Journal of Machine Learning Research 20 (2019) 1-40 Submitted 8/17; Revised 1/19; Published 2/19

A Bootstrap Method for Error Estimation in Randomized Matrix Multiplication

Miles E. Lopes MELOPES@UCDAVIS.EDU
Department of Statistics
University of California at Davis
Davis, CA 95616, USA

Shusen Wang SHUSEN.WANG@STEVENS.EDU
Department of Computer Science
Stevens Institute of Technology
Hoboken, NJ 07030, USA

Michael W. Mahoney MMAHONEY@STAT.BERKELEY.EDU
International Computer Science Institute and Department of Statistics
University of California at Berkeley
Berkeley, CA 94720, USA

Editor: Hui Zou

Abstract

In recent years, randomized methods for numerical linear algebra have received growing interest as a general approach to large-scale problems. Typically, the essential ingredient of these methods is some form of randomized dimension reduction, which accelerates computations, but also creates random approximation error. In this way, the dimension reduction step encodes a tradeoff between cost and accuracy. However, the exact numerical relationship between cost and accuracy is typically unknown, and consequently, it may be difficult for the user to precisely know (1) how accurate a given solution is, or (2) how much computation is needed to achieve a given level of accuracy. In the current paper, we study randomized matrix multiplication (sketching) as a prototype setting for addressing these general problems. As a solution, we develop a bootstrap method for *directly estimating* the accuracy as a function of the reduced dimension (as opposed to deriving worst-case bounds on the accuracy in terms of the reduced dimension). From a computational standpoint, the proposed method does not substantially increase the cost of standard sketching methods, and this is made possible by an "extrapolation" technique. In addition, we provide both theoretical and empirical results to demonstrate the effectiveness of the proposed method.

Keywords: matrix sketching, randomized matrix multiplication, bootstrap methods

1. Introduction

The development of randomized numerical linear algebra (RNLA or RandNLA) has led to a variety of efficient methods for solving large-scale matrix problems, such as matrix multiplication, least-squares approximation, and low-rank matrix factorization, among others (Halko et al., 2011; Mahoney, 2011; Woodruff, 2014; Drineas and Mahoney, 2016). A general feature of these methods is that they apply some form of randomized dimension reduction to an input matrix, which reduces the cost of subsequent computations. In

exchange for the reduced cost, the randomization leads to some error in the resulting solution, and consequently, there is a tradeoff between cost and accuracy.

For many canonical matrix problems, the relationship between cost and accuracy has been the focus of a growing body of theoretical work, and the literature provides many performance guarantees for RNLA methods. In general, these guarantees offer a good qualitative description of how the accuracy depends on factors such as problem size, number of iterations, condition numbers, and so on. Yet, it is also the case that such guarantees tend to be overly pessimistic for any particular problem instance — often because the guarantees are formulated to hold in the worst case among a large class of possible inputs. Likewise, it is often impractical to use such guarantees to determine precisely how accurate a given solution is, or precisely how much computation is needed to achieve a desired level of accuracy.

In light of this situation, it is of interest to develop efficient methods for estimating the exact relationship between the cost and accuracy of RNLA methods on a *problem-specific basis*. Since the literature has been somewhat quiet on this general question, the aim of this paper is to analyze randomized matrix multiplication as a prototype setting, and propose an approach that may be pursued more broadly. (Extensions are discussed at the end of the paper in Section 6.)

1.1. Randomized matrix multiplication

To describe our problem setting, we briefly review the rudiments of randomized matrix multiplication, which is often known as matrix sketching (Drineas et al., 2006a; Mahoney, 2011; Woodruff, 2014). If $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $\mathbf{B} \in \mathbb{R}^{n\times d'}$ are fixed input matrices, then sketching methods are commonly used to approximate $\mathbf{A}^T\mathbf{B}$ in the the regime where $\max\{d, d'\} \ll n$. For instance, this regime corresponds to "big data" applications where $\mathbf{A}$ and $\mathbf{B}$ are data matrices with very large numbers of observations.

As a way of reducing the cost of ordinary matrix multiplication, the main idea of sketching is to compute the product $\tilde{\mathbf{A}}^T\tilde{\mathbf{B}}$ of smaller matrices $\tilde{\mathbf{A}} \in \mathbb{R}^{t\times d}$ and $\tilde{\mathbf{B}} \in \mathbb{R}^{t\times d'}$, for some choice of $t \ll n$. These smaller matrices are referred to as "sketches", and they are generated randomly according to

$$\tilde{\mathbf{A}} := \mathbf{SA} \quad \text{and} \quad \tilde{\mathbf{B}} := \mathbf{SB}, \tag{1}$$

where $\mathbf{S} \in \mathbb{R}^{t\times n}$ is a random "sketching matrix" satisfying the condition

$$\mathbb{E}[\mathbf{S}^T\mathbf{S}] = \mathbf{I}_n, \tag{2}$$

with $\mathbf{I}_n$ being the identity matrix. In particular, the relation (2) implies that the sketched product is an unbiased estimate, $\mathbb{E}[\tilde{\mathbf{A}}^T\tilde{\mathbf{B}}] = \mathbf{A}^T\mathbf{B}$. Most commonly, the matrix $\mathbf{S}$ can be interpreted as acting on $\mathbf{A}$ and $\mathbf{B}$ by sampling their rows, or by randomly projecting their columns. In Section 2, we describe some popular examples of sketching matrices to be considered in our analysis.

1.2. Problem formulation

When sketching is implemented, the choice of the sketch size t plays a central role, since it directly controls the relationship between cost and accuracy. If t is small, then the sketched

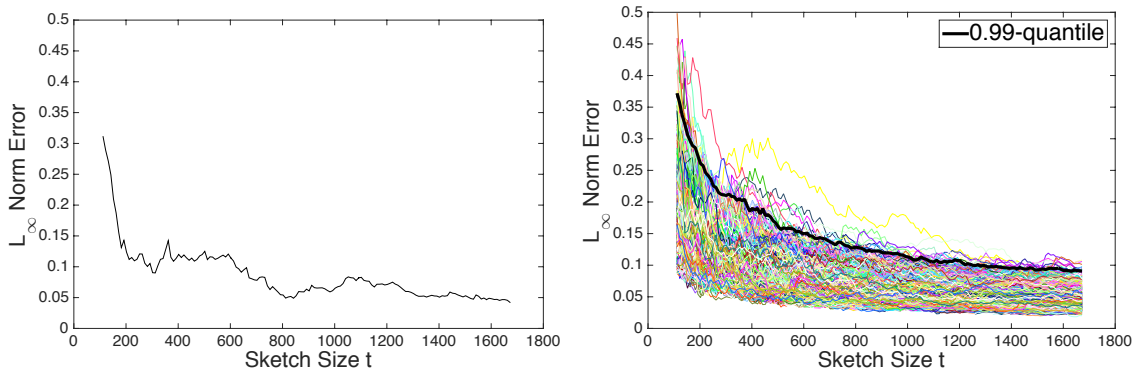


Figure 1: **Left panel:** The curve shows how ε_t fluctuates with varying sketch size t, as rows are added to $\mathbf{S}$, with $\mathbf{A}$ and $\mathbf{B}$ held fixed. (Each row of $\mathbf{A} \in \mathbb{R}^{8{,}124\times 112}$ is a feature vector of the Mushroom dataset (Frank and Asuncion, 2010), and we set $\mathbf{B} = \mathbf{A}$.) The rows of $\mathbf{S}$ were generated randomly from a Gaussian distribution (see Section 2), and the matrix $\mathbf{A}$ was scaled so that $\|\mathbf{A}^T\mathbf{A}\|_\infty = 1$. **Right panel:** There are 1,000 colored curves, each arising from a repetition of the simulation in the left panel. The thick black curve represents $q_{0.99}(t)$.

product $\tilde{\mathbf{A}}^T\tilde{\mathbf{B}}$ may be computed quickly, but it is unlikely to be a good approximation to $\mathbf{A}^T\mathbf{B}$. Conversely, if t is large, then the sketched product is more expensive to compute, but it is more likely to be accurate. For this reason, we will parameterize the relationship between cost and accuracy in terms of t.

Conventionally, the error of an approximate matrix product is measured with a norm, and in particular, we will consider error as measured by the ℓ_∞-norm,

$$\varepsilon_t \;:=\; \left\|\mathbf{A}^T\mathbf{S}^T\mathbf{S}\mathbf{B} - \mathbf{A}^T\mathbf{B}\right\|_\infty, \tag{3}$$

where $\|\mathbf{C}\|_\infty := \max_{i,j}|c_{ij}|$ for a matrix $\mathbf{C} = [c_{ij}]$. (Further background on analysis of ℓ_∞-norm or entry-wise error for matrix multiplication may be found in (Higham, 2002; Drineas et al., 2006a; Demmel et al., 2007; Pagh, 2013), among others.) In the context of sketching, it is crucial to note that ε_t is a random variable, due to the randomness in $\mathbf{S}$. Consequently, it is natural to study the quantiles of ε_t, because they specify the *tightest possible bounds* on ε_t that hold with a prescribed probability. More specifically, for any $\alpha \in (0,1)$, the $(1-\alpha)$-quantile of ε_t is defined as

$$q_{1-\alpha}(t) \;:=\; \inf\left\{q \in [0,\infty) \,\middle|\, \mathbb{P}\big(\varepsilon_t \le q\big) \;\ge\; 1-\alpha\right\}. \tag{4}$$

For example, the quantity $q_{0.99}(t)$ is the tightest upper bound on ε_t that holds with probability at least 0.99. Hence, for any fixed α, the function $q_{1-\alpha}(t)$ represents a precise *tradeoff curve* for relating cost and accuracy. Moreover, the function $q_{1-\alpha}(t)$ is specific to the input matrices $\mathbf{A}$ and $\mathbf{B}$.

To clarify the interpretation of $q_{1-\alpha}(t)$, it is helpful to plot the fluctuations of ε_t. In the left panel of Figure 1, we illustrate a simulation where randomly generated rows are

incrementally added to a sketching matrix $\mathbf{S}$, with $\mathbf{A}$ and $\mathbf{B}$ held fixed. Each time a row is added to $\mathbf{S}$, the sketch size t increases by 1, and we plot the corresponding value of ε_t as t ranges from 100 to 1,700. (Note that the user is typically unable to observe such a curve in practice.) In the right panel, we display 1,000 repetitions of the simulation, with each colored curve corresponding to one repetition. (The variation is due only to the different draws of $\mathbf{S}$.) In particular, the function $q_{0.99}(t)$ is represented by the thick black curve, delineating the top 1% of the colored curves at each value of t.

In essence, the right panel of Figure 1 shows that if the user had knowledge of the (unknown) function $q_{1-\alpha}(t)$, then two important purposes could be served. First, for any fixed value t, the user would have a sharp problem-specific bound on ε_t. Second, for any fixed error tolerance ϵ, the user could select t so that that "just enough" computation is spent in order to achieve $\varepsilon_t \leq \epsilon$ with probability at least $1 - \alpha$.

The estimation problem. The challenge we face is that a naive computation of $q_{1-\alpha}(t)$ by generating samples of ε_t would defeat the purpose of sketching. Indeed, generating samples of ε_t by brute force would require running the sketching method many times, and it would also require computing the entire product $\mathbf{A}^T\mathbf{B}$. Consequently, the technical problem of interest is to develop an efficient way to estimate $q_{1-\alpha}(t)$, without adding much cost to a *single run* of the sketching method.

1.3. Contributions

From a conceptual standpoint, the main novelty of our work is that it bridges two sets of ideas that are ordinarily studied in distinct communities. Namely, we apply the statistical technique of bootstrapping to enhance algorithms for numerical linear algebra. To some extent, this pairing of ideas might seem counterintuitive, since bootstrap methods are sometimes labeled as "computationally intensive", but it will turn out that the cost of bootstrapping can be managed in our context. Another reason our approach is novel is that we use the bootstrap to quantify error in the output of a randomized algorithm, rather than for the usual purpose of quantifying uncertainty arising from data. In this way, our approach harnesses the versatility of bootstrap methods, and we hope that our results in the "use case" of matrix multiplication will encourage broader applications of bootstrap methods in randomized computations. (See also Section 6, and note that in concurrent work, we have pursued similar approaches in the contexts of randomized least-squares and classification algorithms (Lopes et al., 2018b; Lopes, 2019).)

From a technical standpoint, our main contributions are a method for estimating the function $q_{1-\alpha}(t)$, as well as theoretical performance guarantees. Computationally, the proposed method is efficient in the sense that its cost is comparable to a single run of standard sketching methods (see Section 2). This efficiency is made possible by an "extrapolation" technique, which allows us to bootstrap small "initial" sketches with t_0 rows, and inexpensively estimate $q_{1-\alpha}(t)$ at larger values $t \gg t_0$. The empirical performance of the extrapolation technique is also quite encouraging, as discussed in Section 5. Lastly, with regard to theoretical analysis, our proofs circumvent some technical restrictions occurring in the analysis of related bootstrap methods in the statistics literature.

1.4. Related work

Several works have considered the problem of error estimation for randomized matrix computations—mostly in the context of low-rank approximation (Woolfe et al., 2008; Liberty et al., 2007; Halko et al., 2011), least squares (Lopes et al., 2018b), or matrix multiplication (Ar et al., 1993; Sarlós, 2006). With attention to matrix multiplication, the latter two papers offer methods for estimating high-probability bounds on the error $\eta_t := \|\tilde{\mathbf{A}}^T\tilde{\mathbf{B}} - \mathbf{A}^T\mathbf{B}\|$, where $\|\cdot\|$ is either the maximum absolute row sum norm, or the Frobenius norm. At a high level, all of the mentioned papers rely on a common technique, which is to randomly generate a sequence of "test-vectors", say $\mathbf{v}_1, \mathbf{v}_2, \ldots$, and then use the matrix-vector products $\mathbf{w}_i := \tilde{\mathbf{A}}^T\tilde{\mathbf{B}}\mathbf{v}_i - \mathbf{A}^T(\mathbf{B}\mathbf{v}_i)$ to derive an estimated bound, say $\hat{\eta}_t$, for η_t. The origin of this technique may be traced to the classical works (Dixon, 1983; Freivalds, 1979).

Our approach differs from the "test-vector approach" in some essential ways. One difference arises because the bounds on $\hat{\eta}_t$ are generally constructed from the vectors $\{\mathbf{w}_i\}$ using conservative inequalities. By contrast, our approach avoids this conservativeness by *directly estimating* $q_{1-\alpha}(t)$, which is an optimal bound on ε_t in the sense of equation (4).

A second difference deals with computational demands. For example, in order to compute the vectors $\{\mathbf{w}_i\}$ in the test-vector approach, it is necessary to access the full matrices $\mathbf{A}$ and $\mathbf{B}$. On the other hand, our method does not encounter this difficulty, because it only requires access to the much smaller sketches $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{B}}$. Also, in the test-vector approach, the cost to compute each vector $\mathbf{w}_i$ is proportional to the large dimension n, while the cost to compute $\hat{q}_{1-\alpha}(t)$ with our method is *independent* of n. Finally, the test-vector approach can only be used to check if the product $\tilde{\mathbf{A}}^T\tilde{\mathbf{B}}$ is accurate after it has been computed, whereas our approach can be used to dynamically "predict" an appropriate sketch size t from a small "initial" sketching matrix (see Section 3.3).

With regard to the statistics literature, our work builds upon a line of research dealing with "multiplier bootstrap methods" in high-dimensional problems (Chernozhukov et al., 2013, 2014, 2017). Such methods are well-suited to approximating the distributions of statistics such as $\|\bar{\mathbf{x}}\|_\infty$, where $\bar{\mathbf{x}} \in \mathbb{R}^p$ denotes the sample average of n independent mean-zero vectors, with $n \ll p$. More recently, this approach has been substantially extended to other "max type" statistics arising from sample covariance matrices (Chang et al., 2016; Chen, 2018). Nevertheless, the strong results in these works do not readily translate to our context, either because the statistics are substantially different from the ℓ_∞-norm (Chang et al., 2016), or because of technical assumptions (Chen, 2018). For instance, if the results in the latter work are applied to a sample covariance matrix of the form $\frac{1}{n}\sum_{i=1}^n(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top$, where $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ are mean-zero i.i.d. vectors, with $\mathbf{x}_1 = (X_{11}, \ldots, X_{1p})$, then it is necessary to make assumptions such as $\min_{j,k} \mathsf{var}(X_{1j}X_{1k}) \geq c$, for some constant $c > 0$. As this relates to the sketching context, note that the sketched product may be written as $\tilde{\mathbf{A}}^T\tilde{\mathbf{B}} = \frac{1}{t}\sum_{i=1}^t \mathbf{A}^T\mathbf{s}_i\mathbf{s}_i^T\mathbf{B}$, where $\mathbf{s}_1, \ldots, \mathbf{s}_n \in \mathbb{R}^t$ are the rows of $\sqrt{t}\mathbf{S}$. It follows that analogous variance assumptions would lead to conditions on the matrices $\mathbf{A}$ and $\mathbf{B}$ that could be violated if any column of $\mathbf{A}$ or $\mathbf{B}$ has many small entries, or is sparse. By contrast, our results do not rely on such variance assumptions, and we allow the matrices $\mathbf{A}$ and $\mathbf{B}$ to be unrestricted.

At a more technical level, the ability to avoid restrictions on $\mathbf{A}$ and $\mathbf{B}$ comes from our use of the Lévy-Prohorov metric for distributional approximations — which differs from the Kolmogorov metric that has been predominantly used in previous works on multiplier bootstrap methods. More specifically, analyses based on the Kolmogorov metric typically rely on "anti-concentration inequalities" (Chernozhukov et al., 2013, 2015), which ultimately lead to the mentioned variance assumptions. On the other hand, our approach based on the Lévy-Prohorov metric does not require the use of anti-concentration inequalities. Finally it should be mentioned that the techniques used to control the LP metric are related to those that have been developed for bootstrap approximations via coupling inequalities as in Chernozhukov et al. (2016).

Outline. This paper is organized as follows. Section 2 introduces some technical background. Section 3 describes the proposed bootstrap algorithm. Section 4 establishes the main theoretical results, and then numerical performance is illustrated in Section 5. Lastly, conclusions and extensions of the method are presented in Section 6, and all proofs are given in the appendices.

2. Preliminaries

Notation and terminology. The set $\{1, \dots, n\}$ is denoted as $[n]$. The ith standard basis vector is denoted as $\mathbf{e}_i$. If $\mathbf{C} = [c_{ij}]$ is a real matrix, then $\|\mathbf{C}\|_F = (\sum_{i,j} c_{ij}^2)^{1/2}$ is the Frobenius norm, and $\|\mathbf{C}\|_2$ is the spectral norm (maximum singular value). If X is a random variable and $p \geq 1$, we write $\|X\|_p = (\mathbb{E}[|X|^p])^{1/p}$ for the usual L_p norm. If $\psi : [0, \infty) \to [0, \infty)$ is a non-decreasing convex function with $\psi(0) = 0$, then the ψ-Orlicz norm of X is defined as $\|X\|_\psi := \inf\{r > 0 \mid \mathbb{E}[\psi(|X|/r)] \leq 1\}$. In particular, we define $\psi_p(x) := \exp(x^p) - 1$ for $p \geq 1$, and we say that X is sub-Gaussian when $\|X\|_{\psi_2} < \infty$, or sub-exponential when $\|X\|_{\psi_1} < \infty$. In Appendix F, Lemma 9 summarizes the facts about Orlicz norms that will be used.

We will use c to denote a positive absolute constant that may change from line to line. The matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{S}$ are viewed as lying in a sequence of matrices indexed by the tuple (d, d', t, n). For a pair of generic functions f and g, we write $f(d, d', t, n) \lesssim g(d, d', t, n)$ when there is a positive absolute constant c so that $f(d, d', t, n) \leq c\, g(d, d', t, n)$ holds for all large values of d, d', t, and n. Furthermore, if a and b are two quantities that satisfy both $a \lesssim b$ and $b \lesssim a$, then we write $a \asymp b$. Lastly, we do not use the symbols $\lesssim$ or $\asymp$ when relating random variables.

Examples of sketching matrices. Our theoretical results will deal with three common types of sketching matrices, reviewed below.

- *Row sampling.* If $(p_1, \dots, p_n)$ is a probability vector, then $\mathbf{S} \in \mathbb{R}^{t \times n}$ can be constructed by sampling its rows i.i.d. from the set $\{\frac{1}{\sqrt{tp_1}}\mathbf{e}_1, \dots, \frac{1}{\sqrt{tp_n}}\mathbf{e}_n\} \subset \mathbb{R}^n$, where the vector $\frac{1}{\sqrt{tp_i}}\mathbf{e}_i$ is selected with probability p_i. Some of the most well known choices for the sampling probabilities include *uniform sampling*, with $p_i \equiv 1/n$, *length sampling* (Drineas et al., 2006a; Magen and Zouzias, 2011), with

$$p_i = \frac{\|\mathbf{e}_i^T \mathbf{A}\|_2 \|\mathbf{e}_i^T \mathbf{B}\|_2}{\sum_{j=1}^n \|\mathbf{e}_j^T \mathbf{A}\|_2 \|\mathbf{e}_j^T \mathbf{B}\|_2}, \tag{5}$$

and *leverage score sampling*, for which further background may be found in the papers (Drineas et al., 2006b, 2008, 2012).

- *Sub-Gaussian projection.* Gaussian projection is the most well-known random projection method, and is sometimes referred to as the Johnson-Lindenstrauss (JL) transform (Johnson and Lindenstrauss, 1984). In detail, if $\mathbf{G} \in \mathbb{R}^{t\times n}$ is a standard Gaussian matrix, with entries that are i.i.d. samples from $\mathcal{N}(0,1)$, then $\mathbf{S} = \frac{1}{\sqrt{t}}\mathbf{G}$ is a Gaussian projection matrix. More generally, the entries of $\mathbf{G}$ can be drawn i.i.d. from a zero-mean sub-Gaussian distribution, which often leads to similar performance characteristics in RNLA applications.

- *Subsampled randomized Hadamard transform (SRHT).* Let n be a power of 2, and define the Walsh-Hadamard matrix $\mathbf{H}_n$ recursively*

$$\mathbf{H}_n := \begin{pmatrix} \mathbf{H}_{n/2} & \mathbf{H}_{n/2} \\ \mathbf{H}_{n/2} & -\mathbf{H}_{n/2} \end{pmatrix} \quad \text{with} \quad \mathbf{H}_2 := \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

  Next, let $\mathbf{D}_n^\circ \in \mathbb{R}^{n\times n}$ be random diagonal matrix with independent ± 1 Rademacher variables along the diagonal, and let $\mathbf{P} \in \mathbb{R}^{t\times n}$ have rows uniformly sampled from $\{\frac{1}{\sqrt{t/n}}\mathbf{e}_1, \ldots, \frac{1}{\sqrt{t/n}}\mathbf{e}_n\}$. Then, the $t \times n$ matrix

$$\mathbf{S} = \mathbf{P}(\tfrac{1}{\sqrt{n}}\mathbf{H}_n)\mathbf{D}_n^\circ \tag{6}$$

  is called an SRHT matrix. This type of sketching matrix was introduced in the seminal paper (Ailon and Chazelle, 2006), and additional details regarding implementation may be found in the papers (Drineas et al., 2011; Wang, 2015). (The factor $\frac{1}{\sqrt{n}}$ is used so that $\frac{1}{\sqrt{n}}\mathbf{H}_n$ is an orthogonal matrix.) An important property of SRHT matrices is that they can be multiplied with any $n \times d$ matrix in $\mathcal{O}(n \cdot d \cdot \log t)$ time (Ailon and Liberty, 2009), which is faster than the $\mathcal{O}(n \cdot d \cdot t)$ time usually required for a dense sketching matrix.

3. Methodology

Before presenting our method in algorithmic form, we first explain the underlying intuition.

3.1. Intuition for multiplier bootstrap method

If the row vectors of $\sqrt{t}\mathbf{S}$ are denoted $\mathbf{s}_1, \ldots, \mathbf{s}_t \in \mathbb{R}^n$, then $\mathbf{S}^T\mathbf{S}$ may be conveniently expressed as a sample average

$$\mathbf{S}^T\mathbf{S} = \tfrac{1}{t}\textstyle\sum_{i=1}^t \mathbf{s}_i\mathbf{s}_i^T. \tag{7}$$

For row sampling, Gaussian projection, and SRHT, these row vectors satisfy $\mathbb{E}[\mathbf{s}_i\mathbf{s}_i^T] = \mathbf{I}_n$. Consequently, if we define the random $d \times d'$ rank-1 (dyad) matrix

$$\mathbf{D}_i \;=\; \mathbf{A}^T\mathbf{s}_i\mathbf{s}_i^T\mathbf{B}, \tag{8}$$

*The restriction that n is a power of 2 can be relaxed with variants of SRHT matrices (Avron et al., 2010; Boutsidis and Gittens, 2013).

then $\mathbb{E}[\mathbf{D}_i] = \mathbf{A}^T\mathbf{B}$, and it follows that the difference between the sketched and unsketched products can be viewed as a sample average of zero-mean random matrices

$$\mathbf{A}^T\mathbf{S}^T\mathbf{S}\mathbf{B} - \mathbf{A}^T\mathbf{B} \;=\; \tfrac{1}{t}\textstyle\sum_{i=1}^{t}(\mathbf{D}_i - \mathbf{A}^T\mathbf{B}). \tag{9}$$

Furthermore, in the cases of length sampling and Gaussian projection, the matrices $\mathbf{D}_1, \ldots, \mathbf{D}_t$ are independent, and in the case of SRHT sketches, these matrices are "nearly" independent. So, in light of the central limit theorem, it is natural to suspect that the random matrix (9) will be well-approximated (in distribution) by a matrix with Gaussian entries. In particular, if we examine the (j_1, j_2) entry, then we may expect that $\mathbf{e}_{j_1}^T\big(\mathbf{A}^T\mathbf{S}^T\mathbf{S}\mathbf{B} - \mathbf{A}^T\mathbf{B}\big)\mathbf{e}_{j_2}$ will approximately follow the distribution $\mathcal{N}(0, \frac{1}{t}\sigma_{j_1,j_2}^2)$, where the unknown parameter σ_{j_1,j_2}^2 can be estimated with

$$\hat{\sigma}_{j_1,j_2}^2 \;:=\; \tfrac{1}{t}\textstyle\sum_{i=1}^{t}\big(\mathbf{e}_{j_1}^T(\mathbf{D}_i - \mathbf{A}^T\mathbf{S}^T\mathbf{S}\mathbf{B})\mathbf{e}_{j_2}\big)^2.$$

Based on these considerations, the idea of the proposed bootstrap method is to generate a random matrix whose (j_1, j_2) entry is sampled from $\mathcal{N}(0, \frac{1}{t}\hat{\sigma}_{j_1,j_2}^2)$. It turns out that an efficient way of generating such a matrix is to sample i.i.d. random variables $\xi_1, \ldots, \xi_t \sim \mathcal{N}(0,1)$, independent of $\mathbf{S}$, and then compute

$$\tfrac{1}{t}\textstyle\sum_{i=1}^{t}\xi_i\big(\mathbf{D}_i - \mathbf{A}^T\mathbf{S}^T\mathbf{S}\mathbf{B}\big). \tag{10}$$

In other words, if $\mathbf{S}$ is conditioned upon, then the distribution of the (j_1, j_2) entry of the above matrix is exactly $\mathcal{N}(0, \frac{1}{t}\hat{\sigma}_{j_1,j_2}^2)$.[†] Hence, if the matrix (10) is viewed as an "approximate sample" of $\mathbf{A}^T\mathbf{S}^T\mathbf{S}\mathbf{B} - \mathbf{A}^T\mathbf{B}$, then it is natural to use the ℓ_∞-norm of the matrix (10) as an approximate sample of $\varepsilon_t = \|\mathbf{A}^T\mathbf{S}^T\mathbf{S}\mathbf{B} - \mathbf{A}^T\mathbf{B}\|_\infty$. Likewise, if we define the bootstrap sample

$$\varepsilon_t^\star \;:=\; \Big\|\tfrac{1}{t}\textstyle\sum_{i=1}^{t}\xi_i\Big(\mathbf{D}_i - \mathbf{A}^T\mathbf{S}^T\mathbf{S}\mathbf{B}\Big)\Big\|_\infty, \tag{11}$$

then the bootstrap algorithm will generate i.i.d. samples of $\varepsilon_t^\star$, conditionally on $\mathbf{S}$. In turn, the $(1-\alpha)$-quantile of the bootstrap samples, say $\hat{q}_{1-\alpha}(t)$, can be used to estimate $q_{1-\alpha}(t)$.

3.2. Multiplier bootstrap algorithm

We now explain how proposed method can be implemented in just a few lines. This description also reveals the important fact that the algorithm only requires access to the sketches $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{B}}$ (rather than the full matrices $\mathbf{A}$ and $\mathbf{B}$). Although the formula for generating samples of $\varepsilon_t^\star$ given below may appear different from equation (11), it is straightforward to check that these are equivalent. Lastly, the choice of the number of bootstrap samples B will be discussed at the end of subsection 3.3.

[†]It is also possible to show that the *joint* distribution of the entries in the matrix (10) mimics that of $\mathbf{A}^T\mathbf{S}^T\mathbf{S}\mathbf{B} - \mathbf{A}^T\mathbf{B}$, but we omit such details to simplify the discussion.

Algorithm 1 (Multiplier bootstrap for ε_t).
Input: *the number of bootstrap samples B, and the sketches $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{B}}$.*
For $b = 1, \ldots, B$ **do**

1. *Draw an i.i.d. sample $\xi_1, \ldots, \xi_t$ from $\mathcal{N}(0,1)$, independent of $\mathbf{S}$;*
2. *Compute the bootstrap sample $\varepsilon^\star_{t,b} := \left\| \bar{\xi} \cdot (\tilde{\mathbf{A}}^T \tilde{\mathbf{B}}) - \tilde{\mathbf{A}}^T \boldsymbol{\Xi} \tilde{\mathbf{B}} \right\|_\infty$, where $\bar{\xi} := \frac{1}{t} \sum_{i=1}^t \xi_i$ and $\boldsymbol{\Xi} := \mathsf{diag}(\xi_1, \ldots, \xi_t)$.*

Return: $\hat{q}_{1-\alpha}(t) \longleftarrow$ *the $(1-\alpha)$-quantile of the values $\varepsilon^\star_{t,1}, \ldots, \varepsilon^\star_{t,B}$.*

3.3. Saving on computation with extrapolation

In its basic form, the cost of Algorithm 1 is $\mathcal{O}(B \cdot t \cdot d \cdot d')$, which has the favorable property of being independent of the large dimension n. Also, the computation of the samples $\varepsilon^\star_{t,1}, \ldots, \varepsilon^\star_{t,B}$ is embarrassingly parallel, with the cost of each sample being $\mathcal{O}(t \cdot d \cdot d')$. Moreover, due to the way that the quantile $q_{1-\alpha}(t)$ scales with t, it is possible to reduce the cost of Algorithm 1 even further — via the technique of extrapolation (also called Richardson extrapolation) (Sidi, 2003; Brezinski and Zaglia, 2013).

The essential idea of extrapolation is to carry out Algorithm 1 for a modest "initial" sketch size t_0, and then use an initial estimate $\hat{q}_{1-\alpha}(t_0)$ to "look ahead" and predict a larger value t for which $q_{1-\alpha}(t)$ is small enough to satisfy the user's desired level of accuracy. The immediate benefit of this approach is that Algorithm 1 only needs to applied to small "initial versions" of $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{B}}$, each with t_0 rows, which reduces the cost of the algorithm to $\mathcal{O}(B \cdot t_0 \cdot d \cdot d')$. Furthermore, this means that if Algorithm 1 is run in parallel, then it is only necessary to communicate copies of the small initial sketching matrices. (To illustrate the small size of the initial sketching matrices, our experiments include several examples where the ratio t_0/n is approximately 1/100 or less.)

From a theoretical viewpoint, our use of extrapolation is based on the approximation $q_{1-\alpha}(t) \approx \frac{\kappa}{\sqrt{t}}$, where t is sufficiently large, and $\kappa = \kappa(\mathbf{A}, \mathbf{B}, \alpha)$ is an unknown number. A formal justification for this approximation can be made using Proposition 3 in Appendix A, but it is simpler to give an intuitive explanation here. Recall from Section 3.1 that as t becomes large, the (j_1, j_2) entry $[\tilde{\mathbf{A}}^T \tilde{\mathbf{B}} - \mathbf{A}^T \mathbf{B}]_{j_1, j_2}$ should be well-approximated in distribution by a Gaussian random variable of the form $\frac{1}{\sqrt{t}} G_{j_1, j_2}$. In turn, this suggests that ε_t should be well-approximated in distribution by $\frac{1}{\sqrt{t}} \max_{j_1, j_2} |G_{j_1, j_2}|$, which has quantiles that are proportional to $\frac{1}{\sqrt{t}}$.

In order to take advantage of the theoretical scaling $q_{1-\alpha}(t) \approx \frac{\kappa}{\sqrt{t}}$, we may use Algorithm 1 to compute $\hat{q}_{1-\alpha}(t_0)$ with an initial sketch size t_0, and then approximate the value $q_{1-\alpha}(t)$ for $t \gg t_0$ with the following extrapolated estimator

$$\hat{q}^{\,\mathrm{ext}}_{1-\alpha}(t) := \frac{\sqrt{t_0}}{\sqrt{t}} \hat{q}_{1-\alpha}(t_0). \tag{12}$$

Hence, if the user would like to determine a sketch size t so that $q_{1-\alpha}(t) \leq \epsilon$, for some tolerance ϵ, then t should be selected so that $\hat{q}^{\,\mathrm{ext}}_{1-\alpha}(t) \leq \epsilon$, which is equivalent to

$$t \geq \left(\frac{\sqrt{t_0}}{\epsilon} \, \hat{q}_{1-\alpha}(t_0) \right)^2. \tag{13}$$

In our experiments in Section 5, we illustrate some examples where an accurate estimate of $q_{1-\alpha}(t)$ at $t = 10{,}000$ can be obtained from the rule (13) using an initial sketch size $t_0 \approx 500$, yielding a roughly 20-fold speedup on the basic version of Algorithm 1.

Comparison with the cost of sketching. Given that the purpose of Algorithm 1 is to enhance sketching methods, it is important to understand how the added cost of the bootstrap compares to the cost of running sketching methods in the standard way. As a point of reference, we compare with the cost of computing $\mathbf{A}^T\mathbf{S}^T\mathbf{S}\mathbf{B}$ when $\mathbf{S}$ is chosen to be an SRHT matrix, since this is one of the most efficient sketching methods. If we temporarily assume for simplicity that $\mathbf{A}$ and $\mathbf{B}$ are both of size $n \times d$, then it follows from Section 2 that computing $\mathbf{A}^T\mathbf{S}^T\mathbf{S}\mathbf{B}$ has a cost of order $\mathcal{O}(t \cdot d^2 + n \cdot d \cdot \log(t))$. Meanwhile, the cost of running Algorithm 1 with the extrapolation speedup based on an initial sketch size t_0 is $\mathcal{O}(B \cdot t_0 \cdot d^2)$. Consequently, the extra cost of the bootstrap does not exceed the stated cost of sketching when the number of bootstrap samples satisfies

$$B = \mathcal{O}(\tfrac{t}{t_0} + \tfrac{n \log(t)}{d\, t_0}), \tag{14}$$

and in fact, this could be improved further if parallelization of Algorithm 1 is taken into account. It is also important to note that rather small values of B are shown to work well in our experiments, such as $B = 20$. Hence, as long t_0 remains fairly small compared to t, then the condition (14) may be expected to hold, and this is borne out in our experiments. The same reasoning also applies when $n \log(t) \gg d \cdot t_0$, which conforms with the fact that sketching methods are intended to handle situations where n is very large.

3.4. Relation with the non-parametric bootstrap

For readers who are more familiar with the "non-parametric bootstrap" (based on sampling with replacement), the purpose of this short subsection is to explain the relationship with the multiplier bootstrap in Algorithm 1. Indeed, an understanding of this relationship may be helpful, since the non-parametric bootstrap might be viewed as more intuitive, and perhaps easier to generalize to more complex situations. However, it turns out that Algorithm 1 is technically more convenient to analyze, and that is why the paper focuses primarily on Algorithm 1. Meanwhile, from a practical point of view, there is little difference between the two approaches, since both have the same order of computational cost, and in our experience, we have observed essentially the same performance in simulations. Also, the extrapolation technique can be applied to both algorithms in the same way.

To spell out the connection, the only place where Algorithm 1 needs to be changed is in step 1. Rather than choosing the multiplier variables $\xi_1, \ldots, \xi_t$ to be i.i.d. $\mathcal{N}(0,1)$ as in Algorithm 1, the non-parametric bootstrap chooses $\xi_i = \zeta_i - 1$, where $(\zeta_1, \ldots, \zeta_t)$ is a sample from a multinomial distribution, based on tossing t balls into t equally likely bins, where ζ_i is the number of balls in bin i. Hence, the mean and variance of each ξ_i are nearly the same as before, with $\mathbb{E}[\xi_i] = 0$ and $\mathsf{var}(\xi_i) = 1 - 1/t$, but the variables $\xi_1, \ldots, \xi_t$ are no longer independent.

From a more algorithmic viewpoint, it is simple to check that the choice of $\xi_1, \ldots, \xi_t$ based on the multinomial distribution is equivalent to sampling with replacement from the rows of $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{B}}$. The underlying intuition for this approach is based on the fact that for many types of sketching matrices, the rows of $\mathbf{S}$ are i.i.d., which makes the rows of $\tilde{\mathbf{A}}$ i.i.d.,

and likewise for $\tilde{\mathbf{B}}$. Hence, if $\mathbf{S}$ is conditioned upon, then sampling with replacement from the rows of $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{B}}$ imitates the random mechanism that originally generated $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{B}}$.

Algorithm 2 (Non-parametric bootstrap for ε_t).

Input: *the number of samples B, and the sketches $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{B}}$.*

For $b = 1, \ldots, B$ **do**

1. *Draw a vector $(i_1, \ldots, i_t)$ by sampling t numbers with replacement from $\{1, \ldots, t\}$.*
2. *Form matrices $\tilde{\mathbf{A}}^* \in \mathbb{R}^{t \times d}$ and $\tilde{\mathbf{B}}^* \in \mathbb{R}^{t \times d'}$ by selecting (respectively) the rows from $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{B}}$ that are indexed by $(i_1, \ldots, i_t)$.*
3. *Compute the bootstrap sample $\varepsilon^*_{t,b} := \big\|(\tilde{\mathbf{A}}^*)^T(\tilde{\mathbf{B}}^*) - \tilde{\mathbf{A}}^T\tilde{\mathbf{B}}\big\|_\infty$.*

Return: $\hat{q}_{1-\alpha}(t) \longleftarrow$ *the $(1-\alpha)$-quantile of the values $\varepsilon^*_{t,1}, \ldots, \varepsilon^*_{t,B}$.*

4. Main results

Our main results quantify how well the estimate $\hat{q}_{1-\alpha}(t)$ from Algorithm 1 approximates the true value $q_{1-\alpha}(t)$, and this will be done by analyzing how well the distribution of a bootstrap sample $\varepsilon^\star_{t,1}$ approximates the distribution of ε_t. For the purposes of comparing distributions, we will use the Lévy-Prohorov metric, defined below.

Lévy-Prohorov (LP) metric. Let $\mathcal{L}(U)$ denote the distribution of a random variable U, and let $\mathscr{B}$ denote the collection of Borel subsets of $\mathbb{R}$. For any $A \in \mathscr{B}$, and $\delta > 0$, define the δ-neighborhood $A^\delta := \big\{x \in \mathbb{R} \,\big|\, \inf_{y \in A} |x - y| \le \delta\big\}$. Then, for any two random variables U and V, the d_{LP} metric between their distributions is given by

$$d_{\mathrm{LP}}(\mathcal{L}(U), \mathcal{L}(V)) \;:=\; \inf\Big\{\delta > 0 \,\Big|\, \mathbb{P}(U \in A) \le \mathbb{P}(V \in A^\delta) + \delta \;\text{ for all }\; A \in \mathscr{B}\Big\}.$$

The d_{LP} metric is a standard tool for comparing distributions, due to the fact that convergence with respect to d_{LP} is equivalent to convergence in distribution (Huber and Ronchetti, 2009, Theorem 2.9).

Approximating quantiles. An important property of the d_{LP} metric is that if two distributions are close in this metric, then their quantiles are close in the following sense. Recall that if F_U is the distribution function of a random variable U, then the $(1-\alpha)$-quantile of U is the same as the generalized inverse $F_U^{-1}(1-\alpha) := \inf\{q \in [0, \infty) \,|\, F_U(q) \ge 1 - \alpha\}$. Next, suppose that two random variables U and V satisfy

$$d_{\mathrm{LP}}\big(\mathcal{L}(U),\, \mathcal{L}(V)\big) \;\le\; \epsilon,$$

for some $\epsilon \in (0, \alpha)$ with $\alpha \in (0, 1/2)$. Then, the quantiles of U and V are close in the sense that

$$\big|F_U^{-1}(1-\alpha) - F_V^{-1}(1-\alpha)\big| \;\le\; \psi_\alpha(\epsilon), \tag{15}$$

where the function $\psi_\alpha(\epsilon) := F_U^{-1}(1-\alpha+\epsilon) - F_U^{-1}(1-\alpha-\epsilon) + \epsilon$ is strictly monotone, and satisfies $\psi_\alpha(0) = 0$. (For a proof, see Lemma 15 of Appendix F.) In light of this fact, it will

be more convenient to express our results for approximating $q_{1-\alpha}(t)$ in terms of the d_{LP} metric.

4.1. Statements of results

Our main assumption involves three separate cases, corresponding to different choices of the sketching matrix $\mathbf{S}$.

Assumption 1 *The dimensions d and d' satisfy $d \asymp d'$. Also, there is a positive absolute constant $\kappa \geq 1$ such that $d^{1/\kappa} \lesssim t \lesssim d^{\kappa}$, which is to say that neither d nor t grows exponentially with the other. In addition, one of the following sets of conditions holds, involving the parameter $\nu(\mathbf{A}, \mathbf{B}) := \sqrt{\|\mathbf{A}^T\mathbf{A}\|_\infty \|\mathbf{B}^T\mathbf{B}\|_\infty}$.*

(a) (Sub-Gaussian case). The entries of the matrix $\mathbf{S} = [S_{i,j}]$ are zero-mean i.i.d. sub-Gaussian random variables, with $\mathbb{E}[S_{i,j}^2] = \frac{1}{t}$, and $\max_{i,j} \|\sqrt{t} S_{i,j}\|_{\psi_2} \lesssim 1$. Furthermore, $t \gtrsim \nu(\mathbf{A}, \mathbf{B})^{2/3} (\log d)^5$.

(b) (Length sampling case). The matrix $\mathbf{S}$ is generated by length sampling, with the probabilities in equation (5), and also, $t \gtrsim (\|\mathbf{A}\|_F \|\mathbf{B}\|_F)^{2/3} (\log d)^5$.

(c) (SRHT case). The matrix $\mathbf{S}$ is an SRHT matrix as defined in equation (6), and also, $t \gtrsim \nu(\mathbf{A}, \mathbf{B})^{2/3} (\log n)^2 (\log d)^5$.

Clarifications on bootstrap approximation. Before stating our main results below, it is worth clarifying a few technical items. First, since our analysis involves central limit type approximations of $\tilde{\mathbf{A}}^T\tilde{\mathbf{B}} - \mathbf{A}^T\mathbf{B}$ as a sum of t independent matrices, we will rescale the error variables by a factor of $\sqrt{t}$, obtaining

$$Z_t := \sqrt{t}\varepsilon_t, \tag{16}$$

as well as its bootstrap analogue,

$$Z_t^\star := \sqrt{t}\varepsilon_t^\star. \tag{17}$$

With regard to the original problem of estimating the quantile $q_{1-\alpha}(t)$ for ε_t, this rescaling makes no essential difference, since quantiles are homogenous with respect to scaling, and in particular, the $(1-\alpha)$-quantile of Z_t is simply $\sqrt{t} q_{1-\alpha}(t)$.

As a second clarification, recall that the bootstrap method generates samples $\varepsilon_t^\star$ based upon a particular realization of $\mathbf{S}$. For this reason, the bootstrap approximation to $\mathcal{L}(Z_t)$ is the conditional distribution $\mathcal{L}(Z_t^\star | \mathbf{S})$. Consequently, it should be noted that $\mathcal{L}(Z_t^\star | \mathbf{S})$ is a random probability measure, and $d_{\mathrm{LP}}(\mathcal{L}(Z_t)\,,\, \mathcal{L}(Z_t^\star|\mathbf{S}))$ is a random variable, since they both depend on the random matrix $\mathbf{S}$.

Theorem 1 *Let $h(x) = x^{1/2} + x^{3/4}$ for $x \geq 0$. If Assumption 1 (a) holds, then there is an absolute constant $c > 0$ such that the following bound holds with probability at least $1 - \frac{1}{t} - \frac{1}{dd'}$,*

$$d_{\mathrm{LP}}\Big(\mathcal{L}(Z_t)\,,\, \mathcal{L}(Z_t^\star|\mathbf{S})\Big) \;\leq\; \frac{c \cdot h(\nu(\mathbf{A},\mathbf{B})) \cdot \sqrt{\log(d)}}{t^{1/8}}.$$

If Assumption 1 (b) holds, then there is an absolute constant $c > 0$ such that the following bound holds with probability at least $1 - \frac{1}{t} - \frac{1}{dd'}$,

$$d_{\mathrm{LP}}\Big(\mathcal{L}(Z_t)\,,\, \mathcal{L}(Z_t^\star|\mathbf{S})\Big) \;\leq\; \frac{c \cdot h(\|\mathbf{A}\|_F\|\mathbf{B}\|_F) \cdot \sqrt{\log(d)}}{t^{1/8}}.$$

Remarks. A noteworthy property of the bounds is that they are *dimension-free* with respect to the large dimension n. Also, they have a very mild logarithmic dependence on d. With regard to the dependence on t, there are two other important factors to keep in mind. First, the practical performance of the bootstrap method (shown in Section 5) is much better than what the $t^{-1/8}$ rate suggests. Second, the problem of finding the optimal rates of approximation for multiplier bootstrap methods is a largely open problem — even in the simpler setting of bootstrapping the coordinate-wise maximum of vectors (rather than matrices). In the vector context, the literature has focused primarily on the Kolmogorov metric (rather than the LP metric), and some quite recent improvements beyond the $t^{-1/8}$ rate have been developed in Chernozhukov et al. (2017) and Lopes et al. (2018a). However, these works also rely on model assumptions that would lead to additional restrictions on the matrices $\mathbf{A}$ and $\mathbf{B}$ in our setup. Likewise, the problem of extending our results to achieve faster rates or handle other metrics is a natural direction for future work.

The SRHT case. For the case of SRHT matrices, the analogue of Theorem 1 needs to be stated in a slightly different way for technical reasons. From a qualitative standpoint, the results for SRHT and sub-Gaussian matrices turn out to be similar.

The technical issue to be handled is that the rows of an SRHT matrix are not independent, due to their common dependence on the matrix $\mathbf{D}_n^\circ$. Fortunately, this inconvenience can be addressed by conditioning on $\mathbf{D}_n^\circ$. Theoretically, this simplifies the analysis of the bootstrap, since it "decouples" the rows of the SRHT matrix. Meanwhile, if we let $\tilde{q}_{1-\alpha}(t)$ denote the $(1-\alpha)$-quantile of the distribution $\mathcal{L}(\varepsilon_t|\mathbf{D}_n^\circ)$,

$$\tilde{q}_{1-\alpha}(t) \;:=\; \inf\Big\{q \in [0,\infty)\Big|\; \mathbb{P}(\varepsilon_t \leq q\big|\mathbf{D}_n^\circ) \geq 1-\alpha\Big\},$$

then it is simple to check that $\tilde{q}_{1-\alpha}(t)$ acts as a "surrogate" for $q_{1-\alpha}(t)$, since[‡]

$$\begin{aligned}\mathbb{P}(\varepsilon_t \leq \tilde{q}_{1-\alpha}(t)) &= \mathbb{E}\big[\mathbb{P}\big(\varepsilon_t \leq \tilde{q}_{1-\alpha}(t)\big|\mathbf{D}_n^\circ\big)\big] \\ &\geq \mathbb{E}[1-\alpha] \\ &= 1-\alpha.\end{aligned} \tag{18}$$

[‡]It is also possible to show that $\tilde{q}_{1-\alpha}(t)$ fluctuates around $q_{1-\alpha}(t)$. Indeed, if we define the random variable $V := \mathbb{P}(\varepsilon_t \leq q_{1-\alpha}(t)|\mathbf{D}_n^\circ)$, it can be checked that the event $V \geq 1-\alpha$ is equivalent to the event $\tilde{q}_{1-\alpha}(t) \leq q_{1-\alpha}(t)$. Furthermore, if we suppose that $1-\alpha$ lies in the range of the c.d.f. of ε_t, then $\mathbb{E}[V] = 1-\alpha$. In turn, it follows that the event $\tilde{q}_{1-\alpha}(t) \leq q_{1-\alpha}(t)$ occurs when $V \geq \mathbb{E}[V]$, and conversely, the event $\tilde{q}_{1-\alpha}(t) > q_{1-\alpha}(t)$ occurs when $V < \mathbb{E}[V]$.

For this reason, we will view $\tilde{q}_{1-\alpha}(t)$ as the new parameter to estimate (instead of $q_{1-\alpha}(t)$), and accordingly, the aim of the following result is to quantify how well the bootstrap distribution $\mathcal{L}(Z_t^\star|\mathbf{S})$ approximates the conditional distribution $\mathcal{L}(Z_t|\mathbf{D}_n^\circ)$.

Theorem 2 *Let $h(x) = x^{1/2} + x^{3/4}$ for $x \geq 0$. If Assumption 1 (c) holds, then there is an absolute constant $c > 0$ such that the following bound holds with probability at least $1 - \frac{1}{t} - \frac{1}{dd'} - \frac{c}{n}$,*

$$d_{\mathrm{LP}}\Big(\mathcal{L}(Z_t|\mathbf{D}_n^\circ)\,,\ \mathcal{L}(Z_t^\star|\mathbf{S})\Big) \quad \leq \quad \frac{c \cdot h(\nu(\mathbf{A},\mathbf{B})\log(n)) \cdot \sqrt{\log(d)}}{t^{1/8}}.$$

Remarks. Up to a factor involving $\log(n)$, the bound for SRHT matrices matches that for sub-Gaussian matrices. Meanwhile, from a more practical standpoint, our empirical results will show that the bootstrap's performance for SRHT matrices is generally similar to that for both sub-Gaussian and length-sampling matrices.

Further discussion of results. To comment on the role of $\nu(\mathbf{A},\mathbf{B})$ and $\|\mathbf{A}\|_F\|\mathbf{B}\|_F$ in Theorems 1 and 2, it is possible to interpret them as problem-specific "scale parameters". Indeed, it is natural that the bounds on d_{LP} should increase with the scale of $\mathbf{A}$ and $\mathbf{B}$ for the following reason. Namely, if $\mathbf{A}$ or $\mathbf{B}$ is multiplied by a scale factor $\kappa > 0$, then it can be checked that the quantile error $|\hat{q}_{1-\alpha}(t) - q_{1-\alpha}(t)|$ will also change by a factor of κ, and furthermore, the inequality (15) demonstrates a monotone relationship between the sizes of the quantile error and the d_{LP} error. For this reason, the bootstrap may still perform well in relation to the scale of the problem when the magnitudes of the parameters $\nu(\mathbf{A},\mathbf{B})$ and $\|\mathbf{A}\|_F\|\mathbf{B}\|_F$ are large. Alternatively, this idea can be seen by noting that the d_{LP} bounds can be made arbitrarily small by simply changing the units used to measure the entries of $\mathbf{A}$ and $\mathbf{B}$.

Beyond these considerations, it is still of interest to compare the results for different sketching matrices once a particular scaling has been fixed. For concreteness, consider a scaling where the spectral norms of $\mathbf{A}$ and $\mathbf{B}$ satisfy $\|\mathbf{A}\|_2 \asymp \|\mathbf{B}\|_2 \asymp 1$. (As an example, if we view $\mathbf{A}^\top\mathbf{A}$ as a sample covariance matrix, then the condition $\|\mathbf{A}\|_2 \asymp 1$ simply means that the largest principal component score is of order 1.) Under this scaling, it is simple to check that $\nu(\mathbf{A},\mathbf{B}) = \mathcal{O}(1)$, and $\|\mathbf{A}\|_F\|\mathbf{B}\|_F = \mathcal{O}(\sqrt{\mathsf{r}(\mathbf{A})\mathsf{r}(\mathbf{B})})$, where $\mathsf{r}(\mathbf{A}) := \|\mathbf{A}\|_F^2/\|\mathbf{A}\|_2^2$ is the "stable rank". In particular, note that if $\mathbf{A}$ and $\mathbf{B}$ are approximately low rank, as is common in applications, then $\mathsf{r}(\mathbf{A}) \ll d$, and $\mathsf{r}(\mathbf{B}) \ll d'$. Accordingly, we may conclude that if the conditions of Theorems 1 and 2 hold, then bootstrap consistency occurs under the following limits

$$\sqrt{\log(d)}/t^{1/8} = o(1) \quad \text{in the sub-Gaussian case,} \tag{19}$$

$$(\mathsf{r}(\mathbf{A})\mathsf{r}(\mathbf{B}))^{3/8}\sqrt{\log(d)}/t^{1/8} = o(1) \quad \text{in the length-sampling case,} \tag{20}$$

$$\log(n)^{3/4}\sqrt{\log(d)}/t^{1/8} = o(1) \quad \text{in the SRHT case,} \tag{21}$$

where we have used the simplifying assumption that $d \asymp d'$.

5. Experiments

This section outlines a set of experiments for evaluating the performance of Algorithm 1 with the extrapolation speed-up described in Section 3.3. The experiments involved both synthetic and natural matrices, as described below.

Synthetic matrices. In order to generate the matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$ synthetically, we selected the factors of its singular value decomposition $\mathbf{A} = \mathbf{U}\mathsf{diag}(\boldsymbol{\sigma})\mathbf{V}^T$ in the following ways, fixing $n = 30,000$ and $d = 1,000$. In previous work, a number of other experiments in randomized matrix computations have been designed along these lines (Ma et al., 2014; Yang et al., 2016).

The factor $\mathbf{U} \in \mathbb{R}^{n\times d}$ was selected as the Q factor from the reduced QR factorization of a random matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$. The rows of $\mathbf{X}$ were sampled i.i.d. from a multivariate t-distribution, $\boldsymbol{t}_2(\boldsymbol{\mu}, \mathbf{C})$, with 2 degrees of freedom, mean $\boldsymbol{\mu} = \mathbf{0}$, and covariance $c_{ij} = 2 \times 0.5^{|i-j|}$ where $\mathbf{C} = [c_{ij}]$. (This choice causes the matrix $\mathbf{A}$ to have high row-coherence, which is of interest, since this is a challenging case for sampling-based sketching matrices.) Next, the factor $\mathbf{V} \in \mathbb{R}^{d\times d}$ was selected as the Q factor from a QR factorization of a $d \times d$ matrix with i.i.d. $\mathcal{N}(0, 1)$ entries. For the singular values $\boldsymbol{\sigma} \in \mathbb{R}^d_+$, we chose two options, leading to either a low or high stable rank $\mathsf{r}(\mathbf{A}) = \frac{\|\mathbf{A}\|_F^2}{\|\mathbf{A}\|_2^2}$. In the low stable rank case, we put $\sigma_i = 10^{\kappa_i}$ for a set of equally spaced values κ_i between 0 and -6, yielding $\mathsf{r}(\mathbf{A}) = 36.7$. Alternatively, in the high stable rank case, the entries of $\boldsymbol{\sigma}$ were equally spaced between 0.1 and 1, yielding $\mathsf{r}(\mathbf{A}) = 370.1$. Finally, to make all numerical comparisons on a common scale, we normalized $\mathbf{A}$ so that $\|\mathbf{A}^T\mathbf{A}\|_\infty = 1$.

Natural matrices. We also conducted experiments on five natural data matrices $\mathbf{A}$ from the LIBSVM repository Chang and Lin (2011), named 'Connect', 'DNA', 'MNIST', 'Mushrooms', and 'Protein', with the same normalization that was used for the synthetic matrices. These datasets are briefly summarized in Table 1.

Table 1: A summary of the natural datasets.

Dataset	Connect	DNA	MNIST	Mushrooms	Protein
n	67,557	2,000	60,000	8,124	17,766
d	126	180	780	112	356

5.1. Design of experiments

For each matrix $\mathbf{A}$, natural or synthetic, we considered the task of estimating the quantile $q_{0.99}(t)$ for the random sketching error $\varepsilon_t = \left\|\mathbf{A}^T\mathbf{A} - \mathbf{A}^T\mathbf{S}^T\mathbf{S}\mathbf{A}\right\|_\infty$. The sketching matrix $\mathbf{S} \in \mathbb{R}^{t\times n}$ was allowed to be one of three types: Gaussian projection, length-sampling, and SRHT, as described in Section 2.

Ground truth values. The ground truth values for $q_{0.99}(t)$ were constructed in the following way. For each matrix $\mathbf{A}$, a grid of t values was specified, ranging from $d/2$ up to a larger number as high as $10d$ or $20d$, depending on $\mathbf{A}$. Next, for each t value, and for each type of sketching matrix, we used 1,000 realizations of $\mathbf{S} \in \mathbb{R}^{t\times n}$, yielding 1,000 realizations

of the random variable ε_t. In turn, the 0.99 sample quantile of the 1,000 realizations of ε_t was treated as the true value of $q_{0.99}(t)$, and this appears as the black curve in all plots.

Extrapolated estimates. With regard to the bootstrap extrapolation method in Section 3.3, we fixed the value $t_0 = d/2$ as the initial sketch size to extrapolate from. For each $\mathbf{A}$, and each type of sketching matrix, we applied Algorithm 1 to each of the 1,000 realizations of $\tilde{\mathbf{A}} = \mathbf{SA} \in \mathbb{R}^{t_0 \times d}$ generated previously. Each time Algorithm 1 was run, we used the modest choice of $B = 20$ for the number of bootstrap samples. From each set of 20 bootstrap samples, we used the 0.99 sample quantile as the estimate $\hat{q}_{0.99}(t_0)$.[§] Hence, there were 1,000 realizations of $\hat{q}_{0.99}(t_0)$ altogether. Next, we used the scaling rule in equation (12) to obtain 1,000 realizations of the extrapolated estimate $\hat{q}^{\text{ext}}_{0.99}(t)$ for values $t \geq t_0$.

In order to illustrate the variability of the estimate $\hat{q}^{\text{ext}}_{0.99}(t)$ over the 1,000 realizations, we plot three different curves as a function of t. The blue curve represents the average value of $\hat{q}^{\text{ext}}_{0.99}(t)$, while the green and yellow curves respectively correspond to the estimates ranking 100th an 900th out of the 1,000 realizations.

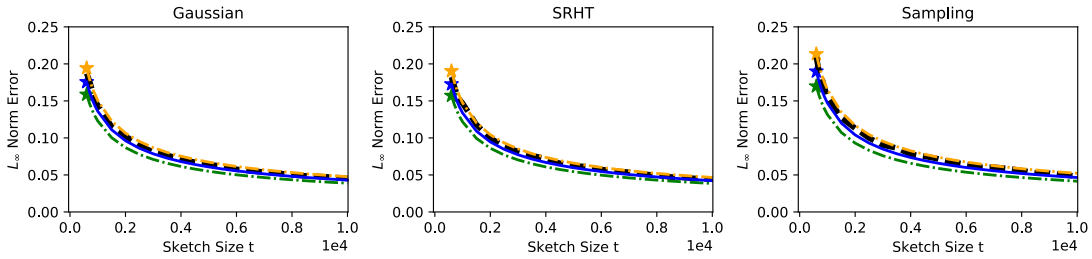
5.2. Comments on numerical results

Overall, the numerical results for the bootstrap extrapolation method are quite encouraging, and to a large extent, the method is accurate across many choices of $\mathbf{A}$ and $\mathbf{S}$. Given that the blue curves representing $\mathbb{E}[\hat{q}^{\text{ext}}_{0.99}(t)]$ are closely aligned with the black curves for $q_{0.99}(t)$, we see that the extrapolated estimate is essentially unbiased. Moreover, the variance of the estimate is fairly low, as indicated by the small gap between the green and yellow curves. The low variance is also notable when considered in light of the fact that only $B = 20$ bootstrap samples are used to construct $\hat{q}^{\text{ext}}_{0.99}(t)$, since the variance should decrease as B becomes larger.

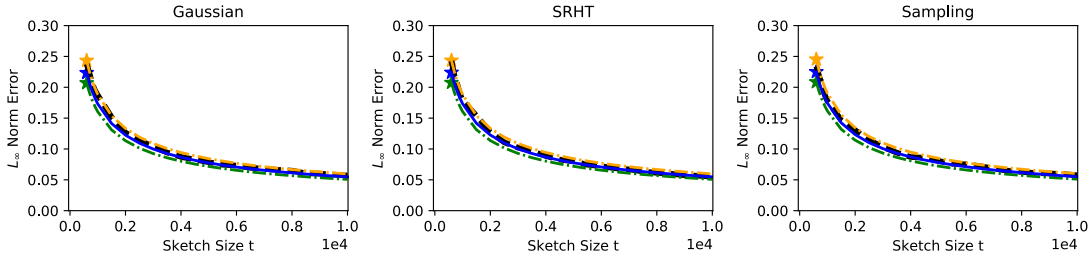
With attention to the extrapolation rule (12), there are two main points to note. First, the plots show that the extrapolation may be initiated at fairly low values of t_0, which are much less than the sketch sizes needed to achieve a small sketching error ε_t. Second, we see that $\hat{q}^{\text{ext}}_{0.99}(t)$ remains accurate for t much larger than t_0, well up to $t = 10,000$ and perhaps even farther. Consequently, the results show that the extrapolation technique is capable of saving quite a bit of computation without much detriment to statistical performance.

To consider the relationship between theory and practice, one basic observation is that all three types of sketching matrices obey roughly similar bounds in Theorems 1 and 2, and indeed, we also see generally similar numerical performance among the three types. At a more fine-grained level however, the Gaussian and SRHT sketching matrices tend to produce estimates $\hat{q}^{\text{ext}}_{0.99}(t)$ with somewhat higher variance than in the case of length sampling. Another difference between theory and simulation, is that the actual performance of the method seems to be better than what the theory suggests — since the estimates are accurate at values of t_0 that are much smaller than what would be expected from the rates in Theorems 1 and 2.

[§]Note that since $19/20 = 0.95$ and $20/20 = 1$, the 0.99 quantile was obtained by an interpolation rule.



(a) Low stable rank data.



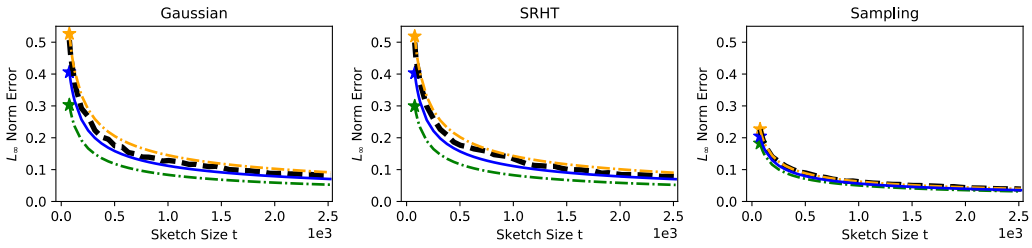
(b) High stable rank data.

Figure 2: **Results for synthetic matrices.** The black line represents $q_{0.99}(t)$ as a function of t. The blue star is the average bootstrap estimate at the initial sketch size $t_0 = d/2 = 500$, and the blue line represents the average extrapolated estimate $\mathbb{E}[\hat{q}^{\text{ext}}_{0.99}(t)]$ derived from the starting value t_0. To display the variability of the estimates, the green and yellow curves correspond to the 100th and 900th largest among the 1,000 realizations of $\hat{q}^{\text{ext}}_{0.99}(t)$ at each t.

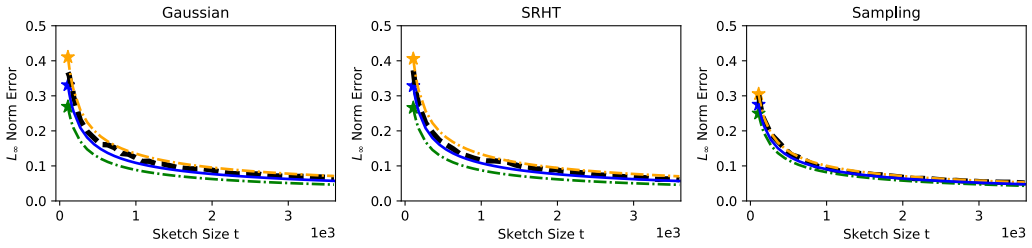
6. Conclusions and extensions

In this paper, we have focused on estimating the quantile $q_{1-\alpha}(t)$ as a way of addressing two fundamental issues in randomized matrix multiplication: (1) knowing how accurate a given sketched product is, and (2) knowing how much computation is needed to achieve a specified degree of accuracy. With regard to methodology, our approach is relatively novel in that it uses the statistical technique of bootstrapping to serve a computational purpose — by quantifying the error of a randomized sketching algorithm. A second important component of our method is the extrapolation technique, which ensures that the cost of estimating $q_{1-\alpha}(t)$ does not substantially increase the overall cost of standard sketching methods. Furthermore, our numerical results show that the extrapolated estimate is quite accurate in a variety of different situations, suggesting that our method may offer a general way to enhance sketching algorithms in practice.

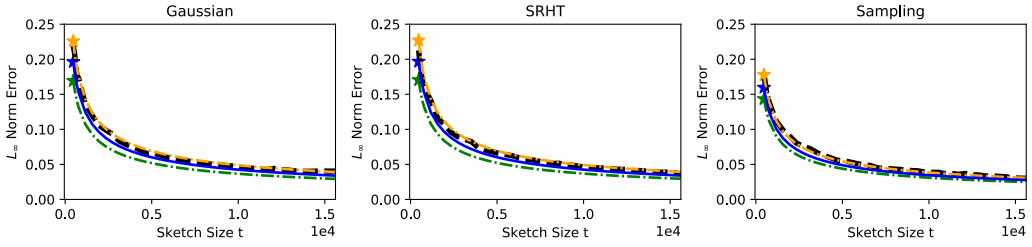
Extensions. More generally, the problems we have addressed for randomized matrix multiplication arise for many other large-scale matrix computations. Hence, it is natural to consider extensions of our approach to more complex settings, and in the remainder of this section, we briefly mention a few possibilities for future study.



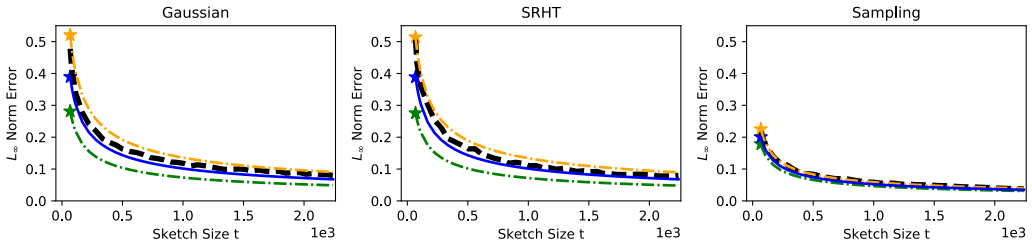
(a) Connect ($n = 67,557$ and $d = 126$).



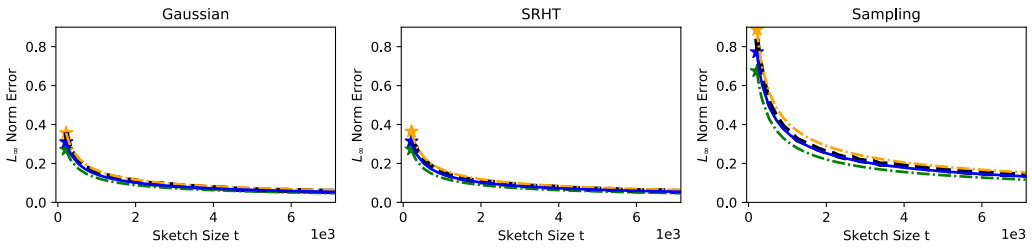
(b) DNA ($n = 2,000$ and $d = 180$).



(c) MNIST ($n = 60,000$ and $d = 780$).



(d) Mushrooms ($n = 8,124$ and $d = 112$).



(e) Protein ($n = 17,766$ and $d = 356$).

Figure 3: **Results for natural matrices.** The results for the natural matrices are plotted in the same way as described in the caption for the results on the synthetic matrices.

At a high level, each of the applications below deals with an object, say Θ, that is difficult to compute, as well as a randomized approximation, say $\widetilde{\Theta}$, that is built from a sketching matrix $\mathbf{S}$ with t rows. Next, if we consider the random error variable

$$\varepsilon_t = \|\widetilde{\Theta} - \Theta\|,$$

for an unspecified norm $\|\cdot\|$, then the problem of estimating the relationship between accuracy and computation can again be viewed as the problem of estimating the quantile function $q_{1-\alpha}(t)$ associated with ε_t. In turn, this leads to the question of how to develop a new bootstrap procedure that can generate approximate samples of ε_t, yielding an estimate $\hat{q}_{1-\alpha}(t)$. However, instead of starting from the multiplier bootstrap (Algorithm 1) as before, it may be conceptually easier to extend the non-parametric bootstrap (Algorithm 2) — because the latter bootstrap can viewed as a "plug-in" procedure that replaces $\mathbf{A}^T\mathbf{B}$ with $\tilde{\mathbf{A}}^T\tilde{\mathbf{B}}$, and replaces $\tilde{\mathbf{A}}^T\tilde{\mathbf{B}}$ with $(\tilde{\mathbf{A}}^*)^T(\tilde{\mathbf{B}}^*)$.

- *Linear regression.* Consider a multi-response linear regression problem, where the rows of $\mathbf{B} \in \mathbb{R}^{n\times d'}$ are response vectors, and the rows of $\mathbf{A} \in \mathbb{R}^{n\times d}$ are input observations. The optimal solution to ℓ_2-regression is given by

$$\mathbf{W}_{\text{opt}} \;=\; \underset{\mathbf{W}\in\mathbb{R}^{d\times d'}}{\text{argmin}} \; \big\|\mathbf{AW} - \mathbf{B}\big\|_F^2 \;=\; (\mathbf{A}^T\mathbf{A})^\dagger\mathbf{A}^T\mathbf{B},$$

  which has $\mathcal{O}(nd^2 + ndd')$ cost. In the case where $\max\{d, d'\} \ll n$, the matrix multiplications are a computational bottleneck, and an approximate solution can be obtained via

$$\widetilde{\mathbf{W}}_{\text{opt}} = (\tilde{\mathbf{A}}^T\tilde{\mathbf{A}})^\dagger(\tilde{\mathbf{A}}^T\tilde{\mathbf{B}}),$$

  which has a cost $\mathcal{O}(td^2 + tdd') + C_{\text{sketch}}$, where C_{sketch} is cost of matrix sketching (Drineas et al., 2006b, 2011, 2012; Clarkson and Woodruff, 2013). In order to estimate the quantile function associated with the error variable $\varepsilon_t = \|\widetilde{\mathbf{W}}_{\text{opt}} - \mathbf{W}_{\text{opt}}\|$, we could consider generating bootstrap samples of the form $\varepsilon_t^* = \|\widetilde{\mathbf{W}}^*_{\text{opt}} - \widetilde{\mathbf{W}}_{\text{opt}}\|$, where $\widetilde{\mathbf{W}}^*_{\text{opt}} = ((\tilde{\mathbf{A}}^*)^T(\tilde{\mathbf{A}}^*))^\dagger(\tilde{\mathbf{A}}^*)^T(\tilde{\mathbf{B}}^*)$. For recent results in the case where $\mathbf{W}$ is a vector, we refer to the paper (Lopes et al., 2018b).

- *Functions of covariance matrices.* If the rows of the matrix $\mathbf{A}$ are viewed as a sample of observations, then inferences on the population covariance structure are often based on functions of the form $\psi(\mathbf{A}^T\mathbf{A})$. For instance, the function $\psi(\mathbf{A}^T\mathbf{A})$ could be the top eigenvector, a set of eigenvalues, the condition number, or a test statistic. In any of these cases, if $\psi(\tilde{\mathbf{A}}^T\tilde{\mathbf{A}})$ is used as a fast approximation (Dasarathy et al., 2015), then the sketching error $\varepsilon_t = \|\psi(\tilde{\mathbf{A}}^T\tilde{\mathbf{A}}) - \psi(\mathbf{A}^T\mathbf{A})\|$ might be bootstrapped using $\varepsilon_t^* = \|\psi((\tilde{\mathbf{A}}^*)^T(\tilde{\mathbf{A}}^*)) - \psi(\tilde{\mathbf{A}}^T\tilde{\mathbf{A}})\|$.

- *Approximate Newton methods.* In large-scale applications, Newton's method is often impractical, since it involves the costly processing of a Hessian matrix. As an example, consider an optimization problem arising in binary classification, where the rows of $\mathbf{X} \in \mathbb{R}^{n\times d}$ are observations $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$, and $y_1, \ldots, y_n \in \{0, 1\}$ are labels. If an ℓ_2-regularized logistic classifier is used, this leads to minimizing the objective

function $f(\mathbf{w}) = \sum_{i=1}^{n} \log\left(1 + e^{-y_i \mathbf{w}^T \mathbf{x}_i}\right) + \frac{\gamma}{2}\|\mathbf{w}\|_2^2$ over coefficient vectors $\mathbf{w}$ in $\mathbb{R}^d$. The associated Newton step, with step size κ, is

$$\mathbf{w} \leftarrow \mathbf{w} - \kappa\, \mathbf{H}^{-1} \nabla f,$$

involving the Hessian

$$\mathbf{H} = \mathbf{A}^T\mathbf{A} + \gamma \mathbf{I}_d, \quad \text{where} \quad \mathbf{A} = \mathsf{diag}\big(1 + e^{y_1 \mathbf{w}^T \mathbf{x}_1}, \ldots, 1 + e^{y_n \mathbf{w}^T \mathbf{x}_n}\big)^{-1} \mathbf{X}.$$

If $d \ll n$, the cost of Newton's method is dominated by the formation of $\mathbf{H}$ at each iteration, and the Hessian matrix can be approximated by the sketched version $\tilde{\mathbf{H}} = \tilde{\mathbf{A}}^T\tilde{\mathbf{A}} + \gamma \mathbf{I}_d$, which reduces the per-iteration cost from $\mathcal{O}(nd^2)$ to $\mathcal{O}(td^2 + nd) + C_{\text{sketch}}$ (Pilanci and Wainwright, 2017; Roosta-Khorasani and Mahoney, 2016; Xu et al., 2016). In this context, the quality of the approximate Newton step could be assessed in terms of the error

$$\varepsilon_t = \|\tilde{\mathbf{H}}^{-1}\nabla f \;-\; \mathbf{H}^{-1}\nabla f\|,$$

and in turn, this might be bootstrapped using $\varepsilon_t^* = \|(\tilde{\mathbf{H}}^*)^{-1}\nabla f \;-\; \tilde{\mathbf{H}}^{-1}\nabla f\|$, where $\tilde{\mathbf{H}}^* = (\tilde{\mathbf{A}}^*)^T(\tilde{\mathbf{A}}^*) + \gamma \mathbf{I}_d$.

Acknowledgments

We thank the anonymous reviewers for their helpful suggestions. MEL thanks the National Science Foundation for partial support under grant DMS-1613218. MWM would like to thank the National Science Foundation, the Army Research Office, and the Defense Advanced Research Projects Agency for providing partial support of this work.

Appendices

Outline of appendices. Appendix A explains the main conceptual ideas underlying the proofs of Theorems 1 and 2. In particular, the proofs of these theorems will be decomposed into two main results: Propositions 3 and 4, which are given in Appendix A.

Appendix B will prove the sub-Gaussian case of Proposition 3, and Appendix C will prove the sub-Gaussian case of Proposition 4. Later on, Appendices D and E, will explain how the arguments can be changed to handle the length-sampling and SRHT cases.

Conventions used in proofs. If either of the matrices $\mathbf{A}$ or $\mathbf{B}$ are $\mathbf{0}$, then ε_t has a trivial point-mass distribution at 0. In this degenerate case, it is simple to check that the bootstrap produces an exact approximation. So, without loss of generality, all proofs are written under the assumption that $\mathbf{A}$ and $\mathbf{B}$ are non-zero. Next, since Assumption 1 is formulated using the $\lesssim$ notation, there is no loss of generality in carrying out calculations under the assumption that all the numbers t, n, d, d' are at least 8, which will ensure that quantities such as $\log(d)$ are greater than 2. Lastly, if a numbered lemma is invoked in the middle of a proof, the lemma may be found in Appendix F.

Appendix A. Gaussian and bootstrap approximations

Section A.1 introduces some notation that helps us to analyze the rescaled sketching error Z_t from the viewpoint of empirical processes. Next, in Section A.2, Theorem 1 will be decomposed into two propositions that compare Z_t and $Z_t^\star$ with the maximum of a suitable Gaussian process. The proofs of these propositions may be found in Appendices B and C.

A.1. Making a link between empirical processes and sketching error

The main idea of our analysis is to view Z_t as the maximum of an empirical process, which we now define. Recall the notation

$$\mathbf{D}_i = \mathbf{A}^T \mathbf{s}_i \mathbf{s}_i^T \mathbf{B} \quad \text{and} \quad \mathbf{M}_i \;=\; \mathbf{A}^T(\mathbf{s}_i\mathbf{s}_i^T - \mathbf{I}_n)\mathbf{B}.$$

Let $\mathbb{G}_t(\cdot)$ be the empirical process that acts on *linear* functions $f : \mathbb{R}^{d\times d'} \to \mathbb{R}$, according to

$$\mathbb{G}_t(f) \;:=\; \frac{1}{\sqrt{t}} \sum_{i=1}^{t} \Big(f(\mathbf{D}_i) - f(\mathbf{A}^T\mathbf{B}) \Big) \;=\; \frac{1}{\sqrt{t}} \sum_{i=1}^{t} f(\mathbf{M}_i).$$

For future reference, we also define the corresponding bootstrap process

$$\mathbb{G}_t^\star(f) \;:=\; \frac{1}{\sqrt{t}} \sum_{i=1}^{t} \xi_i \cdot \Big(f(\mathbf{D}_i) - f\big(\mathbf{A}^T\mathbf{S}^T\mathbf{S}\mathbf{B}\big) \Big),$$

where $\xi_1, \dots, \xi_t$ are i.i.d. $\mathcal{N}(0,1)$ and independent of $\mathbf{S}$.

Next, we define a certain collection $\mathscr{F}$ of linear functions from $\mathbb{R}^{d\times d'}$ to $\mathbb{R}$. Let $j_1 \in [d]$, $j_2 \in [d']$, $s \in \{-1, 1\}$, and $\mathbf{j} := (j_1, j_2, s)$. Then, for any matrix $\mathbf{W} \in \mathbb{R}^{d\times d'}$, we put

$$f_{\mathbf{j}}(\mathbf{W}) \;:=\; s \cdot \mathrm{tr}\big(\mathbf{C}_{\mathbf{j}}^T \mathbf{W}\big),$$

where $\mathbf{C}_{\mathbf{j}} := s\, \mathbf{e}_{j_1} \mathbf{e}_{j_2}^T \in \mathbb{R}^{d\times d'}$ and $\mathbf{e}_{j_1} \in \mathbb{R}^d$, $\mathbf{e}_{j_2} \in \mathbb{R}^{d'}$ are standard basis vectors. In words, the function $f_{\mathbf{j}}$ merely picks out the (j_1, j_2) entry of $\mathbf{W}$, and multiplies by a sign s. Likewise, let $\mathscr{J}$ be the collection of all the triples $\mathbf{j}$, and define the class of linear functions

$$\mathscr{F} \;:=\; \big\{ f_{\mathbf{j}} \mid \mathbf{j} \in \mathscr{J} \big\}.$$

Clearly, $\mathsf{card}(\mathscr{F}) = 2dd'$. Under this definition, it is simple to check that Z_t and $Z_t^\star$, defined in equations (16) and (17), can be expressed as

$$Z_t \;=\; \max_{f_{\mathbf{j}} \in \mathscr{F}} \mathbb{G}_t(f_{\mathbf{j}}), \qquad \text{and} \qquad Z_t^\star \;=\; \max_{f_{\mathbf{j}} \in \mathscr{F}} \mathbb{G}_t^\star(f_{\mathbf{j}}).$$

A.2. Statements of the approximation results

Theorems 1 and 2 are obtained by combining the following two results (Propositions 3 and 4) via the triangle inequality. In essence, these results are based on a comparison with the maximum of a certain Gaussian process. More specifically, let $\mathbb{G} : \mathscr{F} \to \mathbb{R}$ be a zero-mean Gaussian process whose covariance structure is defined according to

$$\begin{aligned} \mathbb{E}\big[\mathbb{G}(f_{\mathbf{j}})\, \mathbb{G}(f_{\mathbf{k}})\big] \;&=\; \mathrm{cov}\big(f_{\mathbf{j}}(\mathbf{D}_1),\, f_{\mathbf{k}}(\mathbf{D}_1)\big) \\ &=\; \mathbb{E}\Big[f_{\mathbf{j}}(\mathbf{D}_1)\, f_{\mathbf{k}}(\mathbf{D}_1)\Big] - f_{\mathbf{j}}(\mathbf{A}^T\mathbf{B})\, f_{\mathbf{k}}(\mathbf{A}^T\mathbf{B}), \end{aligned} \tag{22}$$

for all $\mathbf{j}, \mathbf{k} \in \mathscr{J}$. In turn, define the following random variable as the the maximum of this Gaussian process,

$$Z \;:=\; \max_{f_{\mathbf{j}} \in \mathscr{F}} \mathbb{G}(f_{\mathbf{j}}).$$

In order to handle the case of SRHT matrices, define another zero-mean Gaussian process $\tilde{\mathbb{G}} : \mathscr{F} \to \mathbb{R}$ (conditionally on a fixed realization of $\mathbf{D}_n^\circ$) to have its covariance structure given by

$$\begin{aligned}\mathbb{E}\big[\tilde{\mathbb{G}}(f_{\mathbf{j}})\,\tilde{\mathbb{G}}(f_{\mathbf{k}})\big|\mathbf{D}_n^\circ\big] \;&=\; \operatorname{cov}\big(f_{\mathbf{j}}(\mathbf{D}_1),\, f_{\mathbf{k}}(\mathbf{D}_1)\big|\mathbf{D}_n^\circ\big) \\ &=\; \mathbb{E}\Big[f_{\mathbf{j}}(\mathbf{D}_1)\, f_{\mathbf{k}}(\mathbf{D}_1)\Big|\mathbf{D}_n^\circ\Big] - f_{\mathbf{j}}(\mathbf{A}^T\mathbf{B})\, f_{\mathbf{k}}(\mathbf{A}^T\mathbf{B}), \qquad (23)\end{aligned}$$

and let $\tilde{Z}$ denote the maximum of the process $\tilde{\mathbb{G}}$,

$$\tilde{Z} \;:=\; \max_{f_{\mathbf{j}} \in \mathscr{F}} \tilde{\mathbb{G}}(f_{\mathbf{j}}).$$

We are now in position to state the approximation results.

Proposition 3 (Gaussian approximation) *Under Assumption 1 (a), the following bound holds,*

$$d_{\mathrm{LP}}\big(\mathcal{L}(Z_t)\,,\,\mathcal{L}(Z)\big) \;\leq\; \frac{c\cdot\nu(\mathbf{A},\mathbf{B})^{3/4}\cdot\sqrt{\log(d)}}{t^{1/8}}.$$

Under Assumption 1 (b), the following bound holds,

$$d_{\mathrm{LP}}\big(\mathcal{L}(Z_t)\,,\,\mathcal{L}(Z)\big) \;\leq\; \frac{c\cdot(\|\mathbf{A}\|_F\|\mathbf{B}\|_F)^{3/4}\cdot\sqrt{\log(d)}}{t^{1/8}}.$$

Under Assumption 1 (c), the following bound holds with probability at least $1 - c/n$

$$d_{\mathrm{LP}}\big(\mathcal{L}(Z_t)\,,\,\mathcal{L}(\tilde{Z}|\mathbf{D}_n^\circ)\big) \;\leq\; \frac{c\cdot\nu(\mathbf{A},\mathbf{B})^{3/4}\cdot(\log(n))^{3/4}\cdot\sqrt{\log(d)}}{t^{1/8}}.$$

Proposition 4 (Bootstrap approximation) *If Assumption 1 (a) holds, then the following bound holds with probability at least $1 - \frac{1}{t} - \frac{1}{dd'}$,*

$$d_{\mathrm{LP}}\big(\mathcal{L}(Z)\,,\,\mathcal{L}(Z_t^\star|\mathbf{S})\big) \;\leq\; \frac{c\cdot\nu(\mathbf{A},\mathbf{B})^{1/2}\sqrt{\log(d)}}{t^{1/8}}.$$

If Assumption 1 (b) holds, then the following bound holds with probability at least $1-\frac{1}{t}-\frac{1}{dd'}$,

$$d_{\mathrm{LP}}\big(\mathcal{L}(Z)\,,\,\mathcal{L}(Z_t^\star|\mathbf{S})\big) \;\leq\; \frac{c\cdot(\|\mathbf{A}\|_F\|\mathbf{B}\|_F)^{1/2}\sqrt{\log(d)}}{t^{1/8}}.$$

If Assumption 1 (c) holds, then the following bound holds with probability at least $1 - \frac{1}{t} - \frac{1}{dd'} - \frac{c}{n}$,

$$d_{\mathrm{LP}}\big(\mathcal{L}(\tilde{Z}|\mathbf{D}_n^\circ)\,,\,\mathcal{L}(Z_t^\star|\mathbf{S})\big) \;\leq\; \frac{c\cdot\nu(\mathbf{A},\mathbf{B})^{1/2}\cdot\log(n)^{1/2}\cdot\sqrt{\log(d)}}{t^{1/8}}.$$

Appendix B. Proof of Proposition 3, part (a)

Let $A \subset \mathbb{R}$ be a Borel set. Due to Theorem 3.1 from the paper Chernozhukov et al. (2016), we have for any $\delta > 0$,

$$\mathbb{P}(Z_t \in A) \leq \mathbb{P}(Z \in A^{c\delta}) + \tfrac{c \log^2(d)}{\delta^3 \sqrt{t}} \Big(L_t + K_t(\delta) + J_t(\delta)) \Big), \tag{24}$$

where we define the following non-random quantities

$$L_t \;:=\; \max_{f_{\mathbf{j}} \in \mathscr{F}} \frac{1}{t} \sum_{i=1}^{t} \mathbb{E}\Big[\big| f_{\mathbf{j}}(\mathbf{M}_i) \big|^3 \Big], \tag{25}$$

$$K_t(\delta) \;:=\; \mathbb{E}\left[\max_{f_{\mathbf{j}} \in \mathscr{F}} |f_{\mathbf{j}}(\mathbf{M}_1)|^3 \cdot 1\Big\{ \max_{f_{\mathbf{j}} \in \mathscr{F}} |f_{\mathbf{j}}(\mathbf{M}_1)| > \tfrac{\delta \sqrt{t}}{\log(\mathrm{card}(\mathscr{F}))} \Big\} \right], \tag{26}$$

$$J_t(\delta) \;:=\; \mathbb{E}\left[\max_{f_{\mathbf{j}} \in \mathscr{F}} |\mathbb{G}(f_{\mathbf{j}})|^3 \cdot 1\Big\{ \max_{f_{\mathbf{j}} \in \mathscr{F}} |\mathbb{G}(f_{\mathbf{j}})| > \tfrac{\delta \sqrt{t}}{\log(\mathrm{card}(\mathscr{F}))} \Big\} \right]. \tag{27}$$

The remainder of the proof consists in bounding each of these quantities, and we will establish the following two bounds for all $\delta > 0$,

$$L_t \;\leq\; c\, \nu(\mathbf{A}, \mathbf{B})^3, \tag{28}$$

$$K_t(\delta) + J_t(\delta) \;\leq\; c \Big(\tfrac{\delta \sqrt{t}}{\log(d)} + \log(d)\, \nu(\mathbf{A}, \mathbf{B}) \Big)^3 \cdot \exp\Big(- \tfrac{\delta \sqrt{t}}{c\ \nu(\mathbf{A}, \mathbf{B}) \log^2(d)} \Big). \tag{29}$$

Recall also that $\mathrm{card}(\mathscr{F}) = 2dd'$, and $d \asymp d'$ under Assumption 1.

For the moment, we set aside the task of proving these bounds, and consider the choice of δ. There are two constraints that we would like δ to satisfy. First, we would like to choose δ so that the bounds on L_t and $(K_t(\delta) + J_t(\delta))$ are of the same order. In particular, we desire

$$K_t(\delta) + J_t(\delta) \;\leq\; c\, \nu(\mathbf{A}, \mathbf{B})^3. \tag{30}$$

Second, with regard to line (24) we would like δ to solve the equation

$$\delta \;=\; \tfrac{1}{\delta^3} \tfrac{\log^2(d)\, \nu(\mathbf{A}, \mathbf{B})^3}{\sqrt{t}}, \tag{31}$$

so that the second term in line (24) is of order δ. The idea is that if δ satisfies both of the conditions (30) and (31), then the definition of the d_{LP} metric and line (24) imply

$$d_{\mathrm{LP}}(\mathcal{L}(Z), \mathcal{L}(Z_t)) \;\leq\; c\, \delta.$$

To proceed, consider the choice

$$\delta_0 \;:=\; \tfrac{\log^{1/2}(d)\, \nu(\mathbf{A}, \mathbf{B})^{3/4}}{t^{1/8}},$$

which clearly satisfies line (31). Futhermore, it can be checked that δ_0 also satisfies the constraint (30) under Assumption 1 (a). (The details of verifying this are somewhat tedious and are given in Lemma 16 in Appendix F.)

To finish the proof, it remains to establish the bounds (28) and (29). To handle L_t, note that[¶]

$$\begin{aligned}
\mathbb{E}[|f_{\mathbf{j}}(\mathbf{M}_i)|^3] &= \|f_{\mathbf{j}}(\mathbf{M}_1)\|_3^3 \\
&\leq c\|f_{\mathbf{j}}(\mathbf{M}_1)\|_{\psi_1}^3, \qquad \text{(Lemma 9)} \qquad (32)\\
&\leq c\Big(\frac{\|\mathbf{B}\mathbf{C}_{\mathbf{j}}^T\mathbf{A}^T\|_F^2}{\|\mathbf{B}\mathbf{C}_{\mathbf{j}}^T\mathbf{A}^T\|_2}\Big)^3, \qquad \text{(Lemma 14)} \\
&= c\Big(\|\mathbf{B}\mathbf{C}_{\mathbf{j}}^T\mathbf{A}^T\|_F\Big)^3 \qquad \text{(since } \|\mathbf{H}\|_2 = \|\mathbf{H}\|_F \text{ when } \mathbf{H} \text{ is rank-1)} \\
&= c\Big(\mathrm{tr}\big(\mathbf{B}\mathbf{C}_{\mathbf{j}}^T\mathbf{A}^T\mathbf{A}\mathbf{C}_{\mathbf{j}}\mathbf{B}^T\big)\Big)^{3/2} \\
&= c\Big(\mathbf{e}_{j_1}^T\mathbf{A}^T\mathbf{A}\mathbf{e}_{j_1}\cdot\mathbf{e}_{j_2}^T\mathbf{B}^T\mathbf{B}\mathbf{e}_{j_2}\Big)^{3/2} \\
&\leq c\ \nu(\mathbf{A},\mathbf{B})^3, \qquad (33)
\end{aligned}$$

which proves the claimed bound in line (28).

Next, regarding $K_t(\delta)$, let us consider the random variable

$$\eta := \max_{f_{\mathbf{j}}\in\mathscr{F}} |f_{\mathbf{j}}(\mathbf{M}_1)|.$$

It follows from Lemma 9 (part 4) and Lemma 13 in Appendix F that $K_t(\delta)$ can be bounded in terms of the Orlicz norm $\|\eta\|_{\psi_1}$,

$$K_t(\delta) \leq c\Big(\tfrac{\delta\sqrt{t}}{\log(\mathsf{card}(\mathscr{F}))} + \|\eta\|_{\psi_1}\Big)^3 \cdot \exp(-\tfrac{\delta\sqrt{t}}{\|\eta\|_{\psi_1}\log(\mathsf{card}(\mathscr{F})}).$$

To handle $\|\eta\|_{\psi_1}$, it follows from Lemma 9 (part 3), that

$$\|\eta\|_{\psi_1} \leq c\log(\mathsf{card}(\mathscr{F}))\cdot\max_{f_{\mathbf{j}}\in\mathscr{F}}\|f_{\mathbf{j}}(\mathbf{M}_1)\|_{\psi_1}. \qquad (34)$$

Furthermore, due to the earlier calculation starting at line (32) above,

$$\|f_{\mathbf{j}}(\mathbf{M}_1)\|_{\psi_1} \leq c\ \nu(\mathbf{A},\mathbf{B}). \qquad (35)$$

Combining the last few steps, we conclude that

$$K_t(\delta) \leq c\Big(\tfrac{\delta\sqrt{t}}{\log(\mathrm{card}(\mathscr{F}))} + \log(\mathrm{card}(\mathscr{F}))\nu(\mathbf{A},\mathbf{B})\Big)^3 \cdot \exp\Big(-\tfrac{\delta\sqrt{t}}{c\nu(\mathbf{A},\mathbf{B})\log^2(\mathrm{card}(\mathscr{F}))}\Big). \qquad (36)$$

Lastly, we turn to bounding $J_t(\delta)$. Fortunately, much of the argument for bounding $K_t(\delta)$ can be carried over. Specifically, consider the random variable

$$\zeta := \max_{f_{\mathbf{j}}\in\mathscr{F}} |\mathbb{G}(f_{\mathbf{j}})|.$$

[¶]In this step, we use the assumption that $\|\sqrt{t}S_{i,j}\|_{\psi_2} \leq c$ for all i and j.

Lemma 13 in Appendix F shows that $J_t(\delta)$ can be bounded in terms of $\|\zeta\|_{\psi_1}$,

$$J_t(\delta) \leq c\Big(\tfrac{\delta\sqrt{t}}{\log(\mathrm{card}(\mathscr{F}))} + \|\zeta\|_{\psi_1}\Big)^3 \cdot \exp\Big(-\tfrac{\delta\sqrt{t}}{\|\zeta\|_{\psi_1}\,\log(\mathrm{card}(\mathscr{F}))}\Big).$$

Proceeding in a way that is similar to the bound for $K_t(\delta)$, it follows from part (3) of Lemma 9 that

$$\|\zeta\|_{\psi_1} \leq c\log(\mathsf{card}(\mathscr{F})) \cdot \max_{f_{\mathbf{j}}\in\mathscr{F}} \|\mathbb{G}(f_{\mathbf{j}})\|_{\psi_1}.$$

Furthermore, for every $f_{\mathbf{j}} \in \mathscr{F}$, the facts in Lemma 9 imply

$$\begin{aligned}
\|\mathbb{G}(f_{\mathbf{j}})\|_{\psi_1} &\leq c\,\|\mathbb{G}(f_{\mathbf{j}})\|_{\psi_2} \\
&\leq c\sqrt{\mathsf{var}(\mathbb{G}(f_{\mathbf{j}}))} \\
&= c\sqrt{\mathsf{var}(f_{\mathbf{j}}(\mathbf{D}_1))} \quad \text{(by definition of } \mathbb{G}) \\
&\leq c\|f_{\mathbf{j}}(\mathbf{D}_1)\|_2 \\
&\leq c\|f_{\mathbf{j}}(\mathbf{D}_1)\|_{\psi_1} \qquad (37) \\
&\leq c\,\|f_{\mathbf{j}}(\mathbf{D}_1) - \mathbb{E}[f_{\mathbf{j}}(\mathbf{D}_1)\|_{\psi_1} + c\,|\mathbb{E}[f_{\mathbf{j}}(\mathbf{D}_1)]| \\
&= c\|f_{\mathbf{j}}(\mathbf{M}_1)\|_{\psi_1} + c\,|\mathrm{tr}(\mathbf{B}\mathbf{C}_{\mathbf{j}}^T\mathbf{A})| \\
&\leq c\,\nu(\mathbf{A},\mathbf{B}), \qquad (38)
\end{aligned}$$

where the last step follows from the bounds (32) through (33), and the fact that $|\mathrm{tr}(\mathbf{B}\mathbf{C}_{\mathbf{j}}^T\mathbf{A})| \leq \nu(\mathbf{A},\mathbf{B})$. Consequently, up to a constant factor, $J_t(\delta)$ satisfies the same bound as $K_t(\delta)$ given in line (36), and this proves the claim in line (29).

■

Appendix C. Proof of Proposition 4, part (a)

We will show there is a set of "good" sketching matrices $\mathscr{S} \subset \mathbb{R}^{t\times n}$ with the following two properties. First, a randomly drawn sketching matrix $\mathbf{S}$ is likely to fall in $\mathscr{S}$. Namely,

$$\mathbb{P}(\mathbf{S} \in \mathscr{S}) \geq 1 - \tfrac{1}{t}. \qquad (39)$$

Second, whenever the event $\{\mathbf{S} \in \mathscr{S}\}$ occurs, we have the following bound for any $\delta > 0$ and any Borel set $A \subset \mathbb{R}$,

$$\mathbb{P}\Big(\max_{f_{\mathbf{j}}\in\mathscr{F}} \mathbb{G}_t^\star(f_{\mathbf{j}}) \in A \Big| \mathbf{S}\Big) \;\leq\; \mathbb{P}\Big(\max_{f_{\mathbf{j}}\in\mathscr{F}} \mathbb{G}(f_{\mathbf{j}}) \in A^\delta\Big) + \tfrac{c\,\nu(\mathbf{A},\mathbf{B})\cdot\log(\mathsf{card}(\mathscr{F}))}{\delta\, t^{1/4}}. \qquad (40)$$

If we set δ to the particular choice $\delta_0 := t^{-1/8}\sqrt{\nu(\mathbf{A},\mathbf{B})\cdot\log(\mathsf{card}(\mathscr{F}))}$, then δ_0 solves the equation

$$\delta_0 \;=\; \tfrac{\nu(\mathbf{A},\mathbf{B})\cdot\log(\mathsf{card}(\mathscr{F}))}{\delta_0\, t^{1/4}}.$$

Consequently, by the definition of the d_{LP} metric, this implies that whenever the event $\{\mathbf{S} \in \mathscr{S}\}$ occurs, we have

$$d_{\text{LP}}\big(\mathcal{L}(Z_t^\star|\mathbf{S})\,,\,\mathcal{L}(Z)\big) \;\leq\; c\,t^{-1/8}\,\sqrt{\nu(\mathbf{A},\mathbf{B})\cdot\log(\mathsf{card}(\mathscr{F}))}, \tag{41}$$

and this implies the statement of Proposition 4.

To proceed with the main argument of constructing $\mathscr{S}$ and demonstrating the two properties (39) and (40), it is helpful to think of $\mathbb{G}_t^\star$ (conditionally on $\mathbf{S}$) and $\mathbb{G}$ as Gaussian vectors of dimension $\mathsf{card}(\mathscr{F}) = 2dd'$. From this point of view, we can compare the maxima of these vectors using a result due to Chernozhukov et al. (2016, Theorem 3.2). Under our assumptions, this result implies that for any realization of $\mathbf{S}$, any number $\delta > 0$, and any Borel set $A \subset \mathbb{R}$, we have

$$\mathbb{P}\Big(\max_{f_{\mathbf{j}}\in\mathscr{F}} \mathbb{G}_t^\star(f_{\mathbf{j}}) \in A\Big|\mathbf{S}\Big) \;\;\leq\;\; \mathbb{P}\Big(\max_{f_{\mathbf{j}}\in\mathscr{F}} \mathbb{G}(f_{\mathbf{j}}) \in A^\delta\Big) + \tfrac{c\sqrt{\Delta_t(\mathbf{S})\log(\mathsf{card}(\mathscr{F}))}}{\delta},$$

where we define the following function of $\mathbf{S}$,

$$\Delta_t(\mathbf{S}) \;\;:=\;\; \max_{(f_{\mathbf{j}},f_{\mathbf{k}})\in\mathscr{F}\times\mathscr{F}} \Big|\mathbb{E}\big[\mathbb{G}_t^\star(f_{\mathbf{j}})\mathbb{G}_t^\star(f_{\mathbf{k}})\big|\mathbf{S}\big] - \mathbb{E}\big[\mathbb{G}(f_{\mathbf{j}})\mathbb{G}(f_{\mathbf{k}})\big]\Big|. \tag{42}$$

When referencing Theorem 3.2 from the paper Chernozhukov et al. (2016), note that $\mathbb{E}[\mathbb{G}(f_{\mathbf{j}})] = 0$ and $\mathbb{E}[\mathbb{G}_t^\star(f_{\mathbf{j}})|\mathbf{S}] = 0$ for all $f_{\mathbf{j}} \in \mathscr{F}$. To interpret $\Delta_t(\mathbf{S})$, it may be viewed as the ℓ_∞-distance between the covariance matrices associated with $\mathbb{G}_t^\star$ (conditionally on $\mathbf{S}$) and $\mathbb{G}$.

Using the above notation, we define the set of sketching matrices $\mathscr{S} \subset \mathbb{R}^{n\times t}$ according to

$$\mathbf{S} \in \mathscr{S} \qquad \text{if and only if} \qquad \Delta_t(\mathbf{S}) \leq \tfrac{c}{\sqrt{t}}\cdot\nu(\mathbf{A},\mathbf{B})^2\cdot\log(\mathsf{card}(\mathscr{F})). \tag{43}$$

Based on this definition, it is simple to check that the proof is reduced to showing that the event $\{\mathbf{S} \in \mathscr{S}\}$ occurs with probability at least $1 - \frac{1}{t} - \frac{1}{dd'}$. This is guaranteed by the lemma below.

■

Lemma 5 *Suppose Assumption 1 (a) holds. Then, the event*

$$\Delta_t(\mathbf{S}) \;\leq\; \frac{c}{\sqrt{t}}\cdot\nu(\mathbf{A},\mathbf{B})^2\cdot\log\big(\mathsf{card}(\mathscr{F})\big)$$

occurs with probability at least $1 - \frac{1}{t} - \frac{1}{dd'}$.

Proof We begin by bounding $\Delta_t(\mathbf{S})$ with two other quantities (to be denoted $\Delta_t'(\mathbf{S})$, $\Delta_t''(\mathbf{S})$) that are easier to bound. Using the fact that $\frac{1}{t}\sum_{i=1}^t \mathbf{D}_i = \mathbf{A}^T\mathbf{S}^T\mathbf{S}\mathbf{B}$ it can be checked that

$$\mathbb{E}\big[\mathbb{G}_t^\star(f_{\mathbf{j}})\mathbb{G}_t^\star(f_{\mathbf{k}})\big|\mathbf{S}\big] \;\;=\;\; \Big(\tfrac{1}{t}\textstyle\sum_{i=1}^t f_{\mathbf{j}}(\mathbf{D}_i)f_{\mathbf{k}}(\mathbf{D}_i)\Big) - \Big(\tfrac{1}{t}\textstyle\sum_{i=1}^t f_{\mathbf{j}}(\mathbf{D}_i)\Big)\cdot\Big(\tfrac{1}{t}\textstyle\sum_{i=1}^t f_{\mathbf{k}}(\mathbf{D}_i)\Big).$$

Similarly, recall from line (22) that

$$\mathbb{E}[\mathbb{G}(f_{\mathbf{j}})\mathbb{G}(f_{\mathbf{k}})] \;=\; \mathbb{E}\Big[f_{\mathbf{j}}(\mathbf{D}_1)f_{\mathbf{k}}(\mathbf{D}_1)\Big] - \mathbb{E}[f_{\mathbf{j}}(\mathbf{D}_1)]\cdot\mathbb{E}[f_{\mathbf{k}}(\mathbf{D}_1)].$$

From looking at the last two lines, it is natural to define the following *zero-mean* random variables for any triple $i, \mathbf{j}, \mathbf{k}$,[‖]

$$Q_{i,\mathbf{j},\mathbf{k}} := f_{\mathbf{j}}(\mathbf{D}_i) f_{\mathbf{k}}(\mathbf{D}_i) - \mathbb{E}\Big[f_{\mathbf{j}}(\mathbf{D}_i) f_{\mathbf{k}}(\mathbf{D}_i)\Big],$$

and

$$R_{t,\mathbf{j}} := \tfrac{1}{t} \textstyle\sum_{i=1}^{t} \big(f_{\mathbf{j}}(\mathbf{D}_i) - \mathbb{E}[f_{\mathbf{j}}(\mathbf{D}_i)]\big).$$

Then, some algebra shows that

$$\begin{aligned}\mathbb{E}\big[\mathbb{G}_t^\star(f_{\mathbf{j}})\, \mathbb{G}_t^\star(f_{\mathbf{k}}) \,\big|\, \mathbf{S}\big] - \mathbb{E}\big[\mathbb{G}(f_{\mathbf{j}})\, \mathbb{G}(f_{\mathbf{k}})\big] = \ & \Big(\tfrac{1}{t} \textstyle\sum_{i=1}^{t} Q_{i,\mathbf{j},\mathbf{k}}\Big) - R_{t,\mathbf{j}} \cdot R_{t,\mathbf{k}} \\ & - \mathbb{E}\big[f_{\mathbf{j}}(\mathbf{D}_1)\big] \cdot R_{t,\mathbf{k}} - \mathbb{E}\big[f_{\mathbf{k}}(\mathbf{D}_1)\big] \cdot R_{t,\mathbf{j}}.\end{aligned}$$

So, if we define the quantities

$$\begin{aligned}\Delta_t'(\mathbf{S}) \ &:= \ \max_{(\mathbf{j},\mathbf{k}) \in \mathscr{J} \times \mathscr{J}} \left| \tfrac{1}{t} \sum_{i=1}^{t} Q_{i,\mathbf{j},\mathbf{k}} \right|, \\ \Delta_t''(\mathbf{S}) \ &:= \ \max_{\mathbf{j} \in \mathscr{J}} \big|R_{t,\mathbf{j}}\big|,\end{aligned}$$

then

$$\Delta_t(\mathbf{S}) \ \le \ \Delta_t'(\mathbf{S}) + \Delta_t''(\mathbf{S})^2 + 2\nu(\mathbf{A},\mathbf{B}) \cdot \Delta_t''(\mathbf{S}),$$

where we have made use of the simple bound $|\mathbb{E}[f_{\mathbf{j}}(\mathbf{D}_1)]| \le \|\mathbf{A}^T\mathbf{B}\|_\infty \le \nu(\mathbf{A},\mathbf{B})$. The following lemma establishes tail bounds for $\Delta_t'(\mathbf{S})$ and $\Delta_t''(\mathbf{S})$, which lead to the statement of Proposition 4. ■

Lemma 6 *Suppose Assumption 1 (a) holds. Then, the event*

$$\Delta_t'(\mathbf{S}) \ \le \ \frac{c}{\sqrt{t}} \cdot \nu(\mathbf{A},\mathbf{B})^2 \cdot \log\big(\mathsf{card}(\mathscr{F})\big) \tag{i}$$

occurs with probability at least $1 - \frac{1}{t}$*, and the event*

$$\Delta_t''(\mathbf{S}) \ \le \ \frac{c}{\sqrt{t}} \cdot \nu(\mathbf{A},\mathbf{B}) \cdot \sqrt{\log\big(\mathsf{card}(\mathscr{F})\big)} \tag{ii}$$

occurs with probability at least $1 - \frac{1}{dd'}$.

[‖] Note that $Q_{i,\mathbf{j},\mathbf{k}}$ is a multivariate polynomial of degree-4 in the variables $S_{i,j}$, and so techniques based on moment generating functions, like Chernoff bounds, are not generally applicable to controlling $Q_{i,\mathbf{j},\mathbf{k}}$. For instance, if $X \sim \mathcal{N}(0,1)$, then the variable X^4 does not have a moment generating function. Handling this obstacle is a notable aspect of our analysis.

Proof of Lemma 6 (i). Let $p > 2$. Due to part (3) of Lemma 9 in Appendix F, we have

$$\|\Delta_t'(\mathbf{S})\|_p \;\le\; \big(\mathsf{card}(\mathscr{F})^2\big)^{1/p} \cdot \max_{(\mathbf{j},\mathbf{k})\in\mathscr{J}\times\mathscr{J}} \Big\| \tfrac{1}{t}\textstyle\sum_{i=1}^t Q_{i,\mathbf{j},\mathbf{k}} \Big\|_p. \tag{44}$$

Note that each variable $Q_{i,\mathbf{j},\mathbf{k}}$ has moments of all orders, and when $\mathbf{j}$ and $\mathbf{k}$ are held fixed, the sequence $\{Q_{i,\mathbf{j},\mathbf{k}}\}_{1\le i\le t}$ is i.i.d. For this reason, it is natural to use Rosenthal's inequality to bound the L_p norm of the right side of the previous line. Specifically, the version of Rosenthal's inequality** stated in Lemma 10 in Appendix F leads to

$$\Big\| \tfrac{1}{t}\textstyle\sum_{i=1}^t Q_{i,\mathbf{j},\mathbf{k}} \Big\|_p \le c\cdot \tfrac{p/\log(p)}{t} \cdot \max\left\{ \Big\| \textstyle\sum_{i=1}^t Q_{i,\mathbf{j},\mathbf{k}} \Big\|_2 , \, \big(\textstyle\sum_{i=1}^t \|Q_{i,\mathbf{j},\mathbf{k}}\|_p^p\big)^{1/p} \right\}. \tag{45}$$

The L_2 norm on the right side of Rosenthal's inequality (45) satisfies the bound

$$\begin{aligned}
\Big\| \textstyle\sum_{i=1}^t Q_{i,\mathbf{j},\mathbf{k}} \Big\|_2 &= \sqrt{\mathsf{var}\big(\textstyle\sum_{i=1}^t Q_{i,\mathbf{j},\mathbf{k}}\big)} \\
&= \sqrt{t}\sqrt{\mathsf{var}(Q_{1,\mathbf{j},\mathbf{k}})} \\
&= \sqrt{t}\sqrt{\mathsf{var}\big(f_{\mathbf{j}}(\mathbf{D}_1)\, f_{\mathbf{k}}(\mathbf{D}_1)\big)} \\
&\le \sqrt{t}\Big\| f_{\mathbf{j}}(\mathbf{D}_1)\, f_{\mathbf{k}}(\mathbf{D}_1) \Big\|_2 \\
&\le \sqrt{t}\big\| f_{\mathbf{j}}(\mathbf{D}_1) \big\|_4 \cdot \big\| f_{\mathbf{k}}(\mathbf{D}_1) \big\|_4 \qquad \text{(Cauchy-Schwarz)} \\
&\le c\sqrt{t}\big\| f_{\mathbf{j}}(\mathbf{D}_1) \big\|_{\psi_1} \cdot \big\| f_{\mathbf{k}}(\mathbf{D}_1) \big\|_{\psi_1} \qquad \text{(Lemma 9)} \\
&\le c\,\sqrt{t}\nu(\mathbf{A},\mathbf{B})^2,
\end{aligned}$$

where the last step follows from the fact

$$\|f_{\mathbf{j}}(\mathbf{D}_1)\|_{\psi_1} \le c\,\nu(\mathbf{A},\mathbf{B}), \tag{46}$$

obtained in the bounds (32) through (33).

Next, to handle the L_p norms in the bound (45), observe that

$$\begin{aligned}
\|Q_{1,\mathbf{j},\mathbf{k}}\|_p &\le \big\| f_{\mathbf{j}}(\mathbf{D}_1) f_{\mathbf{k}}(\mathbf{D}_1) \big\|_p + \big| \mathbb{E}\big[f_{\mathbf{j}}(\mathbf{D}_1) f_{\mathbf{k}}(\mathbf{D}_1) \big] \big| \\
&\le 2\big\| f_{\mathbf{j}}(\mathbf{D}_1) \big\|_{2p} \cdot \big\| f_{\mathbf{k}}(\mathbf{D}_1) \big\|_{2p} \qquad \text{(Cauchy-Schwarz)} \\
&\le c\,p^2 \big\| f_{\mathbf{j}}(\mathbf{D}_1) \big\|_{\psi_1} \cdot \big\| f_{\mathbf{k}}(\mathbf{D}_1) \big\|_{\psi_1} \qquad \text{(Lemma 9 in Appendix F)} \\
&\le c\,p^2\,\nu(\mathbf{A},\mathbf{B})^2 \qquad \text{(inequality (46)).}
\end{aligned}$$

Hence, the second term in the Rosenthal bound (45) satisfies

$$\big(\textstyle\sum_{i=1}^t \|Q_{i,\mathbf{j},\mathbf{k}}\|_p^p\big)^{1/p} \;\le\; c\cdot p^2 \cdot t^{1/p} \cdot \nu(\mathbf{A},\mathbf{B})^2,$$

**Here we are using the version of Rosenthal's inequality with the optimal dependence on p. It is a notable aspect of our argument that it makes essential use of this scaling in p.

and as long as the first term in the Rosenthal bound dominates[††], i.e.

$$p^2\ t^{1/p} \lesssim\ t^{1/2} \tag{47}$$

then we conclude that for any $\mathbf{j}$ and $\mathbf{k}$,

$$\left\| \tfrac{1}{t} \textstyle\sum_{i=1}^{t} Q_{i,\mathbf{j},\mathbf{k}} \right\|_p \ \leq\ \tfrac{c\cdot(p/\log(p))\cdot\nu(\mathbf{A},\mathbf{B})^2}{\sqrt{t}}.$$

Since the previous bound does not depend on $\mathbf{j}$ or $\mathbf{k}$, combining it with the first step in line (44) leads to

$$\left\|\Delta_t'(\mathbf{S})\right\|_p \ \leq\ c\cdot(p/\log(p))\cdot \mathsf{card}(\mathscr{F})^{2/p}\cdot\tfrac{\nu(\mathbf{A},\mathbf{B})^2}{\sqrt{t}}.$$

Next, we convert this norm bound into a tail bound. Specifically, if we consider the value

$$x_p\ :=\ c\cdot(p/\log(p))\cdot \mathsf{card}(\mathscr{F})^{2/p}\cdot\tfrac{\nu(\mathbf{A},\mathbf{B})^2}{\sqrt{t}}\cdot t^{1/p}$$

then Markov's inequality gives

$$\mathbb{P}\big(\Delta_t'(\mathbf{S})\ \geq\ x_p\big)\ \leq\ \tfrac{\|\Delta_t'(\mathbf{S})\|_p^p}{x_p^p} \leq \tfrac{1}{t}.$$

Considering the choice of p given by

$$p = \log(\mathsf{card}(\mathscr{F})),$$

and noting that $\mathsf{card}(\mathscr{F})^{1/p} = e$, it follows that under this choice of p,

$$x_p\ \leq\ \left(\tfrac{c\cdot\nu(\mathbf{A},\mathbf{B})^2\cdot\log(\mathsf{card}(\mathscr{F}))}{\sqrt{t}}\right)\cdot\left(\tfrac{t^{1/p}}{\log(p)}\right).$$

Moreover, as long as $t \lesssim \mathsf{card}(\mathscr{F})^{\kappa}$ for some absolute constant $\kappa \geq 1$ (which holds under Assumption 1), then the last factor on the right satisfies

$$\big(\tfrac{t^{1/p}}{\log(p)}\big) \leq \tfrac{(\mathsf{card}(\mathscr{F})^{1/p})^{\kappa}}{\log(p)} = \tfrac{e^{\kappa}}{\log(\log(\mathsf{card}(\mathscr{F}))} \lesssim 1.$$

So, combining the last few steps, there is an absolute constant c such that

$$\mathbb{P}\Big(\Delta_t'(\mathbf{S})\ \geq\ \tfrac{c\cdot\nu(\mathbf{A},\mathbf{B})^2\cdot\log(\mathsf{card}(\mathscr{F}))}{\sqrt{t}}\Big)\ \leq\ \tfrac{1}{t},$$

as needed. ■

[††]Under the choice of $p = \log(\mathsf{card}(\mathscr{F})) = \log(2dd')$ that will be made at the end of this argument, it is straightforward to check that the condition (47) holds under Assumption 1.

Proof of Lemma 6 (ii). Note that for each $i \in [t]$ and $\mathbf{j} \in \mathscr{J}$, we have

$$f_{\mathbf{j}}(\mathbf{D}_i) - \mathbb{E}[f_{\mathbf{j}}(\mathbf{D}_i)] \;=\; f_{\mathbf{j}}(\mathbf{M}_i) \;=\; \mathbf{s}_i^T(\mathbf{B}\mathbf{C}_{\mathbf{j}}^T\mathbf{A}^T)\mathbf{s}_i - \mathrm{tr}(\mathbf{B}\mathbf{C}_{\mathbf{j}}^T\mathbf{A}^T), \tag{48}$$

which is a centered sub-Gaussian quadratic form. Due to the bound (35), we have

$$\Big\| f_{\mathbf{j}}(\mathbf{D}_i) - \mathbb{E}[f_{\mathbf{j}}(\mathbf{D}_i)] \Big\|_{\psi_1} \;\leq\; c\,\nu(\mathbf{A},\mathbf{B}). \tag{49}$$

Furthermore, this can be combined with a standard concentration bound for sums of independent sub-exponential random variables (Lemma 12) to show that for any $r \geq 0$,

$$\mathbb{P}\Big(\Big| \tfrac{1}{t}\textstyle\sum_{i=1}^t f_{\mathbf{j}}(\mathbf{D}_i) - \mathbb{E}[f_{\mathbf{j}}(\mathbf{D}_i)] \Big| \;\geq\; r\,\nu(\mathbf{A},\mathbf{B}) \Big) \;\leq\; 2\exp\Big(-c\cdot t\cdot \min(r^2, r) \Big). \tag{50}$$

Hence, taking a union bound over all $\mathbf{j}$ gives

$$\mathbb{P}\Big(\Delta_t''(\mathbf{S}) \geq\; r\,\nu(\mathbf{A},\mathbf{B}) \Big) \;\leq\; 2\exp\Big(\log(\mathsf{card}(\mathscr{F})) - c\cdot t\cdot \min(r^2, r) \Big). \tag{51}$$

Regarding the choice of r, note that by Assumption 1, we have $\frac{1}{\sqrt{t}}\sqrt{\log(\mathsf{card}(\mathscr{F}))} \lesssim 1$. It follows that there is a sufficiently large absolute constant $c_1 > 0$ such that if we put

$$r = \tfrac{c_1}{\sqrt{t}}\sqrt{\log(\mathsf{card}(\mathscr{F}))},$$

then

$$c\,t\,\min(r^2, r) \geq 2\log(\mathsf{card}(\mathscr{F})),$$

where c is the same as in the bound (51). In turn, this implies

$$\mathbb{P}\Big(\Delta_t''(\mathbf{S}) \;\geq\; \tfrac{c_1}{\sqrt{t}}\sqrt{\log(\mathsf{card}(\mathscr{F}))}\cdot\nu(\mathbf{A},\mathbf{B}) \Big) \;\leq\; 2\exp(-\log(\mathsf{card}(\mathscr{F})) \;= \frac{1}{dd'}, \tag{52}$$

as desired. ∎

Appendix D. Proof of Propositions 3 and 4 in case (b) (length sampling)

In order to carry out the proof Propositions 3 and 4 in the case of length sampling (Assumption 1 (b)), there are only two bounds that need to be updated. Namely, we must derive new bounds on $\|f_{\mathbf{j}}(\mathbf{D}_1) - \mathbb{E}[f_{\mathbf{j}}(\mathbf{D}_1)]\|_{\psi_1}$ and $\|f_{\mathbf{j}}(\mathbf{D}_1)\|_{\psi_1}$ in order to account for the new distributional assumptions in case (b). Both of the new bounds will turn out to be of order $\|\mathbf{A}\|_F\|\mathbf{B}\|_F$, and consequently, the result of the propositions in case (b) will have the same form as in case (a), but with $\|\mathbf{A}\|_F\|\mathbf{B}\|_F$ replacing $\nu(\mathbf{A},\mathbf{B})$.

To derive the bound on $\|f_{\mathbf{j}}(\mathbf{D}_1) - \mathbb{E}[f_{\mathbf{j}}(\mathbf{D}_1)]\|_{\psi_1}$, first note that

$$\begin{aligned} |\mathbb{E}[f_{\mathbf{j}}(\mathbf{D}_1)]| &= |\mathrm{tr}(\mathbf{C}_{\mathbf{j}}^T\mathbf{A}^T\mathbf{B})| \\ &\leq \|\mathbf{A}\|_F\|\mathbf{B}\|_F. \end{aligned}$$

Consequently,

$$\begin{aligned}\|f_{\mathbf{j}}(\mathbf{D}_1) - \mathbb{E}[f_{\mathbf{j}}(\mathbf{D}_1)]\|_{\psi_1} &\leq \|f_{\mathbf{j}}(\mathbf{D}_1)\|_{\psi_1} + \|\mathbb{E}[f_{\mathbf{j}}(\mathbf{D}_1)]\|_{\psi_1} \\ &\leq \|f_{\mathbf{j}}(\mathbf{D}_1)\|_{\psi_1} + c\|\mathbf{A}\|_F\|\mathbf{B}\|_F.\end{aligned} \tag{53}$$

Hence, it remains to show that $\|f_{\mathbf{j}}(\mathbf{D}_1)\|_{\psi_1} \leq c\|\mathbf{A}\|_F\|\mathbf{B}\|_F$, which is the content of Lemma 7 below. ■

Lemma 7 *If* $\mathbf{S}$ *is generated by length sampling with the probabilities in line* (5)*, then for any* $\mathbf{j} \in \mathscr{J}$*, we have the bound*

$$\big\|f_{\mathbf{j}}(\mathbf{D}_1)\big\|_{\psi_1} \leq 2\|\mathbf{A}\|_F\|\mathbf{B}\|_F. \tag{54}$$

Proof By the definition of the ψ_1-Orlicz norm, it suffices to find a value of $r > 0$ so that $\mathbb{E}\big[\exp\big(|f_{\mathbf{j}}(\mathbf{D}_1)|/r\big)\big]$ is at most 2. Due to the Cauchy-Schwarz inequality, the non-zero length-sampling probabilities p_l satisfy

$$\frac{1}{p_l} \leq \frac{\sqrt{\sum_{j=1}^n \|\mathbf{e}_j^T\mathbf{A}\|_2^2}\sqrt{\sum_{j=1}^n \|\mathbf{e}_j^T\mathbf{B}\|_2^2}}{\|\mathbf{e}_l^T\mathbf{A}\|_2\|\mathbf{e}_l^T\mathbf{B}\|_2} = \frac{\|\mathbf{A}\|_F\|\mathbf{B}\|_F}{\|\mathbf{e}_l^T\mathbf{A}\|_2\|\mathbf{e}_l^T\mathbf{B}\|_2}.$$

Consequently, for each $r > 0$ we have

$$\begin{aligned}\mathbb{E}\Big[\exp\Big(\tfrac{|f_{\mathbf{j}}(\mathbf{D}_1)|}{r}\Big)\Big] &= \sum_{l\in[n]:\, p_l>0} p_l \cdot \exp\Big(\tfrac{1}{r}\big|f_{\mathbf{j}}(\tfrac{1}{p_l}\mathbf{A}^T\mathbf{e}_l\mathbf{e}_l^T\mathbf{B})\big|\Big) \\ &\leq \max_{l\in[n]:\, p_l>0} \exp\Big(\tfrac{1}{r}\tfrac{1}{p_l}\big|f_{\mathbf{j}}(\mathbf{A}^T\mathbf{e}_l\mathbf{e}_l^T\mathbf{B})\big|\Big) \\ &\leq \max_{l\in[n]} \exp\Big(\tfrac{1}{r}\|\mathbf{A}\|_F\|\mathbf{B}\|_F\big|\tfrac{\mathbf{e}_l^T\mathbf{B}}{\|\mathbf{e}_l^T\mathbf{B}\|_2}\mathbf{C}_{\mathbf{j}}^T\tfrac{\mathbf{A}^T\mathbf{e}_l}{\|\mathbf{A}^T\mathbf{e}_l\|_2}\big|\Big) \\ &\leq \exp\Big(\tfrac{1}{r}\|\mathbf{A}\|_F\|\mathbf{B}\|_F\ \|\mathbf{C}_{\mathbf{j}}\|_2\Big) \\ &= \exp\Big(\tfrac{1}{r}\|\mathbf{A}\|_F\|\mathbf{B}\|_F\Big).\end{aligned}$$

Hence, if we take $r = 2\|\mathbf{A}\|_F\|\mathbf{B}\|_F$, then the right hand side is at most $e^{1/2} \leq 2$. ■

Appendix E. Proof of Propositions 3 and 4 in case (c) (SRHT)

The steps needed to extend the propositions in the case of SRHT matrices follows the same pattern as in case (b). However, there is a small subtlety insofar as all of the analysis is done conditionally on the matrix of signs $\mathbf{D}_n^\circ$ in the product $\mathbf{S} = \mathbf{P}_n(\frac{1}{\sqrt{n}}\mathbf{H}_n)\mathbf{D}_n^\circ$. Hence, it suffices to bound the ψ_1 Orlicz norm of $f_{\mathbf{j}}(\mathbf{D}_1)$ conditionally on $\mathbf{D}_n^\circ$, as well as the conditional expectation $|\mathbb{E}[f_{\mathbf{j}}(\mathbf{D}_1)|\mathbf{D}_n^\circ]$. Regarding the conditional expectation, it can be checked that $\mathbb{E}[\mathbf{S}^T\mathbf{S}|\mathbf{D}_n^\circ] = \mathbf{I}_n$, and it follows that

$$|\mathbb{E}[f_{\mathbf{j}}(\mathbf{D}_1)|\mathbf{D}_n^\circ]| = |\mathrm{tr}(\mathbf{B}\mathbf{C}_{\mathbf{j}}^T\mathbf{A}^T)| \leq \nu(\mathbf{A},\mathbf{B}).$$

Since we are not aware of a standard notation for a conditional Orlicz norm, we define

$$\big\|f_{\mathbf{j}}(\mathbf{D}_1)\big|\mathbf{D}_n^\circ\big\|_{\psi_1} := \inf\Big\{r>0 \;\Big|\; \mathbb{E}\big[\psi_1\big(|f_{\mathbf{j}}(\mathbf{D}_1)|/r\big)\big|\mathbf{D}_n^\circ\big] \le 1\Big\},$$

which is a random variable, since it is a function of $\mathbf{D}_n^\circ$. The following lemma provides a bound on this quantity, which turns out to be of order $\log(n)\,\nu(\mathbf{A},\mathbf{B})$. For this reason, the SRHT case (c) of Propositions 3 and 4 will have the same form as case (a), but with $\log(n)\,\nu(\mathbf{A},\mathbf{B})$ replacing $\nu(\mathbf{A},\mathbf{B})$. ■

Lemma 8 *If* $\mathbf{S}$ *is an SRHT matrix, then the following bound holds with probability at least* $1-c/n$,

$$\big\|f_{\mathbf{j}}(\mathbf{D}_1)\big|\mathbf{D}_n^\circ\big\|_{\psi_1} \le c\cdot\log(n)\cdot\nu(\mathbf{A},\mathbf{B}). \tag{55}$$

Proof By the definition of the conditional ψ_1-Orlicz norm, it suffices to find a value of $r>0$ so that $\mathbb{E}\big[\exp\big(|f_{\mathbf{j}}(\mathbf{D}_1)|/r\big)\big|\mathbf{D}_n^\circ\big]$ is at most 2 (with the stated probability).

For an SRHT matrix $\mathbf{S}=\mathbf{P}\frac{1}{\sqrt{n}}\mathbf{H}_n\mathbf{D}_n^\circ$, recall that the rows of $\sqrt{t}\mathbf{P}$ are sampled uniformly at random from the set $\{\frac{1}{\sqrt{1/n}}\mathbf{e}_1,\ldots,\frac{1}{\sqrt{1/n}}\mathbf{e}_n\}$. It follows that

$$\begin{aligned}\mathbb{E}\Big[\exp\Big(\tfrac{|f_{\mathbf{j}}(\mathbf{D}_1)|}{r}\Big)\Big|\mathbf{D}_n^\circ\Big] &= \mathbb{E}\Big[\exp\Big(\tfrac{1}{r}\big|\mathrm{tr}(\mathbf{C}_{\mathbf{j}}^T\mathbf{A}^T\mathbf{s}_1\mathbf{s}_1^T\mathbf{B})\big|\Big)\Big|\mathbf{D}_n^\circ\Big]\\ &= \frac{1}{n}\sum_{l=1}^n \exp\Big(\tfrac{1}{r}\big|\mathbf{e}_l^T\big(\mathbf{H}_n\mathbf{D}_n^\circ\big)\mathbf{B}\mathbf{C}_{\mathbf{j}}^T\mathbf{A}^T\big(\mathbf{D}_n^\circ\mathbf{H}_n^T\big)\mathbf{e}_l\big|\Big).\end{aligned}$$

Next, let $\boldsymbol{\varepsilon}_l\in\mathbb{R}^n$ be the lth row of $\mathbf{H}_n\mathbf{D}_n^\circ$, which gives

$$\begin{aligned}\mathbb{E}\Big[\exp\Big(\tfrac{|f_{\mathbf{j}}(\mathbf{D}_1)|}{r}\Big)\Big|\mathbf{D}_n^\circ\Big] &= \frac{1}{n}\sum_{l=1}^n \exp\Big(\big|\tfrac{\boldsymbol{\varepsilon}_l^T(\mathbf{B}\mathbf{C}_{\mathbf{j}}^T\mathbf{A}^T)\boldsymbol{\varepsilon}_l}{r}\big|\Big)\\ &\le \exp\Big(\max_{l\in[n]}\big|\tfrac{\boldsymbol{\varepsilon}_l^T(\mathbf{B}\mathbf{C}_{\mathbf{j}}^T\mathbf{A}^T)\boldsymbol{\varepsilon}_l}{r}\big|\Big).\end{aligned} \tag{56}$$

Recalling that $\mathbf{D}_n^\circ=\mathsf{diag}(\boldsymbol{\varepsilon})$ where $\boldsymbol{\varepsilon}\in\mathbb{R}^n$ is a vector of i.i.d. Rademacher variables, and that all entries of $\mathbf{H}_n$ are ±1, it follows that $\boldsymbol{\varepsilon}_l$ has the same distribution as $\boldsymbol{\varepsilon}$ for each $l\in[n]$. Consequently, each quadratic form $\boldsymbol{\varepsilon}_l^T(\mathbf{B}\mathbf{C}_{\mathbf{j}}^T\mathbf{A}^T)\boldsymbol{\varepsilon}_l$ concentrates around $\mathrm{tr}(\mathbf{B}\mathbf{C}_{\mathbf{j}}^T\mathbf{A}^T)$, and we can use a union bound to control the maximum of these quadratic forms. Note also that the matrix $\mathbf{B}\mathbf{C}_{\mathbf{j}}^T\mathbf{A}^T$ is rank-1, and so $\|\mathbf{B}\mathbf{C}_{\mathbf{j}}^T\mathbf{A}^T\|_2^2=\|\mathbf{B}\mathbf{C}_{\mathbf{j}}^T\mathbf{A}^T\|_F^2$, Hence, by choosing the parameter u to be proportional to $\log(n)\cdot\|\mathbf{B}\mathbf{C}_{\mathbf{j}}^T\mathbf{A}^T\|_F$ in the Hanson-Wright inequality (Lemma 11), and using a union bound, there is an absolute constant $c>0$ such that

$$\mathbb{P}\Big(\max_{l\in[n]}\;\big|\boldsymbol{\varepsilon}_l^T(\mathbf{B}\mathbf{C}_{\mathbf{j}}^T\mathbf{A}^T)\boldsymbol{\varepsilon}_l\big| \;\ge\; \mathrm{tr}(\mathbf{B}\mathbf{C}_{\mathbf{j}}^T\mathbf{A}^T)+c\log(n)\|\mathbf{B}\mathbf{C}_{\mathbf{j}}^T\mathbf{A}^T\|_F\Big) \;\le\; \tfrac{c}{n}. \tag{57}$$

Furthermore, noting that $\mathrm{tr}(\mathbf{B}\mathbf{C}_j^T\mathbf{A}^T)$ and $\|\mathbf{B}\mathbf{C}_{\mathbf{j}}^T\mathbf{A}^T\|_F$ are both at most $\nu(\mathbf{A},\mathbf{B})$, we have

$$\mathbb{P}\Big(\max_{l\in[n]}\;\big|\boldsymbol{\varepsilon}_l^T(\mathbf{B}\mathbf{C}_{\mathbf{j}}^T\mathbf{A}^T)\boldsymbol{\varepsilon}_l\big| \;\ge\; 2c\log(n)\nu(\mathbf{A},\mathbf{B})\Big) \;\le\; \tfrac{c}{n}. \tag{58}$$

Finally, this means that if we take $r = 4c\log(n)\nu(\mathbf{A},\mathbf{B})$ in the bound (56), then the event

$$\mathbb{E}\big[\exp\big(|f_{\mathbf{j}}(\mathbf{D}_1)|/r\big)\big|\mathbf{D}_n^{\circ}\big] \leq e^{1/2}$$

holds with probability at least $1 - \frac{c}{n}$, which completes the proof, since $e^{1/2} \leq 2$. ■

Appendix F. Technical Lemmas

Lemma 9 (Facts about Orlicz norms) *Orlicz norms have the following properties, where c, c_1, and c_2 are positive absolute constants.*

1. *For any random variable X, and any $p \geq 1$,*

$$\|X\|_p \leq cp\|X\|_{\psi_1} \tag{59}$$

$$\|X\|_p \leq c\sqrt{p}\|X\|_{\psi_2} \tag{60}$$

$$\|X\|_{\psi_1} \leq c\|X\|_{\psi_2}. \tag{61}$$

2. *If $X \sim \mathcal{N}(0,\sigma^2)$, then $\|X\|_{\psi_2} \leq c\,\sigma$.*

3. *Let $p \geq 1$. For any sequence of random variables $X_1, \ldots, X_d$,*

$$\Big\| \max_{1\leq j\leq d} X_j \Big\|_p \leq d^{1/p} \max_{1\leq j\leq d} \|X_j\|_p$$

and

$$\Big\| \max_{1\leq j\leq d} X_j \Big\|_{\psi_1} \leq c\log(d) \max_{1\leq j\leq d} \|X_j\|_{\psi_1}$$

4. *Let X be any random variable. Then, for any $x > 0$ and $p \geq 1$, we have*

$$\mathbb{P}\big(|X| \geq x\big) \leq \Big(\tfrac{\|X\|_p}{x}\Big)^p.$$

and

$$\mathbb{P}\big(|X| \geq x\big) \leq c_1 e^{-c_2 x/\|X\|_{\psi_1}}.$$

Proof In part 1, line (59) follows from line 5.11 of Vershynin (2012), line (60) follows from definition 5.13 of Vershynin (2012), and line (61) follows from p.94 of van der Vaart and Wellner (1996). Next, part 2 follows from the definition of the ψ_2-Orlicz norm and the moment generating function for $\mathcal{N}(0,\sigma^2)$. Part 3 is due to Lemma 2.2.2 of van der Vaart and Wellner (1996). Lastly, part 4 follows from Markov's inequality and line 5.14 of Vershynin (2012). ■

Lemma 10 (Rosenthal's inequality with best constants) *Fix any number $p > 2$. Let $Y_1, \ldots, Y_t$ be independent random variables with $\mathbb{E}[Y_i] = 0$ and $\mathbb{E}\big[|Y_i|^p\big] < \infty$ for all $1 \leq i \leq t$. Then,*

$$\big\| \sum_{i=1}^t Y_i \big\|_p \leq c\big(\tfrac{p}{\log(p)}\big) \cdot \max\Big\{ \big\| \sum_{i=1}^t Y_i \big\|_2 \,,\, \big(\sum_{i=1}^t \big\| Y_i \big\|_p^p \big)^{1/p} \Big\}. \tag{62}$$

Proof See the paper Johnson et al. (1985). The statement above differs slightly from the Theorem 4.1 in the paper Johnson et al. (1985), which requires symmetric random variables, but the remark on p.247 of that paper explains why the variables $Y_1, \ldots, Y_t$ need not be symmetric as long as they have mean 0. ■

Lemma 11 (Hanson-Wright inequality) *Let $\mathbf{x} = (X_1, \ldots, X_n)$ be a vector of independent sub-Gaussian random variables with $\mathbb{E}[X_j] = 0$, and $\|X_j\|_{\psi_2} \leq \kappa$ for all $1 \leq j \leq n$. Also, let $\mathbf{H} \in \mathbb{R}^{n \times n}$ be any fixed non-zero matrix. Then, there is an absolute constant $c > 0$ such that for any $u \geq 0$,*

$$\mathbb{P}\Big(\big| \mathbf{x}^T \mathbf{H} \mathbf{x} - \mathbb{E}[\mathbf{x}^T \mathbf{H} \mathbf{x}] \big| \;\geq\; u \Big) \;\leq\; 2 \exp\Big(- c \cdot \min\Big\{ \tfrac{u^2}{\kappa^2 \|\mathbf{H}\|_F^2}, \, \tfrac{u}{\kappa \|\mathbf{H}\|_2} \Big\} \Big). \tag{63}$$

Proof See the paper Rudelson and Vershynin (2013). ■

Lemma 12 (Bernstein inequality for sub-exponential variables) *Let $Y_1, \ldots, Y_t$ be independent random variables with $\mathbb{E}[Y_i] = 0$ and $\|Y_i\|_{\psi_1} \leq \kappa$ for all $1 \leq i \leq t$. Then, there is an absolute constant $c > 0$, such that for any $u \geq 0$,*

$$\mathbb{P}\Big(\Big| \tfrac{1}{t} \sum_{i=1}^t Y_i \Big| \geq \kappa \cdot u \Big) \;\leq\; 2 \exp\Big(- c \cdot t \cdot \min(u^2, u) \Big). \tag{64}$$

Proof See Proposition 16 in Vershynin (2012). ■

Lemma 13 ((Chernozhukov et al., 2016)) *If η is a non-negative random variable, and there are numbers $a, b > 0$ such that*

$$\mathbb{P}(\eta > x) \leq a e^{-x/b},$$

for all $x > 0$, then the following bound holds for all $r > 0$,

$$\mathbb{E}[\eta^3 \cdot 1\{\eta > r\}] \leq 6a(r + b)^3 e^{-r/b}.$$

Proof See Lemma 6.6 in Chernozhukov et al. (2016). ■

Remark. The following lemma may be of independent interest, since it provides an explicit bound on the ψ_1-Orlicz norm of a centered sub-Gaussian quadratic form. Although this bound follows from the Hanson-Wright inequality, we have not seen it stated in the literature.

Lemma 14 *Let $\mathbf{x} = (X_1, \ldots, X_n)$, be independent random variables satisfying $\mathbb{E}[X_j] = 0$ and $\|X_j\|_{\psi_2} \leq \kappa$ for all $1 \leq j \leq n$. Also, let $\mathbf{H} \in \mathbb{R}^{n\times n}$ be a non-zero fixed matrix. Then, there is an absolute constant $c > 0$ such that*

$$\Big\|\mathbf{x}^T\mathbf{H}\mathbf{x} - \mathbb{E}[\mathbf{x}^T\mathbf{H}\mathbf{x}]\Big\|_{\psi_1} \;\leq\; c\kappa^2 \tfrac{\|\mathbf{H}\|_F^2}{\|\mathbf{H}\|_2}.$$

Proof Define the random variable $Q := \mathbf{x}^T\mathbf{H}\mathbf{x} - \mathbb{E}[\mathbf{x}^T\mathbf{H}\mathbf{x}]$. By the definition of the ψ_1-Orlicz norm, it suffices to find a value $r > 0$ such that $\mathbb{E}\big[\exp(|Q|/r)\big] \leq 2$. Using the tail-sum formula, and the change of variable $v = e^{u/r}$, we have

$$\begin{aligned}\mathbb{E}\big[\exp(|Q|/r)\big] &\leq 1 + \int_1^\infty \mathbb{P}(\exp(|Q|/r) > v)dv \\ &= 1 + \tfrac{1}{r}\int_0^\infty \mathbb{P}(|Q| > u)\cdot e^{u/r}du.\end{aligned}$$

Next, we employ the Hanson-Wright inequality (Lemma 11). By considering the "threshold" $u^* := \kappa^2\frac{\|\mathbf{H}\|_F^2}{\|\mathbf{H}\|_2}$, it is helpful to note that the quantities in the exponent of the Hanson-Wright inequality satisfy $\frac{u^2}{\kappa^4\|\mathbf{H}\|_F^2} \leq \frac{u}{\kappa^2\|\mathbf{H}\|_2}$ if and only if $u \leq u^*$. Hence,

$$\begin{aligned}\mathbb{E}\big[\exp(|Q|/r)\big] &\leq 1 + \tfrac{1}{r}\int_0^\infty \exp\Big\{-c\min\big(\tfrac{u^2}{\kappa^4\|\mathbf{H}\|_F^2}, \tfrac{u}{\kappa^2\|\mathbf{H}\|_2}\big)\Big\}\cdot e^{u/r}du \\ &\leq 1 + \tfrac{1}{r}\int_0^{u^*} e^{u/r}du + \tfrac{1}{r}\int_{u^*}^\infty \exp\Big\{-u\big(\tfrac{c}{\kappa^2\|\mathbf{H}\|_2} - \tfrac{1}{r}\big)\Big\}du.\end{aligned}$$

Evaluating the last two integrals directly, if we let $C' := \frac{c}{\kappa^2\|\mathbf{H}\|_2} - \frac{1}{r}$ and choose r so that $C' > 0$, then

$$\begin{aligned}\mathbb{E}\big[\exp(|Q|/r)\big] &\leq e^{u^*/r} + \tfrac{1}{rC'}e^{-u^*C'}, \\ &\leq e^{u^*/r} + \tfrac{1}{\frac{c\cdot r}{\kappa^2\|\mathbf{H}\|_2} - 1}.\end{aligned}$$

Note that the condition $C' > 0$ means that it is necessary to have $r > \frac{1}{c}\kappa^2\|\mathbf{H}\|_2$. To finish the argument, we further require that r is large enough so that (say)

$$\tfrac{u^*}{r} \leq \tfrac{1}{4} \quad \text{and} \quad \tfrac{c\cdot r}{\kappa^2\|\mathbf{H}\|_2} \geq 3, \tag{65}$$

which ensures

$$\mathbb{E}\big[\exp(|Q|/r)\big] \leq e^{1/4} + \tfrac{1}{2} < 2,$$

as desired. Note that the constraints (65) are the same as

$$r \geq 4\kappa^2\tfrac{\|\mathbf{H}\|_F^2}{\|\mathbf{H}\|_2} \quad \text{and} \quad r \geq \tfrac{3}{c}\kappa^2\|\mathbf{H}\|_2.$$

Due to the basic fact that $\|\mathbf{H}\|_2 \leq \|\mathbf{H}\|_F$ for all matrices $\mathbf{H}$, it follows that whenever $r \geq \max(4, \frac{3}{c})\kappa^2 \frac{\|\mathbf{H}\|_F^2}{\|\mathbf{H}\|_2}$, we have $\mathbb{E}\big[\exp(|Q|/r)\big] < 2$. ■

Remark. The following lemma is a basic fact about the d_{LP} metric, but may not be widely known, and so we give a proof. Recall also that we use the generalized inverse $F_V^{-1}(\alpha) := \inf\{z \in \mathbb{R} \,|\, F_V(z) \geq \alpha\}$, where F_V denotes the c.d.f. of V.

Lemma 15 *Fix $\alpha \in (0, 1/2)$ and suppose there is some $\epsilon \in (0, \alpha)$ such that random variables U and V satisfy*

$$d_{\mathrm{LP}}(\mathcal{L}(U), \mathcal{L}(V)) \leq \epsilon.$$

Then, the quantiles of U and V satisfy

$$\big|F_U^{-1}(1-\alpha) - F_V^{-1}(1-\alpha)\big| \;\leq\; \psi_\alpha(\epsilon), \tag{66}$$

where the right side is defined as

$$\psi_\alpha(\epsilon) := F_U^{-1}(1-\alpha+\epsilon) - F_U^{-1}(1-\alpha-\epsilon) + \epsilon.$$

Proof Consider the Lévy metric, defined as

$$d_{\mathrm{L}}(\mathcal{L}(U), \mathcal{L}(V)) := \inf\Big\{\epsilon > 0 \,\Big|\, F_U(x-\epsilon) - \epsilon \leq F_V(x) \leq F_U(x+\epsilon) + \epsilon \text{ for all } x \in \mathbb{R}\Big\}.$$

It is a fact that this metric is always dominated by the d_{LP} metric in the sense that

$$d_{\mathrm{L}}(\mathcal{L}(U), \mathcal{L}(V)) \;\leq\; d_{\mathrm{LP}}(\mathcal{L}(U), \mathcal{L}(V)),$$

for all scalar random variables U and V (Huber and Ronchetti, 2009, p.36). Based on the definition of the d_{L} metric, it is straightforward to check that the following inequalities hold under the assumption of the lemma,

$$F_U^{-1}(1-\alpha-\epsilon) - \epsilon \;\leq\; F_V^{-1}(1-\alpha) \;\leq\; F_U^{-1}(1-\alpha+\epsilon) + \epsilon.$$

(Specifically, consider the choices $x = F_V^{-1}(1-\alpha)$ and $x = F_U^{-1}(1-\alpha+\epsilon)+\epsilon$.) Next, if we subtract $F_U^{-1}(1-\alpha)$ from each side of the inequalities above, and note that $F_U^{-1}(\cdot)$ is non-decreasing, it follows that if we put $a = F_U^{-1}(1-\alpha+\epsilon) - F_U^{-1}(1-\alpha) + \epsilon$ and $b = F_U^{-1}(1-\alpha) - F_U^{-1}(1-\alpha-\epsilon) + \epsilon$, then

$$|F_V^{-1}(\alpha) - F_U^{-1}(\alpha)| \;\leq\; \max\{a, b\} \;\leq\; \psi_\alpha(\epsilon),$$

as needed. ■

Lemma 16 *Under Assumption 1 (a), the quantity*

$$\delta_0 = t^{-1/8} \log^{1/2}(d) \nu(\mathbf{A}, \mathbf{B})^{3/4}$$

satisfies conditions (30) *and* (31).

Proof Consider the number

$$\delta_1(r) := \frac{\log^2(d)\nu(\mathbf{A},\mathbf{B})}{\sqrt{t}} \cdot r,$$

where $r \geq 1$ is a free parameter to be adjusted. Based on the bound (29), it is easy to check that plugging $\delta_1(r)$ into $K_t(\cdot)$ and $J_t(\cdot)$ leads to

$$K_t(\delta_1(r)) + J_t(\delta_1(r)) \leq c \cdot \nu(\mathbf{A},\mathbf{B})^3 \cdot \log(d)^3 \cdot r^3 \cdot \exp(-r/c)$$

and if we take $r \geq c\log(\log(d)^4)$, then

$$K_t(\delta_1(r)) + J_t(\delta_1(r)) \leq c\,\nu(\mathbf{A},\mathbf{B})^3,$$

as desired in (30). Hence, as long as there is a choice of r satisfying

$$r \geq c\log(\log(d)^4) \qquad \text{and} \qquad \delta_1(r) = \delta_0,$$

then $\delta_1(r)$ will satisfy both of the desired constraints (30) and (31). Solving the equation $\delta_1(r) = \delta_0$ gives

$$r = t^{3/8} \cdot \log^{-3/2}(d) \cdot \nu(\mathbf{A},\mathbf{B})^{-1/4},$$

and then the condition $r \geq c\log(\log(d)^4)$ is the same as

$$\begin{aligned} t &\geq \Big(c\,\nu(\mathbf{A},\mathbf{B})^{1/4}\log(d)^{3/2} \cdot \log(\log(d)^4)\Big)^{8/3} \\ &= c\,\nu(\mathbf{A},\mathbf{B})^{2/3}\log(d)^4 \cdot \log(\log(d)^4)^{8/3}, \end{aligned} \tag{67}$$

which holds under Assumption 1 (a). ■

References

N. Ailon and B. Chazelle. Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform. In *Annual ACM Symposium on Theory of Computing (STOC)*, 2006.

N. Ailon and E. Liberty. Fast dimension reduction using Rademacher series on dual BCH codes. *Discrete & Computational Geometry*, 42(4):615–630, 2009.

S. Ar, M. Blum, B. Codenotti, and P. Gemmell. Checking approximate computations over the reals. In *Annual ACM Symposium on Theory of Computing (STOC)*, 1993.

H. Avron, P. Maymounkov, and S. Toledo. Blendenpik: Supercharging lapack's least-squares solver. *SIAM Journal on Scientific Computing*, 32(3):1217–1236, 2010.

C. Boutsidis and A. Gittens. Improved matrix algorithms via the subsampled randomized hadamard transform. *SIAM Journal on Matrix Analysis and Applications*, 34(3):1301–1340, 2013.

C. Brezinski and M. R. Zaglia. *Extrapolation methods: theory and practice*. Elsevier, 2013.

C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011. URL `http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`.

J. Chang, W. Zhou, W.-X. Zhou, and L. Wang. Comparing large covariance matrices under weak conditions on the dependence structure and its application to gene clustering. *Biometrics*, 2016.

X. Chen. Gaussian and bootstrap approximations for high-dimensional u-statistics and their applications. *The Annals of Statistics*, 46(2):642–678, 2018.

V. Chernozhukov, D. Chetverikov, and K. Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786–2819, 2013.

V. Chernozhukov, D. Chetverikov, and K. Kato. Gaussian approximation of suprema of empirical processes. *The Annals of Statistics*, 42(4):1564–1597, 2014.

V. Chernozhukov, D. Chetverikov, and K. Kato. Comparison and anti-concentration bounds for maxima of Gaussian random vectors. *Probability Theory and Related Fields*, 162(1-2): 47–70, 2015.

V. Chernozhukov, D. Chetverikov, and K. Kato. Empirical and multiplier bootstraps for suprema of empirical processes of increasing complexity, and related Gaussian couplings. *Stochastic Processes and their Applications*, 2016.

V. Chernozhukov, D. Chetverikov, and K. Kato. Central limit theorems and bootstrap in high dimensions. *The Annals of Probability*, 45(4):2309–2352, 2017.

K. L. Clarkson and D. P. Woodruff. Low rank approximation and regression in input sparsity time. In *Annual ACM Symposium on theory of computing (STOC)*, 2013.

G. Dasarathy, P. Shah, B. Narayan Bhaskar, and R. D. Nowak. Sketching sparse matrices, covariances, and graphs via tensor products. *IEEE Transactions on Information Theory*, 61(3):1373–1388, 2015.

J. Demmel, I. Dumitriu, O. Holtz, and R. Kleinberg. Fast matrix multiplication is stable. *Numerische Mathematik*, 106(2):199–224, 2007.

J. D. Dixon. Estimating extremal eigenvalues and condition numbers of matrices. *SIAM Journal on Numerical Analysis*, 20(4):812–814, 1983.

P. Drineas and M. W. Mahoney. RandNLA: randomized numerical linear algebra. *Communications of the ACM*, 59(6):80–90, 2016.

P. Drineas, R. Kannan, and M. W. Mahoney. Fast Monte Carlo algorithms for matrices I: Approximating matrix multiplication. *SIAM Journal on Computing*, 36(1):132–157, 2006a.

P. Drineas, M. W. Mahoney, and S. Muthukrishnan. Sampling algorithms for ℓ_2 regression and applications. In *Annual ACM-SIAM Symposium on Discrete Algorithm (SODA)*, 2006b.

P. Drineas, M. W. Mahoney, and S. Muthukrishnan. Relative-error CUR matrix decompositions. *SIAM Journal on Matrix Analysis and Applications*, 30(2):844–881, September 2008.

P. Drineas, M. W. Mahoney, S. Muthukrishnan, and T. Sarlós. Faster least squares approximation. *Numerische Mathematik*, 117(2):219–249, 2011.

P. Drineas, M. Magdon-Ismail, M. W. Mahoney, and D. P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13: 3475–3506, 2012.

A. Frank and A. Asuncion. UCI machine learning repository, 2010. URL `http://archive.ics.uci.edu/ml`.

R. Freivalds. Fast probabilistic algorithms. *Mathematical Foundations of Computer Science*, pages 57–69, 1979.

N. Halko, P.-G. Martinsson, and J. A. Tropp. Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011.

N. J. Higham. *Accuracy and stability of numerical algorithms*. SIAM, 2002.

P. J. Huber and E. M. Ronchetti. *Robust Statistics*. Wiley, 2009.

W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary mathematics*, 26(189-206), 1984.

W. B. Johnson, G. Schechtman, and J. Zinn. Best constants in moment inequalities for linear combinations of independent and exchangeable random variables. *The Annals of Probability*, pages 234–253, 1985.

E. Liberty, F. Woolfe, P.-G. Martinsson, V. Rokhlin, and M. Tygert. Randomized algorithms for the low-rank approximation of matrices. *Proceedings of the National Academy of Sciences*, 104(51):20167–20172, 2007.

M. E. Lopes. Estimating the algorithmic variance of randomized ensembles via the bootstrap. *The Annals of Statistics*, 47(2):1088–1112, 2019.

M. E. Lopes, Z. Lin, and H.-G. Mueller. Bootstrapping max statistics in high dimensions: Near-parametric rates under weak variance decay and application to functional data analysis. *arXiv:1807.04429*, 2018a.

M. E. Lopes, S. Wang, and M. W. Mahoney. Error estimation for randomized least-squares algorithms via the bootstrap. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018b.

P. Ma, M. W. Mahoney, and B. Yu. A statistical perspective on algorithmic leveraging. In *International Conference on Machine Learning (ICML)*, 2014.

A. Magen and A. Zouzias. Low rank matrix-valued Chernoff bounds and approximate matrix multiplication. In *Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2011.

M. W. Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning*, 3(2):123–224, 2011.

R. Pagh. Compressed matrix multiplication. *ACM Transactions on Computation Theory*, 5(3):9, 2013.

M. Pilanci and M. J. Wainwright. Newton sketch: a near linear-time optimization algorithm with linear-quadratic convergence. *SIAM Journal on Optimization*, 27(1):205–245, 2017.

F. Roosta-Khorasani and M. W. Mahoney. Sub-sampled Newton methods II: local convergence rates. *arXiv:1601.04738*, 2016.

M. Rudelson and R. Vershynin. Hanson-Wright inequality and sub-Gaussian concentration. *Electronic Communications in Probability*, 18:9 pp., 2013.

T. Sarlós. Improved approximation algorithms for large matrices via random projections. In *Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2006.

A. Sidi. *Practical Extrapolation Methods: Theory and Applications*. Cambridge University Press, 2003.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing, Theory and Applications*. Cambridge University Press, 2012.

S. Wang. A practical guide to randomized matrix computations with MATLAB implementations. *arXiv:1505.07570*, 2015.

D. P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014.

F. Woolfe, E. Liberty, V. Rokhlin, and M. Tygert. A fast randomized algorithm for the approximation of matrices. *Applied and Computational Harmonic Analysis*, 25(3):335–366, 2008.

P. Xu, J. Yang, F. Roosta-Khorasani, C. Ré, and M. W. Mahoney. Sub-sampled Newton methods with non-uniform sampling. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3000–3008, 2016.

J. Yang, X. Meng, and M. W. Mahoney. Implementing randomized matrix algorithms in parallel and distributed environments. *Proceedings of the IEEE*, 104(1):58–92, 2016.