SIAM J. SCI. COMPUT.
Vol. 36, No. 5, pp. S111–S131

EFFICIENT DIMENSIONALITY REDUCTION FOR CANONICAL CORRELATION ANALYSIS*

HAIM AVRON†, CHRISTOS BOUTSIDIS†,
SIVAN TOLEDO‡, AND ANASTASIOS ZOUZIAS§

Abstract. We present a fast algorithm for approximate canonical correlation analysis (CCA). Given a pair of tall-and-thin matrices, the proposed algorithm first employs a randomized dimensionality reduction transform to reduce the size of the input matrices, and then applies any CCA algorithm to the new pair of matrices. The algorithm computes an approximate CCA to the original pair of matrices with provable guarantees while requiring asymptotically fewer operations than the state-of-the-art exact algorithms.

Key words. principal angles, canonical correlations, randomized algorithms, dimensionality reduction

AMS subject classifications. 15B52, 15A18, 11K45

DOI. 10.1137/130919222

1. Introduction.

Canonical correlation analysis (CCA) [20] is an important technique in statistics, data analysis, and data mining. CCA has been successfully applied in many statistics and machine learning applications, e.g., dimensionality reduction [30], clustering [8], learning of word embeddings [11], sentiment classification [10], discriminant learning [29], and object recognition [22]. In many ways CCA is analogous to principal component analysis (PCA), but instead of analyzing a single dataset (in matrix form), the goal of CCA is to analyze the relation between a pair of datasets (each in matrix form). From a statistical point of view, PCA extracts the maximum covariance directions between elements in a single matrix, whereas CCA finds the direction of maximal correlation between a pair of matrices. From a linear algebraic point of view, CCA measures the similarities between two subspaces (those spanned by the columns of each of the two matrices analyzed). From a geometric point of view, CCA computes the cosine of the *principal angles* between the two subspaces.

There are different ways to define the canonical correlations of a pair of matrices, and all these ways are equivalent [16]. The linear algebraic formulation of Golub and Zha [16], which we present shortly, serves our algorithmic point of view best.

*Received by the editors May 1, 2013; accepted for publication (in revised form) January 14, 2014; published electronically October 30, 2014. An extended abstract of this work appears in *Proceedings of the* 30*th International Conference on Machine Learning*, 2013, pp. 347–355.

http://www.siam.org/journals/sisc/36-5/91922.html

†Business Analytics and Mathematical Sciences, IBM T. J. Watson Research Center, Yorktown Heights, NY 10598 (haimav@us.ibm.com, cboutsi@us.ibm.com). The work of these authors was supported by the XDATA program of the Defense Advanced Research Projects Agency (DARPA), administered through Air Force Research Laboratory contract FA8750-12-C-0323.

‡Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, Israel (stoledo@tau.ac.il). This author's work was supported by grant 1045/09 from the Israel Science Foundation (funded by the Israel Academy of Sciences and Humanities) and by grant 2010231 from the US–Israel Binational Science Foundation.

§Mathematical and Computational Sciences, IBM Zürich Research Lab, Zurich, Switzerland (azo@zurich.ibm.com). This author received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement 259569.

DEFINITION 1.1. *Let* $\mathbf{A} \in \mathbb{R}^{m\times n}$ *and* $\mathbf{B} \in \mathbb{R}^{m\times \ell}$*, and assume that* $p = \mathrm{rank}(\mathbf{A}) \geq \mathrm{rank}(\mathbf{B}) = q$*. The* canonical correlations $\sigma_1(\mathbf{A},\mathbf{B}) \geq \sigma_2(\mathbf{A},\mathbf{B}) \geq \cdots \geq \sigma_q(\mathbf{A},\mathbf{B})$ *of the matrix pair* $(\mathbf{A},\mathbf{B})$ *are defined recursively by the following formula:*

$$\sigma_i(\mathbf{A},\mathbf{B}) = \max_{\mathbf{x}\in\mathcal{A}_i,\mathbf{y}\in\mathcal{B}_i} \sigma(\mathbf{Ax},\mathbf{By}) =: \sigma(\mathbf{Ax}_i,\mathbf{By}_i), \quad i = 1,\ldots,q,$$

where

- $\sigma(\mathbf{u},\mathbf{v}) = |\mathbf{u}^{\mathrm{T}}\mathbf{v}|/(\|\mathbf{u}\|_2\|\mathbf{v}\|_2)$,
- $\mathcal{A}_i = \{\mathbf{x}\in\mathbb{R}^n : \mathbf{Ax} \neq \mathbf{0}, \mathbf{Ax} \perp \{\mathbf{Ax_1},\ldots,\mathbf{Ax_{i-1}}\}\}$,
- $\mathcal{B}_i = \{\mathbf{y}\in\mathbb{R}^\ell : \mathbf{By} \neq \mathbf{0}, \mathbf{By} \perp \{\mathbf{By_1},\ldots,\mathbf{By_{i-1}}\}\}$.

The m*-dimensional unit vectors*

$$\mathbf{Ax}_1/\|\mathbf{Ax}_1\|_2,\ldots,\mathbf{Ax}_q/\|\mathbf{Ax}_q\|_2 \quad \textit{and} \quad \mathbf{By}_1/\|\mathbf{By}_1\|_2,\ldots,\mathbf{By}_q/\|\mathbf{By}_q\|_2$$

are called the canonical *or* principal vectors*. The vectors*

$$\mathbf{x}_1/\|\mathbf{Ax}_1\|_2,\ldots,\mathbf{x}_q/\|\mathbf{Ax}_q\|_2 \quad \textit{and} \quad \mathbf{y}_1/\|\mathbf{By}_1\|_2,\ldots,\mathbf{y}_q/\|\mathbf{By}_q\|_2$$

are called canonical weights *(or* projection vectors*). Here,* $\mathbf{x}_i/\|\mathbf{Ax}_i\|_2 \in \mathbb{R}^n$ *and* $\mathbf{y}_i/\|\mathbf{By}_i\|_2 \in \mathbb{R}^\ell$ *for all* $i = 1:q$*. Note that the canonical weights and the canonical vectors are* not *uniquely defined.*

1.1. Contributions. The main contribution of this article (see Theorem 5.2) is a fast algorithm to compute an approximate CCA. The algorithm computes an approximation to *all* the canonical correlations. It also computes a set of approximate canonical weights with provable guarantees. We show that the proposed algorithm is often asymptotically faster compared to the standard method of Björck and Golub [5]. To the best of our knowledge, this is the *first* subcubic time algorithm for approximate CCA that has provable guarantees.

The proposed algorithm is based on *dimensionality reduction*: given a pair of matrices $(\mathbf{A},\mathbf{B})$, we transform the pair to a new pair $(\hat{\mathbf{A}},\hat{\mathbf{B}})$ that has many fewer rows, and then compute the canonical correlations of the new pair exactly, alongside a set of canonical weights, e.g., using the Björck and Golub algorithm (see section 2.1). We prove that with high probability the canonical correlations of $(\hat{\mathbf{A}},\hat{\mathbf{B}})$ are close to the canonical correlations of $(\mathbf{A},\mathbf{B})$ and that any set of canonical weights of $(\hat{\mathbf{A}},\hat{\mathbf{B}})$ can be used to construct a set of approximately orthogonal canonical vectors of $(\mathbf{A},\mathbf{B})$. The transformation of $(\mathbf{A},\mathbf{B})$ into $(\hat{\mathbf{A}},\hat{\mathbf{B}})$ is done in two steps. First, we apply the *randomized Walsh–Hadamard transform (RHT)* to both $\mathbf{A}$ and $\mathbf{B}$. This is a unitary transformation, so the canonical correlations are preserved exactly. On the other hand, we show that with high probability, the transformed matrices have their "information" equally spread among all the input rows, so now the transformed matrices are amenable to uniform sampling. In the second step, we uniformly sample (without replacement) a sufficiently large set of rows and rescale them to form $(\hat{\mathbf{A}},\hat{\mathbf{B}})$. The combination of RHT and uniform sampling is often called *subsampled randomized Walsh–Hadamard transform (SRHT)* in the literature [33]. Note that other variants of dimensionality reduction [27] might be appropriate as well, but for concreteness we focus on the SRHT. (See section 6 for a discussion of other transforms.)

Our dimensionality reduction scheme is particularly effective when the matrices are tall-and-thin, that is, they have many more rows than columns. Targeting such matrices is natural: in typical CCA applications, columns typically correspond to features or labels and rows correspond to samples or training data. By computing the

CCA on as many instances as possible (as much training data as possible), we get the most reliable estimates of application-relevant quantities. However, in current algorithms adding instances (rows) is expensive, e.g., in the Björck and Golub algorithm we pay $O(n^2 + \ell^2)$ for each new row. Our algorithm allows practitioners to run CCA on huge datasets because we reduce the cost of an extra row to almost $O(n + \ell)$.

Finally, from an empirical point of view, we demonstrate that our algorithm is faster than the standard algorithm in practice by 30% to 60%, even on fairly small matrices (section 7).

1.2. Related work. Dimensionality reduction has been the driving force behind many recent algorithms for accelerating key machine learning and linear algebraic tasks. A representative example is linear regression, i.e., solve the least-squares problem $\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2$, where $\mathbf{b} \in \mathbb{R}^m$. If $m \gg n$, then one can use SRHT to reduce the dimension of $\mathbf{A}$ and $\mathbf{b}$, to form $\hat{\mathbf{A}}$ and $\hat{\mathbf{b}}$, and then solve the small problem $\min_{\mathbf{x}} \|\hat{\mathbf{A}}\mathbf{x} - \hat{\mathbf{b}}\|_2$. This process will return an approximate solution to the original problem [27, 6, 13]. Alternatively, one can observe that $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ and $\hat{\mathbf{A}}^{\mathrm{T}}\hat{\mathbf{A}}$ are "spectrally close," so an effective preconditioner can be obtained via a QR decomposition of $\hat{\mathbf{A}}$ [26, 4]. Other problems that can be accelerated using dimensionality reduction include (i) low-rank matrix approximation [18], (ii) matrix multiplication [27], (iii) K-means clustering [7], and (iv) approximation of matrix coherence and statistical leverage [12].

Our approach uses techniques similar to the algorithms mentioned above. For example, Lemma 2.4 plays a central role in these algorithms as well. However, our analysis requires the use of advanced ideas from matrix perturbation theory and it leads to two new technical lemmas that might be of independent interest: Lemmas 3.4 and 3.5 provide bounds for the singular values of the product of two *different* sampled orthonormal matrices. Previous work only provides bounds for products of the *same* matrix (Lemma 2.4; see also [27, Corollary 11]).

Dimensionality reduction techniques for accelerating CCA have been suggested or used in the past. One common technique is to simply use fewer samples by uniformly sampling the rows. Although this technique might work reasonably well in many instances, it may fail for others unless all rows are sampled. In fact, Theorem 4.1 analyzes uniform sampling and establishes bounds on the required sample size.

Sun, Ceran, and Ye suggest a two-stage approach which involves first solving a least-squares problem, and then using the solution to reduce the problem size [30]. However, their technique involves explicitly factoring one of the two matrices, which takes cubic time. Therefore, their method is especially effective when one of the two matrices has significantly fewer columns than the other. When both matrices have about the same number of columns, there is no asymptotic performance gain. In contrast, our method is subcubic in *any* case.

Finally, it is worth noting that CCA itself has been used for dimensionality reduction [31, 8, 30]. This is not the focus of this article; we suggest a dimensionality reduction technique to accelerate CCA.

2. Preliminaries. We use $i : j$ to denote the set $\{i, \ldots, j\}$, and $[n] = 1 : n$. We use $\mathbf{A}, \mathbf{B}, \ldots$ to denote matrices and $\mathbf{a}, \mathbf{b}, \ldots$ to denote column vectors. $\mathbf{I}_n$ is the $n \times n$ identity matrix; $\mathbf{0}_{m \times n}$ is the $m \times n$ matrix of zeros. We denote the number of nonzero elements in $\mathbf{A}$ by $\mathrm{nnz}(\mathbf{A})$. We denote by $\mathcal{R}(\cdot)$ the column space of its argument matrix. We denote by $[\mathbf{A}; \mathbf{B}]$ the matrix obtained by concatenating the columns of $\mathbf{B}$ next to the columns of $\mathbf{A}$. Given a subset of indices $T \subseteq [m]$, the corresponding

sampling matrix $\mathbf{S}$ is the $|T| \times m$ matrix obtained by discarding from $\mathbf{I}_m$ the rows whose index is not in T. Note that $\mathbf{SA}$ is the matrix obtained by keeping only the rows in $\mathbf{A}$ whose index *appears* in T. A symmetric matrix $\mathbf{A}$ is positive semidefinite (PSD), denoted by $0 \preceq \mathbf{A}$, if $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} \geq 0$ for every vector $\mathbf{x}$. For any two symmetric matrices $\mathbf{X}$ and $\mathbf{Y}$ of the same size, $\mathbf{X} \preceq \mathbf{Y}$ denotes that $\mathbf{Y} - \mathbf{X}$ is a PSD matrix.

We denote the *compact* (or *thin*) SVD of a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ of rank p by $\mathbf{A} = \mathbf{U_A}\mathbf{\Sigma_A}\mathbf{V_A}^{\mathrm{T}}$ with $\mathbf{U_A} \in \mathbb{R}^{m\times p}$, $\mathbf{\Sigma_A} \in \mathbb{R}^{p\times p}$, and $\mathbf{V_A}^{\mathrm{T}} \in \mathbb{R}^{p\times n}$. In this case, we denote the singular values of $\mathbf{A}$ (i.e., the diagonal elements of $\mathbf{\Sigma_A}$) by $\sigma_1(\mathbf{A}) \geq \sigma_2(\mathbf{A}) \geq \cdots \geq \sigma_p(\mathbf{A})$. Sometimes, we also use the *full* SVD of a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ with $\alpha = \min\{m, n\}$ as $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$ with $\mathbf{U} \in \mathbb{R}^{m\times\alpha}$, $\mathbf{\Sigma} \in \mathbb{R}^{\alpha\times\alpha}$, and $\mathbf{V}^{\mathrm{T}} \in \mathbb{R}^{\alpha\times n}$. In this case, we also denote the singular values of $\mathbf{A}$ (i.e., the diagonal elements of $\mathbf{\Sigma}$) by $\sigma_1(\mathbf{A}) \geq \sigma_2(\mathbf{A}) \geq \cdots \geq \sigma_\alpha(\mathbf{A})$. We refer to the thin or full SVD of a matrix simultaneously, and when the two definitions are equivalent or refer to a specific number of singular values of the matrix we do not make any distinction. Finally, the Moore–Penrose pseudoinverse of $\mathbf{A}$ is $\mathbf{A}^+ = \mathbf{V_A}\mathbf{\Sigma_A}^{-1}\mathbf{U_A}^{\mathrm{T}} \in \mathbb{R}^{n\times m}$, where $\mathbf{A} = \mathbf{U_A}\mathbf{\Sigma_A}\mathbf{V_A}^{\mathrm{T}}$ is the thin SVD of $\mathbf{A}$ of rank $p = \mathrm{rank}(\mathbf{A})$.

2.1. The Björck and Golub algorithm. There are quite a few algorithms with which to compute the canonical correlations [16]. One popular method is due to Björck and Golub [5] (see also section 6.4.3 in [17]). It is based on the following observation.

THEOREM 2.1 (Theorem 1 in [5]). *Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ have rank p and $\mathbf{B} \in \mathbb{R}^{m\times\ell}$ have rank q. Assume that the columns of $\mathbf{Q} \in \mathbb{R}^{m\times p}$ ($m \geq p$) and $\mathbf{W} \in \mathbb{R}^{m\times q}$ ($m \geq q$) form an orthonormal basis for the range of $\mathbf{A}$ and $\mathbf{B}$ (respectively). Also, let $p \geq q$. Let $\mathbf{Q}^{\mathrm{T}}\mathbf{W} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$ be the full SVD of $\mathbf{Q}^{\mathrm{T}}\mathbf{W} \in \mathbb{R}^{p\times q}$ with $\mathbf{U} \in \mathbb{R}^{p\times q}$, $\mathbf{\Sigma} \in \mathbb{R}^{q\times q}$, and $\mathbf{V} \in \mathbb{R}^{q\times q}$. The diagonal elements of $\mathbf{\Sigma}$ are the canonical correlations of the matrix pair $(\mathbf{A}, \mathbf{B})$. The canonical vectors are given by the columns of $\mathbf{QU} \in \mathbb{R}^{m\times q}$ (for $\mathbf{A}$) and $\mathbf{WV} \in \mathbb{R}^{m\times q}$ (for $\mathbf{B}$).*

Theorem 2.1 implies that once we have a pair of matrices $\mathbf{Q}$ and $\mathbf{W}$ with orthonormal columns whose column space spans the same column space of $\mathbf{A}$ and $\mathbf{B}$, respectively, then all we need is to compute the SVD of $\mathbf{Q}^{\mathrm{T}}\mathbf{W}$. Björck and Golub suggest the use of QR decompositions to find the matrices $\mathbf{Q}$ and $\mathbf{W}$, but $\mathbf{U_A} \in \mathbb{R}^{m\times p}$ and $\mathbf{U_B} \in \mathbb{R}^{m\times q}$ from the compact SVD of $\mathbf{A}$ and $\mathbf{B}$ will serve as well. Notice that both options require $O(m(n^2 + \ell^2))$ time.

COROLLARY 2.2. *Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ have rank p and $\mathbf{B} \in \mathbb{R}^{m\times\ell}$ have rank $q \leq p$. Let $\mathbf{A} = \mathbf{U_A}\mathbf{\Sigma_A}\mathbf{V_A}^{\mathrm{T}}$ with $\mathbf{U_A} \in \mathbb{R}^{m\times p}$, $\mathbf{\Sigma_A} \in \mathbb{R}^{p\times p}$, and $\mathbf{V_A}^{\mathrm{T}} \in \mathbb{R}^{p\times n}$ be the thin SVD of $\mathbf{A}$. Similarly, let $\mathbf{B} = \mathbf{U_B}\mathbf{\Sigma_B}\mathbf{V_B}^{\mathrm{T}}$ with $\mathbf{U_B} \in \mathbb{R}^{m\times q}$, $\mathbf{\Sigma_B} \in \mathbb{R}^{q\times q}$, and $\mathbf{V_B}^{\mathrm{T}} \in \mathbb{R}^{p\times n}$ be the thin SVD of $\mathbf{B}$. Let $\mathbf{U_A}^{\mathrm{T}}\mathbf{U_B} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$ be the full SVD of $\mathbf{U_A}^{\mathrm{T}}\mathbf{U_B} \in \mathbb{R}^{p\times q}$ with $\mathbf{U} \in \mathbb{R}^{p\times q}$, $\mathbf{\Sigma} \in \mathbb{R}^{q\times q}$, and $\mathbf{V} \in \mathbb{R}^{q\times q}$. Then, for $i \in [q]$, $\sigma_i(\mathbf{A}, \mathbf{B}) = \mathbf{\Sigma}_{ii}$. The canonical vectors are given by the columns of $\mathbf{U_A U} \in \mathbb{R}^{m\times q}$ (for $\mathbf{A}$) and $\mathbf{U_B V} \in \mathbb{R}^{m\times q}$ (for $\mathbf{B}$). The canonical weights are given by the columns of $\mathbf{V_A}\mathbf{\Sigma_A}^{-1}\mathbf{U} \in \mathbb{R}^{n\times q}$ (for $\mathbf{A}$) and $\mathbf{V_B}\mathbf{\Sigma_B}^{-1}\mathbf{V} \in \mathbb{R}^{\ell\times q}$ (for $\mathbf{B}$).*

2.2. Matrix coherence and sampling from an orthonormal matrix. Matrix coherence is a fundamental concept in the analysis of matrix sampling algorithms (e.g. [32, 21]). There are quite a few similar but different ways to define the coherence. In this article we use the following definition.

DEFINITION 2.3. *Given a matrix $\mathbf{A}$ with m rows and rank p, the* coherence *of $\mathbf{A}$ is defined as*

$$\mu(\mathbf{A}) = \max_{i\in[m]} \|\mathbf{e}_i^{\top}\mathbf{U_A}\|_2^2,$$

where $\mathbf{e}_i$ *is the ith standard basis (column) vector of* $\mathbb{R}^m$ *and* $\mathbf{U_A} \in \mathbb{R}^{m \times p}$ *is the U-factor from the thin SVD of* $\mathbf{A}$.

The coherence of a matrix gives information about the localization or uniformity of the elements of an orthonormal basis of the matrix range. Similarly, the coherence is a measure of how close a basis is to sharing a vector with a canonical basis.

Note that the coherence of $\mathbf{A}$ is a property of the column space of $\mathbf{A}$ and does not depend on a particular choice of basis (e.g., the basis described by the columns of $\mathbf{A}$). Therefore, if $\mathcal{R}(\mathbf{A}) = \mathcal{R}(\mathbf{B})$, then $\mu(\mathbf{A}) = \mu(\mathbf{B})$. Furthermore, it is easy to verify that if $\mathcal{R}(\mathbf{A}) \subseteq \mathcal{R}(\mathbf{B})$, then $\mu(\mathbf{A}) \leq \mu(\mathbf{B})$. We also mention that for every matrix $\mathbf{A}$ with m rows, $\operatorname{rank}(\mathbf{A})/m \leq \mu(\mathbf{A}) \leq 1$.

We focus on tall-and-thin matrices, i.e., matrices with (many) more rows than columns. We are interested in dimensionality reduction techniques that (approximately) preserve the singular values of the original matrix. The simplest idea for dimensionality reduction in tall-and-thin matrices is uniform sampling of the rows of the matrix. Coherence measures how susceptible the matrix is to uniform sampling; the following lemma shows that not too many samples are required when the coherence is "small." The bound is almost tight [33, section 3.3].

LEMMA 2.4 (sampling from orthonormal matrix, corollary to Lemma 3.4 in [33]). *Let* $\mathbf{Q} \in \mathbb{R}^{m \times d}$ *have orthonormal columns. Let* $0 < \varepsilon < 1$ *and* $0 < \delta < 1$. *Let* r *be an integer such that* $6\varepsilon^{-2} m \mu(\mathbf{Q}) \ln(3d/\delta) \leq r \leq m$. *Let* T *be a random subset of* $[m]$ *of cardinality* r, *drawn from a uniform distribution over such subsets (without replacement), and let* $\mathbf{S}$ *be the* $|T| \times m$ *sampling matrix corresponding to* T *rescaled by* $\sqrt{m/r}$. *Then, with probability of at least* $1 - \delta$, *for* $i \in [d]$, $\sqrt{1-\varepsilon} \leq \sigma_i(\mathbf{SQ}) \leq \sqrt{1+\varepsilon}$.

Proof. Apply Lemma 3.4 from [33] with the following choice of parameters: $\ell = \alpha M \ln(k/\delta)$, $\alpha = 6/\varepsilon^2$, and $\delta_{tropp} = \eta = \varepsilon$. Here, ℓ, α, M, k, η are the parameters of Lemma 3.4 from [33]; also δ_{tropp} plays the role of δ, an error parameter, of Lemma 3.4 from [33]. ε and δ are from our lemma. ∎

In the above lemma, T is obtained by sampling coordinates from $[m]$ *without* replacement. Similar results can be shown for sampling with replacement or using Bernoulli variables [21].

2.3. Randomized fast unitary transforms. Matrices with high coherence pose a problem for algorithms based on uniform row sampling. (To see this, notice that in Lemma 2.4 the quality of the bound depends on the coherence of $\mathbf{Q}$.) One way to circumvent this problem is to use a coherence-reducing transformation. It is important that this transformation will not change the solution to the problem.

One popular coherence-reducing method is applying a randomized fast unitary transform. The crucial observation is that many problems can be safely transformed using unitary matrices. This is also true for CCA: $\sigma_i(\mathbf{QA}, \mathbf{QB}) = \sigma_i(\mathbf{A}, \mathbf{B})$ if $\mathbf{Q}$ is unitary (i.e., $\mathbf{Q}^{\mathrm{T}}\mathbf{Q}$ is equal to the identity matrix). If the unitary matrix is chosen carefully, it can reduce the coherence as well. However, any fixed unitary matrix will fail to reduce the coherence on some matrices.

The solution is to couple a fixed unitary transform with some randomization. More specifically, the construction is $\mathcal{F} = \mathbf{FD}$, where $\mathbf{D}$ is a random diagonal matrix of size m whose entries are independent Bernoulli random variables that take values $\{+1, -1\}$ with probability $\frac{1}{2}$, and $\mathbf{F}$ is some fixed unitary matrix. An important quantity is the maximum squared element in $\mathbf{F}$ (we denote this quantity with η): for any fixed $\mathbf{X} \in \mathbb{R}^{m \times n}$ it can be shown that with constant probability, $\mu(\mathcal{F}\mathbf{X}) \leq O(\eta n \log(m))$ [4]. So, it is important for η to be small. It is also necessary that $\mathbf{F}$ can

be applied quickly to $\mathbf{X}$. FFT and FFT-like transforms have both these properties and work well in practice due to the availability of high-quality implementations.

A fast unitary transform that has the above two properties is the Walsh–Hadamard transform (WHT), which is defined as follows. Fix an integer $m = 2^h$ for $h = 1, 2, 3, \ldots$. The (nonnormalized) $m \times m$ matrix of the WHT is defined recursively as

$$\mathbf{H}_m = \begin{bmatrix} \mathbf{H}_{m/2} & \mathbf{H}_{m/2} \\ \mathbf{H}_{m/2} & -\mathbf{H}_{m/2} \end{bmatrix} \text{ with } \mathbf{H}_2 = \begin{bmatrix} +1 & +1 \\ +1 & -1 \end{bmatrix}.$$

The $m \times m$ normalized matrix of the WHT is $\mathbf{H} = m^{-\frac{1}{2}}\mathbf{H}_m$.

The recursive nature of the WHT allows us to compute $\mathbf{HX}$ for an $m \times n$ matrix $\mathbf{X}$ in time $O(mn \log(m))$. However, in our case we are interested in $\mathbf{SHX}$, where $\mathbf{S}$ is an r-row sampling matrix. To compute $\mathbf{SHX}$ only $O(mn \log(r))$ operations suffice [2, Theorem 2.1].

The WHT combined with a random diagonal sign matrix is the RHT.

DEFINITION 2.5 (RHT). *Let $m = 2^h$ for some positive integer h. An* RHT *is an $m \times m$ matrix of the form $\mathbf{\Theta} = \mathbf{HD}$, where $\mathbf{D}$ is a random diagonal matrix of size m whose entries are independent Bernoulli random variables that take values $\{+1, -1\}$ with probability $\frac{1}{2}$, and $\mathbf{H}$ is a normalized Walsh–Hadamard matrix of size m.*

For concreteness, our analysis uses the RHT since it has the tightest coherence reducing bound. Our results generalize to other randomized fast unitary transforms (for example, in our experiments we use the discrete Hartley transform), with bounds that differ up to small constants.

LEMMA 2.6 (RHT bounds coherence, Lemma 3.3 in [33]). *Let $\mathbf{A}$ be an $m \times n$ ($m \geq n$, $m = 2^h$ for some positive integer h) matrix, $0 < \delta < 1$, and let $\mathbf{\Theta}$ be an RHT. Then, with probability of at least $1 - \delta$, $\mu(\mathbf{\Theta A}) \leq \frac{1}{m}(\sqrt{n} + \sqrt{8 \ln(m/\delta)})^2$.*

We remark that the original statement and proof of the lemma [33] is for matrices with orthonormal columns. Since we define the coherence of $\mathbf{A}$ with respect to the thin SVD of $\mathbf{A}$, thereby making the coherence a property of the column space and not of the particular basis, we can state in the above a more general result.

3. Perturbation bounds for matrix products. This section states three new technical lemmas which analyze the perturbation of the singular values of the product of a pair of matrices after dimensionality reduction. These lemmas are essential for our analysis in subsequent sections, but they might be of independent interest. We first state three well-known results.

LEMMA 3.1 (see [15, Theorem 3.3]). *Let $\mathbf{\Psi} \in \mathbb{R}^{p \times q}$ and $\mathbf{\Phi} = \mathbf{D_L \Psi D_R}$ with $\mathbf{D_L} \in \mathbb{R}^{p \times p}$ and $\mathbf{D_R} \in \mathbb{R}^{q \times q}$ being nonsingular matrices. Let*

$$\gamma = \max\{\|\mathbf{D_L D_L^T} - \mathbf{I}_p\|_2, \|\mathbf{D_R^T D_R} - \mathbf{I}_q\|_2\}.$$

Then, for all $i = 1, \ldots, \text{rank}(\mathbf{\Psi}) : |\sigma_i(\mathbf{\Phi}) - \sigma_i(\mathbf{\Psi})| \leq \gamma \cdot \sigma_i(\mathbf{\Psi})$.

LEMMA 3.2 (Weyl's inequality for singular values; [19, Corollary 7.3.8]). *Let $\mathbf{\Phi}, \mathbf{\Psi} \in \mathbb{R}^{m \times n}$. Then, for all $i = 1, \ldots, \min(m, n) : |\sigma_i(\mathbf{\Phi}) - \sigma_i(\mathbf{\Psi})| \leq \|\mathbf{\Phi} - \mathbf{\Psi}\|_2$.*

LEMMA 3.3 (conjugating the PSD ordering; Observation 7.7.2 in [19]). *Let $\mathbf{\Phi}, \mathbf{\Psi} \in \mathbb{R}^{n \times n}$ be symmetric matrices with $\mathbf{\Phi} \preceq \mathbf{\Psi}$. Then, for every $n \times m$ matrix $\mathbf{Z} : \mathbf{Z^T \Phi Z} \preceq \mathbf{Z^T \Psi Z}$.*

We now present the new technical lemmas.

LEMMA 3.4. *Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ ($m \geq n$) and $\mathbf{B} \in \mathbb{R}^{m \times \ell}$ ($m \geq \ell$). Define $\mathbf{C} := [\mathbf{A}; \mathbf{B}] \in \mathbb{R}^{m \times (n+\ell)}$, and suppose $\mathbf{C}$ has rank ω, so $\mathbf{U_C} \in \mathbb{R}^{m \times \omega}$, the U-factor from*

the thin SVD of $\mathbf{C}$*, has* ω *columns. Let* $\mathbf{S} \in \mathbb{R}^{r\times m}$ *be any matrix such that* $\sqrt{1-\varepsilon} \le \sigma_\omega(\mathbf{SU_C}) \le \sigma_1(\mathbf{SU_C}) \le \sqrt{1+\varepsilon}$ *for some* $0 < \varepsilon < 1$*. Then, for* $i = 1, \ldots, \min(n, \ell)$*,*

$$\left|\sigma_i\left(\mathbf{A}^{\mathrm{T}}\mathbf{B}\right) - \sigma_i\left(\mathbf{A}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{SB}\right)\right| \le \varepsilon \cdot \|\mathbf{A}\|_2 \cdot \|\mathbf{B}\|_2.$$

In the above, $\sigma_i(\mathbf{A}^{\mathrm{T}}\mathbf{B})$ *are the diagonal entries of the* $\min(n,\ell) \times \min(n,\ell)$ *diagonal matrix from the full SVD of* $\mathbf{A}^{\mathrm{T}}\mathbf{B}$*, and similarly for* $\sigma_i(\mathbf{A}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{SB})$ *and* $\mathbf{A}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{SB}$*.*

Proof. Let $\mathbf{A} = \mathbf{U_A}\mathbf{\Sigma_A}\mathbf{V_A^{\mathrm{T}}}$ with $\mathbf{U_A} \in \mathbb{R}^{m\times p}$, $\mathbf{\Sigma_A} \in \mathbb{R}^{p\times p}$, and $\mathbf{V_A^{\mathrm{T}}} \in \mathbb{R}^{p\times n}$ be the thin SVD of $\mathbf{A}$ with $p = \operatorname{rank}(\mathbf{A})$. Similarly, let $\mathbf{B} = \mathbf{U_B}\mathbf{\Sigma_B}\mathbf{V_B^{\mathrm{T}}}$ with $\mathbf{U_B} \in \mathbb{R}^{m\times q}$, $\mathbf{\Sigma_B} \in \mathbb{R}^{q\times q}$, and $\mathbf{V_B^{\mathrm{T}}} \in \mathbb{R}^{q\times n}$ be the thin SVD of $\mathbf{B}$ with $q = \operatorname{rank}(\mathbf{B})$. Using Weyl's inequality for the singular values of arbitrary matrices (Lemma 3.2) we obtain, for $i = 1, \ldots, \min(n, \ell)$,

$$\begin{aligned}\left|\sigma_i\left(\mathbf{A}^{\mathrm{T}}\mathbf{B}\right) - \sigma_i\left(\mathbf{A}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{SB}\right)\right| &\le \left\|\mathbf{A}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{SB} - \mathbf{A}^{\mathrm{T}}\mathbf{B}\right\| \\ &= \left\|\mathbf{V_A}\mathbf{\Sigma_A}\left(\mathbf{U_A^{\mathrm{T}}}\mathbf{S}^{\mathrm{T}}\mathbf{SU_B} - \mathbf{U_A^{\mathrm{T}}}\mathbf{U_B}\right)\mathbf{\Sigma_B}\mathbf{V_B^{\mathrm{T}}}\right\| \\ &\le \|\mathbf{U_A^{\mathrm{T}}}\mathbf{S}^{\mathrm{T}}\mathbf{SU_B} - \mathbf{U_A^{\mathrm{T}}}\mathbf{U_B}\|_2 \cdot \|\mathbf{A}\|_2 \cdot \|\mathbf{B}\|_2.\end{aligned}$$

Next, we argue that $\|\mathbf{U_A^{\mathrm{T}}}\mathbf{S}^{\mathrm{T}}\mathbf{SU_B} - \mathbf{U_A^{\mathrm{T}}}\mathbf{U_B}\|_2 \le \|\mathbf{U_C^{\mathrm{T}}}\mathbf{S}^{\mathrm{T}}\mathbf{SU_C} - \mathbf{I}_\omega\|_2$. Indeed,

$$\begin{aligned}\|\mathbf{U_A^{\mathrm{T}}}\mathbf{S}^{\mathrm{T}}\mathbf{SU_B} - \mathbf{U_A^{\mathrm{T}}}\mathbf{U_B}\|_2 &= \sup_{\|\mathbf{w}\|_2=1,\ \|\mathbf{z}\|_2=1} |\mathbf{w}^{\mathrm{T}}\mathbf{U_A^{\mathrm{T}}}\mathbf{S}^{\mathrm{T}}\mathbf{SU_B}\mathbf{z} - \mathbf{w}^{\mathrm{T}}\mathbf{U_A^{\mathrm{T}}}\mathbf{U_B}\mathbf{z}| \\ &= \sup_{\|\mathbf{x}\|_2=\|\mathbf{y}\|_2=1,\ \mathbf{x}\in\mathcal{R}(\mathbf{U_A}),\ \mathbf{y}\in\mathcal{R}(\mathbf{U_B})} |\mathbf{x}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{Sy} - \mathbf{x}^{\mathrm{T}}\mathbf{y}| \\ &\le \sup_{\|\mathbf{x}\|_2=\|\mathbf{y}\|_2=1,\ \mathbf{x}\in\mathcal{R}(\mathbf{U_C}),\ \mathbf{y}\in\mathcal{R}(\mathbf{U_B})} |\mathbf{x}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{Sy} - \mathbf{x}^{\mathrm{T}}\mathbf{y}| \\ &\le \sup_{\|\mathbf{x}\|_2=\|y\|_2=1,\ \mathbf{x}\in\mathcal{R}(\mathbf{U_C}),\ \mathbf{y}\in\mathcal{R}(\mathbf{U_C})} |\mathbf{x}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{Sy} - \mathbf{x}^{\mathrm{T}}\mathbf{y}| \\ &= \sup_{\|\mathbf{w}\|_2=1,\ \|\mathbf{z}\|_2=1} |\mathbf{w}^{\mathrm{T}}\mathbf{U_C^{\mathrm{T}}}\mathbf{S}^{\mathrm{T}}\mathbf{SU_C}\mathbf{z} - \mathbf{w}^{\mathrm{T}}\mathbf{U_C^{\mathrm{T}}}\mathbf{U_C}\mathbf{z}| \\ &= \|\mathbf{U_C^{\mathrm{T}}}\mathbf{S}^{\mathrm{T}}\mathbf{SU_C} - \mathbf{I}_\omega\|_2.\end{aligned}$$

In the above, all the equalities follow by the definition of the spectral norm of a matrix, while the two inequalities follow because $\mathcal{R}(\mathbf{U_A}) \subseteq \mathcal{R}(\mathbf{U_C})$ and $\mathcal{R}(\mathbf{U_B}) \subseteq \mathcal{R}(\mathbf{U_C})$. To conclude, recall that we assumed that for $i \in [\omega]$: $1 - \varepsilon \le \lambda_i(\mathbf{U_C^{\mathrm{T}}}\mathbf{S}^{\mathrm{T}}\mathbf{SU_C}) \le 1 + \varepsilon$. ∎

LEMMA 3.5. *Let* $\mathbf{A} \in \mathbb{R}^{m\times n}$ *(*$m \ge n$*) have rank* p *and* $\mathbf{B} \in \mathbb{R}^{m\times \ell}$ *(*$m \ge \ell$*) have rank* q*. Let* $\mathbf{S} \in \mathbb{R}^{r\times m}$ *(*$r \ge p, q$*) be any matrix such that* $\operatorname{rank}(\mathbf{SA}) = \operatorname{rank}(\mathbf{A})$ *and* $\operatorname{rank}(\mathbf{SB}) = \operatorname{rank}(\mathbf{B})$*. Let* $\mathbf{U_A} \in \mathbb{R}^{m\times p}$ *be the* U*-factor from the thin SVD of* $\mathbf{A}$*, and let* $\mathbf{U_B} \in \mathbb{R}^{m\times q}$ *be the* U*-factor from the thin SVD of* $\mathbf{B}$*. Similarly, let* $\mathbf{U_{SA}} \in \mathbb{R}^{r\times p}$ *be the* U*-factor from the thin SVD of* $\mathbf{SA} \in \mathbb{R}^{r\times p}$*, and let* $\mathbf{U_{SB}} \in \mathbb{R}^{r\times q}$ *be the* U*-factor from the thin SVD of* $\mathbf{SB} \in \mathbb{R}^{r\times q}$*. Let all singular values of* $\mathbf{SU_A}$ *and* $\mathbf{SU_B}$ *be inside* $[\sqrt{1-\varepsilon}, \sqrt{1+\varepsilon}]$ *for some* $0 < \varepsilon < 1/2$*, i.e., for all* $i = 1 : p$ *let* $\sqrt{1-\varepsilon} \le \sigma_i(\mathbf{SU_A}) \le \sqrt{1+\varepsilon}$ *and for all* $i = 1 : q$ *let* $\sqrt{1-\varepsilon} \le \sigma_i(\mathbf{SU_B}) \le \sqrt{1+\varepsilon}$*. Then, for* $i = 1, \ldots, \min(p, q)$*,*

$$\left|\sigma_i\left(\mathbf{U_A^{\mathrm{T}}}\mathbf{S}^{\mathrm{T}}\mathbf{SU_B}\right) - \sigma_i\left(\mathbf{U_{SA}^{\mathrm{T}}}\mathbf{U_{SB}}\right)\right| \le 2\varepsilon\left(1 + \varepsilon\right).$$

Here, $\sigma_i(\mathbf{U_A^{\mathrm{T}}}\mathbf{S}^{\mathrm{T}}\mathbf{SU_B})$ *are the diagonal entries of the* $\min(p,q)\times\min(p,q)$ *diagonal matrix from the full SVD of* $\mathbf{U_A^{\mathrm{T}}}\mathbf{S}^{\mathrm{T}}\mathbf{SU_B}$*, and similarly for* $\sigma_i(\mathbf{U_{SA}^{\mathrm{T}}}\mathbf{U_{SB}})$ *and* $\mathbf{U_{SA}^{\mathrm{T}}}\mathbf{U_{SB}}$*.*

Proof. Let $\mathbf{A} = \mathbf{U_A}\mathbf{\Sigma_A}\mathbf{V_A}^{\mathrm{T}}$ with $\mathbf{U_A} \in \mathbb{R}^{m\times p}$, $\mathbf{\Sigma_A} \in \mathbb{R}^{p\times p}$, and $\mathbf{V_A}^{\mathrm{T}} \in \mathbb{R}^{p\times n}$ be the thin SVD of $\mathbf{A}$ with $p = \mathrm{rank}(\mathbf{A})$. Let $\mathbf{B} = \mathbf{U_B}\mathbf{\Sigma_B}\mathbf{V_B}^{\mathrm{T}}$ with $\mathbf{U_B} \in \mathbb{R}^{m\times q}$, $\mathbf{\Sigma_B} \in \mathbb{R}^{q\times q}$, and $\mathbf{V_B}^{\mathrm{T}} \in \mathbb{R}^{q\times n}$ be the thin SVD of $\mathbf{B}$ with $q = \mathrm{rank}(\mathbf{B})$. Let $\mathbf{SA} = \mathbf{U_{SA}}\mathbf{\Sigma_{SA}}\mathbf{V_{SA}}^{\mathrm{T}}$ with $\mathbf{U_{SA}} \in \mathbb{R}^{r\times p}$, $\mathbf{\Sigma_{SA}} \in \mathbb{R}^{p\times p}$, and $\mathbf{V_{SA}}^{\mathrm{T}} \in \mathbb{R}^{p\times n}$ be the thin SVD of $\mathbf{SA}$ with $p = \mathrm{rank}(\mathbf{SA})$. Let $\mathbf{SB} = \mathbf{U_{SB}}\mathbf{\Sigma_{SB}}\mathbf{V_{SB}}^{\mathrm{T}}$ with $\mathbf{U_{SB}} \in \mathbb{R}^{r\times q}$, $\mathbf{\Sigma_{SB}} \in \mathbb{R}^{q\times q}$, and $\mathbf{V_{SB}}^{\mathrm{T}} \in \mathbb{R}^{q\times n}$ be the thin SVD of $\mathbf{SB}$ with $q = \mathrm{rank}(\mathbf{SB})$. Then, for every $i = 1, \ldots, \min(p, q)$, we have

$$\begin{aligned}
\left|\sigma_i\left(\mathbf{U_A}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{SU_B}\right) - \sigma_i\left(\mathbf{U_{SA}}^{\mathrm{T}}\mathbf{U_{SB}}\right)\right| &= \left|\sigma_i\left(\mathbf{\Sigma_A}^{-1}\mathbf{V_A}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{SBV_B}\mathbf{\Sigma_B}^{-1}\right)\right. \\
&\qquad \left. - \sigma_i\left(\mathbf{\Sigma_{SA}}^{-1}\mathbf{V_{SA}}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{SBV_{SB}}\mathbf{\Sigma_{SB}}^{-1}\right)\right| \\
&\leq \gamma \cdot \sigma_i\left(\mathbf{\Sigma_A}^{-1}\mathbf{V_A}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{SBV_B}\mathbf{\Sigma_B}^{-1}\right) \\
&= \gamma \cdot \sigma_i\left(\mathbf{U_A}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{SU_B}\right) \\
&\leq \gamma \cdot \|\mathbf{U_A}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\|_2 \cdot \sigma_i\left(\mathbf{SU_B}\right) \\
&\leq \gamma \cdot (1+\varepsilon)
\end{aligned}$$

with

$$\gamma = \max(\|\mathbf{\Sigma_{SA}}^{-1}\mathbf{V_{SA}}^{\mathrm{T}}\mathbf{V_A}\mathbf{\Sigma_A}^2\mathbf{V_A}^{\mathrm{T}}\mathbf{V_{SA}}\mathbf{\Sigma_{SA}}^{-1} - \mathbf{I}_p\|_2, \|\mathbf{\Sigma_{SB}}^{-1}\mathbf{V_{SB}}^{\mathrm{T}}\mathbf{V_B}\mathbf{\Sigma_B}^2\mathbf{V_B}^{\mathrm{T}}\mathbf{V_{SB}}\mathbf{\Sigma_{SB}}^{-1} - \mathbf{I}_q\|_2).$$

In the above, the first inequality follows using Lemma 3.1: set

$$\begin{aligned}
\mathbf{\Psi} &= \mathbf{\Sigma_A}^{-1}\mathbf{V_A}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{SBV_B}\mathbf{\Sigma_B}^{-1}, \\
\mathbf{D}_L &:= \mathbf{\Sigma_{SA}}^{-1}\mathbf{V_{SA}}^{\mathrm{T}}\mathbf{V_A}\mathbf{\Sigma_A} \quad \text{and} \quad \mathbf{D}_R := \mathbf{\Sigma_B}\mathbf{V_B}^{\mathrm{T}}\mathbf{V_{SB}}\mathbf{\Sigma_{SB}}^{-1}.
\end{aligned}$$

Moreover,

$$\begin{aligned}
\mathbf{D}_L\Psi\mathbf{D}_R &= \left(\mathbf{\Sigma_{SA}}^{-1}\mathbf{V_{SA}}^{\mathrm{T}}\mathbf{V_A}\mathbf{\Sigma_A}\right)\left(\mathbf{\Sigma_A}^{-1}\mathbf{V_A}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{SBV_B}\mathbf{\Sigma_B}^{-1}\right)\left(\mathbf{\Sigma_B}\mathbf{V_B}^{\mathrm{T}}\mathbf{V_{SB}}\mathbf{\Sigma_{SB}}^{-1}\right) \\
&= \mathbf{\Sigma_{SA}}^{-1}\mathbf{V_{SA}}^{\mathrm{T}}\mathbf{V_A}\mathbf{V_A}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{SBV_B}\mathbf{V_B}^{\mathrm{T}}\mathbf{V_{SB}}\mathbf{\Sigma_{SB}}^{-1} \\
&= \mathbf{\Sigma_{SA}}^{-1}\mathbf{V_{SA}}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{SBV_{SB}}\mathbf{\Sigma_{SB}}^{-1},
\end{aligned}$$

since $\mathbf{A} = \mathbf{AV_A}\mathbf{V_A}^{\mathrm{T}}$ and $\mathbf{B} = \mathbf{BV_B}\mathbf{V_B}^{\mathrm{T}}$.

To apply Lemma 3.1 we need to show that $\mathbf{D}_L$ and $\mathbf{D}_R$ are nonsingular. We will prove that $\mathbf{D}_L$ is nonsingular. (The same argument applies to $\mathbf{D}_R$.) $\mathbf{D}_L$ is nonsingular if and only if $\mathbf{V_{SA}}^{\mathrm{T}}\mathbf{V_A}$ is nonsingular. Since $\mathrm{rank}(\mathbf{V_{SA}}) = \mathrm{rank}(\mathbf{V_A})$, it follows that the range of $\mathbf{V_{SA}}$ equals the range of $\mathbf{V_A}$. So, $\mathbf{V_{SA}} = \mathbf{V_A}\mathbf{W}$ for some unitary matrix $\mathbf{W}$ of size p. $\mathbf{V_{SA}}^{\mathrm{T}}\mathbf{V_A} = \mathbf{W}^{\top}$ and $\mathbf{W}$ is nonsingular and so is $\mathbf{D}_L$.

The second inequality follows because for any two matrices $\mathbf{X}, \mathbf{Y}$: $\sigma_i(\mathbf{XY}) \leq \|\mathbf{X}\|_2\sigma_i(\mathbf{Y})$. Finally, in the third inequality we used the fact that $\|\mathbf{U_A}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\|_2 \leq \sqrt{1+\varepsilon}$ and $\sigma_i(\mathbf{SU_B}) \leq \sqrt{1+\varepsilon}$.

We now bound $\|\mathbf{\Sigma_{SA}}^{-1}\mathbf{V_{SA}}^{\mathrm{T}}\mathbf{V_A}\mathbf{\Sigma_A}^2\mathbf{V_A}^{\mathrm{T}}\mathbf{V_{SA}}\mathbf{\Sigma_{SA}}^{-1} - \mathbf{I}_p\|_2$. (The second term in the max expression of γ can be bounded in a similar fashion, so we omit the proof.)

$$\begin{aligned}
&\|\mathbf{\Sigma_{SA}}^{-1}\mathbf{V_{SA}}^{\mathrm{T}}\mathbf{V_A}\mathbf{\Sigma_A}^2\mathbf{V_A}^{\mathrm{T}}\mathbf{V_{SA}}\mathbf{\Sigma_{SA}}^{-1} - \mathbf{I}_p\|_2 \\
&\quad = \|\mathbf{\Sigma_{SA}}^{-1}\mathbf{V_{SA}}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{AV_{SA}}\mathbf{\Sigma_{SA}}^{-1} - \mathbf{I}_p\|_2 \\
&\quad = \|\mathbf{U_{SA}}^{\mathrm{T}}((\mathbf{SA})^{+})^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{A}(\mathbf{SA})^{+}\mathbf{U_{SA}} - \mathbf{U_{SA}}^{\mathrm{T}}\mathbf{U_{SA}}\mathbf{U_{SA}}^{\mathrm{T}}\mathbf{U_{SA}}\|_2
\end{aligned}$$

$$= \|\mathbf{U}_{\mathbf{SA}}^{\mathrm{T}} \left(((\mathbf{SA})^{+})^{\mathrm{T}} \mathbf{A}^{\mathrm{T}} \mathbf{A} (\mathbf{SA})^{+} - \mathbf{U}_{\mathbf{SA}} \mathbf{U}_{\mathbf{SA}}^{\mathrm{T}} \right) \mathbf{U}_{\mathbf{SA}} \|_2$$
$$\leq \|((\mathbf{SA})^{+})^{\mathrm{T}} \mathbf{A}^{\mathrm{T}} \mathbf{A} (\mathbf{SA})^{+} - \mathbf{U}_{\mathbf{SA}} \mathbf{U}_{\mathbf{SA}}^{\mathrm{T}} \|_2,$$

where we used $\mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{V}_{\mathbf{A}} \mathbf{\Sigma}_{\mathbf{A}}^2 \mathbf{V}_{\mathbf{A}}^{\mathrm{T}}$ and $(\mathbf{SA})^{+} \mathbf{U}_{\mathbf{SA}} = \mathbf{V}_{\mathbf{SA}} \mathbf{\Sigma}_{\mathbf{SA}}^{-1}$. Recall that all the singular values of $\mathbf{SU_A}$ are between $\sqrt{1-\varepsilon}$ and $\sqrt{1+\varepsilon}$, so $(1-\varepsilon)\mathbf{I}_p \preceq \mathbf{U}_{\mathbf{A}}^{\mathrm{T}} \mathbf{S}^{\mathrm{T}} \mathbf{S} \mathbf{U}_{\mathbf{A}} \preceq (1+\varepsilon)\mathbf{I}_p$. Conjugating the above PSD ordering with $\mathbf{\Sigma_A V_A^{\mathrm{T}}} (\mathbf{SA})^{+}$ (see Lemma 3.3), it follows that

$$(1-\varepsilon)((\mathbf{SA})^{+})^{\mathrm{T}} \mathbf{A}^{\mathrm{T}} \mathbf{A} (\mathbf{SA})^{+} \preceq \mathbf{U}_{\mathbf{SA}} \mathbf{U}_{\mathbf{SA}}^{\mathrm{T}} \preceq (1+\varepsilon)((\mathbf{SA})^{+})^{\mathrm{T}} \mathbf{A}^{\mathrm{T}} \mathbf{A} (\mathbf{SA})^{+}$$

since $\mathbf{U}_{\mathbf{A}}^{\mathrm{T}} \mathbf{U}_{\mathbf{A}} = \mathbf{I}_p$ and

$$((\mathbf{SA})^{+})^{\mathrm{T}} \mathbf{A}^{\mathrm{T}} \mathbf{S}^{\mathrm{T}} \mathbf{S} \mathbf{A} (\mathbf{SA})^{+} = \mathbf{U}_{\mathbf{SA}} \mathbf{U}_{\mathbf{SA}}^{\mathrm{T}}.$$

Rearranging terms, it follows that

$$\frac{1}{1+\varepsilon} \mathbf{U}_{\mathbf{SA}} \mathbf{U}_{\mathbf{SA}}^{\mathrm{T}} \preceq ((\mathbf{SA})^{+})^{\mathrm{T}} \mathbf{A}^{\mathrm{T}} \mathbf{A} (\mathbf{SA})^{+} \preceq \frac{1}{1-\varepsilon} \mathbf{U}_{\mathbf{SA}} \mathbf{U}_{\mathbf{SA}}^{\mathrm{T}}.$$

Since $0 < \varepsilon < 1/2$, it holds that $\frac{1}{1-\varepsilon} \leq 1 + 2\varepsilon$ and $\frac{1}{1+\varepsilon} \geq 1 - \varepsilon$, hence

$$-2\varepsilon \mathbf{U}_{\mathbf{SA}} \mathbf{U}_{\mathbf{SA}}^{\mathrm{T}} \preceq ((\mathbf{SA})^{+})^{\mathrm{T}} \mathbf{A}^{\mathrm{T}} \mathbf{A} (\mathbf{SA})^{+} - \mathbf{U}_{\mathbf{SA}} \mathbf{U}_{\mathbf{SA}}^{\mathrm{T}} \preceq 2\varepsilon \mathbf{U}_{\mathbf{SA}} \mathbf{U}_{\mathbf{SA}}^{\mathrm{T}}$$

using standard properties of the PSD ordering. This implies that

$$\|((\mathbf{SA})^{+})^{\mathrm{T}} \mathbf{A}^{\mathrm{T}} \mathbf{A} (\mathbf{SA})^{+} - \mathbf{U}_{\mathbf{SA}} \mathbf{U}_{\mathbf{SA}}^{\mathrm{T}} \|_2 \leq 2\varepsilon \|\mathbf{U}_{\mathbf{SA}} \mathbf{U}_{\mathbf{SA}}^{\mathrm{T}} \|_2 = 2\varepsilon.$$

Indeed, let $\mathbf{x}_+$ be the unit eigenvector of the symmetric matrix

$$((\mathbf{SA})^{+})^{\mathrm{T}} \mathbf{A}^{\mathrm{T}} \mathbf{A} (\mathbf{SA})^{+} - \mathbf{U}_{\mathbf{SA}} \mathbf{U}_{\mathbf{SA}}^{\mathrm{T}}$$

corresponding to its maximum eigenvalue. The PSD ordering implies that

$$\lambda_{\max} \left(((\mathbf{SA})^{+})^{\mathrm{T}} \mathbf{A}^{\mathrm{T}} \mathbf{A} (\mathbf{SA})^{+} - \mathbf{U}_{\mathbf{SA}} \mathbf{U}_{\mathbf{SA}}^{\mathrm{T}} \right) \leq 2\varepsilon \mathbf{x}_+^{\mathrm{T}} \mathbf{U}_{\mathbf{SA}} \mathbf{U}_{\mathbf{SA}}^{\mathrm{T}} \mathbf{x}_+ \leq 2\varepsilon \|\mathbf{U}_{\mathbf{SA}} \mathbf{U}_{\mathbf{SA}}^{\mathrm{T}} \|_2 = 2\varepsilon.$$

Similarly, $\lambda_{\min}(((\mathbf{SA})^{+})^{\mathrm{T}} \mathbf{A}^{\mathrm{T}} \mathbf{A} (\mathbf{SA})^{+} - \mathbf{U}_{\mathbf{SA}} \mathbf{U}_{\mathbf{SA}}^{\mathrm{T}}) \geq -2\varepsilon$, which shows the claim. ∎

LEMMA 3.6. *Assume the conditions of Lemma* 3.4. *Then, for all* $\mathbf{w} \in \mathbb{R}^n$ *and* $\mathbf{y} \in \mathbb{R}^\ell$, *we have* $|\mathbf{w}^{\mathrm{T}} \mathbf{A}^{\mathrm{T}} \mathbf{B} \mathbf{y} - \mathbf{w}^{\mathrm{T}} \mathbf{A}^{\mathrm{T}} \mathbf{S}^{\mathrm{T}} \mathbf{S} \mathbf{B} \mathbf{y}| \leq \varepsilon \cdot \|\mathbf{A}\mathbf{w}\|_2 \cdot \|\mathbf{B}\mathbf{y}\|_2$.

Proof. Let $\mathbf{A} = \mathbf{U}_{\mathbf{A}} \mathbf{\Sigma}_{\mathbf{A}} \mathbf{V}_{\mathbf{A}}^{\mathrm{T}}$ with $\mathbf{U}_{\mathbf{A}} \in \mathbb{R}^{m \times p}$, $\mathbf{\Sigma}_{\mathbf{A}} \in \mathbb{R}^{p \times p}$, and $\mathbf{V}_{\mathbf{A}}^{\mathrm{T}} \in \mathbb{R}^{p \times n}$ be the thin SVD of $\mathbf{A}$ with $p = \mathrm{rank}(\mathbf{A})$. Let $\mathbf{B} = \mathbf{U}_{\mathbf{B}} \mathbf{\Sigma}_{\mathbf{B}} \mathbf{V}_{\mathbf{B}}^{\mathrm{T}}$ with $\mathbf{U}_{\mathbf{B}} \in \mathbb{R}^{m \times q}$, $\mathbf{\Sigma}_{\mathbf{B}} \in \mathbb{R}^{q \times q}$, and $\mathbf{V}_{\mathbf{B}}^{\mathrm{T}} \in \mathbb{R}^{q \times n}$ be the thin SVD of $\mathbf{B}$ with $q = \mathrm{rank}(\mathbf{B})$. Let $\mathbf{E} = \mathbf{U}_{\mathbf{A}}^{\mathrm{T}} \mathbf{S}^{\mathrm{T}} \mathbf{S} \mathbf{U}_{\mathbf{B}} - \mathbf{U}_{\mathbf{A}}^{\mathrm{T}} \mathbf{U}_{\mathbf{B}}$. Now,

$$\begin{aligned}
\left| \mathbf{w}^{\mathrm{T}} \mathbf{A}^{\mathrm{T}} \mathbf{B} \mathbf{y} - \mathbf{w}^{\mathrm{T}} \mathbf{A}^{\mathrm{T}} \mathbf{S}^{\mathrm{T}} \mathbf{S} \mathbf{B} \mathbf{y} \right| &= \left| \mathbf{w}^{\mathrm{T}} \mathbf{V}_{\mathbf{A}} \mathbf{\Sigma}_{\mathbf{A}} \mathbf{E} \mathbf{\Sigma}_{\mathbf{B}} \mathbf{V}_{\mathbf{B}}^{\mathrm{T}} \mathbf{y} \right| \\
&\leq \|\mathbf{w}^{\mathrm{T}} \mathbf{V}_{\mathbf{A}} \mathbf{\Sigma}_{\mathbf{A}}\|_2 \|\mathbf{E}\|_2 \|\mathbf{\Sigma}_{\mathbf{B}} \mathbf{V}_{\mathbf{B}}^{\mathrm{T}} \mathbf{y}\|_2 \\
&= \|\mathbf{w}^{\mathrm{T}} \mathbf{V}_{\mathbf{A}} \mathbf{\Sigma}_{\mathbf{A}} \mathbf{U}_{\mathbf{A}}^{\mathrm{T}}\|_2 \|\mathbf{E}\|_2 \|\mathbf{U}_{\mathbf{B}} \mathbf{\Sigma}_{\mathbf{B}} \mathbf{V}_{\mathbf{B}}^{\mathrm{T}} \mathbf{y}\|_2 \\
&= \|\mathbf{w}^{\mathrm{T}} \mathbf{A}^{\mathrm{T}}\|_2 \|\mathbf{E}\|_2 \|\mathbf{B}\mathbf{y}\|_2 \\
&= \|\mathbf{E}\|_2 \|\mathbf{A}\mathbf{w}\|_2 \|\mathbf{B}\mathbf{y}\|_2.
\end{aligned}$$

Now, the proof of Lemma 3.4 ensures that $\|\mathbf{E}\|_2 \leq \varepsilon$. ∎

4. CCA of row sampled pairs. Given $\mathbf{A}$ and $\mathbf{B}$, one straightforward way to accelerate CCA is to sample rows uniformly from both matrices and to compute the CCA of the smaller matrices. In this section we show that if we sample enough rows, then the canonical correlations of the sampled pair are close to the canonical correlations of the original pair. Furthermore, the canonical weights of the sampled pair can be used to find approximate canonical vectors to the original pair. Not surprisingly, the sample size depends on the coherence. More specifically, it depends on the coherence of $[\mathbf{A};\mathbf{B}]$.

THEOREM 4.1. *Suppose $\mathbf{A} \in \mathbb{R}^{m\times n}$ $(m \geq n)$ has rank p and $\mathbf{B} \in \mathbb{R}^{m\times \ell}$ $(m \geq \ell)$ has rank $q \leq p$. Let $0 < \varepsilon < 1/2$ be an accuracy parameter and $0 < \delta < 1$ be a failure probability parameter. Let $\omega = \operatorname{rank}([\mathbf{A};\mathbf{B}]) \leq p + q$. Let r be an integer such that*

$$54\varepsilon^{-2} m\mu([\mathbf{A};\mathbf{B}]) \ln(12\omega/\delta) \leq r \leq m.$$

Let T be a random subset of $[m]$ of cardinality r, drawn from a uniform distribution over such subsets, and let $\mathbf{S} \in \mathbb{R}^{r\times m}$ be the sampling matrix corresponding to T rescaled by $\sqrt{m/r}$. Denote $\hat{\mathbf{A}} = \mathbf{SA} \in \mathbb{R}^{r\times n}$ and $\hat{\mathbf{B}} = \mathbf{SB} \in \mathbb{R}^{r\times \ell}$.

Let $\sigma_1(\hat{\mathbf{A}},\hat{\mathbf{B}}),\ldots,\sigma_q(\hat{\mathbf{A}},\hat{\mathbf{B}})$ be the exact canonical correlations of $(\hat{\mathbf{A}},\hat{\mathbf{B}})$, and let

$$\mathbf{w}_1 = \hat{\mathbf{x}}_1/\|\hat{\mathbf{A}}\hat{\mathbf{x}}_1\|_2, \ldots, \mathbf{w}_q = \hat{\mathbf{x}}_q/\|\hat{\mathbf{A}}\hat{\mathbf{x}}_q\|_2$$

and

$$\mathbf{p}_1 = \hat{\mathbf{y}}_1/\|\hat{\mathbf{B}}\hat{\mathbf{y}}_1\|_2, \ldots, \mathbf{p}_q = \hat{\mathbf{y}}_q/\|\hat{\mathbf{B}}\hat{\mathbf{y}}_q\|_2$$

be the exact canonical weights (or projection vectors) of $(\hat{\mathbf{A}},\hat{\mathbf{B}})$. Here, $\mathbf{w}_i = \hat{\mathbf{x}}_i/\|\mathbf{A}\hat{\mathbf{x}}_i\|_2 \in \mathbb{R}^n$ and $\mathbf{p}_i = \hat{\mathbf{y}}_i/\|\mathbf{B}\hat{\mathbf{y}}_i\|_2 \in \mathbb{R}^\ell$. Also, $\hat{\mathbf{x}}_i \in \mathbb{R}^r$ and $\hat{\mathbf{y}}_i \in \mathbb{R}^r$ are some vectors chosen as in Definition 1.1*, but for the pair $(\hat{\mathbf{A}},\hat{\mathbf{B}})$.*

Then, with probability of at least $1-\delta$ all the following hold simultaneously:

(a) *(approximation of canonical correlations) For every $i = 1,2,\ldots,q$,*

$$|\sigma_i(\mathbf{A},\mathbf{B}) - \sigma_i\left(\hat{\mathbf{A}},\hat{\mathbf{B}}\right)| \leq \varepsilon + 2\varepsilon^2/9 = O(\varepsilon).$$

(b) *(approximate orthonormal bases) The vectors $\{\mathbf{Aw}_i\}_{i\in[q]} \in \mathbb{R}^m$ form an approximately orthonormal basis. That is, for any $c \in [q]$,*

$$\frac{1}{1+\varepsilon/3} \leq \|\mathbf{Aw}_c\|_2^2 \leq \frac{1}{1-\varepsilon/3},$$

and for any $i \neq j$,

$$|\langle \mathbf{Aw}_i,\ \mathbf{Aw}_j\rangle| \leq \frac{\varepsilon}{3-\varepsilon}.$$

In particular, $\|\mathbf{W}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{AW} - \mathbf{I}_q\|_2 \leq 2\varepsilon q/3$, where $\mathbf{W} = [\mathbf{w}_1,\ldots,\mathbf{w}_q]$, and similarly for the set of $\{\mathbf{Bp}_i\}_{i\in[q]} \in \mathbb{R}^m$.

(c) *(approximate correlation) For every $i = 1,2,\ldots,q$,*

$$\frac{\sigma_i(\mathbf{A},\mathbf{B})}{1+\varepsilon/3} - 2\varepsilon \leq \sigma(\mathbf{Aw}_i,\mathbf{Bp}_i) \leq \frac{\sigma_i(\mathbf{A},\mathbf{B})}{1-\varepsilon/3} + 2\varepsilon.$$

Proof. Let $\mathbf{A} = \mathbf{U_A}\mathbf{\Sigma_A}\mathbf{V_A}^{\mathrm{T}}$ with $\mathbf{U_A} \in \mathbb{R}^{m\times p}$, $\mathbf{\Sigma_A} \in \mathbb{R}^{p\times p}$, and $\mathbf{V_A}^{\mathrm{T}} \in \mathbb{R}^{p\times n}$ be the thin SVD of $\mathbf{A}$ with $p = \mathrm{rank}(\mathbf{A})$. Let $\mathbf{B} = \mathbf{U_B}\mathbf{\Sigma_B}\mathbf{V_B}^{\mathrm{T}}$ with $\mathbf{U_B} \in \mathbb{R}^{m\times q}$, $\mathbf{\Sigma_B} \in \mathbb{R}^{q\times q}$, and $\mathbf{V_B}^{\mathrm{T}} \in \mathbb{R}^{q\times n}$ be the thin SVD of $\mathbf{B}$ with $q = \mathrm{rank}(\mathbf{B})$. Let $\mathbf{C} := [\mathbf{U_A}; \mathbf{U_B}] \in \mathbb{R}^{m\times(p+q)}$. Let $\mathbf{C} = \mathbf{U_C}\mathbf{\Sigma_C}\mathbf{V_C}^{\mathrm{T}}$ with $\mathbf{U_C} \in \mathbb{R}^{m\times\omega}$, $\mathbf{\Sigma_C} \in \mathbb{R}^{\omega\times\omega}$, and $\mathbf{V_C}^{\mathrm{T}} \in \mathbb{R}^{\omega\times n}$ be the thin SVD of $\mathbf{C}$ with $\omega = \mathrm{rank}(\mathbf{C}) \geq \mathrm{rank}(\mathbf{B}) \geq \mathrm{rank}(\mathbf{B})$. Lemma 2.4 implies that each of the following three assertions holds with probability of at least $1-\delta/3$; hence, by the union bound, all three events hold simultaneously with probability of at least $1-\delta$:

- For every $r \in [p]$, $\sqrt{1-\varepsilon/3} \leq \sigma_r(\mathbf{SU_A}) \leq \sqrt{1+\varepsilon/3}$.
- For every $k \in [q]$, $\sqrt{1-\varepsilon/3} \leq \sigma_k(\mathbf{SU_B}) \leq \sqrt{1+\varepsilon/3}$.
- For every $h \in [\omega]$, $\sqrt{1-\varepsilon/3} \leq \sigma_h(\mathbf{SU_C}) \leq \sqrt{1+\varepsilon/3}$.

We now show that if indeed all three events hold, then (a)–(c) hold as well.

Proof of (a). Corollary 2.2 implies that for all $i = 1 : \min(p,q)$, $\sigma_i(\mathbf{A},\mathbf{B}) = \sigma_i(\mathbf{U_A}^{\mathrm{T}}\mathbf{U_B})$ and $\sigma_i(\mathbf{SA},\mathbf{SB}) = \sigma_i(\mathbf{U_{SA}}^{\mathrm{T}}\mathbf{U_{SB}})$. Using the triangle inequality we get $|\sigma_i(\mathbf{A},\mathbf{B}) - \sigma_i(\mathbf{SA},\mathbf{SB})| = |\sigma_i(\mathbf{U_A}^{\mathrm{T}}\mathbf{U_B}) - \sigma_i(\mathbf{U_{SA}}^{\mathrm{T}}\mathbf{U_{SB}})| \leq |\sigma_i(\mathbf{U_A}^{\mathrm{T}}\mathbf{U_B}) - \sigma_i(\mathbf{U_A}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{SU_B})| + |\sigma_i(\mathbf{U_A}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{SU_B}) - \sigma_i(\mathbf{U_{SA}}^{\mathrm{T}}\mathbf{U_{SB}})|$.

In the above, $\sigma_i(\mathbf{U_A}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{SU_B})$ are the diagonal elements of the $\min(p,q)\times\min(p,q)$ matrix from the full SVD of $\mathbf{U_A}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{SU_B} \in \mathbb{R}^{p\times q}$, and similarly for $\sigma_i(\mathbf{U_{SA}}^{\mathrm{T}}\mathbf{U_{SB}})$ and $\mathbf{U_{SA}}^{\mathrm{T}}\mathbf{U_{SB}}$. To conclude the proof, use Lemmas 3.4 and 3.5 to bound these two terms, respectively.

Proof of (b). For any $c \in [q]$, $\|\mathbf{Aw}_c\|_2 = \|\mathbf{Aw}_c\|_2/\|\hat{\mathbf{A}}\mathbf{w}_c\|_2$, since $\|\hat{\mathbf{A}}\mathbf{w}_c\|_2 = 1$. Now Lemma 3.6 implies the first inequality. For any $i \neq j$

$$\begin{aligned}
|\langle \mathbf{Aw}_i,\ \mathbf{Aw}_j\rangle| &\leq |\mathbf{w}_i^{\mathrm{T}}\hat{\mathbf{A}}^{\mathrm{T}}\hat{\mathbf{A}}\mathbf{w}_j| + |\mathbf{w}_i^{\mathrm{T}}(\hat{\mathbf{A}}^{\mathrm{T}}\hat{\mathbf{A}} - \mathbf{A}^{\mathrm{T}}\mathbf{A})\mathbf{w}_j| \\
&= |\mathbf{w}_i^{\mathrm{T}}(\hat{\mathbf{A}}^{\mathrm{T}}\hat{\mathbf{A}} - \mathbf{A}^{\mathrm{T}}\mathbf{A})\mathbf{w}_j| \\
&\leq \frac{\varepsilon}{3}\|\mathbf{Aw}_i\|_2\|\mathbf{Aw}_j\|_2 \\
&\leq \frac{\varepsilon/3}{1-\epsilon/3}\|\hat{\mathbf{A}}\mathbf{w}_i\|_2\|\hat{\mathbf{A}}\mathbf{w}_j\|_2 \\
&= \frac{\varepsilon}{3-\epsilon}.
\end{aligned}$$

In the above, we used the triangle inequality, the fact that the $\mathbf{w}_i$'s are the canonical weights of $\hat{\mathbf{A}}$, and Lemma 3.6.

The norm bound follows since the maximum entry (in absolute value) of the matrix $\mathbf{W}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{AW} - \mathbf{I}_q$ is at most $\max\{2\varepsilon/3, \varepsilon/3\} = 2\varepsilon/3$.

Proof of (c). We prove only the upper bound. The lower bound is similar, and we omit it.

$$\begin{aligned}
\sigma(\mathbf{Aw}_i, \mathbf{Bp}_i) &= \frac{\langle \mathbf{Aw}_i,\ \mathbf{Bp}_i\rangle}{\|\mathbf{Aw}_i\|_2\|\mathbf{Bp}_i\|_2} \ \leq\ \frac{1}{1-\varepsilon/3}\cdot\langle \mathbf{Aw}_i,\ \mathbf{Bp}_i\rangle \\
&= \frac{1}{1-\varepsilon/3}\cdot\left(\left\langle \hat{\mathbf{A}}\mathbf{w}_i,\ \hat{\mathbf{B}}\mathbf{p}_i\right\rangle + \mathbf{w}_i^{\mathrm{T}}\left(\mathbf{A}^{\mathrm{T}}\mathbf{B} - \hat{\mathbf{A}}^{\mathrm{T}}\hat{\mathbf{B}}\right)\mathbf{p}_i\right) \\
&\leq \frac{\sigma\left(\hat{\mathbf{A}}\mathbf{w}_i, \hat{\mathbf{B}}\mathbf{p}_i\right)}{1-\varepsilon/3} + \frac{\varepsilon/3}{1-\varepsilon/3}\cdot\|\mathbf{Aw}_i\|_2\cdot\|\mathbf{Bp}_i\|_2 \\
&\leq \frac{\sigma\left(\hat{\mathbf{A}}\mathbf{w}_i, \hat{\mathbf{B}}\mathbf{p}_i\right)}{1-\varepsilon/3} + \frac{\varepsilon/3}{(1-\varepsilon/3)^2} \ =\ \frac{\sigma_i\left(\hat{\mathbf{A}}, \hat{\mathbf{B}}\right)}{1-\varepsilon/3} + \frac{\varepsilon/3}{(1-\varepsilon/3)^2}
\end{aligned}$$

$$\leq \frac{\sigma_i(\mathbf{A},\mathbf{B})}{1-\varepsilon/3} + \frac{\varepsilon + 2\varepsilon^2/9}{1-\varepsilon/3} + \frac{\varepsilon/3}{(1-\varepsilon/3)^2}$$
$$\leq \frac{\sigma_i(\mathbf{A},\mathbf{B})}{1-\varepsilon/3} + 2\varepsilon.$$

In the above, the first equality follows by the definition of $\sigma(\cdot,\cdot)$, whereas the first inequality holds since $1/\|\mathbf{A}\mathbf{w}_i\|_2^2 \leq 1+\epsilon/3$ and $1/\|\mathbf{B}\mathbf{p}_i\|_2^2 \leq 1+\epsilon/3$ from part (b) and then using the fact that $1+\epsilon/3 \leq 1/(1-\epsilon/3)$. The second inequality follows from Lemma 3.6 applied on $\mathbf{w}_i^{\mathrm{T}}(\mathbf{A}^{\mathrm{T}}\mathbf{B} - \hat{\mathbf{A}}^{\mathrm{T}}\hat{\mathbf{B}})\mathbf{p}_i$, and the third inequality follows since $(1-\epsilon)\|\mathbf{A}\mathbf{w}_i\|_2^2 \leq \|\hat{\mathbf{A}}\mathbf{w}_i\|_2^2 = 1$ (the same holds for $\mathbf{B}\mathbf{p}_i$). The fourth inequality follows by part (a), and the last inequality follows by elementary algebraic manipulations and the assumption $\varepsilon < 1/2$. □

5. Fast approximate CCA. First, we define "approximate CCA."

DEFINITION 5.1 (approximate CCA). *For $0 \leq \eta \leq 1$, an η-*approximate CCA of $(\mathbf{A},\mathbf{B})$ *is a set of positive numbers $\hat{\sigma}_1,\ldots,\hat{\sigma}_q$ together with a set of vectors $\mathbf{w}_1,\ldots,\mathbf{w}_q \in \mathbb{R}^{m\times n}$ (for $\mathbf{A}\in\mathbb{R}^n$ of rank p) and a set of vectors $\mathbf{p}_1,\ldots,\mathbf{p}_q \in \mathbb{R}^\ell$ (for $\mathbf{B}\in\mathbb{R}^{m\times\ell}$ of rank $q \leq p$), such that the following hold:*

(a) *For every $i \in [q]$,*

$$|\sigma_i(\mathbf{A},\mathbf{B}) - \hat{\sigma}_i| \leq \eta.$$

(b) *For every $i \in [q]$,*

$$|\|\mathbf{A}\mathbf{w}_i\|_2^2 - 1| \leq \eta,$$

and for $i \neq j$,

$$|\langle \mathbf{A}\mathbf{w}_i,\ \mathbf{A}\mathbf{w}_j\rangle| \leq \eta.$$

In particular, $\|\mathbf{W}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{W} - \mathbf{I}_q\| \leq \eta q$, where $\mathbf{W} = [\mathbf{w}_1,\ldots,\mathbf{w}_q]$, and similarly for the set of $\{\mathbf{B}\mathbf{p}_i : i \in [q]\}$.

(c) *For every $i \in [q]$,*

$$|\sigma_i(\mathbf{A},\mathbf{B}) - \sigma(\mathbf{A}\mathbf{w}_i,\mathbf{B}\mathbf{p}_i)| \leq \eta.$$

We are now ready to present our fast algorithm for approximate CCA of a pair of tall-and-thin matrices. Algorithm 1 gives the pseudocode description of our algorithm.

The analysis in the previous section (Theorem 4.1) shows that if we sample enough rows, the canonical correlations and weights of the sampled matrices are an $O(\varepsilon)$-approximate CCA of $(\mathbf{A},\mathbf{B})$. However, to turn this observation into a concrete algorithm we need an upper bound on the coherence of $[\mathbf{A};\mathbf{B}]$. It is conceivable that in certain scenarios such an upper bound might be known in advance, or that it can be computed quickly [12]. However, even if we know the coherence, it might be as large as one, which will imply that sampling the entire matrix is needed.

To circumvent this problem, our algorithm uses the RHT to reduce the coherence of the matrix pair before sampling rows from it. That is, instead of sampling rows from $(\mathbf{A},\mathbf{B})$ we sample rows from $(\mathbf{\Theta}\mathbf{A},\mathbf{\Theta}\mathbf{B})$, where $\mathbf{\Theta}$ is an RHT matrix (Definition 2.5). This unitary transformation bounds the coherence with high probability, so we can use Theorem 4.1 to compute the number of rows required for an $O(\varepsilon)$-approximate CCA. We now sample the transformed pair $(\mathbf{\Theta}\mathbf{A},\mathbf{\Theta}\mathbf{B})$ to obtain $(\hat{\mathbf{A}},\hat{\mathbf{B}})$. Now the canonical correlations and weights of $(\hat{\mathbf{A}},\hat{\mathbf{B}})$ are computed and returned.

ALGORITHM 1. Fast approximate CCA.

1: **Input:** $\mathbf{A} \in \mathbb{R}^{m\times n}$ of rank at most p, $\mathbf{B} \in \mathbb{R}^{m\times \ell}$ of rank at most q, $0 < \varepsilon < 1/2$, and δ ($n \geq l$, $p \geq q$). Here $m = 2^h$ for some positive integer $h > 0$.
2: $r \longleftarrow \min(54\varepsilon^{-2}[\sqrt{n+\ell} + \sqrt{8\ln(12m/\delta)}]^2 \ln(3(n+\ell)/\delta), m)$.
3: Let $\mathbf{S}$ be the sampling matrix of a random subset of $[m]$ of cardinality r (uniform distribution).
4: Draw a random diagonal matrix $\mathbf{D}$ of size m with ± 1 on its diagonal with equal probability.
5: $\hat{\mathbf{A}} \longleftarrow \mathbf{SH} \cdot (\mathbf{DA})$ using fast subsampled WHT (see section 2.3).
6: $\hat{\mathbf{B}} \longleftarrow \mathbf{SH} \cdot (\mathbf{DB})$ using fast subsampled WHT (see section 2.3).
7: Compute and return the canonical correlations and the canonical weights of $(\hat{\mathbf{A}}, \hat{\mathbf{B}})$ (e.g., using Björck and Golub's algorithm) as an approximate CCA to $(\mathbf{A}, \mathbf{B})$.

THEOREM 5.2. *With probability of at least* $1-\delta$, *Algorithm* 1 *returns an* $O(\varepsilon)$*-approximate CCA of* $(\mathbf{A}, \mathbf{B})$. *Assuming Björck and Golub's algorithm is used in line* 7, *Algorithm* 1 *runs in time* $O(\min\{mn\ln m + \varepsilon^{-2}[\sqrt{n} + \sqrt{\ln(m/\delta)}]^2 \ln(n/\delta)n^2, mn^2\})$.

Proof. Lemma 2.6 ensures that with probability of at least $1-\delta/2$, $\mu([\mathbf{\Theta A}; \mathbf{\Theta B}]) \leq \frac{1}{m}(\sqrt{n+\ell} + \sqrt{8\ln(3m/\delta)})^2$. Assuming that the last inequality holds, Theorem 4.1 ensures that with probability of at least $1-\delta/2$, the canonical correlations and weights of $(\hat{\mathbf{A}}, \hat{\mathbf{B}})$ form an $O(\varepsilon)$-approximate CCA of $(\mathbf{\Theta A}, \mathbf{\Theta B})$. By the union bound, both events hold together with probability of at least $1-\delta$. The RHT transforms applied to $\mathbf{A}$ and $\mathbf{B}$ are unitary, so for every η, an η-approximate CCA of $(\mathbf{\Theta A}, \mathbf{\Theta B})$ is also an η-approximate CCA of $(\mathbf{A}, \mathbf{B})$ (and vice versa).

Running time analysis. Step 2 takes $O(1)$ operations. Step 3 requires $O(r)$ operations. Step 4 requires $O(m)$ operations. Step 5 involves the multiplication of $\mathbf{A}$ with $\mathbf{SHD}$ from the left. Computing $\mathbf{DA}$ requires $O(mn)$ time. Multiplying $\mathbf{SH}$ by $\mathbf{DA}$ using fast subsampled WHT requires $O(mn\ln r)$ time, as explained in section 2.3. Similarly, step 6 requires $O(m\ell\ln r)$ operations. Finally, step 7 takes $O(rn\ell + r(n^2 + \ell^2))$ time. Assuming that $n \geq \ell$, the total running time is $O(rn^2 + mn\ln(r))$. Plugging in the value for r, and using the fact that $r \leq m$, establishes our running time bound. ∎

We remark that while Algorithm 1 is often faster than Björck and Golub's algorithm, it is not guaranteed to always be so. In fact, if $r = m$, then the algorithm is slower due to additional operations. Obviously, since $r = \Omega((n+\ell)\ln(n+\ell))$, the algorithm is beneficial only if m is very large while $n + \ell$ is small.

6. Fast approximate CCA using other transforms. Our discussion so far has focused on the case in which we reduce the dimensions of $\mathbf{A}$ and $\mathbf{B}$ via the SRHT. In recent years, several similar transforms have been suggested by various researchers. For example, one can use the fast Johnson–Lindenstraus method of Ailon and Chazelle [1]. This transform leads to an approximate CCA algorithm with a similar additive error guarantee and running time as in Theorem 5.2.

Recently, Clarkson and Woodruff described a transform that is particularly appealing if the input matrices $\mathbf{A}$ and $\mathbf{B}$ are sparse [9]. We present this transform in the following lemma along with theoretical guarantees similar to those of Lemma 2.4. The following lemma is due to [23] (see also [24]), which analyzed the transform originally due to [9]. We only employ the lemma due to Meng and Mahoney [23] because it provides explicit constants compared to the original result due to Clarkson and Woodruff, which is stated in asymptotic notation [9, Theorem 19].

ALGORITHM 2. Fast approximate CCA using the Clarskon–Woodruff transform [9].

1: **Input:** $\mathbf{A} \in \mathbb{R}^{m\times n}$ of rank at most p, $\mathbf{B} \in \mathbb{R}^{m\times \ell}$ of rank at most q, $0 < \varepsilon < 1/3$, and δ $(n \geq l,\, p \geq q)$.

2: $r \longleftarrow \min(\frac{243((n+\ell)^2+n+\ell)}{\varepsilon^2\delta}, m)$.

3: Let $\mathbf{S}$ be an $r \times m$ matrix constructed as follows: it has each column chosen independently and uniformly from the r standard basis vectors of $\mathbb{R}^r$.

4: Draw a random diagonal matrix $\mathbf{D}$ of size m with ± 1 on its diagonal with equal probability.

5: $\hat{\mathbf{A}} \longleftarrow \mathbf{S} \cdot (\mathbf{DA})$.

6: $\hat{\mathbf{B}} \longleftarrow \mathbf{S} \cdot (\mathbf{DB})$.

7: Compute and return the canonical correlations and the canonical weights of $(\hat{\mathbf{A}}, \hat{\mathbf{B}})$ (e.g., using Björck and Golub's algorithm) as an approximate CCA to $(\mathbf{A}, \mathbf{B})$.

LEMMA 6.1 (Theorem 1 in [23] with ε, δ replaced with $\varepsilon/3$, $\delta/3$, respectively). *Given any matrix $\mathbf{X} \in \mathbb{R}^{m\times d}$ with $m \gg d$, accuracy parameter $0 < \varepsilon < 1/3$, and failure probability parameter $0 < \delta < 1$, let $r \geq \lceil (243(d^2+d))/(\varepsilon^2\delta) \rceil$. Construct an $r \times m$ matrix $\mathbf{\Omega}$ as follows: $\mathbf{\Omega} = \mathbf{SD}$, where $\mathbf{S} \in \mathbb{R}^{r\times m}$ has each column chosen independently and uniformly from the r standard basis vectors of $\mathbb{R}^r$ and $\mathbf{D} \in \mathbb{R}^{m\times m}$ is a diagonal matrix with diagonal entries chosen independently and uniformly from $\{+1, -1\}$. Then with probability at least $1-\delta/3$, for every $j \in [d]$, $\sqrt{1-\varepsilon/3}\cdot\sigma_j(\mathbf{X}) \leq \sigma_j(\mathbf{\Omega X}) \leq \sqrt{1+\varepsilon/3}\cdot\sigma_j(\mathbf{X})$. Moreover, $\mathbf{\Omega X}$ can be calculated in $O(\mathrm{nnz}(\mathbf{X}))$ time.*

Similar to Theorem 4.1, we have the following theorem.

THEOREM 6.2. *Suppose $\mathbf{A} \in \mathbb{R}^{m\times n}$ $(m \geq n)$ has rank p and $\mathbf{B} \in \mathbb{R}^{m\times \ell}$ $(m \geq \ell)$ has rank $q \leq p$. Let $0 < \varepsilon < 1/3$ be an accuracy parameter and $0 < \delta < 1$ be a failure probability parameter. Let $\omega = \mathrm{rank}([\mathbf{A}; \mathbf{B}]) \leq p+q$. Let r be an integer such that $\frac{243(\omega^2+\omega)}{\varepsilon^2\delta} \leq r \leq m$. Let $\mathbf{\Omega} \in \mathbb{R}^{r\times m}$ be constructed as in Lemma 6.1. Denote $\hat{\mathbf{A}} = \mathbf{\Omega A} \in \mathbb{R}^{r\times n}$ and $\hat{\mathbf{B}} = \mathbf{\Omega B} \in \mathbb{R}^{r\times \ell}$. Let $\sigma_1(\hat{\mathbf{A}}, \hat{\mathbf{B}}), \ldots, \sigma_q(\hat{\mathbf{A}}, \hat{\mathbf{B}})$ be the exact canonical correlations of $(\hat{\mathbf{A}}, \hat{\mathbf{B}})$, and let $\mathbf{w}_1 = \hat{\mathbf{x}}_1/\|\hat{\mathbf{A}}\hat{\mathbf{x}}_1\|_2, \ldots, \mathbf{w}_q = \hat{\mathbf{x}}_q/\|\hat{\mathbf{A}}\hat{\mathbf{x}}_q\|_2$, and $\mathbf{p}_1 = \hat{\mathbf{y}}_1/\|\hat{\mathbf{B}}\hat{\mathbf{y}}_1\|_2, \ldots, \mathbf{p}_q = \hat{\mathbf{y}}_q/\|\hat{\mathbf{B}}\hat{\mathbf{y}}_q\|_2$ be the exact canonical weights (or projection vectors) of $(\hat{\mathbf{A}}, \hat{\mathbf{B}})$. Here, $\mathbf{w}_i = \hat{\mathbf{x}}_i/\|\mathbf{A}\hat{\mathbf{x}}_i\|_2 \in \mathbb{R}^n$ and $\mathbf{p}_i = \hat{\mathbf{y}}_i/\|\mathbf{B}\hat{\mathbf{y}}_i\|_2 \in \mathbb{R}^\ell$. Also, $\hat{\mathbf{x}}_i \in \mathbb{R}^r$ and $\hat{\mathbf{y}}_i \in \mathbb{R}^r$ are some vectors chosen as in Definition 1.1 but for $(\hat{\mathbf{A}}, \hat{\mathbf{B}})$. With probability of at least $1-\delta$,* (a), (b), *and* (c) *of Theorem* 4.1 *hold simultaneously.*

Proof. Let $\mathbf{A} = \mathbf{U_A}\mathbf{\Sigma_A}\mathbf{V_A^T}$ with $\mathbf{U_A} \in \mathbb{R}^{m\times p}$, $\mathbf{\Sigma_A} \in \mathbb{R}^{p\times p}$, and $\mathbf{V_A^T} \in \mathbb{R}^{p\times n}$ be the thin SVD of $\mathbf{A}$ with $p = \mathrm{rank}(\mathbf{A})$. Let $\mathbf{B} = \mathbf{U_B}\mathbf{\Sigma_B}\mathbf{V_B^T}$ with $\mathbf{U_B} \in \mathbb{R}^{m\times q}$, $\mathbf{\Sigma_B} \in \mathbb{R}^{q\times q}$, and $\mathbf{V_B^T} \in \mathbb{R}^{q\times n}$ be the thin SVD of $\mathbf{B}$ with $q = \mathrm{rank}(\mathbf{B})$. Let $\mathbf{C} := [\mathbf{U_A}; \mathbf{U_B}]$. Let $\mathbf{C} = \mathbf{U_C}\mathbf{\Sigma_C}\mathbf{V_C^T}$ with $\mathbf{U_C} \in \mathbb{R}^{m\times\omega}$, $\mathbf{\Sigma_C} \in \mathbb{R}^{\omega\times\omega}$, and $\mathbf{V_C^T} \in \mathbb{R}^{\omega\times n}$ be the thin SVD of $\mathbf{C}$ with $\omega = \mathrm{rank}(\mathbf{C})$. Lemma 6.1 implies that each of the following three assertions hold with probability of at least $1-\delta/3$, and hence all three hold simultaneously with probability of at least $1-\delta$:

- For every $r \in [p]$, $\sqrt{1-\varepsilon/3} \leq \sigma_r(\mathbf{\Omega U_A}) \leq \sqrt{1+\varepsilon/3}$.
- For every $k \in [q]$, $\sqrt{1-\varepsilon/3} \leq \sigma_k(\mathbf{\Omega U_B}) \leq \sqrt{1+\varepsilon/3}$.
- For every $h \in [\omega]$, $\sqrt{1-\varepsilon/3} \leq \sigma_h(\mathbf{\Omega U_C}) \leq \sqrt{1+\varepsilon/3}$.

Recall that in the proof of Theorem 4.1 we have shown that if indeed all three hold, then (a)–(c) hold as well. ∎

Finally, similar to Theorem 5.2 we have the following theorem for approximate CCA (see also Algorithm 2).

THEOREM 6.3. *With probability of at least* $1-\delta$*, Algorithm* 2 *returns an* $O(\varepsilon)$*-approximate CCA of* $(\mathbf{A},\mathbf{B})$*. Assuming Björck and Golub's algorithm is used in line* 7*, Algorithm* 2 *runs in time* $O(\min\{m+\text{nnz}(\mathbf{A})+\text{nnz}(\mathbf{B})+n^4\varepsilon^{-2}\delta^{-1}, mn^2\})$.

Proof. The bound is immediate from Theorem 6.2 since $n+\ell \geq \omega$. So, we only need to analyze the running time. Step 2 takes $O(1)$ operations. Step 3 requires $O(m)$ operations. Step 4 requires $O(m)$ operations as well. Step 5 involves the multiplication of $\mathbf{A}$ with $\mathbf{SD}$ from the left. Lemma 6.1 argues that this can be accomplished in $O(\text{nnz}(\mathbf{A}))$ arithmetic operations. Similarly, step 6 requires $O(\text{nnz}(\mathbf{B}))$ operations. Finally, step 7 takes $O(rn\ell + r(n^2+\ell^2))$ arithmetic operations. Assuming that $n \geq \ell$, the total running time is $O(m+\text{nnz}(\mathbf{A})+\text{nnz}(\mathbf{B})+rn^2)$. Plugging the value for $r = (243((n+\ell)^2+n+\ell))/(\varepsilon^2\delta)$ and using again that $n \geq \ell$ establishes the bound. ∎

Again we remark that while Algorithm 2 is often faster than Björck and Golub's algorithm, it is not guaranteed to always be so. In fact, if $r = m$, then the algorithm is slower due to additional operations. Obviously, since $r = \Omega((n+\ell)^2)$ the algorithm is beneficial only if m is very large while $n+\ell$ is small.

6.1. Sufficient properties of a dimension reduction transform. We stress that any matrix Ω that satisfies the three conditions appearing in the beginning of the proof of Theorem 6.2 can be used in our framework.

For example, it might be possible to design a dimensionality reduction algorithm based on "subspace sampling" (also known as "leverage-scores sampling") [14].

7. Experiments. In this section we report the results of a few small-scale experiments. Our experiments are not meant to be exhaustive. However, they do show that our algorithm can be modified slightly to achieve very good running time performance in practice while still producing acceptable approximation results.

Our implementation of Algorithm 1 differs from the pseudocode description in two ways. First, we use

$$r \longleftarrow \min(\varepsilon^{-2}\left[\sqrt{n+\ell}+\sqrt{\ln(m/\delta)}\right]^2 \ln(n+\ell)/\delta), m) \tag{7.1}$$

for setting the sample size in all the following experiments, i.e., we keep the same asymptotic behavior but drop the constants. The constants in Algorithm 1 are rather large, so they might preclude the possibility of beating Björck and Golub's algorithm for reasonable matrix sizes. Our implementation also differs in the choice of the underlying coherence-reduction matrix. Algorithm 1, and the analysis, uses the WHT. However, as we discussed in section 2.3, other Fourier-type transforms will work as well and some of these alternative choices have certain advantages that make them better suited for an actual implementation [4]. Specifically, we use the implementation of the normalized randomized discrete Hartley transform (DHT) in the Blendenpik library [4].[1] The DHT is a matrix $\mathbf{H}_m \in \mathbb{R}^{m\times m}$, where for all i,j: $\mathbf{H}_m(i,j) = \cos(\frac{2\cdot\pi\cdot i\cdot j}{m}) + \sin(\frac{2\cdot\pi\cdot i\cdot j}{m})$. Hence, the normalized randomized DHT is $\sqrt{m/r}\mathbf{DH}_m$, where $\mathbf{D}$ is a random diagonal matrix of size m whose entries are independent Bernoulli random variables that take values $\{+1,-1\}$ with probability $1/2$.

We report the results of three experiments. In each experiment, we run our code five times on a fixed pair of matrices (datasets) $\mathbf{A}$ and $\mathbf{B}$ and compare the different

[1] Available at http://www.mathworks.com/matlabcentral/fileexchange/25241-blendenpik.

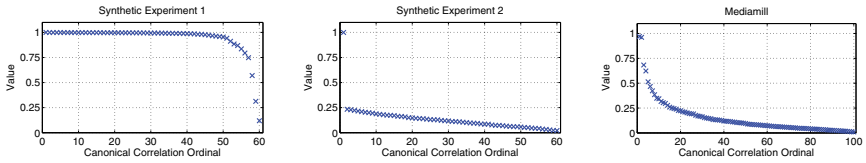

FIG. 1. *The exact canonical correlations.*

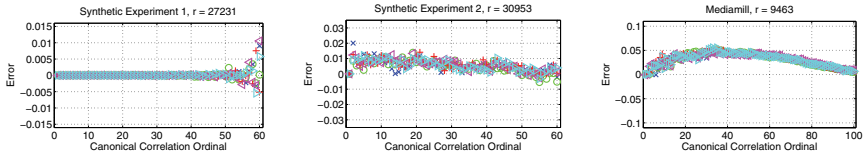

FIG. 2. *Error in approximation of the canonical correlations using Algorithm* 1 *(*r *as in* (7.1)*).*

outputs to the true canonical correlations. The first two experiments involved synthetic datasets, for which we set $\varepsilon = 0.25$ and $\delta = 0.05$. The last experiment was conducted on a real-life dataset, and we used $\varepsilon = 0.5$ and $\delta = 0.2$. All experiments were conducted in a 64-bit version of MATLAB 7.9. We used a Lenovo W520 Thinkpad, Intel Core i7-2760QM CPU running at 2.40 GHz, with 8GB RAM, running Linux 3.5. The measured running times are wall-clock times and were measured using the **ftime** Linux system call.

7.1. Synthetic experiment 1. We first draw five random matrices: three matrices $\mathbf{G}, \mathbf{F}, \mathbf{Z} \in \mathbb{R}^{m \times n}$ with independent entries from the normal distribution, and $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{n \times n}$ with independent entries from the uniform distribution on $[0, 1]$. We now set $\mathbf{A} = \mathbf{GX} + 0.1 \cdot \mathbf{F}$ and $\mathbf{B} = \mathbf{GY} + 0.1 \cdot \mathbf{Z}$. We use the sizes $m = 120{,}000$ and $n = 60$. Conceptually, we first take a random basis (the columns of $\mathbf{G}$), and linearly transform it in two different ways (by multiplying by $\mathbf{X}$ and $\mathbf{Y}$). The transformation does not change the space spanned by the bases. We now add to each base some random noise ($0.1 \cdot \mathbf{F}$ and $0.1 \cdot \mathbf{Z}$). Since both $\mathbf{A}$ and $\mathbf{B}$ essentially span the same column space, only polluted by different noise, we expect $(\mathbf{A}, \mathbf{B})$ to have mostly large canonical correlations (close to 1) but also a few small ones. Indeed, Figure 1(a), which plots the canonical correlations of this pair of matrices, confirms our hypothesis.

Figure 2(a) shows the (signed) error in approximating the canonical correlations using Algorithm 1 (i.e., $\hat{\sigma}_i(\mathbf{A}, \mathbf{B}) - \sigma_i(\mathbf{A}, \mathbf{B})$) with r as in (7.1) ($r = 27{,}231$), in five different runs. The actual error is always an order of magnitude smaller than the input ε; the maximum absolute error is only 0.011. For large canonical correlations the error is much smaller, and the approximated value is very accurate. For smaller correlations, the error starts to get larger, but it is still an order of magnitude smaller than the actual value for the smallest correlation. In Figure 5(a) we plot the maximum error in the correlations as a function of r, when in each experiment we do 30 independent runs. As expected, the error decreases sublinearly as the sample size increases.

Next, we checked whether $\mathbf{AW}$ and $\mathbf{BP}$ are close to having orthogonal columns, where $\mathbf{W}$ and $\mathbf{P}$ contain the canonical weights returned by the proposed algorithm. Figure 3(a) visualizes the entries of $\mathbf{W}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{AW}$ and Figure 4(a) visualizes the entries of $\mathbf{P}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{BP}$ in one of the runs. We see that the diagonal is dominant, and close to 1,

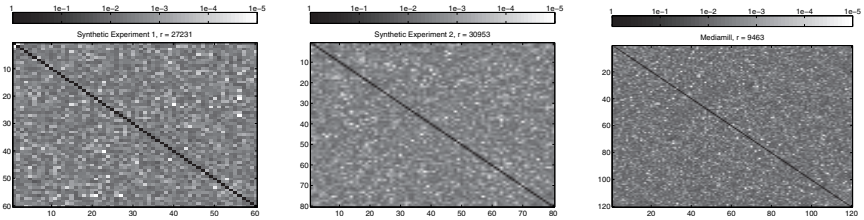

FIG. 3. *Visualization of the absolute value of the enteries in* $\mathbf{W}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{AW}$ *in one of the runs. Color varies between white and black, where black is* 1 *and white is* 10^{-5}.

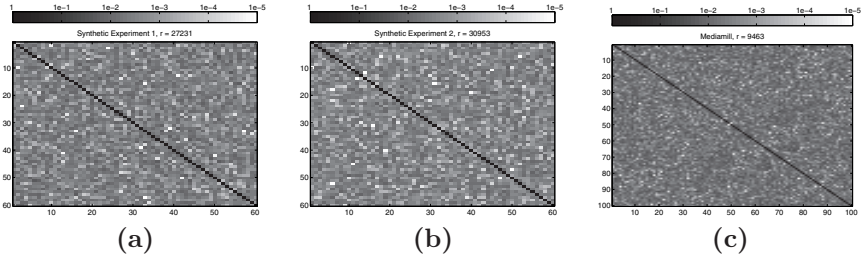

FIG. 4. *Visualization of the absolute value of the enteries in* $\mathbf{P}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{BP}$ *in one of the runs. Color varies between white and black, where black is* 1 *and white is* 10^{-5}.

and the off-diagonal entries are small (but not tiny). The maximum condition number of $\mathbf{AW}$ and $\mathbf{BP}$ we got in the five different runs was 1.18. The maximum value of $\|\mathbf{W}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{AW} - \mathbf{I}_q\|_2$ and $\|\mathbf{P}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{BP} - \mathbf{I}_q\|_2$ we got was 0.096. In Figure 6(a) we plot the maximum value of $\|\mathbf{W}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{AW} - \mathbf{I}_q\|_2$ and $\|\mathbf{P}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{BP} - \mathbf{I}_q\|_2$ as a function of r, when in each experiment we do 30 independent runs. As expected, the deviation from orthogonality decreases sublinearly as the sample size increases.

As for the running time, the proposed algorithm takes about 55% less time than Björck and Golub's algorithm (0.915 seconds versus 2.04 seconds).

7.2. Synthetic experiment 2. We first draw three random matrices. The first matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$ has independent entries from the normal distribution. The second matrix $\mathbf{Y} \in \mathbb{R}^{m\times k}$ has independent entries which take value ± 1 with equal probability. The third matrix $\mathbf{Z} \in \mathbb{R}^{k\times n}$ has independent entries from the uniform distribution on $[0, 1]$. We now set $\mathbf{A} = \mathbf{X} + 0.1 \cdot \mathbf{Y} \cdot (\mathbf{1}_{k\times n} + \mathbf{Z})$ and $\mathbf{B} = \mathbf{Y}$, where $\mathbf{1}_{k\times n}$ is the $k \times n$ all-ones matrix. We use the sizes $m = 80,000$, $n = 80$, and $k = 60$. Here we basically have noise ($\mathbf{B}$) and a matrix polluted with that noise ($\mathbf{A}$). Thus there is some correlation, but really the two subspaces are different; there is one large correlation (almost 1) and all the rest are small (Figure 1(b)).

Figure 2(b) shows the (signed) error in approximating the correlations using Algorithm 1 (i.e., $\hat{\sigma}_i(\mathbf{A}, \mathbf{B}) - \sigma_i(\mathbf{A}, \mathbf{B})$) with r as in (7.1) ($r = 30,953$), in five different runs. The actual error is an order of magnitude smaller than the target ε; the maximum absolute error is only 0.02. Again, for the largest canonical correlation (which is close to 1) the result is very accurate, with tiny errors. For the other correlations it is larger. For tiny correlations the error is about of the same magnitude as the actual value. Interestingly, we observe a bias toward overestimating the correlations. In Figure 5(b) we plot the maximum error in the correlations as a function of r, when

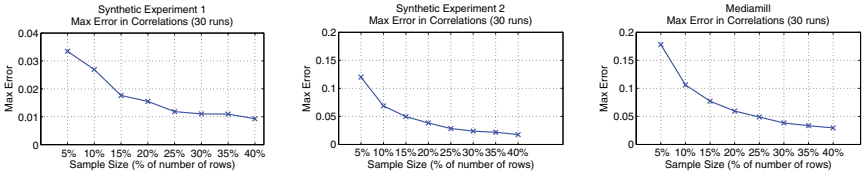

FIG. 5. *Maximum error in the canonical correlation over* 30 *runs as a function of* r.

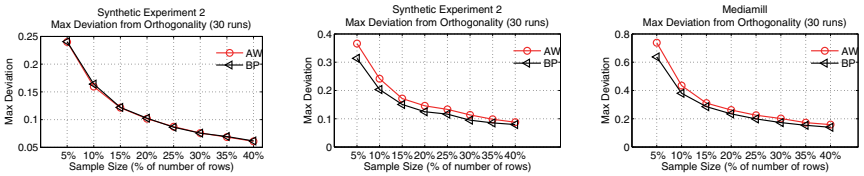

FIG. 6. *Maximum deviation from orthogonality over* 30 *runs as a function of* r.

in each experiment we do 30 independent runs. As expected, the error decreases sublinearly as the sample size increases.

Next, we checked whether $\mathbf{AW}$ and $\mathbf{BP}$ are close to having orthogonal columns, where $\mathbf{W}$ and $\mathbf{P}$ contain the canonical weights returned by the proposed algorithm. Figure 3(b) visualizes the entries of $\mathbf{W}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{AW}$ and Figure 4(b) visualizes the entries of $\mathbf{P}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{BP}$ in one of the runs. We see that the diagonal is dominant, and close to 1, and the off-diagonal entries are small (but not tiny). The maximum condition number of $\mathbf{AW}$ and $\mathbf{BP}$ we got in the five different runs was 1.18. The maximum value of $\|\mathbf{W}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{AW} - \mathbf{I}_q\|_2$ and $\|\mathbf{P}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{BP} - \mathbf{I}_q\|_2$ we got was 0.087. In Figure 6(b) we plot the maximum value of $\|\mathbf{W}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{AW} - \mathbf{I}_q\|_2$ and $\|\mathbf{P}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{BP} - \mathbf{I}_q\|_2$ as a function of r, when in each experiment we do 30 independent runs. As expected, the deviation from orthogonality decreases sublinearly as the sample size increases.

As for the running time, the proposed algorithm takes about 30.5% less time than Björck and Golub's algorithm (0.98 seconds versus 1.41 seconds).

7.3. Real-life dataset: Mediamill. We also tested the proposed algorithm on the annotated video dataset from the Mediamill challenge [28].[2] Combining the training set and the challenge set, 43,907 images are provided, each image a representative keyframe image of a video shot. The dataset provides 120 features for each image, and the set is annotated with 101 labels. The label matrix is rank-deficient with rank 100. Figure 1(c) shows the exact canonical correlations. We see there are a few high correlations with very strong decay afterward.

Figure 2(c) shows the (signed) error in approximating the correlations using Algorithm 1 (i.e., $\hat{\sigma}_i(\mathbf{A},\mathbf{B}) - \sigma_i(\mathbf{A},\mathbf{B})$) with r as in (7.1) ($r = 9,463$), in five different runs. The maximum absolute error is rather small (only 0.055). For the large correlations, which are the more interesting ones in this context, the error is much smaller, so we have a relatively high-accuracy approximation. Again, there is an interesting bias toward overestimating the correlations. In Figure 5(c) we plot the maximum error in the correlations as a function of r, when in each experiment we do

[2]The dataset is publicly available at http://www.csie.ntu.edu.tw/∼cjlin/libsvmtools/datasets/multilabel.html##mediamill.

30 independent runs. As expected, the error decreases sublinearly as the sample size increases.

Next, we checked whether $\mathbf{AW}$ and $\mathbf{BP}$ are close to having orthogonal columns, where $\mathbf{W}$ and $\mathbf{P}$ contain the canonical weights returned by the proposed algorithm. Figure 3(c) visualizes the entries of $\mathbf{W}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{AW}$ and Figure 4(c) visualizes the entries of $\mathbf{P}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{BP}$ in one of the runs. We see that the diagonal is dominant, and close to 1, and the off-diagonal entries are small (but not tiny). The maximum condition number of $\mathbf{AW}$ and $\mathbf{BP}$ we got in the five different runs was 1.51. The maximum value of $\|\mathbf{W}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{AW} - \mathbf{I}_q\|_2$ and $\|\mathbf{P}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{BP} - \mathbf{I}_q\|_2$ we got was 0.24. Both are larger than the previous two examples but still not huge and so indicate some measure of orthogonality. In Figure 6(c) we plot the maximum value of $\|\mathbf{W}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{AW} - \mathbf{I}_q\|_2$ and $\|\mathbf{P}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{BP} - \mathbf{I}_q\|_2$ as a function of r, when in each experiment we do 30 independent runs. As expected, the deviation from orthogonality decreases sublinearly as the sample size increases.

As for the running time, the proposed algorithm is considerably faster than Björck and Golub's algorithm (0.69 seconds versus 2.03 seconds).

7.4. Summary. The experiments are not exhaustive, but they do suggest the following. First, it appears that the theoretical sampling size bounds are rather loose. The algorithm achieves much better approximation errors in practice. Second, there seems to be a connection between the canonical correlation value and the error: for larger correlations the error is smaller. Our bounds fail to capture these phenomena. Finally, the experiments show that the proposed algorithm is faster than Björck and Golub's algorithm *in practice* on both synthetic and real-life datasets, even if they are fairly small. We expect the difference to be much larger on big datasets.

8. Conclusions. We demonstrated that dimensionality reduction via randomized fast unitary transforms leads to faster algorithms for canonical correlation analysis on high-dimensional datasets, beating the seminal SVD-based algorithm of Björck and Golub. The proposed algorithm builds upon a family of similar algorithms which, in recent years, led to similar running time improvements for other classical linear algebraic and machine learning problems: (i) least-squares regression [26, 6, 13, 4], (ii) approximate PCA (via low-rank matrix approximation) [18], (iii) matrix multiplication [27], (v) K-means clustering [7], and (vi) support vector machines [25].

REFERENCES

[1] N. Ailon and B. Chazelle, *Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform*, in Proceedings of the Symposium on Theory of Computing (STOC), 2006, pp. 557–563.

[2] N. Ailon and E. Liberty, *Fast dimension reduction using Rademacher series on dual BCH codes*, in Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA), 2008.

[3] H. Avron, C. Boutsidis, S. Toledo, and A. Zouzias, *Efficient dimensionality reduction for canonical correlation analysis*, in Proceedings of the 30th International Conference on Machine Learning, 2013, pp. 347–355.

[4] H. Avron, P. Maymounkov, and S. Toledo, *Blendenpik: Supercharging LAPACK's least-squares solver*, SIAM J. Sci. Comput., 32 (2010), pp. 1217–1236.

[5] A. Björck and G. H. Golub, *Numerical methods for computing angles between linear subspaces*, Math. Comput., 27 (1973), pp. 579–594.

[6] C. Boutsidis and P. Drineas, *Random projections for the nonnegative least-squares problem*, Linear Algebra Appl., 431 (2009), pp. 760–771.

[7] C. Boutsidis, A. Zouzias, and P. Drineas, *Random projections for k-means clustering*, in Proceedings of Neural Information Processing Systems (NIPS), 2010.

[8] K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan, *Multi-view clustering via canonical correlation analysis*, in Proceedings of International Conference in Machine Learning (ICML), 2009, pp. 129–136.
[9] K. L. Clarkson and D. P. Woodruff, *Low rank approximation and regression in input sparsity time*, in Proceedings of the Symposium on Theory of Computing (STOC), 2013.
[10] P. Dhillon, J. Rodu, D. Foster, and L. Ungar, *Two step CCA: A new spectral method for estimating vector models of words*, in Proceedings of the 29th International Conference on Machine Learning (ICML), 2012.
[11] P. S. Dhillon, D. Foster, and L. Ungar, *Multi-view learning of word embeddings via CCA*, in Proceedings of Neural Information Processing Systems (NIPS), 2011.
[12] P. Drineas, M. Magdon-Ismail, M. W. Mahoney, and D. P. Woodruff, *Fast approximation of matrix coherence and statistical leverage*, in International Conference in Proceedings of Machine Learning (ICML), 2012.
[13] P. Drineas, M. W. Mahoney, S. Muthukrishnan, and T. Sarlós, *Faster least squares approximation*, Numer. Math., 117 (2011), pp. 217–249.
[14] P. Drineas, M. W. Mahoney, and S. Muthukrishnan, *Sampling algorithms for ℓ_2-regression and applications*, in Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA), 2006, pp. 1127–1136.
[15] S. Eisenstat and I. Ipsen, *Relative perturbation techniques for singular value problems*, SIAM J. Numer. Anal., 32 (1995), pp. 1972–1988.
[16] G. H. Golub and H. Zha, *The canonical correlations of matrix pairs and their numerical computation*, IMA Vol. Math. Appl., 69 (1995), pp. 27–27.
[17] G. H. Golub and C. F. Van Loan, *Matrix Computations*, vol. 3, Johns Hopkins University Press, Baltimore, MD, 2012.
[18] N. Halko, P. G. Martinsson, and J. A. Tropp, *Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions*, SIAM Rev., 53 (2011), pp. 217–288.
[19] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.
[20] H. Hotelling, *Relations between two sets of variates*, Biometrika, 28 (1936), pp. 321–377.
[21] I. Ipsen and T. Wentworth, *The Effect of Coherence on Sampling from Matrices with Orthonormal Columns, and Preconditioned Least Squares Problems*, preprint arXiv:1203.4809, 2012.
[22] T.-K. Kim, J. Kittler, and R. Cipolla, *Discriminative learning and recognition of image set classes using canonical correlations*, IEEE Trans. Pattern Anal. Mach. Intell., 29 (2007), pp. 1005–1018.
[23] X. Meng and M. W. Mahoney, *Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression*, in Proceedings of the Symposium on Theory of Computing (STOC), 2013.
[24] J. Nelson and H. L Nguyên, *OSnap: Faster Numerical Linear Algebra Algorithms via Sparser Subspace Embeddings*, preprint, arXiv:1211.1002, 2012.
[25] S. Paul, C. Boutsidis, M. Magdon-Ismail, and P. Drineas, *Random projections for support vector machines*, in Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS), 2013.
[26] V. Rokhlin and M. Tygert, *A fast randomized algorithm for overdetermined linear least-squares regression*, Proc. Nat. Acad. Sci., USA, 105 (2008), pp. 13212–13217.
[27] T. Sarlós, *Improved approximation algorithms for large matrices via random projections*, in Proceedings of the Symposium on Foundations of Computer Science (FOCS), 2006.
[28] C. G. M. Snoek, M. Worring, J. C. van Gemert, J. M. Geusebroek, and A. W. M. Smeulders, *The challenge problem for automated detection of* 101 *semantic concepts in multimedia*, in Proceedings of the ACM International Conference on Multimedia, 2006, pp. 421–430.
[29] Y. Su, Y. Fu, X. Gao, and Q. Tian, *Discriminant learning through multiple principal angles for visual recognition*, IEEE Trans. Image Process., 21 (2012), pp. 1381–1390.
[30] L. Sun, B. Ceran, and J. Ye, *A scalable two-stage approach for a class of dimensionality reduction techniques*, in Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2010, pp. 313–322.
[31] L. Sun, S. Ji, and J. Ye, *A least squares formulation for canonical correlation analysis*, in Proceedings of the International Conference in Machine Learning (ICML), 2008, pp. 1024–1031.

[32] A. Talwalkar and A. Rostamizadeh, *Matrix coherence and the Nyström method*, in Proceedings of the 26th Conference on Uncertainity in Artificial Intelligence, 2010, pp. 572–579.
[33] J. A. Tropp, *Improved analysis of the subsampled randomized Hadamard transform*, Adv. Adapt. Data Anal., special issue, "Sparse Representation of Data and Images" (2011).