Electron. J. Probab. **19** (2014), no. 27, 1–30.
ISSN: 1083-6489 DOI: 10.1214/EJP.v19-2964

Subadditivity of matrix φ-entropy and concentration of random matrices

Richard Y. Chen* Joel A. Tropp†

Abstract

This paper considers a class of entropy functionals defined for random matrices, and it demonstrates that these functionals satisfy a subadditivity property. Several matrix concentration inequalities are derived as an application of this result.

Keywords: Concentration inequalities; entropy method; moment inequalities; random matrix; subadditivity; tensorization.
AMS MSC 2010: Primary 60B20; 60E15, Secondary 60G09; 60F10.
Submitted to EJP on August 13, 2013, final version accepted on February 26, 2014.
Supersedes arXiv:1308.2952.

1 Introduction and Related Work

Entropy and related functions quantify the uncertainty inherent in a probability distribution. Measures of entropy have the property that the total entropy of a "product" is bounded by the sum of the entropies of the "factors." This fundamental fact is called *subadditivity of entropy*, or sometimes *tensorization*, and it drives many applications of entropy. The survey [Lie75] contains a discussion of subadditivity in statistical mechanics, and the monograph [RS13] describes examples in information theory. In this work, we focus on the role of subadditivity of entropy in probability.

1.1 Subadditivity and Concentration

A *concentration inequality* states that a random variable is unlikely to exhibit a significant deviation from its mean value. The current intuition holds that a random variable concentrates whenever it depends smoothly on many independent random variables [Tal96]. Ledoux [Led97, Led01] and Bobkov & Ledoux [BL98] initiated a line of research that uses methods based on entropy to derive concentration inequalities. A few of the many authors who have contributed include Massart [Mas00a, Mas00b], Rio [Rio01], Bousquet [Bou02], and Boucheron et al. [BLM03, BLM09]. See the book [BLM13] for a comprehensive treatment of this theory and its bibliography.

*California Institute of Technology, USA. E-mail: ycchen@caltech.edu
†California Institute of Technology, USA. E-mail: jtropp@cms.caltech.edu

Let us summarize the ideas that lead from entropy to concentration. In this setting, we define the *entropy functional* for each nonnegative, real random variable Z:

$$H(Z) := \mathbb{E}(Z \log Z) - (\mathbb{E}\, Z) \log(\mathbb{E}\, Z). \tag{1.1}$$

Heuristically, $H(Z)$ quantifies our uncertainty about the precise value of Z. We typically consider the situation where $Z = \mathrm{e}^{\theta Y}$ for a zero-mean random variable Y. In this case, we have the identity

$$\log \mathbb{E}\, \mathrm{e}^{\theta Y} = \theta \int_0^\theta \frac{H(\mathrm{e}^{\beta Y})}{\mathbb{E}\, \mathrm{e}^{\beta Y}} \cdot \frac{\mathrm{d}\beta}{\beta^2}. \tag{1.2}$$

Through Markov's inequality, bounds on the left-hand side imply that Y takes a large value with exponentially small probability. Therefore, we might hope to invoke inequalities for the entropy functional H to analyze the fluctuations of Y.

Indeed, the entropy functional exhibits a subadditivity property that allows us to implement this program. Suppose that Z is a function of mutually independent random variables $X_1, \dots, X_n$. We can define conditional entropy functionals

$$H_i(Z) := \mathbb{E}_i(Z \log Z) - (\mathbb{E}_i\, Z) \log(\mathbb{E}_i\, Z)$$

where $\mathbb{E}_i$ denotes the expectation with respect to X_i, holding X_j fixed for $j \neq i$. The conditional entropy H_i reflects the uncertainty about Z that is attributable to our lack of knowledge about X_i. Subadditivity is the nontrivial result that

$$H(Z) \leq \sum\nolimits_{i=1}^{n} \mathbb{E}[H_i(Z)]. \tag{1.3}$$

In other words, our uncertainty about Z does not exceed the total (average) uncertainty due to each X_i individually. Combining the identity (1.2) and the subadditivity property (1.3) with bounds for the conditional entropy, we can establish exponential concentration inequalities for functions of independent random variables.

The idea of considering alternative forms of entropy can be traced at least as far as the work of Rényi [Rén61], Bregman [Brè66], and Csiszár [Csi72]. In the early 2000s, researchers [LO00, Cha04, BBLM05, Cha06] recognized that generalized entropy functionals can exhibit subadditivity properties similar with those of the entropy functional (1.1). Let $\varphi : \mathbb{R}_+ \to \mathbb{R}$ be a convex function. The *φ-entropy functional* is defined for each nonnegative random variable Z by the formula

$$H_\varphi(Z) := \mathbb{E}\, \varphi(Z) - \varphi(\mathbb{E}\, Z).$$

The functional (1.1) derives from the choice $\varphi : t \mapsto t \log t$. Under stringent conditions on φ, it can be shown that the φ-entropy functional satisfies a subadditivity property analogous with (1.3). In particular, the function $\varphi : t \mapsto t^p$ yields a subadditive φ-entropy when $1 \leq p \leq 2$, a fact that leads to polynomial concentration inequalities [BBLM05].

1.2 Subadditivity of Matrix Entropies

The purpose of this paper is to explore the subadditivity properties of entropy functionals defined on matrix-valued random variables. Let $\varphi : \mathbb{R}_+ \to \mathbb{R}$ be a convex function. For a positive-semidefinite (psd) random matrix $\boldsymbol{Z}$, we can consider the *matrix φ-entropy functional*

$$H_\varphi(\boldsymbol{Z}) := \mathbb{E}\, \bar{\mathrm{tr}}\, \varphi(\boldsymbol{Z}) - \bar{\mathrm{tr}}\, \varphi(\mathbb{E}\, \boldsymbol{Z})$$

where φ refers to a standard matrix function and $\bar{\mathrm{tr}}$ denotes the normalized trace. See Section 2.1 for definitions. It may be helpful to note some alternative presentations

of the matrix φ-entropy. First, the expression has the same structure as the scalar entropy (1.1) because

$$H_\varphi(\boldsymbol{Z}) = \mathbb{E}\,\Phi(\boldsymbol{Z}) - \Phi(\mathbb{E}\,\boldsymbol{Z}) \quad \text{where } \Phi := \bar{\operatorname{tr}}\,\varphi \text{ is convex.}$$

Second, we can decompose the matrix entropy as

$$H_\varphi(\boldsymbol{Z}) = \big[\,\mathbb{E}\,\bar{\operatorname{tr}}\,\varphi(\boldsymbol{Z}) - \varphi(\mathbb{E}\,\bar{\operatorname{tr}}\,\boldsymbol{Z})\big] + \big[\varphi(\bar{\operatorname{tr}}\,\mathbb{E}\,\boldsymbol{Z}) - \bar{\operatorname{tr}}\,\varphi(\mathbb{E}\,\boldsymbol{Z})\big].$$

In other words, the matrix entropy quantifies the total loss in two different averaging operations on the matrix.

This work contains two main contributions:

1. We develop conditions on φ which ensure that the matrix φ-entropy is subadditive.
2. We verify these conditions for the functions $\varphi : t \mapsto t \log t$ and $\varphi : t \mapsto t^p$ where $p \in [1, 2]$.

The arguments parallel the analysis of scalar φ-entropies in Boucheron et al. [BBLM05], but the technical difficulties are more formidable in the matrix setting.

There are several areas that may benefit from this investigation.

Random matrix theory In the scalar setting, subadditivity of φ-entropy leads to powerful concentration inequalities. The subadditivity of matrix φ-entropy allows us to adapt these arguments to obtain some concentration inequalities for random matrices. See Section 1.3 for more information.

Convex analysis We derive subadditivity of the matrix φ-entropy functional H_φ from its convexity properties; see Lemma 4.1 et seq. These results may be useful in other contexts. For example, the convexity of scalar φ-entropy plays a role in machine learning [RW11, Sec. 2.5 et seq.].

Operator theory To prove that specific examples of matrix φ-entropy are subadditive, we rely on sophisticated methods from operator theory. In return, the results here may be relevant for problems in operator theory.

Quantum theory In quantum statistical mechanics and quantum information theory, entropies are defined for positive-definite matrices. Subadditivity of the quantum relative entropy function plays an important role in these fields, and this same result is closely connected with subadditivity of the matrix entropy H_φ where $\varphi : t \mapsto t \log t$. As such, subadditivity of other matrix φ-entropies may be relevant for quantum theory.

1.2.1 Related Work

After this paper was written, we learned about a contemporary paper [Han13] of Hansen that contains subadditivity results like the ones here. We also mention the subsequent paper [HZ14], which builds on our work. Detailed references appear below.

For the function $\varphi : t \mapsto t \log t$, we are aware of other precedents for our subadditivity results. In this special case, the subadditivity of H_φ follows from a classical result [Lin73] after a moderate amount of argument. In quantum statistical mechanics, subadditivity of entropy refers to a specific type of inequality for partitioned quantum systems [LR73]; see the lecture notes [Car10] for a recent presentation of these ideas. The paper [HOZ01] contains another type of subadditivity bound.

Finally, we mention the concept of *free entropy*, which is the appropriate generalization of entropy in noncommutative probability. As with other entropy measures, free entropy is subadditive. Arguments based on free entropy can be used to study extremely large matrices that are unitarily invariant. See the survey [Voi02] for further details and references.

1.3 Matrix Concentration Inequalities

A *matrix concentration inequality* provides a bound on the spectral-norm deviation of a random matrix from its mean [AW02, Oli09, Tro11, Tro12b, Min12, MJC$^+$12, PMT13]; see the survey [Tro12c] for an annotated bibliography. These results have already found applications in a wide range of areas, including random graph theory [Oli09, CCT12], randomized linear algebra [DZ11, Tro12a, BG12], and least-squares approximation [CDL13]. There is also a separate line of work that leads to remarkable concentration inequalities for the spectral measure of a random Hermitian matrix; see for example [AGZ10, Sec. 4.4].

In spite of these successes, one frequently encounters random matrices that do not submit to existing techniques. Therefore, the study of matrix concentration remains an active area of investigation. As discussed in Section 1.1, subadditivity of scalar φ-entropy leads to a variety of concentration inequalities [BLM03, BBLM05]. It is natural to ask whether subadditivity of matrix φ-entropy leads to new concentration inequalities for random matrices.

We show that it is indeed possible to adapt scalar arguments based on subadditivity of entropy to the matrix setting, and we obtain some interesting matrix concentration inequalities. On the other hand, this method is not as satisfying as some other approaches to matrix concentration because the resulting bounds involve artificial assumptions. In fact, the matrix concentration inequalities in this work are dominated by the results we can obtain using arguments based on exchangeable pairs [PMT13]. This fact suggests that the subadditivity properties of matrix φ-entropy do not fully capture the behavior of a random matrix.

2 Main Results

In this section, we lay out detailed definitions and statements of our main results on subadditivity of matrix φ-entropy and its application to prove concentration inequalities for random matrices.

2.1 Notation and Background

Let us instate some notation. The set $\mathbb{R}_+$ contains the nonnegative real numbers, and $\mathbb{R}_{++}$ consists of all positive real numbers. We write $\mathbb{M}^d$ for the complex Banach space of $d \times d$ complex matrices, equipped with the usual ℓ_2 operator norm $\|\cdot\|$. The *normalized trace* is the function

$$\bar{\operatorname{tr}}\, \boldsymbol{B} := \frac{1}{d} \sum\nolimits_{j=1}^{d} b_{jj} \quad \text{for } \boldsymbol{B} \in \mathbb{M}^d.$$

The theory can be developed using the standard trace, but additional complications arise.

The set $\mathbb{H}^d$ refers to the real-linear subspace of $d \times d$ Hermitian matrices in $\mathbb{M}^d$. For a matrix $\boldsymbol{A} \in \mathbb{H}^d$, we write $\lambda_{\min}(\boldsymbol{A})$ and $\lambda_{\max}(\boldsymbol{A})$ for the algebraic minimum and maximum eigenvalues. For each interval $I \subset \mathbb{R}$, we define the set of Hermitian matrices whose eigenvalues fall in that interval:

$$\mathbb{H}^d(I) := \{\boldsymbol{A} \in \mathbb{H}^d : [\lambda_{\min}(\boldsymbol{A}), \lambda_{\max}(\boldsymbol{A})] \subset I\}.$$

We also introduce the set $\mathbb{H}^d_+$ of $d \times d$ positive-semidefinite matrices and the set $\mathbb{H}^d_{++}$ of $d \times d$ positive-definite matrices. Curly inequalities refer to the positive-semidefinite order. For example, $\boldsymbol{A} \preccurlyeq \boldsymbol{B}$ means that $\boldsymbol{B} - \boldsymbol{A}$ is positive semidefinite.

Next, let us explain how to extend scalar functions to matrices. Recall that each Hermitian matrix $\boldsymbol{A} \in \mathbb{H}^d$ has a *spectral resolution*

$$\boldsymbol{A} = \sum\nolimits_{i=1}^{d} \lambda_i \boldsymbol{P}_i, \tag{2.1}$$

where $\lambda_1, \dots, \lambda_d$ are the eigenvalues of $\boldsymbol{A}$. The matrices $\boldsymbol{P}_1, \dots, \boldsymbol{P}_d$ are orthogonal projectors that satisfy the orthogonality relations

$$\boldsymbol{P}_i \boldsymbol{P}_j = \delta_{ij} \boldsymbol{P}_j \quad \text{and} \quad \sum\nolimits_{i=1}^{d} \boldsymbol{P}_i = \mathbf{I},$$

where δ_{ij} is the Kronecker delta and $\mathbf{I}$ is the identity matrix. One obtains a standard matrix function by applying a scalar function to the spectrum of a Hermitian matrix.

Definition 2.1 (Standard Matrix Function)**.** *Let $f : I \mapsto \mathbb{R}$ be a function on an interval I of the real line. Suppose that $\boldsymbol{A} \in \mathbb{H}^d(I)$ has the spectral decomposition* (2.1)*. Then*

$$f(\boldsymbol{A}) := \sum\nolimits_{i=1}^{d} f(\lambda_i) \boldsymbol{P}_i.$$

We use lowercase Roman and Greek letters to refer to standard matrix functions. When we apply a familiar real-valued function to an Hermitian matrix, we are referring to the associated standard matrix function. Bold capital letters such as $\boldsymbol{Y}, \boldsymbol{Z}$ denote general matrix functions that are not necessarily standard.

2.2 Subadditivity of Matrix Entropies

In this section, we provide an overview of the theory of matrix φ-entropies. At a high level, our approach has a strong parallel with the work of Boucheron et al. [BBLM05]. Nevertheless, there are interesting differences between the scalar and the matrix setting.

2.2.1 The Class of Matrix Entropies

First, we carve out a class of standard matrix functions that we can use to construct matrix entropies with the same subadditivity properties as their scalar counterparts.

Definition 2.2 (Φ_d Function Class)**.** *Let d be a natural number. The class Φ_d contains each function $\varphi : \mathbb{R}_+ \to \mathbb{R}$ that is either affine or else satisfies the following three conditions.*

1. *φ is convex and continuous at zero.*
2. *φ has two continuous derivatives on $\mathbb{R}_{++}$.*
3. *Define $\psi(t) = \varphi'(t)$ for $t \in \mathbb{R}_{++}$. The derivative $\mathsf{D}\psi$ of the standard matrix function $\psi : \mathbb{H}^d_{++} \to \mathbb{H}^d$ is an invertible linear operator on $\mathbb{H}^d_{++}$, and the map $\boldsymbol{A} \mapsto [\mathsf{D}\psi(\boldsymbol{A})]^{-1}$ is concave with respect to the semidefinite order on operators.*

The technical definitions that support requirement (3) appear in Section 3. For now, we just remark that the scalar equivalent of (3) is the statement that $t \mapsto [\varphi''(t)]^{-1}$ is concave on $\mathbb{R}_{++}$.

The class Φ_1 coincides with the Φ function class considered in [BBLM05]. It can be shown that $\Phi_{d+1} \subseteq \Phi_d$ for each natural number d, so it is appropriate to introduce the *class of matrix entropies*:

$$\Phi_\infty := \bigcap\nolimits_{d=1}^{\infty} \Phi_d$$

This class consists of scalar functions that satisfy the conditions of Definition 2.2 for an arbitrary choice of dimension d. Note that Φ_∞ is a convex cone: it contains all positive multiples and all finite sums of its elements.

In contrast to the scalar setting, it is quite hard to determine what functions are contained in Φ_∞. The main technical achievement of this paper is to demonstrate that the standard entropy and certain power functions belong to the matrix entropy class.

Theorem 2.3 (Elements of the Matrix Entropy Class)**.** *The following functions are members of the Φ_∞ class.*

1. *The standard entropy $t \mapsto t \log t$.*
2. *The power function $t \mapsto t^p$ for each $p \in [1, 2]$.*

The proof of Theorem 2.3 appears in Section 6. The statement about classical entropy can be obtained from standard results in matrix theory after some argument, but the result for power functions demands new effort. In fact, the claim about the classical entropy follows from the result for power functions because of the representation $t \log t = \lim_{p \downarrow 1} (t^p - t)/(p - 1)$.

See the independent work [Han13, Sec. 4] for closely related material. Very recently, Hansen and Zhang have developed an elegant characterization of the matrix entropy class [HZ14].

2.2.2 Matrix φ-Entropy

For each function in the matrix entropy class, we can introduce a generalized entropy functional that measures the amount of fluctuation in a random matrix.

Definition 2.4 (Matrix φ-Entropy)**.** *Let $\varphi \in \Phi_\infty$. Consider a random matrix $\boldsymbol{Z}$ taking values in $\mathbb{H}_+^d$, and assume that $\mathbb{E}\,\|\boldsymbol{Z}\| < \infty$ and $\mathbb{E}\,\|\varphi(\boldsymbol{Z})\| < \infty$. The matrix φ-entropy functional H_φ is*

$$H_\varphi(\boldsymbol{Z}) := \mathbb{E}\,\bar{\mathrm{tr}}\,\varphi(\boldsymbol{Z}) - \bar{\mathrm{tr}}\,\varphi(\mathbb{E}\,\boldsymbol{Z}). \tag{2.2}$$

Similarly, the conditional matrix φ-entropy functional is

$$H_\varphi(\boldsymbol{Z} \,|\, \mathscr{F}) := \mathbb{E}\left[\bar{\mathrm{tr}}\,\varphi(\boldsymbol{Z}) \,|\, \mathscr{F}\right] - \bar{\mathrm{tr}}\,\varphi\big(\mathbb{E}[\boldsymbol{Z} \,|\, \mathscr{F}]\big),$$

where $\mathscr{F}$ is a subalgebra of the master sigma algebra.

For each convex function φ, the trace function $\bar{\mathrm{tr}}\,\varphi : \mathbb{H}_+^d \to \mathbb{R}$ is also convex [Car10, Sec. 2.2]. Therefore, Jensen's inequality implies that the matrix φ-entropy is nonnegative:

$$H_\varphi(\boldsymbol{Z}) \geq 0.$$

For concreteness, here are some basic examples of matrix φ-entropy functionals.

$$\begin{aligned}
&H_\varphi(\boldsymbol{Z}) = \bar{\mathrm{tr}}\left[\mathbb{E}(\boldsymbol{Z} \log \boldsymbol{Z}) - (\mathbb{E}\,\boldsymbol{Z}) \log(\mathbb{E}\,\boldsymbol{Z})\right] && \text{when } \varphi(t) = t \log t.\\
&H_\varphi(\boldsymbol{Z}) = \bar{\mathrm{tr}}\left[\mathbb{E}(\boldsymbol{Z}^p) - (\mathbb{E}\,\boldsymbol{Z})^p\right] && \text{when } \varphi(t) = t^p \text{ for } p \in [1, 2].\\
&H_\varphi(\boldsymbol{Z}) = 0 && \text{when } \varphi \text{ is affine.}
\end{aligned}$$

2.2.3 Subadditivity of Matrix φ-Entropy

The key fact about matrix φ-entropies is that they satisfy a subadditivity property. Let $\boldsymbol{x} := (X_1, \dots, X_n)$ denote a vector of independent random variables taking values in a Polish space, and write $\boldsymbol{x}_{-i}$ for the random vector obtained by deleting the ith entry of $\boldsymbol{x}$.

$$\boldsymbol{x}_{-i} := (X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n).$$

Consider a positive-semidefinite random matrix $\boldsymbol{Z}$ that can be expressed as a measurable function of the random vector $\boldsymbol{x}$.

$$\boldsymbol{Z} := \boldsymbol{Z}(X_1, \dots, X_n) \in \mathbb{H}_+^d.$$

We instate the integrability conditions $\mathbb{E}\,\|\boldsymbol{Z}\| < \infty$ and $\mathbb{E}\,\|\varphi(\boldsymbol{Z})\| < \infty$.

Theorem 2.5 (Subadditivity of Matrix φ-Entropy)**.** *Fix a function $\varphi \in \Phi_\infty$. Under the prevailing assumptions,*

$$H_\varphi(\boldsymbol{Z}) \le \sum\nolimits_{i=1}^n \mathbb{E}\left[H_\varphi(\boldsymbol{Z} \,|\, \boldsymbol{x}_{-i})\right]. \tag{2.3}$$

Typically, we apply Theorem 2.5 by way of a corollary. Let $X_1', \dots, X_n'$ denote independent copies of $X_1, \dots, X_n$, and form the random matrix

$$\boldsymbol{Z}_i' := \boldsymbol{Z}(X_1, \dots, X_{i-1}, X_i', X_{i+1}, \dots, X_n) \in \mathbb{H}_+^d.$$

Then $\boldsymbol{Z}_i'$ and $\boldsymbol{Z}$ are independent and identically distributed, conditional on the sigma algebra generated by $\boldsymbol{x}_{-i}$. In particular, these two random matrices are exchangeable counterparts.

Corollary 2.6 (Entropy Bounds via Exchangeability)**.** *Fix a function $\varphi \in \Phi_\infty$, and write $\psi = \varphi'$. With the prevailing notation,*

$$H_\varphi(\boldsymbol{Z}) \le \frac{1}{2} \sum\nolimits_{i=1}^n \mathbb{E}\,\bar{\mathrm{tr}}\left[(\boldsymbol{Z} - \boldsymbol{Z}_i')(\psi(\boldsymbol{Z}) - \psi(\boldsymbol{Z}_i'))\right].$$

Theorem 2.5 and Corollary 2.6 are matrix counterparts of the foundational results from Boucheron et al. [BBLM05, Sec. 3], which establish that scalar φ-entropies satisfy a similar subadditivity property. We devote Section 4 to the proof of these results.

2.3 Some Matrix Concentration Inequalities

Using Corollary 2.6, we can derive concentration inequalities for random matrices. In contrast to some previous approaches to matrix concentration, we need to place some significant restrictions on the type of random matrices we consider.

Definition 2.7 (Invariance under Signed Permutation)**.** *A random matrix $\boldsymbol{Y} \in \mathbb{H}^d$ is invariant under signed permutation if we have the equality of distribution*

$$\boldsymbol{Y} \sim \boldsymbol{\Pi}^* \boldsymbol{Y} \boldsymbol{\Pi} \quad \textit{for each signed permutation } \boldsymbol{\Pi}.$$

A signed permutation $\boldsymbol{\Pi} \in \mathbb{M}^d$ *is a matrix with the properties that (i) each row and each column contains exactly one nonzero entry and (ii) the nonzero entries only take values $+1$ and -1.*

In particular, consider a random matrix that is invariant under orthogonal conjugation:

$$\boldsymbol{Y} \sim \boldsymbol{U}^* \boldsymbol{Y} \boldsymbol{U} \quad \text{for each orthogonal matrix } \boldsymbol{U}.$$

A matrix that satisfies this condition always verifies the requirement of Definition 2.7. Many classical ensembles, such as the GOE, satisfy this orthogonal invariance condition. Similar remarks apply to random matrices that are invariant under unitary conjugation.

2.3.1 A Bounded Difference Inequality

Let us present an exponential tail bound for a random matrix whose distribution is invariant under signed permutation.

Theorem 2.8 (Bounded Differences)**.** *Let $\boldsymbol{x} := (X_1, \dots, X_n)$ be a vector of independent random variables, and let $\boldsymbol{x}' := (X_1', \dots, X_n')$ be an independent copy of $\boldsymbol{x}$. Consider random matrices*

$$\boldsymbol{Y} := \boldsymbol{Y}(X_1, \dots, X_i, \dots, X_n) \in \mathbb{H}^d \quad \text{and}$$
$$\boldsymbol{Y}_i' := \boldsymbol{Y}(X_1, \dots, X_i', \dots, X_n) \in \mathbb{H}^d \quad \text{for } i = 1, \dots, n.$$

Assume that $\boldsymbol{Y}$ is invariant under signed permutation and that $\|\boldsymbol{Y}\|$ is bounded almost surely. Introduce the variance measure

$$V_{\boldsymbol{Y}} := \sup \left\| \mathbb{E}\left[\sum\nolimits_{i=1}^{n} (\boldsymbol{Y} - \boldsymbol{Y}_i')^2 \,\middle|\, \boldsymbol{x} \right] \right\|, \tag{2.4}$$

where the supremum occurs over all possible values of $\boldsymbol{x}$. For each $t \geq 0$,

$$\mathbb{P}\left\{ \lambda_{\max}(\boldsymbol{Y} - \mathbb{E}\,\boldsymbol{Y}) \geq t \right\} \leq d \cdot \mathrm{e}^{-t^2/(2V_{\boldsymbol{Y}})}, \quad \text{and}$$
$$\mathbb{P}\left\{ \lambda_{\min}(\boldsymbol{Y} - \mathbb{E}\,\boldsymbol{Y}) \leq -t \right\} \leq d \cdot \mathrm{e}^{-t^2/(2V_{\boldsymbol{Y}})}.$$

Theorem 2.8 follows from Corollary 2.6 with the choice $\varphi(t) = t \log t$. See Section 7 for the proof. This result can be viewed as a type of matrix bounded difference inequality. Closely related inequalities already appear in the literature; see [Tro12b, Cor. 7.5], [MJC+12, Cor. 11.1], and [PMT13, Cor. 4.1]. In fact, Theorem 2.8 is dominated by [PMT13, Cor. 4.1], which is not restricted to random matrices that are invariant under signed permutation.

2.3.2 Example: Sample covariance matrices

It may be helpful to sketch a short example that indicates the scope of Theorem 2.8. Consider a random vector of the form

$$\boldsymbol{w} := (\varepsilon_1 W_1, \varepsilon_2 W_2, \dots, \varepsilon_p W_p)^*$$

where (W_k) is an exchangeable family of random variables and (ε_k) is a sequence of independent Rademacher random variables. We also require that the random vector is bounded: $\|\boldsymbol{w}\|^2 \leq B$.

Let $\boldsymbol{w}_1, \dots, \boldsymbol{w}_n$ be iid copies of $\boldsymbol{w}$, and consider the sample covariance matrix

$$\boldsymbol{Y} := \frac{1}{n} \sum\nolimits_{i=1}^{n} \boldsymbol{w}_i \boldsymbol{w}_i^*.$$

Our assumptions on $\boldsymbol{w}$ ensure that $\boldsymbol{Y}$ is invariant under signed permutation and that $\|\boldsymbol{Y}\|$ is bounded. Note that $\mathbb{E}\,\boldsymbol{Y} = c\mathbf{I}$ for a positive number c. It is also easy to check that the variance measure (2.4) satisfies $V_{\boldsymbol{Y}} \leq 2B^2/n$. An application of Theorem 2.8 delivers

$$\mathbb{P}\left\{ \|\boldsymbol{Y} - c\mathbf{I}\| \geq t \right\} \leq 2d \cdot \mathrm{e}^{-nt^2/(4B^2)}.$$

The bound is informative when $c^2 > t^2 > 4B^2 \log(2d)/n$. In other words, the number n of samples should satisfy $n > 4B^2 \log(2d)/c^2$. Modulo constants, this estimate cannot be improved when $\boldsymbol{w}$ has the uniform distribution on $\{\pm\mathbf{e}_1, \dots, \pm\mathbf{e}_p\}$, the set of signed standard basis vectors.

The main result of Rudelson's paper [Rud99] is a concentration bound for sample covariance matrices based on the noncommutative Khintchine inequality [LP86]. Rudelson allows any bounded random vector $\boldsymbol{w}$ with a scalar covariance matrix, and he achieves the same result derived here.

2.3.3 Matrix Moment Bounds

We can also establish moment inequalities for a random matrix whose distribution is invariant under signed permutation.

Theorem 2.9 (Matrix Moment Bound)**.** *Fix a number $q \in \{2, 3, 4, \dots\}$. Let $\boldsymbol{x} := (X_1, \dots, X_n)$ be a vector of independent random variables, and let $\boldsymbol{x}' := (X_1', \dots, X_n')$ be an independent copy of $\boldsymbol{x}$. Consider positive-semidefinite random matrices*

$$\begin{aligned} \boldsymbol{Y} &:= \boldsymbol{Y}(X_1, \dots, X_i, \dots, X_n) \in \mathbb{H}_+^d \quad \text{and} \\ \boldsymbol{Y}_i' &:= \boldsymbol{Y}(X_1, \dots, X_i', \dots, X_n) \in \mathbb{H}_+^d \quad \text{for } i = 1, \dots, n. \end{aligned}$$

Assume that $\boldsymbol{Y}$ is invariant under signed permutation and that $\mathbb{E}(\|\boldsymbol{Y}\|^q) < \infty$. Suppose that there is a constant $c \geq 0$ with the property

$$\boldsymbol{V}_{\boldsymbol{Y}} := \mathbb{E}\left[\sum\nolimits_{i=1}^n (\boldsymbol{Y} - \boldsymbol{Y}_i')^2 \,\middle|\, \boldsymbol{x}\right] \preccurlyeq c\,\boldsymbol{Y}. \tag{2.5}$$

Then the random matrix $\boldsymbol{Y}$ satisfies the moment inequality

$$[\mathbb{E}\,\bar{\mathrm{tr}}(\boldsymbol{Y}^q)]^{1/q} \leq \mathbb{E}\,\bar{\mathrm{tr}}\,\boldsymbol{Y} + \frac{q-1}{2} \cdot c.$$

Theorem 2.9 follows from Corollary 2.6 with the choice $\varphi(t) = t^{q/(q-1)}$. See Section 8 for the proof. This result can be regarded as a matrix extension of a moment inequality for real random variables [BBLM05, Cor. 1]. The paper [PMT13] contains similar moment inequalities for random matrices that need not satisfy the condition of signed permutation invariance. See also [JX03, JX08, JZ11, MJC+12].

2.4 Generalized Subadditivity of Matrix φ-Entropy

Theorem 2.5 is the shadow of a more sophisticated subadditivity property. We outline the simplest form of this more general result. See the lecture notes of Carlen [Car10] for more background on the topics in this section.

We work in the $*$-algebra $\mathbb{M}^d$ of $d \times d$ complex matrices, equipped with the conjugate transpose operation $*$ and the normalized trace inner product $\langle \boldsymbol{A}, \boldsymbol{B} \rangle := \bar{\mathrm{tr}}(\boldsymbol{A}^*\boldsymbol{B})$. We say that a subspace $\mathfrak{A} \subset \mathbb{M}^d$ is a *$*$-subalgebra* when $\mathfrak{A}$ contains the identity matrix, $\mathfrak{A}$ is closed under matrix multiplication, and $\mathfrak{A}$ is closed under conjugate transposition. In other terms, $\mathbf{I} \in \mathfrak{A}$ and $\boldsymbol{AB} \in \mathfrak{A}$ and $\boldsymbol{A}^* \in \mathfrak{A}$ whenever $\boldsymbol{A}, \boldsymbol{B} \in \mathfrak{A}$.

In this setting, there is an elegant notion of conditional expectation. The orthogonal projector $\mathbb{E}_{\mathfrak{A}} : \mathbb{M}^d \to \mathfrak{A}$ onto the $*$-subalgebra $\mathfrak{A}$ is called the *conditional expectation* with respect to the $*$-subalgebra. For $*$-subalgebras $\mathfrak{A}$ and $\mathfrak{B}$, we say that the conditional expectations $\mathbb{E}_{\mathfrak{A}}$ and $\mathbb{E}_{\mathfrak{B}}$ *commute* when

$$(\mathbb{E}_{\mathfrak{A}}\,\mathbb{E}_{\mathfrak{B}})(\boldsymbol{M}) = (\mathbb{E}_{\mathfrak{B}}\,\mathbb{E}_{\mathfrak{A}})(\boldsymbol{M}) \quad \text{for every } \boldsymbol{M} \in \mathbb{M}^d.$$

This construction generalizes the concept of independence in a probability space.

We can define the matrix φ-entropy conditional on a $*$-subalgebra $\mathfrak{A}$:

$$H_\varphi(\boldsymbol{A} \,|\, \mathfrak{A}) := \bar{\mathrm{tr}}[\varphi(\boldsymbol{A}) - \varphi(\mathbb{E}_{\mathfrak{A}}\,\boldsymbol{A})] \quad \text{for } \boldsymbol{A} \in \mathbb{H}_+^d.$$

Note that $\bar{\mathrm{tr}}(\mathbb{E}_{\mathfrak{A}}\,\boldsymbol{A}) = \bar{\mathrm{tr}}\,\boldsymbol{A}$ for each matrix $\boldsymbol{A}$ in $\mathbb{H}_+^d$, so we do not need to include a conditional expectation in the leading term. Let $\mathfrak{A}_1, \dots, \mathfrak{A}_n$ be $*$-subalgebras whose conditional expectations commute. Then we can extend the definition of the matrix φ-entropy to read

$$H_\varphi(\boldsymbol{A} \,|\, \mathfrak{A}_1, \dots, \mathfrak{A}_n) := \bar{\mathrm{tr}}[\varphi(\boldsymbol{A}) - \varphi(\mathbb{E}_{\mathfrak{A}_1} \cdots \mathbb{E}_{\mathfrak{A}_n}\,\boldsymbol{A})] \quad \text{for } \boldsymbol{A} \in \mathbb{H}_+^d.$$

Because of commutativity, the order of the conditional expectations has no effect on the calculation. It turns out that matrix φ-entropy admits the following subadditivity property.

Theorem 2.10 (Subaddivity of Matrix φ-Entropy II)**.** *Fix a function $\varphi \in \Phi_\infty$. Let $\mathfrak{A}_1, \dots, \mathfrak{A}_n$ be $*$-subalgebras of $\mathbb{M}^d$ whose conditional expectations commute. Then*

$$H_\varphi(\boldsymbol{A} \,|\, \mathfrak{A}_1, \dots, \mathfrak{A}_n) \le \sum\nolimits_{i=1}^n H_\varphi(\boldsymbol{A} \,|\, \mathfrak{A}_i) \quad \textit{for } \boldsymbol{A} \in \mathbb{H}_+^d.$$

We omit the proof of this result. The argument involves considerations similar with Theorem 2.5, but it requires an extra dose of operator theory. The work in this paper already addresses the more challenging aspects of the proof. Note that the case $\varphi : t \mapsto t \log t$ is essentially a consequence of the classical results in [Lin73].

Theorem 2.10 can be seen as a formal extension of the subadditivity of matrix φ-entropy expressed in Theorem 2.5. To see why, let $\Omega := \Omega_1 \times \cdots \times \Omega_n$ be a product probability space. The space $L_2(\Omega; \mathbb{M}^d)$ of random matrices is a $*$-algebra with the normalized trace functional $\mathbb{E}\,\bar{\mathrm{tr}}$. For each $i = 1, \dots, n$, we can form a $*$-subalgebra $\mathfrak{A}_i$ consisting of the random matrices that do not depend on the ith factor Ω_i of the product. The conditional expectation $\mathbb{E}_{\mathfrak{A}_i}$ simply integrates out the ith random variable. By independence, the family of conditional expectations $\mathbb{E}_{\mathfrak{A}_1}, \dots, \mathbb{E}_{\mathfrak{A}_n}$ commutes. Using this dictionary, compare the statement of Theorem 2.10 with Theorem 2.5.

3 Operators and Functions acting on Matrices

This work involves a substantial amount of operator theory. This section contains a short treatment of the basic facts. See [Bha97, Bha07] for a more complete introduction.

3.1 Linear Operators on Matrices

Let $\mathbb{C}^d$ be the complex Hilbert space of dimension d, equipped with the standard inner product $\langle \boldsymbol{a}, \boldsymbol{b} \rangle := \boldsymbol{a}^* \boldsymbol{b}$. We usually identify $\mathbb{M}^d$ with $\mathbb{B}(\mathbb{C}^d)$, the complex Banach space of linear operators acting on $\mathbb{C}^d$, equipped with the ℓ_2 operator norm $\|\cdot\|$.

We can also endow $\mathbb{M}^d$ with the normalized trace inner product $\langle \boldsymbol{A}, \boldsymbol{B} \rangle := \bar{\mathrm{tr}}(\boldsymbol{A}^* \boldsymbol{B})$ to form a Hilbert space. As a Hilbert space, $\mathbb{M}^d$ is isometrically isomorphic with $\mathbb{C}^{d^2}$. Let $\mathbb{B}(\mathbb{M}^d)$ denote the complex Banach space of linear operators that map the Hilbert space $\mathbb{M}^d$ into itself, equipped with the induced operator norm. The Banach space $\mathbb{B}(\mathbb{M}^d)$ is isometrically isomorphic with the Banach space $\mathbb{M}^{d^2}$.

As a consequence of this construction, every concept from matrix analysis has an immediate analog for linear operators on matrices. An operator $\mathsf{T} \in \mathbb{B}(\mathbb{M}^d)$ is *self-adjoint* when

$$\langle \boldsymbol{A}, \mathsf{T}(\boldsymbol{B}) \rangle = \langle \mathsf{T}(\boldsymbol{A}), \boldsymbol{B} \rangle \quad \text{for all } \boldsymbol{A}, \boldsymbol{B} \in \mathbb{B}(\mathbb{M}^d).$$

A self-adjoint operator $\mathsf{T} \in \mathbb{B}(\mathbb{M}^d)$ is *positive semidefinite* when

$$\langle \boldsymbol{A}, \mathsf{T}(\boldsymbol{A}) \rangle \ge 0 \quad \text{for all } \boldsymbol{A} \in \mathbb{M}^d.$$

For self-adjoint operators $\mathsf{S}, \mathsf{T} \in \mathbb{B}(\mathbb{M}^d)$, the notation $\mathsf{S} \preccurlyeq \mathsf{T}$ means that $\mathsf{T} - \mathsf{S}$ is positive semidefinite.

Each self-adjoint matrix operator $\mathsf{T} \in \mathbb{B}(\mathbb{M}^d)$ has a spectral resolution of the form

$$\mathsf{T} = \sum\nolimits_{i=1}^{d^2} \lambda_i \mathsf{P}_i \tag{3.1}$$

where $\lambda_1, \dots, \lambda_{d^2}$ are the eigenvalues of T and the spectral projectors $\mathsf{P}_1, \dots, \mathsf{P}_{d^2}$ are positive-semidefinite operators that satisfy

$$\mathsf{P}_i \mathsf{P}_j = \delta_{ij} \mathsf{P}_j \quad \text{and} \quad \sum\nolimits_{i=1}^{d^2} \mathsf{P}_i = \mathsf{I},$$

where δ_{ij} is the Kronecker delta and I is the identity operator. As in the matrix case, a self-adjoint operator with nonnegative eigenvalues is the same thing as a positive-semidefinite operator.

We can extend a scalar function $f : I \to \mathbb{R}$ on an interval I of the real line to obtain a standard operator function. Indeed, if T has the spectral resolution (3.1) and the eigenvalues of T fall in the interval I, we define

$$f(\mathsf{T}) := \sum\nolimits_{i=1}^{d^2} f(\lambda_i) \mathsf{P}_i.$$

This definition, of course, parallels the definition for matrices.

3.2 Monotonicity and Convexity

Let X and Y be sets of self-adjoint operators, such as $\mathbb{H}^d(I)$ or the set of self-adjoint operators in $\mathbb{B}(\mathbb{M}^d)$. We can introduce notions of monotonicity and convexity for a general function $\Psi : X \to Y$ using the semidefinite order on the spaces of operators.

Definition 3.1 (Monotone Operator-Valued Function)**.** *The function $\Psi : X \to Y$ is* monotone *when*

$$\mathsf{S} \preccurlyeq \mathsf{T} \quad \Longrightarrow \quad \Psi(\mathsf{S}) \preccurlyeq \Psi(\mathsf{T}) \quad \textit{for all } \mathsf{S}, \mathsf{T} \in X.$$

Definition 3.2 (Convex Operator-Valued Function)**.** *The function $\Psi : X \to Y$ is* convex *when X is a convex set and*

$$\Psi(\alpha \mathsf{S} + \bar{\alpha} \mathsf{T}) \preccurlyeq \alpha \cdot \Psi(\mathsf{S}) + \bar{\alpha} \cdot \Psi(\mathsf{T}) \quad \textit{for all } \alpha \in [0, 1] \textit{ and all } \mathsf{S}, \mathsf{T} \in X.$$

We have written $\bar{\alpha} := 1 - \alpha$. The function Ψ is concave *when $-\Psi$ is convex.*

The convexity of an operator-valued function Ψ is equivalent with a Jensen-type relation:

$$\Psi(\mathbb{E}\, \mathsf{X}) \preccurlyeq \mathbb{E}\, \Psi(\mathsf{X}) \tag{3.2}$$

whenever X is an integrable random operator taking values in X.

In particular, we can apply these definitions to standard matrix and operator functions. Let I be an interval of the real line. We say that the function $f : I \to \mathbb{R}$ is *operator monotone* when the lifted map $f : \mathbb{H}^d(I) \to \mathbb{H}^d$ is monotone for each natural number d. Likewise, the function $f : I \to \mathbb{R}$ is *operator convex* when the lifted map $f : \mathbb{H}^d(I) \to \mathbb{H}^d$ is convex for each natural number d.

Although scalar monotonicity and convexity are quite common, they are much rarer in the matrix setting [Bha97, Chap. 4]. For present purposes, we note that the power functions $t \mapsto t^p$ with $p \in [0, 1]$ are operator monotone and operator concave. The power functions $t \mapsto t^p$ with $p \in [1, 2]$ and the standard entropy $t \mapsto t \log t$ are all operator convex.

3.3 The Derivative of a Vector-Valued Function

The definition of the Φ_∞ function class involves a requirement that a certain standard matrix function is differentiable. For completeness, we include the background needed to interpret this condition.

Definition 3.3 (Derivative of a Vector-Valued Function)**.** *Let X and Y be Banach spaces, and let U be an open subset of X. A function $\boldsymbol{F} : U \to Y$ is* differentiable *at a point $\boldsymbol{A} \in U$ if there exists a bounded linear operator $\mathsf{T} : X \to Y$ for which*

$$\lim_{\boldsymbol{B} \to \mathbf{0}} \frac{\|\boldsymbol{F}(\boldsymbol{A}+\boldsymbol{B}) - \boldsymbol{F}(\boldsymbol{A}) - \mathsf{T}(\boldsymbol{B})\|_Y}{\|\boldsymbol{B}\|_X} = 0.$$

When $\boldsymbol{F}$ is differentiable at $\boldsymbol{A}$, the operator T is called the derivative *of $\boldsymbol{F}$ at $\boldsymbol{A}$, and we define $\mathsf{D}\boldsymbol{F}(\boldsymbol{A}) := \mathsf{T}$.*

The derivative and the directional derivative have the following relationship:

$$\frac{\mathrm{d}}{\mathrm{d}s}\boldsymbol{F}(\boldsymbol{A}+s\boldsymbol{B})\Big|_{s=0} = \mathsf{D}\boldsymbol{F}(\boldsymbol{A})(\boldsymbol{B}). \tag{3.3}$$

In Section 6.2, we present an explicit formula for the derivative of a standard matrix function.

4 Subadditivity of Matrix φ-Entropy

In this section, we establish Theorem 2.5, which states that the matrix φ-entropy is subadditive for every function in the Φ_∞ class. This result depends on a variational representation for the matrix φ-entropy that appears in Section 4.1. We use the variational formula to derive a Jensen-type inequality in Section 4.2. The proof of Theorem 2.5 appears in Section 4.3.

4.1 Representation of Matrix φ-Entropy as a Supremum

The fundamental fact behind the subadditivity theorem is a representation of the matrix φ-entropy as a supremum of affine functions.

Lemma 4.1 (Supremum Representation for Entropy)**.** *Fix a function $\varphi \in \Phi_\infty$, and introduce the scalar derivative $\psi = \varphi'$. Suppose that $\boldsymbol{Z}$ is a random positive-semidefinite matrix for which $\|\boldsymbol{Z}\|$ and $\|\varphi(\boldsymbol{Z})\|$ are integrable. Then*

$$H_\varphi(\boldsymbol{Z}) = \sup_{\boldsymbol{T}} \mathbb{E}\,\bar{\mathrm{tr}}\left[(\psi(\boldsymbol{T}) - \psi(\mathbb{E}\,\boldsymbol{T}))(\boldsymbol{Z}-\boldsymbol{T}) + \varphi(\boldsymbol{T}) - \varphi(\mathbb{E}\,\boldsymbol{T})\right]. \tag{4.1}$$

The range of the supremum contains each random positive-definite matrix $\boldsymbol{T}$ for which $\|\boldsymbol{T}\|$ and $\|\varphi(\boldsymbol{T})\|$ are integrable. In particular, the matrix φ-entropy H_φ can be written in the dual form

$$H_\varphi(\boldsymbol{Z}) = \sup_{\boldsymbol{T}} \mathbb{E}\,\bar{\mathrm{tr}}\left[\boldsymbol{\Upsilon}_1(\boldsymbol{T}) \cdot \boldsymbol{Z} + \boldsymbol{\Upsilon}_2(\boldsymbol{T})\right], \tag{4.2}$$

where $\boldsymbol{\Upsilon}_i : \mathbb{H}_+^d \to \mathbb{H}^d$ for $i = 1, 2$.

This result implies that H_φ is a convex function on the space of random positive-semidefinite matrices. The dual representation of H_φ is well suited for establishing a form of Jensen's inequality, Lemma 4.3, which is the main ingredient in the proof of the subadditivity property, Theorem 2.5.

It may be valuable to see some particular instances of the dual representation of the matrix φ-entropy:

$$H_\varphi(\boldsymbol{Z}) = \sup_{\boldsymbol{T}} \mathbb{E}\,\bar{\mathrm{tr}}\left[(\log \boldsymbol{T} - \log(\mathbb{E}\,\boldsymbol{T})) \cdot \boldsymbol{Z}\right] \qquad \text{when } \varphi(t) = t\log t.$$

$$H_\varphi(\boldsymbol{Z}) = \sup_{\boldsymbol{T}} \mathbb{E}\,\bar{\mathrm{tr}}\left[p(\boldsymbol{T}^{p-1} - (\mathbb{E}\,\boldsymbol{T})^{p-1}) \cdot \boldsymbol{Z} - (p-1)(\boldsymbol{T}^p - (\mathbb{E}\,\boldsymbol{T})^p)\right] \quad \text{when } \varphi(t) = t^p \text{ for } p \in [1,2].$$

The first formula is the matrix version of a well-known variational principle for the classical entropy, cf. [BBLM05, p. 525]. In the matrix setting, this result can be derived from the joint convexity of quantum relative entropy [Lin73].

4.1.1 The Convexity Lemma

To establish the variational formula, we require a convexity result for a quadratic form connected with the function φ.

Lemma 4.2. *Fix a function $\varphi \in \Phi_\infty$, and let $\psi = \varphi'$. Suppose that $\boldsymbol{Y}$ is a random matrix taking values in $\mathbb{H}^d_+$, and let $\boldsymbol{K}$ be a random matrix taking values in $\mathbb{M}^d$. Assume that $\|\boldsymbol{Y}\|$ and $\|\boldsymbol{K}\|$ are integrable. Then*

$$\mathbb{E}\,\langle \boldsymbol{K},\, \mathsf{D}\psi(\boldsymbol{Y})(\boldsymbol{K})\rangle \geq \langle (\mathbb{E}\,\boldsymbol{K}),\, \mathsf{D}\psi(\mathbb{E}\,\boldsymbol{Y})(\mathbb{E}\,\boldsymbol{K})\rangle$$

Proof. The proof hinges on a basic convexity property of quadratic forms. Define a map that takes a matrix $\boldsymbol{A}$ in $\mathbb{H}^d$ and a positive-definite operator T on $\mathbb{M}^d$ to a nonnegative number:

$$\mathscr{Q} : (\boldsymbol{A}, \mathsf{T}) \mapsto \left\langle \boldsymbol{A},\, \mathsf{T}^{-1}(\boldsymbol{A})\right\rangle.$$

We assert that the function $\mathscr{Q}$ is convex. Indeed, the same result is well known when $\boldsymbol{A}$ and T are replaced by a vector and a positive-definite matrix [Bha07, Exer. 1.5.1], and the extension is immediate from the isometric isomorphism between operators and matrices.

Recall that the Φ_∞ class requires $\boldsymbol{A} \mapsto [\mathsf{D}\psi(\boldsymbol{A})]^{-1}$ to be a concave map on $\mathbb{H}^d_{++}$. With these observations at hand, we can make the following calculation:

$$\begin{aligned}
\mathbb{E}\,\langle \boldsymbol{K},\, \mathsf{D}\psi(\boldsymbol{Y})(\boldsymbol{K})\rangle &= \mathbb{E}\left\langle \boldsymbol{K},\, ([\mathsf{D}\psi(\boldsymbol{Y})]^{-1})^{-1}(\boldsymbol{K})\right\rangle \\
&\geq \left\langle (\mathbb{E}\,\boldsymbol{K}),\, (\mathbb{E}[\mathsf{D}\psi(\boldsymbol{Y})]^{-1})^{-1}(\mathbb{E}\,\boldsymbol{K})\right\rangle \\
&\geq \left\langle (\mathbb{E}\,\boldsymbol{K}),\, ([\mathsf{D}\psi(\mathbb{E}\,\boldsymbol{Y})]^{-1})^{-1}(\mathbb{E}\,\boldsymbol{K})\right\rangle \\
&= \langle (\mathbb{E}\,\boldsymbol{K}),\, \mathsf{D}\psi(\mathbb{E}\,\boldsymbol{Y})(\mathbb{E}\,\boldsymbol{K})\rangle\,.
\end{aligned}$$

We obtain the second relation when we apply Jensen's inequality to the convex function $\mathscr{Q}$. The third relation depends on the semidefinite Jensen inequality (3.2) for the concave function $\boldsymbol{A} \mapsto [\mathsf{D}\psi(\boldsymbol{A})]^{-1}$, coupled with the fact [Bha97, Prop. V.1.6] that the operator inverse reverses the semidefinite order. □

4.1.2 Proof of Lemma 4.1

The argument parallels the proof of [BBLM05, Lem. 1]. We begin with some reductions. The case where φ is an affine function is immediate, so we may require the derivative $\psi = \varphi'$ to be non-constant. By approximation, we may also assume that the random matrix $\boldsymbol{Z}$ is strictly positive definite.

[Indeed, since φ is continuous on $\mathbb{R}_+$, the Dominated Convergence Theorem implies that the matrix φ-entropy H_φ is continuous on the set containing each positive-semidefinite random matrix $\boldsymbol{Y}$ where $\|\boldsymbol{Y}\|$ and $\|\varphi(\boldsymbol{Y})\|$ are integrable. Therefore, we can approximate a positive-semidefinite random matrix $\boldsymbol{Z}$ by a sequence $\{\boldsymbol{Y}_n\}$ of positive-definite random matrices where $\boldsymbol{Y}_n \to \boldsymbol{Z}$ and be confident that $H_\varphi(\boldsymbol{Y}_n) \to H_\varphi(\boldsymbol{Z})$.]

When $\boldsymbol{T} = \boldsymbol{Z}$, the argument of the supremum in (4.1) equals $H_\varphi(\boldsymbol{Z})$. Therefore, our burden is to verify the inequality

$$H_\varphi(\boldsymbol{Z}) \geq \mathbb{E}\,\bar{\mathrm{tr}}\left[(\psi(\boldsymbol{T}) - \psi(\mathbb{E}\,\boldsymbol{T}))(\boldsymbol{Z} - \boldsymbol{T}) + \mathbb{E}\,\varphi(\boldsymbol{T}) - \varphi(\mathbb{E}\,\boldsymbol{T})\right] \tag{4.3}$$

for each random positive-definite matrix $\boldsymbol{T}$ that satisfies the same integrability requirements as $\boldsymbol{Z}$. For simplicity, we assume that the eigenvalues of both $\boldsymbol{Z}$ and $\boldsymbol{T}$ are bounded and bounded away from zero. See Appendix A for the extension to the general case.

We use an interpolation argument to establish (4.3). Define the family of random matrices

$$\boldsymbol{T}_s := (1-s)\cdot\boldsymbol{Z} + s\cdot\boldsymbol{T} \quad \text{for } s\in[0,1].$$

Introduce the real-valued function

$$F(s) := \mathbb{E}\,\bar{\mathrm{tr}}\left[(\psi(\boldsymbol{T}_s) - \psi(\mathbb{E}\,\boldsymbol{T}_s))\cdot(\boldsymbol{Z}-\boldsymbol{T}_s)\right] + H_\varphi(\boldsymbol{T}_s).$$

Observe that $F(0) = H_\varphi(\boldsymbol{Z})$, while $F(1)$ coincides with the right-hand side of (4.3). Therefore, to establish (4.3), it suffices to show that the function $F(s)$ is weakly decreasing on the interval $[0,1]$.

We intend to prove that $F'(s)\le 0$ for $s\in[0,1]$. Since $\boldsymbol{Z}-\boldsymbol{T}_s = -s\cdot(\boldsymbol{T}-\boldsymbol{Z})$, we can rewrite the function F in the form

$$F(s) = -s\cdot\mathbb{E}\,\bar{\mathrm{tr}}\left[(\psi(\boldsymbol{T}_s) - \psi(\mathbb{E}\,\boldsymbol{T}_s))\cdot(\boldsymbol{T}-\boldsymbol{Z})\right] + \mathbb{E}\,\bar{\mathrm{tr}}\left[\varphi(\boldsymbol{T}_s) - \varphi(\mathbb{E}\,\boldsymbol{T}_s))\right]. \tag{4.4}$$

We differentiate the function F to obtain

$$\begin{aligned} F'(s) = -s\cdot\mathbb{E}\,\bar{\mathrm{tr}}\left[\mathsf{D}\psi(\boldsymbol{T}_s)(\boldsymbol{T}-\boldsymbol{Z})\cdot(\boldsymbol{T}-\boldsymbol{Z})\right] + s\cdot\bar{\mathrm{tr}}\left[\mathsf{D}\psi(\mathbb{E}\,\boldsymbol{T}_s)(\mathbb{E}(\boldsymbol{T}-\boldsymbol{Z}))\cdot(\mathbb{E}(\boldsymbol{T}-\boldsymbol{Z}))\right] \\ - \mathbb{E}\,\bar{\mathrm{tr}}\left[(\psi(\boldsymbol{T}_s) - \psi(\mathbb{E}\,\boldsymbol{T}_s))\cdot(\boldsymbol{T}-\boldsymbol{Z})\right] + \mathbb{E}\,\bar{\mathrm{tr}}\left[(\psi(\boldsymbol{T}_s) - \psi(\mathbb{E}\,\boldsymbol{T}_s))\cdot(\boldsymbol{T}-\boldsymbol{Z})\right]. \end{aligned} \tag{4.5}$$

To handle the first term in (4.4), we applied the product rule, the rule (3.3) for directional derivatives, and the expression $\mathrm{d}\boldsymbol{T}_s/\mathrm{d}s = \boldsymbol{T}-\boldsymbol{Z}$. We used the identity $\mathsf{D}\,\mathrm{tr}\,\varphi(\boldsymbol{A}) = \psi(\boldsymbol{A})$ to differentiate the second term. We also relied on the Dominated Convergence Theorem to pass derivatives through expectations, which is justified because φ and ψ are continuously differentiable on $\mathbb{H}^d_{++}$ and the eigenvalues of the random matrices are bounded and bounded away from zero. Now, the last two terms in (4.5) cancel, and we can rewrite the first two terms using the trace inner product:

$$F'(s) = s\cdot\left[\,\langle(\mathbb{E}(\boldsymbol{T}-\boldsymbol{Z})),\,\mathsf{D}\psi(\mathbb{E}\,\boldsymbol{T}_s)(\mathbb{E}(\boldsymbol{T}-\boldsymbol{Z}))\rangle - \mathbb{E}\,\langle(\boldsymbol{T}-\boldsymbol{Z}),\,\mathsf{D}\psi(\boldsymbol{T}_s)(\boldsymbol{T}-\boldsymbol{Z})\rangle\,\right].$$

Invoke Lemma 4.2 to conclude that $F'(s)\le 0$ for $s\in[0,1]$.

4.2 A Conditional Jensen Inequality

The variational inequality in Lemma 4.1 leads directly to a Jensen inequality for the matrix φ-entropy.

Lemma 4.3 (Conditional Jensen Inequality)**.** *Fix a function $\varphi\in\Phi_\infty$. Suppose that (X_1,X_2) is a pair of independent random variables taking values in a Polish space, and let $\boldsymbol{Z}=\boldsymbol{Z}(X_1,X_2)$ be a random positive-semidefinite matrix for which $\|\boldsymbol{Z}\|$ and $\|\varphi(\boldsymbol{Z})\|$ are integrable. Then*

$$H_\varphi\left(\mathbb{E}_1\,\boldsymbol{Z}\right) \le \mathbb{E}_1\,H_\varphi\left(\boldsymbol{Z}\,|\,X_1\right),$$

where $\mathbb{E}_1$ is the expectation with respect to the first variable X_1.

Proof. Let $\mathbb{E}_2$ denote the expectation with respect to the second variable X_2. The result is a simple consequence of the dual representation (4.2) of the matrix φ-entropy:

$$H_\varphi\left(\mathbb{E}_1\,\boldsymbol{Z}\right) = \sup_{\boldsymbol{T}}\ \mathbb{E}_2\,\bar{\mathrm{tr}}\left[\boldsymbol{\Upsilon}_1\big(\boldsymbol{T}(X_2)\big)\cdot(\mathbb{E}_1\,\boldsymbol{Z}) + \boldsymbol{\Upsilon}_2\big(\boldsymbol{T}(X_2)\big)\right]. \tag{4.6}$$

We have written $\boldsymbol{T}(X_2)$ to emphasize that this matrix depends only on the randomness in X_2. To control (4.6), we apply Fubini's theorem to interchange the order of $\mathbb{E}_1$ and

$\mathbb{E}_2$, and then we exploit the convexity of the supremum to draw out the expectation $\mathbb{E}_1$.

$$\begin{aligned} H_\varphi(\mathbb{E}_1 \boldsymbol{Z}) &= \sup_{\boldsymbol{T}} \mathbb{E}_1 \mathbb{E}_2 \bar{\operatorname{tr}}[\boldsymbol{\Upsilon}_1(\boldsymbol{T}(X_2)) \cdot \boldsymbol{Z} + \boldsymbol{\Upsilon}_2(\boldsymbol{T}(X_2))] \\ &\le \mathbb{E}_1 \sup_{\boldsymbol{T}} \mathbb{E}_2 \bar{\operatorname{tr}}[\boldsymbol{\Upsilon}_1(\boldsymbol{T}(X_2)) \cdot \boldsymbol{Z} + \boldsymbol{\Upsilon}_2(\boldsymbol{T}(X_2)] \\ &= \mathbb{E}_1 \sup_{\boldsymbol{T}} \mathbb{E}\left[\bar{\operatorname{tr}}[\boldsymbol{\Upsilon}_1(\boldsymbol{T}(X_2)) \cdot \boldsymbol{Z} + \boldsymbol{\Upsilon}_2(\boldsymbol{T}(X_2)] \,|\, X_1\right] \\ &= \mathbb{E}_1 H_\varphi(\boldsymbol{Z} \,|\, X_1). \end{aligned}$$

The last relation is the duality formula (4.2), applied conditionally. □

4.3 Proof of Theorem 2.5

We are now prepared to establish the main result on subadditivity of matrix φ-entropy. This theorem is a direct consequence of the conditional Jensen inequality, Lemma 4.3. In this argument, we write $\mathbb{E}_i$ for the expectation with respect to the variable X_i. Using the notation from Section 2.2.3, we see that $\mathbb{E}_i = \mathbb{E}[\cdot \,|\, \boldsymbol{x}_{-i}]$.

First, separate the matrix φ-entropy into two parts by adding and subtracting terms:

$$\begin{aligned} H_\varphi(\boldsymbol{Z}) &= \mathbb{E}\,\bar{\operatorname{tr}}[\varphi(\boldsymbol{Z}) - \varphi(\mathbb{E}_1 \boldsymbol{Z}) + \varphi(\mathbb{E}_1 \boldsymbol{Z}) - \varphi(\mathbb{E}\,\boldsymbol{Z})]. \\ &= \mathbb{E}\left[\mathbb{E}_1 \bar{\operatorname{tr}}[\varphi(\boldsymbol{Z}) - \varphi(\mathbb{E}_1 \boldsymbol{Z})]\right] + \mathbb{E}\,\bar{\operatorname{tr}}[\varphi(\mathbb{E}_1 \boldsymbol{Z}) - \varphi(\mathbb{E}\,\mathbb{E}_1 \boldsymbol{Z})]. \end{aligned} \tag{4.7}$$

We can rewrite this expression as

$$\begin{aligned} H_\varphi(\boldsymbol{Z}) &= \mathbb{E}\, H_\varphi(\boldsymbol{Z} \,|\, \boldsymbol{x}_{-1}) + H_\varphi(\mathbb{E}_1 \boldsymbol{Z}) \\ &\le \mathbb{E}\, H_\varphi(\boldsymbol{Z} \,|\, \boldsymbol{x}_{-1}) + \mathbb{E}_1 H_\varphi(\boldsymbol{Z} \,|\, X_1). \end{aligned} \tag{4.8}$$

The inequality follows from Lemma 4.3 because $\boldsymbol{Z} = \boldsymbol{Z}(X_1, \boldsymbol{x}_{-1})$ where X_1 and $\boldsymbol{x}_{-1}$ are independent random variables.

The first term on the right-hand side of (4.8) coincides with the first summand on the right-hand side of the subadditivity inequality (2.3). We must argue that the remaining summands are contained in the second term on the right-hand side of (4.8). Repeating the argument in the previous paragraph, conditioning on X_1, we obtain

$$H_\varphi(\boldsymbol{Z} \,|\, X_1) \le \mathbb{E}\left[H_\varphi(\boldsymbol{Z} \,|\, \boldsymbol{x}_{-2}) \,|\, X_1\right] + \mathbb{E}_2 H_\varphi(\boldsymbol{Z} \,|\, X_1, X_2).$$

Substituting this expression into (4.8), we obtain

$$H_\varphi(\boldsymbol{Z}) \le \sum\nolimits_{i=1}^{2} \mathbb{E}\, H_\varphi(\boldsymbol{Z} \,|\, \boldsymbol{x}_{-i}) + \mathbb{E}_1 \mathbb{E}_2 H_\varphi(\boldsymbol{Z} \,|\, X_1, X_2).$$

Continuing in this fashion, we arrive at the subadditivity inequality (2.3):

$$H_\varphi(\boldsymbol{Z}) \le \sum\nolimits_{i=1}^{n} \mathbb{E}\, H_\varphi(\boldsymbol{Z} \,|\, \boldsymbol{x}_{-i}).$$

This completes the proof of Theorem 2.5.

5 Entropy Bounds via Exchangeability

In this section, we derive Corollary 2.6, which uses exchangeable pairs to bound the conditional entropies that appear in Theorem 2.5. This result follows from another variational representation of the matrix φ-entropy.

5.1 Representation of the Matrix φ-Entropy as an Infimum

In this section, we present another formula for the matrix φ-entropy.

Lemma 5.1 (Infimum Representation for Entropy)**.** *Fix a function $\varphi \in \Phi_\infty$, and let $\psi = \varphi'$. Assume that $\boldsymbol{Z}$ is a random positive-semidefinite matrix where $\|\boldsymbol{Z}\|$ and $\|\varphi(\boldsymbol{Z})\|$ are integrable. Then*

$$H_\varphi(\boldsymbol{Z}) = \inf_{\boldsymbol{A} \in \mathbb{H}^d_+} \mathbb{E}\,\bar{\mathrm{tr}}\,[\varphi(\boldsymbol{Z}) - \varphi(\boldsymbol{A}) - (\boldsymbol{Z} - \boldsymbol{A}) \cdot \psi(\boldsymbol{A})]\,. \tag{5.1}$$

Let $\boldsymbol{Z}'$ be an independent copy of $\boldsymbol{Z}$. Then

$$H_\varphi(\boldsymbol{Z}) \le \frac{1}{2} \cdot \mathbb{E}\,\bar{\mathrm{tr}}\,[(\boldsymbol{Z} - \boldsymbol{Z}')(\psi(\boldsymbol{Z}) - \psi(\boldsymbol{Z}'))]\,. \tag{5.2}$$

We require a familiar trace inequality [Car10, Thm. 2.11]. This bound simply restates the fact that a convex function lies above its tangents.

Proposition 5.2 (Klein's Inequality)**.** *Let $f : I \to \mathbb{R}$ be a differentiable convex function on an interval I of the real line. Then*

$$\bar{\mathrm{tr}}\,\big[f(\boldsymbol{B}) - f(\boldsymbol{A}) - (\boldsymbol{B} - \boldsymbol{A}) \cdot f'(\boldsymbol{A})\big] \ge 0 \quad \text{for all } \boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}^d(I).$$

With Klein's inequality at hand, the variational inequality follows quickly.

Proof of Lemma 5.1. Every function $\varphi \in \Phi_\infty$ is convex and differentiable, so Proposition 5.2 with $\boldsymbol{B} = \mathbb{E}\,\boldsymbol{Z}$ implies that

$$\bar{\mathrm{tr}}\,[-\varphi(\mathbb{E}\,\boldsymbol{Z})] \le \bar{\mathrm{tr}}\,[-\varphi(\boldsymbol{A}) - (\mathbb{E}\,\boldsymbol{Z} - \boldsymbol{A}) \cdot \psi(\boldsymbol{A})]$$

for each fixed matrix $\boldsymbol{A} \in \mathbb{H}^d_+$. Substitute this bound into the definition (2.2) of the matrix φ-entropy, and draw the expectation out of the trace to reach

$$H_\varphi(\boldsymbol{Z}) \le \mathbb{E}\,\bar{\mathrm{tr}}\,[\varphi(\boldsymbol{Z}) - \varphi(\boldsymbol{A}) - (\boldsymbol{Z} - \boldsymbol{A}) \cdot \psi(\boldsymbol{A})]. \tag{5.3}$$

The inequality (5.3) becomes an equality when $\boldsymbol{A} = \mathbb{E}\,\boldsymbol{Z}$, which establishes the variational representation (5.1).

The symmetrized bound (5.2) follows from an exchangeability argument. Select $\boldsymbol{A} = \boldsymbol{Z}'$ in the expression (5.3), and apply the fact that $\mathbb{E}\,\varphi(\boldsymbol{Z}) = \mathbb{E}\,\varphi(\boldsymbol{Z}')$ to obtain

$$H_\varphi(\boldsymbol{Z}) \le -\,\mathbb{E}\,\bar{\mathrm{tr}}\,[(\boldsymbol{Z} - \boldsymbol{Z}') \cdot \psi(\boldsymbol{Z}')]. \tag{5.4}$$

Since $\boldsymbol{Z}$ and $\boldsymbol{Z}'$ are exchangeable, we can also bound the matrix φ-entropy as

$$H_\varphi(\boldsymbol{Z}) \le -\,\mathbb{E}\,\bar{\mathrm{tr}}\,[(\boldsymbol{Z}' - \boldsymbol{Z}) \cdot \psi(\boldsymbol{Z})]. \tag{5.5}$$

Take the average of the two bounds (5.4) and (5.5) to reach the desired inequality (5.2). □

In the scalar case, stronger bounds are available. For a function $\varphi \in \Phi_1$,

$$\varphi(b) - \varphi(a) - (b - a)\varphi'(a) \le (b - a)(\varphi'(b) - \varphi'(a)) \quad \text{for all } a, b \ge 0.$$

See [Cha06, Lem. 4.2] for details.

5.2 Proof of Corollary 2.6

Lemma 5.1 leads to a succinct proof of Corollary 2.6. We continue to use the notation from Section 2.2.3. Apply the inequality (5.2) conditionally to control the conditional matrix φ-entropy:

$$H_\varphi(\boldsymbol{Z} \,|\, \boldsymbol{x}_{-i}) \le \frac{1}{2} \cdot \mathbb{E}\, \bar{\mathrm{tr}} \left[(\boldsymbol{Z} - \boldsymbol{Z}_i')(\psi(\boldsymbol{Z}) - \psi(\boldsymbol{Z}_i')) \,|\, \boldsymbol{x}_{-i} \right] \tag{5.6}$$

because $\boldsymbol{Z}_i'$ and $\boldsymbol{Z}$ are conditionally iid, given $\boldsymbol{x}_{-i}$. Take the expectation on both sides of (5.6), and invoke the tower property of conditional expectation:

$$\mathbb{E}\, H_\varphi(\boldsymbol{Z} \,|\, \boldsymbol{x}_{-i}) \le \frac{1}{2} \cdot \mathbb{E}\, \bar{\mathrm{tr}} \left[(\boldsymbol{Z} - \boldsymbol{Z}_i')(\psi(\boldsymbol{Z}) - \psi(\boldsymbol{Z}_i')) \right]. \tag{5.7}$$

To complete the proof, substitute (5.7) into the right-hand side of the bound (2.3) from the subadditivity result, Theorem 2.5.

6 Members of the Φ_∞ function class

In this section, we demonstrate that the classical entropy and certain power functions belong to the Φ_∞ function class. The main challenge is to verify that $\boldsymbol{A} \mapsto [\mathrm{D}\psi(\boldsymbol{A})]^{-1}$ is a concave operator-valued map. We establish this result for the classical entropy in Section 6.4 and for the power function in Section 6.5. See the independent work [Han13, Sec. 4] for closely related results.

6.1 Tensor Product Operators

First, we explain the tensor product construction of an operator. The tensor product will allow us to represent the derivative of a standard matrix function compactly.

Definition 6.1 (Tensor Product)**.** *Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}^d$. The operator $\boldsymbol{A} \otimes \boldsymbol{B} \in \mathbb{B}(\mathbb{M}^d)$ is defined by the relation*

$$(\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{M}) = \boldsymbol{A}\boldsymbol{M}\boldsymbol{B} \quad \text{for each } \boldsymbol{M} \in \mathbb{M}^d. \tag{6.1}$$

The operator $\boldsymbol{A} \otimes \boldsymbol{B}$ is self-adjoint because we assume the factors are Hermitian matrices.

Suppose that $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}^d$ are Hermitian matrices with spectral resolutions

$$\boldsymbol{A} = \sum\nolimits_{i=1}^d \lambda_i \boldsymbol{P}_i \quad \text{and} \quad \boldsymbol{B} = \sum\nolimits_{j=1}^d \mu_j \boldsymbol{Q}_j. \tag{6.2}$$

Then the tensor product $\boldsymbol{A} \otimes \boldsymbol{B}$ has the spectral resolution

$$\boldsymbol{A} \otimes \boldsymbol{B} = \sum\nolimits_{i,j=1}^d \lambda_i \mu_j \boldsymbol{P}_i \otimes \boldsymbol{Q}_j.$$

In particular, the tensor product of two positive-definite matrices is a positive-definite operator.

6.2 The Derivative of a Standard Matrix Function

Next, we present some classical results on the derivative of a standard matrix function. See [Bha97, Sec. V.3] for further details.

Definition 6.2 (Divided Difference)**.** *Let $f : I \to \mathbb{R}$ be a continuously differentiable function on an interval I of the real line. The* first divided difference *is the map $f^{[1]} : \mathbb{R}^2 \to \mathbb{R}$ defined by*

$$f^{[1]}(\lambda, \mu) := \begin{cases} f'(\lambda), & \lambda = \mu. \\ \dfrac{f(\lambda) - f(\mu)}{\lambda - \mu}, & \lambda \neq \mu, \end{cases}$$

We also require the Hermite representation of the divided difference:

$$f^{[1]}(\lambda, \mu) = \int_0^1 f'(\tau\lambda + \bar{\tau}\mu)\,\mathrm{d}\tau, \tag{6.3}$$

where we have written $\bar{\tau} := 1 - \tau$.

The following result gives an explicit expression for the derivative of a standard matrix function in terms of a divided difference.

Proposition 6.3 (Daleckiĭ–Kreĭn Formula)**.** *Let* $f : I \to \mathbb{R}$ *be a continuously differentiable function of an interval* I *of the real line. Suppose that* $\boldsymbol{A} \in \mathbb{H}^d(I)$ *is a diagonal matrix with* $\boldsymbol{A} = \operatorname{diag}(a_1, \dots, a_d)$. *The derivative* $\mathsf{D}f(\boldsymbol{A}) \in \mathbb{B}(\mathbb{M}^d)$, *and*

$$\mathsf{D}f(\boldsymbol{A})(\boldsymbol{H}) = f^{[1]}(\boldsymbol{A}) \odot \boldsymbol{H} \quad \text{for } \boldsymbol{H} \in \mathbb{M}^d,$$

where $\odot$ *denotes the Schur (i.e., componentwise) product and* $f^{[1]}(\boldsymbol{A})$ *refers to the matrix of divided differences:*

$$\big[f^{[1]}(\boldsymbol{A})\big]_{ij} = f^{[1]}(a_i, a_j) \quad \text{for } i, j = 1, \dots, d.$$

6.3 Operator Means

Our approach also relies on the concept of an operator mean. The following definition is due to Kubo & Ando [KA80].

Definition 6.4 (Operator Mean)**.** *Let* $f : \mathbb{R}_{++} \to \mathbb{R}_{++}$ *be an operator concave function that satisfies* $f(1) = 1$. *Fix a natural number* d. *Let* S *and* T *be positive-definite operators in* $\mathbb{B}(\mathbb{M}^d)$. *We define the mean of the operators:*

$$\mathsf{M}_f(\mathsf{S}, \mathsf{T}) := \mathsf{T}^{1/2} \cdot f(\mathsf{T}^{-1/2}\mathsf{S}\mathsf{T}^{-1/2}) \cdot \mathsf{T}^{1/2} \in \mathbb{B}(\mathbb{M}^d).$$

When S *and* T *commute, the formula simplifies to*

$$\mathsf{M}_f(\mathsf{S}, \mathsf{T}) = \mathsf{T} \cdot f(\mathsf{S}\mathsf{T}^{-1}).$$

A few examples may be helpful. The function $f(s) = (1 + s)/2$ represents the usual arithmetic mean, the function $f(s) = s^{1/2}$ gives the geometric mean, and the function $f(s) = 2s/(1 + s)$ yields the harmonic mean. Operator means have a concavity property, which was established in the paper [KA80].

Proposition 6.5 (Operator Means are Concave)**.** *Let* $f : \mathbb{R}_{++} \to \mathbb{R}_{++}$ *be an operator monotone function with* $f(1) = 1$. *Fix a natural number* d. *Suppose that* $\mathsf{S}_1, \mathsf{S}_2, \mathsf{T}_1, \mathsf{T}_2$ *are positive-definite operators in* $\mathbb{B}(\mathbb{M}^d)$. *Then*

$$\alpha \cdot \mathsf{M}_f(\mathsf{S}_1, \mathsf{T}_1) + \bar{\alpha} \cdot \mathsf{M}_f(\mathsf{S}_2, \mathsf{T}_2) \preccurlyeq \mathsf{M}_f(\alpha\mathsf{S}_1 + \bar{\alpha}\mathsf{S}_2, \alpha\mathsf{T}_1 + \bar{\alpha}\mathsf{T}_2)$$

for $\alpha \in [0, 1]$ *and* $\bar{\alpha} = 1 - \alpha$.

6.4 Entropy

In this section, we demonstrate that the standard entropy function is a member of the Φ_∞ function class.

Theorem 6.6. *The function* $\varphi : t \mapsto t \log t - t$ *is a member of the* Φ_∞ *class.*

This result immediately implies Theorem 2.3(1), which states that $t \mapsto t \log t$ belongs to Φ_∞. Indeed, the matrix entropy class contains all affine functions and all finite sums of its elements.

Theorem 6.6 follows easily from (deep) classical results because the variational representation of the standard entropy from Lemma 4.1 is equivalent with the joint convexity of quantum relative entropy [Lin73]. Instead of pursuing this idea, we present an argument that parallels the approach we use to study the power function. Some of the calculations below also appear in [Lie73, Proof of Cor. 2.1], albeit in compressed form.

Proof. Fix a positive integer d. We plan to show that the function $\varphi : t \mapsto t \log t - t$ is a member of the class Φ_d. Evidently, φ is continuous and convex on $\mathbb{R}_+$, and it has two continuous derivatives on $\mathbb{R}_{++}$. It remains to verify the concavity condition for the second derivative.

Write $\psi(t) = \varphi'(t) = \log t$, and let $\boldsymbol{A} \in \mathbb{H}^d_{++}$. Without loss of generality, we may choose a basis where $\boldsymbol{A} = \operatorname{diag}(a_1, \dots, a_d)$. The Daleckiĭ–Kreĭn formula, Proposition 6.3, tells us

$$\mathsf{D}\psi(\boldsymbol{A})(\boldsymbol{H}) = \psi^{[1]}(\boldsymbol{A}) \odot \boldsymbol{H} = \left[\psi^{[1]}(a_i, a_j) \cdot h_{ij}\right]_{ij}.$$

As an operator, the derivative acts by Schur multiplication. This formula also makes it clear that the inverse of this operator acts by Schur multiplication:

$$[\mathsf{D}\psi(\boldsymbol{A})]^{-1}(\boldsymbol{H}) = \left[\frac{1}{\psi^{[1]}(a_i, a_j)} \cdot h_{ij}\right]_{ij}.$$

Using the Hermite representation (6.3) of the first divided difference of $t \mapsto \mathrm{e}^t$, we find

$$\frac{1}{\psi^{[1]}(\mu, \lambda)} = \frac{\lambda - \mu}{\log \lambda - \log \mu} = \int_0^1 \mathrm{e}^{\tau \log \lambda + \bar{\tau} \log \mu} \,\mathrm{d}\tau = \int_0^1 \lambda^\tau \mu^{\bar{\tau}} \,\mathrm{d}\tau.$$

The latter calculation assumes that $\mu \neq \lambda$; it extends to the case $\mu = \lambda$ because both sides of the identity are continuous. As a consequence,

$$[\mathsf{D}\psi(\boldsymbol{A})]^{-1}(\boldsymbol{H}) = \int_0^1 \left[a_i^\tau h_{ij} a_j^{\bar{\tau}}\right]_{ij} \mathrm{d}\tau = \int_0^1 \boldsymbol{A}^\tau \boldsymbol{H} \boldsymbol{A}^{\bar{\tau}} \,\mathrm{d}\tau = \int_0^1 (\boldsymbol{A}^\tau \otimes \boldsymbol{A}^{\bar{\tau}})(\boldsymbol{H}) \,\mathrm{d}\tau.$$

We discover the expression

$$[\mathsf{D}\psi(\boldsymbol{A})]^{-1} = \int_0^1 \boldsymbol{A}^\tau \otimes \boldsymbol{A}^{\bar{\tau}} \,\mathrm{d}\tau. \tag{6.4}$$

This formula is correct for every positive-definite matrix.

For each $\tau \in [0, 1]$, consider the operator monotone function $f : t \mapsto t^\tau$ defined on $\mathbb{R}_+$. Since $f(1) = 1$, we can construct the operator mean M_f associated with the function f. Note that $\boldsymbol{A} \otimes \mathbf{I}$ and $\mathbf{I} \otimes \boldsymbol{A}$ are commuting positive operators. Thus,

$$\mathsf{M}_f(\boldsymbol{A} \otimes \mathbf{I}, \mathbf{I} \otimes \boldsymbol{A}) = (\mathbf{I} \otimes \boldsymbol{A}) \cdot f((\boldsymbol{A} \otimes \mathbf{I})(\mathbf{I} \otimes \boldsymbol{A})^{-1}) = \boldsymbol{A}^\tau \otimes \boldsymbol{A}^{\bar{\tau}}.$$

The map $\boldsymbol{A} \mapsto (\boldsymbol{A} \otimes \mathbf{I}, \mathbf{I} \otimes \boldsymbol{A})$ is linear, so Proposition 6.5 guarantees that $\boldsymbol{A} \mapsto \boldsymbol{A}^\tau \otimes \boldsymbol{A}^{\bar{\tau}}$ is concave for each $\tau \in [0, 1]$. This result is usually called the Lieb Concavity Theorem [Bha97, Thm. IX.6.1]. Combine this fact with the integral representation (6.4) to reach the conclusion that $\boldsymbol{A} \mapsto [\mathsf{D}\psi(\boldsymbol{A})]^{-1}$ is a concave map on the cone $\mathbb{H}^d_{++}$ of positive-definite matrices. □

6.5 Power Functions

In this section, we prove that certain power functions belong to the Φ_∞ function class.

Theorem 6.7. *For each $p \in [0,1]$, the function $\varphi : t \mapsto t^{p+1}/(p+1)$ is a member of the Φ_∞ class.*

This result immediately implies Theorem 2.3(2), which states that $t \mapsto t^{p+1}$ belongs to the class Φ_∞. Indeed, the matrix entropy class contains all positive multiples of its elements.

The proof of Theorem 6.7 follows the same path as Theorem 6.6, but it is somewhat more involved. First, we derive an expression for the function $\boldsymbol{A} \mapsto [\mathsf{D}\psi(\boldsymbol{A})]^{-1}$ where $\psi = \varphi'$.

Lemma 6.8. *Fix $p \in (0,1]$, and let $\psi(t) = t^p$ for $t \geq 0$. For each matrix $\boldsymbol{A} \in \mathbb{H}_+^d$,*

$$[\mathsf{D}\psi(\boldsymbol{A})]^{-1} = \frac{1}{p}\int_0^1 (\tau \cdot \boldsymbol{A}^p \otimes \mathbf{I} + \bar{\tau} \cdot \mathbf{I} \otimes \boldsymbol{A}^p)^{(1-p)/p}\,\mathrm{d}\tau, \tag{6.5}$$

where $\bar{\tau} := 1 - \tau$.

Proof. As before, we may assume without loss of generality that the matrix $\boldsymbol{A} = \operatorname{diag}(a_1, \dots, a_d)$. Using the Daleckiĭ–Kreĭn formula, Proposition 6.3, we see that

$$[\mathsf{D}\psi(\boldsymbol{A})]^{-1}(\boldsymbol{H}) = \left[\frac{1}{\psi^{[1]}(a_i, a_j)} \cdot h_{ij}\right].$$

The Hermite representation (6.3) of the first divided difference of $t \mapsto t^{1/p}$ gives

$$\frac{1}{\psi^{[1]}(\mu,\lambda)} = \frac{\mu - \lambda}{\mu^p - \lambda^p} = \frac{1}{p}\int_0^1 (\tau \cdot \lambda^p + \bar{\tau} \cdot \mu^p)^{(1-p)/p}\,\mathrm{d}\tau =: g(\lambda, \mu).$$

We use continuity to verify that the latter calculation remains valid when $\mu = \lambda$. Using this function g, we can identify a compact representation of the operator:

$$[\mathsf{D}\psi(\boldsymbol{A})]^{-1}(\boldsymbol{H}) = \sum\nolimits_{ij} g(a_i, a_j) h_{ij} \mathbf{E}_{ij} = \left[\sum\nolimits_{ij} g(a_i, a_j)(\mathbf{E}_{ii} \otimes \mathbf{E}_{jj})\right](\boldsymbol{H}),$$

where we write $\mathbf{E}_{ij}$ for the matrix with a one in the (i,j) position and zeros elsewhere. It remains to verify that the bracket coincides with the expression (6.5). Indeed,

$$\begin{aligned}
\sum\nolimits_{ij} g(a_i, a_j)(\mathbf{E}_{ii} \otimes \mathbf{E}_{jj}) &= \frac{1}{p}\int_0^1 \sum\nolimits_{ij} (\tau \cdot a_i^p + \bar{\tau} \cdot a_j^p)^{(1-p)/p}\,(\mathbf{E}_{ii} \otimes \mathbf{E}_{jj})\,\mathrm{d}\tau \\
&= \frac{1}{p}\int_0^1 \left[\sum\nolimits_{ij} (\tau \cdot a_i^p + \bar{\tau} \cdot a_j^p)(\mathbf{E}_{ii} \otimes \mathbf{E}_{jj})\right]^{(1-p)/p} \mathrm{d}\tau \\
&= \frac{1}{p}\int_0^1 (\tau \cdot \boldsymbol{A}^p \otimes \mathbf{I} + \bar{\tau} \cdot \mathbf{I} \otimes \boldsymbol{A}^p)^{(1-p)/p}\,\mathrm{d}\tau.
\end{aligned}$$

The second relation follows from the definition of the standard operator function associated with $t \mapsto t^{(1-p)/p}$. To confirm that the third line equals the second, expand the matrices $\boldsymbol{A} = \sum_i a_i \mathbf{E}_{ii}$ and $\mathbf{I} = \sum_j \mathbf{E}_{jj}$ and invoke the bilinearity of the tensor product. □

Proof of Theorem 6.7. We are now prepared to prove that certain power functions belong to the Φ_∞ function class. Fix an exponent $p \in [0,1]$, and let d be a fixed positive

integer. We intend to show that the function $\varphi(t) = t^{p+1}/(p+1)$ belongs to the Φ_d class. When $p = 0$, the function φ is affine, so we may assume that $p > 0$. It is clear that φ is continuous and convex on $\mathbb{R}_+$, and φ has two continuous derivatives on $\mathbb{R}_{++}$. It remains to verify that the second derivative has the required concavity property.

Let $\psi(t) = \varphi'(t) = t^p$ for $t \geq 0$, and consider a matrix $\boldsymbol{A} \in \mathbb{H}^d_{++}$. Lemma 6.8 demonstrates that

$$[\mathsf{D}\psi(\boldsymbol{A})]^{-1} = \frac{1}{p} \int_0^1 (\tau \cdot \boldsymbol{A}^p \otimes \mathbf{I} + \bar{\tau} \cdot \mathbf{I} \otimes \boldsymbol{A}^p)^{(1/p)(1-p)} \,\mathrm{d}\tau, \tag{6.6}$$

where we maintain the usage $\bar{\tau} := 1 - \tau$. For each $\tau \in [0,1]$, the scalar function $a \mapsto \tau a + \bar{\tau}$ is operator monotone because it is affine and increasing. On account of the result [And79, Cor. 4.3], the function

$$f : a \mapsto (\tau \cdot a^p + \bar{\tau})^{1/p}$$

is also operator monotone. A short calculation shows that $f(1) = 1$. Therefore, we can use f to construct an operator mean M_f. Since $\boldsymbol{A} \otimes \mathbf{I}$ and $\mathbf{I} \otimes \boldsymbol{A}$ are commuting positive operators, we have

$$\mathsf{M}_f(\boldsymbol{A} \otimes \mathbf{I}, \mathbf{I} \otimes \boldsymbol{A}) = (\mathbf{I} \otimes \boldsymbol{A}) \cdot f((\boldsymbol{A} \otimes \mathbf{I})(\mathbf{I} \otimes \boldsymbol{A})^{-1}) = \left(\tau \cdot \boldsymbol{A}^p \otimes \mathbf{I} + \bar{\tau} \cdot \mathbf{I} \otimes \boldsymbol{A}^p\right)^{1/p}.$$

The map $\boldsymbol{A} \mapsto (\boldsymbol{A} \otimes \mathbf{I}, \mathbf{I} \otimes \boldsymbol{A})$ is linear, so Proposition 6.5 ensures that

$$\boldsymbol{A} \mapsto (\tau \cdot \boldsymbol{A}^p \otimes \mathbf{I} + \bar{\tau} \cdot \mathbf{I} \otimes \boldsymbol{A}^p)^{1/p} \tag{6.7}$$

is a concave map.

We are now prepared to check that (6.6) defines a concave operator. Let $\boldsymbol{S}, \boldsymbol{T}$ be arbitrary positive-definite matrices, and choose $\alpha \in [0,1]$. Suppose that $\boldsymbol{Z}$ is the random matrix that takes value $\boldsymbol{S}$ with probability α and value $\boldsymbol{T}$ with probability $1-\alpha$. For each $\tau \in [0,1]$, we compute

$$\begin{aligned}\mathbb{E}\left[(\tau \cdot \boldsymbol{Z}^p \otimes \mathbf{I} + \bar{\tau} \cdot \mathbf{I} \otimes \boldsymbol{Z}^p)^{1/p}\right]^{1-p} &\preccurlyeq \left[\mathbb{E}\,(\tau \cdot \boldsymbol{Z}^p \otimes \mathbf{I} + \bar{\tau} \cdot \mathbf{I} \otimes \boldsymbol{Z}^p)^{1/p}\right]^{1-p} \\ &\preccurlyeq \left[\left(\tau \cdot (\mathbb{E}\,\boldsymbol{Z})^p \otimes \mathbf{I} + \bar{\tau} \cdot \mathbf{I} \otimes (\mathbb{E}\,\boldsymbol{Z})^p\right)^{1/p}\right]^{1-p}.\end{aligned}$$

The first relation holds because $t \mapsto t^{1-p}$ is operator concave [Bha97, Thm. V.1.9 and Thm. V.2.5]. To obtain the second relation, we apply the concavity property of the map (6.7), followed by the fact that $t \mapsto t^{1-p}$ is operator monotone [Bha97, Thm. V.1.9]. This calculation establishes the claim that

$$\boldsymbol{A} \mapsto \left(\tau \cdot \boldsymbol{A}^p \otimes \mathbf{I} + \bar{\tau} \cdot \mathbf{I} \otimes \boldsymbol{A}^p\right)^{(1-p)/p}$$

is concave on $\mathbb{H}^d_{++}$ for each $\tau \in [0,1]$. In view of the integral representation (6.6), we may conclude that $\boldsymbol{A} \mapsto [\mathsf{D}\psi(\boldsymbol{A})]^{-1}$ is concave on the cone $\mathbb{H}^d_{++}$ of positive-definite matrices. □

7 A Bounded Difference Inequality for Random Matrices

In this section, we prove Theorem 2.8, a bounded difference inequality for a random matrix whose distribution is invariant under signed permutation. We begin with some preliminaries that support the proof, and we establish the main result in Section 7.2.

7.1 Preliminaries

First, we describe how to compute the expectation of a function of a random matrix whose distribution is invariant under signed permutation. See Definition 2.7 for a reminder of what this requirement means.

Lemma 7.1. *Let $f : I \to \mathbb{R}$ be a function on an interval I of the real line. Assume that $\boldsymbol{X} \in \mathbb{H}^d(I)$ is a random matrix whose distribution is invariant under signed permutation. Then*

$$\mathbb{E}\, f(\boldsymbol{X}) = \bar{\mathrm{tr}}[\mathbb{E}\, f(\boldsymbol{X})] \cdot \mathbf{I}.$$

Proof. Let $\boldsymbol{\Pi} \in \mathbb{H}^d$ be an arbitrary signed permutation matrix. Observe that

$$\mathbb{E}\, f(\boldsymbol{X}) = \mathbb{E}\, f(\boldsymbol{\Pi}^* \boldsymbol{X} \boldsymbol{\Pi}) = \boldsymbol{\Pi}^* [\mathbb{E}\, f(\boldsymbol{X})] \boldsymbol{\Pi}. \tag{7.1}$$

The first relation holds because the distribution of $\boldsymbol{X}$ is invariant under conjugation by $\boldsymbol{\Pi}$. The second relation follows from the definition of a standard matrix function and the fact that $\boldsymbol{\Pi}$ is unitary. We may average (7.1) over $\boldsymbol{\Pi}$ drawn from the uniform distribution on the set of signed permutation matrices. A direct calculation shows that the resulting matrix is diagonal, and its diagonal entries are identically equal to $\bar{\mathrm{tr}}[\mathbb{E}\, f(\boldsymbol{X})]$. □

We also require a trace inequality that is related to the mean value theorem. This result specializes [MJC+12, Lem. 3.4].

Proposition 7.2 (Mean Value Trace Inequality)**.** *Let $f : I \to \mathbb{R}$ be a function on an interval I of the real line whose derivative f' is convex. For all $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}^d(I)$,*

$$\bar{\mathrm{tr}}[(\boldsymbol{A} - \boldsymbol{B})(f(\boldsymbol{A}) - f(\boldsymbol{B}))] \le \frac{1}{2}\, \bar{\mathrm{tr}}[(\boldsymbol{A} - \boldsymbol{B})^2 \cdot (f'(\boldsymbol{A}) + f'(\boldsymbol{B}))].$$

7.2 Proof of Theorem 2.8

The argument proceeds in three steps. First, we present some elements of the matrix Laplace transform method. Second, we use the subaddivity of matrix φ-entropy to deduce a differential inequality for the trace moment generating function of the random matrix. Finally, we explain how to integrate the differential inequality to obtain the concentration result.

7.2.1 The Matrix Laplace Transform Method

We begin with a matrix extension of the moment generating function (mgf), which has played a major role in recent work on matrix concentration.

Definition 7.3 (Trace Mgf)**.** *Let $\boldsymbol{Y}$ be a random Hermitian matrix. The* normalized trace moment generating function *of $\boldsymbol{Y}$ is defined as*

$$m(\theta) := m_{\boldsymbol{Y}}(\theta) := \mathbb{E}\, \bar{\mathrm{tr}}\, \mathrm{e}^{\theta \boldsymbol{Y}} \quad \text{for } \theta \in \mathbb{R}.$$

The expectation need not exist for all values of θ.

The following proposition explains how the trace mgf can be used to study the maximum eigenvalue of a random Hermitian matrix [Tro11, Prop. 3.1].

Proposition 7.4 (Matrix Laplace Transform Method)**.** *Let $\boldsymbol{Y} \in \mathbb{H}^d$ be a random matrix with normalized trace mgf $m(\theta) := \bar{\mathrm{tr}}\, \mathrm{e}^{\theta \boldsymbol{Y}}$. For each $t \in \mathbb{R}$,*

$$\mathbb{P}\{\lambda_{\max}(\boldsymbol{Y}) \ge t\} \le d \cdot \inf_{\theta > 0} \mathrm{e}^{-\theta t + \log m(\theta)}.$$

7.2.2 A Differential Inequality for the Trace Mgf

Suppose that $\boldsymbol{Y} \in \mathbb{H}^d$ is a random Hermitian matrix that depends on a random vector $\boldsymbol{x} := (X_1, \dots, X_n)$. We require the distribution of $\boldsymbol{Y}$ to be invariant under signed permutations, and we insist that $\|\boldsymbol{Y}\|$ is bounded. Without loss of generality, assume that $\boldsymbol{Y}$ has zero mean. Throughout the argument, we let the notation of Section 2.2.3 and Theorem 2.8 prevail.

Let us explain how to use the subadditivity of matrix φ-entropy to derive a differential inequality for the trace mgf. Consider the function $\varphi(t) = t \log t$, which belongs to the Φ_∞ class because of Theorem 2.3(1). Introduce the random positive-definite matrix $\boldsymbol{Z} := \mathrm{e}^{\theta \boldsymbol{Y}}$, where $\theta > 0$. We write out an expression for the matrix φ-entropy of $\boldsymbol{Z}$:

$$\begin{aligned} H_\varphi(\boldsymbol{Z}) &= \mathbb{E}\,\bar{\mathrm{tr}}[\varphi(\boldsymbol{Z}) - \varphi(\mathbb{E}\,\boldsymbol{Z})] \\ &= \mathbb{E}\,\bar{\mathrm{tr}}\left[(\theta\boldsymbol{Y})\mathrm{e}^{\theta\boldsymbol{Y}} - \mathrm{e}^{\theta\boldsymbol{Y}} \log \mathbb{E}\,\mathrm{e}^{\theta\boldsymbol{Y}}\right] \\ &= \theta \cdot \mathbb{E}\,\bar{\mathrm{tr}}\left[\boldsymbol{Y}\mathrm{e}^{\theta\boldsymbol{Y}}\right] - (\mathbb{E}\,\bar{\mathrm{tr}}\,\mathrm{e}^{\theta\boldsymbol{Y}}) \log(\mathbb{E}\,\bar{\mathrm{tr}}\,\mathrm{e}^{\theta\boldsymbol{Y}}) \\ &= \theta m'(\theta) - m(\theta)\log m(\theta). \end{aligned} \tag{7.2}$$

In the third line, we have applied Lemma 7.1 to the logarithm in the second term, relying on the fact that $\boldsymbol{Y}$ is invariant under signed permutations. To reach the last line, we recognize that $m'(\theta) = \mathbb{E}\,\bar{\mathrm{tr}}(\boldsymbol{Y}\mathrm{e}^{\theta\boldsymbol{Y}})$. We have used the boundedness of $\|\boldsymbol{Y}\|$ to justify this derivative calculation.

Corollary 2.6 provides an upper bound for the matrix φ-entropy. Define the derivative $\psi(t) = \varphi'(t) = 1 + \log t$. Then

$$\begin{aligned} H_\varphi(\boldsymbol{Z}) &\le \frac{1}{2}\sum\nolimits_{i=1}^n \mathbb{E}\,\bar{\mathrm{tr}}\left[(\boldsymbol{Z} - \boldsymbol{Z}_i')(\psi(\boldsymbol{Z}) - \psi(\boldsymbol{Z}_i'))\right] \\ &= \frac{\theta}{2}\sum\nolimits_{i=1}^n \mathbb{E}\,\bar{\mathrm{tr}}\left[(\mathrm{e}^{\theta\boldsymbol{Y}} - \mathrm{e}^{\theta\boldsymbol{Y}_i'})(\boldsymbol{Y} - \boldsymbol{Y}_i')\right]. \end{aligned}$$

Consider the function $f : t \mapsto \mathrm{e}^{\theta t}$. Its derivative $f' : t \mapsto \theta\mathrm{e}^{\theta t}$ is convex because $\theta > 0$, so Proposition 7.2 delivers the bound

$$\begin{aligned} H_\varphi(\boldsymbol{Z}) &\le \frac{\theta^2}{4}\sum\nolimits_{i=1}^n \mathbb{E}\,\bar{\mathrm{tr}}\left[(\mathrm{e}^{\theta\boldsymbol{Y}} + \mathrm{e}^{\theta\boldsymbol{Y}_i'})(\boldsymbol{Y} - \boldsymbol{Y}_i')^2\right] \\ &= \frac{\theta^2}{2}\sum\nolimits_{i=1}^n \mathbb{E}\,\bar{\mathrm{tr}}\left[\mathrm{e}^{\theta\boldsymbol{Y}}(\boldsymbol{Y} - \boldsymbol{Y}_i')^2\right] \\ &= \frac{\theta^2}{2}\sum\nolimits_{i=1}^n \mathbb{E}\,\bar{\mathrm{tr}}\left[\mathrm{e}^{\theta\boldsymbol{Y}} \cdot \mathbb{E}[(\boldsymbol{Y} - \boldsymbol{Y}_i')^2 \,|\, \boldsymbol{x}]\right]. \end{aligned}$$

The second relation follows from the fact that $\boldsymbol{Y}$ and $\boldsymbol{Y}_i'$ are exchangeable, conditional on $\boldsymbol{x}_{-i}$. The last line is just the tower property of conditional expectation, combined with the observation that $\boldsymbol{Y}$ is a function of $\boldsymbol{x}$. To continue, we simplify the expression and make some additional bounds.

$$\begin{aligned} H_\varphi(\boldsymbol{Z}) &\le \frac{\theta^2}{2}\,\mathbb{E}\,\bar{\mathrm{tr}}\left[\mathrm{e}^{\theta\boldsymbol{Y}} \cdot \sum\nolimits_{i=1}^n \mathbb{E}[(\boldsymbol{Y} - \boldsymbol{Y}_i')^2 \,|\, \boldsymbol{x}]\right] \\ &\le \frac{\theta^2}{2}(\mathbb{E}\,\bar{\mathrm{tr}}\,\mathrm{e}^{\theta\boldsymbol{Y}})\left\|\sum\nolimits_{i=1}^n \mathbb{E}[(\boldsymbol{Y} - \boldsymbol{Y}_i')^2 \,|\, \boldsymbol{x}]\right\| \\ &\le \frac{\theta^2 V_{\boldsymbol{Y}}}{2} \cdot m(\theta). \end{aligned} \tag{7.3}$$

The second relation follows from a standard trace inequality and the observation that $\mathrm{e}^{\theta\boldsymbol{Y}}$ is positive definite. Last, we identify the variance measure $V_{\boldsymbol{Y}}$ defined in (2.4) and the trace mgf $m(\theta)$.

Combine the expression (7.2) with the inequality (7.3) to arrive at the estimate

$$\theta m'(\theta) - m(\theta) \log m(\theta) \le \frac{\theta^2 V_{\boldsymbol{Y}}}{2} \cdot m(\theta) \quad \text{for } \theta > 0. \tag{7.4}$$

We can use this differential inequality to obtain bounds on the trace mgf $m(\theta)$.

7.2.3 Solving the Differential Inequality

Rearrange the differential inequality (7.4) to obtain

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\left[\frac{\log m(\theta)}{\theta}\right] = \frac{m'(\theta)}{\theta m(\theta)} - \frac{\log m(\theta)}{\theta^2} \le \frac{V_{\boldsymbol{Y}}}{2}. \tag{7.5}$$

The l'Hôpital rule allows us to calculate the value of $\theta^{-1} \log m(\theta)$ at zero. Since $m(0) = 1$,

$$\lim_{\theta \to 0} \frac{\log m(\theta)}{\theta} = \lim_{\theta \to 0} \frac{m'(\theta)}{m(\theta)} = \lim_{\theta \to 0} \frac{\mathbb{E}\,\bar{\mathrm{tr}}(\boldsymbol{Y} \mathrm{e}^{\theta \boldsymbol{Y}})}{\mathbb{E}\,\bar{\mathrm{tr}}\,\mathrm{e}^{\theta \boldsymbol{Y}}} = \mathbb{E}\,\bar{\mathrm{tr}}\,\boldsymbol{Y} = 0.$$

This is where we use the hypothesis that $\boldsymbol{Y}$ has mean zero. Now, we integrate (7.5) from zero to some positive value θ to find that the trace mgf satisfies

$$\frac{\log m(\theta)}{\theta} \le \frac{\theta V_{\boldsymbol{Y}}}{2} \quad \text{when } \theta > 0. \tag{7.6}$$

The approach in this section is usually referred to as the Herbst argument [Led99].

7.2.4 The Laplace Transform Argument

We are now prepared to finish the argument. Combine the matrix Laplace transform method, Proposition 7.4, with the trace mgf bound (7.6) to reach

$$\mathbb{P}\left\{\lambda_{\max}(\boldsymbol{Y}) \ge t\right\} \le d \cdot \inf_{\theta > 0} \mathrm{e}^{-\theta t + \log m(\theta)} \le d \cdot \inf_{\theta > 0} \mathrm{e}^{-\theta t + \theta^2 V_{\boldsymbol{Y}}/2} = d \cdot \mathrm{e}^{-t^2/(2V_{\boldsymbol{Y}})}. \tag{7.7}$$

To obtain the result for the minimum eigenvalue, we note that

$$\mathbb{P}\left\{\lambda_{\min}(\boldsymbol{Y}) \le -t\right\} = \mathbb{P}\left\{\lambda_{\max}(-\boldsymbol{Y}) \ge t\right\} \le d \cdot \mathrm{e}^{-t^2/(2V_{\boldsymbol{Y}})}.$$

The inequality follows when we apply (7.7) to the random matrix $-\boldsymbol{Y}$. This completes the proof of Theorem 2.8.

8 Moment Inequalities for Random Matrices with Bounded Differences

In this section, we prove Theorem 2.9, which gives information about the moments of a random matrix that satisfies a kind of self-bounding property.

Proof of Theorem 2.9. Fix a number $q \in \{2, 3, 4, \dots\}$. Suppose that $\boldsymbol{Y} \in \mathbb{H}_+^d$ is a random positive-semidefinite matrix that depends on a random vector $\boldsymbol{x} := (X_1, \dots, X_n)$. We require the distribution of $\boldsymbol{Y}$ to be invariant under signed permutations, and we assume that $\mathbb{E}(\|\boldsymbol{Y}\|^q) < \infty$. The notation of Section 2.2.3 and Theorem 2.9 remains in force.

Let us explain how the subadditivity of matrix φ-entropy leads to a bound on the qth trace moment of $\boldsymbol{Y}$. Consider the power function $\varphi(t) = t^{q/(q-1)}$. Theorem 6.7 ensures

that $\varphi \in \Phi_\infty$ because $q/(q-1) \in (1,2]$. Introduce the random positive-semidefinite matrix $\boldsymbol{Z} := \boldsymbol{Y}^{q-1}$. Then

$$\begin{aligned} H_\varphi(\boldsymbol{Z}) &= \mathbb{E}\,\bar{\mathrm{tr}}\left[\varphi(\boldsymbol{Z}) - \varphi(\mathbb{E}\,\boldsymbol{Z})\right] \\ &= \mathbb{E}\,\bar{\mathrm{tr}}(\boldsymbol{Y}^q) - \bar{\mathrm{tr}}\left[(\mathbb{E}(\boldsymbol{Y}^{q-1}))^{q/(q-1)}\right] \\ &= \mathbb{E}\,\bar{\mathrm{tr}}(\boldsymbol{Y}^q) - \left[\mathbb{E}\,\bar{\mathrm{tr}}(\boldsymbol{Y}^{q-1})\right]^{q/(q-1)}. \end{aligned} \tag{8.1}$$

The transition to the last line requires Lemma 7.1.

Corollary 2.6 provides an upper bound for the matrix φ-entropy. Define the derivative $\psi(t) = \varphi'(t) = (q/(q-1)) \cdot t^{1/(q-1)}$. We have

$$\begin{aligned} H_\varphi(\boldsymbol{Z}) &\le \frac{1}{2}\sum\nolimits_{i=1}^n \mathbb{E}\,\bar{\mathrm{tr}}\left[(\boldsymbol{Z} - \boldsymbol{Z}_i')(\psi(\boldsymbol{Z}) - \psi(\boldsymbol{Z}_i'))\right] \\ &= \frac{q}{2(q-1)}\sum\nolimits_{i=1}^n \mathbb{E}\,\bar{\mathrm{tr}}\left[\left(\boldsymbol{Y}^{q-1} - (\boldsymbol{Y}_i')^{q-1}\right)(\boldsymbol{Y} - \boldsymbol{Y}_i')\right] \end{aligned}$$

The function $f : t \mapsto t^{q-1}$ has the derivative $f' : t \mapsto (q-1)t^{q-2}$, which is convex because $q \in \{2,3,4,\dots\}$. Therefore, the mean value trace inequality, Proposition 7.2, delivers the bound

$$\begin{aligned} H_\varphi(\boldsymbol{Z}) &\le \frac{q}{4}\sum\nolimits_{i=1}^n \mathbb{E}\,\bar{\mathrm{tr}}\left[\left(\boldsymbol{Y}^{q-2} + (\boldsymbol{Y}_i')^{q-2}\right)(\boldsymbol{Y} - \boldsymbol{Y}_i')^2\right] \\ &= \frac{q}{2}\sum\nolimits_{i=1}^n \mathbb{E}\,\bar{\mathrm{tr}}\left[\boldsymbol{Y}^{q-2}(\boldsymbol{Y} - \boldsymbol{Y}_i')^2\right] \\ &= \frac{q}{2}\sum\nolimits_{i=1}^n \mathbb{E}\,\bar{\mathrm{tr}}\left[\boldsymbol{Y}^{q-2}\,\mathbb{E}[(\boldsymbol{Y} - \boldsymbol{Y}_i')^2 \,|\, \boldsymbol{x}]\right]. \end{aligned}$$

The second identity holds because $\boldsymbol{Y}$ and $\boldsymbol{Y}_i'$ are exchangeable, conditional on $\boldsymbol{x}_{-i}$. The last line follows from the tower property of conditional expectation. We simplify this expression as follows.

$$\begin{aligned} H_\varphi(\boldsymbol{Z}) &\le \frac{q}{2}\,\mathbb{E}\,\bar{\mathrm{tr}}\left[\boldsymbol{Y}^{q-2} \cdot \sum\nolimits_{i=1}^n \mathbb{E}[(\boldsymbol{Y} - \boldsymbol{Y}_i')^2 \,|\, \boldsymbol{x}]\right] \\ &\le \frac{q}{2}\,\mathbb{E}\,\bar{\mathrm{tr}}\left[\boldsymbol{Y}^{q-2} \cdot c\boldsymbol{Y}\right] \\ &= \frac{cq}{2}\,\mathbb{E}\,\bar{\mathrm{tr}}(\boldsymbol{Y}^{q-1}). \end{aligned} \tag{8.2}$$

The second inequality derives from the hypothesis (2.5) that $\boldsymbol{V}_{\boldsymbol{Y}} \preccurlyeq c\boldsymbol{Y}$. Note that this bound requires the fact that $\boldsymbol{Y}^{q-2}$ is positive semidefinite.

Combine the expression (8.1) for the matrix φ-entropy with the upper bound (8.2) to achieve the estimate

$$\mathbb{E}\,\bar{\mathrm{tr}}(\boldsymbol{Y}^q) - \left[\mathbb{E}\,\bar{\mathrm{tr}}(\boldsymbol{Y}^{q-1})\right]^{q/(q-1)} \le \frac{cq}{2}\,\mathbb{E}\,\bar{\mathrm{tr}}(\boldsymbol{Y}^{q-1}).$$

Rewrite this bound, and invoke the numerical fact $1 + aq \le (1+a)^q$ to obtain

$$\begin{aligned} \mathbb{E}\,\bar{\mathrm{tr}}(\boldsymbol{Y}^q) &\le \left[\mathbb{E}\,\bar{\mathrm{tr}}(\boldsymbol{Y}^{q-1})\right]^{q/(q-1)} \left(1 + \frac{cq/2}{\left[\mathbb{E}\,\bar{\mathrm{tr}}(\boldsymbol{Y}^{q-1})\right]^{1/q-1}}\right) \\ &\le \left[\mathbb{E}\,\bar{\mathrm{tr}}(\boldsymbol{Y}^{q-1})\right]^{q/(q-1)} \left(1 + \frac{c/2}{\left[\mathbb{E}\,\bar{\mathrm{tr}}(\boldsymbol{Y}^{q-1})\right]^{1/q-1}}\right)^q. \end{aligned}$$

Extract the qth root from both sides to reach

$$\left[\mathbb{E}\,\bar{\mathrm{tr}}(\boldsymbol{Y}^q)\right]^{1/q} \le \left[\mathbb{E}\,\bar{\mathrm{tr}}(\boldsymbol{Y}^{q-1})\right]^{1/(q-1)} + \frac{c}{2}.$$

We have compared the qth trace moment of $\boldsymbol{Y}$ with the $(q-1)$th trace moment. Proceeding by iteration, we arrive at

$$\big[\mathbb{E}\,\bar{\mathrm{tr}}(\boldsymbol{Y}^q)\big]^{1/q} \le \mathbb{E}\,\bar{\mathrm{tr}}\,\boldsymbol{Y} + \frac{q-1}{2}\cdot c.$$

This observation completes the proof of Theorem 2.9. □

A Lemma 4.1, The General Case

In this appendix, we explain how to prove Lemma 4.1 in full generality. The argument calls for a simple but powerful result, known as the generalized Klein inequality [Pet94, Prop. 3], which allows us to lift a large class of scalar inequalities to matrices.

Proposition A.1 (Generalized Klein Inequality)**.** *For each $k = 1, \dots, n$, suppose that $f_k : I_1 \to \mathbb{R}$ and $g_k : I_2 \to \mathbb{R}$ are functions on intervals I_1 and I_2 of the real line. Suppose that*

$$\sum\nolimits_{k=1}^{n} f_k(a)\, g_k(b) \ge 0 \quad \text{for all } a \in I_1 \text{ and } b \in I_2.$$

Then, for each natural number d,

$$\sum\nolimits_{k=1}^{n} \bar{\mathrm{tr}}[f_k(\boldsymbol{A})\, g_k(\boldsymbol{B})] \ge 0 \quad \text{for all } \boldsymbol{A} \in \mathbb{H}^d(I_1) \text{ and } \boldsymbol{B} \in \mathbb{H}^d(I_2).$$

Proof of Lemma 4.1, General Case. We retain the notation from Lemma 4.1. In particular, we assume that $\boldsymbol{Z}$ is a random positive-definite matrix for which $\|\boldsymbol{Z}\|$ and $\|\varphi(\boldsymbol{Z})\|$ are both integrable. We also assume that $\boldsymbol{T}$ is a random positive-definite matrix with $\|\boldsymbol{T}\|$ and $\|\varphi(\boldsymbol{T})\|$ integrable.

For $n \in \mathbb{N}$, define the function $l_n(a) := (a \vee 1/n) \wedge n$, where $\vee$ denotes the maximum operator and $\wedge$ denotes the minimum operator. Consider the random matrices $\boldsymbol{Z}_n := l_n(\boldsymbol{T})$ and $\boldsymbol{T}_k := l_k(\boldsymbol{T})$ for each $k, n \in \mathbb{N}$. These matrices have eigenvalues that are bounded and bounded away from zero, so these entities satisfy the inequality (4.3) we have already established.

$$H_\varphi(\boldsymbol{Z}_n) \ge \mathbb{E}\,\bar{\mathrm{tr}}\,\big[(\psi(\boldsymbol{T}_k) - \psi(\mathbb{E}\,\boldsymbol{T}_k))(\boldsymbol{Z}_n - \boldsymbol{T}_k) + \mathbb{E}\,\varphi(\boldsymbol{T}_k - \varphi(\mathbb{E}\,\boldsymbol{T}_k)\big].$$

Rearrange the terms in this inequality to obtain

$$\mathbb{E}\,\bar{\mathrm{tr}}\,\boldsymbol{\Gamma}(\boldsymbol{Z}_n, \boldsymbol{T}_k) \ge \bar{\mathrm{tr}}\,\big[-\psi(\mathbb{E}\,\boldsymbol{T}_k)(\mathbb{E}\,\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{T}_k) - \varphi(\mathbb{E}\,\boldsymbol{T}_k) + \varphi(\mathbb{E}\,\boldsymbol{Z}_n)\big], \tag{A.1}$$

where we have introduced the function

$$\boldsymbol{\Gamma}(\boldsymbol{A}, \boldsymbol{B}) := \varphi(\boldsymbol{A}) - \varphi(\boldsymbol{B}) - (\boldsymbol{A} - \boldsymbol{B})\psi(\boldsymbol{B}) \quad \text{for } \boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}^d_{++}.$$

To complete the proof of Lemma 4.1, we must develop the bound

$$\mathbb{E}\,\bar{\mathrm{tr}}\,\boldsymbol{\Gamma}(\boldsymbol{Z}, \boldsymbol{T}) \ge \bar{\mathrm{tr}}\,\big[-\psi(\mathbb{E}\,\boldsymbol{T})(\mathbb{E}\,\boldsymbol{Z} - \mathbb{E}\,\boldsymbol{T}) - \varphi(\mathbb{E}\,\boldsymbol{T}) + \varphi(\mathbb{E}\,\boldsymbol{Z})\big] \tag{A.2}$$

by driving $k, n \to \infty$ in (A.1).

Let us begin with the right-hand side of (A.1). We have the sure limit $\boldsymbol{Z}_n \to \boldsymbol{Z}$. Therefore, the Dominated Convergence Theorem guarantees that $\mathbb{E}\,\boldsymbol{Z}_n \to \mathbb{E}\,\boldsymbol{Z}$ because $\|\boldsymbol{Z}\|$ is integrable and $\|\boldsymbol{Z}_n\| \le \|\boldsymbol{Z}\|$. Likewise, $\mathbb{E}\,\boldsymbol{T}_k \to \mathbb{E}\,\boldsymbol{T}$. The functions φ and ψ are continuous, so the limit of the right-hand side of (A.1) satisfies

$$\begin{aligned}\bar{\mathrm{tr}}\,\big[-\psi(\mathbb{E}\,\boldsymbol{T}_k)(\mathbb{E}\,\boldsymbol{Z}_n - \mathbb{E}\,\boldsymbol{T}_k) - \varphi(\mathbb{E}\,\boldsymbol{T}_k) + \varphi(\mathbb{E}\,\boldsymbol{Z}_n)\big] \\ \to \bar{\mathrm{tr}}\,\big[-\psi(\mathbb{E}\,\boldsymbol{T})(\mathbb{E}\,\boldsymbol{Z} - \mathbb{E}\,\boldsymbol{T}) - \varphi(\mathbb{E}\,\boldsymbol{T}) + \varphi(\mathbb{E}\,\boldsymbol{Z})\big].\end{aligned} \tag{A.3}$$

This expression coincides with the right-hand side of (A.2).

Taking the limit of the left-hand side of (A.1) is more involved because the function ψ may grow quickly at zero and infinity. We accomplish our goal in two steps. First, we take the limit as $n \to \infty$. Afterward, we take the limit as $k \to \infty$.

Introduce the nonnegative function

$$\gamma(z,t) := \varphi(z) - \varphi(t) - (z-t)\psi(t) \quad \text{for } z,t > 0.$$

Boucheron et al. [BBLM05, p. 525] establish that

$$\gamma(l_n(z), l_k(t)) \le \gamma(1, l_k(t)) + \gamma(z, l_k(t)) \quad \text{for } z,t > 0. \tag{A.4}$$

The generalized Klein inequality, Proposition A.1, can be applied (with due diligence) to extend (A.4) to matrices. In particular,

$$\bar{\operatorname{tr}}\, \boldsymbol{\Gamma}(\boldsymbol{Z}_n, \boldsymbol{T}_k) = \bar{\operatorname{tr}}\, \boldsymbol{\Gamma}(l_n(\boldsymbol{Z}), l_k(\boldsymbol{T})) \le \bar{\operatorname{tr}}[\boldsymbol{\Gamma}(\mathbf{I}, l_k(\boldsymbol{T})) + \boldsymbol{\Gamma}(\boldsymbol{Z}, l_k(\boldsymbol{T}))] = \bar{\operatorname{tr}}[\boldsymbol{\Gamma}(\mathbf{I}, \boldsymbol{T}_k) + \boldsymbol{\Gamma}(\boldsymbol{Z}, \boldsymbol{T}_k)].$$

Observe that the right-hand side of this inequality is integrable. Indeed, all of the quantities involving $\boldsymbol{T}_k$ are uniformly bounded because the eigenvalues of $\boldsymbol{T}_k$ fall in the range $[k^{-1}, k]$ and the functions φ and ψ are continuous on this interval. The terms involving $\boldsymbol{Z}$ may not be bounded, but they are integrable because $\|\boldsymbol{Z}\|$ and $\|\varphi(\boldsymbol{Z})\|$ are integrable. We may now apply the Dominated Convergence Theorem to take the limit:

$$\mathbb{E}\, \bar{\operatorname{tr}}\, \boldsymbol{\Gamma}(\boldsymbol{Z}_n, \boldsymbol{T}_k) \to \mathbb{E}\, \bar{\operatorname{tr}}\, \boldsymbol{\Gamma}(\boldsymbol{Z}, \boldsymbol{T}_k) \quad \text{as } n \to \infty, \tag{A.5}$$

where we rely again on the sure limit $\boldsymbol{Z}_n \to \boldsymbol{Z}$ as $n \to \infty$.

Boucheron et al. also establish that

$$\gamma(z, l_k(t)) \le \gamma(z, 1) + \gamma(z, t) \quad \text{for } z,t > 0.$$

The generalized Klein inequality, Proposition A.1, ensures that

$$\bar{\operatorname{tr}}\, \boldsymbol{\Gamma}(\boldsymbol{Z}, \boldsymbol{T}_k) \le \bar{\operatorname{tr}}[\boldsymbol{\Gamma}(\boldsymbol{Z}, \mathbf{I}) + \boldsymbol{\Gamma}(\boldsymbol{Z}, \boldsymbol{T})].$$

We may assume that the second term on the right-hand side is integrable or else the desired inequality (A.2) would be vacuous. The first term is integrable because $\|\boldsymbol{Z}\|$ and $\|\varphi(\boldsymbol{Z})\|$ are integrable. Therefore, we may apply the Dominated Convergence Theorem:

$$\mathbb{E}\, \bar{\operatorname{tr}}\, \boldsymbol{\Gamma}(\boldsymbol{Z}, \boldsymbol{T}_k) \to \mathbb{E}\, \bar{\operatorname{tr}}\, \boldsymbol{\Gamma}(\boldsymbol{Z}, \boldsymbol{T}) \quad \text{as } k \to \infty, \tag{A.6}$$

where we rely again on the sure limit $\boldsymbol{T}_k \to \boldsymbol{T}$ as $k \to \infty$.

In summary, the limits (A.5) and (A.6) provide that $\mathbb{E}\, \bar{\operatorname{tr}}\, \boldsymbol{\Gamma}(\boldsymbol{Z}_n, \boldsymbol{T}_k) \to \mathbb{E}\, \bar{\operatorname{tr}}\, \boldsymbol{\Gamma}(\boldsymbol{Z}, \boldsymbol{T})$ as $k, n \to \infty$. In view of the limit (A.3), we have completed the proof of (A.2). □

References

[AGZ10] G. W. Anderson, A. Guionnet, and O. Zeitouni. *An introduction to random matrices*, volume 118 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2010.

[And79] T. Ando. Concavity of certain maps on positive definite matrices and applications to Hadamard products. *Linear Algebra Appl.*, 26:203–241, 1979. MR-0535686

[AW02] R. Ahlswede and A. Winter. Strong converse for identification via quantum channels. *IEEE Trans. Inform. Theory*, 48(3):569–579, 2002. MR-1889969

[BBLM05] S. Boucheron, O. Bousquet, G. Lugosi, and P. Massart. Moment inequalities for functions of independent random variables. *Ann. Probab.*, 33(2):514–560, 2005. MR-2123200

[BG12] C. Boutsidis and A. Gittens. Improved matrix algorithms via the subsampled randomized Hadamard transform. Available at arXiv:1204.0062, 2012. MR-3101094

[Bha97] R. Bhatia. *Matrix Analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1997. MR-1477662

[Bha07] R. Bhatia. *Positive Definite Matrices*. Princeton Series in Applied Mathematics. Princeton University Press, Princeton, NJ, 2007. MR-2284176

[BL98] S. G. Bobkov and M. Ledoux. On modified logarithmic Sobolev inequalities for Bernoulli and Poisson measures. *J. Funct. Anal.*, 156(2):347–365, 1998. MR-1636948

[BLM03] S. Boucheron, G. Lugosi, and P. Massart. Concentration inequalities using the entropy method. *Ann. Probab.*, 31(3):1583–1614, 2003. MR-1989444

[BLM09] S. Boucheron, G. Lugosi, and P. Massart. On concentration of self-bounding functions. *Electron. J. Probab.*, 14:no. 64, 1884–1899, 2009. MR-2540852

[BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities*. Oxford University Press, 2013.

[Bou02] O. Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Math. Acad. Sci. Paris*, 334(6):495–500, 2002. MR-1890640

[Brè66] L. M. Brègman. Relaxation method for finding a common point of convex sets and its application to optimization problems. *Dokl. Akad. Nauk SSSR*, 171:1019–1022, 1966. MR-0210291

[Car10] E. Carlen. Trace inequalities and quantum entropy: an introductory course. In *Entropy and the quantum*, volume 529 of *Contemp. Math.*, pages 73–140. Amer. Math. Soc., Providence, RI, 2010. MR-2681769

[CCT12] K. Chaudhuri, F. Chung, and A. Tsiatas. Spectral clustering of graphs with general degrees in the extended planted partition model. *Journal of Machine Learning Research 2012*, pages 1–23, 2012.

[CDL13] A. Cohen, M. A. Davenport, and D. Leviatan. On the stability and accuracy of least squares approximations. *Found. Comput. Math.*, 2013. MR-3105946

[Cha04] D. Chafaï. Entropies, convexity, and functional inequalities: on Φ-entropies and Φ-Sobolev inequalities. *J. Math. Kyoto Univ.*, 44(2):325–363, 2004. MR-2081075

[Cha06] D. Chafaï. Binomial-Poisson entropic inequalities and the $M/M/\infty$ queue. *ESAIM Probab. Stat.*, 10:317–339 (electronic), 2006. MR-2247924

[Csi72] I. Csiszár. A class of measures of informativity of observation channels. *Period. Math. Hungar.*, 2:191–213, 1972. Collection of articles dedicated to the memory of Alfréd Rényi, I. MR-0335152

[DZ11] P. Drineas and A. Zouzias. A note on element-wise matrix sparsification via a matrix-valued Bernstein inequality. *Inform. Process. Lett.*, 111:385–389, 2011. MR-2760960

[Han13] F. Hansen. Trace functions with applications in quantum physics. *J. Stat. Phys.*, 2013. MR-3163550

[HOZ01] W. Hebisch, R. Olkiewicz, and B. Zegarlinski. On upper bound for the quantum entropy. *Linear Algebra Appl.*, 329:89–96, 2001. MR-1822224

[HZ14] F. Hansen and Z. Zhang. Characterisation of matrix entropies. Available at arXiv:1402.2118, Feb. 2014.

[JX03] M. Junge and Q. Xu. Noncommutative Burkholder/Rosenthal inequalities. *Ann. Probab.*, 31(2):948–995, 2003. MR-1964955

[JX08] M. Junge and Q. Xu. Noncommutative Burkholder/Rosenthal inequalities II: Applications. *Israel J. Math.*, 167:227–282, 2008. MR-2448025

[JZ11] M. Junge and Q. Zheng. Noncommutative Bennett and Rosenthal inequalities. Available at arXiv:1111.1027, Nov. 2011.

[KA80] F. Kubo and T. Ando. Means of positive linear operators. *Math. Ann.*, 246(3):205–224, 1979/80. MR-0563399

[Led99] M. Ledoux. Concentration of measure and logarithmic Sobolev inequalities. In *Séminaire de Probabilités, XXXIII*, volume 1709 of *Lecture Notes in Math.*, pages 120–216. Springer, Berlin, 1999. MR-1767995

[Led01] M. Ledoux. *The Concentration of Measure Phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2001. MR-1849347

[Led97] M. Ledoux. On Talagrand's deviation inequalities for product measures. *ESAIM Probab. Statist.*, 1:63–87 (electronic), 1995/97. MR-1399224

[Lie73] E. H. Lieb. Convex trace functions and the Wigner-Yanase-Dyson conjecture. *Advances in Math.*, 11:267–288, 1973. MR-0332080

[Lie75] E. H. Lieb. Some convexity and subadditivity properties of entropy. *Bull. Amer. Math. Soc.*, 81:1–13, 1975. MR-0356797

[Lin73] G. Lindblad. Entropy, information, and quantum measurements. *Commun. Math. Phys.*, 33:305–322, 1973. MR-0337240

[LO00] R. Latała and K. Oleszkiewicz. Between Sobolev and Poincaré. In *Geometric aspects of functional analysis*, volume 1745 of *Lecture Notes in Math.*, pages 147–168. Springer, Berlin, 2000. MR-1796718

[LP86] F. Lust-Piquard. Inégalités de Khintchine dans C_p $(1 < p < \infty)$. *C. R. Acad. Sci. Paris Sér. I Math.*, 303(7):289–292, 1986. MR-0859804

[LR73] E. H. Lieb and M. B. Ruskai. Proof of the strong subadditivity of quantum-mechanical entropy. *J. Math. Phys.*, 14(12):1938–1941, 1973. MR-0345558

[Mas00a] P. Massart. About the constants in Talagrand's concentration inequalities for empirical processes. *Ann. Probab.*, 28(2):863–884, 2000. MR-1782276

[Mas00b] P. Massart. Some applications of concentration inequalities to statistics. *Ann. Fac. Sci. Toulouse Math. (6)*, 9(2):245–303, 2000. Probability theory. MR-1813803

[Min12] S. Minsker. On some extensions of Bernstein's inequality for self-adjoint operators. Available at arXiv:1112.5448, Jan. 2012.

[MJC+12] L. Mackey, M. I. Jordan, R. Y. Chen, B. Farrell, and J. A. Tropp. Matrix concentration inequalities via the method of exchangeable pairs. Available at arXiv:1201.6002, 2012.

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

[Pet94] D. Petz. A survey of certain trace inequalities. In *Functional analysis and operator theory (Warsaw, 1992)*, volume 30 of *Banach Center Publ.*, pages 287–298. Polish Acad. Sci., Warsaw, 1994. MR-1285615

[PMT13] D. Paulin, L. Mackey, and J. A. Tropp. Deriving matrix concentration inequalities from kernel couplings. Available at arXiv:1305.0612, May 2013.

[Rén61] A. Rényi. On measures of entropy and information. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I*, pages 547–561. Univ. California Press, Berkeley, Calif., 1961. MR-0132570

[Rio01] E. Rio. Inégalités de concentration pour les processus empiriques de classes de parties. *Probab. Theory Related Fields*, 119(2):163–175, 2001. MR-1818244

[RS13] M. Raginsky and I. Sason. Concentration of measure inequalities in information theory, communications and coding. *Foundations and Trends in Communications and Information Theory*, 10(1–2):1–246, 2013. Available at arXiv:1212.4663.

[Rud99] M. Rudelson. Random vectors in the isotropic position. *J. Funct. Anal.*, 164(1):60–72, 1999. MR-1694526

[RW11] M. D. Reid and R. C. Williamson. Information, divergence and risk for binary experiments. *J. Mach. Learn. Res.*, 12:731–817, 2011. MR-2786911

[Tal96] M. Talagrand. A new look at independence. *Ann. Probab.*, 24(1):1–34, 1996. MR-1387624

[Tro11] J. A. Tropp. Freedman's inequality for matrix martingales. *Electron. Commun. Probab.*, 16:262–270, 2011. MR-2802042

[Tro12a] J. A. Tropp. Improved analysis of the subsampled randomized Hadamard transform. *Adv. Adapt. Data Anal.*, 3(1–2):115–126, 2012. MR-2835584

[Tro12b] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12(4):389–434, 2012. MR-2946459

[Tro12c] J. A. Tropp. User-friendly tools for random matrices: An introduction. Available at `http://users.cms.caltech.edu/~jtropp/notes/Tro12-User-Friendly-Tools-NIPS.pdf`, Dec. 2012.

[Voi02] D. Voiculescu. Free entropy. *Bull. London Math. Soc.*, 34(3):257–278, 2002. MR-1887698

Acknowledgments. RYC and JAT are with the Department of Computing and Mathematical Sciences, California Institute of Technology. JAT gratefully acknowledges support from ONR awards N00014-08-1-0883 and N00014-11-1002, AFOSR award FA9550-09-1-0643, and a Sloan Research Fellowship. JAT also wishes to thank the Moore Foundation.