

# Hashing embeddings of optimal dimension, with applications to linear least squares

Coralia Cartis<sup>\*,†</sup>

Jan Fiala<sup>\*,‡</sup>

Zhen Shao<sup>\*,§</sup>

May 24, 2021

## Abstract

The aim of this paper is two-fold: firstly, to present subspace embedding properties for  $s$ -hashing sketching matrices, with  $s \geq 1$ , that are optimal in the projection dimension  $m$  of the sketch, namely,  $m = \mathcal{O}(d)$ , where  $d$  is the dimension of the subspace. A diverse set of results are presented that address the case when the input matrix has sufficiently low coherence (thus removing the  $\log^2 d$  factor dependence in  $m$ , in the low-coherence result of Bourgain et al (2015) at the expense of a smaller coherence requirement); how this coherence changes with the number  $s$  of column nonzeros (allowing a scaling of  $\sqrt{s}$  of the coherence bound), or is reduced through suitable transformations (when considering hashed-instead of subsampled- coherence reducing transformations such as randomised Hadamard). Secondly, we apply these general hashing sketching results to the special case of Linear Least Squares (LLS), and develop Ski-LLS, a generic software package for these problems, that builds upon and improves the Blendenpik solver on dense input and the (sequential) LSRN performance on sparse problems. In addition to the hashing sketching improvements, we add suitable linear algebra tools for rank-deficient and for sparse problems that lead Ski-LLS to outperform not only sketching-based routines on randomly generated input, but also state of the art direct solver SPQR and iterative code HSL on certain subsets of the sparse Florida matrix collection; namely, on least squares problems that are significantly overdetermined, or moderately sparse, or difficult.

**Keywords:** sketching techniques, sparse random matrices, linear least-squares, iterative methods, preconditioning, sparse matrices, mathematical software.

**Mathematics Subject Classification:** 65K05, 93E24, 65F08, 65F10, 65F20, 65F50, 62J05

## 1 Introduction

Over the past fifteen years, sketching techniques have proved to be useful tools for improving the computational efficiency and scalability of linear algebra and optimization techniques, such as of methods for solving least-squares, sums of functions and low-rank matrix problems [32, 45]. The celebrated Johnson-Lindenstrauss Lemma [24] and subsequent results use a carefully-chosen random matrix  $S \in \mathbb{R}^{m \times n}$ ,  $m \ll n$ , to sample/project the column vectors of a matrix  $A \in \mathbb{R}^{n \times d}$ , with  $n \geq d$  to lower dimensions, while approximately preserving the length of vectors in this column space; this quality of  $S$  (and of its associated distribution) is captured by the (oblivious) subspace embedding property [45] (see Definition 3). Sketching has found a variety of uses,

---

\*The order of the authors is alphabetical; the third author is the primary contributor.

<sup>†</sup>Mathematical Institute, University of Oxford, Radcliffe Observatory Quarter, Woodstock Road, Oxford, OX2 6GG, United Kingdom ([cartis@maths.ox.ac.uk](mailto:cartis@maths.ox.ac.uk)). This author's work was supported by the Alan Turing Institute for Data Science, London, UK.

<sup>‡</sup>Numerical Algorithms Group Ltd, 30 St Giles', Oxford, OX1 3LE, United Kingdom ([jan.fiala@nag.co.uk](mailto:jan.fiala@nag.co.uk)).

<sup>§</sup>Mathematical Institute, University of Oxford, Radcliffe Observatory Quarter, Woodstock Road, Oxford, OX2 6GG, United Kingdom ([shaoz@maths.ox.ac.uk](mailto:shaoz@maths.ox.ac.uk)). This author's work was supported by the EPSRC Centre For Doctoral Training in Industrially Focused Mathematical Modelling (EP/L015803/1) in collaboration with the Numerical Algorithms Group Ltd.

such as in the solution of linear least squares problems and low-rank matrix approximations, in subsampling data points and reducing the dimension of the parameter space in training tasks for machine learning systems and imaging, and more.

Gaussian matrices have been shown to have optimal subspace embedding properties in terms of the size  $m$  of the sketch, allowing  $m = \mathcal{O}(d)$  for any given matrix  $A \in \mathbb{R}^{n \times d}$  (with high probability). Due to their density, however, the computational cost of calculating the sketch  $SA$  is often prohibitively expensive, potentially limiting its use in large-scale contexts. To alleviate these deficiencies, as well as an alternative to subsampling techniques, sparse random matrix ensembles have been proposed in the randomized linear algebra literature by Clarkson and Woodruff [8]; namely,  $s$ -hashing matrices with  $s$  (fixed) non-zeros per column,  $s \geq 1$ . Hashing matrices have been shown empirically to be almost as good as Gaussian matrices for sketching purposes [10], but their theoretical embedding guarantees require at least  $m = \mathcal{O}(d^2)$  rows for the sketch  $S$  [39]. Furthermore, numerical evidence also shows that using  $s \geq 2$  instead of  $s = 1$  in  $S$  leads to a more accurately sketched input, especially when this input is sparse. Here, in our theoretical developments, we aim to make precise and quantify these numerical observations; namely, we show that subspace embedding properties can be shown for  $s$ -hashing matrices with  $s \geq 1$  that have the optimal dimension dependence  $m = \mathcal{O}(d)$  on some classes of inputs.

Sketching results have direct implications on the efficiency of solving Linear Least Squares (LLS) problems, which can be written as the optimization problem,

$$\min_{x \in \mathbb{R}^d} \|Ax - b\|_2^2, \quad (1)$$

where  $A \in \mathbb{R}^{n \times d}$  is a given data matrix that has (unknown) rank  $r$ ,  $b \in \mathbb{R}^n$  is the vector of observations, and  $n \geq d \geq r$ . The sketched matrix  $SA$  is used to either directly compute an approximate solution to (1) or to generate a high-quality preconditioner for the iterative solution of (1); the latter has been the basis of state-of-the-art randomized linear algebra codes such as Blendenpik [3] and LSRN [36], where the latter improves on the former by exploiting input sparsity, parallelization and allowing rank-deficiency of the input matrix. Here, we propose a generic solver called Ski-LLS (SKetchIng for Linear Least Squares) that builds upon and extends the Blendenpik/(serial) LSRN frameworks and judiciously uses hashing sketching with one or more nonzeros and efficiently addresses both dense and sparse inputs; extensive numerical results are presented.

#### Existing literature: sparse random ensembles and subspace embedding properties

Various sparse and structured random matrices have been proposed, in an attempt to reduce the computational cost of using dense Gaussian sketching. As an obvious choice, subsampling matrices that have one nonzero per row have been used in [14] but do not have good subspace embedding properties unless sampling is done in a non-uniform way using computationally expensive probabilities.

The fast Johnson-Lindenstrauss transform (FJLT) proposed by Ailon and Chazelle [1] is a structured matrix, needing  $\mathcal{O}(nd(\log(d) + \log \log n))$  operations to apply to a given  $A \in \mathbb{R}^{n \times d}$ . To ensure subspace embedding properties of the sketched input, the sketching matrix  $S$  is required to have  $m = \mathcal{O}\left((\sqrt{d} + \sqrt{\log n})^2 \log(d)\right)$  [44, 2]. Clarkson and Woodruff [8] proposed the use of 1-hashing matrices, with one non-zero ( $\pm 1$ ) per column in random rows. Such a random matrix takes  $\mathcal{O}(nnz(A))$  operations to be applied to  $A$ , where  $nnz(A)$  denotes the number of nonzeros in  $A$ . It needs  $m = \Theta(d^2)$  rows to be a subspace embedding [35, 39, 38]. Cohen [9] and Nelson and Nguyen [37] have shown that increasing number of non-zeros per column in hashing matrices, namely, using  $s$ -hashing matrices with  $s > 1$ , leads to a reduced requirement in the number of rows  $m$  of the sketching matrix. Bourgain et al [6] further showed that if the coherence of the input matrix – a measure of the non-uniformity of its rows – is sufficiently low, 1-hashing sketching requires fewer rows, namely  $m = \mathcal{O}(d \log^2 d)$ ; see Table 1. When  $s$ -hashing matrices are employed, the coherence requirement can be relaxed by a factor of  $\sqrt{s}$  for similar  $m$  requirements [6]. A recent paper on tensor subspace embeddings [20] can be particularised to vector subspace embeddings, leading to a matrix distribution of matrices  $S$  with  $m = \mathcal{O}(d \log^4 d)$  that requires  $\mathcal{O}(n \log n)$  operations to apply to any vector.

In [7], a ‘stable’ 1-hashing matrix is proposed, for which each row has approximately the same number of non-zeros, and that has good JL embedding properties with the optimal  $m = \mathcal{O}(\epsilon^{-2} \log(1/\delta))$  for such projections. Furthermore, [28] proposed learning the positions and values of non-zero entries in 1-hashing matrices by assuming the data comes from a fixed distribution. Our contributions to the embedding properties of  $s$ -hashing matrices with  $s \geq 1$  are summarized below.

**Existing literature: efficient algorithms for LLS problems employing sketching** Theoretical understanding of sketching properties led to the creation of novel algorithms for diverse problem classes, including LLS in (1). In particular, Sarlos [43] first proposed directly sketching problem (1) and solving the reduced problem, while Rokhlin [42] introduced the idea of using sketching for the construction of a suitable preconditioner, to help with solving (1) via iterative means; this has been further successfully implemented in the state of the art solvers that use sketching, Blendenpik [3] and LSRN [36]. Various sketching matrices have been used in algorithms for (1): Blendenpik uses a variant of FJLT; LSRN [36], Gaussian sketching; Iyer [22] and Dahiya[10] experimented with 1-hashing. Recently, [21] implemented Blendenpik in a distributed computing environment, showing the advantages of using sketching over LAPACK routines on huge-size LLS problems with dense matrices.

Large scale LLS problems may alternatively be solved by first-order methods, such as stochastic gradient or block coordinates, that use sketching directly in subselecting the gradient terms or components that form the search direction [27, 26, 30, 19, 31]. Sketching curvature information has also been investigated for problem (1) such as in [25], while [46] proposed a gradient-based sampling method. A nice overview of these methods can be found in [29]. However, such first order approaches are beyond our scope here, as we are searching for high-accuracy solutions to problem (1), extending the Blendenpik and LSRN frameworks.

LLS problems have been the focus of the numerical linear algebra community for several decades now; techniques abound, and we refer the interested reader to Chapter 9 in [40] for a brief introduction, and to classical textbooks [5] for an extensive coverage. For a more recent benchmarking paper that addresses recent developments in preconditioned iterative methods, see [18]; we use one of the most competitive codes in [18] for the benchmarking of our solver.

**Summary of our contributions** In terms of theoretical contributions to **subspace embedding properties** of  $s$ -hashing matrices, we have the following main results.

- We provide a subspace embedding result for 1-hashing matrices  $S \in \mathbb{R}^{m \times n}$  that has the optimal choice of embedding dimension,  $m = \mathcal{O}(d)$ , when applied to a(ny) given  $A \in \mathbb{R}^{n \times d}$  with sufficiently small coherence (see Theorem 2). This result improves upon Bourgain et al [6], where  $m = \mathcal{O}(d \log^2 d)$ , but where a more relaxed coherence requirement is imposed; see Table 1. Our optimal dimension dependence comes at the expense of a stricter coherence requirement, reducing the problem class we can address.
- We cascade our result for 1-hashing to  $s$ -hashing matrices with  $s > 1$ , achieving an embedding result with a sketching matrix of size  $m = \mathcal{O}(d)$  when applied to a larger input matrix class, whose coherence is allowed to increase by at most a factor of  $\sqrt{s}$ ; see Theorem 3.
- Instead of using subsampled coherence-reduction transformations, we propose to replace the subsampled aspect with hashing, leading to novel hashed- variants of any such transformations; Figure 1 gives a simple illustration of the benefits of our proposal. Then, in the case of Randomised Hadamard Transform (RHT), we show that for its 1-hashed variant, H-RHT, a subspace embedding result holds that has optimal dimension dependence, namely,  $m = \mathcal{O}(d)$ , when applied to any  $n \times d$  input matrix with  $n > d^3$  (that is sufficiently overdetermined); see Theorem 6 which improves corresponding results for subsampled RHT [44].
- Finally, we propose a new  $s$ -hashing variant that coincides with (the usual) 1-hashing when  $s = 1$  but may differ when  $s > 1$ ; in particular, this variant has at most  $s$  nonzeros per

column, as opposed to  $s$ -hashing that has precisely  $s$  nonzero entries. We show a general subspace embedding result that allows translating any subspace embedding property of 1-hashing sketching and input matrix with coherence at most (some/any)  $\mu > 0$  to a similar guarantee for the  $s$ -hashing variant and input of coherence at most  $\mu\sqrt{s}$ ; see Theorem 4. Our result relies on a single, intuitive proof, and is not tied to any particular proof of embedding properties in the case of 1-hashing; thus if the latter improves, so will the guarantees for the  $s$ -hashing variant, immediately.

We note that the  $d$  dependence in our embedding results is replaced by  $r$  dependence, where  $r$  is the rank of the embedded matrix  $A$ , as this accurately captures the true data dimension.

Regarding our **algorithmic contributions**, we introduce and analyse a general framework (Algorithm 1) for solving (1) that includes Blendenpik and LSRN, while allowing arbitrary choice of sketching and a wide range of rank-revealing factorizations in order to build a preconditioner using the sketched matrix  $SA$ . We show that a minimal residual solution of (1) is generated by Algorithm 1 when a rank-revealing factorization is calculated for  $SA$ . If the latter is replaced by a complete orthogonal factorization of  $SA$ , then the minimal norm solution of (1) is obtained. Our **software contribution** is Ski-LLS, a C++ implementation of Algorithm 1 that allows and distinguishes between dense and sparse input and makes extensive use of  $s$ -hashing matrices for the sketching step. When  $A$  in (1) is dense, Ski-LLS employs the following changes/improvements to Blendenpik: choice of sketching is the 1-*hashed* DHT transform instead of subsampled; a randomised column pivoted QR factorization [34] of  $SA$  instead of just QR, allowing rank-deficient input. The latter factorization is less computationally expensive than the SVD choice in LSRN. Similarly, in the case of sparse input  $A$ , we let  $S$  be an  $s$ -hashing matrix (with  $s = 2$ ) and use a rank-revealing sparse QR factorization for  $SA$  (SPQR [11]), which leads to improved performance over (serial) LSRN and other state of the art LLS solvers. In particular, Ski-LLS is more than 10 times faster than LSRN for sparse Gaussian inputs. We extensively compare Ski-LLS with LSRN, SPQR and the iterative approach in HSL, on the Florida Matrix Collection [12], and find that Ski-LLS is competitive on significantly-overdetermined or ill-conditioned inputs.

**Structure of the paper** The necessary technical background is given in Section 2. In Section 3, we state and prove our theorem on the coherence requirement needed to use 1-hashing with  $m = O(d)$  as a subspace embedding. In Section 4, we show how increasing the number of non-zeros per column from 1 to  $s$  relaxes the coherence requirement for  $s$ -hashing matrices by a factor of  $\sqrt{s}$ . In the same section, we also consider hashed coherence reduction transformations and show their embeddings properties. The algorithmic framework that uses sketching for rank-deficient linear least squares is introduced and analysed in Section 5. In Section 6 we introduce our solver Ski-LLS and discuss its key features and implementation details. We show extensive numerical experiments for both dense and sparse matrices  $A$  and demonstrate the competitiveness of our solver in Section 7.

Throughout the paper, we let  $\langle \cdot, \cdot \rangle$  and  $\|\cdot\|_2$  denote the usual Euclidean inner product and norm, respectively, and  $\|\cdot\|_\infty$ , the  $l_\infty$  norm. Also, for some  $n \in \mathbb{N}$ ,  $[n] = \{1, 2, \dots, n\}$ . For a(ny) symmetric positive definite matrix  $\overline{W}$ , we define the norm  $\|x\|_{\overline{W}} := x^T \overline{W} x$ , for all  $x$ , as the norm induced by  $\overline{W}$ . The notation  $\Theta(\cdot)$  denotes both lower and upper bounds of the respective order.  $\Omega(\cdot)$  denotes a lower bound of the respective order.

## 2 Technical Background

In this section, we review some important concepts and their properties that we then use throughout the paper. We employ several variants of the notion of random embeddings for finite or infinite sets, as we define next.

### 2.1 Random embeddings

We start with a very general concept of embedding a (finite or infinite) number of points; throughout, we let  $\epsilon \in (0, 1)$  be the user-chosen/arbitrary error tolerance in the embeddings and  $n, k \in \mathbb{N}$ .

**Definition 1** (Generalised JL<sup>1</sup> embedding [45]). *A generalised  $\epsilon$ -JL embedding for a set  $Y \subseteq \mathbb{R}^n$  is a matrix  $S \in \mathbb{R}^{m \times n}$  such that*

$$-\epsilon \|y_i\|_2 \cdot \|y_j\|_2 \leq \langle Sy_i, Sy_j \rangle - \langle y_i, y_j \rangle \leq \epsilon \|y_i\|_2 \cdot \|y_j\|_2, \quad \text{for all } y_i, y_j \in Y. \quad (2)$$

If we let  $y_i = y_j$  in (2), we recover the common notion of an  $\epsilon$ -JL embedding, that approximately preserves the length of vectors in a given set.

**Definition 2** (JL embedding [45]). *An  $\epsilon$ -JL embedding for a set  $Y \subseteq \mathbb{R}^n$  is a matrix  $S \in \mathbb{R}^{m \times n}$  such that*

$$(1 - \epsilon) \|y\|_2^2 \leq \|Sy\|_2^2 \leq (1 + \epsilon) \|y\|_2^2 \quad \text{for all } y \in Y. \quad (3)$$

Often, in the above definitions, the set  $Y = \{y_1, \dots, y_k\}$  is a finite collection of vectors in  $\mathbb{R}^n$ . But an infinite number of points may also be embedded, such as in the case when  $Y$  is an entire subspace. Then, an embedding approximately preserves pairwise distances between any points in the column space of a matrix  $B \in \mathbb{R}^{n \times k}$ .

**Definition 3** ( $\epsilon$ -subspace embedding [45]). *An  $\epsilon$ -subspace embedding for a matrix  $B \in \mathbb{R}^{n \times k}$  is a matrix  $S \in \mathbb{R}^{m \times n}$  such that*

$$(1 - \epsilon) \|y\|_2^2 \leq \|Sy\|_2^2 \leq (1 + \epsilon) \|y\|_2^2 \quad \text{for all } y \in Y = \{y : y = Bz, z \in \mathbb{R}^k\}. \quad (4)$$

In other words,  $S$  is an  $\epsilon$ -subspace embedding for  $B$  if and only if  $S$  is an  $\epsilon$ -JL embedding for the column subspace  $Y$  of  $B$ .

Oblivious embeddings are matrix distributions such that given a(ny) subset/column subspace of vectors in  $\mathbb{R}^n$ , a random matrix drawn from such a distribution is an embedding for these vectors with high probability. We let  $1 - \delta \in [0, 1]$  denote a(ny) success probability of an embedding.

**Definition 4** (Oblivious embedding [45, 43]). *A distribution  $\mathcal{S}$  on  $S \in \mathbb{R}^{m \times n}$  is an  $(\epsilon, \delta)$ -oblivious embedding if given a fixed/arbitrary set of vectors, we have that, with probability at least  $1 - \delta$ , a matrix  $S$  from the distribution is an  $\epsilon$ -embedding for these vectors.*

Using the above definitions of embeddings, we have distributions that are *oblivious JL-embeddings* for a(ny) given/fixed set  $Y$  of some vectors  $y \in \mathbb{R}^n$ , and distributions that are *oblivious subspace embeddings* for a(ny) given/fixed matrix  $B \in \mathbb{R}^{n \times k}$  (and for the corresponding subspace  $Y$  of its columns). We note that depending on the quantities being embedded, in addition to  $\epsilon$  and  $\delta$  dependencies, the size  $m$  of  $S$  may depend on  $n$  and the ‘dimension’ of the embedded sets; for example, in the case of a finite set  $Y$  of  $k$  vectors in  $\mathbb{R}^n$ ,  $m$  additionally may depend on  $k$  while in the subspace embedding case,  $m$  may depend on the rank  $r$  of  $B$ .

## 2.2 Generic properties of subspace embeddings

A necessary condition for a matrix  $S$  to be an  $\epsilon$ -subspace embedding for a given matrix is that the sketched matrix has the same rank. The proofs of the next two lemmas are provided in the appendix.

**Lemma 1.** *If the matrix  $S$  is an  $\epsilon$ -subspace embedding for a given matrix  $B$  for some  $\epsilon \in (0, 1)$ , then  $\text{rank}(SB) = \text{rank}(B)$ , where  $\text{rank}(\cdot)$  denotes the rank of the argument matrix.*

Given any matrix  $A \in \mathbb{R}^{n \times d}$  of rank  $r$ , the compact singular value decomposition (SVD) of  $A$  provides a perfect subspace embedding. In particular, let

$$A = U \Sigma V^T, \quad (5)$$

where  $U \in \mathbb{R}^{n \times r}$  with orthonormal columns,  $\Sigma \in \mathbb{R}^{r \times r}$  is diagonal matrix with strictly positive diagonal entries, and  $V \in \mathbb{R}^{d \times r}$  with orthonormal columns [16]. Then the matrix  $U^T$  is a  $\epsilon$ -subspace embedding for  $A$  for any  $\epsilon \in (0, 1)$ .

Next we connect the embedding properties of  $S$  for  $A$  with those for  $U$  in (5), using a proof technique in Woodruff [45].

---

<sup>1</sup>Note that ‘JL’ stands for Johnson-Lindenstrauss, recalling their pioneering lemma [24].

**Lemma 2.** Let  $A \in \mathbb{R}^{n \times d}$  with rank  $r$  and SVD-decomposition factor  $U \in \mathbb{R}^{n \times r}$  defined in (5), and let  $\epsilon \in (0, 1)$ . Then the following equivalences hold:

- (i) a matrix  $S$  is an  $\epsilon$ -subspace embedding for  $A$  if and only if  $S$  is an  $\epsilon$ -subspace embedding for  $U$ , namely,

$$(1 - \epsilon)\|Uz\|_2^2 \leq \|SUz\|_2^2 \leq (1 + \epsilon)\|Uz\|_2^2, \quad \text{for all } z \in \mathbb{R}^r. \quad (6)$$

- (ii) A matrix  $S$  is an  $\epsilon$ -subspace embedding for  $A$  if and only if for all  $z \in \mathbb{R}^r$  with  $\|z\|_2 = 1$ , we have<sup>2</sup>

$$(1 - \epsilon)\|Uz\|_2^2 \leq \|SUz\|_2^2 \leq (1 + \epsilon)\|Uz\|_2^2. \quad (7)$$

**Remark 1.** Lemma 2 shows that to obtain a subspace embedding for an  $n \times d$  matrix  $A$  it is sufficient (and necessary) to embed correctly its left-singular matrix that has rank  $r$ . Thus, the dependence on  $d$  in subspace embedding results can be replaced by dependence on  $r$ , the rank of the input matrix  $A$ . As rank deficient matrices  $A$  are important in this paper, we opt to state our results in terms of their  $r$  dependency (instead of  $d$ ).

The matrix  $U$  in (5) can be seen as the ideal ‘sketching’ matrix for  $A$ ; however, there is not much computational gain in doing this as computing the compact SVD has similar complexity as computing a minimal residual solution to (1) directly.

### 2.3 Sparse matrix distributions and their embeddings properties

In terms of optimal embedding properties, it is well known that (dense) scaled Gaussian matrices  $S$  with  $m = \mathcal{O}(\epsilon^{-2}(r + \log(1/\delta)))$  provide an  $(\epsilon, \delta)$ -oblivious subspace embedding for  $n \times d$  matrices  $A$  of rank  $r$  [45]. However, the computational cost of the matrix-matrix product  $SA$  is  $\mathcal{O}(nd^2)$ , which is similar to the complexity of solving the original LLS problem (1); thus it seems difficult to achieve computational gains by calculating a sketched solution of (1) in this case. In order to improve the computational cost of using sketching for solving LLS problems, and to help preserve input sparsity (when  $A$  is sparse), sparse random matrices have been proposed, namely, such as random matrices with one non-zero per row. However, uniformly sampling rows of  $A$  (and entries of  $b$  in (1)) may miss choosing some (possibly important) row/entry. A more robust proposal, both theoretically and numerically, is to use hashing matrices, with one (or more) nonzero entries per column, which when applied to  $A$  (and  $b$ ), captures all rows of  $A$  (and entries of  $b$ ) by adding two (or more) rows/entries with randomised signs.

**Definition 5.** [8]  $S \in \mathbb{R}^{m \times n}$  is a  $s$ -hashing matrix if independently for each  $j \in [n]$ , we sample without replacement  $i_1, i_2, \dots, i_s \in [m]$  uniformly at random and let  $S_{i_k j} = \pm 1/\sqrt{s}$  for  $k = 1, 2, \dots, s$ .

It follows that when  $s = 1$ ,  $S \in \mathbb{R}^{m \times n}$  is a 1-hashing matrix if independently for each  $j \in [n]$ , we sample  $i \in [m]$  uniformly at random and let  $S_{ij} = \pm 1$ <sup>3</sup>.

Still, in general, the optimal dimension dependence present in Gaussian subspace embeddings cannot be replicated even for hashing distributions, as our next example illustrates.

**Example 1** (The 1-hashing matrix distributions fails to yield an oblivious subspace embedding with  $m = \mathcal{O}(r)$ ). Consider the matrix

$$A = \begin{pmatrix} I_{r \times r} & 0 \\ 0 & 0 \end{pmatrix} \in \mathbb{R}^{n \times d}. \quad (8)$$

If  $S$  is a 1-hashing matrix with  $m = \mathcal{O}(r)$ , then  $SA = (S_1 \ 0)$ , where the  $S_1$  block contains the first  $r$  columns of  $S$ . To ensure that the rank of  $A$  is preserved (cf. Lemma 1), a necessary condition for  $S_1$  to have rank  $r$  is that the  $r$  non-zeros of  $S_1$  are in different rows. Since, by

<sup>2</sup>We note that since  $\|z\|_2 = 1$  and  $U$  has orthonormal columns,  $\|Uz\|_2 = \|z\|_2 = 1$  in (7).

<sup>3</sup>The random signs are so as to ensure that in expectation, the matrix  $S$  preserves the norm of a vector  $x \in \mathbb{R}^n$ .

definition, the respective row is chosen independently and uniformly at random for each column of  $S$ , the probability of  $S_1$  having rank  $r$  is no greater than

$$\left(1 - \frac{1}{m}\right) \cdot \left(1 - \frac{2}{m}\right) \cdot \dots \cdot \left(1 - \frac{r-1}{m}\right) \leq e^{-\frac{1}{m} - \frac{2}{m} - \dots - \frac{r-1}{m}} = e^{-\frac{r(r-1)}{2m}}, \quad (9)$$

For the probability<sup>4</sup> (9) to be at least  $1/2$ , we must have  $m \geq \frac{r(r-1)}{2 \log(2)}$ .

The above example improves upon the lower bound in Nelson et al. [38] by slightly relaxing the requirements on  $m$  and  $n$ <sup>5</sup>. We note that in the order of  $r$  (or equivalently<sup>6</sup>,  $d$ ), the lower bound  $m = \mathcal{O}(r^2)$  for 1-hashing matches the upper bound given in Nelson and Nguyen [37], Meng and Mahoney [35].

When  $S$  is an  $s$ -hashing matrix, with  $s > 1$ , the tight bound  $m = \Theta(r^2)$  can be improved to  $m = \Theta(r \log r)$  for  $s$  sufficiently large. In particular, Cohen [9] derived a general upper bound that implies, for example, subspace embedding properties of  $s$ -hashing matrices provided  $m = \mathcal{O}(d \log d)$  and  $s = \mathcal{O}(\log d)$ ; the value of  $s$  may be further reduced to a constant (that is not equal to 1) at the expense of increasing  $m$  and worsening its dependence of  $d$ . A lower bound for guaranteeing oblivious embedding properties of  $s$ -hashing matrices is given in [39]. Thus we can see that for  $s$ -hashing (and especially for 1-hashing) matrices, their general subspace embedding properties are suboptimal in terms of the dependence of  $m$  on  $d$  when compared to the Gaussian sketching results. To improve the embedding properties of hashing matrices, we must focus on special structure input matrices.

### 2.3.1 Coherence-dependent embedding properties of sparse random matrices

A feature of the problematic matrix (8) is that its rows are separated into two groups, with the first  $r$  rows containing all the information. If the rows of  $A$  were more ‘uniform’ in the sense of equally important in terms of relevant information content, hashing may perform better as a sketching matrix. Interestingly, it is not the uniformity of the rows of  $A$  but the uniformity of the rows of  $U$ , the left singular matrix from the compact SVD of  $A$ , that plays an important role. The concept of coherence is a useful proxy for the uniformity of the rows of  $U$  and  $A$ <sup>7</sup>.

**Definition 6.** (Matrix coherence [32]) The coherence of a matrix  $A \in \mathbb{R}^{n \times d}$ , denoted  $\mu(A)$ , is the largest Euclidean norm of the rows of  $U$  defined in (5). Namely,

$$\mu(A) = \max_{i \in [n]} \|U_i\|_2, \quad (10)$$

where  $U_i$  denotes the  $i$ th row of  $U$ .

Some useful properties follow.

**Lemma 3.** Let  $A \in \mathbb{R}^{n \times d}$  have rank  $r \leq d \leq n$ . Then

$$\sqrt{\frac{r}{n}} \leq \mu(A) \leq 1. \quad (11)$$

Furthermore, if  $\mu(A) = \sqrt{\frac{r}{n}}$ , then  $\|U_i\|_2 = \sqrt{\frac{r}{n}}$  for all  $i \in [n]$  where  $U$  is defined in (5).

<sup>4</sup>The argument in the example relating to 1-hashing sketching is related to the birthday paradox, as mentioned (but not proved) in Nelson and Nguyen [38].

<sup>5</sup>We note that in fact, [37] considers a more general set up, namely, any matrix distribution with column sparsity one.

<sup>6</sup>See Remark 1.

<sup>7</sup>We note that sampling matrices were shown to have good subspace embedding properties for input matrices with low coherence [1, 44]. Even if the coherence is minimal, the size of the sampling matrix has a  $d \log d$  dependence where the  $\log d$  term cannot be removed due to the coupon collector problem [44].

*Proof.* Since the matrix  $U \in \mathbb{R}^{n \times r}$  has orthonormal columns, we have that

$$\sum_{i=1}^n \|U_i\|_2^2 = r. \quad (12)$$

Therefore the maximum 2-norm of  $U$  must not be less than  $\sqrt{\frac{r}{n}}$ , and thus  $\mu(A) \geq \sqrt{\frac{r}{n}}$ . Furthermore, if  $\mu(A) = \sqrt{\frac{r}{n}}$ , then (12) implies  $\|U_i\|_2 = \sqrt{\frac{r}{n}}$  for all  $i \in [n]$ .

Next, by expanding the set of columns of  $U$  to a basis of  $\mathbb{R}^n$ , there exists  $U_f \in \mathbb{R}^{n \times n}$  such that  $U_f = (U \hat{U})$  orthogonal where  $\hat{U} \in \mathbb{R}^{n \times (n-d)}$  has orthonormal columns. The 2-norm of  $i$ th row of  $U$  is bounded above by the 2-norm of  $i$ th row of  $U_f$ , which is one. Hence  $\mu(A) \leq 1$ .  $\square$

We note that for  $A$  in (8), we have  $\mu(A) = 1$ . The maximal coherence of this matrix sheds some light on the ensuing poor embedding properties we noticed in Example 1.

Bourgain et al [6] gives a general result that captures the coherence-restricted subspace embedding properties of  $s$ -hashing matrices.

**Theorem 1** (Bourgain et al [6]). *Let  $A \in \mathbb{R}^{n \times d}$  with coherence  $\mu(A)$  and rank  $r$ ; and let  $0 < \epsilon, \delta < 1$ . Assume also that*

$$m \geq c_1 \max \{ \delta^{-1}, (r + \log m) [\min \{ \log^2(r/\epsilon), \log^2(m) \} + r \log(1/\delta)] \epsilon^{-2} \} \quad (13)$$

$$\text{and } s \geq c_2 [\log(m) \log(1/\delta) \min \{ \log^2(r/\epsilon), \log^2(m) \} + \log^2(1/\delta)] \mu(A)^2 \epsilon^{-2}, \quad (14)$$

where  $c_1$  and  $c_2$  are positive constants. Then a(ny)  $s$ -hashing matrix  $S \in \mathbb{R}^{m \times n}$  is an  $\epsilon$ -subspace embedding for  $A$  with probability at least  $1 - \delta$ .

Substituting  $s = 1$  in (14), we can use the above Theorem to deduce an upper bound  $\mu$  of acceptable coherence values of the input matrix  $A$ , namely,

$$\mu(A) \leq c_2^{-1/2} \epsilon [\log(m) \log(1/\delta) \min \{ \log^2(r/\epsilon), \log^2(m) \} + \log^2(1/\delta)]^{-1/2} := \mu. \quad (15)$$

Thus Theorem 1 implies that the distribution of 1-hashing matrices with  $m$  satisfying (13) is an oblivious subspace embedding for any input matrix  $A$  with  $\mu(A) \leq \mu$ , where  $\mu$  is defined in (15).

### 2.3.2 Non-uniformity of vectors and their relation to embedding properties of sparse random matrices

In order to prove some of our main results, we need a corresponding notion of coherence of vectors, in order to be able to measure the ‘importance’ of their respective entries; this is captured by the so-called non-uniformity of a vector.

**Definition 7** (Non-uniformity of a vector). *Given  $x \in \mathbb{R}^n$ , the non-uniformity of  $x$ ,  $\nu(x)$ , is defined as*

$$\nu(x) = \frac{\|x\|_\infty}{\|x\|_2}. \quad (16)$$

We note that for any vector  $x \in \mathbb{R}^n$ , we have  $\frac{1}{\sqrt{n}} \leq \nu(x) \leq 1$ .

**Lemma 4.** *Given  $A \in \mathbb{R}^{n \times d}$ , let  $y = Ax$  for some  $x \in \mathbb{R}^d$ . Then*

$$\nu(y) \leq \mu(A). \quad (17)$$

The next lemmas are crucial to our results in the next section; the proof of the first lemma can be found in the respective paper [15], while the proofs of Lemmas 6 and Lemma 7 are given in the appendix.

We also note, in subsequent results, the presence of *problem-independent constants*, also called absolute constants that will be implicitly or explicitly defined, depending on the context.



Our convention here is as expected, that the same notation denotes the same constant across all results in the paper.

The following expression will be needed in our results,

$$\bar{\nu}(\epsilon, \delta) := C_1 \sqrt{\epsilon} \min \left\{ \frac{\log(E/\epsilon)}{\log(1/\delta)}, \sqrt{\frac{\log(E)}{\log(1/\delta)}} \right\}, \quad (18)$$

where  $\epsilon, \delta \in (0, 1)$  and  $E, C_1 > 0$ .

**Lemma 5** ([15], Theorem 2). *Suppose that  $\epsilon, \delta \in (0, 1)$ , and  $E$  satisfies  $C \leq E < \frac{2}{\delta \log(1/\delta)}$ , where  $C > 0$  and  $C_1$  are problem-independent constants. Let  $m \leq n \in \mathbb{N}$  with  $m \geq E\epsilon^{-2} \log(1/\delta)$ .*

*Then, for any  $x \in \mathbb{R}^n$  with*

$$\nu(x) \leq \bar{\nu}(\epsilon, \delta), \quad (19)$$

*where  $\bar{\nu}(\epsilon, \delta)$  is defined in (18), a randomly generated 1-hashing matrix  $S \in \mathbb{R}^{m \times n}$  is an  $\epsilon$ -JL embedding for  $\{x\}$  with probability at least  $1 - \delta$ .*

**Lemma 6.** *Let  $\epsilon, \delta \in (0, 1)$ ,  $\nu \in (0, 1]$  and  $m, n \in \mathbb{N}$ . Let  $\mathcal{S}$  be a distribution of  $m \times n$  random matrices. Suppose that for any given  $y$  with  $\nu(y) \leq \nu$ , a matrix  $S \in \mathbb{R}^{m \times n}$  randomly drawn from  $\mathcal{S}$  is an  $\epsilon$ -JL embedding for  $\{y\}$  with probability at least  $1 - \delta$ . Then for any given set  $Y \subseteq \mathbb{R}^n$  with  $\max_{y \in Y} \nu(y) \leq \nu$  and cardinality  $|Y| \leq 1/\sqrt{\delta}$ , a matrix  $S$  randomly drawn from  $\mathcal{S}$  is an  $\epsilon$ -JL embedding for  $Y$  with probability at least  $1 - |Y|\delta$ .*

**Lemma 7.** *Let  $\epsilon \in (0, 1)$ , and  $Y \subseteq \mathbb{R}^n$  be a finite set such that  $\|y\|_2 = 1$  for each  $y \in Y$ . Define*

$$Y_+ = \{y_1 + y_2 : y_1, y_2 \in Y\} \quad (20)$$

$$Y_- = \{y_1 - y_2 : y_1, y_2 \in Y\}. \quad (21)$$

*If  $S \in \mathbb{R}^{m \times n}$  is an  $\epsilon$ -JL embedding for  $\{Y_+ \cup Y_-\}$ , then  $S$  is a generalised  $\epsilon$ -JL embedding for  $Y$ .*

### 3 Hashing sketching with $m = \mathcal{O}(r)$

Our first result shows that if the coherence of the input matrix is sufficiently low, the distribution of 1-hashing matrices with  $m = \mathcal{O}(r)$  is an  $(\epsilon, \delta)$ -oblivious subspace embedding.

The following expression will be useful later,

$$\bar{\mu}(\epsilon, \delta) := C_1 \sqrt{C_2} \sqrt{\epsilon} \min \left\{ \frac{\log(E/(C_2\epsilon))}{4r + \log(1/\delta)}, \sqrt{\frac{\log(E)}{4r + \log(1/\delta)}} \right\}, \quad (22)$$

where  $\epsilon, \delta \in (0, 1)$ ,  $r, E > 0$  are to be chosen/defined depending on the context, and  $C_1, C_2 > 0$  are problem-independent constants.

**Theorem 2.** *Suppose that  $\epsilon, \delta \in (0, 1)$ ,  $r \leq d \leq n, m \leq n \in \mathbb{N}^+$ ,  $E > 0$  satisfy*

$$C \leq E \leq \frac{2e^{4r}}{[4r + \log(1/\delta)] \delta}, \quad (23)$$

$$m \geq EC_2^{-2} \epsilon^{-2} [4r + \log(1/\delta)], \quad (24)$$

*where  $C > 0$  and  $C_1, C_2 > 0$  are problem-independent constants. Then for any matrix  $A \in \mathbb{R}^{n \times d}$  with rank  $r$  and*

$$\mu(A) \leq \bar{\mu}(\epsilon, \delta), \quad (25)$$

*where  $\bar{\mu}(\epsilon, \delta)$  is defined in (22), a randomly generated 1-hashing matrix  $S \in \mathbb{R}^{m \times n}$  is an  $\epsilon$ -subspace embedding for  $A$  with probability at least  $1 - \delta$ .*

The proof of Theorem 2 relies on the fact that the coherence of the input matrix gives a bound on the non-uniformity of the entries for all vectors in its column space (Lemma 4), adapting standard arguments in [45] involving set covers, which are defined next and proved in the appendix.

**Definition 8.** A  $\gamma$ -cover of a set  $M$  is a subset  $N \subseteq M$  with the property that given any point  $y \in M$ , there exists  $w \in N$  such that  $\|y - w\|_2 \leq \gamma$ .

Consider a given real matrix  $U \in \mathbb{R}^{n \times r}$  with orthonormal columns, and let

$$M := \{Uz \in \mathbb{R}^n : z \in \mathbb{R}^r, \|z\|_2 = 1\}. \quad (26)$$

The next two lemmas show the existence of a  $\gamma$ -net  $N$  for  $M$ , and connect generalised JL embeddings for  $N$  with JL embeddings for  $M$ ; their proofs can be found in the appendix.

**Lemma 8.** Let  $0 < \gamma < 1$ ,  $U \in \mathbb{R}^{n \times r}$  have orthonormal columns and  $M$  be defined in (26). Then there exists a  $\gamma$ -cover  $N$  of  $M$  such that  $|N| \leq (1 + \frac{2}{\gamma})^r$ .

**Lemma 9.** Let  $\epsilon, \gamma \in (0, 1)$ ,  $U \in \mathbb{R}^{n \times d}$ ,  $M \subseteq \mathbb{R}^n$  associated with  $U$  be defined in (26). Suppose  $N \subseteq M$  is a  $\gamma$ -cover of  $M$  and  $S \in \mathbb{R}^{m \times n}$  is a generalised  $\epsilon_1$ -JL embedding for  $N$ , where  $\epsilon_1 = \frac{(1-\gamma)(1-\gamma^2)}{1+2\gamma-\gamma^2}\epsilon$ . Then  $S$  is an  $\epsilon$ -JL embedding for  $M$ .

We are ready to prove Theorem 2.

*Proof of Theorem 2.* Let  $A \in \mathbb{R}^{n \times d}$  with rank  $r$  and satisfying (25). Let  $U \in \mathbb{R}^{n \times r}$  be an SVD factor of  $A$  as defined in (5), which by definition of coherence, implies

$$\mu(U) = \mu(A) \leq \bar{\mu}(\epsilon, \delta), \quad (27)$$

where  $\bar{\mu}(\epsilon, \delta)$  is defined in (22). We let  $\gamma, \epsilon_1, \delta_1 \in (0, 1)$  be defined as

$$\gamma = \frac{2}{e^2 - 1}, \quad C_2 = \frac{(1-\gamma)(1-\gamma^2)}{1+2\gamma-\gamma^2}, \quad \epsilon_1 = C_2\epsilon \quad \text{and} \quad \delta_1 = e^{-4r}\delta, \quad (28)$$

and note that  $C_2 \in (0, 1)$  and

$$\bar{\nu}(\epsilon_1, \delta_1) = \bar{\mu}(\epsilon, \delta), \quad (29)$$

where  $\bar{\nu}(\cdot, \cdot)$  is defined in (18). Let  $M \in \mathbb{R}^n$  be associated to  $U$  as in (26) and let  $N \subseteq M$  be the  $\gamma$ -cover of  $M$  as guaranteed by Lemma 8, with  $\gamma$  defined in (28) which implies that  $|N| \leq e^{2r}$ .

Let  $S \in \mathbb{R}^{m \times n}$  be a randomly generated 1-hashing matrix with  $m \geq E\epsilon_1^{-2} \log(1/\delta_1) = EC_2^{-2}\epsilon^{-2}[4r + \log(1/\delta)]$ , where to obtain the last equality, we used (28).

To show that the sketching matrix  $S$  is an  $\epsilon$ -subspace embedding for  $A$  (with probability at least  $1 - \delta$ ), it is sufficient to show that  $S$  is an  $\epsilon_1$ -generalised JL embedding for  $N \subseteq M$  (with probability at least  $1 - \delta$ ). To see this, recall (26) and Lemma 2(ii) which show that  $S$  is an  $\epsilon$ -subspace embedding for  $A$  if and only if  $S$  is an  $\epsilon$ -JL embedding for  $M$ . Our sufficiency claim now follows by invoking Lemma 9 for  $S$ ,  $N$  and  $M$ .

We are left with considering in detail the cover set  $N = \{y_1, y_2, \dots, y_{|N|}\}$  and the following useful ensuing sets

$$\begin{aligned} N_+ &= \{y_i + y_j : i, j \in [1, |N|]\} \quad \text{and} \quad N_- = \{y_i - y_j : i, j \in [1, |N|]\}, \\ N_-^{(1)} &= \{y_i - y_j : 1 \leq i < j \leq |N|\} \quad \text{and} \quad N_-^{(2)} = \{y_i - y_j : 1 \leq j < i \leq |N|\}. \end{aligned}$$

Now let  $Y := N_+ \cup N_-^{(1)}$  and show that

$$\nu(y) \leq \bar{\nu}(\epsilon_1, \delta_1) \quad \text{for all} \quad y \in Y. \quad (30)$$

To see this, assume first that  $y = y_i + y_j \in N_+$ , with  $y_i, y_j \in N \subseteq M$ . Thus there exist  $z_i, z_j \in \mathbb{R}^r$  such that  $y_i = Uz_i$  and  $y_j = Uz_j$ , and so  $y = U(z_i + z_j)$ . Using Lemma 4,  $\nu(y) \leq \mu(U) = \mu(A)$ ,

which together with (27) and (29), gives (30) for points  $y \in N_+$ ; the proof for  $y \in N_-^{(1)}$  follows similarly.

Lemma 5 with  $(\epsilon, \delta) := (\epsilon_1, \delta_1)$  provides that for any  $x \in \mathbb{R}^n$  with  $\nu(x) \leq \bar{\nu}(\epsilon_1, \delta_1)$ ,  $S$  is an  $\epsilon_1$ -JL embedding for  $\{x\}$  with probability at least  $1 - \delta_1$ . This and (30) imply that the conditions of Lemma 6 are satisfied for  $Y = N_+ \cup N_-^{(1)}$ , from which we conclude that  $S$  is an  $\epsilon_1$ -JL embedding for  $Y$  with probability at least  $1 - |Y|\delta_1$ . Note that

$$|Y| \leq |N_+| + |N_-^{(1)}| \leq \frac{1}{2}|N|(|N|+1) + \frac{1}{2}|N|(|N|-1) = |N|^2.$$

This, the definition of  $\delta_1$  in (28) and  $|N| \leq e^{2r}$  imply that  $1 - |Y|\delta_1 \geq 1 - \delta$ . Therefore  $S$  is an  $\epsilon_1$ -JL embedding for  $N_+ \cup N_-^{(1)}$  with probability at least  $1 - \delta$ .

Finally, Definition 2 of JL-embeddings implies that the sign of the embedded vector is irrelevant and that  $\{0\}$  is always embedded, and so if  $S$  is an  $\epsilon_1$ -JL embedding for  $N_+ \cup N_-^{(1)}$ , it is also an  $\epsilon_1$ -JL embedding for  $N_+ \cup N_-$ . Lemma 7 now provides us with the desired result that then,  $S$  is a generalised  $\epsilon_1$ -JL embedding for  $N$ .  $\square$

Next we discuss the results in Theorem 2.

**Conditions for a well-defined coherence requirement** While our result guarantees optimal dimensionality reduction for the sketched matrix, using a very sparse 1-hashing matrix for the sketch, it imposes implicit restrictions on the number  $n$  of rows of  $A$ . Recalling (11), we note that condition (25) is well-defined when

$$\sqrt{\frac{r}{n}} \leq \bar{\mu}(\epsilon, \delta). \quad (31)$$

Using the definition of  $\bar{\mu}(\epsilon, \delta)$  in (22) and assuming reasonably that  $\log(1/\delta) = \mathcal{O}(r)$ , we have the lower bound

$$\bar{\mu}(\epsilon, \delta) \geq C_1 \sqrt{C_2} \sqrt{\epsilon} \frac{\min \{ \log(E/(C_2\epsilon)), \sqrt{\log E} \}}{4r + \log(1/\delta)},$$

and so (31) is satisfied if

$$n \geq \frac{r(4r + \log(1/\delta))^2}{C_1^2 C_2 \epsilon \min \{ \log^2(E/(C_2\epsilon)), \log E \}} = \mathcal{O} \left( \frac{r^3}{\epsilon \log^2(\epsilon)} \right). \quad (32)$$

**Comparison with data-independent bounds** Existing results show that  $m = \Theta(r^2)$  is both necessary and sufficient in order to secure an oblivious subspace embedding property for 1-hashing matrices with no restriction on the coherence of the input matrix [38, 37, 35]. Aside from requiring more projected rows than in Theorem 2, these results implicitly impose  $n \geq \mathcal{O}(r^2)$  for the size/rank of data matrix in order to secure meaningful dimensionality reduction.

**Comparison with data-dependent bounds** To the best of our knowledge, the only data-dependent result subspace embedding result for hashing matrices is [6] (see Theorem 1). From (13), we have that  $m \geq c_1 r \min \{ \log^2(r/\epsilon), \log^2(m) \} \epsilon^{-2}$  and hence  $m = \Omega(r \log^2 r)$ ; while Theorem 2 only needs  $m = \mathcal{O}(r)$ . However, the coherence requirement on  $A$  in Theorem 1 is weaker than (25) and so [6] applies to a wider range of inputs at the expense of a larger value of  $m$  required for the sketching matrix.

**Summary and look ahead** Table 1 summarises existing results and we see stricter coherence assumptions lead to improved dimensionality reduction properties. In the next section, we investigate relaxing coherence requirements by using hashing matrices with increased column sparsity ( $s$ -hashing) and coherence reduction transformations.

Table 1: Summary of results for 1-hashing

Result	$\mu$ (coherence of $A$ )	$m$ (size of sketching $S$ )
[35]	–	$\Theta(r^2)$
[6]	$\mathcal{O}(\log^{-3/2}(r))$	$\mathcal{O}(r \log^2(r))$
Theorem 2	$\mathcal{O}(r^{-1})$	$\mathcal{O}(r)$

## 4 Relaxing the coherence requirement using $s$ -hashing matrices

This section investigates the embedding properties of  $s$ -hashing matrices when  $s \geq 1$ . Indeed, [6] shows that  $s$ -hashing relaxes their particular coherence requirement by  $\sqrt{s}$ . Theorem 3 presents a similar result for our particular coherence requirement (25) that again guarantees embedding properties for  $m = \mathcal{O}(r)$ . Then we present a new  $s$ -hashing variant that allows us to give a general result showing that (any) subspace embedding properties of 1-hashing matrices immediately translate into similar properties for these  $s$ -hashing matrices when applied to a larger class of data matrices, with larger coherence. A simplified embedding result with  $m = \mathcal{O}(r)$  is then deduced for this  $s$ -hashing variant.

Numerical benefits of  $s$ -hashing (for improved preconditioning) are investigated in later sections; see Figures 2 and 3 for example.

### 4.1 The embedding properties of $s$ -hashing matrices

Our next result shows that using  $s$ -hashing (Definition 5) relaxes the particular coherence requirement in Theorem 2 by  $\sqrt{s}$ .

**Theorem 3.** *Let  $r \leq d \leq n \in \mathbb{N}^+$ . Let  $C_1, C_2, C_3, C_M, C_\nu, C_s > 0$  be problem-independent constants. Suppose that  $\epsilon, \delta \in (0, C_3)$ ,  $m, s \in \mathbb{N}^+$  and  $E > 0$  satisfy<sup>8</sup>*

$$1 \leq s \leq C_s C_2^{-1} \epsilon^{-1} [4r + \log(1/\delta)], \quad (33)$$

$$C_M \leq E \leq C_2^2 \epsilon^2 s [4r + \log(1/\delta)]^{-1} e^{C_s (C_2 \epsilon s)^{-1} [4r + \log(1/\delta)]}, \quad (34)$$

$$m \geq \{EC_2^{-2} \epsilon^{-2} [4r + \log(1/\delta)], se\}. \quad (35)$$

*Then for any matrix  $A \in \mathbb{R}^{n \times d}$  with rank  $r$  and  $\mu(A) \leq \sqrt{s} C_\nu C_1^{-1} \bar{\mu}(\epsilon, \delta)$ , where  $\bar{\mu}(\epsilon, \delta)$  is defined in (22), a randomly generated  $s$ -hashing matrix  $S \in \mathbb{R}^{m \times n}$  is an  $\epsilon$ -subspace embedding for  $A$  with probability at least  $1 - \delta$ .*

Theorem 3 parallels Theorem 2; and its proof relies on the following lemma which parallels Lemma 5.

**Lemma 10** ([23], Theorem 1.5). *Let  $C_1, C_3, C_M, C_\nu, C_s > 0$  be problem-independent constants. Suppose that  $\epsilon, \delta \in (0, C_3)$ ,  $m, s \in \mathbb{N}^+$ ,  $E \in \mathbb{R}$  satisfy*

$$\begin{aligned} 1 &\leq s \leq C_s \epsilon^{-1} \log(1/\delta), \\ C_M &\leq E < \epsilon^2 s \log^{-1}(1/\delta) e^{C_s (\epsilon s)^{-1} \log(1/\delta)}, \\ m &\geq \max \{E \epsilon^{-2} \log(1/\delta), se\}. \end{aligned}$$

*Then for any  $x \in \mathbb{R}^n$  with  $\nu(x) \leq \sqrt{s} C_\nu C_1^{-1} \bar{\nu}(\epsilon, \delta)$ , where  $\bar{\nu}(\epsilon, \delta)$  is defined in (18), a randomly generated  $s$ -hashing matrix  $S \in \mathbb{R}^{m \times n}$  is an  $\epsilon$ -JL embedding for  $\{x\}$  with probability at least  $1 - \delta$ .*

The proof of Theorem 3 follows the same argument as Theorem 2, replacing 1-hashing with  $s$ -hashing and using Lemma 10 instead of Lemma 5. We omit the details.

<sup>8</sup>Note that the expressions of the lower bounds in (24) and (35) are identical apart from the choice of  $E$  and the condition  $m \geq se$ .

## 4.2 A general embedding property for an $s$ -hashing variant

Note that in both Theorem 1 and Theorem 3, allowing column sparsity of hashing matrices to increase from 1 to  $s$  results in coherence requirements being relaxed by  $\sqrt{s}$ . We introduce an  $s$ -hashing variant that allows us to generalise this result.

**Definition 9.** We say  $T \in \mathbb{R}^{m \times n}$  is an  $s$ -hashing variant matrix if independently for each  $j \in [n]$ , we sample with replacement  $i_1, i_2, \dots, i_s \in [m]$  uniformly at random and add  $\pm 1/\sqrt{s}$  to  $T_{i_k j}$ , where  $k = 1, 2, \dots, s$ .<sup>9</sup>

Both  $s$ -hashing and  $s$ -hashing variant matrices reduce to 1-hashing matrices when  $s = 1$ . For  $s \geq 1$ , the  $s$ -hashing variant has at most  $s$  non-zeros per column, while the usual  $s$ -hashing matrix has precisely  $s$  nonzero entries per same column.

The next lemma connects  $s$ -hashing variant matrices to 1-hashing matrices.

**Lemma 11.** An  $s$ -hashing variant matrix  $T \in \mathbb{R}^{m \times n}$  (as in Definition 9) could alternatively be generated by calculating  $T = \frac{1}{\sqrt{s}} [S^{(1)} + S^{(2)} + \dots + S^{(s)}]$ , where  $S^{(k)} \in \mathbb{R}^{m \times n}$  are independent 1-hashing matrices for  $k = 1, 2, \dots, s$ .

*Proof.* In Definition 9, an  $s$ -hashing variant matrix  $T$  is generated by the following procedure:

```

for  $j = 1, 2, \dots, n$  do
  for  $k = 1, 2, \dots, s$  do
    Sample  $i_k \in [m]$  uniformly at random and add  $\pm 1/\sqrt{s}$  to  $T_{i_k, j}$ .

```

Due to the independence of the entries, the 'for' loops in the above routine can be swapped, leading to the equivalent formulation,

```

for  $k = 1, 2, \dots, s$  do
  for  $j = 1, 2, \dots, n$  do
    Sample  $i_k \in [m]$  uniformly at random and add  $\pm 1/\sqrt{s}$  to  $T_{i_k, j}$ .

```

For each  $k \leq s$ , the 'for' loop over  $j$  in the above routine generates an independent random 1-hashing matrix  $S^{(k)}$  and adds  $(1/\sqrt{s}) S^{(k)}$  to  $T$ . □

We are ready to state and prove the main result in this section.

**Theorem 4.** Let  $s, r \leq d \leq n \in \mathbb{N}^+$ ,  $\epsilon, \delta \in (0, 1)$ . Suppose that  $m \in \mathbb{N}^+$  is chosen such that the distribution of 1-hashing matrices  $S \in \mathbb{R}^{m \times ns}$  is an  $(\epsilon, \delta)$ -oblivious subspace embedding for any matrix  $B \in \mathbb{R}^{ns \times d}$  with rank  $r$  and  $\mu(B) \leq \mu$  for some  $\mu > 0$ . Then the distribution of  $s$ -hashing variant matrices  $T \in \mathbb{R}^{m \times n}$  is an  $(\epsilon, \delta)$ -oblivious subspace embedding for any matrix  $A \in \mathbb{R}^{n \times d}$  with rank  $r$  and  $\mu(A) \leq \mu\sqrt{s}$ .

*Proof.* Applying Lemma 11, we let

$$T = \frac{1}{\sqrt{s}} [S^{(1)} + S^{(2)} + \dots + S^{(s)}] \quad (36)$$

be a randomly generated  $s$ -hashing variant matrix where  $S^{(k)} \in \mathbb{R}^{m \times ns}$  are independent 1-hashing matrices,  $k \in \{1, \dots, s\}$ . Let  $A \in \mathbb{R}^{n \times d}$  with rank  $r$  and with  $\mu(A) \leq \mu\sqrt{s}$ ; let  $U \in \mathbb{R}^{n \times r}$  be an SVD-factor of  $A$  as defined in (5). Let

$$W = \frac{1}{\sqrt{s}} \begin{pmatrix} U \\ \vdots \\ U \end{pmatrix} \in \mathbb{R}^{ns \times r}. \quad (37)$$

As  $U$  has orthonormal columns, the matrix  $W$  also has orthonormal columns and hence the coherence of  $W$  coincides with the largest Euclidean norm of its rows

$$\mu(W) = \frac{1}{\sqrt{s}} \mu(U) = \frac{1}{\sqrt{s}} \mu(A) \leq \mu. \quad (38)$$

---

<sup>9</sup>We add  $\pm 1/\sqrt{s}$  to  $T_{i_k j}$  because we may have  $i_k = i_l$  for some  $l < k$ , as we have sampled with replacement.

Let  $S = (S^{(1)} \dots S^{(s)}) \in \mathbb{R}^{m \times ns}$ . We note that the  $j$ -th column of  $S$  is generated by sampling  $i \in [m]$  and setting  $S_{ij} = \pm 1$ . Moreover, as  $S^{(k)}$ ,  $k \in \{1, \dots, s\}$ , are independent, the sampled entries are independent. Therefore,  $S$  is distributed as a 1-hashing matrix. Furthermore, due to our assumption on the distribution of 1-hashing matrices,  $m$  is chosen such that  $S \in \mathbb{R}^{m \times ns}$  is an  $(\epsilon, \delta)$ -oblivious subspace embedding for  $(ns) \times r$  matrices of coherence at most  $\mu$ . Applying this to input matrix  $W$ , we have that with probability at least  $1 - \delta$ ,

$$(1 - \epsilon)\|z\|_2^2 = (1 - \epsilon)\|Wz\|_2^2 \leq \|SWz\|_2^2 \leq (1 + \epsilon)\|Wz\|_2^2 = (1 + \epsilon)\|z\|_2^2, \quad (39)$$

for all  $z \in \mathbb{R}^r$ , where in the equality signs, we used that  $W$  has orthonormal columns. On the other hand, we have that

$$SW = \frac{1}{\sqrt{s}} (S^{(1)} \dots S^{(s)}) \begin{pmatrix} U \\ \vdots \\ U \end{pmatrix} = \frac{1}{\sqrt{s}} [S^{(1)}U + S^{(2)}U + \dots + S^{(s)}U] = TU.$$

This and (39) provide that, with probability at least  $1 - \delta$ ,

$$(1 - \epsilon)\|z\|_2^2 \leq \|TUz\|_2^2 \leq (1 + \epsilon)\|z\|_2^2, \quad (40)$$

which implies that  $T$  is an  $\epsilon$ -subspace embedding for  $A$  by Lemma 2.  $\square$

Theorem 2 and Theorem 4 imply an  $s$ -hashing variant version of Theorem 3.

**Theorem 5.** *Suppose that  $\epsilon, \delta \in (0, 1)$ ,  $s, r \leq d \leq n, m \leq n \in \mathbb{N}^+$ ,  $E > 0$  satisfy (23) and (24). Then for any matrix  $A \in \mathbb{R}^{n \times d}$  with rank  $r$  and  $\mu(A) \leq \bar{\mu}(\epsilon, \delta)\sqrt{s}$ , where  $\bar{\mu}(\epsilon, \delta)$  is defined in (22), a randomly generated  $s$ -hashing variant matrix  $S \in \mathbb{R}^{m \times ns}$  is an  $\epsilon$ -subspace embedding for  $A$  with probability at least  $1 - \delta$ .*

*Proof.* Theorem 2 implies that the distribution of 1-hashing matrices  $S \in \mathbb{R}^{m \times ns}$  is an  $(\epsilon, \delta)$ -oblivious subspace embedding for any matrix  $B \in \mathbb{R}^{ns \times d}$  with rank  $r$  and  $\mu(B) \leq \bar{\mu}(\epsilon, \delta)$ . We also note that this result is invariant to the number of rows in  $B$  (as long as the column size of  $S$  matches the row count of  $B$ ), and so the expressions for  $m$ ,  $\bar{\mu}(\epsilon, \delta)$  and the constants therein remain unchanged.

Theorem 4 then provides that the distribution of  $s$ -hashing variant matrices  $S \in \mathbb{R}^{m \times ns}$  is an  $(\epsilon, \delta)$ -oblivious subspace embedding for any matrix  $A \in \mathbb{R}^{n \times d}$  with rank  $r$  and  $\mu(A) \leq \bar{\mu}(\epsilon, \delta)\sqrt{s}$ ; the desired result follows.  $\square$

Theorem 3 and Theorem 5 provide similar results, and we find that the latter provides simpler constant expressions (such as for  $E$ ).

### 4.3 The Hashed-Randomised-Hadamard-Transform sketching

Here we consider the Randomised-Hadamard-Transform [1], to be applied to the input matrix  $A$  before sketching, as another approach that allows reducing the coherence requirements under which good subspace embedding properties can be guaranteed. It is common to use the Subsampled-RHT (SHRT) [1], but the size of the sketch needs to be at least  $\mathcal{O}(r \log r)$ ; this prompts us to consider using hashing instead of subsampling in this context (as well), and obtain an optimal order sketching bound. Figure 2 illustrates numerically the benefit of HRHT sketching for preconditioning compared to SRHT.

**Definition 10.** *A Hashed-Randomised-Hadamard-Transform (HRHT) is an  $m \times n$  matrix of the form  $S = S_h H D$  with  $m \leq n$ , where*

- $D$  is a random  $n \times n$  diagonal matrix with  $\pm 1$  independent entries.
- $H$  is an  $n \times n$  Walsh-Hadamard matrix defined by

$$H_{ij} = n^{-1/2} (-1)^{\langle (i-1)_2, (j-1)_2 \rangle}, \quad (41)$$

where  $(i-1)_2, (j-1)_2$  are binary representation vectors of the numbers  $(i-1), (j-1)$  respectively<sup>10</sup>.

---

<sup>10</sup>For example,  $(3)_2 = (1, 1)$ .

- $S_h$  is a random  $m \times n$   $s$ -hashing or  $s$ -hashing variant matrix, independent of  $D$ .

Our next results show that if the input matrix is sufficiently over-determined, the distribution of  $HRHT$  matrices with optimal sketching size and either choice of  $S_h$ , is an  $(\epsilon, \delta)$ -oblivious subspace embedding.

**Theorem 6** ( $s$ -hashing version).  $r \leq d \leq n \in \mathbb{N}^+$ . Let  $C_1, C_2, C_3, C_M, C_\nu, C_s > 0$  be problem-independent constants. Suppose that  $\epsilon, \delta \in (0, C_3)$ ,  $m, s \in \mathbb{N}^+$  and  $E > 0$  satisfy (33), (34) and (35). Let  $\delta_1 \in (0, 1)$  and suppose further that

$$n \geq \frac{\left(\sqrt{r} + \sqrt{8 \log(n/\delta_1)}\right)^2}{s C_\nu^2 C_1^{-2} \bar{\mu}(\epsilon, \delta)^2}, \quad (42)$$

where  $\bar{\mu}(\epsilon, \delta)$  is defined in (22). Then for any matrix  $A \in \mathbb{R}^{n \times d}$  with rank  $r$ , an  $HRHT$  matrix  $S \in \mathbb{R}^{m \times n}$  with an  $s$ -hashing matrix  $S_h$ , is an  $\epsilon$ -subspace embedding for  $A$  with probability at least  $(1 - \delta)(1 - \delta_1)$ .

**Theorem 7** ( $s$ -hashing variant distribution). Suppose that  $\epsilon, \delta \in (0, 1)$ ,  $r \leq d \leq n, m \leq n \in \mathbb{N}$ ,  $E > 0$  satisfy (23) and (24). Let  $\delta_1 \in (0, 1)$  and suppose further that

$$n \geq \frac{\left(\sqrt{r} + \sqrt{8 \log(n/\delta_1)}\right)^2}{s \bar{\mu}(\epsilon, \delta)^2}, \quad (43)$$

where  $\bar{\mu}(\epsilon, \delta)$  is defined in (22). Then for any matrix  $A \in \mathbb{R}^{n \times d}$  with rank  $r$ , an  $HRHT$  matrix  $S \in \mathbb{R}^{m \times n}$  with an  $s$ -hashing variant matrix  $S_h$ , is an  $\epsilon$ -subspace embedding for  $A$  with probability at least  $1 - \delta - \delta_1 + \delta\delta_1$ .

The proof of Theorem 6 and Theorem 7 relies on the analysis in [44] of Randomised-Hadamard-Transforms, which are shown to reduce the coherence of any given matrix with high probability.

**Lemma 12.** [44] Let  $r \leq n \in \mathbb{N}^+$  and  $U \in \mathbb{R}^{n \times r}$  have orthonormal columns. Suppose that  $H, D$  are defined in Definition 10 and  $\delta_1 \in (0, 1)$ . Then

$$\mu(HDU) \leq \sqrt{\frac{r}{n}} + \sqrt{\frac{8 \log(n/\delta_1)}{n}}$$

with probability at least  $1 - \delta_1$ .

We are ready to prove Theorem 6.

*Proof of Theorem 6 and Theorem 7.* Let  $A = U\Sigma V$  be defined in (5),  $S = S_h H D$  be an  $HRHT$  matrix. Define the following events:

- $B_1 = \left\{ \mu(HDU) \leq \sqrt{r/n} + \sqrt{8 \log(n/\delta_1)/n} \right\}$ ,
- $B_2 = \left\{ \mu(HDA) \leq \sqrt{r/n} + \sqrt{8 \log(n/\delta_1)/n} \right\}$ ,
- $B_3 = \{ \mu(HDA) \leq \hat{\mu}(s, \epsilon, \delta) \}$ ,
- $B_4 = \{ S_h \text{ is an } \epsilon\text{-subspace embedding for } HDA \}$ ,
- $B_5 = \{ S_h H D \text{ is an } \epsilon\text{-subspace embedding for } A \}$ ,

where  $\hat{\mu}(s, \epsilon, \delta) = \sqrt{s} C_\nu C_1^{-1} \bar{\mu}(\epsilon, \delta)$  if  $S_h$  is an  $s$ -hashing matrix and  $\hat{\mu}(s, \epsilon, \delta) = \sqrt{s} \bar{\mu}(\epsilon, \delta)$  if  $S_h$  is an  $s$ -hashing variant matrix, and where  $\bar{\mu}(\epsilon, \delta)$  is defined in (22).

Observe that  $B_4$  implies  $B_5$  because  $B_4$  gives

$$(1 - \epsilon) \|Ax\|^2 \leq (1 - \epsilon) \|HDAx\|^2 \leq \|S_h HDAx\|^2 \leq (1 + \epsilon) \|HDAx\|^2 \leq (1 + \epsilon) \|Ax\|^2, \quad (44)$$

where the first and the last equality follows from  $HD$  being orthogonal. Moreover, observe that  $B_1 = B_2$  because  $\mu(HDA) = \max_i \|(HDU)_i\|_2 = \mu(HDU)$ , where the first equality follows from

$HDA = (H DU)\Sigma V^T$  being an SVD of  $HDA$ . Furthermore,  $B_2$  implies  $B_3$  due to (42) in the  $s$ -hashing case; and (43) in the  $s$ -hashing variant case.

Thus  $\mathbb{P}(B_5) \geq \mathbb{P}(B_4) = \mathbb{P}(B_4|B_3)\mathbb{P}(B_3) \geq \mathbb{P}(B_4|B_3)\mathbb{P}(B_2) = \mathbb{P}(B_4|B_3)\mathbb{P}(B_1)$ . If  $S_h$  is an  $s$ -hashing matrix, Theorem 3 gives  $\mathbb{P}(B_4|B_3) \geq 1 - \delta$ . If  $S_h$  is an  $s$ -hashing variant matrix, Theorem 5 gives  $\mathbb{P}(B_4|B_3) \geq 1 - \delta$ . Therefore in both cases, we have

$$\mathbb{P}(B_5) \geq \mathbb{P}(B_4|B_3)\mathbb{P}(B_1) \geq (1 - \delta)\mathbb{P}(B_1) \geq (1 - \delta)(1 - \delta_1), \quad (45)$$

where the third inequality uses Lemma 12.  $\square$

## 5 Algorithmic framework and analysis for linear least squares

We now turn our attention to the LLS problem (1) we are interested in solving. Building on the Blendenpik [3] and LSRN [36] techniques, we introduce a generic algorithmic framework for (1) that can employ any rank-revealing factorization of  $SA$ , where  $S$  is a(ny) sketching matrix; we then analyse its convergence.

**Algorithm 1** (Generic Sketching Algorithm). *Given  $A \in \mathbb{R}^{n \times d}$  and  $b \in \mathbb{R}^n$ , set positive integers  $m$  and  $it_{max}$ , and accuracy tolerances  $\tau_a$  and  $\tau_r$ , and an  $m \times n$  random matrix distribution  $\mathcal{S}$ .*

1. *Randomly draw a sketching matrix  $S \in \mathbb{R}^{m \times n}$  from  $\mathcal{S}$ , compute the matrix-matrix product  $SA \in \mathbb{R}^{m \times d}$  and the matrix-vector product  $Sb \in \mathbb{R}^m$ .*
2. *Compute a factorization of  $SA$  of the form,*

$$SA = QR\hat{V}^T, \quad (46)$$

where

- $R = \begin{pmatrix} R_{11} & R_{12} \\ 0 & 0 \end{pmatrix} \in \mathbb{R}^{d \times d}$ , where  $R_{11} \in \mathbb{R}^{p \times p}$  is nonsingular.
  - $Q = (Q_1 \quad Q_2) \in \mathbb{R}^{m \times d}$ , where  $Q_1 \in \mathbb{R}^{m \times p}$  and  $Q_2 \in \mathbb{R}^{m \times (d-p)}$  have orthonormal columns.
  - $\hat{V} = (V_1 \quad V_2) \in \mathbb{R}^{d \times d}$  is an orthogonal matrix with  $V_1 \in \mathbb{R}^{d \times p}$ .
3. *Compute  $x_s = V_1 R_{11}^{-1} Q_1^T Sb$ . If  $\|Ax_s - b\|_2 \leq \tau_a$ , terminate with solution  $x_s$ .*
  4. *Else, iteratively, compute*

$$y_\tau \approx \arg \min_{y \in \mathbb{R}^p} \|Wy - b\|_2, \quad (47)$$

where

$$W = AV_1 R_{11}^{-1}, \quad (48)$$

using LSQR [41] with (relative) tolerance  $\tau_r$  and maximum iteration count  $it_{max}$ .

5. *Return  $x_\tau = V_1 R_{11}^{-1} y_\tau$ .*

**Remark 2.** (i) The factorization  $SA = QR\hat{V}^T$  allows column-pivoted QR, or other rank-revealing factorization, complete orthogonal decomposition ( $R_{12} = 0$ ) and the SVD ( $R_{12} = 0$ ,  $R_{11}$  diagonal). It also includes the usual QR factorisation if  $SA$  is full rank; then the  $R_{12}$  block is absent.

- (ii) Often in implementations, the factorization (46) has  $R = \begin{pmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{pmatrix}$ , where  $R_{22} \approx 0$  and is treated as the zero matrix.
- (iii) For computing  $x_s$  in Step 3, we note that in practical implementations,  $R_{11}$  in (46) is upper triangular, enabling efficient calculation of matrix-vector products involving  $R_{11}^{-1}$ ; then, there is no need to form/calculate  $R_{11}^{-1}$  explicitly.
- (iv) For the solution of (47), we use the termination criterion  $\|y - y_*\|_{W^T W} \leq \tau_r \|y - y_*\|_{W^T W}$  in the theoretical analysis, where  $y_*$  is defined in (50). In practical implementations different termination criteria need to be employed (see Section 6).



## 5.1 Analysis of Algorithm 1

Given problem (1), we denote its minimal Euclidean norm solution as follows

$$x_{*,2} = \arg \min_{x^* \in \mathbb{R}^d} \|x_*\|_2 \quad \text{subject to} \quad \|Ax_* - b\|_2 = \min_x \|Ax - b\|_2. \quad (49)$$

and let

$$y_* = \arg \min_{y \in \mathbb{R}^p} \|Wy - b\|_2, \quad \text{where } W \text{ is defined in (48)}. \quad (50)$$

The following two lemmas provide basic properties of Algorithm 1. Their proofs can be found in the Appendix.

**Lemma 13.**  $W \in \mathbb{R}^{n \times p}$  defined in (48) has full rank  $p$ .

**Lemma 14.** In Algorithm 1, if  $S$  is an  $\epsilon$ -subspace embedding for  $A$  for some  $\epsilon \in (0, 1)$ , then  $p = r$  where  $r$  is the rank of  $A$ .

If the LLS problem (1) has a sufficiently small optimal residual, then Algorithm 1 terminates early in Step 3 with the solution  $x_s$  of the sketched problem  $\min \|SAx - Sb\|_2$ ; then, no LSQR iterations are required.

**Lemma 15** (Explicit Sketching Guarantee). *Given problem (1), suppose that the matrix  $S \in \mathbb{R}^{m \times n}$  in Algorithm 1 is an  $\epsilon$ -subspace embedding for the augmented matrix  $(A \ b)$  for some  $0 < \epsilon < 1$ . Then*

$$\|Ax_s - b\|_2 \leq \frac{1 + \epsilon}{1 - \epsilon} \|Ax_* - b\|_2, \quad (51)$$

where  $x_s$  is defined in Step 3 of Algorithm 1 and  $x_*$  is a(ny) solution of (1).

The proof is similar to the result in [45] that shows that any solution of the sketched problem  $\min_x \|SAx - Sb\|_2$  satisfies (51). For completeness, the proof is included in the appendix.

The following technical lemma is needed in the proof of our next theorem.

**Lemma 16.** Let  $A \in \mathbb{R}^{n \times d}$  and  $V_1 \in \mathbb{R}^{d \times p}$  be defined in Algorithm 1. Then  $\ker(V_1^T) \cap \text{range}(A^T) = \{0\}$ , where  $\ker(V_1^T)$  and  $\text{range}(A^T)$  denote the null space of  $V_1^T$  and range subspace generated by the rows of  $A$ , respectively.

Theorem 8 shows that when the LSQR algorithm in Step 4 converges, Algorithm 1 returns a minimal residual solution of (1).

**Theorem 8** (Implicit Sketching Guarantee). *Given problem (1), suppose that the matrix  $S \in \mathbb{R}^{m \times n}$  in Algorithm 1 is an  $\epsilon$ -subspace embedding for the augmented matrix  $(A \ b)$  for some  $0 < \epsilon < 1$ . If  $y_\tau = y_*$  in Step 4 of Algorithm 1 (by setting  $\tau_r := 0$ ), where  $y_*$  is defined in (50), then  $x_\tau$  in Step 5 satisfies  $x_\tau = x_*$ , where  $x_*$  is a solution of (1).*

*Proof.* Using the optimality conditions (normal equations) for the LLS in (50), and  $y_\tau = y_*$ , we deduce  $W^T W y_\tau = W^T b$ , where  $W$  is defined in (50). Substituting the definition of  $x_\tau$  from Step 5 of Algorithm 1, we deduce

$$(R_{11}^{-1})^T V_1^T A^T A x_\tau = (R_{11}^{-1})^T V_1^T A^T b.$$

Multiplying the last displayed equation by  $R_{11}^T$ , we obtain

$$V_1^T (A^T A x_\tau - A^T b) = 0. \quad (52)$$

It follows from (52) that  $A^T A x_\tau - A^T b \in \ker(V_1^T) \cap \text{range}(A^T)$ . But Lemma 16 implies that the latter set intersection only contains the origin, and so  $A^T A x_\tau - A^T b = 0$ ; this and the normal equations for (1) imply that  $x_\tau$  is an optimal solution of (1).  $\square$

The following technical lemma is needed for our next result; it re-states Theorem 3.2 from [36] in the context of Algorithm 1.

**Lemma 17.** [36] *Given problem (1), let  $x_{*,2}$  be its minimal Euclidean norm solution defined in (49) and  $P \in \mathbb{R}^{d \times p}$ , a nonsingular matrix. Let  $x_\tau := Py_\tau$ , where  $y_\tau$  is assumed to be the minimal Euclidean norm solution of  $\min_{y \in \mathbb{R}^p} \|APy - b\|_2$ . Then  $x_\tau = x_{*,2}$  if  $\text{range}(P) = \text{range}(A^T)$ .*

Theorem 9 further guarantees that if  $R_{12} = 0$  in (46) such as when a complete orthogonal factorization is used, then the minimal Euclidean norm solution of (1) is obtained.

**Theorem 9** (Minimal-Euclidean Norm Solution Guarantee). *Given problem (1), suppose that the matrix  $S \in \mathbb{R}^{m \times n}$  in Algorithm 1 is an  $\epsilon$ -subspace embedding for the augmented matrix  $(A \ b)$  for some  $0 < \epsilon < 1$ . If  $R_{12} = 0$  in (46) and  $y_\tau = y_*$  in Step 4 of Algorithm 1 (by setting  $\tau_r := 0$ ), where  $y_*$  is defined in (50), then  $x_\tau$  in Step 5 satisfies  $x_\tau = x_{*,2}$ , where  $x_{*,2}$  is the minimal Euclidean norm solution (49) of (1).*

*Proof.* The result follows from Lemma 17 with  $P := V_1 R_{11}^{-1}$ , provided  $\text{range}(V_1 R_{11}^{-1}) = \text{range}(A^T)$ . To see this, note that

$$\text{range}(V_1 R_{11}^{-1}) = \text{range}(V_1) = \text{range}((SA)^T),$$

where the last equality follows from  $(SA)^T = V_1 R_{11}^T Q_1^T + V_2 R_{12} Q_1^T$  and  $R_{12} = 0$ . Using the SVD decomposition (5) of  $A$ , we further have

$$\text{range}(V_1 R_{11}^{-1}) = \text{range}(A^T S^T) = \text{range}(V \Sigma U^T S^T) = \text{range}(V \Sigma (SU)^T).$$

Since  $S$  is an  $\epsilon$ -subspace embedding for  $A$ , it is also an  $\epsilon$ -subspace embedding for  $U$  by Lemma 2 and therefore by Lemma 1,  $\text{rank}(SU) = \text{rank}(U) = r$ . Since  $SU \in \mathbb{R}^{m \times r}$  has full column rank, we have that  $\text{range}(V \Sigma (SU)^T) = \text{range}(V) = \text{range}(A^T)$ .  $\square$

Theorem 10 gives an iteration complexity bound for the inner solver in Step 4 of Algorithm 1, as well as particularising this result for a special starting point for which an optimality guarantee can be given. It relies crucially on the quality of the preconditioner provided by the sketched factorization in (46), and its proof, that uses standard LSQR results, is given in the appendix.

**Theorem 10** (Rate of convergence). *Given problem (1), suppose that the matrix  $S \in \mathbb{R}^{m \times n}$  in Algorithm 1 is an  $\epsilon$ -subspace embedding for the augmented matrix  $(A \ b)$  for some  $0 < \epsilon < 1$ . Then:*

(i) *Step 4 of Algorithm 1 takes at most*

$$\tau \leq O\left(\frac{|\log \tau_r|}{|\log \epsilon|}\right) \quad (53)$$

*LSQR iterations to return a solution  $y_\tau$  such that*

$$\|y_\tau - y_*\|_{W^T W} \leq \tau_r \|y_0 - y_*\|_{W^T W}, \quad (54)$$

*where  $y_*$  and  $W$  are defined in (50).*

(ii) *If we initialize  $y_0 := Q^T S b$  for the LSQR method in Step 4, then at termination of Algorithm 1, we can further guarantee that*

$$\|Ax_\tau - b\|_2 \leq \left(1 + \frac{2\epsilon\tau_r}{1-\epsilon}\right) \|Ax_* - b\|_2, \quad (55)$$

*where  $x_\tau$  is computed in Step 5 of Algorithm 1 and  $x_*$  is a solution of (1).*

## 6 The Ski-LLS solver for linear least squares

**SK**etching-for-Linear-Least-Squares (Ski-LLS)<sup>11</sup> is a C++ implementation of Algorithm 1 for solving problem (1). We distinguish two cases based on whether the data matrix  $A$  in (1) is stored as a dense matrix or a sparse matrix.

<sup>11</sup>Available at: <https://github.com/numericalalgorithmsgroup/Ski-LLS>

**Dense matrix input** When  $A$  in (1) is stored as a dense matrix<sup>12</sup>, we employ the following implementation of Algorithm 1. We refer here to the resulting Ski-LLS variant as **Ski-LLS-dense**.

1. In Step 1 of Algorithm 1, we let

$$S = S_h F D, \quad (56)$$

where

- (a)  $D$  is a random  $n \times n$  diagonal matrix with  $\pm 1$  independent entries, as in Definition 10.
  - (b)  $F$  is a matrix representing the normalized Discrete Hartley Transform (DHT), defined as  $F_{ij} = \sqrt{1/n} [\cos(2\pi(i-1)(j-1)/n) + \sin(2\pi(i-1)(j-1)/n)]$ <sup>13</sup>. We use the (DHT) implementation in FFTW 3.3.8<sup>14</sup>.
  - (c)  $S_h$  is an  $s$ -hashing matrix, defined in Definition 5. We use the sparse matrix-matrix multiplication routine in SuiteSparse 5.3.0<sup>15</sup> to compute  $S_h \times (FDA)$ .
2. In Step 2 of Algorithm 1, the default factorization for the sketched matrix  $SA$  is the Randomised Column Pivoted QR (R-CPQR)<sup>16</sup> proposed in [34, 33]. When  $A$  is (assumed to be) full rank, we allow a usual QR factorization as in Blendenpik, using DGEQRF from LAPACK for its computation, with the same default values, except the parameter  $rcond$  (defined below) is absent/not needed.
  3. In Step 3 of Algorithm 1, since  $R_{11}$  from R-CPQR is upper triangular, we do not explicitly compute its inverse, but instead, use back-solve from the LAPACK provided by Intel MKL 2019<sup>17</sup>.
  4. In Step 4 of Algorithm 1, we use the LSQR routine<sup>18</sup> to solve (47), as implemented in LSRN [36]. The solution of (47) terminates when

$$\frac{\|W^T(Wy_k - b)\|}{\|W\| \cdot \|Wy_k - b\|} \leq \tau_r \quad (57)$$

where  $W$  is defined in (48); see Section 6 in [41] for a justification of this termination condition.

The choice of sketching matrix (56) in Step 1 of Ski-LLS-dense is novel. It is a hashing variant of the subsampled DHT (SR-DHT, (63)) choice in Blendenpik. In Theorem 6, we made the case that using hashing – instead of sampling – with the randomised Walsh-Hadamard transform (HRHT) has optimal embedding properties in terms of the sketching dimension  $m$ . Here, we aim to show the numerical gains in running time of using hashing (instead of sampling) with coherence-reduction transformations. We focus on an DHT variant (rather than RHT) as the former is more flexible (in terms of allowing any values of  $n$ ), stable and faster according to experience in [3].

The solver employs the following user-chosen parameters, already introduced in Algorithm 1:  $m$ , number of rows of sketching matrix  $S$  (default is  $1.7d$ );  $s$ , number of nonzero entries per column in  $S$  (default is 1); accuracy tolerances  $\tau_a$  (default is  $10^{-8}$ ) and  $\tau_r$  (default is  $10^{-6}$ ); maximum iteration count  $it_{max}$  (default value is  $10^4$ ).

Two more parameters are needed:

<sup>12</sup>Namely, we assume that sufficiently many entries in  $A$  are nonzero that specialised, sparse numerical linear algebra techniques are ineffective.

<sup>13</sup>Here we use the same transform (DHT) as that in Blendenpik for comparison of other components of Ski-LLS-dense, instead of the Walsh-Hadamard transform defined in Definition 10.

<sup>14</sup>Available at <http://www.fftw.org>.

<sup>15</sup>Available at <https://people.engr.tamu.edu/davis/suitesparse.html>.

<sup>16</sup>The implementation can be found at <https://github.com/flame/hqrrp/>. The original code only has a 32-bit integer interface. We wrote a 64-bit integer wrapper as our code has 64-bit integers.

<sup>17</sup>See <https://software.intel.com/content/www/us/en/develop/tools/oneapi/components/onemkl.html>.

<sup>18</sup>Available at <https://web.stanford.edu/group/SOL/software/lshr/>. We fixed some very minor bugs in the code.

- *rcond* (default value is  $10^{-12}$ ), which is a parameter used in Step 2 of Algorithm 1 as follows. R-CPQR computes (46) as  $SA = Q\tilde{R}\tilde{V}^T$ , which is then used to compute  $\tilde{R}_{11}$ , the upper left  $p \times p$  block of  $\tilde{R}$  by letting  $p = \max \{q : \tilde{r}_{q,q} \geq rcond\}$ , where  $\tilde{r}_{q,q}$  is the  $q$ th diagonal entry of  $\tilde{R}$ .
- *wisdom* (default value is 1). The DHT routine we use is faster with pre-tuning, see Blendenpik [4] for a detailed discussion. If the DHT has been pre-tuned, the user needs to set *wisdom* = 1; otherwise *wisdom* = 0. In all our experiments, the default is to tune the DHT using the simplest tuning mechanism offered by FFTW (which in our experiments has been very efficient). The cost of this tuning is not included in the reported runs here.

**Sparse matrix input** When  $A$  is stored as a sparse matrix<sup>19</sup>, we employ the following implementation of Algorithm 1. We refer here to the resulting Ski-LLS variant as **Ski-LLS-sparse**.

1. In Step 1 of Algorithm 1, we let  $S$  be an  $s$ -hashing matrix, as in Definition 5.
2. In Step 2 of Algorithm 1, we use the sparse QR factorization (SPQR) proposed in [11] and implemented in SuiteSparse.
3. In Step 3 of Algorithm 1, since  $R_{11}$  from SPQR is upper triangular, we do not explicitly compute its inverse, but instead, use the sparse back-substitution routine from SuiteSparse.
4. In Step 4 of Algorithm 1, we use the LSQR routine as implemented in LSRN, extended to include the employment of a sparse preconditioner and sparse numerical linear algebra tools from SuiteSparse. Again, we terminate the solution of (47) when (57) holds.

The solver employs the following user-chosen parameters, already introduced in Algorithm 1:  $m$  (default value is  $1.4d$ ),  $s$  (default value is 2),  $\tau_a$  (default value is  $10^{-8}$ ),  $\tau_r$  (default value is  $10^{-6}$ ),  $it_{max}$  (default value is  $10^4$ ). We also need

- *ordering* (default value 2) which is a parameter of the SPQR routine that influences the permutation matrix  $\hat{V}$  and the sparsity of  $R$ .<sup>20</sup>
- *rcond<sub>thres</sub>* (default value  $10^{-10}$ ), which checks the conditioning of  $R_{11}$  in (46) computed by SPQR. If the condition number  $\kappa(R_{11}) \geq 1/rcond_{thres}$ , we use the perturbed back-solve for upper triangular linear systems involving  $R_{11}$  with *perturb* (default value  $10^{-10}$ ). In particular, any back-solve involving  $R_{11}$  or its transpose is modified as follows: when divisions by a diagonal entry  $r_{ii}$  of  $R_{11}$  is required ( $1 \leq i \leq p$ ), we divide by  $r_{ii} + perturb$  instead<sup>21</sup>.

A few comments regarding our Ski-LLS implementation are in order, that connect earlier theoretical developments to our implementation choices.

**Subspace embedding properties** Our analysis of Algorithm 1 in the previous section relies crucially on the sketching matrix  $S$  being a subspace embedding. Theorem 6 does not apply specifically to the coherence-reduction DHT we use in Ski-LLS-dense, but it inspired us to explore the practical efficiency of using hashing instead of subsampling in the random transform used in Blendenpik. For sparse matrices, Theorem 3 guarantees the desired embedding properties of  $s$ -hashing matrices when  $A$  has low coherence. However as Figures 18 and 19 indicate, these results may not be tight in that  $s$ -hashing matrices with  $s > 1$  and  $m = \mathcal{O}(d)$  also seem to embed correctly matrices  $A$  with high(er) coherence. In both the dense and sparse Ski-LLS,

<sup>19</sup>Namely, the user stores the data in a sparse matrix format, which is maintained through the algorithm’s run for efficient linear algebra procedures.

<sup>20</sup>Note that this is slightly different from the SPQR default, which uses COLAMD if  $m_2 \leq 2 \cdot n_2$ ; otherwise switches to AMD. Let  $f$  be the flops for  $\text{chol}((S^*P)^*(S^*P))$  with the ordering  $P$  found by AMD. Then if  $f/\text{nnz}(R) \geq 500$  and  $\text{nnz}(R)/\text{nnz}(S) \geq 5$  then use METIS, and take the best ordering found (AMD or METIS), where typically  $m_2 = m$ ,  $n_2 = n$  for  $SA \in \mathbb{R}^{m \times n}$ . By contrast, Ski-LLS by default always uses the AMD ordering.

<sup>21</sup>This is a safeguard when SPQR fails to detect the rank of  $A$ . This happens infrequently [11].

numerical calibration is used to determine the default value of  $m$  – and in the sparse case also  $s$  – such that  $\epsilon$ -subspace embedding of  $A$  is achieved with sufficiently high probability (namely, for all the matrices in the calibration set); see 6 for details.

Even if the distribution of hashing matrices provides an oblivious subspace embedding, there is still a (small) positive probability that for a given  $A$ , a randomly drawn  $S$  fails to sketch  $A$  accurately. Then, if  $A$  has full rank, Ski-LLS can still compute an accurate solution of (1) as the preconditioner  $V_1 R_{11}^{-1}$  is a nonsingular matrix. When  $A$  is rank deficient, and  $m$  has been calibrated as mentioned above, we have not found detrimental evidence of solver failure in that case, just possibly some loss of accuracy; see for example, the Florida Collection test problem ‘landmark’ in Table 2 where the sketched input matrix has lower rank, indicating failure of subspace embedding property, and resulting in slightly lower accuracy in the final residual value.

**Sketching size tuning to ensure subspace embedding properties** We use numerical calibration to determine the default value of  $m$  for Ski-LLS-dense (with and without R-CPQR), as well as the values of both  $m$  and  $s$  for Ski-LLS-sparse. These choices are made such that  $\epsilon$ -subspace embedding properties are achieved with sufficiently high probability for all the random matrices in the respective calibration set (see Test Set 1 and 2 next). Note for example, the typical U-shaped curve in Figure 14: as  $\gamma := m/d$  grows, we have better subspace embedding properties with smaller  $\epsilon$  and hence, fewer LSQR iterations (due to (53)). However the cost of sketching in Ski-LLS’ Step 1 and of the factorization in Step 2 will grow as  $m$  grows. Thus a trade-off is achieved for an approximately median value of  $m$ ; see Appendices B (dense input) and C (sparse input) for the plots of the calibration results for  $m$ , and  $s$  when needed.

**Approximate factorization in Step 2** In both dense and sparse Ski-LLS, the factorization in Step 2 is not guaranteed to be accurate when  $A$  is rank-deficient. This is because R-CPQR, like column-pivoted QR [16], does not guarantee detection of rank. However, the numerical rank is typically correctly determined, in the sense that using the procedure described in the definition of the parameter  $rcond$ , the factorization of  $SA$  will be accurate approximately up to  $rcond$  error. Similarly, SPQR only performs heuristical rank detection, without guarantees. Our encouraging numerical results testing for accuracy, alleviate these shortcomings; see Section 7.3.2.

## 7 Numerical experiments

### 7.1 Test sets

The following linear least squares problems of the form (1) are used to test and benchmark Ski-LLS.

**The vector  $b$**  In each test problem, the vector  $b \in \mathbb{R}^n$  in (1) is chosen to be a vector of all ones.

**The input matrix  $A$**

1. (Test Set 1) Three different types of **random dense matrices** as used by Avron et al [3] to compare Blendenpik with the LAPACK least square solver. They have different ‘non-uniformity’ of the rows.
  - (a) Incoherent dense type, defined by (5) with  $r = d$ , and where  $U \in \mathbb{R}^{n \times d}$ ,  $V \in \mathbb{R}^{d \times d}$  are generated by orthogonalising columns of two independent matrices with i.i.d.  $N(0,1)$  entries, and  $\Sigma \in \mathbb{R}^{d \times d}$  is a diagonal matrix with diagonal entries equally spaced from 1 to  $10^6$ .
  - (b) Semi-coherent dense type, defined by

$$A = \begin{pmatrix} B & 0 \\ 0 & I_{\frac{d}{2}} \end{pmatrix} + 10^{-8} J \in \mathbb{R}^{n \times d}, \quad (58)$$

where  $B$  is an incoherent dense matrix as in 1(a) and  $J \in \mathbb{R}^{n \times d}$  is a matrix of all ones.

(c) Coherent dense type, defined by

$$A = \begin{pmatrix} I_{d \times d} \\ 0 \end{pmatrix} + 10^{-8} J \in \mathbb{R}^{n \times d}, \quad (59)$$

where  $J$  is again a matrix of all ones.

2. (Test Set 2) The following are three different types of **random sparse matrices** with different ‘non-uniformity’ of rows.

(a) Incoherent sparse type, defined by

$$A = \text{sprandn}(n, d, 0.01, 1e - 6), \quad (60)$$

where ‘sprandn’ is a command in MATLAB that generates a matrix with approximately  $0.01nd$  normally distributed non-zero entries and a condition number approximately equal to  $10^6$ .

(b) Semi-coherent sparse type, defined by

$$A = \hat{D}^5 B, \quad (61)$$

where  $B \in \mathbb{R}^{n \times d}$  is an incoherent sparse matrix defined in (60) and  $\hat{D}$  is a diagonal matrix with independent  $N(0, 1)$  entries on the diagonal.

(c) Coherent sparse type, defined by

$$A = \hat{D}^{20} B, \quad (62)$$

where  $B \in \mathbb{R}^{n \times d}$ ,  $\hat{D}$  are the same as in (61).

3. (Test Set 3) A total of 181 matrices from the **Florida (SuiteSparse) matrix collection** [12] such that if the matrix is under-determined, we transpose it to make it over-determined; and once transposed,  $A \in \mathbb{R}^{n \times d}$  must have  $n \geq 30000$  and  $n \geq 2d$ .

**Remark 3.** Note that the ‘coherence’ in Test Sets 1 and 2 above is slightly different than the earlier notion of coherence  $\mu$  in (6). In particular, the former is a more general concept, indicating that the leverage scores of  $A$  are somewhat non-uniform, while the latter only predicates the value of the maximum leverage score. Although ‘coherent dense/sparse’  $A$  tends to have higher values of  $\mu(A)$  than that of ‘incoherent dense/sparse’  $A$ , the value of  $\mu(A)$  may be similar for semi-coherent and coherent test matrices (namely, it is 1 in Test Sets 1 and 2), while their differences lie in that the row norms of  $A$  (and hence of  $U$  in (5)) tend to be more non-uniform than in the semi-coherent case. The coherence terminology for the Sets 1 and 2 is derived from *Blendenpik* and we have maintained it for ease of comparison with earlier results.

## 7.2 Solvers and their parameters

For **dense input**, we compare Ski-LLS-dense to the state-of-the-art sketching solver *Blendenpik* [3]<sup>22</sup>. The row dimension of the sketching matrix in *Blendenpik* is set to  $m = 2.2d$ , which we obtained by calibrating *Blendenpik* as shown in Appendix B. Note that we do *not* use the default value of  $m$  in *Blendenpik* which is 4; this default choice would yield significantly longer runtimes. We also set  $\tau_r = 10^{-6}$ ,  $it_{max} = 10^4$  and  $wisdom = 1$ . The same wisdom data file as in Ski-LLS is used.

For **sparse input**, we compare Ski-LLS-sparse to the following solvers:

<sup>22</sup> Available at <https://github.com/haimav/Blendenpik>. For the sake of a fair comparison, we wrote a C interface and use the same LSQR routine as in Ski-LLS.

1. HSL\_MI35 (LS\_HSL), that uses an incomplete Cholesky factorization of  $A$  to compute a preconditioner for problem (1), before using LSQR.<sup>23</sup> This is a state of the art preconditioned, iterative solver for sparse LLS problems [18, 17]. We use  $\tau_r = 10^{-6}$  and  $it_{max} = 10^4$ .
2. SPQR\_SOLVE (LS\_SPQR), that uses SPQR from Suitesparse to compute a sparse QR factorization of  $A$ , which is exploited to solve (1) directly.<sup>24</sup> This is a state of the art direct solver for sparse LLS problems [11].
3. LSRN, that uses the framework of Algorithm 1, with  $S$  having i.i.d.  $N(0, 1/\sqrt{m})$  entries in Step 1; SVD factorization from Intel LAPACK of the matrix  $SA$  in Step 2; the same LSQR routine as Ski-LLS in Step 4.<sup>25</sup> LSRN has been shown to be efficient for possibly rank-deficient dense and sparse LLS problems in a parallel computing environment [36]. However, parallel techniques are outside the scope of this paper and we therefore run LSRN serially. The parameters are chosen to be  $m = 1.1d$  (calibrated, random test sets) while  $m$  is set to LSRN default otherwise;  $\tau_r = 10^{-6}$ ,  $it_{max} = 10^4$ .

For each solver, the algorithm parameters are set to their default values unless otherwise specified above.

**Compilation and environment for timed experiments** For our numerical studies, unless otherwise mentioned, we use Intel C compiler `icc` with optimisation flag `-O3` to compile all the C code, and Intel Fortran compiler `ifort` with `-O3` to compile Fortran-based code. All code has been compiled in sequential mode and linked with sequential dense/sparse linear algebra libraries provided by Intel MKL, 2019 and Suitesparse 5.3.0. The machine used has Intel(R) Xeon(R) CPU E5-2667 v2 @ 3.30GHz with 8GB RAM. All reported times are wall clock times in seconds.

**Performance profiles** Performance profiles [13] are a popular tool when benchmarking software. In the performance profile to be encountered here, for some of the results, we plot the runtime ratio against the fastest solver on the horizontal axis, in  $\log 2$  scale. For each runtime ratio  $a$ , we have the ratio of problems in the test set  $b$  on the vertical axis such that for a particular solver, the runtime ratio against the best solver is within  $a$  for  $b$  percent of the problems in the test set. For example, the intersect between the performance curve and the y-axis gives the ratio of the problems in the test set such that a particular solver is the fastest.

Given a problem  $A$  from the test set, let  $(r_1, r_2, r_3, r_4)$  be the residual values at the solutions computed by the four solvers, that we compare in the sparse case. And let  $r = \min_{i \in \{1, \dots, 4\}} r_i$ . A solver is declared as having failed on this particular problem if one of the following two conditions holds

1.  $r_i > (1 + \tau_r)r$  and  $r_i > r + \tau_a$ , implying that the residual at the computed solution is neither relatively or absolutely sufficiently close to the residual at the best solution found by (one of the) remaining solvers.
2. The solver takes more than 800 wall clock seconds to compute a solution.

When a solver fails, we set the runtime of the solver to be 9999 seconds on the corresponding problem. Thus a large runtime (ratio) could be due to either an inaccurate or an inefficient solution obtained by the respective solver. We note that for all successful solvers, the runtime is bounded above by 800 seconds so that there can be no confusion whether a solver is successful or not.

<sup>23</sup>See [http://www.hsl.rl.ac.uk/specs/hsl\\_mi35.pdf](http://www.hsl.rl.ac.uk/specs/hsl_mi35.pdf) for a full specification. For a fair comparison, we wrote a C interface and used the same LSQR routine as in Ski-LLS. We also disabled the pre-processing of the data as it was not done for the other solvers. We found that using no scaling and no ordering was more efficient than the default scaling and ordering, and so we chose the former. It may thus be possible that the performance of HSL may improve, however [18] experimented with the use of different scalings and orderings, providing some evidence that the improvement will not be significant.

<sup>24</sup>Available at <https://people.engr.tamu.edu/davis/suitesparse.html>.

<sup>25</sup>Note that LSRN does not contain the Step 3.

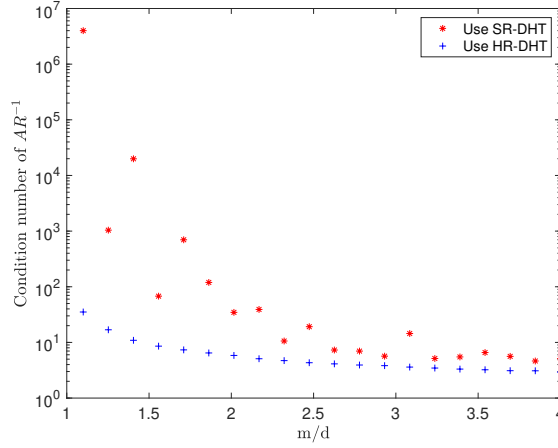


Figure 1: Hashing combined with a randomised Discrete Hartley Transform (DHT) produces more accurate sketched matrix  $SA$  for a given  $m/d$  ratio comparing to sampling combined with a randomised DHT; the accuracy of the sketch is reflected in the quality of the preconditioner  $R$  constructed from the matrix  $SA$ , see (87).

### 7.3 Numerical results

We now present our numerical findings.

#### 7.3.1 Numerical illustrations

The advantages of two key choices of our implementation are illustrated next. We use Matlab for these illustrations as they do not involve runtime comparisons, nor the running of Ski-LLS.

**Using hashing instead of sampling with coherence-reduction transformations** In Figure 1, we generate a random coherent dense matrix  $A \in \mathbb{R}^{4000 \times 400}$  as in (59). For each  $m/d$  (where  $d = 400$ ), we sketch  $A$  using an HR-DHT  $S \in \mathbb{R}^{m \times n}$  defined in (56) and using SR-DHT as in Blendenpik, defined by

$$S = S_s F D, \quad (63)$$

where  $S_s \in \mathbb{R}^{m \times n}$  is a sampling matrix, whose individual rows contain a single non-zero entry at a random column with value 1;  $F, D$  are defined the same as in (56). We then compute a QR factorization without pivoting, of each sketch  $SA = QR$ , and the condition number of  $AR^{-1}$ .

We see that using hashing instead of sampling in randomised DHT allows the use of a smaller  $m$  to reach a given preconditioning quality.

**Using  $s$ -hashing with  $s > 1$  to sketch sparse input** In Figure 2, we let  $A \in \mathbb{R}^{4000 \times 400}$  be a random incoherent sparse matrix as in (60). In Figure 3,  $A$  is defined as in (58) but with  $B \in \mathbb{R}^{n \times d}$  being a random incoherent sparse matrix<sup>26</sup>. Comparing Figure 2 with Figure 3, we see that using  $s$ -hashing matrices with  $s > 1$  is essential in order to obtain a good preconditioner.

<sup>26</sup>We use this type of random sparse matrix instead of one of the types in Test Set 2 as this matrix better showcases the failure of 1-hashing.



	Ski-LLS-Sparse	Ski-LLS-Dense	Blendenpik	LSRN
lp_ship12l	1.28E-09	9.84E-10	NaN	1.01E-08
Franz1	3.99E-09	1.55E-09	9405.535	5.18E-08
GL7d26	3.11E-09	2.33E-09	NaN	1.08E-08
cis-n4c6-b2	1.24E-14	3.82E-14	174.4694	5.54E-15
lp_modszk1	4.93E-09	4.83E-09	NaN	5.48E-08
rel5	1.29E-10	1.92E-10	NaN	1.13E-09
ch5-5-b1	2.20E-10	1.07E-10	11.49551	7.85E-10
n3c5-b2	5.56E-15	1.36E-16	86.50651	1.41E-15
ch4-4-b1	1.48E-12	4.30E-12	283.7672	9.77E-15
n3c5-b1	0	0	9.185009	3.99E-14
n3c4-b1	0	0	43.42909	0
connectus	5.11E-03	NaN	NaN	5.11E-03
landmark	2.83E-10	6.81E-16	NaN	6.55E-11
cis-n4c6-b3	2.83E-09	1.34E-09	2400.632	8.81E-09

Table 2: Difference of residual values generated by the solvers compared to the residual of the truncated-SVD solution, for a range of rank-deficient problems from the Florida matrix collection [12]. The matrices, each given in sparse format, are converted into dense format before applying a dense solver such as Blendenpik. We find that Ski-LLS-dense and Ski-LLS-sparse achieve good residual accuracy on rank-deficient problems, as well as LSRN. Blendenpik is not designed for rank-deficient problems and either returns a large residual or encounters numerical issues/out of memory.

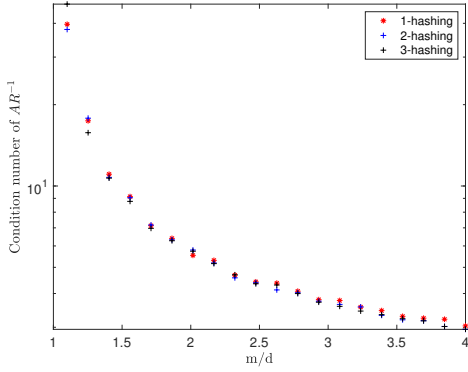


Figure 2: When the data matrix  $A$  is an ill-conditioned sparse Gaussian, using 1, 2, 3-hashing produces similarly good preconditioners.

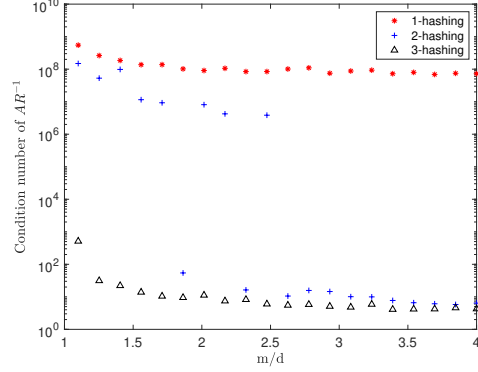


Figure 3: When the data matrix  $A$  has higher coherence, using  $s$ -hashing with  $s > 1$  is crucial to obtaining an acceptable preconditioner.

### 7.3.2 Residual accuracy obtained by Ski-LLS for some rank-deficient problems

Here we check the final residual accuracy (objective decrease in (1)) obtained by Ski-LLS, as an illustration/check that the implementation does not depart much from the theoretical guarantees in Section 5. We choose 14 matrices in the Florida matrix collection that are rank-deficient (Table 6). We use LAPACK’s SVD-based LLS solver (SVD), LSRN, Blendenpik, Ski-LLS-dense and Ski-LLS-sparse on these problems with the residual shown in Table 2. Both Ski-LLS-dense and Ski-LLS-sparse have very good residual accuracy compared to the SVD-based LAPACK solver; also, as expected, Blendenpik is unable to accurately solve rank-deficient problems.

Furthermore, in our large scale numerical study with the Florida matrix collection in Section 7.4, we have also compared the residual values and found that the solution of Ski-LLS-sparse is no-less accurate than the state-of-the-art sparse solvers LS\_SPQR and LS\_HSL.

### 7.3.3 Runtime performance of Ski-LLS on random, full-rank and dense input (Test Set 1)

We compare Ski-LLS-dense (with and without R-CPQR and with default settings) with (calibrated) Blendenpik (as above) on dense matrix input in (1). For each of the sizes shown (on the horizontal axis) in Figures 4, 5 and 6, we generate three types of random matrices as in

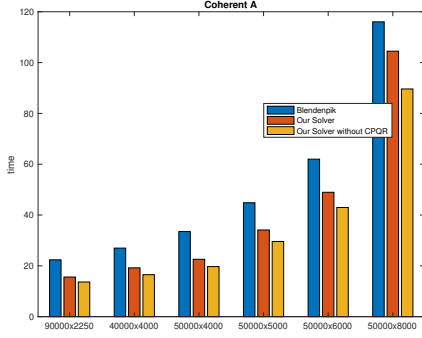


Figure 4: Time (in seconds) taken by solvers applied to problem (1) with coherent, dense and random matrices of various sizes (x-axis).

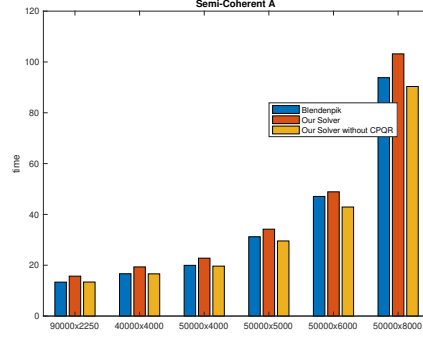


Figure 5: Time (in seconds) taken by solvers applied to problem (1) with semi-coherent, dense and random matrices of various sizes (x-axis).

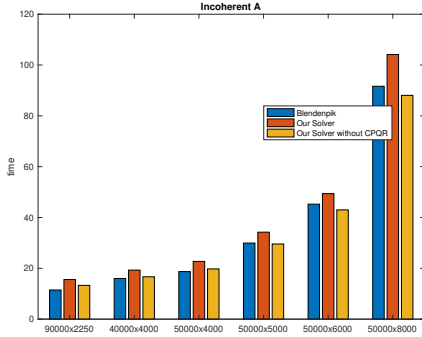


Figure 6: Time (in seconds) taken by solvers applied to problem (1) with incoherent, dense and random matrices of various sizes (x-axis).

Test Set 1. The residual values obtained by the different solvers coincide to six significant digits, indicating all three solvers give an accurate solution of (1).

Our results in Figures 4, 5 and 6 showcase the improvements obtained when using hashing instead of subsampling in the DHT sketching, yielding improved runtimes for Ski-LLS-dense without R-CPQR compared to Blendenpik, especially when the input matrix  $A$  is of coherent type, as in (59). We also see that Ski-LLS-dense with R-CPQR is as fast as Blendenpik on these full-rank and dense problems, while also being able to solve rank-deficient problems (see Table 2).

#### 7.3.4 Runtime performance of Ski-LLS on random, full-rank and sparse input (Test Set 2)

Figures 7, 8 and 9 show the performance of Ski-LLS-sparse compared to LS\_HSL, LS\_SPQR and LSRN on sparse random matrices of different types and sizes. We see Ski-LLS can be up to 10 times faster on this class of data matrix. Tables 3, 4 and 5 in the appendix record the residual values at the solution. We see that our solver can be more accurate than the state-of-the-art LS\_HSL and LS\_SPQR, although LSRN is the most accurate.

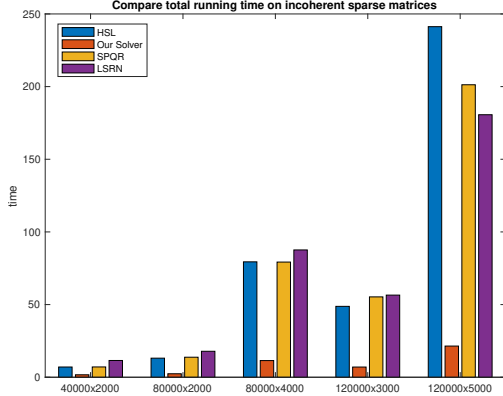


Figure 7: Runtime comparison of Ski-LLS with LS\_HSL, LS\_SPQR and LSRN on randomly generated incoherent sparse matrices of different sizes.

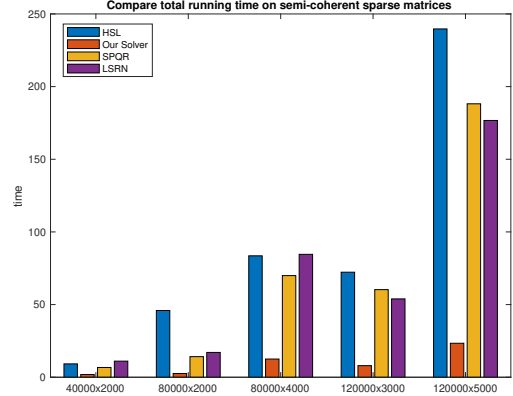


Figure 8: Runtime comparison of Ski-LLS with LS\_HSL, LS\_SPQR and LSRN on randomly generated semi-coherent sparse matrices of different sizes.

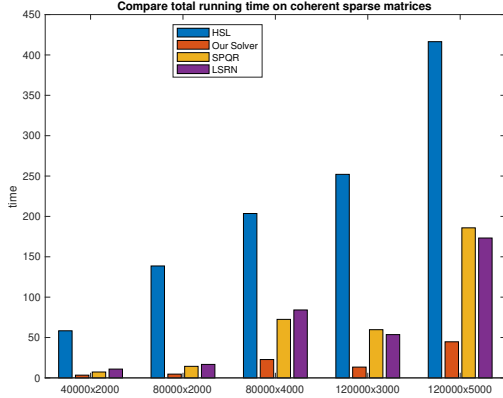


Figure 9: Runtime comparison of Ski-LLS with LS\_HSL, LS\_SPQR and LSRN on randomly generated coherent sparse matrices of different sizes.

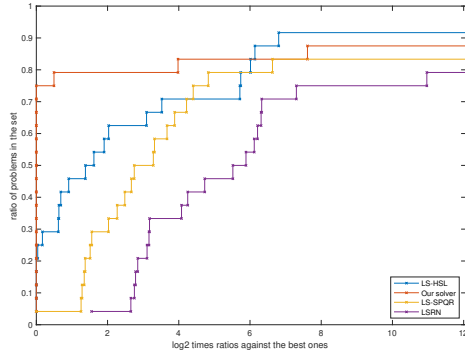


Figure 10: Performance profile comparison of Ski-LLS with LSRN, LS\_HSL and LS\_SPQR for all matrices  $A \in \mathbb{R}^{n \times d}$  in the Florida matrix collection with  $n \geq 30d$ .

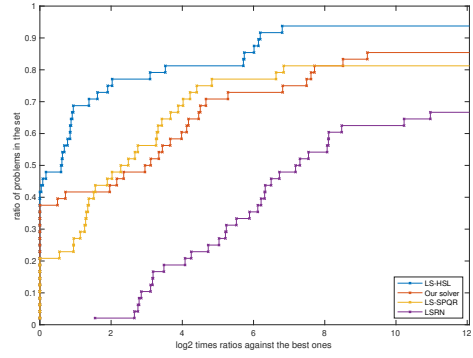


Figure 11: Performance profile comparison of Ski-LLS with LSRN, LS\_HSL and LS\_SPQR for all matrices  $A \in \mathbb{R}^{n \times d}$  in the Florida matrix collection with  $n \geq 10d$ .

## 7.4 Benchmarking Ski-LLS on the Florida Matrix Collection (Test Set 3)

We present the results of benchmarking Ski-LLS-sparse against LSRN, SPQR and HSL on the large scale Test Set 3. Figure 22 shows these results for the entire collection considered here, namely the matrices or their transpose with  $n \geq 2d$ ; we find that neither of the two sketching solvers, Ski-LLS and LSRN, are sufficiently competitive with the direct solver SPQR and the fastest solver, preconditioned iterative one in HSL. This is perhaps unsurprising as we would only really expect sketching solvers to improve state of the art ones when sketching can play a role, and for that either substantially more rows must be present, or some form of difficulty/structure. Indeed, the situation substantially changes when we subselect from the entire Florida set the performance of solvers on significantly overdetermined problems, or ill-conditioned/‘difficult ones’ or moderately sparse ones.

**Highly over-determined matrices in the Florida Matrix Collection** Figure 10 shows Ski-LLS is the fastest in 75% of problems in the Florida matrix collection with  $n \geq 30d$ , outperforming the iterative HSL code (with incomplete Cholesky preconditioning).

**On moderately over-determined Florida collection problems** Figure 11 shows LS\_HSL is the fastest for the largest percentage of problems in the Florida matrix collection with  $n \geq 10d$ . However, Ski-LLS is still competitive and noticeably faster than LSRN. As the proportion of rows versus columns further decreases, however, the profiles would approach those in Figure 22.

**Effect of condition number** Many of the matrices in the Florida matrix condition are well-conditioned. Thus LSQR (without preconditioning) converges in few iterations. Then the benefits of the high quality preconditioner computed by Ski-LLS are lost to the incomplete one in LS\_HSL. However, if the problems are not well-conditioned/‘difficult’, the situation reverses.

Figure 12 shows Ski-LLS is the fastest in more than 50% of the moderately over-determined problems ( $n \geq 10d$ ) if we only consider problems such that it takes LSQR more than 5 seconds to solve.

**Effect of sparsity** Figure 13 shows Ski-LLS is highly competitive, being the fastest in all but one moderately over-determined problems with moderate sparsity ( $\text{nnz}(A) \geq 0.01nd$ ). The effect of increasing the sparsity by allowing ( $\text{nnz}(A) \geq 0.001nd$ ) is shown in Figure 23.

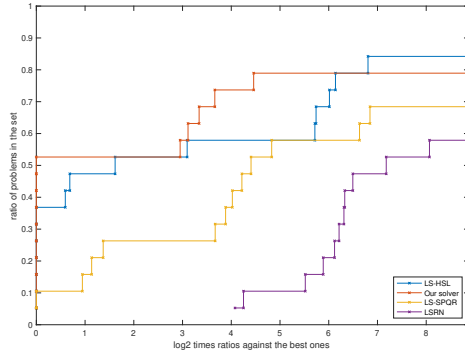


Figure 12: Performance profile comparison of Ski-LLS with LSRN, LS\_HSL and LS\_SPQR for all matrices  $A \in \mathbb{R}^{n \times d}$  in the Florida matrix collection with  $n \geq 10d$  and the unpreconditioned LSQR takes more than 5 seconds to solve.

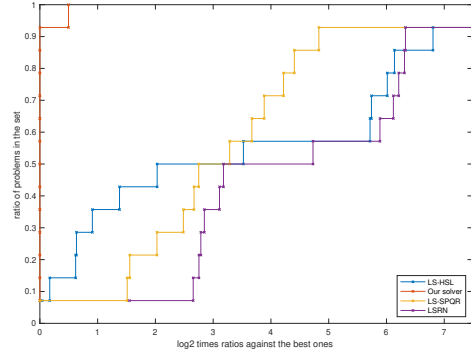


Figure 13: Performance profile comparison of Ski-LLS with LSRN, LS\_HSL and LS\_SPQR for all matrices  $A \in \mathbb{R}^{n \times d}$  in the Florida matrix collection with  $n \geq 10d$  and  $\text{nnz}(A) \geq 0.01nd$ .

## References

- [1] N. Ailon and B. Chazelle. Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform. In *STOC'06: Proceedings of the 38th Annual ACM Symposium on Theory of Computing*, pages 557–563. ACM, New York, 2006.
- [2] N. Ailon and E. Liberty. An almost optimal unrestricted fast Johnson-Lindenstrauss transform. *ACM Trans. Algorithms*, 9(3):Art. 21, 1–12, 2013.
- [3] H. Avron, P. Maymounkov, and S. Toledo. Blendenpik: supercharging Lapack’s least-squares solver. *SIAM J. Sci. Comput.*, 32(3):1217–1236, 2010.
- [4] H. Avron, E. Ng, and S. Toledo. Using perturbed  $QR$  factorizations to solve linear least-squares problems. *SIAM J. Matrix Anal. Appl.*, 31(2):674–693, 2009.
- [5] A. Björck. *Numerical methods for least squares problems*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1996.
- [6] J. Bourgain, S. Dirksen, and J. Nelson. Toward a unified theory of sparse dimensionality reduction in Euclidean space. *Geom. Funct. Anal.*, 25(4):1009–1088, 2015.
- [7] L. Chen, S. Zhou, and J. Ma. Stable sparse subspace embedding for dimensionality reduction. *Knowledge-Based Systems*, 195:105639, 2020.
- [8] K. L. Clarkson and D. P. Woodruff. Low-rank approximation and regression in input sparsity time. *J. ACM*, 63(6):Art. 54, 1–45, 2017.
- [9] M. B. Cohen. Nearly tight oblivious subspace embeddings by trace inequalities. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 278–287. ACM, New York, 2016.
- [10] Y. Dahiya, D. Konomis, and D. P. Woodruff. An empirical evaluation of sketching for numerical linear algebra. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '18*, pages 1292–1300, New York, NY, USA, 2018. Association for Computing Machinery.
- [11] T. A. Davis. Algorithm 915, SuiteSparseQR: multifrontal multithreaded rank-revealing sparse QR factorization. *ACM Trans. Math. Software*, 38(1):Art. 1, 1–22, 2011.
- [12] T. A. Davis and Y. Hu. The University of Florida sparse matrix collection. *ACM Trans. Math. Software*, 38(1):Art. 1, 1–25, 2011.

- [13] E. D. Dolan and J. J. Moré. Benchmarking optimization software with performance profiles. *Mathematical programming*, 91(2):201–213, 2002.
- [14] P. Drineas, M. W. Mahoney, and S. Muthukrishnan. Sampling algorithms for l2 regression and applications. In *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithm*, SODA '06, page 1127–1136, USA, 2006. Society for Industrial and Applied Mathematics.
- [15] C. Freksen, L. Kamma, and K. G. Larsen. Fully understanding the hashing trick. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 5394–5404, Red Hook, NY, USA, 2018. Curran Associates Inc.
- [16] G. H. Golub and C. F. Van Loan. *Matrix computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, third edition, 1996.
- [17] N. Gould and J. Scott. The state-of-the-art of preconditioners for sparse linear least-squares problems: the complete results. Technical report, STFC Rutherford Appleton Laboratory, 2015. Available at [ftp://cuter.rl.ac.uk/pub/nimg/pubs/GoulScot16b\\_toms.pdf](ftp://cuter.rl.ac.uk/pub/nimg/pubs/GoulScot16b_toms.pdf).
- [18] N. Gould and J. Scott. The state-of-the-art of preconditioners for sparse linear least-square problems. *ACM Trans. Math. Software*, 43(4):Art. 36, 1–35, 2017.
- [19] R. M. Gower and P. Richtárik. Randomized iterative methods for linear systems. *SIAM J. Matrix Anal. Appl.*, 36(4):1660–1690, 2015.
- [20] M. A. Iwen, D. Needell, E. Rebrova, and A. Zare. Lower Memory Oblivious (Tensor) Subspace Embeddings with Fewer Random Bits: Modewise Methods for Least Squares. *SIAM J. Matrix Anal. Appl.*, 42(1):376–416, 2021.
- [21] C. Iyer, H. Avron, G. Kollias, Y. Ineichen, C. Carothers, and P. Drineas. A randomized least squares solver for terabyte-sized dense overdetermined systems. *J. Comput. Sci.*, 36:100547, 2019.
- [22] C. Iyer, C. Carothers, and P. Drineas. Randomized sketching for large-scale sparse ridge regression problems. In *2016 7th Workshop on Latest Advances in Scalable Algorithms for Large-Scale Systems (ScalA)*, pages 65–72, 2016.
- [23] M. Jagadeesan. Understanding sparse JL for feature hashing. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
- [24] W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In *Conference in modern analysis and probability (New Haven, Conn., 1982)*, volume 26 of *Contemp. Math.*, pages 189–206. Amer. Math. Soc., Providence, RI, 1984.
- [25] N. Kahale. Least-squares regressions via randomized Hessians. *arXiv e-prints*, page arXiv:2006.01017, June 2020.
- [26] J. Lacotte and M. Pilanci. Faster Least Squares Optimization. *arXiv e-prints*, page arXiv:1911.02675, Nov. 2019.
- [27] J. Lacotte and M. Pilanci. Optimal Randomized First-Order Methods for Least-Squares Problems. *arXiv e-prints*, page arXiv:2002.09488, Feb. 2020.
- [28] S. Liu, T. Liu, A. Vakilian, Y. Wan, and D. P. Woodruff. Extending and Improving Learned CountSketch. *arXiv e-prints*, page arXiv:2007.09890, July 2020.
- [29] N. Loizou. Randomized Iterative Methods for Linear Systems: Momentum, Inexactness and Gossip. *arXiv e-prints*, page arXiv:1909.12176, Sept. 2019.
- [30] N. Loizou and P. Richtárik. Convergence analysis of inexact randomized iterative methods. *SIAM J. Sci. Comput.*, 42(6):A3979–A4016, 2020.

- [31] M. Lopes, S. Wang, and M. Mahoney. Error estimation for randomized least-squares algorithms via the bootstrap. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3217–3226. PMLR, 10–15 Jul 2018.
- [32] M. W. Mahoney. Randomized algorithms for matrices and data. *Found. Trends Mach. Learn.*, 3(2):123–224, Feb. 2011.
- [33] P.-G. Martinsson. Blocked rank-revealing QR factorizations: How randomized sampling can be used to avoid single-vector pivoting. *arXiv e-prints*, page arXiv:1505.08115, May 2015.
- [34] P.-G. Martinsson, G. Quintana Ortí, N. Heavner, and R. van de Geijn. Householder QR factorization with randomization for column pivoting (HQRFP). *SIAM J. Sci. Comput.*, 39(2):C96–C115, 2017.
- [35] X. Meng and M. W. Mahoney. Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. In *STOC’13—Proceedings of the 2013 ACM Symposium on Theory of Computing*, pages 91–100. ACM, New York, 2013.
- [36] X. Meng, M. A. Saunders, and M. W. Mahoney. LSRN: a parallel iterative solver for strongly over- or underdetermined systems. *SIAM J. Sci. Comput.*, 36(2):C95–C118, 2014.
- [37] J. Nelson and H. L. Nguyen. OSNAP: faster numerical linear algebra algorithms via sparser subspace embeddings. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science—FOCS 2013*, pages 117–126. IEEE Computer Soc., Los Alamitos, CA, 2013.
- [38] J. Nelson and H. L. Nguyen. Sparsity lower bounds for dimensionality reducing maps. In *STOC’13—Proceedings of the 2013 ACM Symposium on Theory of Computing*, pages 101–110. ACM, New York, 2013.
- [39] J. Nelson and H. L. Nguyen. Lower bounds for oblivious subspace embeddings. In *Automata, languages, and programming. Part I*, volume 8572 of *Lecture Notes in Comput. Sci.*, pages 883–894. Springer, Heidelberg, 2014.
- [40] J. Nocedal and S. J. Wright. *Numerical optimization*. Springer Series in Operations Research and Financial Engineering. Springer, New York, second edition, 2006.
- [41] C. C. Paige and M. A. Saunders. LSQR: an algorithm for sparse linear equations and sparse least squares. *ACM Trans. Math. Software*, 8(1):43–71, 1982.
- [42] V. Rokhlin and M. Tygert. A fast randomized algorithm for overdetermined linear least-squares regression. *Proc. Natl. Acad. Sci. USA*, 105(36):13212–13217, 2008.
- [43] T. Sarlos. Improved approximation algorithms for large matrices via random projections. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS’06)*, pages 143–152, 2006.
- [44] J. A. Tropp. Improved analysis of the subsampled randomized Hadamard transform. *Adv. Adapt. Data Anal.*, 3(1-2):115–126, 2011.
- [45] D. P. Woodruff. Sketching as a tool for numerical linear algebra. *Found. Trends Theor. Comput. Sci.*, 10(1-2):1–157, 2014.
- [46] R. Zhu. Gradient-based sampling: An adaptive importance sampling for least-squares. In *Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS’16*, pages 406–414, Red Hook, NY, USA, 2016. Curran Associates Inc.

## A Additional proofs

### A.1 Proof of results in Section 2

*Proof of Lemma 1.* Let  $B \in \mathbb{R}^{n \times k}$ . By rank-nullity theorem,  $\text{rank}(B) + \dim \ker(B) = \text{rank}(SB) + \dim \ker(SB) = k$ . Clearly,  $\dim \ker(SB) \geq \dim \ker(B)$ . If the previous inequality is strict, then there exists  $z \in \mathbb{R}^k$  such that  $\|SBz\|_2 = 0$  and  $\|Bz\|_2 > 0$ , contradicting the assumption that  $S$  is an  $\epsilon$ -subspace embedding for  $B$  according to (4).  $\square$

*Proof of Lemma 2.* (i) Let  $A = U\Sigma V^T$  be defined as in (5). If  $S$  is an  $\epsilon$ -subspace embedding for  $A$ , let  $z \in \mathbb{R}^r$  and define  $x = V\Sigma^{-1}z \in \mathbb{R}^d$ . Then we have  $Uz = Ax$  and

$$\|SUz\|_2^2 = \|SAx\|_2^2 \leq (1 + \epsilon)\|Ax\|_2^2 = (1 + \epsilon)\|Uz\|_2^2, \quad (64)$$

where we have used  $Uz = Ax$  and (4). Similarly, we have  $\|SUz\|_2^2 \geq (1 - \epsilon)\|Uz\|_2^2$ . Hence  $S$  is an  $\epsilon$ -subspace embedding for  $U$ .

Conversely, given  $S$  is an  $\epsilon$ -subspace embedding for  $U$ , let  $x \in \mathbb{R}^d$  and  $z = \Sigma V^T x \in \mathbb{R}^r$ . Then we have  $Ax = Uz$ , and  $\|SAx\|_2^2 = \|SUz\|_2^2 \leq (1 + \epsilon)\|Uz\|_2^2 = (1 + \epsilon)\|Ax\|_2^2$ . Similarly  $\|SAx\|_2^2 \geq (1 - \epsilon)\|Ax\|_2^2$ . Hence  $S$  is an  $\epsilon$ -subspace embedding for  $A$ .

(ii) Since the equivalence in (i) holds, note that (6) clearly implies (7). The latter also implies the former if (7) is applied to  $z/\|z\|_2$  for any nonzero  $z \in \mathbb{R}^r$ .  $\square$

*Proof of Lemma 4.* Let  $A = U\Sigma V^T$  be defined as in (5), and let  $z = \Sigma V^T x \in \mathbb{R}^r$ . Then  $y = Ax = Uz$ . Therefore

$$\|y\|_\infty = \|Uz\|_\infty = \max_{1 \leq i \leq n} |\langle U_i, z \rangle| \leq \max_{1 \leq i \leq n} \|U_i\|_2 \|z\|_2 \leq \mu(A) \|z\|_2, \quad (65)$$

where  $U_i$  denotes the  $i^{\text{th}}$  row of  $U$ . Furthermore,  $\|y\|_2 = \|Uz\|_2 = \|z\|_2$  which then implies  $\nu(y) = \|y\|_\infty / \|y\|_2 \leq \mu(A)$ .  $\square$

*Proof of Lemma 6.* Let  $Y = \{y_1, y_2, \dots, y_{|Y|}\}$ . Let  $B_i$  be the event that  $S$  is an  $\epsilon$ -JL embedding for  $y_i \in Y$ . Then  $\mathbb{P}(B_i) \geq 1 - \delta$  by assumption. We have

$$\mathbb{P}(S \text{ is an } \epsilon\text{-JL embedding for } Y) = \mathbb{P}(\cap_i B_i) = 1 - \mathbb{P}((\cap_i B_i)^c) \quad (66)$$

$$= 1 - \mathbb{P}(\cup_i B_i^c) \quad (67)$$

$$\geq 1 - \sum_i \mathbb{P}(B_i^c) \quad (68)$$

$$= 1 - \sum_i [1 - \mathbb{P}(B_i)] \quad (69)$$

$$\geq 1 - \sum_i [1 - (1 - \delta)] = 1 - |Y|\delta. \quad (70)$$

$\square$

*Proof of Lemma 7.* Let  $y_1, y_2 \in Y$ . We have that

$$\begin{aligned} |\langle Sy_1, Sy_2 \rangle - \langle y_1, y_2 \rangle| &= |(\|S(y_1 + y_2)\|^2 - \|S(y_1 - y_2)\|^2) \\ &\quad - (\|(y_1 + y_2)\|^2 - \|(y_1 - y_2)\|^2)|/4 \\ &\leq \epsilon(\|(y_1 + y_2)\|^2 + \|(y_1 - y_2)\|^2)/4 \\ &= \epsilon(\|y_1\|^2 + \|y_2\|^2)/2 \\ &= \epsilon, \end{aligned} \quad (71)$$

where to obtain the inequality, we use that  $S$  is an  $\epsilon$ -JL embedding for  $\{Y_+ \cup Y_-\}$ ; the last equality follows from  $\|y_1\|_2 = \|y_2\|_2 = 1$ .  $\square$



## A.2 Proof of results in Section 3

*Proof of Lemma 8.* Let  $\tilde{M} = \{z \in \mathbb{R}^r : \|z\|_2 = 1\}$ . Let  $\tilde{N} \subseteq \tilde{M}$  be the maximal set such that no two points in  $\tilde{N}$  are within distance  $\gamma$  from each other. Then it follows that the  $r$ -dimensional balls centred at points in  $\tilde{N}$  with radius  $\gamma/2$  are all disjoint and contained in the  $r$ -dimensional ball centred at the origin with radius  $(1 + \gamma/2)$ . Hence

$$\frac{\text{Volume of the } r\text{-dimensional ball centred at the origin with radius } (1 + \gamma/2)}{\text{Total volume of the } r\text{-dimensional balls centred at points in } \tilde{N} \text{ with radius } \gamma/2} \quad (72)$$

$$= \frac{1}{|\tilde{N}|} \frac{(1 + \frac{\gamma}{2})^r}{(\frac{\gamma}{2})^r} \geq 1, \quad (73)$$

which implies  $|\tilde{N}| \leq (1 + \frac{\gamma}{2})^r$ .

Let  $N = \{Uz \in \mathbb{R}^n : z \in \tilde{N}\}$ . Then  $|N| \leq |\tilde{N}| \leq (1 + \frac{\gamma}{2})^r$  and we show  $N$  is a  $\gamma$ -cover for  $M$ . Given  $y_M \in M$ , there exists  $z_M \in \tilde{M}$  such that  $y_M = Uz_M$ . By definition of  $\tilde{N}$ , there must be  $z_N \in \tilde{N}$  such that  $\|z_M - z_N\|_2 \leq \gamma$  as otherwise  $\tilde{N}$  would not be maximal. Let  $y_N = Uz_N \in N$ . Since  $U$  has orthonormal columns, we have  $\|y_M - y_N\|_2 = \|z_M - z_N\|_2 \leq \gamma$ .  $\square$

To prove Lemma 9, we need the following Lemma.

**Lemma A.1.** *Let  $\gamma \in (0, 1)$ ,  $U \in \mathbb{R}^{n \times r}$  having orthonormal columns and  $M \subseteq \mathbb{R}^n$  associated with  $U$  be defined in (26). Let  $N$  be a  $\gamma$ -cover of  $M$ ,  $y \in M$ . Then for any  $k \in \mathbb{N}$ , there exists  $\alpha_0, \alpha_1, \dots, \alpha_k \in \mathbb{R}$ ,  $y_0, y_1, y_2, \dots, y_k \in N$  such that*

$$\|y - \sum_{i=0}^k \alpha_i y_i\|_2 \leq \gamma^{k+1}, \quad (74)$$

$$|\alpha_i| \leq \gamma^i, i = 0, 1, \dots, k. \quad (75)$$

*Proof.* We use induction. Let  $k = 0$ . Then by definition of a  $\gamma$ -cover, there exists  $y_0 \in N$  such that  $\|y - y_0\| < \gamma$ . Letting  $\alpha_0 = 1$ , we have covered the  $k = 0$  case.

Now assume (74) and (75) are true for  $k = K \in \mathbb{N}$ . Namely there exists  $\alpha_0, \alpha_1, \dots, \alpha_K \in \mathbb{R}$ ,  $y_0, y_1, y_2, \dots, y_K \in N$  such that

$$\|y - \sum_{i=0}^K \alpha_i y_i\|_2 \leq \gamma^{K+1} \quad (76)$$

$$|\alpha_i| \leq \gamma^i, i = 0, 1, \dots, K. \quad (77)$$

Because  $y, y_0, y_1, \dots, y_K \in N \subseteq M$ , there exists  $z, z_0, z_1, \dots, z_K \in \mathbb{R}^r$  such that  $y = Uz, y_0 = Uz_0, y_1 = Uz_1, \dots, y_K = Uz_K$  with  $\|z\| = \|z_0\| = \dots = \|z_K\| = 1$ . Therefore

$$\frac{y - \sum_{i=0}^K \alpha_i y_i}{\|y - \sum_{i=0}^K \alpha_i y_i\|_2} = \frac{U(z - \sum_{i=0}^K \alpha_i z_i)}{\|U(z - \sum_{i=0}^K \alpha_i z_i)\|_2} = U \frac{z - \sum_{i=0}^K \alpha_i z_i}{\|z - \sum_{i=0}^K \alpha_i z_i\|_2} \in M, \quad (78)$$

where we have used that the columns of  $U$  are orthonormal.

Since  $N$  is a  $\gamma$ -cover for  $M$ , there exists  $y_{K+1} \in N$  such that

$$\left\| \frac{y - \sum_{i=0}^K \alpha_i y_i}{\|y - \sum_{i=0}^K \alpha_i y_i\|_2} - y_{K+1} \right\|_2 \leq \gamma. \quad (79)$$

Multiplying both sides by  $\alpha_{K+1} := \|y - \sum_{i=0}^K \alpha_i y_i\|_2 \leq \gamma^{K+1}$ , we have

$$\|y - \sum_{i=0}^{K+1} \alpha_i y_i\|_2 \leq \gamma^{K+2}, \quad (80)$$

$$|\alpha_i| \leq \gamma^i, i = 0, 1, \dots, K+1. \quad (81)$$

□

*Proof of Lemma 9.* Let  $y \in M$  and  $k \in \mathbb{N}$ , and consider the approximate representation of  $y$  provided in Lemma A.1, namely, assume that (74) and (75) hold. Then we have

$$\begin{aligned} \|S \sum_{i=0}^k \alpha_i y_i\|_2^2 &= \sum_{i=0}^k \|S \alpha_i y_i\|_2^2 + \sum_{0 \leq i < j \leq k} 2 \langle S \alpha_i y_i, S \alpha_j y_j \rangle \\ &= \sum_{i=0}^k \|S \alpha_i y_i\|_2^2 + \sum_{0 \leq i < j \leq k} 2 \langle \alpha_i y_i, \alpha_j y_j \rangle + \\ &\quad + \left[ \sum_{0 \leq i < j \leq k} 2 \langle S \alpha_i y_i, S \alpha_j y_j \rangle - \sum_{0 \leq i < j \leq k} 2 \langle \alpha_i y_i, \alpha_j y_j \rangle \right] \\ &\leq (1 + \epsilon_1) \sum_{i=0}^k \|\alpha_i y_i\|_2^2 + \sum_{0 \leq i < j \leq k} 2 \langle \alpha_i y_i, \alpha_j y_j \rangle + 2 \sum_{0 \leq i < j \leq k} \epsilon_1 \|\alpha_i y_i\|_2 \|\alpha_j y_j\|_2 \\ &= \left\| \sum_{i=0}^k \alpha_i y_i \right\|_2^2 + \epsilon_1 \left[ \sum_{i=0}^k \|\alpha_i y_i\|_2^2 + 2 \sum_{0 \leq i < j \leq k} \|\alpha_i y_i\|_2 \|\alpha_j y_j\|_2 \right], \end{aligned}$$

where to deduce the inequality, we use that  $S$  is a generalised  $\epsilon_1$ -JL embedding for  $N$ . Using  $\|y_i\|_2 = 1$  and  $|\alpha_i| \leq \gamma^i$ , we have

$$\begin{aligned} \frac{1}{\epsilon_1} \left\{ \left\| S \sum_{i=0}^k \alpha_i y_i \right\|_2^2 - \left\| \sum_{i=0}^k \alpha_i y_i \right\|_2^2 \right\} &= \sum_{i=0}^k \|\alpha_i y_i\|_2^2 + 2 \sum_{0 \leq i < j \leq k} \|\alpha_i y_i\|_2 \|\alpha_j y_j\|_2 \quad (82) \\ &\leq \sum_{i=0}^k \gamma^{2i} + 2 \sum_{0 \leq i < j \leq k} \gamma^i \gamma^j \\ &\leq \frac{1 - \gamma^{2k+2}}{1 - \gamma^2} + \frac{2\gamma(1 - \gamma^{k+1})(1 - \gamma^{k+1})}{(1 - \gamma)(1 - \gamma^2)}, \end{aligned}$$

where we have used

$$\begin{aligned} \sum_{0 \leq i < j \leq k} \gamma^i \gamma^j &= \sum_{i=0}^{k-1} \gamma^i \sum_{j=i+1}^k \gamma^j = \sum_{i=0}^{k-1} \gamma^{2i+1} \sum_{j=0}^{k-i-1} \gamma^j \\ &= \frac{\gamma(1 - \gamma^{k+1})}{1 - \gamma} \sum_{i=0}^{k-1} \gamma^{2i} = \frac{\gamma(1 - \gamma^{k+1})(1 - \gamma^{2k})}{(1 - \gamma)(1 - \gamma^2)}. \end{aligned}$$

Letting  $k \rightarrow \infty$  in (82), we deduce

$$\frac{1}{\epsilon_1} \left\{ \left\| S \sum_{i=0}^{\infty} \alpha_i y_i \right\|_2^2 - \left\| \sum_{i=0}^{\infty} \alpha_i y_i \right\|_2^2 \right\} \leq \frac{1}{1 - \gamma} + \frac{2\gamma}{(1 - \gamma)(1 - \gamma^2)} = \frac{1 + 2\gamma - \gamma^2}{(1 - \gamma)(1 - \gamma^2)},$$

Letting  $k \rightarrow \infty$  in (74) implies  $y = \sum_{i=0}^{\infty} \alpha_i y_i$ , and so the above gives

$$\|S y\|_2^2 - \|y\|_2^2 \leq \epsilon_1 \frac{1 + 2\gamma - \gamma^2}{(1 - \gamma)(1 - \gamma^2)} = \epsilon \|y\|_2^2,$$

where to get the first equality, we used  $\|y\|_2 = 1$  and the definition of  $\epsilon_1$ . The lower bound in the  $\epsilon_1$ -JL embedding follows similarly. □

### A.3 Proof of results in Section 5

*Proof of Lemma 13.* Note  $SW$  has rank  $p$  because  $SW = Q_1$ , where  $Q_1$  is defined in (46). By rank-nullity theorem in  $\mathbb{R}^p$ ,  $\text{rank}(W) + \dim \ker(W) = \text{rank}(SW) + \dim \ker(SW)$  where  $\ker(W)$  denotes the null space of  $W$ ; and since  $\dim \ker(SW) \geq \dim \ker(W)$ , we have that  $\text{rank}(SW) \leq \text{rank}(W)$ . So  $\text{rank}(W) \geq p$ . It follows that  $\text{rank}(W) = p$  because  $W \in \mathbb{R}^{n \times p}$  can have at most rank  $p$ .  $\square$

*Proof of Lemma 14.* Lemma 1 gives  $r = \text{rank}(A) = \text{rank}(SA) = p$ .  $\square$

*Proof of Lemma 15.* We have that  $x_s \in \arg \min \|SAx - Sb\|_2$  by checking the optimality condition  $(SA)^T SAx_s = (SA)^T Sb$ . Hence we have that

$$\|Ax_s - b\|_2^2 \leq \frac{1}{1-\epsilon} \|SAx_s - Sb\|_2^2 \leq \frac{1}{1-\epsilon} \|SAx^* - Sb\|_2^2 \leq \frac{1+\epsilon}{1-\epsilon} \|Ax^* - b\|_2^2, \quad (83)$$

where the first and the last inequality follow from  $S$  being a subspace embedding for  $(A \ b)$ , while the second inequality is due to  $x_s$  minimizing  $\|SAx - Sb\|$ .  $\square$

*Proof of Lemma 16.* Let  $z \in \ker(V_1^T) \cap \text{range}(A^T)$ . Then  $V_1^T z = 0$  and  $z = A^T w$  for some  $w \in \mathbb{R}^n$ . Let  $U, \Sigma, V$  be the SVD factors of  $A$  as defined in (5). Since  $S$  is an  $\epsilon$ -subspace embedding for  $A$ ,  $\text{rank}((SU)^T) = \text{rank}(SU) = \text{rank}(SA) = r$ , where  $r$  is the rank of  $A$  and hence there exists  $\hat{w} \in \mathbb{R}^m$  such that  $(SU)^T \hat{w} = U^T w$ . Note that

$$0 = V_1^T z = V_1^T A^T w = V_1^T V \Sigma U^T w = V_1^T V \Sigma U^T S^T \hat{w} = V_1^T A^T S^T \hat{w} = R_{11}^T Q_1^T \hat{w}, \quad (84)$$

which implies  $Q_1^T \hat{w} = 0$  because  $R_{11}^T$  is nonsingular. It follows that

$$z = A^T w = V \Sigma U^T w = V \Sigma U^T S^T \hat{w} = (SA)^T \hat{w} = V \begin{pmatrix} R_{11}^T Q_1^T \\ R_{12}^T Q_1^T \end{pmatrix} \hat{w} = 0, \quad (85)$$

where we have used  $Q_1^T \hat{w} = 0$  for the last equality.  $\square$

*Proof of Theorem 10.* (i) Using results in [16], LSQR applied to (47) converges as follows

$$\frac{\|y_j - y_*\|_{W^T W}}{\|y_0 - y_*\|_{W^T W}} \leq 2 \left( \frac{\sqrt{\kappa[W^T W]} - 1}{\sqrt{\kappa[W^T W]} + 1} \right)^j, \quad (86)$$

where  $y_j$  denotes the  $j$ th iterate of LSQR and  $\kappa(W^T W)$  refers to the condition number of  $W^T W$ . Since  $S$  is an  $\epsilon$ -subspace embedding for  $A$ , we have that the largest singular value of  $W$  satisfies

$$\sigma_{\max}(W) = \max_{\|y\|=1} \|AV_1 R_{11}^{-1} y\| \leq \sqrt{1+\epsilon} \max_{\|y\|=1} \|SAV_1 R_{11}^{-1} y\| = \sqrt{1+\epsilon} \max_{\|y\|=1} \|Q_1 y\| = \sqrt{1+\epsilon},$$

where we have used that  $SAV_1 R_{11}^{-1} = Q_1$  from (46). Similarly, it can be shown that the smallest singular value of  $W$  satisfies  $\sigma_{\min}(W) \geq \sqrt{1-\epsilon}$ . Hence

$$\kappa(W^T W) \leq \frac{1+\epsilon}{1-\epsilon}. \quad (87)$$

Hence we have

$$\frac{\sqrt{\kappa[W^T W]} - 1}{\sqrt{\kappa[W^T W]} + 1} \leq \frac{\sqrt{1+\epsilon} - \sqrt{1-\epsilon}}{\sqrt{1+\epsilon} + \sqrt{1-\epsilon}} = \frac{(\sqrt{1+\epsilon} - \sqrt{1-\epsilon})(\sqrt{1+\epsilon} + \sqrt{1-\epsilon})}{(\sqrt{1+\epsilon} + \sqrt{1-\epsilon})^2} \leq \epsilon.$$

Thus (86) implies  $\|y_\tau - y_*\|_{W^T W} \leq \tau_r \|y_0 - y_*\|_{W^T W}$  whenever  $\tau \geq \frac{\log(2) + |\log \tau_r|}{|\log \epsilon|}$ .

(ii) If we initialize  $y_0 := Q^T Sb$  for the LSQR method in Step 4, then we have

$$\begin{aligned} \|y_0 - y_*\|_{W^T W} &= \|Ax_s - Ax_*\|_2 = \|Ax_s - b - (Ax_* - b)\|_2 \leq \|Ax_s - b\| + \|Ax_* - b\| \\ &\leq \left( \sqrt{\frac{1+\epsilon}{1-\epsilon}} - 1 \right) \|Ax_* - b\| \leq \frac{2\epsilon}{1-\epsilon} \|Ax_* - b\|, \end{aligned}$$

where we have used  $\sqrt{\frac{1+\epsilon}{1-\epsilon}} \leq \frac{1+\epsilon}{1-\epsilon}$  to get the last inequality. Using part (i), after at most  $\frac{\log(2)+|\log \tau_r|}{|\log \epsilon|}$  LSQR iterations, we have that

$$\|y_\tau - y_*\|_{W^T W} \leq \frac{2\epsilon\tau_r}{1-\epsilon} \|Ax_* - b\|_2. \quad (88)$$

Note that  $\|Ax_\tau - Ax_*\|_2 = \|y_\tau - y^*\|_{W^T W}$ . Using the triangle inequality, we deduce

$$\|Ax_\tau - b\|_2 = \|Ax_\tau - Ax_* + Ax_* - b\|_2 \leq \left(1 + \frac{2\epsilon\tau_r}{1-\epsilon}\right) \|Ax_* - b\|_2. \quad (89)$$

□

## B Calibration of sketching dimension for dense solvers

In Figures 14, 15, 16 and 17, we display the calibration results for finding a good default value for  $m$ , the size of the hashing sketching matrix, in Ski-LLS-dense, Blendenpik and LSRN.

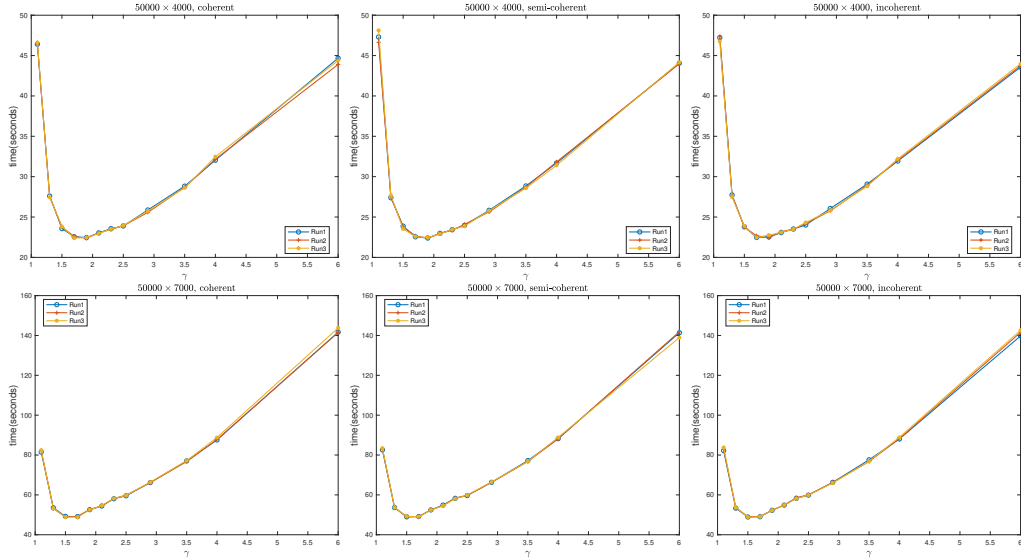


Figure 14: Runtime of Ski-LLS-dense on dense matrices  $A \in \mathbb{R}^{n \times d}$  from Test Set 1 with  $n = 50000$ ,  $d = 4000$  and  $n = 50000$ ,  $d = 7000$  and different values of  $\gamma = m/d$ . For each plot, Ski-LLS-dense is run three times on (the same) randomly generated  $A$ . We see that the runtime has low variance despite the randomness in the solver. We choose  $\gamma = 1.7$  to approximately minimize the runtime across the above plots.

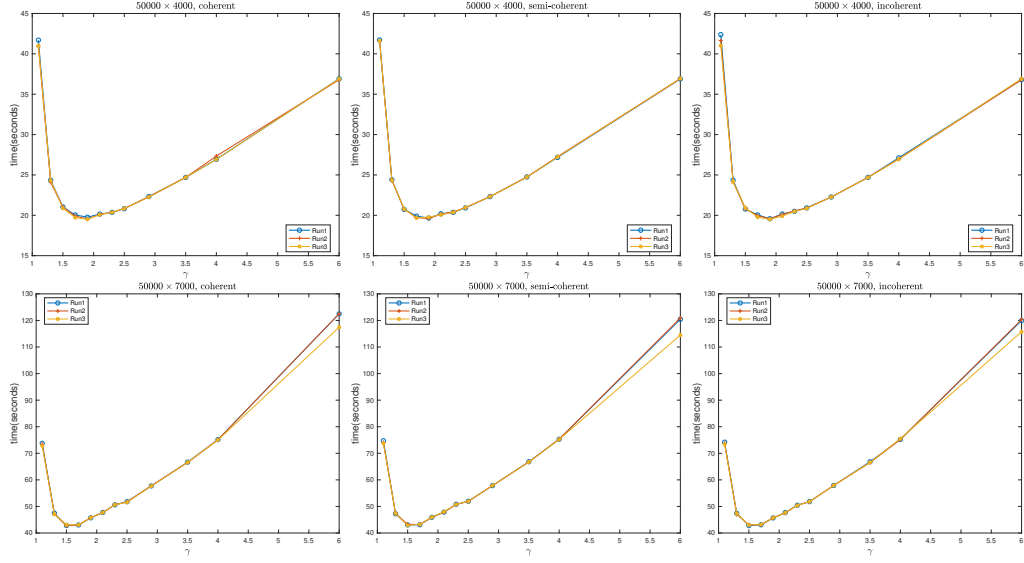


Figure 15: Runtime of Ski-LLS-dense without R-CPQR on dense matrices  $A \in \mathbb{R}^{n \times d}$  from Test Set 1 with  $n = 50000, d = 4000$  and  $n = 50000, d = 7000$  and different values of  $\gamma = m/d$ . For each plot, Ski-LLS-dense without R-CPQR is run three times on (the same) randomly generated  $A$ . We see that the runtime has low variance despite the randomness in the solver. Note that using LAPACK QR instead of R-CPQR results in slightly shorter runtime (according to Figure 14). We choose  $\gamma = 1.7$  to approximately minimize the runtime across the above plots.

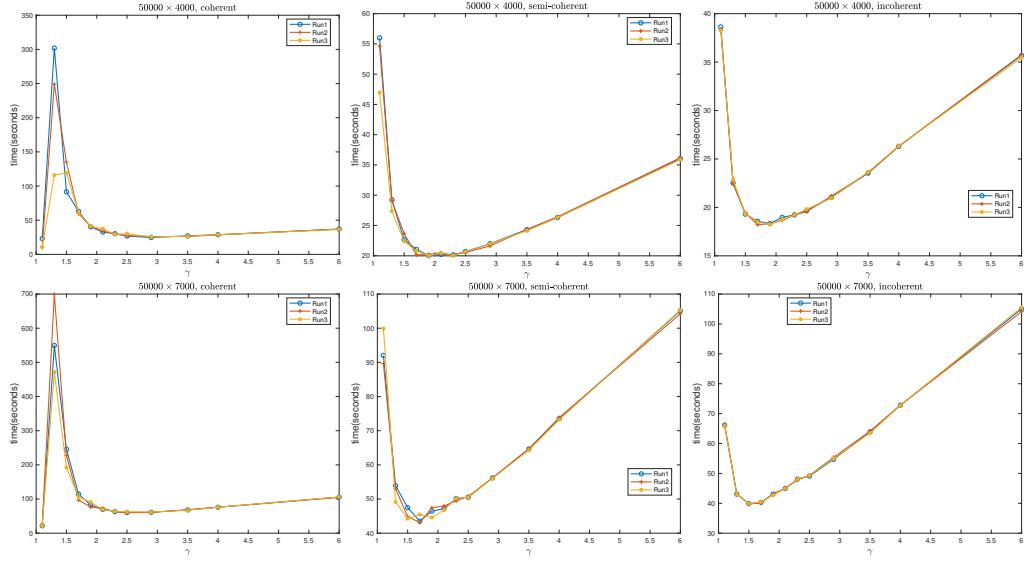


Figure 16: Runtime of Blendenpik on dense matrices  $A \in \mathbb{R}^{n \times d}$  from Test Set 1 with  $n = 50000, d = 4000$  and  $n = 50000, d = 7000$  and different values of  $\gamma = m/d$ . For each plot, Blendenpik is run three times on (the same) randomly generated  $A$ . We see that the runtime has low variance despite the randomness in the solver. Note that Blendenpik handles coherent dense  $A$  significantly less well than Ski-LLS-dense. We choose  $\gamma = 2.2$  to approximately minimize the runtime across the above plots.

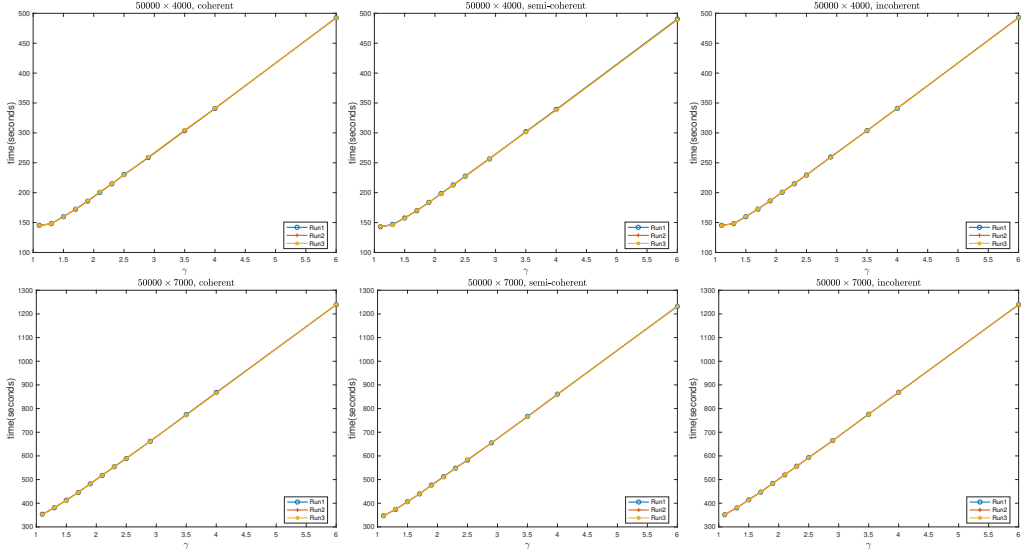


Figure 17: Runtime of LSRN on dense matrices  $A \in \mathbb{R}^{n \times d}$  from Test Set 1 with  $n = 50000, d = 4000$  and  $n = 50000, d = 7000$  and different values of  $\gamma = m/d$ . For each plot, LSRN is run three times on (the same) randomly generated  $A$ . We see that the runtime has low variance despite the randomness in the solver. Note that LSRN runs more than 5 times slower comparing to Blendenpik or Ski-LLS in the serial testing environment, due to the use of SVD and Gaussian sketching. We choose  $\gamma = 1.1$  to approximately minimize the runtime across the above plots.

## C Calibration of sketching dimension and number of hashing entries for sparse solvers

In Figures 18 and Figure 19 we display the calibration results for finding a good default value for  $m$  (the size of the sketching matrix) and  $s$  (the number of nonzero entries per sketching column), for Ski-LLS-sparse; similar plots are given for LSRN in Figure 20 and Figure 21, respectively.

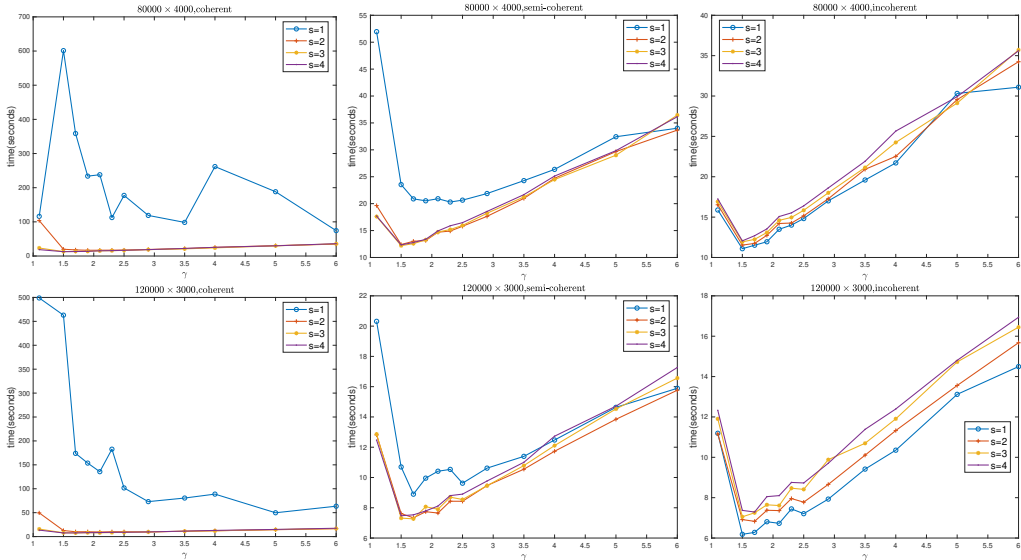


Figure 18: Runtime for Ski-LLS-sparse on sparse  $A \in \mathbb{R}^{n \times d}$  from Test Set 2 with  $n = 80000, d = 4000$  and  $n = 120000, d = 3000$  and different values of  $\gamma = m/d$  using different values of  $s$  and  $\gamma = m/d$ . Using these plots and some experiments with Test Set 3 (Florida collection), we set  $m = 1.4d$  and  $s = 2$  in Ski-LLS-sparse.

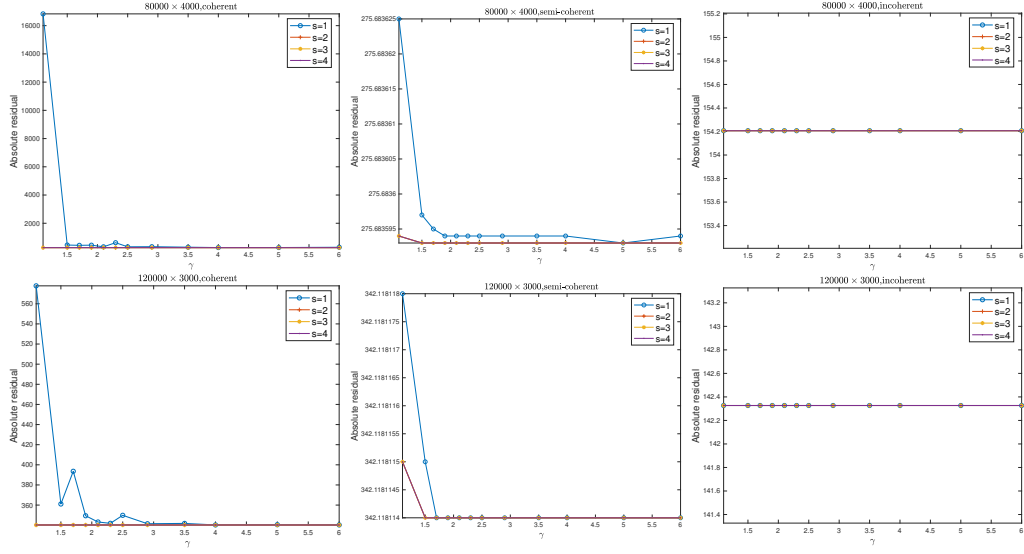


Figure 19: Residual values obtained by Ski-LLS-sparse on the same sparse problems as in Figure 18. Note that using 1-hashing ( $s = 1$ ) results in inaccurate solutions.

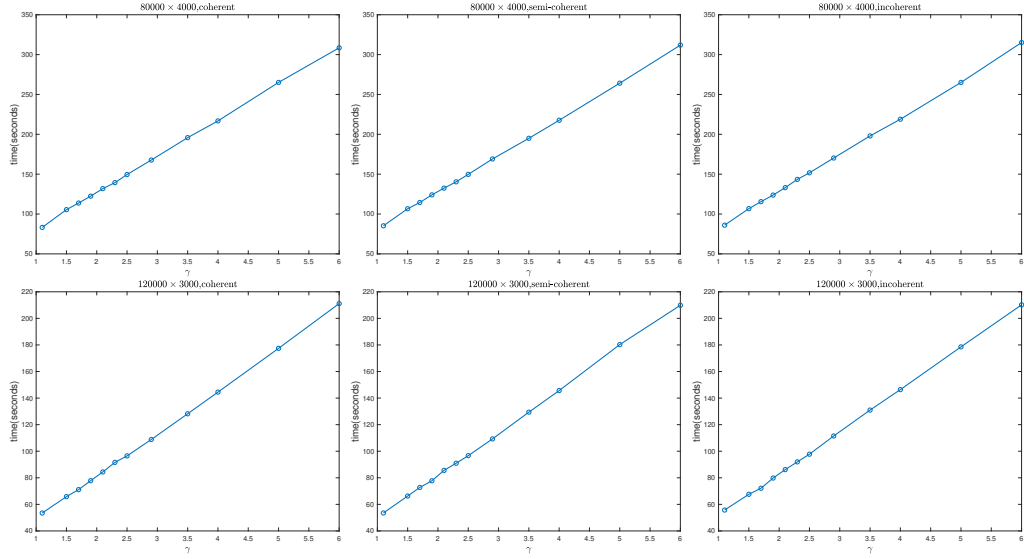


Figure 20: Runtime for LSRN on sparse  $A \in \mathbb{R}^{n \times d}$  from Test Set 2 with  $n = 80000, d = 4000$  and  $n = 120000, d = 3000$  and different values of  $\gamma = m/d$ . The plots indicate that a value of  $m = 1.1d$  is reasonable.

	Ski-LLS	LS_SPQR	LS_HSL	LSRN
$40000 \times 2000$	100.339	100.339	100.339	100.339
$80000 \times 2000$	110.353	110.339	110.496	110.353
$80000 \times 4000$	154.207	154.207	154.247	154.207
$120000 \times 3000$	142.327	142.327	142.461	142.327
$120000 \times 5000$	182.171	182.171	182.208	182.171

Table 3: Residual values obtained by solvers on problems in Figure 7.

	Ski-LLS	LS_SPQR	LS_HSL	LSRN
$40000 \times 2000$	194.714	194.714	194.718	194.714
$80000 \times 2000$	279.352	279.352	279.359	279.352
$80000 \times 4000$	275.684	275.684	275.685	275.684
$120000 \times 3000$	342.118	342.118	342.120	342.118
$120000 \times 5000$	339.129	339.129	339.141	339.129

Table 4: Residual values obtained by solvers on problems in Figure 8.

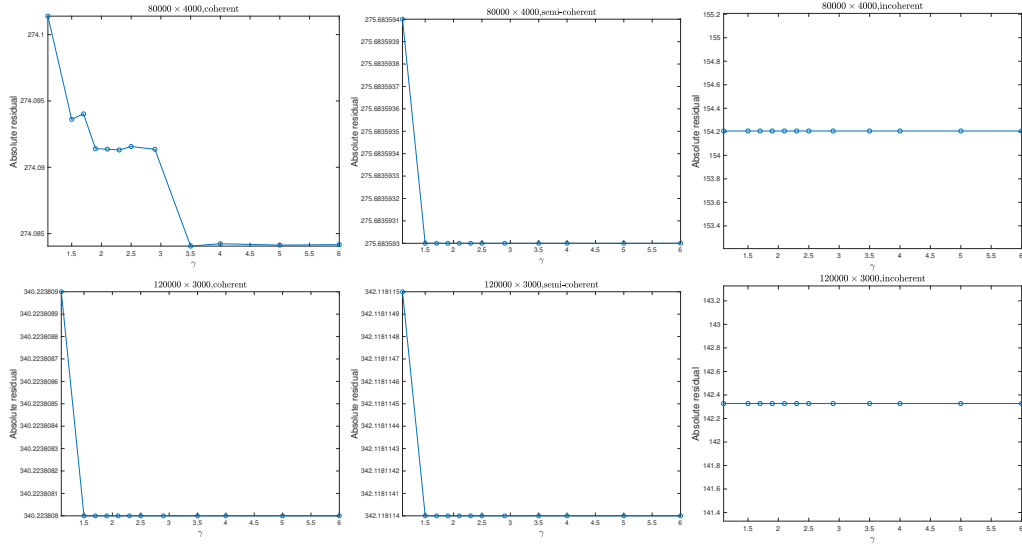


Figure 21: Residual values obtained by LSRN on the same sparse problems as in Figure 20. Note that the reduced accuracy in one of the plot is likely due to the relative tolerance of the LSQR being set at  $10^{-6}$ .

## D Residual accuracy results for randomly generated problems

See Table 3, Table 4, Table 5.

	Ski-LLS	LS_SPQR	LS_HSL	LSRN
$40000 \times 2000$	193.851	194.158	195.586	193.802
$80000 \times 2000$	277.746	277.750	278.561	277.746
$80000 \times 4000$	274.117	274.117	279.763	274.102
$120000 \times 3000$	340.224	340.236	343.114	340.224
$120000 \times 5000$	337.100	337.144	342.291	337.085

Table 5: Residual values obtained by solvers on problems in Figure 9.



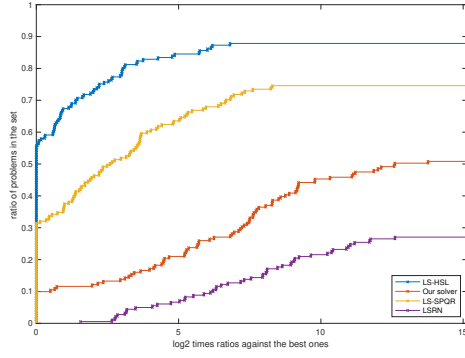


Figure 22: Performance profile comparison of Ski-LLS with LSRN, LS\_HSL and LS\_SPQR for all matrices  $A \in \mathbb{R}^{n \times d}$  in the Florida matrix collection with  $n \geq 2d$ .

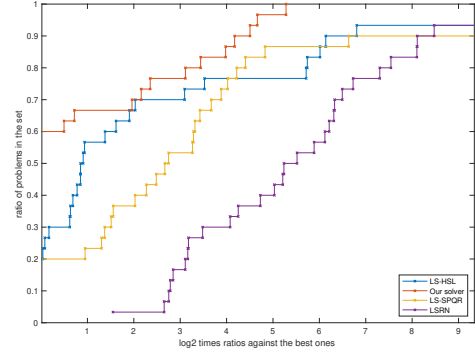


Figure 23: Performance profile comparison of Ski-LLS with LSRN, LS\_HSL and LS\_SPQR for all matrices  $A \in \mathbb{R}^{n \times d}$  in the Florida matrix collection with  $n \geq 10d$  and  $\text{nnz}(A) \geq 0.001nd$ .

	nrow	ncol	rank
lp_ship12l	5533	1151	1042
Franz1	2240	768	755
GL7d26	2798	305	273
cis-n4c6-b2	1330	210	190
lp_modszk1	1620	687	686
rel5	240	35	24
ch5-5-b1	200	25	24
n3c5-b2	120	45	36
ch4-4-b1	72	16	15
n3c5-b1	45	10	9
n3c4-b1	15	6	5
connectus	394792	512	<458
landmark	71952	2704	2671
cis-n4c6-b3	5940	1330	1140

Table 6: Dimensions for the Florida collection problems tested in Table 2

## E Additional Performance Profiles for the Florida Matrix Collection

See Figures 22 and 23.

## F Additional information on the Florida matrix collection

Table 6 gives dimensions for the Florida collection problems tested in Table 2.