Efficient Globally Convergent Stochastic Optimization for Canonical Correlation Analysis

Weiran Wang1∗  Jialei Wang2∗  Dan Garber1  Nathan Srebro1
1Toyota Technological Institute at Chicago  2University of Chicago
{weiranwang,dgarber,nati}@ttic.edu  jialei@uchicago.edu

Abstract

We study the stochastic optimization of canonical correlation analysis (CCA), whose objective is nonconvex and does not decouple over training samples. Although several stochastic gradient based optimization algorithms have been recently proposed to solve this problem, no global convergence guarantee was provided by any of them. Inspired by the alternating least squares/power iterations formulation of CCA, and the shift-and-invert preconditioning method for PCA, we propose two globally convergent meta-algorithms for CCA, both of which transform the original problem into sequences of least squares problems that need only be solved approximately. We instantiate the meta-algorithms with state-of-the-art SGD methods and obtain time complexities that significantly improve upon that of previous work. Experimental results demonstrate their superior performance.

1 Introduction

Canonical correlation analysis (CCA, [1]) and its extensions are ubiquitous techniques in scientific research areas for revealing the common sources of variability in multiple views of the same phenomenon. In CCA, the training set consists of paired observations from two views, denoted \((x_1, y_1), \ldots, (x_N, y_N)\), where \(N\) is the training set size, \(x_i \in \mathbb{R}^{d_x}\) and \(y_i \in \mathbb{R}^{d_y}\) for \(i = 1, \ldots, N\). We also denote the data matrices for each view by \(X = [x_1, \ldots, x_N] \in \mathbb{R}^{d_x \times N}\) and \(Y = [y_1, \ldots, y_N] \in \mathbb{R}^{d_y \times N}\), and \(d := d_x + d_y\). The objective of CCA is to find linear projections of each view such that the correlation between the projections is maximized:

\[
\max_{u,v} \quad u^T \Sigma_{xy} v \\
\text{s.t.} \quad u^T \Sigma_{xx} u = v^T \Sigma_{yy} v = 1
\]

(1)

where \(\Sigma_{xy} = \frac{1}{N} XX^T\) is the cross-covariance matrix, \(\Sigma_{xx} = \frac{1}{N} XX^T + \gamma_x I\) and \(\Sigma_{yy} = \frac{1}{N} YY^T + \gamma_y I\) are the auto-covariance matrices, and \((\gamma_x, \gamma_y) \geq 0\) are regularization parameters [2].

We denote by \((u^*, v^*)\) the global optimum of (1), which can be computed in closed-form. Define

\[
T := \Sigma_{xx}^{-\frac{1}{2}} \Sigma_{xy} \Sigma_{yy}^{-\frac{1}{2}} \in \mathbb{R}^{d_x \times d_y},
\]

(2)

and let \((\phi, \psi)\) be the (unit-length) left and right singular vector pair associated with \(T\)'s largest singular value \(\rho_1\). Then the optimal objective value, i.e., the canonical correlation between the views, is \(\rho_1\), achieved by \((u^*, v^*) = (\Sigma_{xx}^{-\frac{1}{2}} \phi, \Sigma_{yy}^{-\frac{1}{2}} \psi)\). Note that

\[
\rho_1 = \|T\| \leq \left\| \Sigma_{xx}^{-\frac{1}{2}} X \right\| \left\| \Sigma_{yy}^{-\frac{1}{2}} Y \right\| \leq 1.
\]

Furthermore, we are guaranteed to have \(\rho_1 < 1\) if \((\gamma_x, \gamma_y) > 0\).

∗The first two authors contributed equally.
2We assume that \(X\) and \(Y\) are centered at the origin for notational simplicity; if they are not, we can center them as a pre-processing operation.
The main contribution of this paper is the proposal of two globally convergent meta-algorithms for achieving \( \eta \)-suboptimal solution \((u, v)\) to CCA, i.e., \( \min (\|u\Sigma_{xx}u\|_2^2, (v\Sigma_{yy}v)^2) \geq 1 - \eta \). GD=gradient descent, AGD=accelerated GD, SVRG=stochastic variance reduced gradient, ASVRG=accelerated SVRG. Note ASVRG provides speedup over SVRG only when \( \tilde{\kappa} > N \), and we show the dominant term in its complexity.

| Algorithm            | Least squares solver | Time complexity |
|----------------------|----------------------|-----------------|
| AppGrad [3]          | GD                   | \( \tilde{\mathcal{O}} (\eta N \tilde{\kappa}^2 \frac{\rho_1^2}{\rho_2^2} \cdot \log \left( \frac{1}{\eta} \right)) \) (local) |
| CCA Lin [6]          | AGD                  | \( \tilde{\mathcal{O}} (\eta N \sqrt{\kappa} (\frac{\rho_1^2}{\rho_2^2})^2 \cdot \log \left( \frac{1}{\eta} \right)) \) |
| This work: Alternating least squares (ALS) | AGD | \( \tilde{\mathcal{O}} (\eta d N + \tilde{\kappa} (\frac{\rho_1^2}{\rho_2^2})^2 \cdot \log \left( \frac{1}{\eta} \right)) \) |
|                      | SVRG                | \( \tilde{\mathcal{O}} (\eta d N + \tilde{\kappa} (\frac{\rho_1^2}{\rho_2^2})^2 \cdot \log \left( \frac{1}{\eta} \right)) \) |
|                      | ASVRG               | \( \tilde{\mathcal{O}} (\eta d N \tilde{\kappa} \sqrt{\kappa} (\frac{\rho_1^2}{\rho_2^2})^2 \cdot \log \left( \frac{1}{\eta} \right)) \) |
| This work: Shift-and-invert preconditioning (SI) | AGD | \( \tilde{\mathcal{O}} (\eta d N \tilde{\kappa} \sqrt{\kappa} (\frac{\rho_1^2}{\rho_2^2})^2 \cdot \log \left( \frac{1}{\eta} \right)) \) |
|                      | SVRG                | \( \tilde{\mathcal{O}} (\eta d N \tilde{\kappa} \sqrt{\kappa} (\frac{\rho_1^2}{\rho_2^2})^2 \cdot \log \left( \frac{1}{\eta} \right)) \) |
|                      | ASVRG               | \( \tilde{\mathcal{O}} (\eta d N \tilde{\kappa} \sqrt{\kappa} (\frac{\rho_1^2}{\rho_2^2})^2 \cdot \log \left( \frac{1}{\eta} \right)) \) |

For large and high dimensional datasets, it is time and memory consuming to first explicitly form the matrix \( T \) (which requires eigen-decomposition of the covariance matrices) and then compute its singular value decomposition (SVD). For such datasets, it is desirable to develop stochastic algorithms that have efficient updates, converges fast, and takes advantage of the input sparsity. There have been recent attempts to solve (1) based on stochastic gradient descent (SGD) methods [3, 4, 5], but none of these work provides rigorous convergence analysis for their stochastic CCA algorithms.

The main contribution of this paper is the proposal of two globally convergent meta-algorithms for solving (1), namely, alternating least squares (ALS, Algorithm 5.2) and shift-and-invert preconditioning (SI, Algorithm 5), both of which transform the original problem (1) into sequences of least squares problems that need only be solved approximately. We instantiate the meta algorithms with state-of-the-art SGD methods and obtain efficient stochastic optimization algorithms for CCA.

In order to measure the alignments between an approximate solution \((u, v)\) and the optimum \((u^*, v^*)\), we assume that \( T \) has a positive singular value gap \( \Delta := \rho_1 - \rho_2 \in (0, 1] \) so its top left and right singular vector pair is unique (up to a change of sign).

Table 1 summarizes the time complexities of several algorithms for achieving \( \eta \)-suboptimal alignments, where \( \tilde{\kappa} = \max \min (\sigma_{\min}(\Sigma_{xx}), \sigma_{\min}(\Sigma_{yy})) \) is the upper bound of condition numbers of least squares problems solved in all cases. We use the notation \( \tilde{\mathcal{O}}(\cdot) \) to hide poly-logarithmic dependencies (see Sec. 5.1.1 and Sec. 5.2.3 for the hidden factors). Each time complexity may be preferrable in certain regime depending on the parameters of the problem.

**Notations** We use \( \sigma_i(A) \) to denote the \( i \)-th largest singular value of a matrix \( A \), and use \( \sigma_{\max}(A) \) and \( \sigma_{\min}(A) \) to denote the largest and smallest singular values of \( A \) respectively.

### 2 Motivation: Alternating least squares

Our solution to (1) is inspired by the alternating least squares (ALS) formulation of CCA [7] Algorithm 5.2, as shown in Algorithm 6. Let the nonzero singular values of \( T \) be \( 1 \geq \rho_1 \geq \rho_2 \geq \cdots \geq \rho_r > 0 \), where \( r = \text{rank}(T) \leq \min(d_x, d_y) \), and the corresponding (unit-length) left and right singular vector pairs be \((a_1, b_1), \ldots, (a_r, b_r)\), with \( a_1 = \phi \) and \( b_1 = \psi \). Define

\[
C = \begin{bmatrix} 0 & T \newline T^\top & 0 \end{bmatrix} \in \mathbb{R}^{d \times d}. \tag{3}
\]

---

For the ALS meta-algorithm, its enough to consider a per-view conditioning. And when using AGD as the least squares solver, the time complexities depends on \( \sigma_{\max}(\Sigma_{xx}) \) instead, which is less than \( \max_i \|x_i\|^2 \).
Theorem
with corresponding eigenvectors
Then we can equivalently rewrite the steps of Algorithm 1 in the new variables as in
Algorithm 1
Then for
of Algorithm 1 is of the form
It is straightforward to check that the nonzero eigenvalues of \( C \) are:
with corresponding eigenvectors
The key observation is that Algorithm 1 effectively runs a variant of power iterations on \( C \) to extract its top eigenvector. To see this, make the following change of variables
Then we can equivalently rewrite the steps of Algorithm 1 in the new variables as in \{ \} of each line.
Observe that the iterates are updated as follows from step \( t - 1 \) to step \( t \):
Except for the special normalization steps which rescale the two sets of variables separately, Algorithm 1 is very similar to the power iterations \[8\].
We show the convergence rate of ALS below (see its proof in Appendix A). The first measure of progress is the alignment of \( \phi_t \) to \( \phi \) and the alignment of \( \psi_t \) to \( \psi \), i.e., \( (\phi_t, \phi)^2 = (u_t^\top \Sigma_{xx} u^*)^2 \) and \( (\psi_t, \psi)^2 = (v_t^\top \Sigma_{yy} v^*)^2 \). The maximum value for such alignments is 1, achieved when the iterates completely align with the optimal solution. The second natural measure of progress is the objective of \(1\), i.e., \( u_t^\top \Sigma_{xy} v_t \), with the maximum value being \( \rho_1 \).

**Theorem 1** (Convergence of Algorithm 1). Let \( \mu := \min \left( (u_0^\top \Sigma_{xx} u^*)^2, (v_0^\top \Sigma_{yy} v^*)^2 \right) > 0 \). Then for \( t \geq \frac{\mu^2}{\rho_1^2 \rho_2^2 \log \left( \frac{\mu}{\rho_1} \right)} \), we have in Algorithm 1 that \( \min \left( (u_t^\top \Sigma_{xx} u^*)^2, (v_t^\top \Sigma_{yy} v^*)^2 \right) \geq 1 - \eta \), and \( u_t^\top \Sigma_{xy} v_t \geq \rho_1 (1 - 2\eta) \).

**Remarks** We have assumed a nonzero singular value gap in Theorem 1 to obtain linear convergence in both the alignments and the objective. When there exists no singular value gap, the top singular vector pair is not unique and it is no longer meaningful to measure the alignments. Nonetheless, it is possible to extend our proof to obtain sublinear convergence for the objective in this case.

Observe that, besides the steps of normalization to unit length, the basic operation in each iteration of Algorithm 1 is of the form
which is equivalent to solving the following regularized least squares (ridge regression) problem
In the next section, we show that, to maintain the convergence of ALS, it is unnecessary to solve the least squares problems exactly. This enables us to use state-of-the-art SGD methods for solving \(6\) to sufficient accuracy, and to obtain a globally convergent stochastic algorithm for CCA.

\[ \min_{u} \frac{1}{2N} \| u^\top X - v_{t-1}^\top Y \|^2 + \frac{\gamma x}{2} \| u \|^2 \equiv \min_{u} \frac{1}{N} \sum_{i=1}^{N} \left[ \frac{1}{2} \| u^\top x_i - v_{t-1}^\top y_i \|^2 + \frac{\gamma x}{2} \| u \|^2 \right]. \]
Algorithm 2 The alternating least squares (ALS) meta-algorithm for CCA.

**Input:** Data matrices $X \in \mathbb{R}^{d_x \times N}$, $Y \in \mathbb{R}^{d_y \times N}$, regularization parameters $(\gamma_x, \gamma_y)$. Initialize $u_0 \in \mathbb{R}^{d_x}$, $v_0 \in \mathbb{R}^{d_y}$.

\begin{align*}
    u_0 &\leftarrow u_0 / \sqrt{u_0^\top \Sigma_{xx} u_0}, & v_0 &\leftarrow v_0 / \sqrt{v_0^\top \Sigma_{yy} v_0}, & u_0 &\leftarrow u_0, & v_0 &\leftarrow v_0 \\
    \text{for } t = 1, 2, \ldots, T \text{ do} & \quad & \text{end for}
\end{align*}

Solve $\min_u f_t(u) := \frac{1}{2N} \|u^\top X - v_{t-1}^\top Y\|^2 + \frac{\gamma_x}{2} \|u\|^2$ with initialization $u_{t-1}$, and output approximate solution $u_t$ satisfying $f_t(u_t) \leq \min_u f_t(u) + \epsilon$.

Solve $\min_v g_t(v) := \frac{1}{2N} \|v^\top Y - u_{t-1}^\top X\|^2 + \frac{\gamma_y}{2} \|v\|^2$ with initialization $v_{t-1}$, and output approximate solution $v_t$ satisfying $g_t(v_t) \leq \min_v g_t(v) + \epsilon$.

$u_t \leftarrow u_t / \sqrt{u_t^\top \Sigma_{xx} u_t}$, $v_t \leftarrow v_t / \sqrt{v_t^\top \Sigma_{yy} v_t}$

**Output:** $(u_T, v_T)$ is the approximate solution to CCA.

3 Our algorithms

3.1 Algorithm I: Alternating least squares (ALS) with variance reduction

Our first algorithm consists of two nested loops. The outer loop runs inexact power iterations while the inner loop uses advanced stochastic optimization methods, e.g., stochastic variance reduced gradient (SVRG, [2]) to obtain approximate matrix-vector multiplications. A sketch of our algorithm is provided in Algorithm 2. We make the following observations from this algorithm.

**Connection to previous work** At step $t$, if we optimize $f_t(u)$ and $g_t(v)$ crudely by a single batch gradient descent step from the initialization $(\tilde{u}_{t-1}, \tilde{v}_{t-1})$, we obtain the following update rule:

\begin{align*}
    \tilde{u}_t &\leftarrow \tilde{u}_{t-1} - 2\xi X(X^\top \tilde{u}_{t-1} - Y^\top \tilde{v}_{t-1})/N, & u_t &\leftarrow \tilde{u}_t / \sqrt{\tilde{u}_t^\top \Sigma_{xx} \tilde{u}_t} \\
    \tilde{v}_t &\leftarrow \tilde{v}_{t-1} - 2\xi Y(Y^\top \tilde{v}_{t-1} - X^\top \tilde{u}_{t-1})/N, & v_t &\leftarrow \tilde{v}_t / \sqrt{\tilde{v}_t^\top \Sigma_{yy} \tilde{v}_t}
\end{align*}

where $\xi > 0$ is the stepsize (assuming $\gamma_x = \gamma_y = 0$). This coincides with the AppGrad algorithm of [3, Algorithm 3], for which only local convergence is shown. Since the objectives $f_t(u)$ and $g_t(v)$ decouple over training samples, it is convenient to apply SGD methods to them. This observation motivated the stochastic CCA algorithms of [3, 4]. We note however, no global convergence guarantee was shown for these stochastic CCA algorithms, and the key to our convergent algorithm is to solve the least squares problems to sufficient accuracy.

**Warm-start** Observe that for different $t$, the least squares problems $f_t(u)$ only differ in their targets as $v_t$ changes over time. Since $v_{t-1}$ is close to $v_t$ (especially when near convergence), we may use $\tilde{u}_t$ as initialization for minimizing $f_{t+1}(u)$ with an iterative algorithm.

**Normalization** At the end of each outer loop, Algorithm 2 implements exact normalization of the form $u_t \leftarrow u_t / \sqrt{u_t^\top \Sigma_{xx} u_t}$ to ensure the constraints, where $u_t^\top \Sigma_{xx} u_t = \frac{1}{N}(u_t^\top X)(u_t^\top X)^\top + \gamma_x \|u_t\|^2$ requires computing the projection of the training set $u_t^\top X$. However, this does not introduce extra computation because we also compute this projection for the batch gradient used by SVRG (at the beginning of time step $t + 1$). In contrast, the stochastic algorithms of [3, 4] (possibly adaptively) estimate the covariance matrix from a minibatch of training samples and use the estimated covariance for normalization. This is because their algorithms perform normalizations after each update and thus need to avoid computing the projection of the entire training set frequently. But as a result, their inexact normalization steps introduce noise to the algorithms.

**Input sparsity** For high dimensional sparse data (such as those used in natural language processing [10]), an advantage of gradient based methods over the closed-form solution is that the former takes into account the input sparsity. For sparse inputs, the time complexity of our algorithm depends on $\text{nnz}(X, Y)$, i.e., the total number of nonzeros in the inputs instead of $dN$.

**Canonical ridge** When $(\gamma_x, \gamma_y) > 0$, $f_t(u)$ and $g_t(v)$ are guaranteed to be strongly convex due to the $\ell_2$ regularizations, in which case SVRG converges linearly. It is therefore beneficial to use
small nonzero regularization for improved computational efficiency, especially for high dimensional datasets where inputs $X$ and $Y$ are approximately low-rank.

Convergence By the analysis of inexact power iterations where the least squares problems are solved (or the matrix-vector multiplications are computed) only up to necessary accuracy, we provide the following theorem for the convergence of Algorithm 2 (see its proof in Appendix B). The key to our analysis is to bound the distances between the iterates of Algorithm 2 and that of the two algorithms have the same quality.

**Theorem 2** (Convergence of Algorithm 2). Fix $T \geq \lceil \frac{\tilde{\eta}^2}{\rho_1^2 - \rho_2^2} \cdot \frac{\log \left( \frac{2M}{\rho_1^2 - \rho_2^2} \right)}{\tilde{\eta}} \rceil$, and set $\epsilon(T) \leq \frac{n^2 \rho^2}{128} \cdot \left( \frac{2M}{\rho_1^2 - \rho_2^2} \cdot \frac{\log \left( \frac{2M}{\rho_1^2 - \rho_2^2} \right)}{\tilde{\eta}} \right)^2$ in Algorithm 2. Then we have $u^T_T \Sigma_{xx} u_T = v^T_T \Sigma_{yy} v_T = 1$, $\min \left( (u^T_T \Sigma_{xx} u^*)^2, (v^T_T \Sigma_{yy} v^*)^2 \right) \geq 1 - \eta$, and $u^T_T \Sigma_{xy} v_T \geq \rho_1 (1 - 2\eta)$.

3.1.1 Stochastic optimization of regularized least squares We now discuss the inner loop of Algorithm 2 which approximately solves problems of the form (6). Appendix D that the initial suboptimality for minimizing $\tilde{f}(u)$ is

$$\min \left( (u^T_T \Sigma_{xx} u^*)^2, (v^T_T \Sigma_{yy} v^*)^2 \right) \geq 1 - \eta,$$

and we only show the dominant term in the above complexity.

Remarks As mentioned in [14], the acceleration version provides speedup over normal SVRG only when $\kappa_x > N$ and we only show the dominant term in the above complexity.

By combining the iteration complexity of the outer loop (Theorem 2) and the time complexity of the inner loop (Lemma 3), we obtain the total time complexity of $\tilde{O} \left( d (N + \kappa) \left( \frac{\rho^2}{\rho_1^2 - \rho_2^2} \right)^2 \cdot \log \left( \frac{1}{\eta} \right) \right)$ for ALS+SVRG and $\tilde{O} \left( d \sqrt{N} \kappa' \left( \frac{\rho^2}{\rho_1^2 - \rho_2^2} \right)^2 \cdot \log \left( \frac{1}{\eta} \right) \right)$ for ALS+ASVRG, where $\kappa := \max \left( \frac{\max \|x\|^2}{\sigma_{\min}(\Sigma_{xx})}, \frac{\max \|y\|^2}{\sigma_{\min}(\Sigma_{yy})} \right)$ and $\tilde{O}(\cdot)$ hides poly-logarithmic dependences on $\frac{1}{\rho_1}$ and $\frac{1}{\rho_2}$. Our algorithm does not require the initialization to be close to the optimum and converges globally. For comparison, the locally convergent AppGrad has a time complexity \cite{appgrad} of $\tilde{O} \left( d N \kappa' \rho^2 \left( \frac{\rho^2}{\rho_1^2 - \rho_2^2} \cdot \log \left( \frac{1}{\eta} \right) \right) \right)$, where $\kappa' := \max \left( \frac{\sigma_{\max}(\Sigma_{xx})}{\sigma_{\min}(\Sigma_{xx})}, \frac{\sigma_{\max}(\Sigma_{yy})}{\sigma_{\min}(\Sigma_{yy})} \right)$. Note, in this complexity, the dataset size $N$ and the least squares condition number $\kappa'$ are multiplied together because AppGrad essentially uses batch gradient descent as the least squares solver. Within our framework, we can use accelerated gradient descent (AGD, \cite{AGD}) instead and obtain a globally convergent algorithm with a total time complexity of $\tilde{O} \left( d N \sqrt{\kappa' \left( \frac{\rho^2}{\rho_1^2 - \rho_2^2} \right)^2 \cdot \log \left( \frac{1}{\eta} \right)} \right)$.

3.2 Algorithm II: Shift-and-invert preconditioning (SI) with variance reduction The second algorithm is inspired by the shift-and-invert preconditioning method for PCA \cite{PCA}. Instead of running power iterations on $C$ as defined in \cite{PCA}, we will be running power iterations on

$$M_{\lambda} = (\lambda I - C)^{-1} = \begin{bmatrix} \lambda I & -T^\top \\ -T & \lambda I \end{bmatrix}^{-1} \in \mathbb{R}^{d \times d},$$

\textsuperscript{5}The expectation is taken over random sampling of component functions. High probability error bounds can be obtained using the Markov’s inequality.
Theorem
We have the following convergence guarantee for Phase I (see its proof in Appendix F) and it proceeds in two phases.

Algorithm 3. Then the constraints, and this is taken care of in Phase II.

3.2.1 Phase I: shift-and-invert preconditioning for eigenvectors of $\mathbf{M}_\lambda$

Using an estimate of the singular value gap $\Delta$ and starting from an over-estimate of $\rho_1$ ($1 + \tilde{\Delta}$ suffices), the algorithm gradually shrinks $\lambda_{(s)}$ towards $\rho_1$ by crudely estimating the leading eigenvector/eigenvalues of each $\mathbf{M}_{\lambda_{(s)}}$ along the way and shrinking the gap $\lambda_{(s)} - \rho_1$, until we reach an $\lambda_{(f)} \in (\rho_1, \rho_1 + c(\rho_1 - \rho_2))$ where $c \sim O(1)$. Afterwards, the algorithm fixes $\lambda_{(f)}$ and runs inexact power iterations on $\mathbf{M}_{\lambda_{(f)}}$ to obtain an accurate estimate of its leading eigenvector. Note in this phase, power iterations implicitly operate on the concatenated variables $\frac{1}{\sqrt{2}} \left[ \frac{\mathbf{u}_t}{\mathbf{v}_t} \right]$ in $\mathbb{R}^d$ (but without ever computing $\Sigma_{xx}^{+}$ and $\Sigma_{yy}^{+}$).

Matrix-vector multiplication The matrix-vector multiplications in Phase I have the form

$$
\begin{bmatrix}
\mathbf{u}_t \\
\mathbf{v}_t
\end{bmatrix} \leftarrow \frac{1}{\sqrt{2}} \left[ \mathbf{u}^\top \mathbf{v}^\top \right] \begin{bmatrix}
\lambda \mathbf{S}_{xx} - \mathbf{S}_{xy} \\
- \mathbf{S}_{xy}^\top \lambda \mathbf{S}_{yy}
\end{bmatrix}^{-1} \begin{bmatrix}
\mathbf{S}_{xx} \\
\mathbf{S}_{yy}
\end{bmatrix} \begin{bmatrix}
\mathbf{u} \\
\mathbf{v}
\end{bmatrix} - \mathbf{u}^\top \mathbf{S}_{xx} \mathbf{u}_{t-1} - \mathbf{v}^\top \mathbf{S}_{yy} \mathbf{v}_{t-1}.
$$

And as in ALS, this least squares problem can be further written as finite-sum:

$$
\min_{\mathbf{u}, \mathbf{v}} \ h_t(\mathbf{u}, \mathbf{v}) = \frac{1}{N} \sum_{i=1}^{N} h_t^i(\mathbf{u}, \mathbf{v}) \quad \text{where}
$$

$$
h_t^i(\mathbf{u}, \mathbf{v}) = \frac{1}{2} \left[ \mathbf{u}^\top \mathbf{v}^\top \right] \begin{bmatrix}
\lambda (\mathbf{x}_i \mathbf{x}_i^\top + \gamma_i \mathbf{I}) \\
- \mathbf{y}_i \mathbf{y}_i^\top \\
\end{bmatrix} \begin{bmatrix}
\mathbf{u} \\
\mathbf{v}
\end{bmatrix} - \mathbf{u}^\top \mathbf{S}_{xx} \mathbf{u}_{t-1} - \mathbf{v}^\top \mathbf{S}_{yy} \mathbf{v}_{t-1}.
$$

We could directly apply SGD methods to this problem as before.

Normalization The normalization steps in Phase I have the form

$$
\begin{bmatrix}
\mathbf{u}_t \\
\mathbf{v}_t
\end{bmatrix} \leftarrow \frac{1}{\sqrt{2}} \left[ \frac{\mathbf{u}_t}{\mathbf{v}_t} \right] \sqrt{\mathbf{u}_t^\top \mathbf{S}_{xx} \mathbf{u}_t + \mathbf{v}_t^\top \mathbf{S}_{yy} \mathbf{v}_t},
$$

and so the following remains true for the normalized iterates in Phase I:

$$
\mathbf{u}_t^\top \mathbf{S}_{xx} \mathbf{u}_t + \mathbf{v}_t^\top \mathbf{S}_{yy} \mathbf{v}_t = 2, \quad \text{for} \quad t = 1, \ldots, T.
$$

Unlike the normalizations in ALS, the iterates $\mathbf{u}_t$ and $\mathbf{v}_t$ in Phase I do not satisfy the original CCA constraints, and this is taken care of in Phase II.

We have the following convergence guarantee for Phase I (see its proof in Appendix G).

Theorem 4 (Convergence of Algorithm 3 Phase I). Let $\Delta = \rho_1 - \rho_2 \in (0, 1]$, and $\bar{\mu} := \frac{1}{4} \left( \mathbf{u}_0^\top \mathbf{S}_{xx} \mathbf{u}^* + \mathbf{v}_0^\top \mathbf{S}_{yy} \mathbf{v}^* \right)^2 > 0$, and $\bar{\Delta} \in [c_1 \Delta, c_2 \Delta]$ where $0 < c_1 \leq c_2 \leq 1$. Set $m_1 = \lceil 8 \log \left( \frac{16}{\bar{\mu}} \right) \rceil$, $m_2 = \lceil 2 \log \left( \frac{128}{\bar{\mu}^2} \right) \rceil$, and $\bar{\mu} \leq \min \left( \frac{1}{3084} \left( \frac{\bar{\Delta}}{18} \right)^{m_1-1}, \frac{\bar{\mu}^2}{4} \left( \frac{\bar{\Delta}}{18} \right)^{m_2-1} \right)$ in Algorithm 3. Then the $(\mathbf{u}_T, \mathbf{v}_T)$ output by Phase I of Algorithm 3 satisfies

$$
\frac{1}{4} (\mathbf{u}_T^\top \mathbf{S}_{xx} \mathbf{u}^* + \mathbf{v}_T^\top \mathbf{S}_{yy} \mathbf{v}^*)^2 \geq 1 - \frac{\eta_T^2}{64}, \quad \text{and}
$$

and the number of calls to the least squares solver of $h_t(\mathbf{u}, \mathbf{v})$ is $O \left( \log \left( \frac{1}{\bar{\mu}} \right) \log \left( \frac{1}{\bar{\mu}^2} \right) + \log \left( \frac{1}{\bar{\mu}^{m_2}} \right) \right)$.
3.2.2 Phase II: final normalization

In order to satisfy the CCA constraints, we perform a last normalization

\[ \hat{u} \leftarrow u_T / \sqrt{u_T^T \Sigma_{xx} u_T}, \quad \hat{v} \leftarrow v_T / \sqrt{v_T^T \Sigma_{yy} v_T}. \]  

(12)

And we output \((\hat{u}, \hat{v})\) as our final approximate solution to \((1)\). We show that this step does not cause much loss in the alignments, as stated below (see it proof in Appendix G).

**Theorem 5** (Convergence of Algorithm 3, Phase II). Let Phase I of Algorithm 3 outputs \((u_T, v_T)\) that satisfy \((13)\). Then after \((12)\), we obtain an approximate solution \((\hat{u}, \hat{v})\) to \((1)\) such that
\[
\hat{u}^T \Sigma_{xx} \hat{u} = \hat{v}^T \Sigma_{yy} \hat{v} = 1, \min \left( (\hat{u}^T \Sigma_{xx} u)^2, (\hat{v}^T \Sigma_{yy} v)^2 \right) \geq 1 - \eta, \text{and } \hat{u}^T \Sigma_{xy} \hat{v} \geq \rho_1 (1 - 2\eta).
\]

3.2.3 Time complexity

We have shown in Theorem 4 that Phase I only approximately solves a small number of instances of \((9)\). The normalization steps \((10)\) require computing the projections of the traning set which are reused for computing batch gradients of \((9)\). The final normalization \((12)\) is done only once and costs \(O(dN)\). Therefore, the time complexity of our algorithm mainly comes from solving the least squares problems \((10)\) using SGD methods in a blackbox fashion. And the time complexity for SGD methods depends on the condition number of \((9)\). Denote
\[
Q_\lambda = \begin{bmatrix}
\lambda \Sigma_{xx} & - \Sigma_{xy} \\
- \Sigma_{xy} & \Sigma_{yy}
\end{bmatrix} = \begin{bmatrix}
\Sigma_{xx} & \frac{1}{\lambda} T \\
\frac{1}{\lambda} T & - \Sigma_{yy}
\end{bmatrix} \begin{bmatrix}
\Sigma_{xx} & \frac{1}{\lambda} T \\
\frac{1}{\lambda} T & - \Sigma_{yy}
\end{bmatrix}
\]

(13)

It is clear that
\[
\begin{align*}
s_{\max}(Q_\lambda) & \leq (\lambda + \rho_1) \cdot \max (s_{\max}(\Sigma_{xx}), s_{\max}(\Sigma_{yy})) \\
s_{\min}(Q_\lambda) & \geq (\lambda - \rho_1) \cdot \min (s_{\min}(\Sigma_{xx}), s_{\min}(\Sigma_{yy})).
\end{align*}
\]

We have shown in the proof of Theorem 4 that \(\frac{\lambda + \rho_1}{\lambda - \rho_1} \leq \frac{9}{4} \leq \frac{9}{\epsilon_1} \epsilon_2\) throughout Algorithm 3 (cf. Lemma 10 Appendix F.2), and thus the condition number for AGD is \(\frac{s_{\max}(Q_{\lambda})}{s_{\min}(Q_{\lambda})} \leq \frac{9}{\epsilon_1} \epsilon_2 \kappa'\), where \(\kappa' := \max (s_{\max}(\Sigma_{xx}), s_{\max}(\Sigma_{yy}))\). For SVRG/ASVRG, the relevant condition number depends on the gradient Lipschitz constant of individual components. We show in Appendix H (Lemma 12) that the relevant condition number is at most \(\frac{9}{\epsilon_1} \epsilon_2 \frac{1}{\rho_1 - \rho_2} \tilde{\kappa}\), where \(\tilde{\kappa} := \max (\max (\|x\|^2, \|y\|^2), \min (s_{\min}(\Sigma_{xx}), s_{\min}(\Sigma_{yy})))\). An interesting issue for SVRG/ASVRG is that, depending on the value of \(\lambda\), the independent components \(h_1^i(u, v)\) may be nonconvex. If \(\lambda \geq 1\), each component is still guaranteed to be convex; otherwise, some components might be non-convex, with the overall average \(\frac{1}{N} \sum_{i=1}^{N} h_1^i\) \(u, v\) being convex. In the later case, we use the modified analysis of SVRG [10 Appendix B] for its time complexity. We use warm-start in SI as in ALS, and the initial suboptimality for each subproblem can be bounded similarly.

The total time complexities of our SI meta-algorithm are given in Table 1 Note that \(\tilde{\kappa}\) (or \(\tilde{\kappa}'\)) and \(\frac{1}{\rho_1 - \rho_2}\) are multiplied together, giving the effective condition number. When using SVRG as the least squares solver, we obtain the total time complexity of \(\tilde{O} \left( \frac{d N + \tilde{\kappa}}{\rho_1 - \rho_2} \cdot \log^2 \left( \frac{1}{\eta} \right) \right) \) if all components are convex, and \(\tilde{O} \left( d N \sqrt{\tilde{\kappa}} \sqrt{\frac{1}{\rho_1 - \rho_2}} \cdot \log^2 \left( \frac{1}{\eta} \right) \right) \) otherwise. When using ASVRG, we have \(\tilde{O} \left( d N \sqrt{\tilde{\kappa}'} \sqrt{\frac{1}{\rho_1 - \rho_2}} \cdot \log^2 \left( \frac{1}{\eta} \right) \right) \) if all components are convex, and \(\tilde{O} \left( d N \sqrt{\tilde{\kappa}'} \sqrt{\frac{1}{\rho_1 - \rho_2}} \cdot \log^2 \left( \frac{1}{\eta} \right) \right) \) otherwise. Here \(\tilde{O}(\cdot)\) hides poly-logarithmic dependences on \(\frac{1}{\rho_1 - \rho_2}\) and \(\frac{1}{\eta}\). It is remarkable that the SI meta-algorithm is able to separate the dependence of dataset size \(N\) from other parameters in the time complexities.

**Parallel work** In a parallel work [6], the authors independently proposed a similar ALS algorithm and they solve the least squares problems using AGD. The time complexity of their algorithm for extracting the first canonical correlation is \(\tilde{O} \left( d N \sqrt{\kappa'} \sqrt{\frac{1}{\rho_1 - \rho_2}} \cdot \log \left( \frac{1}{\eta} \right) \right)\), which has linear dependence on \(\frac{\rho_2}{\rho_1 - \rho_2} \log \left( \frac{1}{\eta} \right)\) (so their algorithm is linearly convergent, but our complexity for ALS+AGD has quadratic dependence on this factor), but typically worse dependence on \(N\) and \(\kappa'\) (see remarks in Section 3.1.1). Moreover, our SI algorithm tends to significantly outperform ALS theoretically and empirically. It is future work to remove extra \(\log \left( \frac{1}{\eta} \right)\) dependence in our analysis.

*Our arxiv preprint for the ALS meta-algorithm was posted before their paper got accepted by ICML 2016.*
We make the following observations from the results. First, the proposed stochastic algorithms significantly outperform batch gradient based methods AppGrad/CCALin. This is because the least squares condition numbers for these datasets are large, and SVRG enables us to decouple dependences on the dataset size $N$ and the condition number $\kappa$ in the time complexity. Second, SI-VR converges faster than ALS-VR as it further decouples the dependence on $N$ and the singular value gap of $\mathbf{T}$. Third, inexact normalizations keep the $\text{s-AppGrad}$ algorithm from converging to an accurate solution. Finally, ASVRG improves over SVRG when the the condition number is large.

Figure 1: Comparison of suboptimality vs. # passes for different algorithms. For each dataset and regularization parameters $(\gamma_x, \gamma_y)$, we give $\kappa' = \max \left( \frac{\sigma_{\max}(\Sigma_{xx})}{\sigma_{\min}(\Sigma_{xx})}, \frac{\sigma_{\max}(\Sigma_{yy})}{\sigma_{\min}(\Sigma_{yy})} \right)$ and $\delta = \frac{\rho^2}{\rho_{L-1}^2}$.

Extension to multi-dimensional projections To extend our algorithms to $L$-dimensional projections, we can extract the dimensions sequentially and remove the explained correlation from $\Sigma_{xy}$ each time we extract a new dimension [13]. For the ALS meta-algorithm, a cleaner approach is to extract the $L$ dimensions simultaneously using (inexact) orthogonal iterations [8], in which case the subproblems become multi-dimensional regressions and our normalization steps are of the form $\mathbf{U}_i \leftarrow \mathbf{U}_i (\mathbf{U}_i^T \Sigma_{xx} \mathbf{U}_i)^{-1/2}$ (the same normalization is used by [3, 4]). Such normalization involves the eigenvalue decomposition of a $L \times L$ matrix and can be solved exactly as we typically look for low dimensional projections. Our analysis for $L = 1$ can be extended to this scenario and the convergence rate of ALS will depend on the gap between $\rho_L$ and $\rho_{L-1}$.

4 Experiments

We demonstrate the proposed algorithms, namely ALS-VR, ALS-AVR, SI-VR, and SI-AVR, abbreviated as “meta-algorithm – least squares solver” (VR for SVRG, and AVR for ASVRG) on three real-world datasets: Mediamill [19] ($N = 3 \times 10^5$), JW11 [20] ($N = 3 \times 10^4$), and MNIST [21] ($N = 6 \times 10^4$). We compare our algorithms with batch AppGrad and its stochastic version s-AppGrad [8], as well as the CCALin algorithm in parallel work [6]. For each algorithm, we compare the canonical correlation estimated by the iterates at different number of passes over the data with that of the exact solution by SVD. For each dataset, we vary the regularization parameters $\gamma_x = \gamma_y$ over $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$ to vary the least squares condition numbers, and larger regularization leads to better conditioning. We plot the suboptimality in objective vs. # passes for each algorithm in Figure 1. Experimental details (e.g. SVRG parameters) are given in Appendix I.
References

[1] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

[2] H. D. Vinod. Canonical ridge and econometrics of joint production. *J. Econometrics*, 1976.

[3] Z. Ma, Y. Lu, and D. Foster. Finding linear structure in large datasets with scalable canonical correlation analysis. In *ICML*, 2015.

[4] W. Wang, R. Arora, N. Srebro, and K. Livescu. Stochastic optimization for deep CCA via nonlinear orthogonal iterations. In *ALLERTON*, 2015.

[5] B. Xie, Y. Liang, and L. Song. Scale up nonlinear component analysis with doubly stochastic gradients. In *NIPS*, 2015.

[6] R. Ge, C. Jin, S. Kakade, P. Netrapalli, and A. Sidford. Efficient algorithms for large-scale generalized eigenvector computation and canonical correlation analysis. arXiv, April 13 2016.

[7] G. Golub and H. Zha. *Linear Algebra for Signal Processing*, chapter The Canonical Correlations of Matrix Pairs and their Numerical Computation, pages 27–49. 1995.

[8] G. Golub and C. van Loan. *Matrix Computations*. third edition, 1996.

[9] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *NIPS*, 2013.

[10] Y. Lu and D. Foster. Large scale canonical correlation analysis with iterative least squares. In *NIPS*, 2014.

[11] M. Schmidt, N. Le Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. Technical Report HAL 00860051, École Normale Supérieure, 2013.

[12] S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 2013.

[13] R. Frostig, R. Ge, S. Kakade, and A. Sidford. Un-regularizing: Approximate proximal point and faster stochastic algorithms for empirical risk minimization. In *ICML*, 2015.

[14] H. Lin, J. Mairal, and Z. Harchaoui. A universal catalyst for first-order optimization. In *NIPS*, 2015.

[15] Y. Nesterov. *Introductory Lectures on Convex Optimization. A Basic Course*. Springer, 2004.

[16] D. Garber and E. Hazan. Fast and simple PCA via convex optimization. arXiv, 2015.

[17] C. Jin, S. Kakade, C. Musco, P. Netrapalli, and A. Sidford. Robust shift-and-invert preconditioning: Faster and more sample efficient algorithms for eigenvector computation. 2015.

[18] D. Witten, R. Tibshirani, and T. Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 2009.

[19] C. Snoek, M. Worrall, J. van Gemert, J. Geusebroek, and A. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *MULTIMEDIA*, 2006.

[20] J. Westbury. *X-Ray Microbeam Speech Production Database User’s Handbook*, 1994.

[21] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324, 1998.

[22] M. Warmuth and D. Kuzmin. Randomized online PCA algorithms with regret bounds that are logarithmic in the dimension. *Journal of Machine Learning Research*, 2008.

[23] R. Arora, A. Cotter, K. Livescu, and N. Srebro. Stochastic optimization for PCA and PLS. In *ALLERTON*, 2012.

[24] A. Balsubramani, S. Dasgupta, and Y. Freund. The fast convergence of incremental PCA. In *NIPS*, 2013.

[25] O. Shamir. A stochastic PCA and SVD algorithm with an exponential convergence rate. In *ICML*, 2015.

[26] F. Yger, M. Berar, G. Gasso, and A. Rakotomamonjy. Adaptive canonical correlation analysis based on matrix manifolds. In *ICML*, 2012.
A Proof of Theorem I

Proof. It is easy to see that by the end of the first iteration of Algorithm I, \( \tilde{\psi}_1 \) and \( \psi_1 \) lie in the span of \( \{b_i\}_{i=1} \), while \( \tilde{\phi}_1 \) and \( \phi_1 \) lie in the span of \( \{a_i\}_{i=1} \). And therefore they remain in these spaces for all \( t \geq 1 \).

Let us first focus on \( \phi_t \). For \( t \geq 2 \), we observe that

\[
\phi_t = T\psi_{t-1} / \|\phi_t\| = TT^T\phi_{t-2} / (\|\phi_t\| \cdot \|\psi_{t-1}\|).
\]

Since \( \|\phi_{t-2}\| = \|\phi_t\| = 1 \), it is equivalent to using the following updates:

\[
\phi_t \leftarrow TT^T\phi_{t-2}, \quad \phi_t \leftarrow \phi_t / \|\phi_t\|.
\]

This indicates that, Algorithm I runs the standard power iterations on \( TT^T \) to generate the \( \{\phi_t\}_{t \geq 1} \) sequence for every two steps.

(i) For \( t = 2, 4, \ldots \), we have \( \phi_t = \frac{(TT^T)\hat{\psi}_0}{\|TT^T\hat{\psi}_0\|} \). Let \( M = TT^T \), whose nonzero eigenvalues are \( \rho_1^2 \geq \rho_2^2 \geq \cdots \geq \rho_r^2 > 0 \), with corresponding eigenvectors \( a_1, \ldots, a_r \). Then, for \( i = 1, \ldots, r \),

\[
(a_i^T \phi_t)^2 = \frac{(a_i^T M^2 \phi_0)^2}{\|M^2 \phi_0\|^2} = \frac{(a_i^T M^2 \phi_0)^2}{\phi_0^TM^t \phi_0} = \frac{(\rho_i^2 a_i^T \phi_0)^2}{\sum_{j=1}^r \rho_j^2 (a_i^T \phi_0)^2} = \frac{(a_i^T \phi_0)^2}{\sum_{j=1}^r \rho_j^2 (a_i^T \phi_0)^2} \leq \frac{(a_i^T \phi_0)^2}{(a_i^T \phi_0)^2} \exp \left(-\frac{\rho_i^2 - \rho_j^2}{\rho_i^2} t \right).
\]

(ii) For \( t = 1, 3, \ldots \), we have \( \phi_t = \frac{(TT^T)^{t-1} T \psi_0}{\|TT^T\)^{t-1} T \psi_0\|} \). Let \( N = T^T T \), whose nonzero eigenvalues are \( \rho_1^2 \geq \rho_2^2 \geq \cdots \geq \rho_r^2 > 0 \), with corresponding eigenvectors \( b_1, \ldots, b_r \). Then, for \( i = 1, \ldots, r \),

\[
(a_i^T \phi_t)^2 = \frac{(a_i^T (TT^T)^{t-1} T \psi_0)^2}{\|TT^T\)^{t-1} T \psi_0\|^2} = \frac{(a_i^T (TT^T)^{t-1} T \psi_0)^2}{\psi_0^T N^t \psi_0} = \frac{(\rho_i^2 b_i^T \psi_0)^2}{\sum_{j=1}^r \rho_j^2 (b_i^T \psi_0)^2} \leq \frac{(b_i^T \psi_0)^2}{(b_i^T \psi_0)^2} \exp \left(-\frac{\rho_i^2 - \rho_j^2}{\rho_i^2} t \right).
\]

Given \( \delta \in (0, 1) \), define \( S(\delta) = \{ i : \rho_i^2 > (1 - \delta) \rho_1^2 \} \). For \( \delta_1, \delta_2 \in (0, 1) \), define

\[
T(\delta_1, \delta_2) := \left\lfloor \frac{1}{\delta_1} \log \left( \frac{1}{\mu \delta_2} \right) \right\rfloor.
\]

For all \( i \notin S(\delta_1) \), when \( t > T(\delta_1, \delta_2) \), it holds that \( (a_i^T \phi_t)^2 \leq \delta_2 (a_i^T \phi_0)^2 \) if \( t \) is even, and \( (a_i^T \phi_t)^2 \leq \delta_2 (b_i^T \psi_0)^2 \) if \( t \) is odd. In both cases, we have \( \sum_{i \in S(\delta_1)} (a_i^T \phi_t)^2 \geq 1 - \delta_2 \).

When there exists a positive singular value gap, i.e., \( \rho_1 > \rho_2 > 0 \), set \( \delta_1 = (\rho_1^2 - \rho_2^2) / \rho_1^2 \) and thus \( S(\delta_1) = 1 \). Furthermore, set \( \delta_2 = \eta \) and we obtain \( (a_i^T \phi_t)^2 \geq 1 - \eta \).
The proof for $\psi_t$ is completely analogous. To obtain the bound on the objective, we have

$$u_t^\top \Sigma_{xx} v_t = \phi_t^\top T \psi_t = \rho_1 (\phi_t^\top a_1) (\psi_t^\top b_1) + \sum_{i=2}^{r} \rho_i (\phi_t^\top a_i) (\psi_t^\top b_i) \geq \rho_1 (\phi_t^\top a_1) (\psi_t^\top b_1) - \rho_1 \sum_{i=2}^{r} \phi_t^\top a_i \psi_t^\top b_i \geq \rho_1 (1 - \eta) - \rho_1 \sqrt{\sum_{i=2}^{r} (\phi_t^\top a_i)^2} \sqrt{\sum_{i=2}^{r} (\psi_t^\top b_i)^2} \geq \rho_1 (1 - \eta) - \rho_1 \eta = \rho_1 (1 - 2\eta),$$

where we have used the Cauchy-Schwartz inequality in the second inequality. □

**B Proof of Theorem 2**

From now on, we distinguish the iterates of our stochastic algorithm (Algorithm 2) from the iterates of the exact power iterations (Algorithm 1) and denote the latter with asterisks, i.e., $\tilde{u}_t$ and $\tilde{v}_t$ for the unnormalized iterates and $u_t^*$ and $v_t^*$ for the normalized iterates. We denote the exact optimum of $f_t(u)$ and $g_t(v)$ by $\bar{u}_t$ and $\bar{v}_t$, respectively.

The following lemma bounds the distance between the iterates of inexact and exact power iterations.

**Lemma 6.** Assume that Algorithm 1 and Algorithm 2 start with the same initialization, i.e., $\tilde{u}_0 = \bar{u}_0$ and $\tilde{v}_0 = \bar{v}_0$. Then, for $t \geq 1$, the unnormalized iterates of Algorithm 2 satisfy

$$\max \left( \left\| \Sigma_{xx}^{t} \tilde{u}_t - \Sigma_{xx}^{t} \bar{u}_t \right\|, \left\| \Sigma_{yy}^{t} \tilde{v}_t - \Sigma_{yy}^{t} \bar{v}_t \right\| \right) \leq \tilde{S}_t,$$

where

$$\tilde{S}_t := \sqrt{2\epsilon} \frac{(2\rho_1/\rho_r)^t - 1}{(2\rho_1/\rho_r) - 1}.$$

Furthermore, for $t \geq 1$, the normalized iterates of Algorithm 2 satisfy

$$\max \left( \left\| \Sigma_{xx}^{t} \tilde{u}_t - \Sigma_{xx}^{t} \bar{u}_t \right\|, \left\| \Sigma_{yy}^{t} \tilde{v}_t - \Sigma_{yy}^{t} \bar{v}_t \right\| \right) \leq \tilde{S}_t \leq \frac{2\tilde{S}_t}{\rho_r}.$$

**Proof.** We focus on the $\{\tilde{u}_t\}_{t \geq 0}$ and $\{u_t\}_{t \geq 0}$ sequences below; the proof for $\{\tilde{v}_t\}_{t \geq 0}$ and $\{v_t\}_{t \geq 0}$ is completely analogous.

We prove the bound for unnormalized iterates by induction. First, the case for $t = 1$ holds trivially. For $t \geq 2$, we can bound the error of the unnormalized iterates using the exact solution to $f_t(u)$:

$$\left\| \Sigma_{xx}^{t-1} \tilde{u}_t - \Sigma_{xx}^{t-1} \bar{u}_t \right\| \leq \left\| \Sigma_{xx}^{t-1} \tilde{u}_t - \Sigma_{xx}^{t-1} \bar{u}_t \right\| + \left\| \Sigma_{xx}^{t-1} \bar{u}_t - \Sigma_{xx}^{t-1} \tilde{u}_t \right\| \leq \epsilon. \quad (14)$$

For the first term of (14), notice $f_t(u)$ is a quadratic function with minimum achieved at $\tilde{u}_t = \Sigma_{xx}^{-1} \Sigma_{xy} v_{t-1}$. For the approximate solution $\tilde{u}_t$, we have

$$f_t(\tilde{u}_t) - f_t(\bar{u}_t) = \frac{1}{2} (\tilde{u}_t - \bar{u}_t)^\top \Sigma_{xx} (\tilde{u}_t - \bar{u}_t) = \frac{1}{2} \left\| \Sigma_{xx}^{t-1} \tilde{u}_t - \Sigma_{xx}^{t-1} \bar{u}_t \right\|^2 \leq \epsilon.$$

It then follows that

$$\left\| \Sigma_{xx}^{t-1} \tilde{u}_t - \Sigma_{xx}^{t-1} \bar{u}_t \right\| \leq \sqrt{2\epsilon}.$$

The second term of (14) is concerned with the error due to inexact target in the least squares problem $f_t(u)$ as $v_{t-1}$ is different from $v_{t-1}^*$. We can bound it as

$$\left\| \Sigma_{xx}^{t-1} \tilde{u}_t - \Sigma_{xx}^{t-1} \bar{u}_t \right\| = \left\| \Sigma_{xx}^{t-1} \Sigma_{xy} v_{t-1} - \Sigma_{xx}^{t-1} \Sigma_{xy} v_{t-1}^* \right\| = \left\| (\Sigma_{xx}^{-1} \Sigma_{xy} \Sigma_{yy}^{-1}) (\Sigma_{yy}^{t-1} v_{t-1} - v_{t-1}^*) \right\| \leq \|T\| \left( \Sigma_{yy}^{t-1} v_{t-1} - \Sigma_{yy}^{t-1} v_{t-1}^* \right) = \rho_1 \left( \Sigma_{yy}^{t-1} v_{t-1} - \Sigma_{yy}^{t-1} v_{t-1}^* \right). \quad (15)$$
In view of the update rule of our algorithm and the triangle inequality, we have
\begin{equation}
\left\| \Sigma_{yy}^T \hat{v}_t - \Sigma_{yy}^T \hat{v}_t^{*} \right\| \\
\leq \left\| \frac{1}{\Sigma_{yy}^T \hat{v}_t} - \frac{1}{\Sigma_{yy}^T \hat{v}_t^{*}} \right\| + \left\| \frac{1}{\Sigma_{yy}^T \hat{v}_t} - \frac{1}{\Sigma_{yy}^T \hat{v}_t^{*}} \right\|
\end{equation}

We now bound \( \left\| \Sigma_{yy}^T \hat{v}_t^{*} \right\| \) from below. Since \( t \geq 2 \), we have
\[ \Sigma_{yy}^T \hat{v}_t = \Sigma_{yy}^T \Sigma_{xx}^{-1} \Sigma_{xy}^T u_{t-2} = \left( \Sigma_{yy}^T \Sigma_{xy}^T \Sigma_{xx}^{-1} \right) \left( \Sigma_{xx}^T u_{t-2} \right) = T^T \left( \Sigma_{xx}^T u_{t-2} \right). \]
Now, \( \Sigma_{xx}^T u_{t-2} \) corresponds to \( \phi_{t-2} \) in Algorithm[1] which has unit length and lies in the span of \( \{a_1, \ldots, a_n\} \), so we have
\[ \left\| \Sigma_{yy}^T \hat{v}_t \right\| = \left\| T^T \phi_{t-2} \right\| \geq \rho_r. \]
Combining (14), (15) and (16) gives
\[ \left\| \Sigma_{xx}^T \hat{u}_t - \Sigma_{xx}^T \hat{u}_t^{*} \right\| \leq \sqrt{2e} + \frac{2\rho_1}{\rho_r} \cdot \tilde{S}_{t-1} = \sqrt{2e} + \frac{2\rho_1}{\rho_r} \cdot \sqrt{2e} \left( \frac{2\rho_1}{\rho_r} \right)^{t-1} - 1 = \tilde{S}_t. \]
The bound for normalized iterates follows from (16).

**Proof of Theorem[2]** We prove the theorem by relating the iterates of inexact power iterations to those of exact power iterations.

Assume the same initialization as in Lemma[6]. First observe that
\begin{align*}
(u_t^T \Sigma_{xx} u_t^*)^2 &= \left( (u_t^T \Sigma_{xx} u_t^* + (u_t - u_t^*)^T \Sigma_{xx} u_t^*)^2 \\
&\geq \left( (u_t^T \Sigma_{xx} u_t^*)^2 + 2 (u_t^T \Sigma_{xx} u_t^*) (u_t - u_t^*)^T \Sigma_{xx} u_t^* \right) \\
&\geq \left( (u_t^T \Sigma_{xx} u_t^*)^2 - 2 \left( \Sigma_{xx}^T (u_t - u_t^*) \right)^T \Sigma_{xx} u_t^* \right) \\
&\geq \left( (u_t^T \Sigma_{xx} u_t^*)^2 - 2 \left\| \Sigma_{xx}^T u_t - \Sigma_{xx}^T u_t^{*} \right\| \right)
\end{align*}

where we have used the fact that \( \left\| \Sigma_{xx}^T u_t \right\| = \left\| \Sigma_{xx}^T u_t^* \right\| = \left\| \Sigma_{xx}^T u_t^* \right\| = 1 \) and the Cauchy-Schwartz inequality in the last two steps.

Applying Theorem[1] with \( T \geq \left\lceil \frac{\sqrt{2e}}{\rho_r} \log \left( \frac{2\rho_1}{\rho_r} \right) \right\rceil \), we have that \( (u_T^T \Sigma_{xx} u_T^*)^2 \geq 1 - \eta/2 \).
On the other hand, in view of Lemma[6], we have for the specified \( \epsilon \) value in Algorithm[2] that \( \left\| \Sigma_{xx}^T u_T - \Sigma_{xx}^T u_T^* \right\| \leq \eta = \eta_T = \eta/4 \). Plugging these two bounds into (17) gives the desired result.

The proof for \( v_T \) is completely analogous.
C  SVRG for minimizing $f(u)$

We provide the pseudo-code of SVRG for solving the least squares problem (6) below.

SVRG for $\min_u f(u) := \frac{1}{N} \sum_{i=1}^{N} \left( \frac{1}{2} \|u^\top x_i - v^\top y_i\|^2 + \frac{2}{\eta} \|u\|^2 \right)$.

**Input:** Stepsize $\xi$.
- Initialize $u_{(0)} \in \mathbb{R}^{d_x}$.
- for $j = 1, 2, \ldots, M$ do
  - $w_0 \leftarrow u_{(j-1)}$
  - Evaluate the batch gradient $\nabla f(w_0) = X(X^\top w_0 - Y^\top v)/N + \gamma_x w_0$
  - for $t = 1, 2, \ldots, m$ do
    - Randomly pick $i_t$ from $\{1, \ldots, N\}$
    - $w_t \leftarrow w_{t-1} - \xi \left((x_i x_i^\top + \gamma_x I)(w_{t-1} - w_0) + \nabla f(w_0)\right)$
  - end for
  - $u_{(j)} \leftarrow w_t$ for randomly chosen $t \in \{1, \ldots, m\}$.
- end for

**Output:** $u_{(M)}$ is the approximate solution.

D  Initial suboptimality of warm-starts in Algorithm 2

At time step $t$, we initialize the least squares problem $f_t(u)$ with the unnormalized iterate $\bar{u}_{t-1}$ from the previous time step. We now bound the suboptimality of this initialization. Observe that the minimum of $f_t(u)$ is achieved by $\bar{u}_t = \Sigma_{xx}^{-1}\Sigma_{xy}v_{t-1}$, and that

$$f_t(\bar{u}_{t-1}) - f_t(\bar{u}_t) = \frac{1}{2} (\bar{u}_{t-1} - \bar{u}_t)^\top \Sigma_{xx} (\bar{u}_{t-1} - \bar{u}_t) = \frac{1}{2} \|\Sigma_{xx}^{\frac{1}{2}} \bar{u}_{t-1} - \Sigma_{xx}^{\frac{1}{2}} \bar{u}_t\|^2.$$  

Applying the triangle inequality, we have for $t = 1$ that

$$\|\Sigma_{xx}^{\frac{1}{2}} \bar{u}_0 - \Sigma_{xx}^{\frac{1}{2}} \bar{u}_1\| \leq \|\Sigma_{xx}^{\frac{1}{2}} \bar{u}_0\| + \|\Sigma_{xx}^{\frac{1}{2}} \bar{u}_1\| \leq 1 + \|\Sigma_{xx}^{\frac{1}{2}} \Sigma_{xx}^{-1}\Sigma_{xy}v_0\| = 1 + \rho_1 \leq 2$$  

where we have used facts that $\|\Sigma_{yy}^{\frac{1}{2}} \bar{u}_0\| = \|\Sigma_{yy}^{\frac{1}{2}} v_0\| = 1$ due to the initial normalization.

And we have for $t \geq 2$ that

$$\|\Sigma_{xx}^{\frac{1}{2}} \bar{u}_{t-1} - \Sigma_{xx}^{\frac{1}{2}} \bar{u}_t\| \leq \|\Sigma_{xx}^{\frac{1}{2}} \bar{u}_{t-1} - \Sigma_{xx}^{\frac{1}{2}} \bar{u}_t\| + \|\Sigma_{xx}^{\frac{1}{2}} \bar{u}_{t-1} - \Sigma_{xx}^{\frac{1}{2}} \bar{u}_t\| \leq \sqrt{2} + \|\Sigma_{xx}^{\frac{1}{2}} \Sigma_{xx}^{-1}\Sigma_{xy}v_{t-2} - \Sigma_{xx}^{\frac{1}{2}} \Sigma_{xx}^{-1}\Sigma_{xy}v_{t-1}\| = \sqrt{2} + \|\Sigma_{yy}^{\frac{1}{2}} v_{t-2} - \Sigma_{yy}^{\frac{1}{2}} v_{t-1}\| \leq \sqrt{2} + \|\Sigma_{yy}^{\frac{1}{2}} v_{t-2} - \Sigma_{yy}^{\frac{1}{2}} v_{t-1}\| \leq \sqrt{2} + \rho_2 \leq \sqrt{2} + 2$$

where we have used the fact that $\|\Sigma_{yy}^{\frac{1}{2}} v_{t-2}\| = \|\Sigma_{yy}^{\frac{1}{2}} v_{t-1}\| = 1$ in the last inequality.

Therefore, for all $t \geq 1$, the ratio between initial suboptimality and required accuracy is

$$\frac{f_t(\bar{u}_{t-1}) - f_t(\bar{u}_t)}{\epsilon} \sim \frac{2}{\epsilon}.$$
E  The shift-and-invert preconditioning (SI) algorithm for CCA

Our shift-and-invert preconditioning (SI) meta-algorithm is detailed in Algorithm 3.

**Algorithm 3** The shift-and-invert preconditioning meta-algorithm for CCA.

**Input:** Data matrices $X$, $Y$, regularization parameters $(\gamma_x, \gamma_y)$, an estimate $\Delta$ for $\Delta = \rho_1 - \rho_2$.

1. Initialize $\tilde{u}_0 \in \mathbb{R}^{d_x}$, $\tilde{v}_0 \in \mathbb{R}^{d_y}$
2. $u_0 \leftarrow \tilde{u}_0 / \sqrt{u_0^\top \Sigma_{xx} u_0}$, $v_0 \leftarrow \tilde{v}_0 / \sqrt{v_0^\top \Sigma_{yy} v_0}$

**Phase I: Shift-and-invert preconditioning for eigenvectors of $M_\lambda$**

$s \leftarrow 0$, $\lambda_{(0)} \leftarrow 1 + \Delta$

repeat

for $t = (s-1)m_1 + 1, \ldots, sm_1$ do

- Optimize the least squares problem

$$\min_{u, v} h_t(u, v) := \frac{1}{2} \left[ u^\top v \right] \left[ \begin{array}{cc} \lambda_{(s-1)} \Sigma_{xx} & -\Sigma_{xy} \\ -\Sigma_{xy}^\top & \lambda_{(s-1)} \Sigma_{yy} \end{array} \right] \left[ \begin{array}{c} u \\ v \end{array} \right] - u^\top \Sigma_{xx} u_{t-1} - v^\top \Sigma_{yy} v_{t-1}$$

and output an approximate solution $(\tilde{u}_t, \tilde{v}_t)$ satisfying $h_t(\tilde{u}_t, \tilde{v}_t) \leq \min_{u, v} h_t(u, v) + \bar{\epsilon}$.

- Normalization: $\left[ \frac{u_t}{v_t} \right] \leftarrow \sqrt{2} \left[ \frac{u_t}{v_t} \right] / \sqrt{u_t^\top \Sigma_{xx} u_t + v_t^\top \Sigma_{yy} v_t}$

end for

- Optimize the least squares problem

$$\min_w l_s(w) := \frac{1}{2} w^\top \left[ \begin{array}{cc} \lambda_{(s-1)} \Sigma_{xx} & -\Sigma_{xy} \\ -\Sigma_{xy}^\top & \lambda_{(s-1)} \Sigma_{yy} \end{array} \right] w - w^\top \left[ \begin{array}{c} \Sigma_{xx} u_{sm_1} \\ \Sigma_{yy} v_{sm_1} \end{array} \right]$$

and output an approximate solution $w_s$ satisfying $l_s(w_s) \leq \min_w l_s(w) + \bar{\epsilon}$.

$$\Delta_s \leftarrow \frac{1}{2} \cdot \left[ \frac{u_t^\top v_{sm_1}}{\Sigma_{yy}} \right] \frac{1}{\sqrt{w_s^2 - 2 \bar{\epsilon}/\Delta}}$$

until $\Delta_{(s)} \leq \Delta$

$
\lambda_{(f)} \leftarrow \lambda_{(s)}$

for $t = sm_1 + 1, sm_1 + 2, \ldots, sm_1 + m_2$ do

- Optimize the least squares problem

$$\min_{u, v} h_t(u, v) := \frac{1}{2} \left[ u^\top v \right] \left[ \begin{array}{cc} \lambda_{(f)} \Sigma_{xx} & -\Sigma_{xy} \\ -\Sigma_{xy}^\top & \lambda_{(f)} \Sigma_{yy} \end{array} \right] \left[ \begin{array}{c} u \\ v \end{array} \right] - u^\top \Sigma_{xx} u_{t-1} - v^\top \Sigma_{yy} v_{t-1}$$

and output an approximate solution $(\tilde{u}_t, \tilde{v}_t)$ satisfying $h_t(\tilde{u}_t, \tilde{v}_t) \leq \min_{u, v} h_t(u, v) + \bar{\epsilon}$.

- Normalization: $\left[ \frac{u_t}{v_t} \right] \leftarrow \sqrt{2} \left[ \frac{u_t}{v_t} \right] / \sqrt{u_t^\top \Sigma_{xx} u_t + v_t^\top \Sigma_{yy} v_t}$

end for

**Phase II: Final normalization**

$T \leftarrow sm_1 + m_2$, $\tilde{u} \leftarrow u_T / \sqrt{u_T^\top \Sigma_{xx} u_T}$, $\tilde{v} \leftarrow v_T / \sqrt{v_T^\top \Sigma_{yy} v_T}$

**Output:** $(\tilde{u}, \tilde{v})$ is the approximate solution to CCA.

F  Proof of Theorem 4

The proof of Theorem 4 closely follows that of [16, Theorem 4.2]. And we will need a few lemmas on the convergence of inexact power iterations.
F.1 Auxiliary lemmas

Define the condition number of $M_\lambda$ as

$$\kappa_\lambda := \frac{\sigma_1(M_\lambda)}{\sigma_d(M_\lambda)} = \frac{\lambda - \rho_1}{\lambda + \rho_1},$$

and the inverse relative spectral gap of $M_\lambda$ as

$$\delta_\lambda := \frac{\sigma_1(M_\lambda)}{\sigma_1(M_\lambda) - \sigma_2(M_\lambda)} = \frac{\lambda - \rho_2}{\lambda + \rho_2}.$$

The first lemma states the convergence of exact power iterations, paralleling [16, Theorem A.1].

**Lemma 7** (Convergence of exact power iterations). Fix $\alpha > 0$. For the exact power iterations on $M_\lambda$ where

$$\begin{align*}
\left[ \begin{array}{c}
\tilde{u}_t^* \\
\tilde{v}_t^*
\end{array} \right] &\leftarrow \left[ \begin{array}{cc}
\Lambda \Sigma_{xx} & -\Sigma_{xy} \\
-\Sigma_{xy} & \Lambda \Sigma_{yy}
\end{array} \right]^{-1} \left[ \begin{array}{c}
\Sigma_{xx} \\
\Sigma_{yy}
\end{array} \right] \left[ \begin{array}{c}
u_{t-1}^* \\
v_{t-1}^*
\end{array} \right], \\
\left[ \begin{array}{c}
u_t^* \\
v_t^*
\end{array} \right] &\leftarrow \sqrt{2} \left[ \begin{array}{c}
u_t^* \\
v_t^*
\end{array} \right] \sqrt{\left( \tilde{u}_t^* \right)^\top \Sigma_{xx} \tilde{u}_t^* + (\tilde{v}_t^*)^\top \Sigma_{yy} \tilde{v}_t^*},
\end{align*}$$

and $\mu := \frac{1}{4} \left( \left( (u_0^*)^\top \Sigma_{xx} u^* + (v_0^*)^\top \Sigma_{yy} v^* \right)^2 > 0$, we have

- (crude regime)
  $$\frac{1}{2} \left( (u_0^*)^\top \Sigma_{xx} u^* + (v_0^*)^\top \Sigma_{yy} v^* \right)^2 \geq 1 - \alpha$$

  for $t \geq \left\lceil \frac{\log \left( \frac{2}{\mu} \right)}{\alpha} \right\rceil$.
- (accurate regime)
  $$\frac{1}{4} \left( (u_0^*)^\top \Sigma_{xx} u^* + (v_0^*)^\top \Sigma_{yy} v^* \right)^2 \geq 1 - \alpha$$

  for $t \geq \left\lceil \frac{\log \left( \frac{2}{\mu} \right)}{\alpha} \right\rceil$.

The second lemma bounds the distances between the iterates of inexact and exact power iterations, paralleling [16, Lemma 4.1]. Recall that the $(\tilde{u}_t, \tilde{v}_t)$ in Algorithm 3 satisfies $h_t(\tilde{u}_t, \tilde{v}_t) \leq \min_{u,v} h_t(u, v) + \bar{\epsilon}$. Let $(\tilde{u}_t, \tilde{v}_t)$ be the exact minimum of $h_t$. Then we have

$$h_t(\tilde{u}_t, \tilde{v}_t) - h_t(\tilde{u}_t, \tilde{v}_t) \leq \min_{u,v} h_t(u, v) + \bar{\epsilon}.$$

**Lemma 8** (Power iterations with inexact matrix-vector multiplications). Consider the inexact power iterations on $M_\lambda$ where

$$\begin{align*}
\left[ \begin{array}{c}
u_t \\
v_t
\end{array} \right] &\leftarrow \sqrt{2} \left[ \begin{array}{c}
u_t \\
v_t
\end{array} \right] \sqrt{\left( \tilde{u}_t \right)^\top \Sigma_{xx} \tilde{u}_t + \tilde{v}_t^\top \Sigma_{yy} \tilde{v}_t},
\end{align*}$$

for $t = 1, \ldots, m$. 

15
Compare these iterates with those of the exact power iterations described in Lemma 7 using the same initialization $\mathbf{u}_0 = \mathbf{u}_0^0$, $\mathbf{v}_0 = \mathbf{v}_0^0$. Then, for $t \geq 0$, the unnormalized iterates satisfy

$$
\left\| \frac{1}{\sqrt{2}} \begin{bmatrix} \Sigma_{xx}^{\frac{1}{2}} \mathbf{u}_t \\ \Sigma_{yy}^{\frac{1}{2}} \mathbf{v}_t \end{bmatrix} - \frac{1}{\sqrt{2}} \begin{bmatrix} \Sigma_{xx}^{\frac{1}{2}} \mathbf{u}_t^* \\ \Sigma_{yy}^{\frac{1}{2}} \mathbf{v}_t^* \end{bmatrix} \right\| \leq \tilde{R}_t
$$

where

$$
\tilde{R}_t := \sqrt{\sigma_1(M_\lambda)} \cdot \frac{(2\kappa_\lambda)^t - 1}{2\kappa_\lambda - 1},
$$

while the normalized iterates satisfy

$$
\left\| \frac{1}{\sqrt{2}} \begin{bmatrix} \Sigma_{xx}^{\frac{1}{2}} \mathbf{u}_t \\ \Sigma_{yy}^{\frac{1}{2}} \mathbf{v}_t \end{bmatrix} - \frac{1}{\sqrt{2}} \begin{bmatrix} \Sigma_{xx}^{\frac{1}{2}} \mathbf{u}_t^* \\ \Sigma_{yy}^{\frac{1}{2}} \mathbf{v}_t^* \end{bmatrix} \right\| \leq R_t := \frac{2\tilde{R}_t}{\sigma_d(M_\lambda)}.
$$

The third lemma states the convergence of inexact power iterations, paralleling [16, Theorem 4.1].

**Lemma 9** (Convergence of inexact power iterations). Fix $\alpha > 0$. Consider the inexact power iterations described in Lemma 8.

- **(crude regime)** Let $t_1 = \lceil \frac{2}{\alpha} \log \left( \frac{4}{\mu_\alpha} \right) \rceil$. Fix $T \geq t_1$, and set $\bar{\epsilon}(T) = \frac{\alpha^2 \sigma_d(M_\lambda)}{64\kappa_\lambda} (2\kappa_\lambda - 1)^2 t_1$. Then we have

$$
\frac{1}{2} \begin{bmatrix} \mathbf{u}_t^\top \Sigma_{xx}^{\frac{1}{2}} \mathbf{u}_t \\ \mathbf{v}_t^\top \Sigma_{yy}^{\frac{1}{2}} \mathbf{v}_t \end{bmatrix} M_\lambda \begin{bmatrix} \Sigma_{xx}^{\frac{1}{2}} \mathbf{u}_T \\ \Sigma_{yy}^{\frac{1}{2}} \mathbf{v}_T \end{bmatrix} \geq (1 - \alpha) \cdot \sigma_1(M_\lambda).
$$

- **(accurate regime)** Let $t_2 = \lceil \frac{8}{\alpha} \log \left( \frac{2}{\mu_\alpha} \right) \rceil$. Fix $T \geq t_2$, and set $\bar{\epsilon}(T) = \frac{\alpha^2 \sigma_d(M_\lambda)}{64\kappa_\lambda} (2\kappa_\lambda - 1)^2 t_2$. Then we have

$$
\frac{1}{4} \left( \mathbf{u}_T^\top \Sigma_{xx} \mathbf{u}_T + \mathbf{v}_T^\top \Sigma_{yy} \mathbf{v}_T \right)^2 \geq 1 - \alpha.
$$

For brevity, let us define the following short-hands:

$$
\tilde{r}_t = \frac{1}{\sqrt{2}} \begin{bmatrix} \Sigma_{xx}^{\frac{1}{2}} \mathbf{u}_t \\ \Sigma_{yy}^{\frac{1}{2}} \mathbf{v}_t \end{bmatrix}, \quad r_t = \frac{1}{\sqrt{2}} \begin{bmatrix} \Sigma_{xx}^{\frac{1}{2}} \mathbf{u}_t \\ \Sigma_{yy}^{\frac{1}{2}} \mathbf{v}_t \end{bmatrix}, \quad \tilde{r}_t = \frac{1}{\sqrt{2}} \begin{bmatrix} \Sigma_{xx}^{\frac{1}{2}} \mathbf{u}_t \\ \Sigma_{yy}^{\frac{1}{2}} \mathbf{v}_t \end{bmatrix},
$$

$$
\tilde{r}_t^* = \frac{1}{\sqrt{2}} \begin{bmatrix} \Sigma_{xx}^{\frac{1}{2}} \mathbf{u}_t^* \\ \Sigma_{yy}^{\frac{1}{2}} \mathbf{v}_t^* \end{bmatrix}, \quad r_t^* = \frac{1}{\sqrt{2}} \begin{bmatrix} \Sigma_{xx}^{\frac{1}{2}} \mathbf{u}_t^* \\ \Sigma_{yy}^{\frac{1}{2}} \mathbf{v}_t^* \end{bmatrix}, \quad r_t^* = \frac{1}{\sqrt{2}} \begin{bmatrix} \Sigma_{xx}^{\frac{1}{2}} \mathbf{u}_t^* \\ \Sigma_{yy}^{\frac{1}{2}} \mathbf{v}_t^* \end{bmatrix}.
$$

All these vectors are in $\mathbb{R}^d$ and have length 1.

Observe that the matrix-vector multiplication is equivalent to

$$
\begin{bmatrix} \Sigma_{xx}^{\frac{1}{2}} \mathbf{u}_t \\ \Sigma_{yy}^{\frac{1}{2}} \mathbf{v}_t \end{bmatrix} \leftarrow \begin{bmatrix} \Sigma_{xx}^{\frac{1}{2}} \\ \Sigma_{yy}^{\frac{1}{2}} \end{bmatrix} \left( \begin{bmatrix} \lambda \Sigma_{xx}^{\frac{1}{2}} \mathbf{u}_t^{\top} - \Sigma_{xy}^{\frac{1}{2}} \mathbf{v}_t^{\top} \\ -\Sigma_{xy}^{\frac{1}{2}} \mathbf{u}_t^{\top} \lambda \Sigma_{yy}^{\frac{1}{2}} \end{bmatrix} -1 \right) \begin{bmatrix} \Sigma_{xx}^{\frac{1}{2}} \\ \Sigma_{yy}^{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \Sigma_{xx}^{\frac{1}{2}} \mathbf{u}_t^{\top} - \Sigma_{xy}^{\frac{1}{2}} \mathbf{v}_t^{\top} \\ -\Sigma_{xy}^{\frac{1}{2}} \mathbf{u}_t^{\top} \lambda \Sigma_{yy}^{\frac{1}{2}} \end{bmatrix},
$$

and

$$
\begin{bmatrix} \Sigma_{xx}^{\frac{1}{2}} \\ \Sigma_{yy}^{\frac{1}{2}} \end{bmatrix} \leftarrow \begin{bmatrix} \lambda \Sigma_{xx}^{\frac{1}{2}} \mathbf{u}_t^{\top} - \Sigma_{xy}^{\frac{1}{2}} \mathbf{v}_t^{\top} \\ -\Sigma_{xy}^{\frac{1}{2}} \mathbf{u}_t^{\top} \lambda \Sigma_{yy}^{\frac{1}{2}} \end{bmatrix} -1 \begin{bmatrix} \Sigma_{xx}^{\frac{1}{2}} \\ \Sigma_{yy}^{\frac{1}{2}} \end{bmatrix} \left( \begin{bmatrix} \lambda \Sigma_{xx}^{\frac{1}{2}} \mathbf{u}_t^{\top} - \Sigma_{xy}^{\frac{1}{2}} \mathbf{v}_t^{\top} \\ -\Sigma_{xy}^{\frac{1}{2}} \mathbf{u}_t^{\top} \lambda \Sigma_{yy}^{\frac{1}{2}} \end{bmatrix} -1 \right) \begin{bmatrix} \Sigma_{xx}^{\frac{1}{2}} \\ \Sigma_{yy}^{\frac{1}{2}} \end{bmatrix}. 
$$

Notice that the matrix $M_\lambda$ is equal to

$$
M_\lambda = \begin{bmatrix} \Sigma_{xx}^{\frac{1}{2}} & -\Sigma_{xy}^{\frac{1}{2}} \\ -\Sigma_{xy}^{\frac{1}{2}} & \Sigma_{yy}^{\frac{1}{2}} \end{bmatrix} = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}.
$$
Then the updates for exact power iterations can be written as

\[ \tilde{r}_t^* \leftarrow M_\Lambda r_{t-1}^*, \quad r_t^* \leftarrow \tilde{r}_t^*/\|\tilde{r}_t^*\|, \quad t = 1, \ldots, \]

and the updates for inexact power iterations can be written as

\[ \bar{r}_t \approx M_\Lambda r_{t-1}, \quad r_t \leftarrow \bar{r}_t/\|\bar{r}_t\|, \quad t = 1, \ldots. \]

Note we have according to (18) that

\[ \bar{\varepsilon} \geq (\bar{r}_t - \bar{r}_t)\top M_\Lambda^{-1} (\bar{r}_t - \bar{r}_t) \geq \sigma_d(M_\Lambda^{-1}) \cdot \|\bar{r}_t - \bar{r}_t\| \]

or equivalently

\[ \|\bar{r}_t - \bar{r}_t\| \leq \sqrt{\sigma_1(M_\Lambda)} \cdot \bar{\varepsilon}. \quad (19) \]

**Proof of Lemma**

Recall that the eigenvectors of \( M_\Lambda \) are:

\[ \lambda_1 := \frac{1}{\lambda - \rho_1} > \lambda_2 := \frac{1}{\lambda - \rho_2} \geq \cdots \geq \lambda_{d-1} := \frac{1}{\lambda + \rho_2} \geq \lambda_d := \frac{1}{\lambda + \rho_1}, \]

with corresponding eigenvectors

\[ e_1 = r^* = \frac{1}{\sqrt{2}} \begin{bmatrix} a_1 \\ b_1 \end{bmatrix}, \quad e_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} a_2 \\ b_2 \end{bmatrix}, \ldots, \quad e_{d-1} = \frac{1}{\sqrt{2}} \begin{bmatrix} a_2 \\ -b_2 \end{bmatrix}, \quad e_d = \frac{1}{\sqrt{2}} \begin{bmatrix} a_3 \\ -b_1 \end{bmatrix}. \]

By the update rule of exact power iterations, it holds that for \( i = 1, \ldots, d \) that

\[ (e_i\top r_t^*)^2 = \frac{(e_i\top M_\Lambda^t r_0^*)^2}{\|M_\Lambda^t r_0^*\|^2} = \frac{(e_i\top M_\Lambda^t r_0^*)^2}{(r_0^*)\top M_\Lambda^t M_\Lambda r_0^*} = \frac{(\lambda'^1 e_i\top r_0^*)^2}{\sum_{j=1}^d \lambda_j^t (e_j\top r_0^*)^2} = \frac{(e_i\top r_0^*)^2}{\sum_{j=1}^d \left(\frac{\lambda_j}{\lambda_1}\right)^{2t} (e_j\top r_0^*)^2} \]

\[ \leq \frac{(e_i\top r_0^*)^2}{\left(\frac{\lambda_1}{\lambda_i}\right)^{2t} (e_i\top r_0^*)^2} = \frac{(e_i\top r_0^*)^2}{\left(\frac{\lambda_i}{\lambda_1}\right)^{2t}} = \frac{(e_i\top r_0^*)^2}{\mu} \cdot \exp \left(-2 \frac{\lambda_i - \lambda_1}{\lambda_1} t \right). \]

Given \( \delta \in (0, 1) \), define \( S(\delta) = \{ i : \lambda_i > (1 - \delta) \lambda_1 \} \). For \( \delta_1, \delta_2 \in (0, 1) \), define

\[ T(\delta_1, \delta_2) := \left[ 1 - \frac{1}{2\delta_1} \log \left( \frac{1}{\mu \delta_2} \right) \right]. \]

For all \( i \notin S(\delta_1) \), when \( t > T(\delta_1, \delta_2) \), it holds that \( (e_i\top r_t^*)^2 \leq \delta_2 (e_i\top r_0^*)^2 \), and thus in particular

\[ \sum_{i \in S(\alpha/2)} (e_i\top r_t^*)^2 \geq 1 - \delta_2. \]

Part one (crude regime) of the lemma now follows by noticing that, by setting \( \delta_1 = \delta_2 = \frac{\alpha}{2} \) we have that for \( t \geq T \left( \frac{\alpha}{2}, \frac{\alpha}{2} \right) \), it holds that

\[ (r_t^*)\top M_\Lambda r_t^* = \sum_{i=1}^d \lambda_i (e_i\top r_t^*)^2 \geq \sum_{i \in S(\alpha/2)} \left(1 - \frac{\alpha}{2}\right) \lambda_1 (e_i\top r_t^*)^2 \geq \left(1 - \frac{\alpha}{2}\right)^2 \lambda_1 \geq (1 - \alpha) \lambda_1. \]

For the second part (accurate regime) of the lemma, note that \( S \left( \frac{\Lambda - \lambda_1}{\lambda_1}, \alpha \right) = \{1\} \). Thus for all \( t \geq T \left( \frac{\Lambda - \lambda_1}{\lambda_1}, \alpha \right) \), it holds that \( (e_1\top r_t^*)^2 \geq 1 - \alpha. \)
**Proof of Lemma 8** We prove the bound for unnormalized iterates by induction. The case for \( t = 1 \) holds trivially. For \( t \geq 2 \), we can bound the error of the unnormalized iterates using the exact solution to \( h_t \):

\[
\|\tilde{r}_t - \tilde{r}_t^*\| \leq \|\tilde{r}_t - \hat{r}_t\| + \|\hat{r}_t - \tilde{r}_t^*\|.
\]  
(20)

The second term of (20) is concerned with the error due to inexact target in the least squares problem \( h_t(u, v) \) as \( \begin{bmatrix} u_{t-1} \\ v_{t-1} \end{bmatrix} \) is different from \( \begin{bmatrix} u_{t-1}^* \\ v_{t-1}^* \end{bmatrix} \). We can bound this term as

\[
\|\tilde{r}_t - \tilde{r}_t^*\| = \|M_{\lambda}r_{t-1} - M_{\lambda}r_{t-1}^*\| \leq \|M_{\lambda}\| \|r_{t-1} - r_{t-1}^*\| = \sigma_1(M_{\lambda}) \|r_{t-1} - r_{t-1}^*\|.
\]  
(21)

In view of the update rule of our algorithm and the triangle inequality, we have

\[
\|r_{t-1} - r_{t-1}^*\| \\
\leq \left(\frac{\|\tilde{r}_t - \hat{r}_t\|}{\|\tilde{r}_t - \hat{r}_t\|} \frac{1}{\|\tilde{r}_t - \hat{r}_t\|} + \frac{1}{\|\tilde{r}_t - \hat{r}_t\|} \|\tilde{r}_t - \hat{r}_t\| \right) \\
= \|r_{t-1} - r_{t-1}^*\| \|\tilde{r}_t - \hat{r}_t\| + \frac{1}{\|\tilde{r}_t - \hat{r}_t\|} \|\tilde{r}_t - \hat{r}_t\| \\
\leq \frac{2}{\|\tilde{r}_t - \hat{r}_t\|} \|r_{t-1} - r_{t-1}^*\| \leq \frac{2}{\|\tilde{r}_t - \hat{r}_t\|} \|\tilde{r}_t - \hat{r}_t\|.
\]  
(22)

For \( t \geq 2 \), we have \( \tilde{r}_t^* = M_{\lambda}r_{t-2}^* \) and \( \|r_{t-2}^*\| = 1 \), and thus

\[\|r_{t-1}^*\| \geq \sigma_d(M_{\lambda}).\]

Combining (20), (21) and (22) gives

\[\|\tilde{r}_t - \tilde{r}_t^*\| \leq \sqrt{\sigma_1(M_{\lambda})} \cdot e + 2\kappa_{\lambda} \tilde{R}_{t-1} = \tilde{R}_t.\]

The bound for normalized iterates follows from (22).

**Proof of Lemma 9** For the first item (crude regime), observe that

\[
r_t^T M_{\lambda} r_t = (r_t^*)^T M_{\lambda} r_t^* + (r_t^*)^T M_{\lambda} r_t^* - r_t^T M_{\lambda} r_t.
\]  
(23)

and that

\[
\| (r_t^*)^T M_{\lambda} (r_t^* - r_t^T M_{\lambda} r_t) \| = \left| (M_{\lambda}^2 r_t^* + M_{\lambda} r_t^* - M_{\lambda}^2 r_t^*)^T (M_{\lambda}^2 r_t^* - M_{\lambda}^2 r_t^*) \right| \\
\leq \| M_{\lambda}^2 r_t^* + M_{\lambda} r_t^* \| \| M_{\lambda}^2 r_t^* - M_{\lambda}^2 r_t^* \| \\
\leq \| M_{\lambda}^2 \| \| r_t^* + r_t \| \| M_{\lambda}^2 \| \| r_t^* - r_t \| \\
\leq \| M_{\lambda} \| (\| r_t^* \| + \| r_t \|) \| r_t^* - r_t \| \\
= 2\sigma_1(M_{\lambda}) \| r_t^* - r_t \|.
\]

Our choices of \( T \) and \( \tilde{c} \) make sure that \( (r_T^*)^T M_{\lambda} r_T^* \geq (1 - \frac{\alpha}{2}) \cdot \sigma_1(M_{\lambda}) \) by Lemma 7 and that \( \| r_T^* - r_T \| \leq R_T = \alpha/4 \) by Lemma 8. Continuing from (23), we have

\[
r_T^T M_{\lambda} r_T \geq \left( 1 - \frac{\alpha}{2} \right) \cdot \sigma_1(M_{\lambda}) - \frac{\alpha}{2} \cdot \sigma_1(M_{\lambda}) = (1 - \alpha) \cdot \sigma_1(M_{\lambda}).
\]

For the second item (accurate regime), observe that

\[
(r_T^* r_T^* - r_T^* r_T - r_T^* r_T) \geq (r_T^* - r_T^* r_T)^2 - 2 \| r_T - r_T^* \|.  
\]  
(24)

Our choices of \( T \) and \( \tilde{c} \) make sure that \( (r_T^* r_T^*)^2 \geq 1 - \frac{\alpha}{2} \) by Lemma 7 and that \( \| r_T^* - r_T \| \leq R_T = \alpha/4 \) by Lemma 8. Continuing from (24), we have

\[
(r_T^* r_T^*)^2 \geq 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha.
\]

\]
F.2 Iteration complexity of Algorithm \[3\]

Observe that, the for loops within the repeat-until loop, as well as the final for loop in Algorithm \[3\] are running inexact power iterations on \(M_{\lambda_{(s)}}\) and \(M_{\lambda_{(f)}}\) for \(m_1\) and \(m_2\) inexact matrix-vector multiplication respectively. And the convergence of inexact power iterations is provided by Lemma \[8\] for each iteration of the repeat-until loop, we work in the crude regime and only require \(r_{sm_1}\) to give a constant multiple estimate of \(M_{\lambda_{(s)}}\). The lemma below shows an important property of \(\Delta_s\) which is used to locate \(\lambda_{(f)}\), and the number of iterations needed to reach \(\lambda_{(f)}\).

**Lemma 10** (Iteration complexity of the repeat-until loop in Algorithm \[3\]). Suppose that \(\tilde{\Delta} \in [c_1\Delta, c_2\Delta]\) where \(c_2 \leq 1\). Set \(m_1 = \lceil 8\log\left(\frac{16}{\mu^2}\right) \rceil\) and \(\tilde{\epsilon} \leq \frac{1}{3084}\left(\frac{\Delta}{18}\right)^{m_1-1}\) in Algorithm \[3\].

Then for all \(s \geq 1\) it holds that

\[
\frac{1}{2}(\lambda_{(s-1)} - \rho_1) \leq \Delta_s \leq \lambda_{(s-1)} - \rho_1,
\]

upon exiting this loop, the \(\lambda_{(f)}\) satisfies

\[
\rho_1 + \frac{\tilde{\Delta}}{4} \leq \lambda_{(f)} \leq \rho_1 + \frac{3\tilde{\Delta}}{2},
\]

and the number of iterations run by the repeat-until loop is \(\log\left(\frac{1}{\tilde{\epsilon}}\right)\).

**Proof.** Let \(\sigma\) be an upper bound of all \(\sigma_1(M_{\lambda_{(s)}})\) used in the repeat-until loop, i.e.,

\[
\sigma \geq \sigma_1(M_{\lambda_{(s)}}), \quad s = 1, 2, \ldots.
\]

And suppose for now that throughout the loop, \(\tilde{\epsilon}\) satisfies

\[
\sqrt{\sigma \tilde{\epsilon}} \leq \frac{\sigma_1(M_{\lambda_{(s-1)}})}{8}.
\]

Set \(\alpha = \frac{1}{4}\) in Lemma \[8\] (crude regime), and with our choice of \(m_1\) and

\[
\tilde{\epsilon} \leq \frac{\sigma_d(M_{\lambda_{(s)}})}{1024\kappa_{\lambda_{(s)}}} \left(\frac{2\kappa_{\lambda_{(s)}} - 1}{(2\kappa_{\lambda_{(s)}})^{m_1} - 1}\right)^2,
\]

we have

\[
\vec{r}_{sm_1} M_{\lambda_{(s-1)}} \vec{r}_{sm_1} \geq \frac{3}{4} \sigma_1(M_{\lambda_{(s-1)}}) \cdot \tilde{\epsilon}.
\]

In view of the definition of the vector \(w_s\) in Algorithm \[3\] and following the same argument in \[18\], we have

\[
\left\| \frac{z_s}{\sqrt{2}} - M_{\lambda_{(s-1)}} \vec{r}_{sm_1} \right\| \leq \sqrt{\sigma_1(M_{\lambda_{(s-1)}})} \cdot \tilde{\epsilon}
\]

where \(z_s = \begin{bmatrix} \Sigma_{xx}^{\frac{1}{2}} & 0 \\ 0 & \Sigma_{yy}^{\frac{1}{2}} \end{bmatrix} w_s\).

Then for every iteration of the repeat-until loop, it holds that

\[
\frac{1}{2} \begin{bmatrix} u_{sm_1}^T & v_{sm_1}^T \end{bmatrix} \begin{bmatrix} \Sigma_{xx} & 0 \\ 0 & \Sigma_{yy} \end{bmatrix} w_s = \vec{r}_{sm_1}^T \begin{bmatrix} \frac{z_s}{\sqrt{2}} \end{bmatrix} = \vec{r}_{sm_1}^T M_{\lambda_{(s-1)}} \vec{r}_{sm_1} + \vec{r}_{sm_1}^T \left( \frac{z_s}{\sqrt{2}} - M_{\lambda_{(s-1)}} \vec{r}_{sm_1} \right)
\]

\[
\leq \begin{bmatrix} \vec{r}_{sm_1} M_{\lambda_{(s-1)}} \vec{r}_{sm_1} & -\sqrt{\sigma_1(M_{\lambda_{(s-1)}})} \cdot \tilde{\epsilon} \end{bmatrix} + \begin{bmatrix} \vec{r}_{sm_1} M_{\lambda_{(s-1)}} \vec{r}_{sm_1} & + \sqrt{\sigma_1(M_{\lambda_{(s-1)}})} \cdot \tilde{\epsilon} \end{bmatrix}
\]

\[
\leq \begin{bmatrix} \vec{r}_{sm_1} M_{\lambda_{(s-1)}} \vec{r}_{sm_1} & -\sqrt{\sigma_1(M_{\lambda_{(s-1)}})} \cdot \tilde{\epsilon} \end{bmatrix} + \begin{bmatrix} \vec{r}_{sm_1} M_{\lambda_{(s-1)}} \vec{r}_{sm_1} & + \sqrt{\sigma_1(M_{\lambda_{(s-1)}})} \cdot \tilde{\epsilon} \end{bmatrix}.
\]
where we have used the Cauchy-Schwartz inequality in the second step.

In view of (26) and (28), it follows that

\[
\frac{1}{2} \left[ u_{sm1}^T v_{sm1}^T \right] \begin{bmatrix} \Sigma_{xx} & \Sigma_{yy} \end{bmatrix} w_s - \sqrt{\sigma e} \\
\in \begin{bmatrix} r_{sm1}^T M_{\lambda(s-1)} r_{sm1} - 2 \sqrt{\sigma e}, & r_{sm1}^T M_{\lambda(s-1)} r_{sm1} \end{bmatrix} \\
\in \left[ \frac{1}{2} \sigma_1 (M_{\lambda(s-1)}), \sigma_1 (M_{\lambda(s-1)}) \right].
\]

By the definition of \( \Delta_s \) in Algorithm 3 and the fact that \( \sigma_1 (M_{\lambda(s-1)}) = \frac{1}{\lambda(s-1) - \rho_1} \), we have

\[
\Delta_s = \frac{1}{2} \left[ u_{sm1}^T v_{sm1}^T \right] \begin{bmatrix} \Sigma_{xx} & \Sigma_{yy} \end{bmatrix} w_s - \sqrt{\sigma e} \in \left[ \frac{1}{2} (\lambda(s-1) - \rho_1), \lambda(s-1) - \rho_1 \right]. \tag{29}
\]

And as a result,

\[
\lambda(s) = \lambda(s-1) - \frac{\Delta_s}{2} = \lambda(s-1) - \frac{1}{2} (\lambda(s-1) - \rho_1) = \frac{\lambda(s-1) + \rho_1}{2},
\]

and thus by induction (note \( \lambda(0) \geq \rho_1 \)) we have \( \lambda(s) \geq \rho_1 \) throughout the repeat-until loop. From (29) we also obtain

\[
\lambda(s) - \rho_1 = \lambda(s-1) - \rho_1 - \frac{\Delta_s}{2} \leq \lambda(s-1) - \rho_1 - \frac{1}{4} (\lambda(s-1) - \rho_1) = \frac{3}{4} (\lambda(s-1) - \rho_1).
\]

To sum up, \( \lambda(s) \) approaches \( \rho_1 \) from above and the gap between \( \lambda(s) \) and \( \rho_1 \) reduces at the geometric rate of \( \frac{3}{4} \). Thus after at most \( t_3 = \lceil \log_{3/4} \left( \frac{\lambda(s)}{\rho_1} \right) \rceil \sim O \left( \log \left( \frac{1}{\Delta} \right) \right) \) iterations, we reach a \( \lambda(t_3) \) such that \( \lambda(t_3) - \rho_1 \leq \tilde{\Delta} \). And in view of (29), the repeat-until loop exits in the next iteration. Hence, the overall number of iterations is at most \( t_3 + 1 = O \left( \frac{1}{\Delta} \right) \).

We now analyze \( \lambda(f) \) and derive the interval it lies in. Note that \( \Delta_f \leq \tilde{\Delta} \) and \( \Delta_{f-1} > \tilde{\Delta} \) by the exiting condition. In view of (29), we have

\[
\lambda(f) - \rho_1 = \lambda(f-1) - \rho_1 - \frac{\Delta_f}{2} \leq 2 \Delta_f - \frac{\Delta_f}{2} = \frac{3 \Delta_f}{2} \leq \frac{3 \tilde{\Delta}}{2}.
\]

On the other hand,

\[
\lambda(f) - \rho_1 = \lambda(f-1) - \rho_1 - \frac{\Delta_f}{2} \geq \lambda(f-1) - \rho_1 - \frac{1}{4} (\lambda(f-1) - \rho_1) = \frac{1}{4} (\lambda(f-1) - \rho_1) \tag{30}.
\]

If \( f = 1 \), then by our choice of \( \lambda(0) \) we have that \( \lambda(f) - \rho_1 \geq \tilde{\Delta} \). Otherwise, by unfolding (30) one more time, we have that

\[
\lambda(f) - \rho_1 \geq \frac{1}{4} (\lambda(f-2) - \rho_1) \geq \frac{\Delta_{f-2}}{4} \geq \frac{\tilde{\Delta}}{4}.
\]

Thus in both case, we have that \( \lambda(f) - \rho_1 \geq \frac{\tilde{\Delta}}{4} \) holds.

It remains to give an explicit bound on \( \tilde{\epsilon} \) based on the two requirements (26) and (27). Since the \( \lambda(s) \) values are monotonically non-increasing and lower-bounded by \( \rho_1 + \frac{\tilde{\Delta}}{4} \), we have

\[
\max_s \sigma_1 (M_{\lambda(s)}) = \sigma_1 (M_{\lambda(f)}) = \frac{1}{\lambda(f) - \rho_1} \leq \frac{4}{\tilde{\Delta}} =: \varpi,
\]

and

\[
\min_s \sigma_1 (M_{\lambda(s)}) = \sigma_1 (M_{\lambda(0)}) = \frac{1}{\lambda(0) - \rho_1} = \frac{1}{1 + \Delta - \rho_1} \geq 1 + \frac{1 - c_2 \Delta}{c_2} \geq 1 + \frac{1 - c_2 \Delta}{c_2} \Delta := \varrho.
\]

20
We now derive a lower bound of the right hand side of (27). Notice

\[ \sigma_d(M_{\lambda(s)}) \geq \sigma_d(M_{\lambda(0)}) = \frac{1}{\lambda(0) + \rho_1} = \frac{1}{1 + \Delta + \rho_1} \geq \frac{1}{3}. \]

As a result, we have

\[ \left( \frac{\varphi}{8\sqrt{\sigma}} \right)^2 = \left( \frac{1 + \frac{c_3}{c_2} \frac{\Delta}{\sigma}}{64 \cdot \frac{\Delta}{\sigma}} \right)^2 \geq \frac{1}{64 \cdot \frac{\Delta}{\sigma}} = \frac{\Delta}{256} \geq \hat{\varepsilon}. \] (31)

We now derive a lower bound of the right hand side of (35). Notice

\[ \kappa_{\lambda(s)} = \frac{\lambda(s) + \rho_1}{\lambda(s) - \rho_1} = 1 + \frac{2 \rho_1}{\lambda(s) - \rho_1} \leq 1 + 2 \rho_1 \sigma / \Delta \leq 1 + 2 \sigma \leq \frac{9}{\Delta}. \] (32)

On the other hand,

\[ \sigma_d(M_{\lambda(s)}) \geq \sigma_d(M_{\lambda(0)}) = \frac{1}{\lambda(0) + \rho_1} = \frac{1}{1 + \Delta + \rho_1} \geq \frac{1}{3}. \]

In view of (32), we work in the accurate regime of power iterations. Recall that we have proved in (32) that

\[ \kappa_{\lambda(f)} = \frac{\lambda(f) + \rho_1}{\lambda(f) - \rho_1} = 1 + \frac{2 \rho_1}{\lambda(f) - \rho_1} \leq 1 + \frac{2 \rho_1}{\lambda(f) - \rho_1} \leq 1 + \frac{2 \sigma}{\Delta} \leq \frac{16}{\Delta}. \]

In algorithm (3), we set

\[ \eta = \frac{\eta^2}{64} \] (34)

Proof: Notice when \( \lambda = \rho_1 + c(\rho_1 - \rho_2) \), we have

\[ \delta(M_{\lambda}) = \frac{\sigma_1(M_{\lambda})}{\sigma_1(M_{\lambda}) - \sigma_2(M_{\lambda})} = \frac{\lambda - \rho_2}{\lambda - \rho_1 - \rho_2} = \frac{\rho_1 + c(\rho_1 - \rho_2) - \rho_2}{\rho_1 - \rho_2} = c + 1. \]

In view of (32), \( \lambda_{(f)} - \rho_1 \leq \frac{3}{2} \Delta \leq \frac{3c_3}{2} \Delta \leq \frac{3}{2} \Delta \), and thus \( \delta(M_{\lambda_{(f)}}) \leq \frac{3}{2} \Delta \). Set \( \alpha = \frac{\eta^2}{64} \) in Lemma 8 (accurate regime), and with our choice of \( m_2 \) and

\[ \hat{\varepsilon} \leq \frac{\eta^4 \cdot \sigma_d(M_{\lambda(f)})}{64^3 \cdot \kappa_{\lambda(f)}} \left( \frac{2 \kappa_{\lambda(f)} - 1}{(2 \kappa_{\lambda(f)})^{m_2} - 1} \right)^2, \] (35)

we are guaranteed to obtained the desired alignment.

We now give a lower bound of the right hand side of (35). First,

\[ \sigma_d(M_{\lambda_{(f)}}) = \frac{1}{\lambda_{(f)} + \rho_1} \geq \frac{1}{\rho_1 + \frac{3}{2} \Delta + \rho_1} \geq \frac{1}{4}. \]

Recall that we have proved in (32) that \( \kappa_{\lambda_{(f)}} \leq \frac{9}{\Delta} \). Following a derivation similar to that of (35), we have

\[ \frac{\eta^4 \cdot \sigma_d(M_{\lambda_{(f)}})}{64^3 \cdot \kappa_{\lambda(f)}} \left( \frac{2 \kappa_{\lambda_{(f)}} - 1}{(2 \kappa_{\lambda_{(f)})}^{m_2} - 1} \right)^2 \geq \frac{\eta^4}{4^{10}} \left( \frac{\Delta}{18} \right)^{m_2} \] (36)
Proof of Theorem. As shown in Lemma 11, the repeat-until loop runs $O \left( \log \left( \frac{1}{\Delta} \right) \right) \sim O \left( \log \left( \frac{1}{\alpha} \right) \right)$ iterations, and inside each iteration, we run $m_1$ approximate matrix-vector multiplications. On the other hand, the final for loop runs $m_2$ approximate matrix-vector multiplications. By the definitions of $m_1$ and $m_2$, the total number of invocations of approximate matrix-vector multiplications/least squares problems is

$$m_1 \cdot \log \left( \frac{1}{\Delta} \right) + m_2 \sim O \left( \log \left( \frac{1}{\mu} \right) \log \left( \frac{1}{\Delta} \right) + \log \left( \frac{1}{\mu \eta^2} \right) \right) \sim \tilde{O}(1).$$

$\square$

G Proof of Theorem 5

Proof. Notice that the eigenvectors of $M_1$ form an orthonormal bases of $\mathbb{R}^{d_x+d_y}$. Thus when (34) holds, i.e., the alignment between $\begin{bmatrix} \Sigma_{xx}^\frac{1}{2} \hat{u}_T \\ \Sigma_{yy}^\frac{1}{2} \hat{v}_T \end{bmatrix}$ and the tailing eigenvector $\begin{bmatrix} \Sigma_{xx}^\frac{1}{2} u^* \\ -\Sigma_{yy}^\frac{1}{2} v^* \end{bmatrix}$ has to be small:

$$(u_T^T \Sigma_{xx} u^* - v_T^T \Sigma_{yy} v^*)^2 \leq \frac{\eta^2}{16}. \quad (37)$$

From (37) and (54), we have respectively

$$-\frac{\eta}{4} \leq |u_T^T \Sigma_{xx} u^*| - |v_T^T \Sigma_{yy} v^*| \leq \frac{\eta}{4},$$

$$|u_T^T \Sigma_{xx} u^*| + |v_T^T \Sigma_{yy} v^*| \geq 2 \sqrt{1 - \frac{\eta^2}{64}} \geq 2 \left( 1 - \frac{\eta}{8} \right)$$

where we have used the fact that $\sqrt{1 - x} \geq 1 - \sqrt{x}$ for $x \in [0, 1]$ in the second inequality. Averaging the above two inequalities gives

$$|u_T^T \Sigma_{xx} u^*| \geq 1 - \frac{\eta}{4}, \quad |v_T^T \Sigma_{yy} v^*| \geq 1 - \frac{\eta}{4}.$$

Finally,

$$(\hat{u}^T \Sigma_{xx} u^*)^2 + (\hat{v}^T \Sigma_{yy} v^*)^2 \geq (u_T^T \Sigma_{xx} u_T)^2 + (v_T^T \Sigma_{yy} v_T)^2$$

$$\geq (1 - \frac{\eta}{4})^2 \left( \frac{1}{u_T^T \Sigma_{xx} u_T} + \frac{1}{v_T^T \Sigma_{yy} v_T} \right)$$

$$\geq (1 - \frac{\eta}{4})^2 \left( \frac{4}{u_T^T \Sigma_{xx} u_T + v_T^T \Sigma_{yy} v_T} \right)$$

$$\geq 2 \left( 1 - \frac{\eta}{2} \right) = 2 - \eta$$

where we have used the fact that $\frac{1}{x} + \frac{1}{y} \geq \frac{4}{x+y}$ in the first inequality, and (10) in the second inequality. Then the theorem follows from the fact that $(\hat{u}^T \Sigma_{xx} u^*)^2$ and $(\hat{v}^T \Sigma_{yy} v^*)^2$ can be at most 1. $\square$

H Condition number of $h_t$ for SVRG

Lemma 12. Throughout Algorithm 3 the condition number of $h_t$ for SVRG is at most $\frac{g}{c} \tilde{\kappa}$, where

$$\tilde{\kappa} := \max_i \max_j \left( \frac{\|x_i\|^2, \|y_i\|^2}{\min_j (\sigma_{\min}(\Sigma_{xx}), \sigma_{\min}(\Sigma_{yy}))} \right).$$
Proof. The gradient Lipschitz constant of $h^i_t(u,v)$ is bounded by the largest eigenvalue (in absolute value) of its Hessian:

$$Q^i_\lambda = \begin{bmatrix} \lambda x_i x_i^T & -x_i y_i^T \\ -y_i x_i^T & \lambda y_i y_i^T \end{bmatrix},$$

and the largest eigenvalue is defined as

$$\max_{g_x \in \mathbb{R}^{d_x}, g_y \in \mathbb{R}^{d_y}} \beta := \left| g_x^T Q^i_\lambda g_y \right| \quad \text{s.t.} \quad \|g_x\|^2 + \|g_y\|^2 = 1.$$

We have

$$\beta = |\lambda (g_x^T x_i)^2 + \lambda (g_y^T y_i)^2 - 2 (g_x^T x_i) (g_y^T y_i)|$$

$$\leq \lambda (g_x^T x_i)^2 + \lambda (g_y^T y_i)^2 + 2 |g_x^T x_i| |g_y^T y_i|$$

$$= (\lambda + 1) \left( (g_x^T x_i)^2 + (g_y^T y_i)^2 \right)$$

$$\leq (\lambda + 1) \left( \|g_x\|^2 \|x_i\|^2 + \|g_y\|^2 \|y_i\|^2 \right)$$

where we have used the Cauchy-Schwartz inequality and the constraint in the third and the last inequality respectively.

It only remains to bound $\frac{\lambda + 1}{\lambda - \rho}$. Note that we have shown in Lemma 10 that $\lambda \geq \rho_1 + \frac{\Delta}{4}$ throughout Algorithm 3 and thus

$$\frac{\lambda + 1}{\lambda - \rho} = 1 + \frac{1 + \rho}{\lambda - \rho} \leq 1 + \frac{2}{\lambda - \rho} \leq 1 + \frac{4}{\Delta} \leq \frac{9}{\Delta} \leq \frac{9/c_1}{\Delta}. $$


I. More details of the experiments

The statistics of these datasets are summarized in Table 2. These datasets have also been used by [3, 4] for demonstrating their stochastic CCA algorithms.

| Dataset     | Description                        | $d_x$ | $d_y$ | $N$  |
|-------------|------------------------------------|-------|-------|------|
| Mediamill   | Image and its labels               | 100   | 120   | 30,000 |
| JW11        | Acoustic and articulation measurements | 273   | 112   | 30,000 |
| MNIST       | Left and right halves of images    | 392   | 392   | 60,000 |

We now provide additional details for the experiments. For $s$-AppGrad, both gradient and normalization steps are estimated with mini-batches of 100 samples (the authors of [3] suggest that the mini-batch size shall be at least the same magnitude as the dimensionality of the CCA projection). For SI-VR and SI-AVR, within the repeat-until loop, we apply SVRG with $M = 2$ epochs to approximately find the top eigenvector $w_s$, and SVRG with $M = 2$ epochs to approximately calculate its top eigenvalue of $M_{\lambda(s)}$ as $w_s^T M_{\lambda(s)} w_s$. We exit the repeat-until loop when $\Delta_s \leq 0.06$. Afterwards, for the fixed $\lambda(f)$, we apply SVRG to solve every least squares problem with $M = 4$ epochs. Each epoch of SVRG includes a batch gradient evaluation and $m = N$ stochastic gradient steps. We set the step size according to the smoothness for each least squares solver, i.e., $\frac{1}{\max_i \|x_i\|^2}$ for GD/AGD in AppGrad/s/AppGrad/CCALin, and $\frac{1}{\max_i \|x_i\|^2}$ for SVRG/ASVRG in our algorithms.

---

7We omit the regularization terms, which are typically very small, to have concise expressions.
J Other related work

Recent years have witnessed continuous efforts to scale up fundamental methods such as principal component analysis (PCA) and partial least squares with stochastic/online updates \cite{22, 23, 24, 25, 5, 16, 17}. But as pointed out by \cite{23}, the CCA objective is more challenging due to the constraints. \cite{26} proposed an adaptive CCA algorithm with efficient online updates based on matrix manifolds defined by the constraints. However, the goal of their algorithm is anomaly detection for streaming data with a varying distribution, rather than to optimize the CCA objective on a given dataset. Similar to our algorithms, the stochastic CCA algorithms of \cite{3, 4} are motivated by the ALS formulation. \cite{5} proposed a stochastic algorithm based on the Lagrangian formulation of the objective \cite{1}. None of these online/stochastic algorithms have rigorous global convergence guarantee.