Statistics and Computing (2023) 33:34
https://doi.org/10.1007/s11222-022-10148-5

On randomized sketching algorithms and the Tracy–Widom law

Daniel Ahfock[1] · **William J. Astle**[1] · **Sylvia Richardson**[1]

Received: 12 December 2021 / Accepted: 3 September 2022 / Published online: 19 January 2023

Abstract
There is an increasing body of work exploring the integration of random projection into algorithms for numerical linear algebra. The primary motivation is to reduce the overall computational cost of processing large datasets. A suitably chosen random projection can be used to embed the original dataset in a lower-dimensional space such that key properties of the original dataset are retained. These algorithms are often referred to as sketching algorithms, as the projected dataset can be used as a compressed representation of the full dataset. We show that random matrix theory, in particular the Tracy–Widom law, is useful for describing the operating characteristics of sketching algorithms in the tall-data regime when the sample size n is much greater than the number of variables d. Asymptotic large sample results are of particular interest as this is the regime where sketching is most useful for data compression. In particular, we develop asymptotic approximations for the success rate in generating random subspace embeddings and the convergence probability of iterative sketching algorithms. We test a number of sketching algorithms on real large high-dimensional datasets and find that the asymptotic expressions give accurate predictions of the empirical performance.

Keywords Sketching · Random matrix theory · Random projection

1 Introduction

Sketching is a probabilistic data compression technique that makes use of random projection (Cormode 2011; Mahoney 2011; Woodruff 2014). Suppose interest lies in a $n \times d$ dataset $\boldsymbol{A}$. When n and or d are large, typical data analysis tasks will involve a heavy numerical computing load. This computational burden can be a practical obstacle for statistical learning with Big Data. When the sample size n is the computational bottleneck, sketching algorithms use a linear random projection to create a smaller sketched dataset of size $k \times d$, where $k \ll n$. The random projection can be represented as a $k \times n$ random matrix $\boldsymbol{S}$, and the sketched dataset $\widetilde{\boldsymbol{A}}$ is generated through the linear embedding $\widetilde{\boldsymbol{A}} = \boldsymbol{S}\boldsymbol{A}$. The smaller sketched dataset $\widetilde{\boldsymbol{A}}$ is used as a surrogate for the full dataset $\boldsymbol{A}$ within numerical routines. Through a judicious choice of the distribution on the random sketching matrix $\boldsymbol{S}$, it is often possible to bound the error that is introduced stochastically into calculations given the use of the randomized approximation $\widetilde{\boldsymbol{A}}$ in place of $\boldsymbol{A}$

✉ Daniel Ahfock
d.ahfock@uq.edu.au

[1] MRC Biostatistics Unit, University of Cambridge, Cambridge, UK

The selected distribution of the random sketching matrix $\boldsymbol{S}$ can be divided into two categories, data-oblivious sketches, where the distribution is not a function of the source data $\boldsymbol{A}$, and data-aware sketches, where the distribution is a function of $\boldsymbol{A}$. There are also hybrid approaches where a sketching matrix $\boldsymbol{S}$ is formed by taking $\boldsymbol{S} = \tilde{\boldsymbol{S}}\boldsymbol{A}^{\mathsf{T}}$ for some data-oblivious sketch $\tilde{\boldsymbol{S}}$. The majority of data-aware sketches perform weighted sampling with replacement, and are closely connected to finite population survey sampling methods (Ma et al. 2015; Quiroz et al. 2018). The analysis of data-oblivious sketches requires different methods to data-aware sketches, as there are no clear ties to finite-population subsampling. In general, data-oblivious sketches generate a dataset of k pseudo-observations, where each instance in the compressed representation $\widetilde{\boldsymbol{A}}$ has no exact counterpart in the original source dataset $\boldsymbol{A}$.

Three important data-oblivious sketches are the Gaussian sketch, the Hadamard sketch and the Clarkson–Woodruff sketch. The Gaussian sketch is the simplest of these, where each element in the $k \times n$ matrix $\boldsymbol{S}$ is an independent sample from a $N(0, 1/k)$ distribution. The Hadamard sketch uses structured elements for fast matrix multiplication, and the Clarkson–Woodruff uses sparsity in $\boldsymbol{S}$ for efficient computation of the sketched dataset. Other sketches that make use of

sparsity include the OSNAP (Nelson and Nguyên 2013) and LESS embeddings (Derezinski et al. 2021). The comparative performance between distributions on $\boldsymbol{S}$ is of interest, as there is a trade-off between the computational cost of calculating $\widetilde{\boldsymbol{A}}$ and the fidelity of the approximation $\widetilde{\boldsymbol{A}}$ with respect to original $\boldsymbol{A}$ when choosing the type of sketch. Our results help to establish guidelines for selecting the sketching distribution.

Sketching algorithms are typically framed using stochastic (δ, ϵ) error bounds, where the algorithm is shown to attain $(1 \pm \epsilon)$ accuracy with probability at least $1 - \delta$ (Woodruff 2014). These notions are made more precise in Sect. 2. Existing bounds are typically developed from a worst-case non-asymptotic viewpoint (Mahoney 2011; Woodruff 2014; Tropp 2011). We take a different approach, and use random matrix theory to develop asymptotic approximations to the success probability given the sketching distortion factor ϵ. Recent work has demonstrated the usefulness of random matrix theory to characterize the convergence rate of sketching-based iterative optimisation algorithms (Lacotte et al. 2020; Lacotte and Pilanci 2020).

Our main result is an asymptotic expression for the probability that a Gaussian based sketching algorithm satisfies general $(1 \pm \epsilon)$ probabilistic error bounds in terms of the Tracy–Widom law (Theorem 1), which describes the distribution of the extreme eigenvalues of large random matrices (Tracy and Widom 1994; Johnstone 2001). We then identify regularity conditions where other data-oblivious projections are expected to demonstrate the same limiting behavior (Theorem 3). If the motivation for using a sketching algorithm is data compression due to large n, the asymptotic approximations are of particular interest as they become more accurate as the computational benefits afforded by the use of a sketching algorithm increase in tandem. Empirical work has found that the quality of results can be consistent across the choice of random projections (Venkatasubramanian and Wang 2011; Le et al. 2013; Dahiya et al. 2018), and our results shed some light on this issue. An application is to determine the convergence probability when sketching is used in iterative least-squares optimisation. We test the asymptotic theory and find good agreement on datasets with large sample sizes $n \gg d$. Our theoretical and empirical results show that random matrix theory has an important role in the analysis of data-oblivious sketching algorithms for data compression.

2 Sketching

2.1 Data-oblivious sketches

As mentioned, a key component in a sketching algorithm is the distribution on $\boldsymbol{S}$.

- The uniform sketch, which implements subsampling uniformly with replacement followed by a rescaling step. The Uniform projection can be represented as $\boldsymbol{S} = \sqrt{n/k}\boldsymbol{\Phi}$. The random matrix $\boldsymbol{\Phi}$ subsamples k rows of $\boldsymbol{A}$ with replacement. Element $\Phi_{r,i} = 1$ if observation i in the source dataset is selected in the rth subsampling round $(r = 1, \ldots, k;\ i = 1 \ldots, n)$. The uniform sketch can be implemented in $O(k)$ time.
- The Gaussian sketch, which is formed by independently sampling each element of $\boldsymbol{S}$ from a $N(0, 1/k)$ distribution. Computation of the sketched data is $O(ndk)$.
- The Hadamard sketch is a structured random matrix (Ailon and Chazelle 2009). The sketching matrix is formed as $\boldsymbol{S} = \boldsymbol{\Phi}\boldsymbol{H}\boldsymbol{D}/\sqrt{k}$, where $\boldsymbol{\Phi}$ is a $k \times n$ matrix and $\boldsymbol{H}$ and $\boldsymbol{D}$ are both $n \times n$ matrices. The fixed matrix $\boldsymbol{H}$ is a Hadamard matrix of order n. A Hadamard matrix is a square matrix with elements that are either $+1$ or -1 and orthogonal rows. Hadamard matrices do not exist for all integers n, the source dataset can be padded with zeroes so that a conformable Hadamard matrix is available. The random matrix $\boldsymbol{D}$ is a diagonal matrix where each of the n diagonal entries is an independent Rademacher random variable. The random matrix $\boldsymbol{\Phi}$ subsamples k rows of $\boldsymbol{H}$ with replacement. The structure of the Hadamard sketch allows for fast matrix multiplication, reducing the complexity of the calculation of the sketched dataset relative to the Gaussian sketch, to $O(nd \log k)$ operations.
- The Clarkson–Woodruff sketch is a sparse random matrix (Clarkson and Woodruff 2013). The projection can be represented as the product of two independent random matrices, $\boldsymbol{S} = \boldsymbol{\Gamma}\boldsymbol{D}$, where $\boldsymbol{\Gamma}$ is a random $k \times n$ matrix and $\boldsymbol{D}$ is a random $n \times n$ matrix. The matrix $\boldsymbol{\Gamma}$ is initialized as a matrix of zeros. In each column, independently, one entry is selected and set to $+1$. The matrix $\boldsymbol{D}$ is a diagonal matrix where each of the n diagonal entries is an independent Rademacher random variable. This results in a sparse $\boldsymbol{S}$, where there is only one nonzero entry per column. The sparsity of the Clarkson–Woodruff sketch speeds up matrix multiplication, dropping the complexity of generating the sketched dataset to $O(nd)$.

The Gaussian sketch was central to early work on sketching algorithms (Sarlos 2006). The drawback of the Gaussian sketch is that computation of the sketched data is quite demanding, taking $O(ndk)$ operations. As such, there has been work on designing more computationally efficient random projections.

Sketch quality is commonly measured using ϵ-subspace embeddings (Woodruff (2014, Chapter 2), Meng and Mahoney 2013, Yang et al. 2015). These are defined below.

Definition 1 **ϵ-subspace embedding**

For a given $n \times d$ matrix $\boldsymbol{A}$, we call a $k \times n$ matrix $\boldsymbol{S}$ an ϵ-subspace embedding for $\boldsymbol{A}$, if for all vectors $\boldsymbol{z} \in \mathbb{R}^d$

$$(1-\epsilon)||\boldsymbol{Az}||_2^2 \leq ||\boldsymbol{SAz}||_2^2 \leq (1+\epsilon)||\boldsymbol{Az}||_2^2.$$

An ϵ-subspace preserves the linear structure of the original dataset up to a multiplicative $(1 \pm \epsilon)$ factor. Broadly speaking, the covariance matrix of the sketched dataset $\widetilde{\boldsymbol{A}} = \boldsymbol{SA}$ is similar to the covariance matrix of the source dataset $\boldsymbol{A}$ if ϵ is small. Mathematical arguments show that the sketched dataset is a good surrogate for many linear statistical methods if the sketching matrix $\boldsymbol{S}$ is an ϵ-subspace embedding for the original dataset, with ϵ sufficiently small (Woodruff 2014). Suitable ranges for ϵ depend on the task of interest and structural properties of the source dataset (Mahoney and Drineas 2016).

The Gaussian, Hadamard and Clarkson–Woodruff projections are popular data-oblivious projections as it is possible to argue that they produce ϵ-subspace embeddings with high probability for an arbitrary data matrix $\boldsymbol{A}$. It is considerably more difficult to establish universal worst case bounds for the uniform projection (Drineas et al. 2006; Ma et al. 2015). We include the uniform projection in our discussion as it is a useful baseline. Results for sub-Gaussian sketches (Nelson and Nguyên 2013) and LESS embeddings (Derezinski et al. 2021) are also included for comparison. Table 1 summarises some key properties of different sketching matrices.

2.2 Sketching algorithms

Sketching algorithms have been proposed for key linear statistical methods such as low rank matrix approximation, principal components analysis, linear discriminant analysis and ordinary least squares regression (Mahoney 2011; Woodruff 2014; Erichson et al. 2016; Falcone et al. 2021). Sketching has also been investigated for Bayesian posterior approximation (Bardenet and Maillard 2015; Geppert et al. 2017). A common thread throughout these works is the reliance on the generation of an ϵ-subspace embedding. In general, ϵ serves an approximation tolerance parameter, with smaller ϵ guaranteeing higher fidelity to exact calculation with respect to some divergence measure.

An example application of sketching is ordinary least squares regression (Sarlos 2006). The sketched responses and predictors are defined as $\widetilde{\boldsymbol{y}} = \boldsymbol{Sy}, \widetilde{\boldsymbol{X}} = \boldsymbol{SX}$. Let $\boldsymbol{\beta}_F = \text{argmin}_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X\beta}\|_2^2$, $\boldsymbol{\beta}_S = \text{argmin}_{\boldsymbol{\beta}} \|\widetilde{\boldsymbol{y}} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta}\|_2^2$, and $RSS_F = \|\boldsymbol{y} - \boldsymbol{X\beta}_F\|_2^2$. It is possible to establish the concrete bounds, that if $\boldsymbol{S}$ is an ϵ-subspace embedding for $\boldsymbol{A} = (\boldsymbol{y}, \boldsymbol{X})$ (Sarlos 2006), then

$$\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_F\|_2^2 \leq \frac{\epsilon^2}{\sigma_{\min}^2(\boldsymbol{X})} RSS_F,$$

where $\sigma_{\min}(\boldsymbol{X})$ represents the smallest singular value of the design matrix $\boldsymbol{X}$. If ϵ is very small, then $\boldsymbol{\beta}_S$ is a good approximation to $\boldsymbol{\beta}_F$.

Given the central role of ϵ-subspace embeddings (Definition 1), the success probability,

$$\Pr(\boldsymbol{S} \text{ is an } \epsilon\text{-subspace embedding for } \boldsymbol{A}) \tag{1}$$

is thus an important descriptive measure of the uncertainty attached to the randomized algorithm. The probability statement is over the random sketching matrix $\boldsymbol{S}$ with the dataset $\boldsymbol{A}$ treated as fixed. The embedding probability is difficult to characterize precisely using existing theory (Venkatasubramanian and Wang 2011). The bounds in Table 1 only give qualitative guidance about the embedding probability. Users will benefit from more prescriptive results in order to choose the sketch size k, and the type of sketch for applications (Grellmann et al. 2016; Geppert et al. 2017; Ahfock et al. 2020; Falcone et al. 2021).

Another use for sketching is in iterative solvers for ordinary least squares regression. A sketch $\widetilde{\boldsymbol{X}} = \boldsymbol{SX}$ can be used to generate a random preconditioner, $(\widetilde{\boldsymbol{X}}^\mathsf{T}\widetilde{\boldsymbol{X}})^{-1}$, that is then applied to the normal equations $\boldsymbol{X}^\mathsf{T}\boldsymbol{X\beta} = \boldsymbol{X}^\mathsf{T}\boldsymbol{y}$. The approach with a single sketched preconditioner is analysed in Pilanci and Wainwright (2016) and referred to as a Hessian sketch. Given some initial value $\boldsymbol{\beta}^{(0)}$, the iteration is defined as

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + (\widetilde{\boldsymbol{X}}^\mathsf{T}\widetilde{\boldsymbol{X}})^{-1}\boldsymbol{X}^\mathsf{T}(\boldsymbol{y} - \boldsymbol{X\beta}^{(t)}). \tag{2}$$

If $\widetilde{\boldsymbol{X}}^\mathsf{T}\widetilde{\boldsymbol{X}} = \boldsymbol{X}^\mathsf{T}\boldsymbol{X}$ the iteration will converge in a single step. The degree of noise in the preconditioner will be influenced by the sketch size k. A sufficient condition for convergence of the iteration (2) is that $\boldsymbol{S}$ is an ϵ-subspace embedding for $\boldsymbol{X}$ with $\epsilon < 0.5$ (Pilanci and Wainwright 2016). As is typical with randomized algorithms, we accept some failure probability in order to relax the computational demands. It is of interest to develop expressions for the failure probability of the algorithm as a function of the sketch size k, as this can give useful guidelines in practice. It is possible to establish worst case bounds using the results in Table 1, however we will aim to give a point estimate of the probability. Although it is possible to improve on the iteration (2) using acceleration methods (Meng et al. 2014; Dahiya et al. 2018; Lacotte et al. 2020), we focus on the basic iteration to introduce our asymptotic techniques.

2.3 Operating characteristics

Let the singular value decomposition of the source dataset be given by $\boldsymbol{A} = \boldsymbol{UDV}^\mathsf{T}$. Let $\sigma_{\min}(\boldsymbol{M})$ and $\sigma_{\max}(\boldsymbol{M})$ denote the minimum and maximum singular values respectively, of a matrix $\boldsymbol{M}$. Likewise, let $\lambda_{\min}(\boldsymbol{M})$ and $\lambda_{\max}(\boldsymbol{M})$ denote the

Table 1 Properties of different sketching matrices (see Woodruff 2014 and Derezinski et al. 2021 and the references therein)

Sketch	Sketching time	Required sketch size k
Gaussian	$O(ndk)$	$O((d+\log(1/\delta))/\epsilon^2)$
Hadamard	$O(nd\log k)$	$O((\sqrt{d}+\sqrt{\log n})^2(\log(d/\delta))/\epsilon^2)$
Clarkson–Woodruff	$O(nd)$	$O(d^2/(\delta\epsilon^2))$
Uniform	$O(k)$	–
Sub-Gaussian	$O(ndk)$	$O((d+\log(1/\delta))/\epsilon^2)$
LESS	$O(nd\log n + kd^2)$	$O((d\log(d/\delta))/\epsilon^2)$

The third column refers to the necessary sketch size k to obtain an ϵ-subspace embedding for an arbitrary $n \times d$ source dataset with at least probability $(1-\delta)$

minimum and maximum eigenvalues of a matrix $\boldsymbol{M}$. It is possible to show

$$\begin{aligned}&\Pr(\boldsymbol{S} \text{ is an } \epsilon\text{-subspace embedding for } \boldsymbol{A})\\ &\quad = \Pr(\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U}) \le \epsilon),\end{aligned} \tag{3}$$

where $\boldsymbol{U}$ is the $n \times d$ matrix of left singular vectors of the source data matrix $\boldsymbol{A}$ (Woodruff 2014). Now as

$$\begin{aligned}\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U}) &= \max(|\lambda_{\max}(\boldsymbol{I}_d - \boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U})|,\\ &\qquad |\lambda_{\min}(\boldsymbol{I}_d - \boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U})|)\\ &= \max(|1-\lambda_{\min}(\boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U})|,\\ &\qquad |1-\lambda_{\max}(\boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U})|),\end{aligned} \tag{4}$$

the extreme eigenvalues of $\boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U}$ are the critical factor in generating ϵ-subspace embeddings. The convergence behavior of the basic iteration (2) is also tied to the eigenvalues of $\boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U}$ where $\boldsymbol{A} = \boldsymbol{X}$. Providing that $(\widetilde{\boldsymbol{X}}^\mathsf{T}\widetilde{\boldsymbol{X}})$ is of rank d, the maximum eigenvalue satisfies

$$\lambda_{\max}((\widetilde{\boldsymbol{X}}^\mathsf{T}\widetilde{\boldsymbol{X}})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{X}) = \lambda_{\max}((\boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U})^{-1}).$$

From standard results on iterative solvers (Hageman and Young 2012), a necessary and sufficient condition for the iteration to converge is $\lim_{t\to\infty}\|\boldsymbol{\beta}_F - \boldsymbol{\beta}^{(t)}\|_2 = 0$ if and only if $\lambda_{\max}((\widetilde{\boldsymbol{X}}^\mathsf{T}\widetilde{\boldsymbol{X}})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{X}) < 2$. The probability of convergence can then be expressed as

$$\Pr\left(\lim_{t\to\infty}\|\boldsymbol{\beta}_F - \boldsymbol{\beta}^{(t)}\|_2 = 0\right) = \Pr(\lambda_{\min}(\boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U}) > 0.5). \tag{5}$$

Most existing results on the probabilities (3) and (5) are finite sample lower bounds (Tropp 2011; Nelson and Nguyên 2013; Meng 2014). Worst case bounds can be conservative in practice, and there is value in developing other methods to characterize the performance of randomized algorithms (Halko et al. 2011; Raskutti and Mahoney 2014; Lopes et al. 2018; Dobriban and Liu 2018). The embedding probability (3) and the convergence probability (5) are related to the extreme eigenvalues of $\boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U}$. In Sect. 3 we study this distribution for the Gaussian sketch and develop a Tracy–Widom approximation. The approximation is then extended to the Clarkson–Woodruff and Hadamard sketches in Sect. 4.

3 Gaussian sketch

3.1 Exact representations

Meng (2014, Sect. 2.3) notes that when using a Gaussian sketch, it is instructive to consider directly the distribution of the random variable $\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U})$ to study the embedding probability (3). Consider an arbitrary $n \times d$ data matrix $\boldsymbol{A}$. As $\boldsymbol{S}$ is a matrix of independent Gaussians with mean zero and variance $1/k$, $\boldsymbol{S}\boldsymbol{U}$ is a $k \times d$ matrix of where each row has a $N(0, \boldsymbol{I}_d/k)$ distribution. It follows from the definition of a Wishart distribution that

$$\boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U} \sim \text{Wishart}\,(k, \boldsymbol{I}_d/k)\,.$$

The key term $\boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U}$ is in some sense a pivotal quantity, as its distribution is invariant to the actual values of the data matrix $\boldsymbol{A}$. When using a Gaussian sketch, the probability of obtaining an ϵ-subspace embedding has no dependence on the number of original observations n, or on the values in the data matrix $\boldsymbol{A}$. This is a useful property for a data-oblivious sketch, as it is possible to develop universal performance guarantees that will hold for any possible source dataset. This invariance property is also noted in Meng (2014), although the derivation is different.

Let us define the random matrix $\boldsymbol{W} \sim \text{Wishart}(k, \boldsymbol{I}_d/k)$. The success probability of interest can then be expressed in terms of the extreme eigenvalues of the Wishart distribution The embedding probability of interest has the representation:

$$\begin{aligned}&\Pr(\boldsymbol{S} \text{ is an } \epsilon\text{-subspace embedding for } \boldsymbol{A})\\ &\quad = \Pr\left(|1-\lambda_{\min}(\boldsymbol{W})| \le \epsilon, |1-\lambda_{\max}(\boldsymbol{W})| \le \epsilon\right).\end{aligned} \tag{6}$$

where we have made use of the expression for the maximum singular value (4).

It is difficult to obtain a mathematically tractable expression for the embedding probability as it involves the joint distribution of the extreme eigenvalues (Chiani 2017). Meng forms a lower bound on the probability (6) using concentration results on the eigenvalues of the Wishart distribution.

The convergence probability (5), can also be related to the eigenvalues of the Wishart distribution. Assuming $k \geq d$, the matrix $\widetilde{\boldsymbol{X}}^{\mathsf{T}}\widetilde{\boldsymbol{X}}$ has full rank with probability one. As such, using the same pivotal quantity $\boldsymbol{U}^{\mathsf{T}}\boldsymbol{S}^{\mathsf{T}}\boldsymbol{S}\boldsymbol{U}$ as before,

$$\Pr\left(\lim_{t\to\infty} \|\boldsymbol{\beta}_F - \boldsymbol{\beta}^{(t)}\|_2 = 0\right) = \Pr(\lambda_{\min}(\boldsymbol{W}) > 0.5), \tag{7}$$

where $\boldsymbol{W} \sim \text{Wishart}(k, \boldsymbol{I}_d/k)$. The convergence probability (7) has no dependence on the specific response vector $\boldsymbol{y}$ or design matrix $\boldsymbol{X}$ under consideration. Problem invariance is a highly desirable property for a randomized iterative solver (Roosta-Khorasani and Mahoney 2016; Lacotte et al. 2020). Both the embedding probability and the convergence probability are related to the extreme eigenvalues of the Wishart distribution. The extreme eigenvalues of Wishart random matrices are a well studied topic in random matrix theory (Edelman 1988), and we can make use of existing results to analyse the operating characteristics of sketching algorithms. In the following section we develop approximations to the embedding probability and the convergence probability in the asymptotic regime:

$$n, d, k \to \infty, \quad n \gg k, \quad d/k \to \alpha \in (0, 1]. \tag{8}$$

The regime (8) can be viewed as an interesting stress test for sketching algorithms for data compression. A key feature of the bounds in Table 1 for the embedding probability is that there is either no dependence or weak dependence on the sample size n. Working in the regime where $n \gg k$ is natural to demonstrate the effectiveness of the sketching algorithm. Allowing the number of variables d to grow with n allows for the difficulty of the compression task to increase with n. Fixing the variables to sketch size ratio d/k is important to ensure that estimates derived from the sketched dataset remain stable. The benefit in adopting this regime is the ability to obtain explicit estimates for the embedding probability that are easily computable.

3.2 Random matrix theory

Random matrix theory involves the analysis of large random matrices (Bai and Silverstein 2010). The Tracy–Widom law is an important result in the study of the extreme eigenvalue statistics (Tracy and Widom 1994). Johnstone (2001) showed that Tracy–Widom law gives the asymptotic distribution of the maximum eigenvalue of a Wishart$(k, \boldsymbol{I}_d/k)$ matrix after appropriate centering and scaling. In subsequent work Ma (2012) showed that the rate of convergence could be improved from $O(d^{-1/3})$ to $O(d^{-2/3})$ by using different centering and scaling constants than in Johnstone (2001). We build from the convergence results given by Ma.

The `R` package `RMTstat` contains a number of functions for working with the Tracy–Widom distribution (Johnstone et al. 2014). The main application of the Tracy–Widom law to statistical inference has been its use in hypothesis testing in high-dimensional statistical models (Johnstone 2006; Bai and Silverstein 2010). The Tracy–Widom law has also been demonstrated to be a universal law for the extreme eigenvalues for a wide range of random matrices beyond the Wishart (Bao et al. 2015). To the best of our knowledge, the connection to sketching algorithms has not been explored in great depth. The Tracy–Widom law can be used to approximate the embedding probability (3).

Theorem 1 *Suppose we have an arbitrary $n \times d$ data matrix $\boldsymbol{A}$ where $n > d$ and $\boldsymbol{A}$ is of rank d. Furthermore assume we take a Gaussian sketch of size k. Consider the limit in n, k and d, such that $d/k \to \alpha$ with $\alpha \in (0, 1]$. Define centering and scaling constants $\mu_{k,d}$ and $\sigma_{k,d}$ as*

$$\mu_{k,d} = k^{-1}(\sqrt{k - 1/2} + \sqrt{d - 1/2})^2,$$
$$\sigma_{k,d} = \frac{k^{-1}(\sqrt{k - 1/2} + \sqrt{d - 1/2})}{\left(1/\sqrt{k - 1/2} + 1/\sqrt{d - 1/2}\right)^{1/3}}.$$

Set $Z \sim F_1$ where F_1 is the Tracy–Widom distribution. Let $\psi_{n,k,d}$ give the exact embedding probability and let $\widehat{\psi}_{n,k,d}$ give the asymptotic approximation to the embedding probability:

$$\psi_{n,k,d} = \Pr\left(\boldsymbol{S} \text{ is an } \epsilon\text{-subspace embedding for } \boldsymbol{A}\right),$$
$$\widehat{\psi}_{n,k,d} = \Pr\left(Z \leq \frac{\epsilon + 1 - \mu_{k,d}}{\sigma_{k,d}}\right).$$

Then asymptotically in n, d and k, for any $\epsilon > 0$,

$$\lim_{n,d,k\to\infty} \left|\psi_{n,k,d} - \widehat{\psi}_{n,k,d}\right| = 0$$

Furthermore, for even d, $\left|\psi_{n,k,d} - \widehat{\psi}_{n,k,d}\right| = O(d^{-2/3})$.

The proof is given in the supplementary material.

The convergence probability of the iterative algorithm (5) can also be approximated using the Tracy–Widom law.

Theorem 2 *Suppose we have an arbitrary $n \times d$ data matrix $\boldsymbol{A}$ where $n > d$ and $\boldsymbol{A}$ is of rank d. Furthermore, assume we take a Gaussian sketch of size k. Consider the limit in n, k and d, such that $d/k \to \alpha$ with $\alpha \in (0, 1]$. Set*

$$\mu_{k,d} = (\sqrt{k - 1/2} - \sqrt{d - 1/2})^2,$$

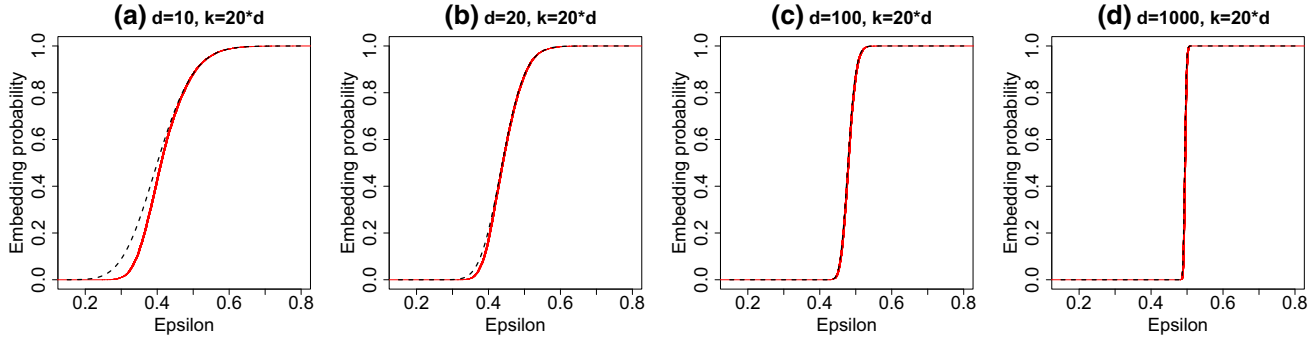

Fig. 1 Accuracy of Tracy–Widom approximation for embedding probability (6) for the Gaussian sketch. The dashed black line gives the asymptotic limit, the solid red line gives the empirical probability. When $d \geq 20$ the approximation given in Theorem 1 is very accurate. (Color figure online)

$$\sigma_{k,d} = (\sqrt{k-1/2} - \sqrt{d-1/2}) \left(\frac{1}{\sqrt{k-1/2}} - \frac{1}{\sqrt{d-1/2}} \right)^{1/3},$$

and define the following centering and scaling constants $\tau_{k,d} = \sigma_{k,d}/\mu_{k,d}$, $\nu_{k,d} = \log(\mu_{k,d}) - \log k - \tau_{k,d}^2/8$. *Set* $Z \sim F_1$, *where* F_1 *is the Tracy–Widom distribution. Let* $\gamma_{n,k,d}$ *give the exact convergence probability, and* $\widehat{\gamma}_{n,k,d}$ *give the asymptotic approximation to the convergence probability:*

$$\gamma_{n,k,d} = \Pr\left(\lim_{t\to\infty} \|\boldsymbol{\beta}_F - \boldsymbol{\beta}^{(t)}\|_2 = 0 \right),$$
$$\widehat{\gamma}_{n,k,d} = \Pr\left(Z \leq \frac{\nu_{k,d} - \log(1/2)}{\tau_{k,d}} \right).$$

Then for all starting values $\boldsymbol{\beta}^{(0)}$, *asymptotically in* n, d *and* k,

$$\lim_{n,d,k\to\infty} \left| \gamma_{n,k,d} - \widehat{\gamma}_{n,k,d} \right| = 0.$$

Furthermore, for even d, $\left| \gamma_{n,k,d} - \widehat{\gamma}_{n,k,d} \right| = O(d^{-2/3})$.

The proof is given in the supplementary material.

The embedding probability for the Gaussian sketch can be estimated by simulating $\boldsymbol{W} \sim \text{Wishart}(k, \boldsymbol{I}_d/k)$ and using the empirical distribution of the random variable $\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{W})$. To assess the accuracy of the approximation in Theorem 1, we generated $B = 10,000$ random Wishart matrices $\boldsymbol{W}^{[1]}, \ldots, \boldsymbol{W}^{[B]}$. For each simulated matrix $\boldsymbol{W}^{[b]}$ we computed the distortion factor $\epsilon^{[b]} = \sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{W}^{[b]})$ for $b = 1, \ldots, B$. The simulated distortion factors $\epsilon^{[1]}, \ldots, \epsilon^{[B]}$ were used to give a Monte Carlo estimate of the embedding probability:

$$\widehat{\Pr}(\boldsymbol{S} \text{ is an } \epsilon\text{-subspace embedding for } \boldsymbol{A}) = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}(\epsilon^{[b]} \leq \epsilon). \tag{9}$$

We used the `ARPACK` library (Lehoucq et al. 1998) to compute the maximum singular values $\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{W}^{[b]})$. The estimated embedding probabilities are displayed in Fig. 1 for different dimensions d. The sketch size to variables ratio, k/d, was held fixed at 20. The solid red line shows the empirical probability of obtaining an ϵ-subspace embedding. The dashed black line gives the Tracy–Widom approximation given in Theorem 1. The agreement is consistently good over dimensions d, and the range of sketch sizes k that were considered.

4 Computationally efficient sketches

4.1 Asymptotics for the extreme eigenvalues

Asymptotic methods are useful to analyse data-oblivious sketches that do not admit interpretable finite sample distributions (Li et al. 2006; Ahfock et al. 2020; Lacotte et al. 2020). Here we describe the limiting behavior of the sketched algorithms for fixed k and d as the number of source observations n increases.

Under an assumption on the limiting leverage scores of the source data matrix, we can establish a limit theorem for the Hadamard and Clarkson–Woodruff sketches.

Assumption 1 Define the singular value decomposition of the $n \times d$ source dataset as $\boldsymbol{A}_{(n)} = \boldsymbol{U}_{(n)} \boldsymbol{D}_{(n)} \boldsymbol{V}_{(n)}^{\mathsf{T}}$. Let $\boldsymbol{u}_{(n)i}^{\mathsf{T}}$ give the ith row in $\boldsymbol{U}_{(n)}$. Assume that the maximum leverage score tends to zero, that is

$$\lim_{n\to\infty} \max_{i=1,\ldots,n} \|\boldsymbol{u}_{(n)i}\|_2^2 = 0.$$

Assuming that $\boldsymbol{A}_{(n)}$ is of rank d, the leverage scores have an important standardization property in that

$$\sum_{i=1}^{n} \|\boldsymbol{u}_{(n)i}\|_2^2 = d. \tag{10}$$

Assumption 1 represents an asymptotic negligibility condition on the significance of any single observation. The

same assumption is made in the analysis of high-dimensional regression models in Huber (1973, Proposition 2.2). Similar to the Lindeberg-Feller condition (Van Der Vaart 1998), Assumption 1 requires that the contribution of any single observation to the total variance in the source dataset (10) is arbitrarily small for sufficiently large values of n. Assumption 1 is expected to hold if there are no extreme outliers in the source dataset.

The asymptotic probability of obtaining an ϵ-subspace embedding for the Hadamard and Clarkson–Woodruff sketches can be related to the Wishart distribution.

Theorem 3 *Consider a sequence of arbitrary $n \times d$ data matrices $\boldsymbol{A}_{(n)}$, where each data matrix is of rank d, and d is fixed. Let $\boldsymbol{A}_{(n)} = \boldsymbol{U}_{(n)}\boldsymbol{D}_{(n)}\boldsymbol{V}_{(n)}^{\mathsf{T}}$ represent the singular value decomposition of $\boldsymbol{A}_{(n)}$. Let $\boldsymbol{S}_{(n)}$ be a $k \times n$ Hadamard or Clarkson–Woodruff sketching matrix where k is also fixed. Suppose that Assumption 1 is satisfied. Then as n tends to infinity with k and d fixed,*

$$\lim_{n\to\infty} \Pr\left(\boldsymbol{S}_{(n)} \text{ is an } \epsilon\text{-subspace embedding for } \boldsymbol{A}_{(n)}\right) = \Pr\left(\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{W}) \leq \epsilon\right),$$

where $\boldsymbol{W} \sim \text{Wishart}(k, \boldsymbol{I}_d/k)$.

The proof is given in the supplementary material.

Theorem 3 states the the embedding probability for the Hadamard and Clarkson–Woodruff sketches converges to that of the Gaussian sketch as $n \to \infty$. Therefore, Theorem 1 can also be used to approximate the embedding probability. Empirical studies have shown that the Hadamard and Clarkson–Woodruff sketches can give similar quality results to the Gaussian projection (Venkatasubramanian and Wang 2011; Le et al. 2013; Dahiya et al. 2018). Theorem 3 helps to characterize situations where this phenomenon is expected to be observed.

Remark 1 The same line of proof used in Theorem 3 can be used to show that the convergence probability of (2) using the Hadamard and Clarkson–Woodruff projections converges to that of the Gaussian sketch under Assumption 1. Theorem 2 also gives an asymptotic approximation for the Hadamard and Clarkson–Woodruff sketches.

It remains to establish a formal limit theorem in terms of the Tracy–Widom distribution for the Hadamard and Clarkson–Woodruff sketches. The proof of Theorem 3 treats k and d as fixed, with only n being taken to infinity. It is possible that Assumption 1 on the leverage scores will remain sufficient in the expanding dimension scenario. For any d, the maximum leverage score must be greater than the average leverage score,

$$\max_{i=1,\ldots,n} \|\boldsymbol{u}_{(n)i}\|_2^2 \geq \frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{u}_{(n)i}\|_2^2 = \frac{d}{n}.$$

If we maintain that Assumption 1 holds on the leverage scores as $n, d, k \to \infty$, this implies that $d/n \to 0$. As we have assumed that our primary motivation for sketching is data compression when $n \gg d$, we feel that analysis in the asymptotic regime $d/n \to 0$ is reasonable for this use-case setting. The asymptotic approximations developed here are recommended for applications of sketching in tall-data problems where $n \gg d$.

The key result is that the Hadamard and Clarkson–Woodruff sketches behave like the Gaussian projection for large n, with k and d fixed. If the Tracy–Widom approximation in Theorem 1 is good for finite k and d with the Gaussian sketch, then it should hold well for the Hadamard and Clarkson–Woodruff projections for n sufficiently large.

4.2 Uniform sketch

It is considerably more difficult to approximate the embedding probability for the uniform sketch compared to the other data-oblivious projections. Vershynin (2010) provides a bound for the uniform sketch that is useful for comparative purposes.

Theorem 4 (Vershynin (2010), Theorem 5.41) *Consider an $n \times d$ matrix $\boldsymbol{U}$ such that $\boldsymbol{U}^{\mathsf{T}}\boldsymbol{U} = \boldsymbol{I}_d$. Let $\boldsymbol{u}_i^{\mathsf{T}}$ represent the i-th row in $\boldsymbol{U}$ for $i = 1, \ldots, n$. Let r give an upper bound on the leverage scores, so*

$$\max_{i=1,\ldots,n} \|\boldsymbol{u}_i\|_2^2 \leq r.$$

Let $\boldsymbol{S}$ be a uniform sketch of size k. Then for every $t \geq 0$, with probability at least $1 - 2d\exp(-ct^2)$ one has

$$1 - t\sqrt{\frac{rn}{k}} \leq \sigma_{\min}(\boldsymbol{SU}) \leq \sigma_{\max}(\boldsymbol{SU}) \leq 1 + t\sqrt{\frac{rn}{k}},$$

where $c > 0$ is an absolute constant.

Theorem 4 is a minor reformulation of the result presented in Vershynin (2010), this is elaborated on in the supplementary material.

Theorem 4 can be used to give a lower bound on the probability of obtaining an ϵ-subspace embedding. Both Theorems 4 and 3 involve the maximum leverage score. Holding k and d fixed, in order for the bound in Theorem 4 to remain controlled as the sample size n increases, the maximum leverage score r must decrease at a sufficient rate. In contrast, Assumption 1 does not enforce a rate of decay on the maximum leverage score, only that it eventually tends to zero as $n \to \infty$. This suggests that the uniform projection could be more sensitive to the maximum leverage score than the Gaussian, Hadamard and Clarkson–Woodruff projections.

4.3 Asymptotics for the empirical spectral distribution

An alternative approach to estimate the embedding probability is to use the limiting empirical spectral distribution of $\boldsymbol{M}_d = \boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U}$. For a random Hermitian matrix $\boldsymbol{M}_d$ of size $d \times d$, the empirical spectral distribution of $\boldsymbol{M}_d$ is the cumulative distribution function of its eigenvalues $\lambda_1 \leq \cdots \leq \lambda_d$, i.e, $F_{M_d}(x) := \frac{1}{d}\sum_{j=1}^{d} \mathbb{1}(\lambda_j \leq x)$ for $x \in \mathbb{R}$.

Lacotte et al. (2020) derive the limiting empirical spectral distribution of $\boldsymbol{M}_d = \boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U}$ for the Hadamard sketch in the asymptotic regime where $\lim_{n\to\infty} d/n = \gamma \in (0, 1)$, $\lim_{n\to\infty} k/n = \xi \in (\gamma, 1)$ and $\lim_{n\to\infty} d/k = \alpha$. The extreme eigenvalues of $\boldsymbol{M}_d = \boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U}$ under the Hadamard sketch converge pointwise to (Lacotte and Pilanci 2020)

$$\lambda_{\min}(\boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U}) = (\sqrt{1-\gamma} - \sqrt{(1-\xi)\alpha})^2. \tag{11}$$

$$\lambda_{\max}(\boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U}) = (\sqrt{1-\gamma} + \sqrt{(1-\xi)\alpha})^2. \tag{12}$$

The results (11) and (12) imply the convergence result for the maximum singular value

$$\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{U}^\mathsf{T}\boldsymbol{S}^\mathsf{T}\boldsymbol{S}\boldsymbol{U}) = \max(|1 - (\sqrt{1-\gamma} - \sqrt{(1-\xi)\alpha})^2|, \\ |1 - (\sqrt{1-\gamma} + \sqrt{(1-\xi)\alpha})^2|).$$

The resulting approximation to the embedding probability is then

$$\Pr(\boldsymbol{S} \text{ is an } \epsilon\text{-subspace embedding}) = \begin{cases} 1 \text{ if } \epsilon \geq \sigma^* \\ 0 \text{ if } \epsilon < \sigma^*, \end{cases} \tag{13}$$

where $\sigma^* = \max(|1 - (\sqrt{1-\gamma} - \sqrt{(1-\xi)\alpha})^2|, |1 - (\sqrt{1-\gamma} + \sqrt{(1-\xi)\alpha})^2|)$.

We estimated the embedding probability using the Hadamard sketch on simulated data with $d = 50, k = 1000$ and $n \in \{5000, 10000, 50000, 100000\}$ over 1000 sketches. Each row in the source dataset was an independent draw from a $N(0, \Sigma)$ distribution where $\Sigma_{ij} = \rho^{|i-j|}$ and $\rho = 0.5$. Lacotte et al. (2020) consider a slight variant of the Hadamard sketch where the subsampling matrix Φ is constructed using subsampling without replacement. In the simulation, the Hadamard sketch was implemented using subsampling without replacement. Figure 2 compares the empirical embedding probability (solid red line) to the to the Tracy–Widom approximation in Theorem 1 (black dashed line) and the empirical spectral distribution approximation in (13) (step function).

When d/n is large, the approximation to the embedding probability (13) suggests that the Hadamard sketch will perform better than is predicted by the Tracy–Widom law. In panel (a) of Figure 2 where $d = 50, n = 5000$, the empirical embedding probability for the Hadamard sketch is shifted to the left compared to the Tracy–Widom limit, which indicates superior performance. As n increases and $d/n \to 0$, the Tracy–Widom approximation becomes more accurate as predicted by Theorem 3. In panel (d) of Fig. 2 where $d = 50, n = 100,000$ there is close agreement between the empirical embedding probability and the Tracy–Widom limit.

5 Data application

5.1 ϵ-subspace embedding

We tested the theory on a large genetic dataset of European ancestry participants in UK Biobank. The covariate data consists of genotypes at $p = 1032$ genetic variants in the Protein Kinase C Epsilon (PKCε) gene on $n = 407,779$ subjects. Variants were filtered to have minor allele frequency of greater than one percent. The response variable was haemoglobin concentration adjusted for age, sex and technical covariates. The region was chosen as many associations with haemoglobin concentration were discovered in a genome-wide scan using univariable models; these associations were with variants with different allele frequencies, suggesting multiple distinct causal variants in the region. We also considered a subset of this dataset with $p = 130$ representative markers identified by hierarchical clustering. When including the intercept and response, the PKCε subset has $n = 407,779, d = 132$, and the full PKCε dataset has $n = 407,779, d = 1034$.

The full PKCε dataset is of moderate size, so it was feasible to take the singular value decomposition of the full $n \times d$ dataset $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\mathsf{T}$. Given the singular value decomposition we ran an oracle procedure to estimate the exact embedding probability. We generated B sketching matrices $\boldsymbol{S}^{[1]}, \ldots, \boldsymbol{S}^{[B]}$. These were used to compute $\epsilon^{[b]} = \sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{U}^\mathsf{T}\boldsymbol{S}^{[b]\mathsf{T}}\boldsymbol{S}^{[b]}\boldsymbol{U})$ for $b = 1, \ldots, B$ and give an estimated embedding probability as in (9). When working with the full PKCε dataset we simulated directly from the matrix normal distribution $\widetilde{\boldsymbol{U}} \sim \text{MN}(\boldsymbol{I}_k, \boldsymbol{I}_d/k)$ for the Gaussian sketch, rather than computing the matrix multiplication $\boldsymbol{S}\boldsymbol{U}$. We took $B = 1000$ sketches of the PKCε subset, and $B = 100$ sketches of the full PKCε dataset using the uniform, Gaussian, Hadamard and Clarkson–Woodruff projections, with $k = 20 \times d$.

Figure 3 shows the empirical and theoretical embedding probabilities for the PKCε subset ($n = 407,779, d = 132$) for each type of sketch. The observed and theoretical curves match well for the Gaussian, Hadamard and Clarkson–Woodruff projection. The uniform projection performs worse than the other data-oblivious random projections, as larger values of ϵ indicate weaker approximation bounds. The uniform projection does not satisfy a central limit theorem for

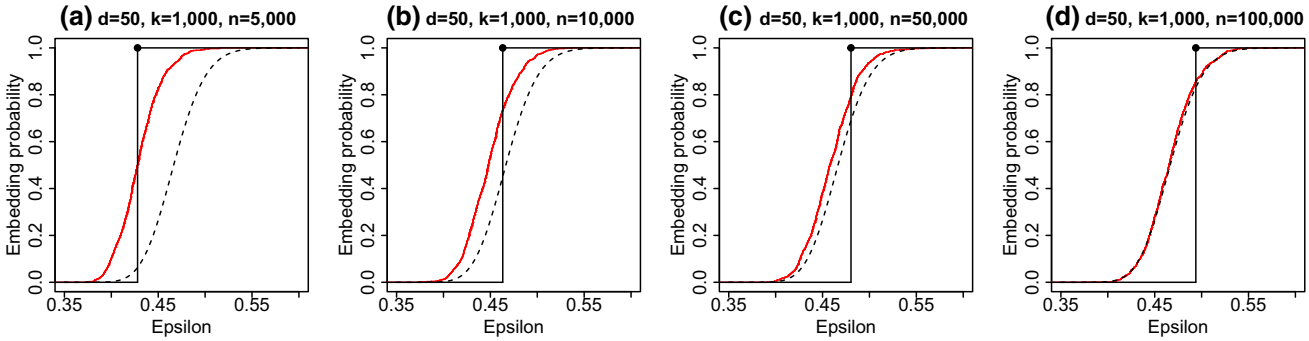

Fig. 2 Embedding probability for the Hadamard sketch. The solid red line gives the empirical probability. The dashed black line gives the Tracy–Widom approximation to the embedding probability. The step function represents the approximation to the embedding probability using the limiting empirical spectral distribution. (Color figure online) (13)

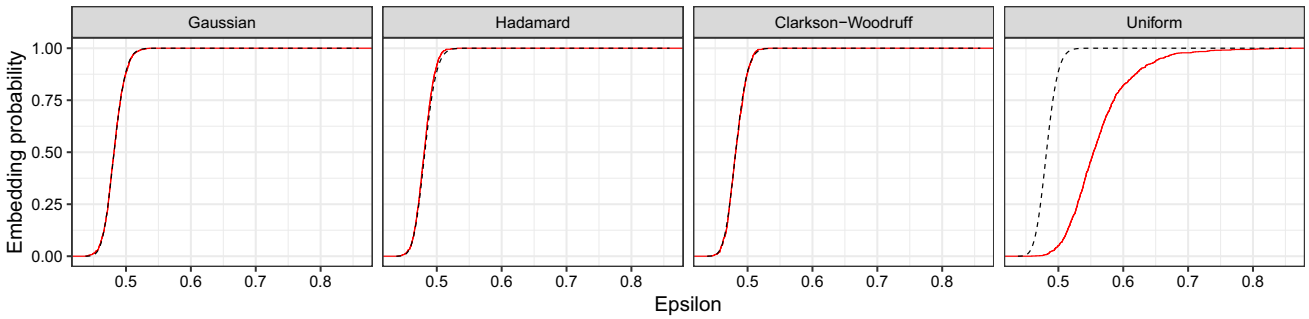

Fig. 3 Analysis of subset of PKCε dataset ($n = 407,779, d = 132$) with $B = 1000$ sketches of size $k = 20d$. The dashed black line and the solid red line gives the theoretical and empirical embedding probabilities respectively. The Tracy–Widom approximation is accurate for the Gaussian, Hadamard and Clarkson–Woodruff sketches. (Color figure online)

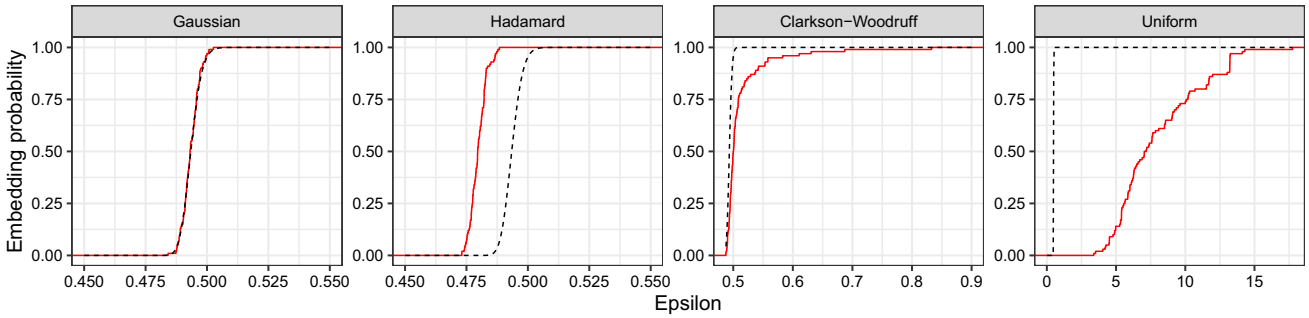

Fig. 4 Analysis of full PKCε dataset ($n = 407,779, d = 1,034$) with $B = 100$ sketches of size $k = 20d$. The x-axis is different in each panel.The dashed black line and the solid red line gives the theoretical and empirical embedding probabilities respectively. The Uniform projection is much less successful at generating ϵ-subspace embeddings than the other data-oblivious projections. (Color figure online)

fixed k, so we do not necessarily expect the Tracy–Widom law to give a good approximation for the uniform projection.

Figure 4 shows the empirical and theoretical embedding probabilities for the full PKCε dataset ($n = 407,779, d = 1032$) for each type of sketch. The Tracy–Widom approximation is accurate for the Gaussian sketch, but there are some deviations for the Hadamard and the Clarkson–Woodruff sketch. The empirical cdf for the Hadamard sketch (red) is to the left of the theoretical value (black), indicating smaller values of ϵ than predicted. This phenomenon is to be expected given the results on the extreme eigenvalues for the Hadamard sketch developed in Lacotte and Pilanci (2020). The distribution of ϵ has a longer right tail under the Clarkson–Woodruff sketch than is predicted by the Tracy–Widom law.

The deviation from the Tracy–Widom limit in Fig. 4 could be because the finite sample approximation is poor. Theorem 3 suggests that the Hadamard and Clarkson–Woodruff

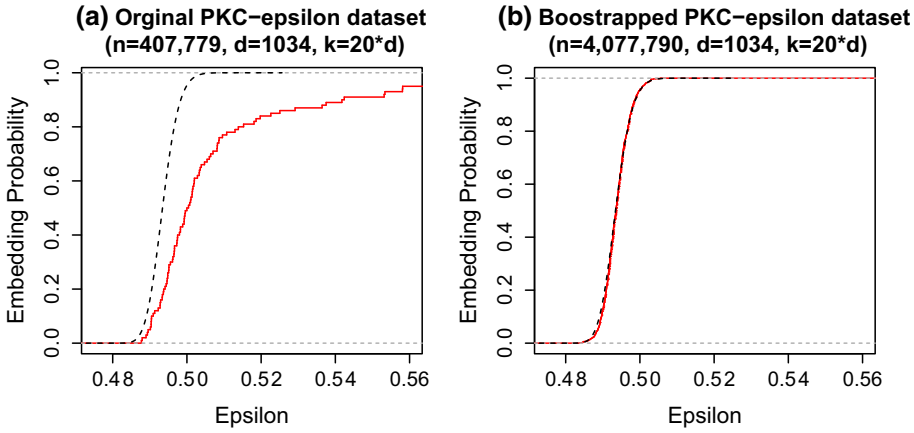

Fig. 5 Comparison of results on the original PKCε dataset ($n = 407,779$) and the bootstrapped larger PKCϵ dataset ($n = 4,077,790$). The dashed black line and the solid red line gives the theoretical and empirical embedding probabilities respectively. As expected from Theorem 3, the accuracy of the Tracy–Widom increases with n. (Color figure online)

Table 2 Mean sketching time (seconds) over ten sketches for each dataset

Projection	Subset ($p = 132$)	Full ($p = 1034$)
Gaussian	769	–
Hadamard	17.2	156
Clarkson–Woodruff	1.33	21
Uniform	0.03	2.8

The Gaussian sketch is considerably slower than the Hadamard and Clarkson–Woodruff sketches on the subset as is expected from Table 1

projections behave like the Gaussian sketch for n sufficiently large with respect to d. To test this we bootstrapped the full PKCε dataset to be ten times its original size. The bootstrapped PKCε dataset has $n = 4,077,790, d = 1034$. We took one thousand sketches of size $k = 20 \times d$ using the Clarkson–Woodruff projection and ran the oracle procedure of computing $\epsilon^{[b]} = \sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{U}^\top \boldsymbol{S}^{[b]\top} \boldsymbol{S}^{[b]} \boldsymbol{U})$ for each sketch. Figure 5 compares the distribution of $\sigma_{\max}(\boldsymbol{I}_d - \boldsymbol{U}^\top \boldsymbol{S}^\top \boldsymbol{S} \boldsymbol{U})$ using Clarkson–Woodruff projection on the original dataset and on the large bootstrapped dataset. As n increases we expect the quality of the Tracy–Widom approximation to improve. Panel (a) of Fig. 5 compares the theoretical to the simulation results on the original dataset. The Clarkson–Woodruff projection shows greater variance than expected. Panel (b) compares the theoretical to the simulation results on the bootstrapped dataset. In (b) there is very good agreement between the empirical distribution and the theoretical distribution. It seems that for this dataset $n \approx 400,000$ is not big enough for the large sample asymptotics to kick in. At $n \approx 4$ million the Tracy–Widom approximation is very good. As mentioned earlier, our motivation for using a sketching algorithm to perform data compression with tall datasets $n \gg d$. This example highlights that the asymptotic approximations become more accurate as the sample size n grows and the computational incentives for using sketching increase in parallel (Table 2).

5.2 Iterative optimisation

We considered iterative least-squares optimisation using the song year dataset available from the UCI machine learning repository. The dataset has $n = 515,344$ observations, $p = 90$ covariates, and year of song release as the response. We assessed the convergence probability by running the iteration (2) with the sketched preconditioner. The initial parameter estimate $\boldsymbol{\beta}^{(0)}$ was a vector of zeros. The iteration was run for 2000 steps, with convergence being declared if the gradient norm condition $\|\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{(t)})\|_2 < 10^{-6}$ was satisfied at any time step t. This convergence criterion was used instead of $\|\boldsymbol{\beta}_F - \boldsymbol{\beta}^{(t)}\|_2$ as $\boldsymbol{\beta}_F$ will not be known in practice. This was repeated one hundred times for each of the random projections discussed in Sect. 2.1 using different sketch sizes k. Figure 6 compares the empirical (solid red points) and theoretical convergence probabilities (dashed black line) against the sketch size k. The point-ranges represent 95% confidence intervals. The Gaussian, Hadamard and Clarkson–Woodruff show near identical behavior, and the empirical convergence probabilities closely match the theoretical predictions using Theorem 2. The uniform sketch was much less successful in generating preconditioners, the algorithm did not show convergence in any replication at each sketch size k. In this example, the additional computational cost of the Gaussian, Hadamard and Clarkson–Woodruff sketches compared to the Uniform subsampling has clear benefits.

6 Conclusion

The analysis of the asymptotic behavior of common data-oblivious random projections revealed an important connection to the Tracy–Widom law. The probability of attaining an ϵ-subspace embedding (Definition 1) is an integral descriptive measure for many sketching algorithms. The asymptotic embedding probability can approximated using the Tracy–Widom law for the Gaussian, Hadamard and Clarkson–

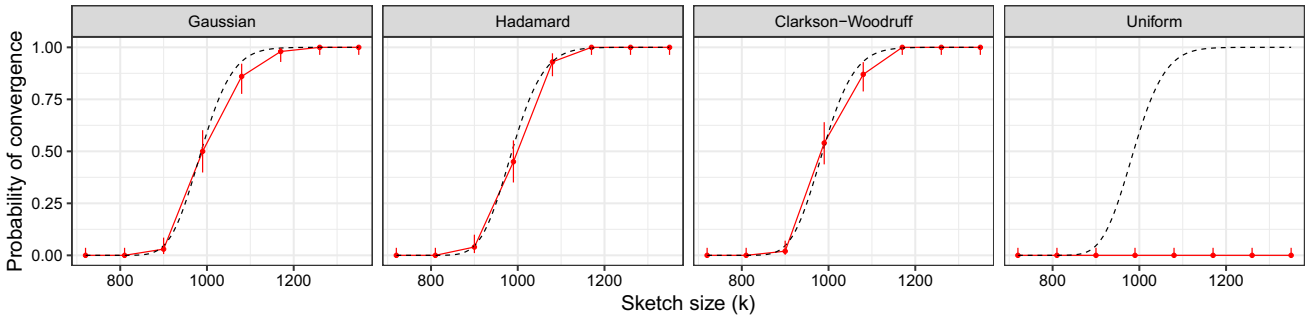

Fig. 6 Convergence probability on year dataset ($n = 515,344, d = 91$). Red solid points show the empirical convergence probability over $B = 100$ sketches. The black dashed line gives the theoretical convergence probability using Theorem 2. The Tracy–Widom approximation is accurate for the Gaussian, Hadamard and Clarkson–Woodruff sketches. The uniform sketch fails to generate useful preconditioners. (Color figure online)

Woodruff sketches. The Tracy–Widom law can also be used to estimate the convergence probability for iterative schemes with a sketched preconditioner. We have tested the predictions empirically and seen close agreement. The majority of existing results for sketching algorithms have been established using non-asymptotic tools. Asymptotic results are a useful complement that can provide answers to important questions that are difficult to address concretely in a finite dimensional framework.

There was a stark contrast between the performance of the basic uniform projection and the other data-oblivious projections (Gaussian, Hadamard and Clarkson–Woddruff) in the data application. The Hadmard and Clarkson–Woodruff projections are expected to behave like the Gaussian projection under mild regularity conditions on the maximum leverage score. We observed this phenomenon when n/d was large, as is required by Theorem 3. The Hadamard and Clarkson–Woodruff projections are substantially more computationally efficient than the Gaussian projection (recall Table 1), so their universal limiting behavior implies that the trade-off between computation time and performance guarantees is asymptotically negligible in the regime (8).

The Tracy–Widom law has found many applications in high-dimensional statistics and probability (Edelman and Wang 2013), and we have shown that it useful for describing the asymptotic behavior of sketching algorithms. The asymptotic behaviour with respect to large n is of practical interest, as this is the regime where sketching is attractive as a data compression technique. The universal behavior of high-dimensional random matrices has practical and theoretical consequences for randomized algorithms that use linear dimension reduction (Dobriban and Liu 2018; Lacotte et al. 2020).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11222-022-10148-5.

Funding This work has been conducted using the UK Biobank resource under applications number 13745. SR was supported by funding from the UKRI Medical Research Council (MC_UU_00002/10) and the Alan Turing Institute (TU/B/000092). WJA was supported by NHS Blood and Transplant and the NIHR BTRU in Donor Health and Genomics NIHR (BTRU-2014-10024).

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

References

Ahfock, D.C., Astle, W.J., Richardson, S.: Statistical properties of sketching algorithms. Biometrika **108**(2), 283–297 (2020)

Ailon, N., Chazelle, B.: The fast Johnson Lindenstrauss transform and approximate nearest neighbors. SIAM J. Comput. **39**(1), 302–322 (2009)

Bai, Z., Silverstein, J.W.: Spectral Analysis of Large Dimensional Random Matrices, 2nd edn. Springer, New York (2010)

Bao, Z., Pan, G., Zhou, W.: Universality for the largest eigenvalue of sample covariance matrices with general population. Ann. Stat. **43**(1), 382–421 (2015)

Bardenet, R., Maillard, O.A.: A note on replacing uniform subsampling by random projections in MCMC for linear regression of tall datasets. HAL preprint 01248841 (2015)

Chiani, M.: On the probability that all eigenvalues of Gaussian, Wishart, and double Wishart random matrices lie within an interval. IEEE Trans. Inf. Theory **63**(7), 4521–4531 (2017)

Clarkson, K.L., Woodruff, D.P.: Low rank approximation and regression in input sparsity time. In: Proceedings of the Forty-Fifth Annual

ACM Symposium on Theory of Computing, pp. 81–90. ACM (2013)
Cormode, G.: Sketch techniques for approximate query processing. Found. Trends Databases (2011)
Dahiya, Y., Konomis, D., Woodruff, D.P.: An empirical evaluation of sketching for numerical linear algebra. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1292–1300. ACM (2018)
Derezinski, M., Liao, Z., Dobriban, E., Mahoney, M.: Sparse sketches with small inversion bias. In: Conference on Learning Theory, pp. 1467–1510. PMLR (2021)
Dobriban, E., Liu, S.: A new theory for sketching in linear regression. arXiv preprint arXiv:1810.06089 (2018)
Drineas, P., Mahoney, M.W., Muthukrishnan, S.: Sampling algorithms for l2 regression and applications. In: Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 1127–1136. Society for Industrial and Applied Mathematics (2006)
Edelman, A.: Eigenvalues and condition numbers of random matrices. SIAM J. Matrix Anal. Appl. **9**(4), 543–560 (1988)
Edelman, A., Wang, Y.: Random matrix theory and its innovative applications. In: Advances in Applied Mathematics, Modeling, and Computational Science, pp. 91–116. Springer, Cham (2013)
Erichson, N.B., Voronin, S., Brunton, S.L., Kutz, J.N.: Randomized Matrix Decompositions using R. arXiv preprint p. arXiv:1608.02148 (2016)
Falcone, R., Anderlucci, L., Montanari, A.: Matrix sketching for supervised classification with imbalanced classes. Data Min. Knowl. Discov. 1–35 (2021)
Geppert, L.N., Ickstadt, K., Munteanu, A., Quedenfeld, J., Sohler, C.: Random projections for Bayesian regression. Stat. Comput. **27**(1), 79–101 (2017)
Grellmann, C., Neumann, J., Bitzer, S., Kovacs, P., Tönjes, A., Westlye, L.T., Andreassen, O.A., Stumvoll, M., Villringer, A., Horstmann, A.: Random projection for fast and efficient multivariate correlation analysis of high-dimensional data: a new approach. Front. Genet. **7**, 102 (2016)
Hageman, L., Young, D.: Applied Iterative Methods. Dover, Illinois (2012)
Halko, N., Martinsson, P.G., Tropp, J.A.: Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. SIAM Rev. **53**(2), 217–288 (2011)
Huber, P.J.: Robust regression: asymptotics, conjectures and Monte Carlo. Ann. Stat. **1**(5), 799–821 (1973)
Johnstone, I.M.: On the distribution of the largest eigenvalue in principal components analysis. Ann. Stat. **29**(2), 295–327 (2001)
Johnstone, I.M .: High dimensional statistical inference and random matrices. arXiv preprint arXiv:math/0611589 (2006)
Johnstone, I.M., Ma, Z., Perry, P.O., Shahram, M.: RMTstat: distributions, statistics and tests derived from random matrix theory (2014). R package version 0.3
Lacotte, J., Liu, S., Dobriban, E., Pilanci, M.: Limiting spectrum of randomized Hadamard transform and optimal iterative sketching methods. arXiv preprint arXiv:2002.00864 (2020)
Lacotte, J., Pilanci, M.: Optimal randomized first-order methods for least-squares problems. In: Proceedings of the 37th International Conference on Machine Learning. JMLR (2020)
Le, Q., Sarlós, T., Smola, A.: Fastfood-computing hilbert space expansions in loglinear time. In: International Conference on Machine Learning, pp. 244–252. (2013)
Lehoucq, R.B., Sorensen, D.C., Yang, C.: ARPACK users' guide: solution of large-scale eigenvalue problems with implicitly restarted Arnoldi methods, vol. 6. In: SIAM (1998)
Li, P., Hastie, T.J., Church, K.W.: Very sparse random projections. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 287–296. ACM (2006)
Lopes, M.E., Wang, S., Mahoney, M.W.: Error estimation for randomized least-squares algorithms via the bootstrap. arXiv preprint arXiv:1803.08021 (2018)
Ma, P., Mahoney, M.W., Yu, B.: A statistical perspective on algorithmic leveraging. J. Mach. Learn. Res. **16**(1), 861–911 (2015)
Ma, Z.: Accuracy of the Tracy-Widom limits for the extreme eigenvalues in white Wishart matrices. Bernoulli **18**(1), 322–359 (2012)
Mahoney, M.: Randomized algorithms for matrices and data. Found. Trends Mach. Learn. **3**(2), 123–224 (2011)
Mahoney, M., Drineas, P.: Structural properties underlying high-quality Randomized Numerical Linear Algebra algorithms. In: Buhlmann, P., Drineas, P., Kane, M., van de Laan, M. (eds.) Handbook of Big Data, pp. 137–154. Chapman and Hall, London (2016)
Meng, X.: Randomized Algorithms for Large-scale Strongly Overdetermined Linear Regression Problems. Ph.D. thesis, Stanford University, Stanford, California, United States (2014)
Meng, X., Mahoney, M.M.: Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. In: Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing, pp. 91–100. ACM (2013)
Meng, X., Saunders, M.A., Mahoney, M.W.: LSRN: a parallel iterative solver for strongly over-or underdetermined systems. SIAM J. Sci. Comput. 36(2), C95–C118 (2014)
Nelson, J., NguyêNelson, J., Nguyên, H.L.: OSNAP: Faster numerical linear algebra algorithms via sparser subspace embeddings. In: 54th Annual IEEE Symposium on the Foundations of Computer Science, pp. 117–126. IEEE (2013)
Pilanci, M., Wainwright, M.J.: Iterative Hessian sketch: fast and accurate solution approximation for constrained least-squares. J. Mach. Learn. Res. **17**(1), 1842–1879 (2016)
Quiroz, M., Villani, M., Kohn, R., Tran, M.N., Dang, K.D.: Subsampling MCMC-an introduction for the survey statistician. Sankhya A **80**(1), 33–69 (2018)
Raskutti, G., Mahoney, M.: A statistical perspective on randomized sketching for ordinary least-squares. arXiv preprint arXiv:1406.5986 (2014)
Roosta-Khorasani, F., Mahoney, M.W.: Sub-sampled Newton methods I: globally convergent algorithms. arXiv preprint arXiv:1601.04737 (2016)
Sarlos, T.: Improved approximation algorithms for large matrices via random projections. In: 47th Annual IEEE Symposium on Foundations of Computer Science, pp. 143–152. IEEE (2006)
Tracy, C.A., Widom, H.: Level-spacing distributions and the airy kernel. Commun. Math. Phys. **159**(1), 151–174 (1994)
Tropp, J.A.: Improved analysis of the subsampled randomized Hadamard transform. Adv. Adapt. Data Anal. **3**, 115–126 (2011)
Van Der Vaart, A.: Asymptotic Statistics. Cambridge University Press, Cambridge (1998)
Venkatasubramanian, S., Wang, Q.: The Johnson-Lindenstrauss transform: an empirical study. In: 2011 Proceedings of the Thirteenth Workshop on Algorithm Engineering and Experiments, pp. 164–173. SIAM (2011)
Vershynin, R.: Introduction to the non-asymptotic analysis of random matrices. arXiv preprint arXiv:1011.3027 (2010)
Woodruff, D.P.: Sketching as a tool for numerical linear algebra. Found. Trends Theor. Comput. Sci. **10**(1–2), 1–157 (2014)
Yang, J., Meng, X., Mahoney, M.W.: Implementing randomized matrix algorithms in parallel and distributed environments. arXiv preprint arXiv:1502.03032 (2015)