X_n$. The exact evaluation of our bound requires computing traces and operator norms of combinations of the rank-one operators $Y_i = \phi(X_i) \otimes \phi(X_i)$. By utilizing the kernel trick, these operator-theoretic quantities can be computed strictly from scalar evaluations of K. For instance, the empirical trace proxy $Z_i = \text{tr}(Y_i')$ in our framework is exactly $1 - k(X_{2i-1}, X_{2i})^2$. Furthermore, the operator norm of the intermediate empirical variance Σ_n is equivalent to the maximum eigenvalue of the matrix product KM, where M is an $n \times n$ block-diagonal matrix whose entries depend strictly on $k(X_{2i-1}, X_{2i})$.

This allows us to exactly evaluate the data-driven confidence radius proposed in Theorem 5.5. We utilize the Gaussian RBF kernel k_{RBF} ($\gamma = 0.1$) applied to 10-dimensional input data $X_i \in \mathbb{R}^{10}$. We sample the inputs from three distinct distributions: $\mathcal{N}(0, I_{10})$, $\mathcal{U}[-1, 1]^{10}$, and a coordinate-wise $\text{Exp}(1)$ distribution. For the oracle baseline, the intrinsic dimension $\text{tr}(V)/\|V\|$ for each distribution was estimated offline via a high-density Monte Carlo simulation ($N = 2,000$), utilizing the property that for shift-invariant kernels k with $k(x, x) = 1$, the true variance operator $V = C - C^2$ has eigenvalues $\gamma_j = \lambda_j(1 - \lambda_j)$, where λ_j are the true population eigenvalues of $C = \mathbb{E}Y_i$. We use the probability partition $\delta_1 = \delta(n-2)/n$, $\delta_2 = \delta_3 = \delta/n$.

D.3 Higher-dimensional noncommuting updates

For each $r \in \{3, 5, 10, 20\}$, draw independent Rademacher signs ϵ_{ij} and set $u_i = (\epsilon_{i1}, \dots, \epsilon_{ir})/\sqrt{r}$. The dense rank-one observations are $X_i = Q_r u_i u_i^\top Q_r^\top$, where Q_r comprises the first r columns of a fixed 1000×20 orthonormal matrix obtained by QR decomposition of a Gaussian matrix. Thus $c = B = 1$, $C = \mathbb{E}X_i = Q_r Q_r^\top/r$, and $\Sigma = \mathbb{E}[(X_i - C)^2]$ has r nonzero eigenvalues $(r-1)/r^2$, giving exact effective rank r. We use $n = 20{,}000$ and 40 independent trials per rank.

Computation in the active r coordinates preserves products, traces, norms, and commutators, while MEB 1 and MEB 2 retain the ambient dimension $d = 1000$. An ambient bound explicitly restricted to the known active subspace could instead use $d = r$; the comparison here measures the cost of retaining the ambient representation. Across all $n-1$ adjacent pairs per trial, the nonzero-commutator frequencies are 74.90%, 93.77%, 75.13%, and 82.32% in rank order. They are evaluated exactly from $0 < |\epsilon_i^\top \epsilon_{i+1}| < r$, without a floating-point threshold or resampling commuting pairs.

OEB is smaller than both ambient baselines throughout Table 1; at rank 10, the reductions in mean normalized radius are 27.6% relative to MEB 1 and 17.5% relative to MEB 2.

Table 1: Noncommuting rank-one updates: mean radius divided by the exact intrinsic oracle, $\pm$ one Monte Carlo standard error ($d = 1000$, $n = 20{,}000$, $\delta = 0.05$, 40 trials per rank).

Effective rank	OEB	MEB 1	MEB 2
3	1.0475 ± 0.0006	1.6646 ± 0.0009	1.5369 ± 0.0003
5	1.0757 ± 0.0003	1.6173 ± 0.0005	1.4665 ± 0.0009
10	1.1493 ± 0.0003	1.5880 ± 0.0005	1.3923 ± 0.0008
20	1.2807 ± 0.0002	1.5966 ± 0.0005	1.3393 ± 0.0006

D.4 Finite-rank RKHS truncations

On the product space $\{-1, +1\}^{\mathbb{N}}$ with independent fair coordinates, take $p_j = 0.1(0.9)^{j-1}$, $\phi(x)_j = \sqrt{p_j}x_j$, and $k(x, y) = \sum_{j\geq 1} p_j x_j y_j$. The coordinate functions are orthonormal under the input law, with Mercer weights p_j, and $k(x, x) = 1$. For truncation rank m, use $X_i = \phi_m(x_i) \otimes \phi_m(x_i)$ without renormalizing the truncated features. The exact envelopes and population operators are

$$c_m = 1 - 0.9^m, \qquad B_m = c_m^2, \qquad C_m = \operatorname{diag}(p_1, \ldots, p_m),$$
$$\Sigma_m = c_m C_m - C_m^2 = \operatorname{diag}\{p_j(c_m - p_j)\}_{j=1}^m.$$

All methods use $c = c_m$, $B = B_m$, and the exact oracle for that truncation; the ambient baselines use $d = m$. These radii cover C_m, not the untruncated C. We generate 100 independent $10{,}000 \times 256$ sign samples and use their leading columns for $m = 32, 64, 128, 256$, so all four truncations share the same samples. The effective rank tends to $200/19 \approx 10.5263$ as $m \to \infty$.

Table 2: Rademacher-feature RKHS truncations: mean radius divided by the exact truncation-specific intrinsic oracle, $\pm$ one Monte Carlo standard error ($n = 10{,}000$, $\delta = 0.05$, 100 shared Monte Carlo samples).

Truncation rank	OEB	MEB 1	MEB 2
32	1.1983 ± 0.0001	1.3298 ± 0.0002	1.1266 ± 0.0002
64	1.2046 ± 0.0001	1.3986 ± 0.0002	1.1791 ± 0.0002
128	1.2049 ± 0.0001	1.4653 ± 0.0002	1.2322 ± 0.0002
256	1.2049 ± 0.0001	1.5299 ± 0.0002	1.2836 ± 0.0002

From rank 32 to 256, the mean normalized OEB radius changes by 0.6%, while MEB 1 and MEB 2 increase by 15.0% and 13.9%. MEB 2 is tighter at ranks 32 and 64, but OEB crosses below it between 64 and 128. At rank 256, OEB is 21.2% smaller than MEB 1 and 6.1% smaller than MEB 2. These comparisons demonstrate adaptation across finite truncations without implying uniform finite-sample dominance.

Bound evaluation for both additional experiments. Set $\delta = 0.05$ and $\delta_1 = \delta(n-2)/n$, $\delta_2 = \delta_3 = \delta/n$. Since $0 \preceq X_i, C \preceq cI$, the centered norm envelope is c. Accordingly, OEB uses (16) with c replacing $2c$ only in its final linear term; the paired-estimator envelopes $2B$, $4Bc^2$, and $2c^2$ are unchanged. For $\sigma^2 = \|\Sigma\|$ and $d_{\text{eff}} = \operatorname{tr}(\Sigma)/\sigma^2$, the exact intrinsic Bernstein oracle is

$$r_{\text{int}} = \sqrt{\frac{2\sigma^2}{n} \log \frac{2d_{\text{eff}}}{\delta}} + \frac{c}{3n} \log \frac{2d_{\text{eff}}}{\delta}.$$

Here "exact" refers to the analytically known population moments, not to an optimal radius or an inversion of Bennett's bound. The ambient oracle replaces d_{eff} by d. MEB 1 and MEB 2 implement Wang and Ramdas [43, Theorem 3.1 and Corollary 4.3] on X_i/c, rescaled by c, with failure probability $\delta/2$ per tail. OEB and MEB 1 use the sample mean; MEB 2 uses its predictable-weight mean. Its weights use the past empirical variance, initialized at $I/4$ for normalized observations, with variance floor $5\log(2d/\delta)/n$. All methods receive identical observations within each condition.

Reproducibility and uncertainty reporting. Tables 1 and 2 use the completed run with master seed 20260919 (NumPy PCG64/SeedSequence). Noncommuting condition/trial streams use $[20260919, 0, r, t]$, RKHS trial streams use $[20260919, 1, 0, t]$, and the embedding uses $[20260919, 2]$,

with zero-based t. RKHS signs are generated in a $256 \times n$ array and transposed before taking leading columns. Table entries are arithmetic means $\pm s/\sqrt{T}$, with Bessel-corrected trial standard deviation s and $T = 40$ or 100; they quantify Monte Carlo uncertainty in the mean, not confidence coverage. Shared RKHS ranks are not treated as independent trials. The package includes all 560 condition-trial records, the exact oracles, both estimator errors, coverage indicators, and a manifest with source/output hashes.

In `rank_one/`, install `requirements.txt` and run `python -m unittest discover -s tests -v`. Reproduce both experiments with `python run_experiments.py -jobs 4 -output my_results`. The recorded run used Python 3.12.14, NumPy 2.3.5, SciPy 1.16.3, and four CPU worker processes with one numerical-library thread each, taking 19.93 minutes on macOS 15.7.9. The iterative symmetric eigensolver uses tolerance 10^{-11}, checks residuals, and falls back to a dense solve; `-solver dense` selects the reference implementation. CPU model and peak memory were not recorded.

E Finite-rank reduction and trace domains

We establish here the trace arguments used in Theorems 4.1 and 5.2. We work under the operator and MGF assumptions of Theorem 4.1, with $S_k = \sum_{i \le k} X_i$, $V_n = \sum_{i \le n} \Delta V_i$, and $\varphi(u) = \cosh(u) - 1$.

The trace inequalities in Theorems 3.1 and 3.2 are stated for matrices. In an infinite-dimensional space, even the identity has infinite trace, so we must first check which expressions can be traced before passing from matrices to operators. The centered functions $\varphi(A) = \cosh(A) - I$ and $e^{\psi V} - I$ allow us to make this passage.

Specifically, for θ in the MGF domain, we establish the following points in order:

1. The centered functions have finite traces, and finite-rank approximation recovers $\mathrm{tr}\varphi(\theta S_k)$.
2. The MGF assumption provides the integrability needed for the compressed MGFs and their logarithms to converge in trace norm.
3. Lieb's inequality passes to the full space, yielding $\mathbb{E}\mathrm{tr}\varphi(\theta S_n) \le \mathrm{tr}(e^{\psi(|\theta|)V_n} - I) < \infty$.
4. The process $(\mathrm{tr}\varphi(\theta S_k))_{k=0}^n$ is an integrable nonnegative submartingale, so Doob's maximal inequality applies.

Details of step 1. We begin by explaining why these centered traces are finite. Write $\|T\|_1 = \mathrm{tr}|T|$, where $|T| = (T^*T)^{1/2}$, for the trace norm. An operator is trace-class when this norm is finite. For self-adjoint Hilbert-Schmidt A, its square is trace-class, and

$$\varphi(A) = \sum_{j \ge 1} \frac{A^{2j}}{(2j)!}, \qquad \|A^{2j}\|_1 \le \|A\|^{2j-2} \|A\|_{\mathrm{HS}}^2. \tag{34}$$

Here we use $A^{2j} = A^{2j-2}A^2$ and $\|BC\|_1 \le \|B\|\|C\|_1$, with $\|A^2\|_1 = \mathrm{tr}(A^2) = \|A\|_{\mathrm{HS}}^2$. Consequently, this series converges in trace norm. The absence of constant and linear terms is what allows the Hilbert-Schmidt assumption to suffice. Similarly, if $V \succeq 0$ is trace-class and $\psi \ge 0$ is finite, then $\|V^j\|_1 \le \|V\|^{j-1}\mathrm{tr}(V)$ shows that $e^{\psi V} - I$ is trace-class. All infinite-dimensional traces below are taken after this centering. Trace-norm convergence is useful here because $|\mathrm{tr}C - \mathrm{tr}D| \le \|C - D\|_1$ for trace-class C, D.

To reduce the infinite-dimensional case to matrices, choose an orthonormal basis of $\mathbb{H}$ and let P_m project onto its first m vectors, following Minsker [27]. These projections increase to the identity: $P_m x \to x$ for every $x \in \mathbb{H}$. Define

$$X_{i,m} = P_m X_i P_m, \qquad S_{k,m} = \sum_{i \le k} X_{i,m} = P_m S_k P_m.$$

On $P_m\mathbb{H}$, the $X_{i,m}$ are finite-dimensional self-adjoint matrices and retain independence and zero mean. We also view them as operators on $\mathbb{H}$ by setting them equal to zero on $(P_m\mathbb{H})^\perp$. Since $\varphi(0) = 0$, their centered functions are zero there as well, so their matrix traces agree with their traces on the full space.

For a self-adjoint Hilbert-Schmidt operator A, the compression $A_m = P_m A P_m$ retains the first m rows and columns of its matrix in the chosen basis. Square summability of the matrix entries gives $\|A_m - A\|_{\mathrm{HS}} \to 0$, with $\|A_m\|_{\mathrm{HS}} \le \|A\|_{\mathrm{HS}}$. To see how this yields convergence of centered traces, use the product inequality $\|BC\|_1 \le \|B\|_{\mathrm{HS}}\|C\|_{\mathrm{HS}}$ to obtain

$$\|A_m^2 - A^2\|_1 \le \big(\|A_m\|_{\mathrm{HS}} + \|A\|_{\mathrm{HS}}\big)\|A_m - A\|_{\mathrm{HS}} \longrightarrow 0.$$

For higher powers, the identity

$$A_m^q - A^q = \sum_{\ell=0}^{q-1} A_m^{\ell}(A_m - A)A^{q-1-\ell}, \qquad q \ge 2,$$

has $A_m - A$ and at least one further Hilbert-Schmidt factor in each term. The same product inequality therefore gives trace-norm convergence of every power of degree at least two. The series bound in (34) is uniform in m, since $\|A_m\| \le \|A\|$. We can consequently pass the limit through the power series for φ. In particular, for every fixed θ and $k \le n$,

$$\|\varphi(\theta S_{k,m}) - \varphi(\theta S_k)\|_1 \longrightarrow 0 \quad \text{almost surely.} \tag{35}$$

The same reasoning applies to $e^{\pm\theta A} - I \mp \theta A$, whose power series also starts at degree two.

Details of step 2. Passing to expectations requires an integrable bound in addition to the samplewise convergence in (35). Fix $\theta \in (-\theta_{\max}, \theta_{\max})$ and put $M_i^{\pm} = \mathbb{E}e^{\pm\theta X_i}$. The scalar inequality $e^u \ge 1 + u$, applied by spectral calculus and then averaged, gives $M_i^{\pm} \succeq I$ because $\mathbb{E}X_i = 0$. Thus the MGF assumption implies

$$0 \preceq \log M_i^{\pm} \preceq \psi(|\theta|)\Delta V_i.$$

The upper bound is trace-class, so $\log M_i^{\pm}$ is also trace-class. The centered-exponential argument following (34) now gives that $M_i^{\pm} - I = e^{\log M_i^{\pm}} - I$ is trace-class. Since $\varphi(\theta X_i) \succeq 0$, Tonelli's theorem, applied to its diagonal entries in an orthonormal basis, yields

$$\mathbb{E}\mathrm{tr}\varphi(\theta X_i) = \frac{1}{2}\mathrm{tr}\big[(M_i^{+} - I) + (M_i^{-} - I)\big] < \infty. \tag{36}$$

We next use the integrability in (36) to control the compressed MGFs. Define the nonnegative convex functions $f_{\pm}(u) = e^{\pm\theta u} - 1 \mp \theta u$, which satisfy $f_{+}(u) + f_{-}(u) = 2\varphi(\theta u)$. The following consequence of scalar Jensen explains why compression does not increase their traces. For any nonnegative convex f with $f(0) = 0$ and trace-class $f(A)$, choose an orthonormal eigenbasis $(u_j)_{j=1}^m$ of A_m on $P_m\mathbb{H}$. Its eigenvalues are $\langle u_j, Au_j\rangle$. Expressing u_j in an eigenbasis of A writes this inner product as a weighted average of the eigenvalues of A, with nonnegative weights summing to one. Applying scalar Jensen to each of these averages and then summing gives

$$\mathrm{tr}f(A_m) = \sum_{j=1}^{m} f(\langle u_j, Au_j\rangle) \le \sum_{j=1}^{m}\langle u_j, f(A)u_j\rangle \le \mathrm{tr}f(A). \tag{37}$$

The last inequality uses $f(A) \succeq 0$. Applied to $f_{\pm}$ and $A = X_i$, this gives

$$0 \le \mathrm{tr}f_{\pm}(X_{i,m}) \le \mathrm{tr}f_{\pm}(X_i) \le 2\mathrm{tr}\varphi(\theta X_i).$$

Together with the trace-norm convergence of $f_{\pm}(X_{i,m})$ established after (35), this supplies the integrable bound $\|f_{\pm}(X_{i,m}) - f_{\pm}(X_i)\|_1 \le 4\mathrm{tr}\varphi(\theta X_i)$. Dominated convergence therefore allows us to pass expectations through the compression limit.

More explicitly, let $M_{i,m}^{\pm} = \mathbb{E}e^{\pm\theta X_{i,m}}$, extended by the identity on $(P_m\mathbb{H})^{\perp}$. Zero mean gives $M_{i,m}^{\pm} - I = \mathbb{E}f_{\pm}(X_{i,m})$ and $M_i^{\pm} - I = \mathbb{E}f_{\pm}(X_i)$, hence, by dominated convergence,

$$\|M_{i,m}^{\pm} - M_i^{\pm}\|_1 \le \mathbb{E}\|f_{\pm}(X_{i,m}) - f_{\pm}(X_i)\|_1 \longrightarrow 0.$$

Lieb's theorem involves the logarithms of these MGFs, so we also need their convergence. For $C, D \succeq I$ with trace-class difference, the integral representation of $\log$ and the resolvent identity give

$$\log D - \log C = \int_0^{\infty} (C + tI)^{-1}(D - C)(D + tI)^{-1}\, dt,$$

$$\|\log D - \log C\|_1 \le \|D - C\|_1 \int_0^\infty (1+t)^{-2}\, dt = \|D - C\|_1.$$

Here the resolvent norms are at most $(1+t)^{-1}$ because $C, D \succeq I$. Both $M_{i,m}^{\pm}$ and $M_i^{\pm}$ have this lower bound. Consequently, their summed log MGFs, or cumulants, satisfy

$$K_m^{\pm} := \sum_{i=1}^n \log M_{i,m}^{\pm} \longrightarrow K^{\pm} := \sum_{i=1}^n \log M_i^{\pm} \quad \text{in trace norm.} \tag{38}$$

Details of step 3. We now apply Lieb's concavity theorem (Theorem 3.2) in finite dimension. Independence allows us to condition on one increment at a time and replace its exponential by its MGF, giving $\mathbb{E}\mathrm{tr} e^{\pm\theta S_{n,m}} \le \mathrm{tr} e^{K_m^{\pm}}$ on $P_m\mathbb{H}$. Subtracting the finite trace of the identity on this subspace and averaging the two signs yields

$$\mathbb{E}\mathrm{tr}\varphi(\theta S_{n,m}) \le \frac{1}{2}\mathrm{tr}\big[(e^{K_m^+} - I) + (e^{K_m^-} - I)\big]. \tag{39}$$

To pass the right-hand side of (39) to the limit, the exponential power series gives

$$\|(e^{K_m^{\pm}} - I) - (e^{K^{\pm}} - I)\|_1 \le e^{\max\{\|K_m^{\pm}\|, \|K^{\pm}\|\}} \|K_m^{\pm} - K^{\pm}\|_1 \longrightarrow 0.$$

The exponential factor is bounded because the cumulants converge in trace norm by (38), so the traces on the right-hand side of (39) converge. On the left-hand side, the almost sure trace convergence following from (35) and nonnegativity permit Fatou's lemma. Combining these two limiting arguments with (39), we obtain

$$\begin{aligned}
\mathbb{E}\mathrm{tr}\varphi(\theta S_n) &\le \liminf_{m\to\infty} \mathbb{E}\mathrm{tr}\varphi(\theta S_{n,m}) \\
&\le \lim_{m\to\infty} \frac{1}{2}\mathrm{tr}\big[(e^{K_m^+} - I) + (e^{K_m^-} - I)\big] \\
&= \frac{1}{2}\mathrm{tr}\big[(e^{K^+} - I) + (e^{K^-} - I)\big] \\
&\le \mathrm{tr}\big(e^{\psi(|\theta|)V_n} - I\big) < \infty.
\end{aligned} \tag{40}$$

Here the first inequality is Fatou's lemma, the second follows from the finite-dimensional Lieb bound (39) at each m, and the equality follows from the trace-norm convergence of the centered exponentials. Thus the centered trace bound obtained from Lieb's concavity, averaged over the two signs, passes from the finite-dimensional subspaces to the Hilbert space.

The final inequality in (40) then uses the MGF assumption (4) on the original operators: $0 \preceq K^{\pm} \preceq \psi(|\theta|)V_n$. By the min-max characterization, their eigenvalues, ordered from largest to smallest, satisfy $\lambda_j(K^{\pm}) \le \psi(|\theta|)\lambda_j(V_n)$. The centered exponentials are positive and trace-class by step 1, so their traces are the sums of $e^{\lambda_j(K^{\pm})} - 1$ and $e^{\psi(|\theta|)\lambda_j(V_n)} - 1$, respectively. Scalar monotonicity compares each pair of summands and hence their sums. Averaging the resulting bounds for K^+ and K^- gives the final comparison in (40). Repeating the argument with the first k increments proves $\mathbb{E}\mathrm{tr}\varphi(\theta S_k) < \infty$ for every $k \le n$.

Details of step 4. Finally, we establish the submartingale property required by Doob's inequality. Let $\mathcal{F}_k = \sigma(X_1, \dots, X_k)$, $T_{k,m} = \mathrm{tr}\varphi(\theta S_{k,m})$, and $T_k = \mathrm{tr}\varphi(\theta S_k)$. The compression trace bound (37) gives $0 \le T_{k,m} \le T_k$, so these variables are integrable by (40). Conditional trace Jensen (Theorem 3.1) and $\mathbb{E}[X_{k+1,m} \mid \mathcal{F}_k] = 0$ give, as in Lemma 3.3,

$$\mathbb{E}[T_{k+1,m} \mid \mathcal{F}_k] \ge \mathrm{tr}\varphi\big(\theta\mathbb{E}[S_{k+1,m} \mid \mathcal{F}_k]\big) = T_{k,m}. \tag{41}$$

To pass (41) to the full space, apply the scalar Jensen compression inequality (37) to the nested subspaces $P_m\mathbb{H} \subset P_{m+1}\mathbb{H}$ and the function $u \mapsto \varphi(\theta u)$. It gives $T_{k,m} \le T_{k,m+1}$. The trace convergence from (35) identifies their limit as T_k, so conditional monotone convergence yields

$$\mathbb{E}[T_{k+1} \mid \mathcal{F}_k] = \lim_{m\to\infty} \mathbb{E}[T_{k+1,m} \mid \mathcal{F}_k] \ge \lim_{m\to\infty} T_{k,m} = T_k.$$

Since T_k is nonnegative and integrable by (40), this proves that $(T_k)_{k=0}^n$ is a nonnegative submartingale. We have therefore established both ingredients needed by the master theorems: the trace MGF bound and the validity of Doob's maximal inequality in the full space.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction explicitly state the scope (bounded, self-adjoint Hilbert-Schmidt operators), the core assumptions (independent sequences), and accurately reflect the main theoretical contributions (intrinsic-dimension empirical Bennett and Bernstein inequalities).

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We explicitly define the structural boundaries of our theorems (e.g., Assumption 5.1 requiring uniformly bounded operators and independent sequences) and discuss the limitation of independence in the conclusion, identifying it as an avenue for future work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Every theorem, lemma, and corollary explicitly references the required conditions (e.g., Assumption 5.1). Complete, rigorous mathematical proofs for all results are provided in Appendix B, with the finite-rank reduction in Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Appendix D describes the synthetic models, bound parameters, and evaluation procedures. The accompanying `empirical_bernstein_matrix` repository supplies code and execution instructions for all experiments; `rank_one/` includes pinned dependencies, fixed seeds, and trial-level results for Tables 1 and 2.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The accompanying `empirical_bernstein_matrix` repository provides simulation and plotting scripts, dependencies, and execution instructions for Figures 1 and 2 and Tables 1 and 2. Recorded outputs are included, with trial-level results for both tables.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Appendix D specifies the data-generating models, dimensions, sample sizes, trial counts, confidence allocations, boundedness constants, and baseline settings. The additional RKHS experiment explicitly states its geometric weights, shared samples, and absence of renormalization after truncation.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: Figure 1 reports central 95% empirical intervals across 50 trials. Tables 1 and 2 report means $\pm$ one Monte Carlo standard error, computed as the Bessel-corrected standard deviation divided by the square root of 40 or 100 trials; these quantify simulation uncertainty rather than confidence coverage.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Appendix D.4 records the additional experiments' software environment, four single-threaded CPU workers, and elapsed time of 19.93 minutes. CPU model, peak memory, and complete resource measurements for the original figures were not recorded.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This is foundational theoretical research in probability and statistics and strictly conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This work focuses on foundational mathematical theory (operator-valued concentration inequalities). It is not tied to any particular deployment or application that poses direct societal impacts or foreseeable negative consequences.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release any datasets, models, or technologies that carry a risk of misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [N/A]

Justification: The paper relies entirely on synthetically generated data for its simulations. No existing datasets, models, or third-party code assets were used.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper does not introduce or release new data or model assets. The supplement documents the reproduction scripts and synthetic trial-level outputs for the additional experiments.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The research does not involve crowdsourcing or human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The research does not involve human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: Large language models were not used as a component of the core originality of this research.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.