Journal of Machine Learning Research 20 (2019) 1-46 Submitted 2/18; Published 10/19

Stochastic Canonical Correlation Analysis

Chao Gao CHAOGAO@GALTON.UCHICAGO.EDU
University of Chicago
Chicago, IL 60637, USA

Dan Garber DANGAR@TECHNION.AC.IL
Technion – Israel Institute of Technology
Haifa, 3200003, Israel

Nathan Srebro NATI@TTIC.EDU
Toyota Technological Institute at Chicago
Chicago, IL 60637, USA

Jialei Wang JIALEI@UCHICAGO.EDU
University of Chicago
Chicago, IL 60637, USA

Weiran Wang WEIRANWANG@TTIC.EDU
Toyota Technological Institute at Chicago
Chicago, IL 60637, USA

Editor: John Shawe-Taylor

Abstract

We study the sample complexity of canonical correlation analysis (CCA), i.e., the number of samples needed to estimate the population canonical correlation and directions up to arbitrarily small error. With mild assumptions on the data distribution, we show that in order to achieve ϵ-suboptimality in a properly defined measure of alignment between the estimated canonical directions and the population solution, we can solve the empirical objective exactly with $N(\epsilon, \Delta, \gamma)$ samples, where Δ is the singular value gap of the whitened cross-covariance matrix and $1/\gamma$ is an upper bound of the condition number of auto-covariance matrices. Moreover, we can achieve the same learning accuracy by drawing the same level of samples and solving the empirical objective approximately with a stochastic optimization algorithm; this algorithm is based on the shift-and-invert power iterations and only needs to process the dataset for $\mathcal{O}\left(\log \frac{1}{\epsilon}\right)$ passes. Finally, we show that, given an estimate of the canonical correlation, the streaming version of the shift-and-invert power iterations achieves the same learning accuracy with the same level of sample complexity, by processing the data only once.

Keywords: Canonical correlation analysis, sample complexity, shift-and-invert preconditioning, streaming CCA

1. Introduction

Let $\mathbf{x} \in \mathbb{R}^{d_x}$ and $\mathbf{y} \in \mathbb{R}^{d_y}$ be two random vectors with a joint probability distribution $P(\mathbf{x}, \mathbf{y})$. The objective of CCA (Hotelling, 1936) in the population setting is to find $\mathbf{u} \in \mathbb{R}^{d_x}$ and $\mathbf{v} \in \mathbb{R}^{d_y}$ such that projections of the random variables onto these directions are

maximally correlated:[1]

$$\max_{\mathbf{u},\mathbf{v}} \frac{\mathbb{E}[(\mathbf{u}^\top \mathbf{x})(\mathbf{v}^\top \mathbf{y})]}{\sqrt{\mathbb{E}[(\mathbf{u}^\top \mathbf{x})^2]}\sqrt{\mathbb{E}[(\mathbf{v}^\top \mathbf{y})^2]}}. \tag{1}$$

This objective can be written in the equivalent constrained form

$$\max_{\mathbf{u},\mathbf{v}} \mathbf{u}^\top \mathbf{E}_{xy}\mathbf{v} \quad \text{s.t.} \quad \mathbf{u}^\top \mathbf{E}_{xx}\mathbf{u} = \mathbf{v}^\top \mathbf{E}_{yy}\mathbf{v} = 1 \tag{2}$$

where the cross- and auto-covariance matrices are defined as

$$\mathbf{E}_{xy} = \mathbb{E}[\mathbf{x}\mathbf{y}^\top], \qquad \mathbf{E}_{xx} = \mathbb{E}[\mathbf{x}\mathbf{x}^\top], \qquad \mathbf{E}_{yy} = \mathbb{E}[\mathbf{y}\mathbf{y}^\top]. \tag{3}$$

The global optimum of (2), denoted by $(\mathbf{u}^*, \mathbf{v}^*)$, can be computed in closed-form. Define

$$\mathbf{T} := \mathbf{E}_{xx}^{-\frac{1}{2}} \mathbf{E}_{xy} \mathbf{E}_{yy}^{-\frac{1}{2}} \in \mathbb{R}^{d_x \times d_y}, \tag{4}$$

and let $(\mathbf{a}_1, \mathbf{b}_1)$ be the (unit-length) top left and right singular vector pair associated with $\mathbf{T}$'s largest singular value $\rho_1 = \sigma_1(\mathbf{T})$. Then the optimal objective value, i.e., the canonical correlation between $\mathbf{x}$ and $\mathbf{y}$, is $\rho_1 \leq 1$ (see Lemma 20), achieved by $(\mathbf{u}^*, \mathbf{v}^*) = (\mathbf{E}_{xx}^{-\frac{1}{2}}\mathbf{a}_1, \mathbf{E}_{yy}^{-\frac{1}{2}}\mathbf{b}_1)$.

In practice, we do not have access to the population covariance matrices, but observe samples pairs $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_N, \mathbf{y}_N)$ drawn from $P(\mathbf{x}, \mathbf{y})$. In this paper, we are concerned with both the number of samples $N(\epsilon)$ needed to approximately solve (2), and the time complexity for obtaining the approximate solution. Note that the CCA objective is not a stochastic convex program due to the ratio form (1), and standard stochastic approximation methods do not apply (Arora et al., 2012). Globally convergent stochastic optimization of CCA has long been a challenge even for the empirical objective, and attracted continuous effort (Lu and Foster, 2014; Ma et al., 2015; Wang and Livescu, 2016), until the recent breakthrough by Ge et al. (2016); Wang et al. (2016). And our understanding of the stochastic objective, e.g., the existence of an efficient algorithm and the sample complexity, has been very limited.

Our contributions The contributions of our paper are summarized as follows.

- First, we provide the ERM sample complexity of CCA. We show that in order to achieve ϵ-suboptimality in the alignment between the estimated canonical directions and the population solution (relative to the population covariances, see Section 2), we can solve the empirical objective exactly with $N(\epsilon, \Delta, \gamma)$ samples where Δ is the singular value gap of the whitened cross-covariance and $1/\gamma$ is a upper bound of the condition number of the auto-covariance, for several general classes of distributions widely used in statistics and machine learning.

1. For simplicity (especially for the streaming setting), we assume that $\mathbb{E}[\mathbf{x}] = \mathbf{0}$ and $\mathbb{E}[\mathbf{y}] = \mathbf{0}$. Nonzero means can be easily handled in the ERM approach (see Remark 4).

- Second, to alleviate the high computational complexity of exactly solving the empirical objective, we show that we can achieve the same learning accuracy by drawing the same level of samples and solving the empirical objective approximately with the stochastic optimization algorithm of Wang et al. (2016). This algorithm is based on the shift-and-invert power iterations (Saad, 1992; Garber and Hazan, 2015; Garber et al., 2016). We provide tightened analysis of the algorithm's time complexity, removing an extra $\log \frac{1}{\epsilon}$ factor from the complexity given by Wang et al. (2016). Our analysis shows that asymptotically it suffices to process the sample set for $\mathcal{O}\left(\log \frac{1}{\epsilon}\right)$ passes. While near-linear runtime in the required number of samples is known and achieved for convex learning problems using SGD, no such result was estabilished for the nonconvex CCA objective previously.
- Third, we show that the streaming version of shift-and-invert power iterations achieves the same learning accuracy with the same level of sample complexity, given a good estimate of the canonical correlation. This approach requires only $\mathcal{O}(d)$ memory where $d := d_x + d_y$ is the input dimensionality, and thus further alleviates the memory cost of solving the empirical objective. This addresses the challenge of the existence of a stochastic algorithm for CCA proposed by Arora et al. (2012).

Notation We use $\sigma_i(\mathbf{A})$ to denote the i-th largest singular value of a matrix $\mathbf{A}$, and use $\sigma_{\max}(\mathbf{A})$ and $\sigma_{\min}(\mathbf{A})$ to denote the largest and smallest singular values of $\mathbf{A}$ respectively. We use $\|\cdot\|$ to denote the spectral norm of a matrix or the ℓ_2-norm of a vector. For a positive definite matrix $\mathbf{M}$, the vector norm $\|\cdot\|_{\mathbf{M}}$ is defined as $\|\mathbf{w}\|_{\mathbf{M}} = \sqrt{\mathbf{w}^\top \mathbf{M}\mathbf{w}}$ for any $\mathbf{w}$. We use C and C' to denote universal constants that are independent of problem parameters, and their specific values may vary among appearances. We hide poly-logarithmic dependencies in the notation $\tilde{\mathcal{O}}(\cdot)$.

2. Problem setup

Assumptions We assume the following properties of the input random variables.

1. **Bounded covariances**: Eigenvalues of population auto-covariance matrices are bounded:[2]

$$\max\left(\|\mathbf{E}_{xx}\|, \|\mathbf{E}_{yy}\|\right) \le 1,$$
$$\gamma := \min\left(\sigma_{\min}(\mathbf{E}_{xx}), \sigma_{\min}(\mathbf{E}_{yy})\right) > 0.$$

   Hence $\mathbf{E}_{xx}$ and $\mathbf{E}_{yy}$ are invertible with condition numbers bounded by $1/\gamma$.

2. **Singular value gap**: For the purpose of learning the canonical directions $(\mathbf{u}^*, \mathbf{v}^*)$, we assume that there exists a positive singular value gap $\Delta := \sigma_1(\mathbf{T}) - \sigma_2(\mathbf{T}) \in (0, 1)$, such that the top left- and right-singular vector pair of $\mathbf{T}$ is uniquely defined.

Distribution classes In this paper, we analyze three input distribution classes commonly used in the statistics and machine learning literature. Let

$$\mathbf{z} = \begin{bmatrix} \mathbf{E}_{xx} & \mathbf{E}_{xy} \\ \mathbf{E}_{xy}^\top & \mathbf{E}_{yy} \end{bmatrix}^{-\frac{1}{2}} \cdot \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \in \mathbb{R}^d, \tag{5}$$

2. CCA is invariant to linear transformations of the inputs, so we could always rescale the data.

the distribution classes are defined with $(\mathbf{x}, \mathbf{y}, \mathbf{z})$ as follows.

- **(Sub-Gaussian)** Let $\mathbf{z}$ be isotropic and sub-Gaussian, that is, $\mathbb{E}\left[\mathbf{z}\mathbf{z}^\top\right] = \mathbf{I}$ and there exists constant $C > 0$ such that $\mathbb{P}\left(\left|\mathbf{q}^\top \mathbf{z}\right| > t\right) \leq \exp(-Ct^2)$ for any unit vector $\mathbf{q}$.

- **(Regular polynomial-tail, Srivastava and Vershynin, 2013)** Let $\mathbf{z}$ be isotropic and regular polynomial-tail, that is, $\mathbb{E}\left[\mathbf{z}\mathbf{z}^\top\right] = \mathbf{I}$ and there exist constants $r > 1, C > 0$ such that $\mathbb{P}\left(\|\mathbf{V}\mathbf{z}\|^2 > t\right) \leq Ct^{-1-r}$ for any orthogonal projection $\mathbf{V}$ in $\mathbb{R}^d$ and any $t > C \cdot \operatorname{rank}(\mathbf{V})$. Note that this class is general and only implies the existence of a $(4+\delta)$-moment condition for some $\delta > 0$.

- **(Bounded)** Let $\mathbf{x}$ and $\mathbf{y}$ be bounded and in particular $\sup\left(\|\mathbf{x}\|^2, \|\mathbf{y}\|^2\right) \leq 1$ (which implies $\max\left(\|\mathbf{E}_{xx}\|, \|\mathbf{E}_{yy}\|\right) \leq 1$ as in Assumption 1).

As shown later, these classes satisfy the same concentration property, allowing us to study them (and potentially other distributions) in a unified framework.

Measure of error For an estimate $(\mathbf{u}, \mathbf{v})$ of the optimal solution to (2), which need not be correctly normalized (i.e., they may not satisfy the constraints of (2)), we can always define $(\overline{\mathbf{u}}, \overline{\mathbf{v}}) := \left(\frac{\mathbf{u}}{\|\mathbf{E}_{xx}^{\frac{1}{2}}\mathbf{u}\|}, \frac{\mathbf{v}}{\|\mathbf{E}_{yy}^{\frac{1}{2}}\mathbf{v}\|}\right)$ as the correctly normalized version. And we can measure the quality of these directions by the alignment (cosine of the angle) between $\left(\frac{1}{\sqrt{2}}\begin{bmatrix}\mathbf{E}_{xx}^{\frac{1}{2}}\overline{\mathbf{u}} \\ \mathbf{E}_{yy}^{\frac{1}{2}}\overline{\mathbf{v}}\end{bmatrix}, \frac{1}{\sqrt{2}}\begin{bmatrix}\mathbf{E}_{xx}^{\frac{1}{2}}\mathbf{u}^* \\ \mathbf{E}_{yy}^{\frac{1}{2}}\mathbf{v}^*\end{bmatrix}\right)$, or the sum of alignment between $\left(\mathbf{E}_{xx}^{\frac{1}{2}}\overline{\mathbf{u}}, \mathbf{E}_{xx}^{\frac{1}{2}}\mathbf{u}^*\right)$ and alignment between $\left(\mathbf{E}_{yy}^{\frac{1}{2}}\overline{\mathbf{v}}, \mathbf{E}_{yy}^{\frac{1}{2}}\mathbf{v}^*\right)$ (all vectors have unit length):

$$\operatorname{align}\left((\mathbf{u}, \mathbf{v}); (\mathbf{u}^*, \mathbf{v}^*)\right) := \frac{1}{2}\left(\frac{\mathbf{u}^\top \mathbf{E}_{xx}\mathbf{u}^*}{\|\mathbf{E}_{xx}^{\frac{1}{2}}\mathbf{u}\|} + \frac{\mathbf{v}^\top \mathbf{E}_{yy}\mathbf{v}^*}{\|\mathbf{E}_{yy}^{\frac{1}{2}}\mathbf{v}\|}\right).$$

This measure of alignment is invariant to the lengths of $\mathbf{u}$ and $\mathbf{v}$, and achieves the maximum of 1 if $(\mathbf{u}, \mathbf{v})$ lie in the same direction as $(\mathbf{u}^*, \mathbf{v}^*)$. Intuitively, this measure respects the geometry imposed by the CCA constraints that the projections of each view have unit length. As we will show later, this measure is also closely related to the learning guarantee we can achieve with power iterations. Moreover, high alignment implies accurate estimate of the canonical correlation.

Lemma 1 *Let $\eta \in (0, 1)$. If $\operatorname{align}\left((\mathbf{u}, \mathbf{v}); (\mathbf{u}^*, \mathbf{v}^*)\right) \geq 1 - \frac{\eta}{8}$, then*

$$\frac{\mathbf{u}^\top \mathbf{E}_{xy}\mathbf{v}}{\sqrt{\mathbf{u}^\top \mathbf{E}_{xx}\mathbf{u}}\sqrt{\mathbf{v}^\top \mathbf{E}_{yy}\mathbf{v}}} \geq \rho_1(1 - \eta).$$

All proofs are deferred to the appendix.

3. The sample complexity of ERM

One approach to address this problem is empirical risk minization (ERM): We draw N samples $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$ from $P(\mathbf{x}, \mathbf{y})$ and solve the empirical version of (2):

$$\max_{\mathbf{u},\mathbf{v}} \quad \mathbf{u}^\top \mathbf{\Sigma}_{xy} \mathbf{v} \quad \text{s.t.} \quad \mathbf{u}^\top \mathbf{\Sigma}_{xx} \mathbf{u} = \mathbf{v}^\top \mathbf{\Sigma}_{yy} \mathbf{v} = 1 \tag{6}$$

where the empirical covariance matrices are defined as

$$\mathbf{\Sigma}_{xy} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i \mathbf{y}_i^\top, \quad \mathbf{\Sigma}_{xx} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i^\top, \quad \mathbf{\Sigma}_{yy} = \frac{1}{N} \sum_{i=1}^N \mathbf{y}_i \mathbf{y}_i^\top. \tag{7}$$

Similarly, define the empirical version of $\mathbf{T}$ as

$$\widehat{\mathbf{T}} := \mathbf{\Sigma}_{xx}^{-\frac{1}{2}} \mathbf{\Sigma}_{xy} \mathbf{\Sigma}_{yy}^{-\frac{1}{2}} \in \mathbb{R}^{d_x \times d_y}. \tag{8}$$

We will approximate the population canonical correation and directions based on solution to the above empirical objective.

Before going to the detailed analysis, we highlight the key property that enable us to study different input distributions in a unified manner. In fact this property is the only place we handle the stochasticity of data in studying ERM.

Proposition 2 (Concentration property) *For any $\nu > 0$, with sufficiently large sample sizes $N_0(\nu)$, the following inequality is satisfied with high probability by sub-Gaussian, regular polynomial-tail, and bounded random variables:* [3]

$$\max \left(\|\mathbf{E}_{xx}^{-\frac{1}{2}} \mathbf{\Sigma}_{xx} \mathbf{E}_{xx}^{-\frac{1}{2}} - \mathbf{I}\|, \ \|\mathbf{E}_{yy}^{-\frac{1}{2}} \mathbf{\Sigma}_{yy} \mathbf{E}_{yy}^{-\frac{1}{2}} - \mathbf{I}\|, \ \|\mathbf{E}_{xx}^{-\frac{1}{2}} (\mathbf{\Sigma}_{xy} - \mathbf{E}_{xy}) \mathbf{E}_{yy}^{-\frac{1}{2}}\| \right) \leq \nu. \tag{9}$$

We provide detailed bounds on $N_0(\nu)$ for different distributions in Lemma 3.

Roadmap for this section We proceed to analyze the sample complexities, eventually obtained in Theorem 9. We first analyze the concentration property of different classes in Lemma 3, and provide the number of samples needed to guarantee small perturbation between $\widehat{\mathbf{T}}$ and $\mathbf{T}$ in Lemma 6, which by the Weyl's inequality (Horn and Johnson, 1986) provides the sample complexity for learning canonical correlations (regardless of the existence of a singular value gap for $\mathbf{T}$). Then by the perturbation of singular vectors and after fixing the issue of normalization, we obtain guarantees for the alignment between the estimated and the optimal canonical directions.

3.1. Approximating the canonical correlation

We first discuss the error of approximating ρ_1 by $\widehat{\rho}_1 = \sigma_1(\widehat{\mathbf{T}})$. Observe that, although the empirical covariance matrices are unbiased estimates of their population counterparts, we do *not* have $\mathbb{E}[\widehat{\mathbf{T}}] = \mathbf{T}$ due to the nonlinear operations (matrix multiplication, inverse, and square root) involved in computing $\mathbf{T}$. Nonetheless, we can provide approximation

3. We refrain from specifying the failure probability as it only adds additional mild dependences to our results.

guarantee based on concentrations. We will separate the probabilistic property of data—the concentration property in Proposition 2—from the deterministic error analysis, and we show below that it is satisfied by distributions considered here.

Lemma 3 *Let Assumption 1 hold for the random variables. Then the concentration property* (9) *is satisfied with high probability, if*

$$\begin{aligned} N_0(\nu) &\geq C' \frac{d}{\nu^2} && \text{for the sub-Gaussian class,} \\ N_0(\nu) &\geq C' \frac{d}{\nu^{2(1+r^{-1})}} && \text{for the polynomial-tail class,} \\ N_0(\nu) &\geq C \frac{1}{\nu^2 \gamma^2} && \text{for the bounded class.} \end{aligned}$$

Remark 4 *When* $(\mathbf{x}, \mathbf{y})$ *have nonzero means, we use the unbiased estimate of covariance matrices* $\mathbf{\Sigma}_{xy} = \frac{\sum_{i=1}^N (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{y}_i - \bar{\mathbf{y}})^\top}{N-1}$, $\mathbf{\Sigma}_{xx} = \frac{\sum_{i=1}^N (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top}{N-1}$, *and* $\mathbf{\Sigma}_{yy} = \frac{\sum_{i=1}^N (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^\top}{N-1}$ *instead of those in* (7)*, where* $\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^N \mathbf{x}_i$ *and* $\bar{\mathbf{y}} = \frac{1}{N}\sum_{i=1}^N \mathbf{y}_i$*. We have similar concentration results, and all results in Sections 3 and 4 still apply.*

We will decompose the difference $\mathbf{T} - \widehat{\mathbf{T}}$ and apply the above concentration results. In the decomposition, we need to bound terms of the form $\mathbf{E}_{xx}^{-\frac{1}{2}} \mathbf{\Sigma}_{xx}^{\frac{1}{2}} - \mathbf{I}$. Such bounds can be derived from our assumption on $\left\| \mathbf{E}_{xx}^{-\frac{1}{2}} \mathbf{\Sigma}_{xx} \mathbf{E}_{xx}^{-\frac{1}{2}} - \mathbf{I} \right\|$ using Lemma 5 below. This lemma is derived from the main result of Mathias (1997), with extra effort taken to understand the size of perturbation for which higher order error terms can be safely ignored.

Lemma 5 (Perturbation of matrix square root) *Let* $\mathbf{H} \in \mathbb{R}^{d \times d}$ *be positive definite, with eigenvalues in the range* $[\sigma_{\min}, \sigma_{\max}]$ *for some* $\sigma_{\min} > 0$*. Let* $\mathbf{\Theta} \in \mathbb{R}^{d \times d}$ *be Hermitian, satisfying* $\left\| \mathbf{H}^{-\frac{1}{2}} \mathbf{\Theta} \mathbf{H}^{-\frac{1}{2}} \right\| = 1$*. Then for* $\zeta \leq \frac{3}{4} \sigma_{\max}^{-2} \sigma_{\min}^2$*, we have*

$$\left\| (\mathbf{H} + \zeta \cdot \mathbf{\Theta})^{\frac{1}{2}} \mathbf{H}^{-\frac{1}{2}} - \mathbf{I} \right\| \leq C_d \cdot \zeta$$

where $C_d = \mathcal{O}(\log d)$ *is independent of* ζ*.*

Lemma 6 *Assume that we draw* N *samples* $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$ *independently from the underlying joint distribution* $P(\mathbf{x}, \mathbf{y})$ *for computing the sample covariance matrices in* (7)*, and* $P(\mathbf{x}, \mathbf{y})$ *satisfies Assumption 1 and the concentration property* (9)*. Then for* $\nu \leq \frac{1}{4}\gamma^2$*, we have*

$$|\widehat{\rho}_1 - \rho_1| \leq \left\| \mathbf{T} - \widehat{\mathbf{T}} \right\| \leq 4 C_d \cdot \nu$$

where C_d *is the same constant in Lemma 5.*

We note that the requirement of $\nu = \mathcal{O}(\gamma^2)$ is not too constraining, since the size of the perturbation ν is closely related to the statistical error, and we are mainly interested in the regime of the statistical error going to zero. It is then straightforward to combine Lemma 3 and Lemma 6 to obtain the following sample complexities for the three distribution classes.

Corollary 7 (Sample complexity for learning canonical correlation by ERM) *Let* $\epsilon' \in (0,1)$ *and* $\epsilon' \leq C_d\gamma^2$. *Then for* $N \geq N_0\left(\frac{\epsilon'}{4C_d}\right)$, *i.e,*

$$\begin{aligned} N &\geq C\frac{d\log^2 d}{\epsilon'^2} && \text{for the sub-Gaussian class,} \\ N &\geq C\frac{d\log^{2(1+r^{-1})} d}{\epsilon'^{2(1+r^{-1})}} && \text{for the polynomial-tail class,} \\ N &\geq C\frac{\log^2 d}{\epsilon'^2\gamma^2} && \text{for the bounded class,} \end{aligned}$$

we have with high probability that $|\widehat{\rho}_1 - \rho_1| \leq \epsilon'$.

Remark 8 *Due to better concentration properties, the sample complexity for the sub-Gaussian and regular polynomial-tail classes are independent of the condition number* $\frac{1}{\gamma}$ *of the auto-covariances.*

Comparison to Arora et al. (2017) In a parallel work by Arora et al. (2017), the authors studied the top-k stochastic CCA for bounded inputs, and proposed stochastic approximation-type algorithms with $\tilde{\mathcal{O}}\left(\frac{1}{\epsilon'^2\gamma^2}\right)$ sample-complexity upper bound for approximating the top canonical correlation. We note, however, their stochastic algorithms are derived from the convex relaxation of stochastic CCA, which lifts the original problem into the space of matrices in $\mathbb{R}^{d_x\times d_y}$ and requires a whitening operation (multiplying each fresh sample by $\mathbf{\Sigma}_{xx}^{-\frac{1}{2}}$ or $\mathbf{\Sigma}_{yy}^{-\frac{1}{2}}$) and a projection operation (onto the set of low 2-norm and nuclear-norm matrices) in each iteration, which are inefficient in high dimensions. Our work studies three different classes of input distributions in a uniform manner[4], with the goal of matching the statistical limits for the Gaussian inputs (see Section 5.2). The algorithms we provide in the next sections require only elementary vector operations and thus more practical for high dimensional data.

3.2. Approximating the canonical directions

We now discuss the error in learning $(\mathbf{u}^*, \mathbf{v}^*)$ by ERM, when $\mathbf{T}$ has a singular value gap $\Delta > 0$. Let the nonzero singular values of $\mathbf{T}$ be $1 \geq \rho_1 \geq \rho_2 \geq \cdots \geq \rho_r$, where $r = \mathrm{rank}(\mathbf{T}) \leq \min(d_x, d_y)$, and the corresponding (unit-length) singular vector pairs be $(\mathbf{a}_1, \mathbf{b}_1), \ldots, (\mathbf{a}_r, \mathbf{b}_r)$. Define

$$\mathbf{C} = \begin{bmatrix} \mathbf{0} & \mathbf{T} \\ \mathbf{T}^\top & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{d\times d}. \tag{10}$$

The eigenvalues of $\mathbf{C}$ are

$$\rho_1 \geq \cdots \geq \rho_r > 0 = \cdots = 0 > -\rho_r \geq \cdots \geq -\rho_1,$$

4. If we only study the case of bounded inputs, we can bypass Lemma 5 and the $\log d$ dependence in our bound can be reduced.

with corresponding unit eigenvectors

$$\frac{1}{\sqrt{2}}\begin{bmatrix}\mathbf{a}_1\\ \mathbf{b}_1\end{bmatrix},\ \ldots,\ \frac{1}{\sqrt{2}}\begin{bmatrix}\mathbf{a}_r\\ \mathbf{b}_r\end{bmatrix},\ \ldots,\ \frac{1}{\sqrt{2}}\begin{bmatrix}\mathbf{a}_r\\ -\mathbf{b}_r\end{bmatrix},\ \ldots,\ \frac{1}{\sqrt{2}}\begin{bmatrix}\mathbf{a}_1\\ -\mathbf{b}_1\end{bmatrix}.$$

Thus, learning canonical directions $(\mathbf{u}^*, \mathbf{v}^*)$ reduces to learning the top eigenvector of $\mathbf{C}$.

We denote the empirical version of $\mathbf{C}$ by $\widehat{\mathbf{C}}$, and the singular vector pairs of $\widehat{\mathbf{T}}$ by $\{(\widehat{\mathbf{a}}_i, \widehat{\mathbf{b}}_i)\}$. Due to the block structure of $\mathbf{C}$ and $\widehat{\mathbf{C}}$, we have $\left\|\mathbf{C}-\widehat{\mathbf{C}}\right\| = \left\|\mathbf{T}-\widehat{\mathbf{T}}\right\|$. Let the ERM solution be $(\widehat{\mathbf{u}}, \widehat{\mathbf{v}}) = \left(\mathbf{\Sigma}_{xx}^{-\frac{1}{2}}\widehat{\mathbf{a}}_1, \mathbf{\Sigma}_{yy}^{-\frac{1}{2}}\widehat{\mathbf{b}}_1\right)$, which satisfy $\left\|\mathbf{\Sigma}_{xx}^{\frac{1}{2}}\widehat{\mathbf{u}}\right\| = \left\|\mathbf{\Sigma}_{yy}^{\frac{1}{2}}\widehat{\mathbf{v}}\right\| = 1$. We now state the sample complexity for learning the canonical directions by ERM.

Theorem 9 *Let $\epsilon \in (0,1)$ and $\epsilon \le \frac{16C_d^2\gamma^4}{\Delta^2}$. Then for $N \ge N_0\left(\frac{\sqrt{\epsilon}\Delta}{16C_d}\right)$, i.e.,*

$$N \ge C\frac{d\log^2 d}{\epsilon\Delta^2} \qquad \textit{for the sub-Gaussian class,}$$

$$N \ge C\frac{d\log^{2(1+r^{-1})} d}{\epsilon^{(1+r^{-1})}\Delta^2} \qquad \textit{for the regular polynomial-tail class,}$$

$$N \ge C\frac{\log^2 d}{\epsilon\Delta^2\gamma^2} \qquad \textit{for the bounded class,}$$

we have with high probability that $\mathrm{align}\left((\widehat{\mathbf{u}}, \widehat{\mathbf{v}}); (\mathbf{u}^*, \mathbf{v}^*)\right) \ge 1-\epsilon$.

Proof sketch The proof of Theorem 9 consists of two steps. We first bound the error between $\widehat{\mathbf{C}}$'s top eigenvector $\frac{1}{\sqrt{2}}\begin{bmatrix}\mathbf{\Sigma}_{xx}^{\frac{1}{2}}\widehat{\mathbf{u}}\\ \mathbf{\Sigma}_{yy}^{\frac{1}{2}}\widehat{\mathbf{v}}\end{bmatrix}$ and $\mathbf{C}$'s top eigenvector $\frac{1}{\sqrt{2}}\begin{bmatrix}\mathbf{E}_{xx}^{\frac{1}{2}}\mathbf{u}^*\\ \mathbf{E}_{yy}^{\frac{1}{2}}\mathbf{v}^*\end{bmatrix}$ using a standard result on perturbation of eigenvectors, namely the Davis-Kahan $\sin\theta$ theorem (Davis and Kahan, 1970) which states $\sin^2\theta \le \frac{\left\|\mathbf{C}-\widehat{\mathbf{C}}\right\|^2}{\Delta^2} \le \frac{\epsilon'^2}{\Delta^2}$ where θ is the angle between top eigenvectors of $\mathbf{C}$ and $\widehat{\mathbf{C}}$. We then show that $\frac{1}{\sqrt{2}}\begin{bmatrix}\mathbf{\Sigma}_{xx}^{\frac{1}{2}}\widehat{\mathbf{u}}\\ \mathbf{\Sigma}_{yy}^{\frac{1}{2}}\widehat{\mathbf{v}}\end{bmatrix}$ is very close to the "correctly normalized" $\frac{1}{\sqrt{2}}\begin{bmatrix}\mathbf{E}_{xx}^{\frac{1}{2}}\widehat{\mathbf{u}}/\|\mathbf{E}_{xx}^{\frac{1}{2}}\widehat{\mathbf{u}}\|\\ \mathbf{E}_{yy}^{\frac{1}{2}}\widehat{\mathbf{v}}/\|\mathbf{E}_{yy}^{\frac{1}{2}}\widehat{\mathbf{v}}\|\end{bmatrix}$, so the later still aligns well with the population solution.

Comparison to prior analysis For the sub-Gaussian class, the tightest analysis of the sample complexity upper bound we are aware of was by Gao et al. (2017). However, their proof relies on the assumption that $\rho_2 = o(\rho_1)$, i.e., they require that $\rho_2 \ll \rho_1$. In contrast, we do not require this assumption, and our bound is sharp in terms of the gap $\Delta = \rho_1 - \rho_2$. Up to the $\log^2 d$ factor, our ERM sample complexity for the same loss matches the minimax lower bound $\frac{d}{\epsilon\Delta^2}$ given by Gao et al. (2017) (see also Section 5.2).

4. Stochastic optimization for ERM

A disadvantage of the empirical risk minimization approach is that it can be time and memory consuming. To obtain the exact solution to (6), we need to explicitly form and

store the covariance matrices and to compute their singular value decompositions (SVDs); these steps have a time complexity of $\mathcal{O}(Nd^2 + d^3)$ and a memory complexity of $\mathcal{O}(d^2)$.

In this section, we study the stochastic optimization of the empirical objective, and show that the computational complexity is low: We just need to process a large enough dataset (with the same level of samples as ERM requires) nearly constant times in order to achieve small error with respect to the population objective. The basic algorithm we use here is the shift-and-invert meta-algorithm proposed by Wang et al. (2016). However, in this section we provide refined analysis of the algorithm's time complexity than that provided by Wang et al. (2016). We show that, using a better measure of progress and careful initializations for each least squares problem, the algorithm enjoys linear convergence (see Theorem 12), i.e., the time complexity for achieving η-suboptimalilty in the empirical objective depends on $\log \frac{1}{\eta}$, whereas the result of Wang et al. (2016) has a dependence of $\log^2 \frac{1}{\eta}$.

We also note that the recent work of Allen-Zhu and Li (2016) and Allen-Zhu and Li (2017) have extended the ERM problem to extracting the top $k \geq 1$ pairs of canonical directions, and applied the technique of peeling/deflation together with shift-and-invert. However, their convergence rate for the fist pair of canonical directions does not improve that of Wang et al. (2016).[5] As mentioned above, our result strictly improves that of Wang et al. (2016), and in particular replaces the $\tilde{\mathcal{O}}(\cdot)$ notation with the $\mathcal{O}(\cdot)$ notation in total runtime, achieving true linear convergence.

Roadmap for this section We first introduce the shift-and-invert power iterations and provide its iteration complexity, assuming that each matrix-vector multiplication or equivalently a convex least squares problem is solved to sufficient accuracy (Lemma 10). We then show each least squares can be warm-started using rescaled estimates from the previous iteration (Lemma 11). Finally, we plug in the time complexity of SVRG for each subproblem, and give runtime complexities for each distribution class which have different "condition numbers" (Corollary 13).

The condition numbers depend on, among other things, the smallest eigenvalues of the covariance matrices, which are bounded away from zero as discussed below.

Eigenvalues of empirical covariance According to the analysis of ERM from previous section, we have been working in the regime that the concentration parameter in (9) satisfies $\nu \leq \frac{\gamma^2}{4} \leq \frac{\gamma}{2}$. Thus in view of Assumption 1, we have with high probability that

$$\|\mathbf{\Sigma}_{xx} - \mathbf{E}_{xx}\| = \left\|\mathbf{E}_{xx}^{\frac{1}{2}}(\mathbf{E}_{xx}^{-\frac{1}{2}}\mathbf{\Sigma}_{xx}\mathbf{E}_{xx}^{-\frac{1}{2}} - \mathbf{I})\mathbf{E}_{xx}^{\frac{1}{2}}\right\| \leq \|\mathbf{E}_{xx}\| \cdot \left\|\mathbf{E}_{xx}^{-\frac{1}{2}}\mathbf{\Sigma}_{xx}\mathbf{E}_{xx}^{-\frac{1}{2}} - \mathbf{I}\right\| \leq \frac{\gamma}{2}$$

and similarly $\|\mathbf{\Sigma}_{yy} - \mathbf{E}_{yy}\| \leq \frac{\gamma}{2}$. According to Weyl's inequality, these inequalities make sure eigenvalues of $\mathbf{\Sigma}_{xx}$ and $\mathbf{\Sigma}_{yy}$ lie in $[\frac{\gamma}{2}, 1+\frac{\gamma}{2}]$, and consequently the involved subproblems are strongly-convex and can be solved efficiently.

5. See the second and third last lines of Table 1, and the last paragraph of Section 1.2 in Allen-Zhu and Li (2017): "Our running time matches that of [29] when $k = 1$".

4.1. Shift-and-invert power iterations

Our algorithm runs the shift-and-invert power iterations on the following matrix

$$\widehat{\mathbf{M}}_\lambda = \left(\lambda \mathbf{I} - \widehat{\mathbf{C}}\right)^{-1} = \begin{bmatrix} \lambda \mathbf{I} & -\widehat{\mathbf{T}} \\ -\widehat{\mathbf{T}}^\top & \lambda \mathbf{I} \end{bmatrix}^{-1} \tag{11}$$

where $\lambda > \widehat{\rho}_1$. It is straightforward to see that $\widehat{\mathbf{M}}_\lambda$ is positive definite with eigenvalues

$$\frac{1}{\lambda - \widehat{\rho}_1} \geq \cdots \geq \frac{1}{\lambda - \widehat{\rho}_r} \geq \cdots \geq \frac{1}{\lambda + \widehat{\rho}_r} \geq \cdots \geq \frac{1}{\lambda + \widehat{\rho}_1},$$

and has the same set of eigenvectors as $\widehat{\mathbf{C}}$.

Assume that there exists a singular value gap for $\widehat{\mathbf{T}}$ (this can be guaranteed by drawing sufficiently many samples so that the singular values of $\widehat{\mathbf{T}}$ are within a fraction of the gap Δ of $\mathbf{T}$), denoted as $\widehat{\Delta} = \widehat{\rho}_1 - \widehat{\rho}_2$. The key observation is that, as opposed to running power iterations on $\widehat{\mathbf{C}}$ (which is essentially done by Ge et al. 2016), $\widehat{\mathbf{M}}_\lambda$ has a large eigenvalue gap when $\lambda = \widehat{\rho}_1 + c(\widehat{\rho}_1 - \widehat{\rho}_2)$ with $c = \mathcal{O}(1)$, and thus power iterations on $\widehat{\mathbf{M}}_\lambda$ converge more quickly. In particular, we assume for now the availability of an estimated eigenvalue λ such that $\lambda - \widehat{\rho}_1 \in [l\widehat{\Delta}, u\widehat{\Delta}]$ where $0 < l < u < 1$; locating such a λ is discussed later in Remark 14.

Define

$$\widehat{\mathbf{A}}_\lambda := \begin{bmatrix} \lambda \mathbf{\Sigma}_{xx} & -\mathbf{\Sigma}_{xy} \\ -\mathbf{\Sigma}_{xy}^\top & \lambda \mathbf{\Sigma}_{yy} \end{bmatrix}, \qquad \widehat{\mathbf{B}} := \begin{bmatrix} \mathbf{\Sigma}_{xx} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{yy} \end{bmatrix},$$

and we have $\widehat{\mathbf{M}}_\lambda = \widehat{\mathbf{B}}^{\frac{1}{2}} \widehat{\mathbf{A}}_\lambda^{-1} \widehat{\mathbf{B}}^{\frac{1}{2}}$. And by the relationship $\widehat{\mathbf{A}}_\lambda = \widehat{\mathbf{B}}^{\frac{1}{2}} \widehat{\mathbf{M}}_\lambda^{-1} \widehat{\mathbf{B}}^{\frac{1}{2}}$, eigenvalues of $\widehat{\mathbf{A}}_\lambda$ are bounded:

$$\sigma_{\max}\left(\widehat{\mathbf{A}}_\lambda\right) \leq \sigma_{\max}\left(\widehat{\mathbf{M}}_\lambda^{-1}\right) \cdot \sigma_{\max}\left(\widehat{\mathbf{B}}\right) \leq (\lambda + \widehat{\rho}_1)(1 + \frac{\gamma}{2}),$$
$$\sigma_{\min}\left(\widehat{\mathbf{A}}_\lambda\right) \geq \sigma_{\min}\left(\widehat{\mathbf{M}}_\lambda^{-1}\right) \cdot \sigma_{\min}\left(\widehat{\mathbf{B}}\right) \geq (\lambda - \widehat{\rho}_1)\gamma/2.$$

It is convenient to study the convergence in the concatenated variables

$$\mathbf{w}_t := \frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{u}_t \\ \mathbf{v}_t \end{bmatrix}, \quad \mathbf{r}_t := \widehat{\mathbf{B}}^{\frac{1}{2}} \mathbf{w}_t = \frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{\Sigma}_{xx}^{\frac{1}{2}} \mathbf{u}_t \\ \mathbf{\Sigma}_{yy}^{\frac{1}{2}} \mathbf{v}_t \end{bmatrix}.$$

Define the following quantities using the ERM solution

$$\widehat{\mathbf{w}} := \frac{1}{\sqrt{2}} \begin{bmatrix} \widehat{\mathbf{u}} \\ \widehat{\mathbf{v}} \end{bmatrix}, \qquad \widehat{\mathbf{r}} := \widehat{\mathbf{B}}^{\frac{1}{2}} \widehat{\mathbf{w}} = \frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{\Sigma}_{xx}^{\frac{1}{2}} \widehat{\mathbf{u}} \\ \mathbf{\Sigma}_{yy}^{\frac{1}{2}} \widehat{\mathbf{v}} \end{bmatrix},$$

which satisfy $\widehat{\mathbf{w}}^\top \widehat{\mathbf{B}} \widehat{\mathbf{w}} = 1$ and $\widehat{\mathbf{r}}^\top \widehat{\mathbf{r}} = 1$ respectively.

4.2. Convergence of inexact shift-and-invert

Our algorithm iteratively applies the approximate matrix-vector multiplications: for $t = 0, 1, \ldots$

$$\mathbf{r}_{t+1} \approx \widehat{\mathbf{M}}_\lambda \mathbf{r}_t, \qquad \Longleftrightarrow \qquad \mathbf{w}_{t+1} \approx \widehat{\mathbf{A}}_\lambda^{-1} \widehat{\mathbf{B}} \mathbf{w}_t. \tag{12}$$

This equivalence allows us to directly work with $(\mathbf{u}_t, \mathbf{v}_t)$ and avoids computing $\mathbf{\Sigma}_{xx}^{\frac{1}{2}}$ or $\mathbf{\Sigma}_{yy}^{\frac{1}{2}}$ explicitly. Note that we do not perform normalizations of the form $\mathbf{w}_t \leftarrow \mathbf{w}_t / \left\| \widehat{\mathbf{B}}^{\frac{1}{2}} \mathbf{w}_t \right\|$ at each iteration as done by Wang et al. (2016) (Phase-I of their SI meta-algorithm); the length of each iterate is irrelevant for the purpose of optimizing the alignment between vectors and we could always perform the normalization in the end to satisfy the length constants. Exact power iterations is known to converge linearly when there exist an eigenvalue gap (Golub and van Loan, 1996).

The matrix-vector multiplication $\widehat{\mathbf{A}}_\lambda^{-1} \widehat{\mathbf{B}} \mathbf{w}_t$ is equivalent to solving the least squares problem

$$\min_{\mathbf{w}} \; f_{t+1}(\mathbf{w}) := \frac{1}{2} \mathbf{w}^\top \widehat{\mathbf{A}}_\lambda \mathbf{w} - \mathbf{w}^\top \widehat{\mathbf{B}} \mathbf{w}_t \tag{13}$$

whose unique solution is $\mathbf{w}_{t+1}^* = \widehat{\mathbf{A}}_\lambda^{-1} \widehat{\mathbf{B}} \mathbf{w}_t$ with the optimal objective $f_{t+1}^* = -\frac{1}{2} \mathbf{w}_t^\top \widehat{\mathbf{B}} \widehat{\mathbf{A}}_\lambda^{-1} \widehat{\mathbf{B}} \mathbf{w}_t$. Of course, solving the problem exactly is costly and we will apply stochastic gradient methods to it. We will show that, when the least squares problems are solved accurately enough, the iterates are of the same quality as those of the exact solutions and enjoys linear convergence.

We begin by introducing the measure of progress for the iterates. Denote the eigenvalues of $\widehat{\mathbf{M}}_\lambda$ by $\beta_1 \geq \beta_2 \geq \cdots \geq \beta_d$, with corresponding eigenvectors $\mathbf{p}_1, \ldots, \mathbf{p}_d$ forming an orthonormal basis of $\mathbb{R}^d$. Recall that $\mathbf{p}_1 = \widehat{\mathbf{r}}$, $\mathbf{p}_i^\top \widehat{\mathbf{M}}_\lambda \mathbf{p}_i = \beta_i$ for $i = 1, \ldots, d$, and $\mathbf{p}_i^\top \widehat{\mathbf{M}}_\lambda \mathbf{p}_j = 0$ for $i \neq j$.

We therefore can write each iterate as a linear combination of the eigenvectors: $\frac{\mathbf{r}_t}{\|\mathbf{r}_t\|} = \sum_{i=1}^d \xi_{ti} \mathbf{p}_i$, where $\xi_{ti} = \frac{\mathbf{r}_t^\top \mathbf{p}_i}{\|\mathbf{r}_t\|}$ for $i = 1, \ldots, d$, and $\sum_{i=1}^d \xi_{ti}^2 = 1$. The potential function we use to evaluate the progress of each iteration is

$$G(\mathbf{r}_t) = \frac{\left\| \mathbf{P}_\perp \frac{\mathbf{r}_t}{\|\mathbf{r}_t\|} \right\|_{\widehat{\mathbf{M}}_\lambda^{-1}}}{\left\| \mathbf{P}_\parallel \frac{\mathbf{r}_t}{\|\mathbf{r}_t\|} \right\|_{\widehat{\mathbf{M}}_\lambda^{-1}}} = \frac{\sqrt{\sum_{i=2}^d \xi_{ti}^2 / \beta_i}}{\sqrt{\xi_{t1}^2 / \beta_1}},$$

where $\mathbf{P}_\perp$ and $\mathbf{P}_\parallel$ denote projections onto the subspaces perpendicular and parallel to $\widehat{\mathbf{r}}$ respectively.

The same potential function was used by Garber et al. (2016) for analyzing the convergence of shift-and-invert for PCA. The potential function is invariant to the length of $\mathbf{r}_t$, and is equivalent to the criterion $|\tan\theta_t| := \frac{\sqrt{\sum_{i=2}^d \xi_{ti}^2}}{\sqrt{\xi_{t1}^2}}$ where θ_t is the angle between $\mathbf{r}_t$ and $\widehat{\mathbf{r}}$: in the following sense:

$$|\sin\theta_t| = \sqrt{\sum_{i=2}^d \xi_{ti}^2} \leq \sqrt{\frac{\beta_1}{\beta_2}} |\tan\theta_t| \leq G(\mathbf{r}_t) \leq \sqrt{\frac{\beta_1}{\beta_d}} |\tan\theta_t| .$$

The lemma below shows that under the iterative scheme (12), $\{G(\mathbf{r}_t)\}_{t=1,\dots}$ converges linearly to 0.

Lemma 10 *Let $\eta \in (0,1)$. Assume that for each approximate matrix-vector multiplication, we solve the least squares problem so accurately that the approximate solution $\mathbf{w}_{t+1}$ satisfies*

$$\epsilon_t := \frac{f_{t+1}(\mathbf{w}_{t+1}) - f^*_{t+1}}{\mathbf{w}_t^\top \widehat{\mathbf{B}} \mathbf{w}_t} \leq \min\left(\sum_{i=2}^{d} \xi_{ti}^2/\beta_i,\ \xi_{t1}^2/\beta_1\right) \cdot \frac{(\beta_1 - \beta_2)^2}{32}. \tag{14}$$

Let $T = \lceil \log_{\frac{7}{5}}\left(\frac{G(\mathbf{r}_0)}{\eta}\right)\rceil$. Then we have $|\sin\theta_t| \leq G(\mathbf{r}_t) \leq \eta$ for all $t \geq T$.

4.3. Bounding initial error for least squares

It is natural to use an initialization of the form $\alpha \mathbf{w}_t$ for minimizing $f_{t+1}(\mathbf{w})$. The following lemma provides the optimal α and the resulting initial suboptimality, see detailed analysis in Appendix D.2.

Lemma 11 (Warm start for least squares) *Initializing $\min_{\mathbf{w}} f_{t+1}(\mathbf{w})$ with $\alpha_t^* \mathbf{w}_t$ where $\alpha_t^* = \frac{\mathbf{w}_t^\top \widehat{\mathbf{B}} \mathbf{w}_t}{\mathbf{w}_t^\top \widehat{\mathbf{A}}_\lambda \mathbf{w}_t}$, it suffices to set the ratio between the initial and the final error to be $64 \cdot \max(1, G(\mathbf{r}_t))$ so that (14) is satisfied.*

This result indicates that in the converging stage ($G(\mathbf{r}_t) \leq 1$), we just need to set the ratio between the initial and the final error to the constant 64 (and set it to be the constant $64G(\mathbf{r}_0)$ before that). This will ensure that the time complexity of least squares has no dependence on the final error ϵ.

4.4. Solving the least squares by SGD

The least squares objective (13) is the sum of N functions: $f_{t+1}(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^N f_{t+1}^i(\mathbf{w})$ where

$$f_{t+1}^i(\mathbf{w}) = \frac{1}{2}\mathbf{w}^\top \begin{bmatrix} \lambda \mathbf{x}_i \mathbf{x}_i^\top & -\mathbf{x}_i \mathbf{y}_i^\top \\ -\mathbf{y}_i \mathbf{x}_i^\top & \lambda \mathbf{y}_i \mathbf{y}_i^\top \end{bmatrix} \mathbf{w} - \mathbf{w}^\top \begin{bmatrix} \mathbf{\Sigma}_{xx} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{yy} \end{bmatrix} \mathbf{w}_t. \tag{15}$$

There has been much recent progress on developping linearly convergent stochastic algorithms for solving finite-sum problems. We use SVRG (Johnson and Zhang, 2013) here due to its algorithmic simplicity and memory efficiency; in the next section, we will be using the "online" version of SVRG for stochastic CCA in the streaming setting. Note that although $f_{t+1}(\mathbf{w})$ is convex, each component f_{t+1}^i may not be convex.

We provide the time complexity of SVRG for this case (based on Garber and Hazan, 2015, Appendix B), as well as the "condition number" for the three classes of distributions in Appendix D.3 and D.4 respectively.

4.5. Total time complexity

We first provide the runtime for solving the empirical objective using the (offline) shift-and-invert CCA algorithm.

Theorem 12 *Let $\eta \in (0,1)$. Draw N samples for ERM such that $\sigma_{\min}(\mathbf{\Sigma}_{xx}) \geq \frac{\gamma}{2}$ and $\sigma_{\min}(\mathbf{\Sigma}_{yy})) \geq \frac{\gamma}{2}$. Initialize $\mathbf{w}_0 = \frac{\tilde{\mathbf{w}}_0}{\sqrt{\tilde{\mathbf{w}}_0^\top \widehat{\mathbf{B}} \tilde{\mathbf{w}}_0}}$ where entries of $\tilde{\mathbf{w}}_0 \in \mathbb{R}^d$ are randomly sampled from the standard Gaussian distribution. Then with high probability, offline shift-and-invert outputs an $(\mathbf{u}_T, \mathbf{v}_T)$ satisfying $\min\left(\frac{\mathbf{u}_T{}^\top \mathbf{\Sigma}_{xx} \widehat{\mathbf{u}}}{\|\mathbf{\Sigma}_{xx}^{\frac{1}{2}} \mathbf{u}_T\|}, \frac{\mathbf{v}_T{}^\top \mathbf{\Sigma}_{yy} \widehat{\mathbf{v}}}{\|\mathbf{\Sigma}_{yy}^{\frac{1}{2}} \mathbf{v}_T\|}\right) \geq 1 - \eta$ in total time*

$$\mathcal{O}\left(d\left(N + \frac{d^2}{\widehat{\Delta}^2\gamma^2}\right)\log\frac{d}{\widehat{\Delta}\gamma}\log\frac{d}{\widehat{\Delta}\gamma\eta}\right) \qquad \textit{for sub-Gaussian/polynomial-tail,}$$

$$\mathcal{O}\left(d\left(N + \frac{1}{\widehat{\Delta}^2\gamma^2}\right)\log\frac{d}{\widehat{\Delta}\gamma}\log\frac{d}{\widehat{\Delta}\gamma\eta}\right) \qquad \textit{for the bounded class.}$$

We have already shown in Theorem 9 that the ERM solution aligns well with the population solution. By drawing slighly more samples and requiring our algorithm to find an approximate solution that aligns well with the ERM solution, we can guarantee high alignment for the approximate solution as shown in the following corollary.

Corollary 13 *Let $\epsilon \in (0,1)$ and $\epsilon \leq \frac{64C_d^2\gamma^4}{\Delta^2}$. Draw $N = N_0\left(\frac{\sqrt{\epsilon}\Delta}{32C_d}\right)$ samples for the ERM objective, and use the initialization strategy in Theorem 12. Then with high probability, the total time for offline shift-and-invert to output $(\mathbf{u}_T, \mathbf{v}_T)$ with align$((\mathbf{u}_T, \mathbf{v}_T); (\mathbf{u}^*, \mathbf{v}^*)) \geq 1-\epsilon$ is*

$$\mathcal{O}\left(d\left(\frac{d\log^2 d}{\epsilon\Delta^2} + \frac{d^2}{\Delta^2\gamma^2}\right)\log\frac{d}{\Delta\gamma}\log\frac{d}{\Delta\gamma\epsilon}\right) \qquad \textit{for sub-Gaussian,}$$

$$\mathcal{O}\left(d\left(\frac{d\log^{2(1+r^{-1})} d}{\epsilon^{(1+r^{-1})}\Delta^2} + \frac{d^2}{\Delta^2\gamma^2}\right)\log\frac{d}{\Delta\gamma}\log\frac{d}{\Delta\gamma\epsilon}\right) \qquad \textit{for polynomial-tail,}$$

$$\mathcal{O}\left(d\left(\frac{\log^2 d}{\epsilon\Delta^2\gamma^2} + \frac{1}{\Delta^2\gamma^2}\right)\log\frac{d}{\Delta\gamma}\log\frac{d}{\Delta\gamma\epsilon}\right) \qquad \textit{for the bounded class.}$$

The ϵ-dependent term is near-linear in the ERM sample complexity $N(\epsilon, \Delta, \gamma)$ and is also the dominant term in the total runtime (when $\epsilon = o(\gamma^2)$ for the first two classes). For sub-Gaussian/regular polynomial-tail classes, we incur an undesirable d^2 dependence for the least squares problem's condition number (see more details in Appendix D.3), mainly due to weak concentration regarding the data norm (we have stronger concentration for the streaming setting discussed next). One can alleviate the issue of large condition number using accelerated SVRG (Lin et al., 2015).

Remark 14 *We have assumed so far the availability of $\lambda = \widehat{\rho}_1 + c(\widehat{\rho}_1 - \widehat{\rho}_2)$ with $c = \mathcal{O}(1)$ for shift-and-invert to work. There exists an efficient algorithm for locating such an λ, see the* ***repeat-until*** *loop of Algorithm 3 in Wang et al. (2016). This procedure computes $\mathcal{O}\left(\log\frac{1}{\Delta}\right)$ approximate matrix-vector multiplications, and its time complexity does not depend on ϵ as we only want to achieve good estimate of the top eigenvalue (and not the top eigenvector). So the cost of locating λ is not dominant in the total runtime.*

5. Streaming shift-and-invert CCA

A disadvantage of the ERM approach is that we need to store all the samples in order to go through the dataset multiple times. We now study the shift-and-invert algorithms in the streaming setting in which we draw samples from the underlying distribution $P(\mathbf{x}, \mathbf{y})$ and process them once. Clearly, the streaming approach requires only $\mathcal{O}(d)$ memory.

In this section, we assume the availability of a $\lambda = \rho_1 + c\Delta$, where $0 < c < 1$.[6] Our algorithm is the same as in the ERM case, except that we now directly work with the population covariances through fresh samples instead of their empirical estimates. With slight abuse of notation, we use $(\mathbf{A}_\lambda, \mathbf{B}, \mathbf{M}_\lambda)$ to denote the population version of $(\widehat{\mathbf{A}}_\lambda, \widehat{\mathbf{B}}, \widehat{\mathbf{M}}_\lambda)$:

$$\mathbf{A}_\lambda := \begin{bmatrix} \lambda \mathbf{E}_{xx} & -\mathbf{E}_{xy} \\ -\mathbf{E}_{xy}^\top & \lambda \mathbf{E}_{yy} \end{bmatrix}, \qquad \mathbf{B} := \begin{bmatrix} \mathbf{E}_{xx} & \mathbf{0} \\ \mathbf{0} & \mathbf{E}_{yy} \end{bmatrix}, \qquad \mathbf{M}_\lambda = \mathbf{B}^{\frac{1}{2}} \mathbf{A}_\lambda^{-1} \mathbf{B}^{\frac{1}{2}},$$

use $\{(\beta_i, \mathbf{p}_i)\}_{i=1}^d$ to denote the eigensystem of $\mathbf{M}_\lambda$, and use $(\mathbf{u}_t, \mathbf{v}_t)$ as well as

$$\mathbf{w}_t = \frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{u}_t \\ \mathbf{v}_t \end{bmatrix}, \quad \mathbf{r}_t = \mathbf{B}^{\frac{1}{2}} \mathbf{w}_t = \frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{E}_{xx}^{\frac{1}{2}} \mathbf{u}_t \\ \mathbf{E}_{yy}^{\frac{1}{2}} \mathbf{v}_t \end{bmatrix},$$

$t = 0, \ldots$ to denote the iterates of our algorithm. Also, define ξ_{ti}, θ_t and $G(\mathbf{r}_t)$ similarly as in Section 4.

Handling normalizations It is sufficient to achieve high alignment between

$$\frac{\mathbf{r}_T}{\|\mathbf{r}_T\|} = \begin{bmatrix} \mathbf{E}_{xx}^{\frac{1}{2}} \mathbf{u}_T \\ \mathbf{E}_{yy}^{\frac{1}{2}} \mathbf{v}_T \end{bmatrix} \Big/ \sqrt{\mathbf{u}_T^\top \mathbf{E}_{xx} \mathbf{u}_T + \mathbf{v}_T^\top \mathbf{E}_{yy} \mathbf{v}_T}$$

where $(\mathbf{u}, \mathbf{v})$ are normalized jointly, and $\mathbf{r}^* = \frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{E}_{xx}^{\frac{1}{2}} \mathbf{u}^* \\ \mathbf{E}_{yy}^{\frac{1}{2}} \mathbf{v}^* \end{bmatrix}$ where $(\mathbf{u}, \mathbf{v})$ are normalized separately. According to the lemma below, this would imply high alignment between $\frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{E}_{xx}^{\frac{1}{2}} \mathbf{u}_T / \|\mathbf{E}_{xx}^{\frac{1}{2}} \mathbf{u}_T\| \\ \mathbf{E}_{yy}^{\frac{1}{2}} \mathbf{v}_T / \|\mathbf{E}_{yy}^{\frac{1}{2}} \mathbf{v}_T\| \end{bmatrix}$ and $\mathbf{r}^*$ which is our final goal.

Lemma 15 (Conversion from joint alignment to separate alignment) *Let $\eta \in (0, 1)$. If the output $(\mathbf{u}_T, \mathbf{v}_T)$ of our online shift-and-invert algorithm satisfy that*

$$\frac{1}{\sqrt{2}} \cdot \frac{\mathbf{u}^{*\top} \mathbf{E}_{xx} \mathbf{u}_T + \mathbf{v}^{*\top} \mathbf{E}_{yy} \mathbf{v}_T}{\sqrt{\mathbf{u}_T^\top \mathbf{E}_{xx} \mathbf{u}_T + \mathbf{v}_T^\top \mathbf{E}_{yy} \mathbf{v}_T}} \geq 1 - \frac{\eta}{4},$$

we also have

$$\mathit{align}\left((\mathbf{u}_T, \mathbf{v}_T); (\mathbf{u}^*, \mathbf{v}^*)\right) = \frac{1}{2} \left(\frac{\mathbf{u}^{*\top} \mathbf{E}_{xx} \mathbf{u}_T}{\sqrt{\mathbf{u}_T^\top \mathbf{E}_{xx} \mathbf{u}_T}} + \frac{\mathbf{v}^{*\top} \mathbf{E}_{yy} \mathbf{v}_T}{\sqrt{\mathbf{v}_T^\top \mathbf{E}_{yy} \mathbf{v}_T}} \right) \geq 1 - \eta.$$

6. Based on the same intuition given in Remark 14, we believe that a procedure similar to that of Wang et al. (2016) also works in the streaming setting and the cost in locating λ is not dominant, although we do not have a formal analysis.

Algorithm 1 Streaming SVRG for $\min_{\mathbf{w}} f(\mathbf{w})$.

Input: Initialization $\mathbf{w}^0 = \mathbf{0}$, stepsize scaling factor $s = \frac{1}{352}$, (μ, S, σ^2) are respectively the strong convexity, streaming smoothness, and streaming variance given in Lemma 16.

for $\tau = 1, \ldots, \Gamma$ **do**

$\bar{\mathbf{z}} \leftarrow \mathbf{w}^{\tau-1}$

$m_\tau \leftarrow \lceil \frac{44^2 S}{\mu} \rceil, \quad k_\tau \leftarrow \max\left(\lceil \frac{44S}{\mu} \rceil, \lceil \frac{20\sigma^2 \cdot 2^{\tau-1}}{\beta_1 \|\mathbf{r}_t\|^2} \rceil \right)$

Draw k_τ samples $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_{k_\tau}, \mathbf{y}_{k_\tau})$ and estimate the batch gradient

$$\mathbf{g} \leftarrow \frac{1}{k_\tau} \sum\nolimits_{i=1}^{k_\tau} \nabla \phi(\bar{\mathbf{z}}; \mathbf{x}_i, \mathbf{y}_i)$$

Sample $\widetilde{m}_\tau$ uniformly at random from $\{1, \ldots, m_\tau\}$

$\mathbf{z} \leftarrow \bar{\mathbf{z}}$

for $i = 1, \ldots, \widetilde{m}_\tau$ **do**

Draw sample $(\mathbf{x}_i, \mathbf{y}_i)$

$\mathbf{z} \leftarrow \mathbf{z} - \frac{s}{S} \left(\nabla \phi(\mathbf{z}; \mathbf{x}_i, \mathbf{y}_i) - \nabla \phi(\bar{\mathbf{z}}; \mathbf{x}_i, \mathbf{y}_i) + \mathbf{g} \right)$

end for

$\mathbf{w}^\tau \leftarrow \mathbf{z}$

end for

Output: Return $\mathbf{w}^\Gamma$ as the approximate solution.

Note that Lemma 15 improves over a similar result by Wang et al. (2016, Theorem 5), which requires the joint alignment to be $\mathcal{O}(\eta^2)$-suboptimal for the separate alignment to be $\mathcal{O}(\eta)$-suboptimal.

5.1. Solving least squares by streaming SVRG

Turning to the streaming algorithm, the least squares problem at iteration $t+1$, is now a stochastic program:

$$\min_{\mathbf{w}} \; f_{t+1}(\mathbf{w}) = \frac{1}{2} \mathbf{w}^\top \mathbf{A}_\lambda \mathbf{w} - \mathbf{w}^\top \mathbf{B} \mathbf{w}_t = \mathbb{E}\left[\phi_{t+1}(\mathbf{w}; \mathbf{x}, \mathbf{y}) \right]$$

where $\phi_{t+1}(\mathbf{w}; \mathbf{x}, \mathbf{y}) := \frac{1}{2} \mathbf{w}^\top \begin{bmatrix} \lambda \mathbf{x}\mathbf{x}^\top & -\mathbf{x}\mathbf{y}^\top \\ -\mathbf{y}\mathbf{x}^\top & \lambda \mathbf{y}\mathbf{y}^\top \end{bmatrix} \mathbf{w} - \mathbf{w}^\top \begin{bmatrix} \mathbf{x}\mathbf{x}^\top & \mathbf{0} \\ \mathbf{0} & \mathbf{y}\mathbf{y}^\top \end{bmatrix} \mathbf{w}_t$, and the expectation is computed over $P(\mathbf{x}, \mathbf{y})$. The optimal solution to this stochastic program is $\mathbf{w}_{t+1}^* = \mathbf{A}_\lambda^{-1} \mathbf{B} \mathbf{w}_t$.

Due to the high sample complexity of accurately estimating $\alpha_t^* = \frac{\mathbf{w}_t^\top \mathbf{B} \mathbf{w}_t}{\mathbf{w}_t^\top \mathbf{A}_\lambda \mathbf{w}_t}$ in the streaming setting, we instead initialize each linear systems with the zero vector. With this initialization, we have

$$f_{t+1}(\mathbf{0}) - f_{t+1}^* = 0 - \left(-\frac{1}{2} \mathbf{w}_t^\top \mathbf{B} \mathbf{A}_\lambda^{-1} \mathbf{B} \mathbf{w}_t \right) = \frac{\mathbf{r}_t^\top \mathbf{M}_\lambda \mathbf{r}_t}{2} \leq \frac{\beta_1 \|\mathbf{r}_t\|^2}{2}. \tag{16}$$

We then solve the linear system with the streaming SVRG algorithm proposed by Frostig et al. (2015), as detailed in Algorithm 1. This is the same approach taken by Garber et al.

(2016) for streaming PCA, and our analysis follows the same structure. Streaming SVRG is a natural choice here since it is the "online" version of the SVRG algorithm for optimizing empirical objectives and enjoys the same algorithmic simplicity and low computational complexity. Moreover, for stochastic least squares problems, streaming SVRG is shown to have the same sample complexity as solving the ERM problem (Frostig et al., 2015), which aligns well with our goal of an overall sample efficient algorithm. With this choice, the final algorithm is very similar to the stochastic optimization algorithm in Section 4, except that fresh samples are used for each update.

To analyze the sample complexity of streaming SVRG, we first calculate the streaming smoothness and streaming variance parameters for the three classes of distributions.

Lemma 16 (Parameters of streaming SVRG) *For any $\mathbf{w}, \mathbf{w}' \in \mathbb{R}^d$, we have*

- *Strong convexity:*

$$f_{t+1}(\mathbf{w}) \geq f_{t+1}(\mathbf{w}') + \left\langle \nabla f_{t+1}(\mathbf{w}'), \mathbf{w} - \mathbf{w}' \right\rangle + \frac{\mu}{2} \left\| \mathbf{w} - \mathbf{w}' \right\|^2,$$

- *Streaming smoothness:*

$$\mathbb{E}\left[\left\| \nabla \phi_{t+1}(\mathbf{w}) - \nabla \phi_{t+1}(\mathbf{w}^*_{t+1}) \right\|^2 \right] \leq 2S \left(f_{t+1}(\mathbf{w}) - f^*_{t+1} \right),$$

- *Streaming variance:*

$$\mathbb{E}\left[\frac{1}{2} \left\| \nabla \phi(\mathbf{w}^*_{t+1}) \right\|^2_{\left(\nabla^2 f(\mathbf{w}^*_{t+1})\right)^{-1}} \right] \leq \sigma^2.$$

Here $\mu := \frac{\gamma}{\beta_1} \geq C\Delta\gamma$ for some $C > 0$, and

$$S = \mathcal{O}\left(\frac{d\beta_1}{\gamma} \right), \quad \sigma^2 = \mathcal{O}\left(d\beta_1^3 \left\| \mathbf{r}_t \right\|^2 \right)$$

for sub-Gaussian/regular polynomial-tail classes, and

$$S = \mathcal{O}\left(\frac{\beta_1}{\gamma} \right), \quad \sigma^2 = \mathcal{O}\left(\frac{\beta_1^3 \left\| \mathbf{r}_t \right\|^2}{\gamma^2} \right)$$

for the bounded class.

The proof of Lemma 16 is somewhat technical for the sub-Gaussian/regular polynomial-tail classes, which repeatedly applies the concentration properties of these two classes. But this lemma is the key for the sample complexity of our streaming algorithm to match the lower bound in the case of sub-Gaussian inputs: since we always draw fresh samples in the streaming setting, the "condition number" S/μ for these two classes depend on d only linearly (as opposed to quadratically in approximate ERM). These quantities determine the number of samples to be used in each round τ of Algorithm 1: m_τ is on the order of the condition number, and with m_τ stochastic updates, one can reduce the suboptimality by a constant factor in each round; κ_τ has to eventually increase geometrically to make sure the variance is reduced at the same pace.

Based on these quantities, we can apply the structural result of Frostig et al. (2015) and give the sampling complexity for driving the final suboptimality to η_t times the initial suboptimality in (16).

Lemma 17 (Sample complexity of streaming SVRG for least squares) *Let $\eta_t \in (0,1)$. Applying streaming-SVRG in Algorithm 1 to $\min_{\mathbf{w}} f_{t+1}(\mathbf{w})$ with initialization $\mathbf{0}$, we have*

$$\mathbb{E}\left[f_{t+1}(\mathbf{w}^\tau) - f_{t+1}^*\right] \leq \eta_t \left(\frac{\beta_1 \left\|\mathbf{r}_t\right\|^2}{2}\right)$$

for $\tau \geq \Gamma = \mathcal{O}\left(\log \frac{1}{\eta_t}\right)$. The sample complexity of the first Γ iterations is $\mathcal{O}\left(\frac{d}{\Delta^2 \eta_t} + \frac{d}{\Delta^2 \gamma^2} \log \frac{1}{\eta_t}\right)$ for the sub-Gaussian/regular polynomial-tail classes, and $\mathcal{O}\left(\frac{1}{\Delta^2 \gamma^2 \eta_t}\right)$ for the bounded class.

Based on the linear convergence of shift-and-invert, we need only solve $\mathcal{O}\left(\log \frac{1}{\epsilon}\right)$ linear systems, and we can bound $\frac{1}{\eta}$ by a geometrically increasing series where the last term is $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ (so the sum of this truncated series is still $\mathcal{O}\left(\frac{1}{\epsilon}\right)$). This results in the following total sample complexity.

Theorem 18 (Total sample complexity of streaming shift-and-invert CCA) *Let $\epsilon \in (0,1)$. After solving $T = \mathcal{O}\left(\log \frac{1}{\epsilon}\right)$ linear systems to sufficient accuracy, streaming shift-and-invert CCA algorithm outputs $(\mathbf{u}_T, \mathbf{v}_T)$ with $\mathrm{align}\left((\mathbf{u}_T, \mathbf{v}_T); (\mathbf{u}^*, \mathbf{v}^*)\right) \geq 1 - \epsilon$. Our algorithm processes each sample in $\mathcal{O}(d)$ time, and has a total sample complexity of*

$$\mathcal{O}\left(\frac{d}{\epsilon \Delta^2} + \frac{d}{\Delta^2 \gamma^2} \log^2 \frac{1}{\epsilon}\right) \qquad \text{for the sub-Gaussian/regular polynomial-tail classes,}$$

$$\mathcal{O}\left(\frac{1}{\epsilon \Delta^2 \gamma^2}\right) \qquad \text{for the bounded class.}$$

Interestingly, the sample complexity of our streaming CCA algorithm (assuming the parameter λ) improves over that of ERM we showed in Theorem 9: it removes small $\log d$ factors for all classes, and most remarkably achieves polynomial improvement in ϵ for the regular polynomial-tail class. This is due to the fact that the sample complexity of streaming SVRG basically only uses the moments, and does not require concentration of the whole covariance in Lemma 3. As a result, it is not clear if our analysis of ERM is the tightest possible.

5.2. Lower bound for Gaussian inputs

Consider the following Gaussian distribution named *single canonical pair model* (Chen et al., 2013):

$$\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{I} & \Delta \boldsymbol{\phi} \boldsymbol{\psi}^\top \\ \Delta \boldsymbol{\psi} \boldsymbol{\phi}^\top & \mathbf{I} \end{bmatrix}\right), \tag{17}$$

where $\|\boldsymbol{\phi}\| = \|\boldsymbol{\psi}\| = 1$. It is straightforward to check that $\mathbf{T} = \mathbf{E}_{xy} = \Delta \boldsymbol{\phi} \boldsymbol{\psi}^\top$ for such a distribution. Observe that $\mathbf{T}$ is of rank one and has a singular value gap Δ, and the single pair of canonical directions are $(\mathbf{u}^*, \mathbf{v}^*) = (\boldsymbol{\phi}, \boldsymbol{\psi})$. Denote this class of model by $\mathcal{F}(d_x, d_y, \Delta)$. We have the following minimax lower bound for CCA under this model, which is an application of the result of Gao et al. (2017) for sparse CCA (by using rank $r = 1$ and hard sparsity, i.e., $q = 0$ and sparsity level d in their Theorem 3.2).

Table 1: Summary of sample, time (measured in floating point operations), and memory complexities of different approaches, in terms of (d, Δ, ϵ), for stochastic CCA with Gaussian inputs. We give the dominant term in complexities as $\epsilon \to 0$. Note that the time complexity of exact ERM is dominated by forming the eigen-system, while the memory complexity of ERM is dominated by saving the dataset.

Method	Sample	Time	Memory
Exact ERM	$\tilde{\mathcal{O}}\left(\frac{d}{\epsilon\Delta^2}\right)$	$\tilde{\mathcal{O}}\left(\frac{d^3}{\epsilon\Delta^2}\right)$	$\tilde{\mathcal{O}}\left(\frac{d^2}{\epsilon\Delta^2}\right)$
Approximate ERM by shift-and-invert	$\tilde{\mathcal{O}}\left(\frac{d}{\epsilon\Delta^2}\right)$	$\tilde{\mathcal{O}}\left(\frac{d^2}{\epsilon\Delta^2}\right)$	$\tilde{\mathcal{O}}\left(\frac{d^2}{\epsilon\Delta^2}\right)$
Streaming shift-and-invert (assuming close init. for λ)	$\mathcal{O}\left(\frac{d}{\epsilon\Delta^2}\right)$	$\mathcal{O}\left(\frac{d^2}{\epsilon\Delta^2}\right)$	$\mathcal{O}(d)$

Lemma 19 (Lower bound for single canonical pair model) *Suppose the data is generated by the single canonical pair model. Let* $(\mathbf{u}, \mathbf{v})$ *be some estimate of the canonical directions* $(\mathbf{u}^*, \mathbf{v}^*)$ *based on* N *samples. Then, there is a universal constant* C*, so that for* N *sufficiently large, we have:*

$$\inf_{\mathbf{u},\mathbf{v}} \sup_{\substack{\mathbf{u}^*,\mathbf{v}^* \in \\ \mathcal{F}(d_x,d_y,\Delta)}} \mathbb{E}\left[1 - \mathit{align}\left((\mathbf{u}_T, \mathbf{v}_T); (\mathbf{u}^*, \mathbf{v}^*)\right)\right] \geq C\frac{d}{\Delta^2 N}.$$

This lemma implies that, to estimate the canonical directions up to ϵ-suboptimality in our measure of alignment, we expect to use at least $\mathcal{O}\left(\frac{d}{\epsilon\Delta^2}\right)$ samples. We therefore observe that, for Gaussian inputs, the sample complexity of the our streaming algorithm matches that of the minimax rate of CCA, up to small factors.

In Table 1, we collect the complexities of different approaches, namely exact optimization of ERM (Section 3), stochastic optimization of ERM with shift-and-invert (Section 4), and streaming shift-and-invert (Section 5). We observe that while all three approaches are sample efficient (up to small factors), stochastic and streaming algorithms are more efficient in time and memory.

6. Conclusion

In this paper, we have studied the sample complexity of population CCA for several classes of input distributions, and proposed sample-efficient algorithms for learning the first pair of canonical directions. While the original problem is nonconvex, we exploit its structure as an eigenvalue problem, and analyze the statistical performance of the shift-and-invert power iterations.

Based on the deflation/peeling scheme (Allen-Zhu and Li, 2016, 2017) for eigenvalue problems, our results shall be extended to extracting the top-k canonical direction pairs. Our algorithms also apply to related eigenvalue problems in machine learning, such as partial least squares (Chen et al., 2017) and linear discriminant analysis (Bach and Jordan, 2005), which are special versions of CCA with the population covariances being identity

(i.e., $\mathbf{E}_{xx} = \mathbf{E}_{yy} = \mathbf{I}$) and $\mathbf{y}$ being one-hot representations for class labels respectively. It is an interesting question if our general approach can be adapted to study the statistical performance of the kernel extension of CCA (Fukumizu et al., 2007).

Acknowledgement

Research partially supported by NSF BIGDATA award 1546462.

References

Zeyuan Allen-Zhu and Yuanzhi Li. LazySVD: Even faster SVD decomposition yet without agonizing pain. In *Advances in Neural Information Processing Systems*, 2016.

Zeyuan Allen-Zhu and Yuanzhi Li. Doubly accelerated methods for faster CCA and generalized eigendecomposition. In *Proc. of the International Conference on Machine Learning*, 2017.

Raman Arora, Andy Cotter, Karen Livescu, and Nati Srebro. Stochastic optimization for PCA and PLS. In *50th Annual Allerton Conference on Communication, Control, and Computing*, 2012.

Raman Arora, Teodor V. Marinov, Poorya Mianjy, and Nathan Srebro. Stochastic approximation for canonical correlation analysis. In *Advances in Neural Information Processing Systems*, 2017.

Sanjeev Arora, Satish Rao, and Umesh Vazirani. Expander flows, geometric embeddings and graph partitioning. *Journal of the ACM*, (2), 2009.

Francis R. Bach and Michael I. Jordan. A probabilistic interpretation of canonical correlation analysis. Technical Report 688, Department of Statistics, University of California, Berkeley, April 21 2005.

Marcus Carlsson. Perturbation theory for the matrix square root and matrix modulus. arXiv:1810.01464 [math.FA], October 2 2018.

Mengjie Chen, Chao Gao, Zhao Ren, and Harrison H. Zhou. Sparse CCA via precision adjusted iterative thresholding. arXiv:1311.6186 [math.ST], November 24 2013.

Zhehui Chen, Lin F. Yang, Chris J. Li, and Tuo Zhao. Dropping convexity for more efficient and scalable online multiview learning. In *Proc. of the International Conference on Machine Learning*, 2017.

Chandler Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation III. *SIAM Journal of Numerical Analysis*, 7(1):1–46, 1970.

Roy Frostig, Rong Ge, Sham M. Kakade, and Aaron Sidford. Competing with the empirical risk minimizer in a single pass. In *Conference on Learning Theory (COLT)*, pages 728–763, 2015.

Kenji Fukumizu, Francis R. Bach, and Arthur Gretton. Statistical consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research*, 8:361–383, February 2007.

Chao Gao, Zongming Ma, and Harrison H. Zhou. Sparse CCA: Adaptive estimation and computational barriers. *Annals of Statistics*, 45(5):2074–2101, 2017.

Dan Garber and Elad Hazan. Fast and simple PCA via convex optimization. arXiv:1509.05647 [math.OC], November25 2015.

Dan Garber, Elad Hazan, Chi Jin, Sham M. Kakade, Cameron Musco, Praneeth Netrapalli, and Aaron Sidford. Faster eigenvector computation via shift-and-invert preconditioning. In *Proc. of the International Conference on Machine Learning*, 2016.

Rong Ge, Chi Jin, Sham M. Kakade, Praneeth Netrapalli, and Aaron Sidford. Efficient algorithms for large-scale generalized eigenvector computation and canonical correlation analysis. In *Proc. of the International Conference on Machine Learning*, 2016.

Eckart Gekeler. On the pointwise matrix product and the mean value theorem. *Linear Algebra and its Applications*, 35:183–191, 1981.

Gene H. Golub and Charles F. van Loan. *Matrix Computations*. Johns Hopkins University Press, 1996.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 1986.

Roger A. Horn and Charles R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1991.

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

Daniel Hsu, Sham Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electron. Commun. Probab.*, 17(52):1–6, 2012.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, 2013.

Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. A universal catalyst for first-order optimization. In *Advances in Neural Information Processing Systems*, 2015.

Yichao Lu and Dean P. Foster. Large scale canonical correlation analysis with iterative least squares. In *Advances in Neural Information Processing Systems*, 2014.

Zhuang Ma, Yichao Lu, and Dean Foster. Finding linear structure in large datasets with scalable canonical correlation analysis. In *Proc. of the International Conference on Machine Learning*, 2015.

Roy Mathias. A bound for the matrix square root with application to eigenvector perturbation. *SIAM J. Matrix Anal. and Apps.*, 18(4):861–867, 1997.

Yousef Saad. *Numerical Methods for Large Eigenvalue Problems*. Manchester University Press, 1992.

Nikhil Srivastava and Roman Vershynin. Covariance estimation for distributions with $2+\varepsilon$-moments. *Annals of Probability*, 41(5):3081–3111, 2013.

Roman Vershynin. *Compressed Sensing: Theory and Applications*, chapter Introduction to the Non-asymptotic Analysis of Random Matrices. Cambridge University Press, 2012.

Weiran Wang and Karen Livescu. Large-scale approximate kernel canonical correlation analysis. In *Proc. of the International Conference on Learning Representations*, 2016.

Weiran Wang, Jialei Wang, Dan Garber, and Nathan Srebro. Globally convergent stochastic optimization for canonical correlation analysis. In *Advances in Neural Information Processing Systems*, 2016.

Lin Xiao and Tong Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

Appendix A. Auxiliary Lemmas

Lemma 20 *The population canonical correlation is bounded by* 1*, i.e.,*

$$\rho_1 = \sigma_1(\mathbf{T}) \leq 1.$$

Proof By the Cauchy-Schwarz inequality of random variables, we have

$$\rho_1 = \mathbb{E}[({\mathbf{u}^*}^\top \mathbf{x})({\mathbf{v}^*}^\top \mathbf{y})] \leq \sqrt{\mathbb{E}[({\mathbf{u}^*}^\top \mathbf{x})^2]} \cdot \sqrt{\mathbb{E}[({\mathbf{v}^*}^\top \mathbf{y})^2]} = \sqrt{{\mathbf{u}^*}^\top \mathbf{E}_{xx} \mathbf{u}} \cdot \sqrt{{\mathbf{v}^*}^\top \mathbf{E}_{yy} \mathbf{v}} = 1.$$

■

Lemma 21 (Distance between normalized vectors) *For two nonzero vectors* $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$*, we have*

$$\left\| \frac{\mathbf{a}}{\|\mathbf{a}\|} - \frac{\mathbf{b}}{\|\mathbf{b}\|} \right\| \leq 2 \frac{\|\mathbf{a} - \mathbf{b}\|}{\|\mathbf{a}\|}.$$

Proof By direct calculation, we have

$$\begin{aligned}
\left\| \frac{\mathbf{a}}{\|\mathbf{a}\|} - \frac{\mathbf{b}}{\|\mathbf{b}\|} \right\| &\leq \left\| \frac{\mathbf{a}}{\|\mathbf{a}\|} - \frac{\mathbf{b}}{\|\mathbf{a}\|} \right\| + \left\| \frac{\mathbf{b}}{\|\mathbf{a}\|} - \frac{\mathbf{b}}{\|\mathbf{b}\|} \right\| \\
&= \frac{\|\mathbf{a} - \mathbf{b}\|}{\|\mathbf{a}\|} + \|\mathbf{b}\| \cdot \frac{|\|\mathbf{a}\| - \|\mathbf{b}\||}{\|\mathbf{a}\| \, \|\mathbf{b}\|} \\
&\leq \frac{\|\mathbf{a} - \mathbf{b}\|}{\|\mathbf{a}\|} + \frac{\|\mathbf{a} - \mathbf{b}\|}{\|\mathbf{a}\|} \\
&= 2 \frac{\|\mathbf{a} - \mathbf{b}\|}{\|\mathbf{a}\|}
\end{aligned}$$

where we have used the triangle inequality in the two inequalities. ■

Lemma 22 (Conversion from joint alignment to separate alignment) *Let $\eta \in \left(0, \frac{1}{4}\right)$. Consider the four nonzero vectors $\mathbf{a}, \mathbf{x} \in \mathbb{R}^{d_x}$ and $\mathbf{b}, \mathbf{y} \in \mathbb{R}^{d_y}$ such that $\|\mathbf{a}\| = \|\mathbf{b}\| = 1$. If*

$$\frac{1}{\sqrt{2}} \cdot \frac{\mathbf{a}^\top \mathbf{x} + \mathbf{b}^\top \mathbf{y}}{\sqrt{\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2}} \geq 1 - \eta, \tag{18}$$

we also have

$$\frac{1}{2}\left(\left|\frac{\mathbf{a}^\top \mathbf{x}}{\|\mathbf{x}\|}\right| + \left|\frac{\mathbf{b}^\top \mathbf{y}}{\|\mathbf{y}\|}\right|\right) \geq 1 - 4\eta.$$

Proof By the Cauchy-Schwarz inequality, we have

$$\begin{aligned}\frac{\mathbf{a}^\top \mathbf{x} + \mathbf{b}^\top \mathbf{y}}{\sqrt{\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2}} &= \frac{\mathbf{a}^\top \mathbf{x}}{\|\mathbf{x}\|} \cdot \frac{\|\mathbf{x}\|}{\sqrt{\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2}} + \frac{\mathbf{b}^\top \mathbf{y}}{\|\mathbf{y}\|} \cdot \frac{\|\mathbf{y}\|}{\sqrt{\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2}} \\ &\leq \sqrt{\left(\frac{\mathbf{a}^\top \mathbf{x}}{\|\mathbf{x}\|}\right)^2 + \left(\frac{\mathbf{b}^\top \mathbf{y}}{\|\mathbf{y}\|}\right)^2}.\end{aligned}$$

Thus according to (18), we obtain

$$\left(\frac{\mathbf{a}^\top \mathbf{x}}{\|\mathbf{x}\|}\right)^2 + \left(\frac{\mathbf{b}^\top \mathbf{y}}{\|\mathbf{y}\|}\right)^2 \geq 2(1-\eta)^2 \geq 2 - 4\eta.$$

Since $\left(\frac{\mathbf{b}^\top \mathbf{y}}{\|\mathbf{y}\|}\right)^2 \leq 1$, this implies

$$\left|\frac{\mathbf{a}^\top \mathbf{x}}{\|\mathbf{x}\|}\right| \geq \sqrt{1 - 4\eta} \geq 1 - 4\eta$$

where the last step is due to the fact that $\sqrt{x} \geq x$ for $x \in (0, 1)$. Similarly we have $\left|\frac{\mathbf{b}^\top \mathbf{y}}{\|\mathbf{y}\|}\right| \geq 1 - 4\eta$. Then the theorem follows. ■

Lemma 23 (Moment inequalities of sub-Gaussian and regular polynomial-tail random vectors) *Let $\mathbf{z} \in \mathbb{R}^d$ be isotropic and sub-Gaussian or regular polynomial-tail (see their definitions in Lemma 3). Then for some constant $C' > 0$, we have*

$$\mathbb{E}\,\|\mathbf{z}\|^2 \leq d, \qquad \mathbb{E}\,\|\mathbf{z}\|^4 \leq C' d^2, \qquad \mathbb{E}\left|\mathbf{q}^\top \mathbf{z}\right|^4 \leq C'$$

where $\mathbf{q}$ is any unit vector.

Proof Sub-Gaussian case The first bound is by $\mathbb{E}\left\|\mathbf{z}\right\|^2 = \mathbb{E}\operatorname{tr}\left(\mathbf{z}\mathbf{z}^\top\right) = \operatorname{tr}\left(I\right) = d$. To prove the second one, note that according to Theorem 2.1 in Hsu et al. (2012), we have

$$\mathbb{P}\left(\left\|\mathbf{z}\right\|^2 > C_1(d+t)\right) < e^{-t}$$

for all $t > 0$. Therefore

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{z}\right\|^4 &= \int_0^\infty \mathbb{P}\left(\left\|\mathbf{z}\right\|^4 > s\right) ds \\
&= \int_0^{C_1^2 d^2} \mathbb{P}\left(\left\|\mathbf{z}\right\|^4 > s\right) ds + \int_{C_1^2 d^2}^\infty \mathbb{P}\left(\left\|\mathbf{z}\right\|^4 > s\right) ds \\
&\leq C_1^2 d^2 + \int_{C_1^2 d^2}^\infty \exp\left(-\left(\frac{\sqrt{s}}{C_1} - d\right)\right) ds \\
&\leq C' d^2.
\end{aligned}$$

Lastly,

$$\begin{aligned}
\mathbb{E}\left|\mathbf{q}^\top \mathbf{z}\right|^4 &= \int_0^\infty \mathbb{P}\left(\left|\mathbf{q}^\top \mathbf{z}\right|^4 > s\right) ds \\
&\leq \int_0^\infty e^{-C\sqrt{s}} ds \\
&\leq C'.
\end{aligned}$$

Regular polynomial-tail case The first bound is still by $\mathbb{E}\left\|\mathbf{z}\right\|^2 = \mathbb{E}\operatorname{tr}\left(\mathbf{z}\mathbf{z}^\top\right) = \operatorname{tr}\left(I\right) = d$. When $r > 1$, we have

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{z}\right\|^4 &= \int_0^\infty \mathbb{P}\left(\left\|\mathbf{z}\right\|^4 > s\right) ds \\
&\leq \int_0^{C^2 d^2} \mathbb{P}\left(\left\|\mathbf{z}\right\|^4 > s\right) ds + \int_{C^2 d^2}^\infty \mathbb{P}\left(\left\|\mathbf{z}\right\|^4 > s\right) ds \\
&\leq C^2 d^2 + \int_{C^2 d^2}^\infty C s^{-\frac{1+r}{2}} ds \\
&\leq C' d^2.
\end{aligned}$$

To prove the last bound, take $\mathbf{V} = \mathbf{q}\mathbf{q}^\top$ in the definition of regular polynomial-tail random vectors, and then

$$\mathbb{P}\left(\left|\mathbf{q}^\top \mathbf{z}\right|^2 > t\right) \leq C t^{-1-r},$$

for any $t > C$. We have

$$\begin{aligned}
\mathbb{E}\left|\mathbf{q}^\top \mathbf{z}\right|^4 &= \int_0^\infty \mathbb{P}\left(\left|\mathbf{q}^\top \mathbf{z}\right|^4 > s\right) ds \\
&\leq \int_0^{C^2} \mathbb{P}\left(\left|\mathbf{q}^\top \mathbf{z}\right|^4 > s\right) ds + \int_{C^2}^\infty \mathbb{P}\left(\left|\mathbf{q}^\top \mathbf{z}\right|^4 > s\right) ds \\
&\leq C^2 + \int_{C^2}^\infty C s^{-\frac{1+r}{2}} ds \\
&\leq C'.
\end{aligned}$$

■

Appendix B. Proofs for Section 1

B.1. Proof of Lemma 1

Proof Using the fact that $\frac{\mathbf{u}^\top \mathbf{E}_{xx}\mathbf{u}^*}{\left\|\mathbf{E}_{xx}^{\frac{1}{2}}\mathbf{u}\right\|}$ and $\frac{\mathbf{v}^\top \mathbf{E}_{yy}\mathbf{v}^*}{\left\|\mathbf{E}_{yy}^{\frac{1}{2}}\mathbf{v}\right\|}$ are at most 1, the condition on alignment implies

$$\frac{\mathbf{u}^\top \mathbf{E}_{xx}\mathbf{u}^*}{\left\|\mathbf{E}_{xx}^{\frac{1}{2}}\mathbf{u}\right\|} = \mathbf{a}_1^\top \frac{\mathbf{E}_{xx}^{\frac{1}{2}}\mathbf{u}}{\left\|\mathbf{E}_{xx}^{\frac{1}{2}}\mathbf{u}\right\|} \geq 1 - \frac{\eta}{4}, \qquad \frac{\mathbf{v}^\top \mathbf{E}_{yy}\mathbf{v}^*}{\left\|\mathbf{E}_{yy}^{\frac{1}{2}}\mathbf{v}\right\|} = \mathbf{b}_1^\top \frac{\mathbf{E}_{yy}^{\frac{1}{2}}\mathbf{v}}{\left\|\mathbf{E}_{yy}^{\frac{1}{2}}\mathbf{v}\right\|} \geq 1 - \frac{\eta}{4}.$$

Since $\{\mathbf{a}_i\}_{i=1}^r$ and $\{\mathbf{b}_i\}_{i=1}^r$ are orthonormal, we have

$$\sum_{i=2}^{r} \left(\mathbf{a}_i^\top \frac{\mathbf{E}_{xx}^{\frac{1}{2}}\mathbf{u}}{\left\|\mathbf{E}_{xx}^{\frac{1}{2}}\mathbf{u}\right\|} \right)^2 \leq 1 - \left(1 - \frac{\eta}{4}\right)^2 \leq \frac{\eta}{2},$$

$$\sum_{i=2}^{r} \left(\mathbf{b}_i^\top \frac{\mathbf{E}_{yy}^{\frac{1}{2}}\mathbf{v}}{\left\|\mathbf{E}_{yy}^{\frac{1}{2}}\mathbf{v}\right\|} \right)^2 \leq 1 - \left(1 - \frac{\eta}{4}\right)^2 \leq \frac{\eta}{2}.$$

Observe that

$$\begin{aligned}
\frac{\mathbf{u}^\top \mathbf{E}_{xy}\mathbf{v}}{\sqrt{\mathbf{u}^\top \mathbf{E}_{xx}\mathbf{u}}\sqrt{\mathbf{v}^\top \mathbf{E}_{yy}\mathbf{v}}} &= \frac{(\mathbf{E}_{xx}^{\frac{1}{2}}\mathbf{u})^\top \mathbf{T} (\mathbf{E}_{yy}^{\frac{1}{2}}\mathbf{v})}{\left\|\mathbf{E}_{xx}^{\frac{1}{2}}\mathbf{u}\right\| \left\|\mathbf{E}_{yy}^{\frac{1}{2}}\mathbf{v}\right\|} = \sum_{i=1}^{d} \rho_i \left(\mathbf{a}_i^\top \frac{\mathbf{E}_{xx}^{\frac{1}{2}}\mathbf{u}}{\left\|\mathbf{E}_{xx}^{\frac{1}{2}}\mathbf{u}\right\|} \right) \left(\mathbf{b}_i^\top \frac{\mathbf{E}_{yy}^{\frac{1}{2}}\mathbf{v}}{\left\|\mathbf{E}_{yy}^{\frac{1}{2}}\mathbf{v}\right\|} \right) \\
&\geq \rho_1 \left(\mathbf{a}_i^\top \frac{\mathbf{E}_{xx}^{\frac{1}{2}}\mathbf{u}}{\left\|\mathbf{E}_{xx}^{\frac{1}{2}}\mathbf{u}\right\|} \right) \left(\mathbf{b}_1^\top \frac{\mathbf{E}_{yy}^{\frac{1}{2}}\mathbf{v}}{\left\|\mathbf{E}_{yy}^{\frac{1}{2}}\mathbf{v}\right\|} \right) - \rho_2 \sqrt{\sum_{i=2}^{r} \left(\mathbf{a}_i^\top \frac{\mathbf{E}_{xx}^{\frac{1}{2}}\mathbf{u}}{\left\|\mathbf{E}_{xx}^{\frac{1}{2}}\mathbf{u}\right\|} \right)^2} \sqrt{\sum_{i=2}^{r} \left(\mathbf{b}_i^\top \frac{\mathbf{E}_{yy}^{\frac{1}{2}}\mathbf{v}}{\left\|\mathbf{E}_{yy}^{\frac{1}{2}}\mathbf{v}\right\|} \right)^2} \\
&\geq \rho_1 \left(1 - \frac{\eta}{4}\right)^2 - \rho_1 \cdot \frac{\eta}{2} \geq \rho_1 (1 - \eta)
\end{aligned}$$

where we have used the Cauchy-Schwarz inequality in the first inequality. ■

Appendix C. Proofs for Section 3

C.1. Proof of Lemma 3

Proof Sub-Gaussian/regular polynomial-tail cases Consider the random variable $\mathbf{z}$ defined in (5), and draw i.i.d. samples $\mathbf{z}_1, \ldots, \mathbf{z}_n$ of $\mathbf{z}$. It is known that when the sample

size n is large enough (as specified in the lemma), we have

$$\left\| \frac{1}{N} \sum_{i=1}^{N} \mathbf{z}_i \mathbf{z}_i^\top - \mathbf{I} \right\| \leq \frac{\nu}{2}$$

with high probability for the sub-Gaussian class (Vershynin, 2012) and for the regular polynomial-tail class (Srivastava and Vershynin, 2013), given $N > C' \frac{d}{\nu^2}$ and $N \geq C' \frac{d}{\nu^{2(1+r^{-1})}}$ respectively.

We then turn to bounding the error in each covariance matrix. We note that the covariance of $\mathbf{f} := \begin{bmatrix} \mathbf{E}_{xx}^{-\frac{1}{2}} \mathbf{x} \\ \mathbf{E}_{yy}^{-\frac{1}{2}} \mathbf{y} \end{bmatrix}$ is $\mathbf{\Xi} = \begin{bmatrix} \mathbf{I} & \mathbf{T} \\ \mathbf{T}^\top & \mathbf{I} \end{bmatrix}$ with $\|\mathbf{\Xi}\| = 1 + \rho_1 \leq 2$ (since the eigenvalues of $\mathbf{\Xi}$ are of the form $1 \pm \sigma_i(\mathbf{T})$). On the other hand, define

$$\mathbf{\Phi} := \begin{bmatrix} \mathbf{E}_{xx} & \\ & \mathbf{E}_{yy} \end{bmatrix}^{-\frac{1}{2}} \begin{bmatrix} \mathbf{E}_{xx} & \mathbf{E}_{xy} \\ \mathbf{E}_{xy}^\top & \mathbf{E}_{yy} \end{bmatrix}^{\frac{1}{2}} \qquad \text{satisfying} \quad \mathbf{\Phi}\mathbf{\Phi}^\top = \mathbf{\Xi}$$

and we have $\mathbf{f} = \mathbf{\Phi}\mathbf{z}$ by the definition of $\mathbf{z}$. Furthermore, $\mathbf{f}_i = \mathbf{\Phi}\mathbf{z}_i$, $i = 1, \ldots, N$ are i.i.d. samples of $\mathbf{f}$. Therefore, it holds that

$$\left\| \begin{bmatrix} \frac{1}{N}\sum_{i=1}^N \mathbf{E}_{xx}^{-\frac{1}{2}} \mathbf{x}_i \mathbf{x}_i^\top \mathbf{E}_{xx}^{-\frac{1}{2}} - \mathbf{I} & \frac{1}{N}\sum_{i=1}^N \mathbf{E}_{xx}^{-\frac{1}{2}} \mathbf{x}_i \mathbf{y}_i^\top \mathbf{E}_{yy}^{-\frac{1}{2}} - \mathbf{T} \\ \frac{1}{N}\sum_{i=1}^N \mathbf{E}_{yy}^{-\frac{1}{2}} \mathbf{y}_i \mathbf{x}_i^\top \mathbf{E}_{xx}^{-\frac{1}{2}} - \mathbf{T}^\top & \frac{1}{N}\sum_{i=1}^N \mathbf{E}_{yy}^{-\frac{1}{2}} \mathbf{y}_i \mathbf{y}_i^\top \mathbf{E}_{yy}^{-\frac{1}{2}} - \mathbf{I} \end{bmatrix} \right\| = \left\| \frac{1}{N} \sum_{i=1}^{N} \mathbf{f}_i \mathbf{f}_i^\top - \mathbf{\Xi} \right\|$$

$$= \left\| \mathbf{\Phi} \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{z}_i \mathbf{z}_i^\top - \mathbf{I} \right) \mathbf{\Phi}^\top \right\| \leq \left\| \mathbf{\Phi}\mathbf{\Phi}^\top \right\| \cdot \left\| \frac{1}{N} \sum_{i=1}^{N} \mathbf{z}_i \mathbf{z}_i^\top - \mathbf{I} \right\| = \|\mathbf{\Xi}\| \cdot \left\| \frac{1}{N} \sum_{i=1}^{N} \mathbf{z}_i \mathbf{z}_i^\top - \mathbf{I} \right\| \leq \nu.$$

Since the norm of each block is bounded by the norm of the entire matrix, we conclude that the error in estimating each covariance matrix is bounded by ν, as required by Proposition 2.

Remark 24 *In view of Lemma 23 and the proof technique here, for the sub-Gaussian/regular polynomial-tail cases, the bound of $\|\mathbf{z}\|^2$ leads to a bound for $\|\mathbf{x}\|^2$ and $\|\mathbf{y}\|^2$:*

$$\mathbb{E}(\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2) \leq \|\mathbf{E}_{xx}\| \cdot \mathbb{E} \left\| \mathbf{E}_{xx}^{-\frac{1}{2}} \mathbf{x} \right\|^2 + \|\mathbf{E}_{yy}\| \cdot \mathbb{E} \left\| \mathbf{E}_{yy}^{-\frac{1}{2}} \mathbf{y} \right\|^2 \leq \mathbb{E}\,\|\mathbf{f}\|^2 \leq 2\mathbb{E}\,\|\mathbf{z}\|^2 \leq Cd$$

for some constant $C > 0$, where we have used Assumption 1 in the second inequality. And similarly, we have

$$\mathbb{E}(\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2)^2 \leq \mathbb{E}\,\|\mathbf{f}\|^4 \leq 4\mathbb{E}\,\|\mathbf{z}\|^4 \leq C' d^2$$

for some constant $C' > 0$.

Bounded case Consider the joint covariance matrix

$$\begin{bmatrix} \mathbf{E}_{xx} & \mathbf{E}_{xy} \\ \mathbf{E}_{xy}^\top & \mathbf{E}_{yy} \end{bmatrix} \in \mathbb{R}^{d \times d}$$

which has eigenvalue bounded by 2 due to the assumption that $\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 \leq 2$. Applying Vershynin (2012, Corollary 5.52), we obtain that

$$\left\| \begin{bmatrix} \mathbf{\Sigma}_{xx} & \mathbf{\Sigma}_{xy} \\ \mathbf{\Sigma}_{xy}^\top & \mathbf{\Sigma}_{yy} \end{bmatrix} - \begin{bmatrix} \mathbf{E}_{xx} & \mathbf{E}_{xy} \\ \mathbf{E}_{xy}^\top & \mathbf{E}_{yy} \end{bmatrix} \right\| \leq \nu' \tag{19}$$

with probability at least $1 - d^{-t^2}$ when $N \geq C(t/\nu')^2 \log d$ for some constant $C > 0$. Setting the failure probability $\delta = d^{-t^2}$ gives $t^2 = \frac{\log \frac{1}{\delta}}{\log d}$, and thus we require $N \geq C\frac{1}{\nu'^2} \log \frac{1}{\delta}$ for $1 - \delta$ success probability.

Due to the block structure of the joint covariance matrix, (19) implies

$$\|\mathbf{\Sigma}_{xy} - \mathbf{E}_{xy}\| \leq \nu', \qquad \|\mathbf{\Sigma}_{xx} - \mathbf{E}_{xx}\| \leq \nu', \qquad \|\mathbf{\Sigma}_{yy} - \mathbf{E}_{yy}\| \leq \nu'$$

hold simultaneously.

Now, to satisfy the first inequality of (9), observe that

$$\begin{aligned} \left\| \mathbf{E}_{xx}^{-\frac{1}{2}} \mathbf{\Sigma}_{xx} \mathbf{E}_{xx}^{-\frac{1}{2}} - \mathbf{I} \right\| &= \left\| \mathbf{E}_{xx}^{-\frac{1}{2}} (\mathbf{\Sigma}_{xx} - \mathbf{E}_{xx}) \mathbf{E}_{xx}^{-\frac{1}{2}} \right\| \\ &\leq \left\| \mathbf{E}_{xx}^{-\frac{1}{2}} \right\| \cdot \|\mathbf{\Sigma}_{xx} - \mathbf{E}_{xx}\| \cdot \left\| \mathbf{E}_{xx}^{-\frac{1}{2}} \right\| \\ &\leq \|\mathbf{\Sigma}_{xx} - \mathbf{E}_{xx}\| / \gamma \end{aligned}$$

where we have used the assumption that $\sigma_{\min}(\mathbf{E}_{xx}) \geq \gamma$ in the last inequality. Therefore, we obtain $\left\| \mathbf{E}_{xx}^{-\frac{1}{2}} \mathbf{\Sigma}_{xx} \mathbf{E}_{xx}^{-\frac{1}{2}} - \mathbf{I} \right\| \leq \nu$ by setting $\nu' = \gamma\nu$ in (19), and this yields the $N_0(\nu)$ chosen in the lemma. The other two inequalities of (9) can be obtained analogously. ■

C.2. Proof of Lemma 5

Proof This result can be derived from the main result of Mathias (1997) with a bit of detective work, which is needed to understand the higher order error term. As in Mathias (1997), assume without loss of generality that $\mathbf{H} = \operatorname{diag}(\lambda_1, \ldots, \lambda_d)$ is diagonal. In the proof of Theorem 2, Mathias (1997) applied the Daleckii-Krein formula in his equation (4) with the $\mathcal{O}$-notation, which can be rephrased as (see also Carlsson (2018)[Theorem 2.1]): for ζ in a neighborhood of 0, it holds

$$\varepsilon := \left\| (\mathbf{H} + \zeta\mathbf{\Theta})^{\frac{1}{2}} - \mathbf{H}^{\frac{1}{2}} - \zeta \left[\frac{\lambda_i^{\frac{1}{2}} \lambda_j^{\frac{1}{2}}}{\lambda_i^{\frac{1}{2}} + \lambda_j^{\frac{1}{2}}} \right]_{i,j=1}^{d} \circ (\mathbf{H}^{-\frac{1}{2}} \mathbf{\Theta} \mathbf{H}^{-\frac{1}{2}}) \right\| = \mathcal{O}(\zeta^2)$$

where $\circ$ denotes elementwise (Hadamard) multiplication.

To locate the neighborhood of ζ for which the above is true, we apply the matrix mean value theorem (Gekeler, 1981) with the second order derivative of matrix square root (Horn

and Johnson, 1991, Theorem 6.6.30 and page 549) to obtain:

$$\varepsilon \leq \frac{\zeta^2}{2} \max_{\tau\in[0,\zeta]} \left\| \mathbf{U}(\tau) \Bigg(\sum_{k=1}^{d} \left[\frac{1}{\sqrt{\lambda_i(\tau)} + \sqrt{\lambda_j(\tau)}} \cdot \frac{1}{\sqrt{\lambda_i(\tau)} + \sqrt{\lambda_k(\tau)}} \cdot \frac{1}{\sqrt{\lambda_j(\tau)} + \sqrt{\lambda_k(\tau)}} \right]_{i,j=1}^{d} \right.$$
$$\left. \circ \left[\mathbf{c}_k(\tau) \mathbf{c}_k^\top(\tau) \right] \Bigg) \mathbf{U}(\tau)^\top \right\| \tag{20}$$

where $\mathbf{H} + \tau\mathbf{\Theta} = \mathbf{U}(\tau) \cdot \operatorname{diag}(\lambda_1(\tau), \ldots, \lambda_d(\tau)) \cdot \mathbf{U}(\tau)^\top$ is the eigenvalue decomposition of the perturbation of $\mathbf{H}$, and $\mathbf{c}_k(\tau)$ is the k-th column of $\mathbf{X}(\tau) = \mathbf{U}(\tau)^\top \mathbf{\Theta} \mathbf{U}(\tau)$.

Define $\mathbf{Z}(\tau) = \left[\frac{1}{\sqrt{\lambda_i(\tau)} + \sqrt{\lambda_j(\tau)}} \right]_{i,j=1}^{d}$. The summation enclosed in () of the right hand side of (20) can be written as $\mathbf{Z}(\tau) \circ (\mathbf{Z}(\tau) \circ \mathbf{X}(\tau))^2$. Thus continuing from (20) yields

$$\varepsilon \leq \frac{\zeta^2}{2} \cdot \max_{\tau\in[0,\zeta]} \left\| \mathbf{Z}(\tau) \circ (\mathbf{Z}(\tau) \circ \mathbf{X}(\tau))^2 \right\|.$$

On the one hand, by the assumption that $\|\mathbf{H}\| \leq \sigma_{\max}$, we have

$$\|\mathbf{X}(\tau)\| = \|\mathbf{\Theta}\| = \left\| \mathbf{H}^{\frac{1}{2}} (\mathbf{H}^{-\frac{1}{2}} \mathbf{\Theta} \mathbf{H}^{-\frac{1}{2}}) \mathbf{H}^{\frac{1}{2}} \right\| \leq \|\mathbf{H}\| \cdot \left\| \mathbf{H}^{-\frac{1}{2}} \mathbf{\Theta} \mathbf{H}^{-\frac{1}{2}} \right\| \leq \sigma_{\max}. \tag{21}$$

On the other hand, the matrix $\mathbf{Z}(\tau)$ is positive semidefinite (see Horn and Johnson, 1991, Problem 9, page 348). Using the fact that $\|\mathbf{A} \circ \mathbf{B}\| \leq (\max_i \mathbf{A}_{ii}) \cdot \|\mathbf{B}\|$ for positive semidefinite $\mathbf{A}$ and Hermitian $\mathbf{B}$ (Horn and Johnson, 1991, Theorem 5.5.18), we conclude

$$\varepsilon \leq \frac{\zeta^2 \sigma_{\max}^2}{2} \left[\max_{\tau\in[0,\zeta]} \max_i \frac{1}{2\sqrt{\lambda_i(\tau)}} \right]^3.$$

Let $\zeta \leq \frac{3}{4} \sigma_{\max}^{-1} \sigma_{\min}$. Then in view of (21) and the Weyl's inequality, $\lambda_i(\tau) \geq \frac{\sigma_{\min}}{4}$ for all $\tau \in [0, \zeta]$, and we have $\varepsilon \leq \frac{1}{2} \zeta^2 \sigma_{\max}^2 \sigma_{\min}^{-\frac{3}{2}}$.

To sum up, we have shown so far the following first order approximation: for certain error matrix $\mathcal{E} \in \mathbb{R}^{d\times d}$, it holds

$$(\mathbf{H} + \zeta\mathbf{\Theta})^{\frac{1}{2}} = \mathbf{H}^{\frac{1}{2}} + \zeta \left[\frac{\lambda_i^{\frac{1}{2}} \lambda_j^{\frac{1}{2}}}{\lambda_i^{\frac{1}{2}} + \lambda_j^{\frac{1}{2}}} \right]_{i,j=1}^{d} \circ (\mathbf{H}^{-\frac{1}{2}} \mathbf{\Theta} \mathbf{H}^{-\frac{1}{2}}) + \mathcal{E} \qquad \text{where} \quad \|\mathcal{E}\| \leq \frac{1}{2} \zeta^2 \sigma_{\max}^2 \sigma_{\min}^{-\frac{3}{2}}.$$

Consequently, we have

$$(\mathbf{H} + \zeta\mathbf{\Theta})^{\frac{1}{2}} \mathbf{H}^{-\frac{1}{2}} - \mathbf{I} = \zeta \left[\frac{\lambda_i^{\frac{1}{2}}}{\lambda_i^{\frac{1}{2}} + \lambda_j^{\frac{1}{2}}} \right]_{i,j=1}^{d} \circ (\mathbf{H}^{-\frac{1}{2}} \mathbf{\Theta} \mathbf{H}^{-\frac{1}{2}}) + \mathcal{E}\mathbf{H}^{-\frac{1}{2}}.$$

Mathias (1997) showed that the norm of the first term on the right hand size is of the order $\mathcal{O}(\log d \cdot \zeta)$. Combining this with the fact that $\left\| \mathcal{E}\mathbf{H}^{-\frac{1}{2}} \right\| \leq \frac{1}{2} \zeta^2 \sigma_{\max}^2 \sigma_{\min}^{-2}$, we conclude that $\left\| \mathcal{E}\mathbf{H}^{-\frac{1}{2}} \right\| = \mathcal{O}(\zeta)$ for $\zeta = \mathcal{O}(\sigma_{\max}^{-2} \sigma_{\min}^2)$ and the lemma follows. ■

C.3. Proof of Lemma 6

Proof In view of the Weyl's inequality, we have

$$|\widehat{\rho}_1 - \rho_1| \leq \left\|\widehat{\mathbf{T}} - \mathbf{T}\right\| = \left\|\boldsymbol{\Sigma}_{xx}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-\frac{1}{2}} - \mathbf{E}_{xx}^{-\frac{1}{2}}\mathbf{E}_{xy}\mathbf{E}_{yy}^{-\frac{1}{2}}\right\|. \tag{22}$$

For the right hand side of (22), we have the following decomposition

$$\begin{aligned}
&\boldsymbol{\Sigma}_{xx}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-\frac{1}{2}} - \mathbf{E}_{xx}^{-\frac{1}{2}}\mathbf{E}_{xy}\mathbf{E}_{yy}^{-\frac{1}{2}} \\
=&\left(\boldsymbol{\Sigma}_{xx}^{-\frac{1}{2}} - \mathbf{E}_{xx}^{-\frac{1}{2}}\right)\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-\frac{1}{2}} + \mathbf{E}_{xx}^{-\frac{1}{2}}\left(\boldsymbol{\Sigma}_{xy} - \mathbf{E}_{xy}\right)\boldsymbol{\Sigma}_{yy}^{-\frac{1}{2}} + \mathbf{E}_{xx}^{-\frac{1}{2}}\mathbf{E}_{xy}\left(\boldsymbol{\Sigma}_{yy}^{-\frac{1}{2}} - \mathbf{E}_{yy}^{-\frac{1}{2}}\right).
\end{aligned} \tag{23}$$

By the equality

$$\mathbf{A}^{-\frac{1}{2}} - \mathbf{B}^{-\frac{1}{2}} = \mathbf{B}^{-\frac{1}{2}}\left(\mathbf{B}^{\frac{1}{2}} - \mathbf{A}^{\frac{1}{2}}\right)\mathbf{A}^{-\frac{1}{2}},$$

the first term of the RHS of (23) becomes

$$\left(\boldsymbol{\Sigma}_{xx}^{-\frac{1}{2}} - \mathbf{E}_{xx}^{-\frac{1}{2}}\right)\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-\frac{1}{2}} = \mathbf{E}_{xx}^{-\frac{1}{2}}\left(\mathbf{E}_{xx}^{\frac{1}{2}} - \boldsymbol{\Sigma}_{xx}^{\frac{1}{2}}\right)\boldsymbol{\Sigma}_{xx}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-\frac{1}{2}}.$$

When $\left\|\mathbf{E}_{xx}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{xx}\mathbf{E}_{xx}^{-\frac{1}{2}} - \mathbf{I}\right\| \leq \nu$, according to Lemma 5, we have (by making the identification that $\mathbf{H} = \mathbf{E}_{xx}$, $\zeta = \left\|\mathbf{E}_{xx}^{-\frac{1}{2}}(\boldsymbol{\Sigma}_{xx} - \mathbf{E}_{xx})\mathbf{E}_{xx}^{-\frac{1}{2}}\right\|$, and $\boldsymbol{\Theta} = (\boldsymbol{\Sigma}_{xx} - \mathbf{E}_{xx})/\left\|\mathbf{E}_{xx}^{-\frac{1}{2}}(\boldsymbol{\Sigma}_{xx} - \mathbf{E}_{xx})\mathbf{E}_{xx}^{-\frac{1}{2}}\right\|$)

$$\left\|\mathbf{E}_{xx}^{-\frac{1}{2}}\left(\mathbf{E}_{xx}^{\frac{1}{2}} - \boldsymbol{\Sigma}_{xx}^{\frac{1}{2}}\right)\right\| \leq C_d \cdot \nu$$

for $\nu \leq \frac{3}{4}\gamma^2$. Combining with the fact that $\left\|\boldsymbol{\Sigma}_{xx}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-\frac{1}{2}}\right\| \leq 1$, we have

$$\left\|\left(\boldsymbol{\Sigma}_{xx}^{-\frac{1}{2}} - \mathbf{E}_{xx}^{-\frac{1}{2}}\right)\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-\frac{1}{2}}\right\| \leq C_d \cdot \nu.$$

A similar bound can be obtained for the third term of (23). Observe that when $\left\|\mathbf{E}_{yy}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{yy}\mathbf{E}_{yy}^{-\frac{1}{2}} - \mathbf{I}\right\| \leq \nu < 1$, we have $\|\boldsymbol{\Sigma}_{yy} - \mathbf{E}_{yy}\| \leq \nu$ and all eigenvalues of $\boldsymbol{\Sigma}_{yy}$ lie in $[\gamma - \nu, 1 + \nu]$. Additionally, $\mathbf{E}_{yy}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{yy}\mathbf{E}_{yy}^{-\frac{1}{2}}$ is invertible, and all eigenvalues of $\boldsymbol{\Sigma}_{yy}^{-\frac{1}{2}}\mathbf{E}_{yy}\boldsymbol{\Sigma}_{yy}^{-\frac{1}{2}}$ lie in $[\frac{1}{1+\nu}, \frac{1}{1-\nu}]$, impling that $\left\|\boldsymbol{\Sigma}_{yy}^{-\frac{1}{2}}\mathbf{E}_{yy}\boldsymbol{\Sigma}_{yy}^{-\frac{1}{2}} - \mathbf{I}\right\| \leq \frac{\nu}{1-\nu}$. According to Lemma 5, we have (by making the identification that $\mathbf{H} = \boldsymbol{\Sigma}_{yy}$, $\zeta = \left\|\boldsymbol{\Sigma}_{yy}^{-\frac{1}{2}}(\boldsymbol{\Sigma}_{yy} - \mathbf{E}_{yy})\boldsymbol{\Sigma}_{yy}^{-\frac{1}{2}}\right\|$, and $\boldsymbol{\Theta} = (\boldsymbol{\Sigma}_{yy} - \mathbf{E}_{yy})/\left\|\boldsymbol{\Sigma}_{yy}^{-\frac{1}{2}}(\boldsymbol{\Sigma}_{yy} - \mathbf{E}_{yy})\boldsymbol{\Sigma}_{yy}^{-\frac{1}{2}}\right\|$)

$$\left\|\left(\mathbf{E}_{yy}^{\frac{1}{2}} - \boldsymbol{\Sigma}_{yy}^{\frac{1}{2}}\right)\boldsymbol{\Sigma}_{yy}^{-\frac{1}{2}}\right\| \leq \frac{C_d \cdot \nu}{1 - \nu} \tag{24}$$

for $\nu \leq \frac{3}{4}\left(\frac{\gamma-\nu}{1+\nu}\right)^2$, which is satisfied for $\nu \leq \frac{1}{4}\gamma^2$. Therefore, we can bound the third term of (23) as

$$\begin{aligned}\left\|\mathbf{E}_{xx}^{-\frac{1}{2}}\mathbf{E}_{xy}\left(\mathbf{\Sigma}_{yy}^{-\frac{1}{2}}-\mathbf{E}_{yy}^{-\frac{1}{2}}\right)\right\| &= \left\|\mathbf{E}_{xx}^{-\frac{1}{2}}\mathbf{E}_{xy}\mathbf{E}_{yy}^{-\frac{1}{2}}\left(\mathbf{E}_{yy}^{\frac{1}{2}}-\mathbf{\Sigma}_{yy}^{\frac{1}{2}}\right)\mathbf{\Sigma}_{yy}^{-\frac{1}{2}}\right\| \\ &\leq \left\|\mathbf{E}_{xx}^{-\frac{1}{2}}\mathbf{E}_{xy}\mathbf{E}_{yy}^{-\frac{1}{2}}\right\| \cdot \left\|\left(\mathbf{E}_{yy}^{\frac{1}{2}}-\mathbf{\Sigma}_{yy}^{\frac{1}{2}}\right)\mathbf{\Sigma}_{yy}^{-\frac{1}{2}}\right\| \\ &\leq \frac{C_d\cdot\nu}{1-\nu} \leq 2C_d\cdot\nu\end{aligned}$$

where we have used the fact that $\left\|\mathbf{E}_{xx}^{-\frac{1}{2}}\mathbf{E}_{xy}\mathbf{E}_{yy}^{-\frac{1}{2}}\right\| \leq 1$ by Lemma 20.

For the second term of (23), we have by assumption that

$$\left\|\mathbf{E}_{xx}^{-\frac{1}{2}}\left(\mathbf{\Sigma}_{xy}-\mathbf{E}_{xy}\right)\mathbf{\Sigma}_{yy}^{-\frac{1}{2}}\right\| \leq \nu.$$

Applying the triangle inequality, we obtain from (23) that

$$\left\|\mathbf{\Sigma}_{xx}^{-\frac{1}{2}}\mathbf{\Sigma}_{xy}\mathbf{\Sigma}_{yy}^{-\frac{1}{2}}-\mathbf{E}_{xx}^{-\frac{1}{2}}\mathbf{E}_{xy}\mathbf{E}_{yy}^{-\frac{1}{2}}\right\| \leq 4C_d\cdot\nu. \tag{25}$$

To sum up, it suffices to set $\nu = \frac{\epsilon'}{4C_d}$ to ensure $\left\|\widehat{\mathbf{T}}-\mathbf{T}\right\| \leq \epsilon'$, and this yields the desired sample complexity. ■

C.4. Proof of Theorem 9

Proof Apply Lemma 6 with $\epsilon' = \frac{\sqrt{\epsilon}\Delta}{4}$ where $\epsilon \in (0,1)$ is the desired accuracy in Theorem 9. Since $\left\|\mathbf{T}-\hat{\mathbf{T}}\right\| \leq \epsilon' < \frac{\Delta}{4}$, the eigenvalues of $\hat{\mathbf{T}}$ are within $\frac{\Delta}{4}$ of those of $\mathbf{T}$ due to Weyl's inequality, so there exists a positive singular gap of $\frac{\Delta}{2}$ for the empirical estimate $\hat{\mathbf{T}}$, whose first pair of singular vectors is unique. In view of the off-diagonal structure of $\widehat{\mathbf{C}}$, we observe that $\left\|\mathbf{C}-\hat{\mathbf{C}}\right\| = \left\|\mathbf{T}-\hat{\mathbf{T}}\right\| \leq \epsilon'$ and that the top eigenvector of $\hat{\mathbf{C}}$ is unique. Then with the number of samples given in the theorem ensuring the ϵ' perturbation, according to the Davis-Kahan $\sin\theta$ theorem (Davis and Kahan, 1970), the top eigenvectors of $\mathbf{C}$ and $\widehat{\mathbf{C}}$ are well aligned:

$$\sin^2\theta \leq \frac{\left\|\mathbf{C}-\widehat{\mathbf{C}}\right\|^2}{\Delta^2} \leq \frac{\epsilon'^2}{\Delta^2} = \frac{\epsilon}{16} \tag{26}$$

where θ is the angle between the top eigenvector of $\mathbf{C}$ and that of $\widehat{\mathbf{C}}$. This is equivalent to

$$\left\|\mathbf{E}_{xx}^{\frac{1}{2}}\mathbf{u}^*-\mathbf{\Sigma}_{xx}^{\frac{1}{2}}\widehat{\mathbf{u}}\right\|^2 + \left\|\mathbf{E}_{yy}^{\frac{1}{2}}\mathbf{v}^*-\mathbf{\Sigma}_{yy}^{\frac{1}{2}}\widehat{\mathbf{v}}\right\|^2 \leq \frac{\epsilon}{8} \tag{27}$$

and so $\max\left(\left\|\mathbf{E}_{xx}^{\frac{1}{2}}\mathbf{u}^* - \mathbf{\Sigma}_{xx}^{\frac{1}{2}}\widehat{\mathbf{u}}\right\|^2, \left\|\mathbf{E}_{yy}^{\frac{1}{2}}\mathbf{v}^* - \mathbf{\Sigma}_{yy}^{\frac{1}{2}}\widehat{\mathbf{v}}\right\|^2\right) \leq \frac{\epsilon}{8}$.

In the rest of the proof, we fix the issue of incorrect normalization of $(\widehat{\mathbf{u}}, \widehat{\mathbf{v}})$. Recall we have shown in the proof of Lemma 6 that (see e.g., (24))

$$\left\|\mathbf{I} - \mathbf{E}_{xx}^{\frac{1}{2}}\mathbf{\Sigma}_{xx}^{-\frac{1}{2}}\right\| \leq \epsilon' \leq \frac{\sqrt{\epsilon}}{4}.$$

Consequently, we have

$$\begin{aligned}
\left\|\mathbf{E}_{xx}^{\frac{1}{2}}\mathbf{u}^* - \mathbf{E}_{xx}^{\frac{1}{2}}\widehat{\mathbf{u}}\right\|^2 &= \left\|\mathbf{E}_{xx}^{\frac{1}{2}}\mathbf{u}^* - \left(\mathbf{E}_{xx}^{\frac{1}{2}}\mathbf{\Sigma}_{xx}^{-\frac{1}{2}}\right)\left(\mathbf{\Sigma}_{xx}^{\frac{1}{2}}\widehat{\mathbf{u}}\right)\right\|^2 \\
&\leq \left(\left\|\mathbf{E}_{xx}^{\frac{1}{2}}\mathbf{u}^* - \mathbf{\Sigma}_{xx}^{\frac{1}{2}}\widehat{\mathbf{u}}\right\| + \left\|\left(\mathbf{I} - \mathbf{E}_{xx}^{\frac{1}{2}}\mathbf{\Sigma}_{xx}^{-\frac{1}{2}}\right)\mathbf{\Sigma}_{xx}^{\frac{1}{2}}\widehat{\mathbf{u}}\right\|\right)^2 \\
&\leq 2\left\|\mathbf{E}_{xx}^{\frac{1}{2}}\mathbf{u}^* - \mathbf{\Sigma}_{xx}^{\frac{1}{2}}\widehat{\mathbf{u}}\right\|^2 + 2\left\|\mathbf{I} - \mathbf{E}_{xx}^{\frac{1}{2}}\mathbf{\Sigma}_{xx}^{-\frac{1}{2}}\right\|^2 \\
&\leq \frac{\epsilon}{4} + \frac{\epsilon}{8} \\
&\leq \frac{\epsilon}{2}
\end{aligned}$$

where we have used the facts that $(x+y)^2 \leq 2x^2 + 2y^2$ and $\left\|\mathbf{\Sigma}_{xx}^{\frac{1}{2}}\widehat{\mathbf{u}}\right\| = 1$ in the second inequality.

According to Lemma 21, we then have

$$\left\|\frac{\mathbf{E}_{xx}^{\frac{1}{2}}\mathbf{u}^*}{\left\|\mathbf{E}_{xx}^{\frac{1}{2}}\mathbf{u}^*\right\|} - \frac{\mathbf{E}_{xx}^{\frac{1}{2}}\widehat{\mathbf{u}}}{\left\|\mathbf{E}_{xx}^{\frac{1}{2}}\widehat{\mathbf{u}}\right\|}\right\|^2 \leq \frac{4\left\|\mathbf{E}_{xx}^{\frac{1}{2}}\mathbf{u}^* - \mathbf{E}_{xx}^{\frac{1}{2}}\widehat{\mathbf{u}}\right\|^2}{\left\|\mathbf{E}_{xx}^{\frac{1}{2}}\mathbf{u}^*\right\|^2} = 4\left\|\mathbf{E}_{xx}^{\frac{1}{2}}\mathbf{u}^* - \mathbf{E}_{xx}^{\frac{1}{2}}\widehat{\mathbf{u}}\right\|^2 \leq 2\epsilon$$

and thus the alignment between these two vectors is

$$\frac{\widehat{\mathbf{u}}^\top\mathbf{E}_{xx}\mathbf{u}^*}{\left\|\mathbf{E}_{xx}^{\frac{1}{2}}\widehat{\mathbf{u}}\right\|} = 1 - \frac{1}{2}\left\|\frac{\mathbf{E}_{xx}^{\frac{1}{2}}\mathbf{u}^*}{\left\|\mathbf{E}_{xx}^{\frac{1}{2}}\mathbf{u}^*\right\|} - \frac{\mathbf{E}_{xx}^{\frac{1}{2}}\widehat{\mathbf{u}}}{\left\|\mathbf{E}_{xx}^{\frac{1}{2}}\widehat{\mathbf{u}}\right\|}\right\|^2 \geq 1 - \epsilon.$$

A similar bound is obtained for $\widehat{\mathbf{v}}$:

$$\frac{\widehat{\mathbf{v}}^\top\mathbf{E}_{yy}\mathbf{v}^*}{\left\|\mathbf{E}_{yy}^{\frac{1}{2}}\widehat{\mathbf{v}}\right\|} \geq 1 - \epsilon.$$

Averaging the above two inequalities yields the desired result. Requiring that $\epsilon' = \frac{\sqrt{\epsilon}\Delta}{4} \leq C_d\gamma^2$ as in Corollary 7 leads to the extra condition that $\epsilon \leq \frac{16C_d^2\gamma^4}{\Delta^2}$. ■

Appendix D. Proofs for Section 4

D.1. Proof of Lemma 10

Proof If we obtain an approximate solution $\mathbf{w}_{t+1}$ to (13), such that $f_{t+1}(\mathbf{w}_{t+1}) - f_{t+1}(\mathbf{w}^*_{t+1}) = \epsilon_t(\mathbf{w}_t^\top \widehat{\mathbf{B}}\mathbf{w}_t)$, it holds that

$$\begin{aligned}
\epsilon_t \left\|\widehat{\mathbf{B}}^{\frac{1}{2}}\mathbf{w}_t\right\|^2 &= \frac{1}{2}\left(\mathbf{w}_{t+1} - \mathbf{w}^*_{t+1}\right)^\top \widehat{\mathbf{A}}_\lambda \left(\mathbf{w}_{t+1} - \mathbf{w}^*_{t+1}\right) \\
&= \frac{1}{2}\left(\widehat{\mathbf{B}}^{\frac{1}{2}}\mathbf{w}_{t+1} - \widehat{\mathbf{B}}^{\frac{1}{2}}\mathbf{w}^*_{t+1}\right)^\top \widehat{\mathbf{B}}^{-\frac{1}{2}}\widehat{\mathbf{A}}_\lambda\widehat{\mathbf{B}}^{-\frac{1}{2}}\left(\widehat{\mathbf{B}}^{\frac{1}{2}}\mathbf{w}_{t+1} - \widehat{\mathbf{B}}^{\frac{1}{2}}\mathbf{w}^*_{t+1}\right) \\
&= \frac{1}{2}\left(\mathbf{r}_{t+1} - \mathbf{r}^*_{t+1}\right)^\top \widehat{\mathbf{M}}_\lambda^{-1}\left(\mathbf{r}_{t+1} - \mathbf{r}^*_{t+1}\right) = \frac{1}{2}\left\|\mathbf{r}_{t+1} - \mathbf{r}^*_{t+1}\right\|^2_{\widehat{\mathbf{M}}_\lambda^{-1}},
\end{aligned}$$

or equivalently

$$\left\|\mathbf{r}_{t+1} - \mathbf{r}^*_{t+1}\right\|_{\widehat{\mathbf{M}}_\lambda^{-1}} = \sqrt{2\epsilon_t}\cdot\left\|\mathbf{r}_t\right\|.$$

Note that our choice of ϵ_t is also invariant to the length of $\mathbf{r}_t$ (or whether normalization is performed).

For the exact solution to the linear system, we have

$$\mathbf{r}^*_{t+1} = \widehat{\mathbf{M}}_\lambda\mathbf{r}_t = \|\mathbf{r}_t\|\sum_{i=1}^d \beta_i\xi_{ti}\mathbf{p}_i.$$

As a result, we can bound the numerator and denominator of $G(\mathbf{r}_{t+1})$ respectively:

$$\begin{aligned}
\left\|\mathbf{P}_\perp\frac{\mathbf{r}_{t+1}}{\|\mathbf{r}_{t+1}\|}\right\|_{\widehat{\mathbf{M}}_\lambda^{-1}} &\le \frac{1}{\|\mathbf{r}_{t+1}\|}\left(\left\|\mathbf{P}_\perp\mathbf{r}^*_{t+1}\right\|_{\widehat{\mathbf{M}}_\lambda^{-1}} + \left\|\mathbf{P}_\perp\left(\mathbf{r}_{t+1} - \mathbf{r}^*_{t+1}\right)\right\|_{\widehat{\mathbf{M}}_\lambda^{-1}}\right) \\
&\le \frac{1}{\|\mathbf{r}_{t+1}\|}\left(\left\|\mathbf{P}_\perp\mathbf{r}^*_{t+1}\right\|_{\widehat{\mathbf{M}}_\lambda^{-1}} + \left\|\mathbf{r}_{t+1} - \mathbf{r}^*_{t+1}\right\|_{\widehat{\mathbf{M}}_\lambda^{-1}}\right) \\
&= \frac{\|\mathbf{r}_t\|}{\|\mathbf{r}_{t+1}\|}\left(\sqrt{\sum_{i=2}^d \beta_i\xi_{ti}^2} + \sqrt{2\epsilon_t}\right),
\end{aligned}$$

and

$$\begin{aligned}
\left\|\mathbf{P}_\parallel\frac{\mathbf{r}_{t+1}}{\|\mathbf{r}_{t+1}\|}\right\|_{\widehat{\mathbf{M}}_\lambda^{-1}} &\ge \frac{1}{\|\mathbf{r}_{t+1}\|}\left(\left\|\mathbf{P}_\parallel\mathbf{r}^*_{t+1}\right\|_{\widehat{\mathbf{M}}_\lambda^{-1}} - \left\|\mathbf{P}_\parallel\left(\mathbf{r}_{t+1} - \mathbf{r}^*_{t+1}\right)\right\|_{\widehat{\mathbf{M}}_\lambda^{-1}}\right) \\
&\ge \frac{1}{\|\mathbf{r}_{t+1}\|}\left(\left\|\mathbf{P}_\parallel\mathbf{r}^*_{t+1}\right\|_{\widehat{\mathbf{M}}_\lambda^{-1}} - \left\|\mathbf{r}_{t+1} - \mathbf{r}^*_{t+1}\right\|_{\widehat{\mathbf{M}}_\lambda^{-1}}\right) \\
&= \frac{\|\mathbf{r}_t\|}{\|\mathbf{r}_{t+1}\|}\left(\sqrt{\beta_1\xi_{t1}^2} - \sqrt{2\epsilon_t}\right).
\end{aligned}$$

Consequently, we have

$$G(\mathbf{r}_{t+1}) \le \frac{\sqrt{\sum_{i=2}^d \beta_i \xi_{ti}^2} + \sqrt{2\epsilon_t}}{\sqrt{\beta_1 \xi_{t1}^2} - \sqrt{2\epsilon_t}} \le \frac{\beta_2 \sqrt{\sum_{i=2}^d \xi_{ti}^2/\beta_i} + \sqrt{2\epsilon_t}}{\beta_1 \sqrt{\xi_{t1}^2/\beta_1} - \sqrt{2\epsilon_t}}$$
$$= G(\mathbf{r}_t) \cdot \frac{\beta_2 + \frac{\sqrt{2\epsilon_t}}{\sqrt{\sum_{i=2}^d \xi_{ti}^2/\beta_i}}}{\beta_1 - \frac{\sqrt{2\epsilon_t}}{\sqrt{\xi_{t1}^2/\beta_1}}}.$$

As long as $\sqrt{2\epsilon_t} \le \min\left(\sqrt{\sum_{i=2}^d \xi_{ti}^2/\beta_i},\ \sqrt{\xi_{t1}^2/\beta_1}\right) \cdot \frac{\beta_1-\beta_2}{4}$, i.e.,

$$\epsilon_t \le \min\left(\sum_{i=2}^d \xi_{ti}^2/\beta_i,\ \xi_{t1}^2/\beta_1\right) \cdot \frac{(\beta_1-\beta_2)^2}{32},$$

we are guaranteed that

$$G(\mathbf{r}_{t+1}) \le G(\mathbf{r}_t) \cdot \frac{\beta_1 + 3\beta_2}{3\beta_1 + \beta_2}.$$

Substituting in $\beta_i = \frac{1}{\lambda - \widehat{\rho}_i}$ with $\lambda - \widehat{\rho}_1 \le \widehat{\Delta}$, we obtain that

$$\frac{\beta_1 + 3\beta_2}{3\beta_1 + \beta_2} \le \frac{5}{7} < 1.$$

This means that if (14) holds for each least squares problem, the sequence $\{G(\mathbf{r}_t)\}_{t=0,\dots}$ decreases (at least) at a constant geometric rate of $\frac{5}{7}$. Therefore, the number of inexact matrix-vector multiplications T needed to achieve $|\sin\theta_T| \le \eta$ is $\log_{\frac{7}{5}}\left(\frac{G(\mathbf{r}_0)}{\eta}\right)$. ■

D.2. Bounding the initial error for each least squares

We can minimize the initial suboptimality for the least squares problem f_{t+1} for reducing the time complexity of its solver. It is natural to use an initialization of the form $\alpha\mathbf{w}_t$, a scaled version of the previous iterate, which gives the following objective

$$f_{t+1}(\alpha\mathbf{w}_t) = \frac{(\mathbf{w}_t^\top \widehat{\mathbf{A}}_\lambda \mathbf{w}_t)}{2}\alpha^2 - (\mathbf{w}_t^\top \widehat{\mathbf{B}}\mathbf{w}_t)\alpha.$$

This is a quadratic function of α, and minimizing $f_{t+1}(\alpha\mathbf{w}_t)$ over α gives the optimal scaling $\alpha_t^* = \frac{\mathbf{w}_t^\top \widehat{\mathbf{B}}\mathbf{w}_t}{\mathbf{w}_t^\top \widehat{\mathbf{A}}_\lambda \mathbf{w}_t}$ (and this quantity is also invariant to the length of $\mathbf{w}_t$). Observe that α_t^* naturally measures the quality of $\mathbf{w}_t$: As $\mathbf{w}_t$ converges to $\widehat{\mathbf{w}}$, α_t^* converges to β_1. This initialization technique plays an important role in showing the linear convergence of our algorithm, and was used by Ge et al. (2016) for their standard power iterations (alternating least squares) scheme for CCA.

Proof [Proof of Lemma 11] With the given initialization, we have

$$\begin{aligned} f_{t+1}(\alpha_t^*\mathbf{w}_t) - f_{t+1}^* &\leq f_{t+1}(\beta_1\mathbf{w}_t) - f_{t+1}^* \\ &= \frac{\beta_1^2\mathbf{r}_t^\top\widehat{\mathbf{M}}_\lambda^{-1}\mathbf{r}_t}{2} - \beta_1\mathbf{r}_t^\top\mathbf{r}_t + \frac{\mathbf{r}_t\widehat{\mathbf{M}}_\lambda\mathbf{r}_t}{2} \\ &= \frac{\|\mathbf{r}_t\|^2}{2}\sum_{i=1}^d \xi_{ti}^2\left(\frac{\beta_1^2}{\beta_i} - 2\beta_1 + \beta_i\right) \\ &= \frac{\|\mathbf{r}_t\|^2}{2}\sum_{i=1}^d \frac{\xi_{ti}^2}{\beta_i}\left(\beta_1 - \beta_i\right)^2 \\ &\leq \frac{(\mathbf{w}_t^\top\widehat{\mathbf{B}}\mathbf{w}_t)}{2}\cdot\beta_1^2\sum_{i=2}^d\frac{\xi_{ti}^2}{\beta_i}. \end{aligned}$$

Therefore, in view of (14), it suffices to set the ratio between the initial and the final error of f_{t+1} to

$$\max\left(1, G(\mathbf{r}_t)\right)\cdot\frac{16\beta_1^2}{\left(\beta_1-\beta_2\right)^2}.$$

In the initial phase, $G(\mathbf{r}_t)$ is large, we can set the ratio to be $G(\mathbf{r}_0)\cdot\frac{16\beta_1^2}{(\beta_1-\beta_2)^2}$, until it is reduced to 1 after $\mathcal{O}\left(\log G(\mathbf{r}_0)\right)$ iterations. Afterwards, we can set the ratio to be the constant of $\frac{16\beta_1^2}{(\beta_1-\beta_2)^2}$, until we reach the desired accuracy. Observe that

$$\frac{\beta_1^2}{\left(\beta_1-\beta_2\right)^2} = \left(\frac{\frac{1}{\lambda-\widehat{\rho}_1}}{\frac{1}{\lambda-\widehat{\rho}_1}-\frac{1}{\lambda-\widehat{\rho}_2}}\right)^2 = \left(\frac{\lambda-\widehat{\rho}_2}{\widehat{\rho}_1-\widehat{\rho}_2}\right)^2 \leq (u+1)^2 \leq 4.$$

■

D.3. Time complexity of SVRG for finite sum with nonconvex component

Lemma 25 (Time complexity of SVRG for (15)**)** *With the initialization $\alpha_t^*\mathbf{w}_t$, SVRG outputs an $\mathbf{w}_{t+1}$ such that $f_{t+1}(\mathbf{w}_{t+1}) - f_{t+1}^* \leq \epsilon_t(\mathbf{w}_t^\top\widehat{\mathbf{B}}\mathbf{w}_t)$ in time*

$$\mathcal{O}\left(d(N+\kappa^2)\log\left(64\max\left(G(\mathbf{r}_t),1\right)\right)\cdot\kappa\right),$$

where $\kappa = \frac{\max_i L_i}{\Lambda}$ with L_i being the gradient Lipschitz constant of f_{t+1}^i, and Λ is the strongly-convex constant of f_{t+1}. Futhermore, if we sample each component f_{t+1}^i non-uniformly with probability proportional to L_i^2 for the SVRG stochastic updates, we have instead $\kappa = \sqrt{\frac{\frac{1}{N}\sum_{i=1}^N L_i^2}{\Lambda^2}}$.

Although not explicitly stated by Garber and Hazan (2015), the result for non-uniform sampling is straightforward by a careful investigation of their analysis; we provide detailed proof of this result in Appendix F. The effect of improved dependence on L_i's through non-uniform sampling agrees with related work (Xiao and Zhang, 2014). The purpose of the non-uniform sampling variant is to bound κ^2 with high probability for sub-Gaussian/regular polynomial-tail inputs.

D.4. Bounding the condition number for SVRG

The next lemma upper-bounds the "condition number" κ^2.

Lemma 26 *Solving* $\min_{\mathbf{w}} f_{t+1}(\mathbf{w})$ *using SVRG with non-uniform sampling, we have* $\kappa^2 = \mathcal{O}\left(\frac{d^2}{\widehat{\Delta}^2\gamma^2}\right)$ *for the sub-Gaussian/regular polynomial-tail classes with high probability over the sample set, and* $\kappa^2 = \mathcal{O}\left(\frac{1}{\widehat{\Delta}^2\gamma^2}\right)$ *for the bounded class.*

Proof The gradient Lipschitz constant L_i is bounded by the largest eigenvalue (in absolute value) of its Hessian

$$\mathbf{Q}_\lambda^i = \begin{bmatrix} \lambda\mathbf{x}_i\mathbf{x}_i^\top & -\mathbf{x}_i\mathbf{y}_i^\top \\ -\mathbf{y}_i\mathbf{x}_i^\top & \lambda\mathbf{y}_i\mathbf{y}_i^\top \end{bmatrix},$$

and the largest eigenvalue is defined as

$$\max_{\mathbf{g}_x\in\mathbb{R}^{d_x},\mathbf{g}_y\in\mathbb{R}^{d_y}} \beta := \left|[\mathbf{g}_x^\top,\mathbf{g}_y^\top]\mathbf{Q}_\lambda^i \begin{bmatrix} \mathbf{g}_x \\ \mathbf{g}_y \end{bmatrix}\right| \qquad \text{s.t.} \quad \|\mathbf{g}_x\|^2 + \|\mathbf{g}_y\|^2 = 1.$$

We have

$$\begin{aligned}
\beta &= \left|\lambda(\mathbf{g}_x^\top\mathbf{x}_i)^2 + \lambda(\mathbf{g}_y^\top\mathbf{y}_i)^2 - 2(\mathbf{g}_x^\top\mathbf{x}_i)(\mathbf{g}_y^\top\mathbf{y}_i)\right| \\
&\le \lambda(\mathbf{g}_x^\top\mathbf{x}_i)^2 + \lambda(\mathbf{g}_y^\top\mathbf{y}_i)^2 + 2\left|\mathbf{g}_x^\top\mathbf{x}_i\right|\left|\mathbf{g}_y^\top\mathbf{y}_i\right| \\
&\le \lambda(\mathbf{g}_x^\top\mathbf{x}_i)^2 + \lambda(\mathbf{g}_y^\top\mathbf{y}_i)^2 + (\mathbf{g}_x^\top\mathbf{x}_i)^2 + (\mathbf{g}_y^\top\mathbf{y}_i)^2 \\
&= (\lambda+1)\left((\mathbf{g}_x^\top\mathbf{x}_i)^2 + (\mathbf{g}_y^\top\mathbf{y}_i)\right) \\
&\le (\lambda+1)\left(\|\mathbf{g}_x\|^2\|\mathbf{x}_i\|^2 + \|\mathbf{g}_y\|^2\|\mathbf{y}_i\|^2)\right) \\
&\le (\lambda+1)\cdot\max\left(\|\mathbf{x}_i\|^2, \|\mathbf{y}_i\|^2\right) \\
&\le (\lambda+1)\cdot\left(\|\mathbf{x}_i\|^2 + \|\mathbf{y}_i\|^2\right)
\end{aligned}$$

where we have used the Cauchy-Schwarz inequality in the third inequality.

Note that, for bounded inputs, we have $\|\mathbf{x}_i\|^2 + \|\mathbf{y}_i\|^2 \le 2$ and so $L_i^2 \le 4(\lambda+1)^2$ for all $i = 1,\ldots,N$. For sub-Gaussian/regular polynomial-tail inputs, we have

$$\frac{1}{N}\sum_{i=1}^N L_i^2 \le (\lambda+1)^2\cdot\frac{1}{N}\sum_{i=1}^N\left(\|\mathbf{x}_i\|^2 + \|\mathbf{y}_i\|^2\right)^2 = \mathcal{O}((\lambda+1)^2d^2)$$

with high probability in view of Remark 24.

On the other hand, we have shown that $\Lambda = \sigma_{\min}(\mathbf{A}_\lambda) \ge (\lambda - \widehat{\rho}_1)\gamma/2$. Recalling $\lambda = \widehat{\rho}_1 + c\widehat{\Delta}$ with $c \in (0,1)$, we have $\lambda \le 2$ and $\Lambda \ge c\widehat{\Delta}\gamma/2$. Combining this with the data norm bound above yields the desired result. ■

D.5. Proof of Theorem 12

Proof Since $\frac{{\mathbf{u}_T}^\top \mathbf{\Sigma}_{xx}\widehat{\mathbf{u}}}{\left\|\mathbf{\Sigma}_{xx}^{\frac{1}{2}}\mathbf{u}_T\right\|} \leq 1$ and $\frac{{\mathbf{v}_T}^\top \mathbf{\Sigma}_{yy}\widehat{\mathbf{v}}}{\left\|\mathbf{\Sigma}_{yy}^{\frac{1}{2}}\mathbf{v}_T\right\|} \leq 1$, it suffices to require

$$\frac{{\mathbf{u}_T}^\top \mathbf{\Sigma}_{xx}\widehat{\mathbf{u}}}{\left\|\mathbf{\Sigma}_{xx}^{\frac{1}{2}}\mathbf{u}_T\right\|} + \frac{{\mathbf{v}_T}^\top \mathbf{\Sigma}_{yy}\widehat{\mathbf{v}}}{\left\|\mathbf{\Sigma}_{yy}^{\frac{1}{2}}\mathbf{v}_T\right\|} \geq 2 - \eta.$$

According to Lemma 22 (making the identification that $\mathbf{a} = \mathbf{\Sigma}_{xx}^{\frac{1}{2}}\widehat{\mathbf{u}}$, $\mathbf{x} = \mathbf{\Sigma}_{xx}^{\frac{1}{2}}\mathbf{u}_T$, $\mathbf{b} = \mathbf{\Sigma}_{yy}^{\frac{1}{2}}\widehat{\mathbf{v}}$, and $\mathbf{y} = \mathbf{\Sigma}_{yy}^{\frac{1}{2}}\mathbf{v}_T$), it then suffices to have

$$\cos\theta_T = \frac{1}{\sqrt{2}} \frac{\widehat{\mathbf{u}}^\top \mathbf{\Sigma}_{xx}\mathbf{u}_T + \widehat{\mathbf{v}}^\top \mathbf{\Sigma}_{yy}\mathbf{v}_T}{\sqrt{\mathbf{u}_T^\top \mathbf{\Sigma}_{xx}\mathbf{u}_T + \mathbf{v}_T^\top \mathbf{\Sigma}_{yy}\mathbf{v}_T}} \geq 1 - \frac{\eta}{8}. \tag{28}$$

Since $\cos\theta_T = \sqrt{1-\sin^2\theta_T} \geq 1 - \sin^2\theta_T$, we just need $|\sin\theta_T| \leq \frac{\sqrt{\eta}}{\sqrt{8}}$, and we ensure it by requiring $G(\mathbf{r}_T) \leq \frac{\sqrt{\eta}}{\sqrt{8}}$. Applying results from the previous sections, we need to solve $\mathcal{O}\left(\log\frac{G(\mathbf{r}_0)}{\eta}\right)$ linear systems, and the time complexity for solving each is at most $\mathcal{O}\left((N+\kappa^2)\log\left(G(\mathbf{r}_0)\cdot\kappa\right)\right)$ for SVRG.

It remains to bound $G(\mathbf{r}_0)$. By the definition of $G(\cdot)$, we have

$$G(\mathbf{r}_0) = \frac{\sqrt{\sum_{i=2}^d \xi_{0i}^2/\beta_i}}{\sqrt{\xi_{01}^2/\beta_1}} \leq \sqrt{\frac{\beta_1}{\beta_d}} \cdot \frac{1}{|\xi_{01}|}$$

where $|\xi_{01}|$ is the alignment between $\mathbf{r}_0$ and $\mathbf{p}_1$ which, by the relationship between $\mathbf{r}_0$ and $\mathbf{w}_0$, satisfy

$$\begin{aligned}
\left|\mathbf{r}_0^\top \mathbf{p}_1\right| &= \frac{\left|\tilde{\mathbf{w}}_0^\top \widehat{\mathbf{B}}^{\frac{1}{2}}\mathbf{p}_1\right|}{\left\|\widehat{\mathbf{B}}^{\frac{1}{2}}\tilde{\mathbf{w}}_0\right\|} \geq \frac{\left|\tilde{\mathbf{w}}_0^\top (\widehat{\mathbf{B}}^{\frac{1}{2}}\mathbf{p}_1)\right|}{\sigma_{\max}(\widehat{\mathbf{B}}^{\frac{1}{2}})\cdot\|\tilde{\mathbf{w}}_0\|} = \frac{\left\|\widehat{\mathbf{B}}^{\frac{1}{2}}\mathbf{p}_1\right\| \cdot \left|(\tilde{\mathbf{w}}_0/\|\tilde{\mathbf{w}}_0\|)^\top (\widehat{\mathbf{B}}^{\frac{1}{2}}\mathbf{p}_1/\left\|\widehat{\mathbf{B}}^{\frac{1}{2}}\mathbf{p}_1\right\|)\right|}{\sigma_{\max}(\widehat{\mathbf{B}}^{\frac{1}{2}})} \\
&\geq \frac{\sigma_{\min}(\widehat{\mathbf{B}}^{\frac{1}{2}})}{\sigma_{\max}(\widehat{\mathbf{B}}^{\frac{1}{2}})} \cdot \left|\left(\frac{\tilde{\mathbf{w}}_0}{\|\tilde{\mathbf{w}}_0\|}\right)^\top \left(\frac{\widehat{\mathbf{B}}^{\frac{1}{2}}\mathbf{p}_1}{\left\|\widehat{\mathbf{B}}^{\frac{1}{2}}\mathbf{p}_1\right\|}\right)\right|.
\end{aligned}$$

According to the way $\tilde{\mathbf{w}}_0$ is initialized and Arora et al. (2009, Lemma 5), we have with probability at least $1-C$ that $\left|\left(\frac{\tilde{\mathbf{w}}_0}{\|\tilde{\mathbf{w}}_0\|}\right)^\top \left(\frac{\widehat{\mathbf{B}}^{\frac{1}{2}}\mathbf{p}_1}{\left\|\widehat{\mathbf{B}}^{\frac{1}{2}}\mathbf{p}_1\right\|}\right)\right| \geq \frac{C}{\sqrt{d}}$. On the other hand, we have $\frac{\beta_1}{\beta_d} = \mathcal{O}(\frac{1}{\widehat{\Delta}})$ and $\frac{\sigma_{\max}(\widehat{\mathbf{B}}^{\frac{1}{2}})}{\sigma_{\min}(\widehat{\mathbf{B}}^{\frac{1}{2}})} = \mathcal{O}\left(\frac{1}{\sqrt{\gamma}}\right)$. Combining these results yields that $G(\mathbf{r}_0) = \mathcal{O}\left(\sqrt{\frac{d}{\widehat{\Delta}\gamma}}\right)$ with high probability. Then the theorem follows. ■

D.6. Proof of Corollary 13

Proof Denote $\widetilde{\mathbf{r}} := \frac{1}{\sqrt{2}} \begin{bmatrix} \boldsymbol{\Sigma}_{xx}^{\frac{1}{2}} \mathbf{u}_T / \left\| \boldsymbol{\Sigma}_{xx}^{\frac{1}{2}} \mathbf{u}_T \right\| \\ \boldsymbol{\Sigma}_{yy}^{\frac{1}{2}} \mathbf{v}_T / \left\| \boldsymbol{\Sigma}_{yy}^{\frac{1}{2}} \mathbf{v}_T \right\| \end{bmatrix}$, with $\|\widetilde{\mathbf{r}}\| = 1$. Assume without loss of generality that $\left\| \boldsymbol{\Sigma}_{xx}^{\frac{1}{2}} \mathbf{u}_T \right\| = \left\| \boldsymbol{\Sigma}_{yy}^{\frac{1}{2}} \mathbf{v}_T \right\| = 1$; this does not affect our measure of alignment, and can be ensured by a final (separate) normalization step with cost $\mathcal{O}(Nd)$ (Wang et al., 2016).

Apply Lemma 6 with $\epsilon' = \frac{\sqrt{\epsilon}\Delta}{8}$; requiring that $\epsilon' = \frac{\sqrt{\epsilon}\Delta}{8} \leq C_d \gamma^2$ as in Corollary 7 leads to the extra condition that $\epsilon \leq \frac{64 C_d^2 \gamma^4}{\Delta^2}$.

With the specified sample complexity, we have that with high probability

$$\left\| \mathbf{T} - \widehat{\mathbf{T}} \right\| \leq \frac{\sqrt{\epsilon}\Delta}{8} \leq \frac{\Delta}{8}. \tag{29}$$

In view of the Weyl's inequality, (29) implies that $\widehat{\Delta} \geq \frac{3\Delta}{4}$.

Let $\mathbf{r}^* = \frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{E}_{xx}^{\frac{1}{2}} \mathbf{u}^* \\ \mathbf{E}_{yy}^{\frac{1}{2}} \mathbf{v}^* \end{bmatrix}$ be the top eigenvector of $\mathbf{C}$. And recall $\widehat{\mathbf{r}} := \frac{1}{\sqrt{2}} \begin{bmatrix} \boldsymbol{\Sigma}_{xx}^{\frac{1}{2}} \widehat{\mathbf{u}} \\ \boldsymbol{\Sigma}_{yy}^{\frac{1}{2}} \widehat{\mathbf{v}} \end{bmatrix}$ is the top eigenvector of $\widehat{\mathbf{C}}$. According to the Davis-Kahan $\sin\theta$ theorem (Davis and Kahan, 1970), with the number of samples given in the theorem, the top eigenvectors of $\mathbf{C}$ and $\widehat{\mathbf{C}}$ are well aligned:

$$\sin^2\theta \leq \frac{\left\| \mathbf{C} - \widehat{\mathbf{C}} \right\|^2}{\Delta^2} \leq \frac{\epsilon}{64}$$

where θ is the angle between $\mathbf{r}^*$ and $\widehat{\mathbf{r}}$. This implies that

$$\widehat{\mathbf{r}}^\top \mathbf{r}^* = \cos\theta = \sqrt{1 - \sin^2\theta} \geq 1 - \sin^2\theta \geq 1 - \frac{\epsilon}{64}.$$

We now show that the theorem follows if we manage to solve the ERM objective so accurately that

$$\widetilde{\mathbf{r}}^\top \widehat{\mathbf{r}} = \frac{1}{2} \left(\frac{\widehat{\mathbf{u}}^\top \boldsymbol{\Sigma}_{xx} \mathbf{u}_T}{\sqrt{\mathbf{u}_T^\top \boldsymbol{\Sigma}_{xx} \mathbf{u}_T}} + \frac{\widehat{\mathbf{v}}^\top \boldsymbol{\Sigma}_{yy} \mathbf{v}_T}{\sqrt{\mathbf{v}_T^\top \boldsymbol{\Sigma}_{yy} \mathbf{v}_T}} \right) \geq 1 - \frac{\epsilon^2}{8192}. \tag{30}$$

To see this, first observe that (30) implies

$$\|\widetilde{\mathbf{r}} - \widehat{\mathbf{r}}\| = \sqrt{2 - 2(\widetilde{\mathbf{r}}^\top \widehat{\mathbf{r}})} \leq \frac{\epsilon}{64},$$

and as a result

$$\widetilde{\mathbf{r}}^\top \mathbf{r}^* \geq \widehat{\mathbf{r}}^\top \mathbf{r}^* - \left| (\widetilde{\mathbf{r}} - \widehat{\mathbf{r}})^\top \mathbf{r}^* \right| \geq \widehat{\mathbf{r}}^\top \mathbf{r}^* - \|\widetilde{\mathbf{r}} - \widehat{\mathbf{r}}\| \geq 1 - \frac{\epsilon}{32}.$$

Consequently, we have

$$\frac{1}{2}\left(\left\|\mathbf{E}_{xx}^{\frac{1}{2}}\mathbf{u}^* - \mathbf{\Sigma}_{xx}^{\frac{1}{2}}\mathbf{u}_T\right\|^2 + \left\|\mathbf{E}_{yy}^{\frac{1}{2}}\mathbf{v}^* - \mathbf{\Sigma}_{yy}^{\frac{1}{2}}\mathbf{v}_T\right\|^2\right) = \|\widetilde{\mathbf{r}} - \mathbf{r}^*\|^2 = 2\left(1 - \widetilde{\mathbf{r}}^\top \mathbf{r}^*\right) \leq \frac{\epsilon}{16}$$

and so $\max\left(\left\|\mathbf{E}_{xx}^{\frac{1}{2}}\mathbf{u}^* - \mathbf{\Sigma}_{xx}^{\frac{1}{2}}\mathbf{u}_T\right\|^2, \left\|\mathbf{E}_{yy}^{\frac{1}{2}}\mathbf{v}^* - \mathbf{\Sigma}_{yy}^{\frac{1}{2}}\mathbf{v}_T\right\|^2\right) \leq \frac{\epsilon}{8}$. We are now in the same situation as (27); we can fix the incorrect normalization of $\widetilde{\mathbf{r}}$ analogously and then our lemma follows.

It remains to show the time complexity to achieve (30). According to Lemma 22, it suffices to have

$$\cos\theta_T = \frac{1}{\sqrt{2}} \frac{\widehat{\mathbf{u}}^\top \mathbf{\Sigma}_{xx}\mathbf{u}_T + \widehat{\mathbf{v}}^\top \mathbf{\Sigma}_{yy}\mathbf{v}_T}{\sqrt{\mathbf{u}_T^\top \mathbf{\Sigma}_{xx}\mathbf{u}_T + \mathbf{v}_T^\top \mathbf{\Sigma}_{yy}\mathbf{v}_T}} \geq 1 - \frac{\epsilon^2}{2^{15}}. \tag{31}$$

In turn, it suffices to have $|\sin\theta_T| \leq \frac{\epsilon}{256}$ and we ensure it by requiring $G(\mathbf{r}_T) \leq \frac{\epsilon}{256}$. We obtain the stated time complexity by applying Theorem 12 with $\eta = \frac{\epsilon}{256}$.

■

Appendix E. Proofs for Section 5

E.1. Proof of Lemma 15

Proof The desired result is a direct consequence of Lemma 22, by making the identification that

$$\mathbf{a} = \mathbf{E}_{xx}^{\frac{1}{2}}\mathbf{u}^*, \qquad \mathbf{x} = \mathbf{E}_{xx}^{\frac{1}{2}}\mathbf{u}_T, \qquad \mathbf{b} = \mathbf{E}_{yy}^{\frac{1}{2}}\mathbf{v}^*, \qquad \mathbf{y} = \mathbf{E}_{yy}^{\frac{1}{2}}\mathbf{v}_T.$$

■

E.2. Parameters of Streaming SVRG for stochastic least squares

We divide Lemma 16 into the following three propositions.

Proposition 27 (Strong convexity) *For any $\mathbf{w}, \mathbf{w}' \in \mathbb{R}^d$, we have*

$$f_{t+1}(\mathbf{w}) \geq f_{t+1}(\mathbf{w}') + \left\langle \nabla f_{t+1}(\mathbf{w}'), \mathbf{w} - \mathbf{w}' \right\rangle + \frac{\mu}{2}\left\|\mathbf{w} - \mathbf{w}'\right\|^2$$

where $\mu := \frac{\gamma}{\beta_1} \geq C\Delta\gamma$ for some $C > 0$.

Proof Just observe that the Hessian of $f_{t+1}(\mathbf{w})$ is $\mathbf{A}_\lambda = \mathbf{B}^{\frac{1}{2}}\mathbf{M}_\lambda^{-1}\mathbf{B}^{\frac{1}{2}}$, whose eigenvalues are bounded from below: $\sigma_{\min}(\mathbf{A}_\lambda) \geq (\lambda - \rho_1) \cdot \sigma_{\min}(\mathbf{B}) = \gamma/\beta_1$. The lemma follows from the assumption that $\lambda = \rho_1 + c\Delta$ for $c \in (0, 1)$. ■

Proposition 28 (Streaming smoothness) *For any $\mathbf{w} \in \mathbb{R}^d$, we have*

$$\mathbb{E}\left[\left\|\nabla\phi_{t+1}(\mathbf{w}) - \nabla\phi_{t+1}(\mathbf{w}^*_{t+1})\right\|^2\right] \le 2S\left(f_{t+1}(\mathbf{w}) - f^*_{t+1}\right)$$

where $S = \mathcal{O}\left(\frac{d\beta_1}{\gamma}\right) = \mathcal{O}\left(\frac{d}{\Delta\gamma}\right)$ for the sub-Gaussian/regular polynomial-tail classes, and $S = \mathcal{O}\left(\frac{\beta_1}{\gamma}\right) = \mathcal{O}\left(\frac{1}{\Delta\gamma}\right)$ for the bounded class.

Proof Observe that

$$\nabla\phi_{t+1}(\mathbf{w}) = \begin{bmatrix} \lambda\mathbf{x}\mathbf{x}^\top & -\mathbf{x}\mathbf{y}^\top \\ -\mathbf{y}\mathbf{x}^\top & \lambda\mathbf{y}\mathbf{y}^\top \end{bmatrix}\mathbf{w} - \begin{bmatrix} \mathbf{x}\mathbf{x}^\top & \\ & \mathbf{y}\mathbf{y}^\top \end{bmatrix}\mathbf{w}_t.$$

As shown in Lemma 26, this gradient function is Lipschitz continuous:

$$\left\|\nabla\phi_{t+1}(\mathbf{w}) - \nabla\phi_{t+1}(\mathbf{w}^*_{t+1})\right\| \le (\lambda+1)\cdot\sup\left(\|\mathbf{x}\|, \|\mathbf{y}\|\right)\cdot\left\|\mathbf{w} - \mathbf{w}^*_{t+1}\right\|.$$

Note that $\lambda \le \rho_1 + u\Delta$ where $\rho_1 \le 1$, $\Delta \le 1$, and $u < 1$ by assumption, and thus $\lambda \le 2$. As a result, we obtain

$$\mathbb{E}\left\|\nabla\phi_{t+1}(\mathbf{w}) - \nabla\phi_{t+1}(\mathbf{w}^*_{t+1})\right\|^2 \le 9\mathbb{E}\left[\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2\right]\cdot\left\|\mathbf{w} - \mathbf{w}^*_{t+1}\right\|^2.$$

For the distributions of $P(\mathbf{x},\mathbf{y})$ considered here, $\mathbb{E}\|\mathbf{x}\|^2$ and $\mathbb{E}\|\mathbf{y}\|^2$ are both $\mathcal{O}(d)$ for the sub-Gaussian/regular polynomial-tail inputs (see Remark 24), and bounded by 1 for the bounded inputs.

On the other hand, according to Lemma 27, we have

$$f(\mathbf{w}) - f(\mathbf{w}^*_{t+1}) \ge C\Delta\gamma\left\|\mathbf{w} - \mathbf{w}^*_{t+1}\right\|^2$$

for some $C > 0$.

Combining the above two inequalities gives the desired result. ■

Proposition 29 (Streaming variance) *We have*

$$\mathbb{E}\left[\frac{1}{2}\left\|\nabla\phi(\mathbf{w}^*_{t+1})\right\|^2_{\left(\nabla^2 f(\mathbf{w}^*_{t+1})\right)^{-1}}\right] \le \sigma^2.$$

where $\sigma^2 = \mathcal{O}\left(d\beta_1^3\|\mathbf{r}_t\|^2\right)$ for the sub-Gaussian/regular polynomial-tail classes, and $\sigma^2 = \mathcal{O}\left(\frac{\beta_1^3\|\mathbf{r}_t\|^2}{\gamma^2}\right)$ for the bounded class.

Proof Observe that $\mathbf{w}^*_{t+1} = \mathbf{A}_\lambda^{-1}\mathbf{B}\mathbf{w}_t$ and

$$\nabla\phi(\mathbf{w}^*_{t+1}) = \left(\begin{bmatrix} \lambda\mathbf{x}\mathbf{x}^\top & -\mathbf{x}\mathbf{y}^\top \\ -\mathbf{y}\mathbf{x}^\top & \lambda\mathbf{y}\mathbf{y}^\top \end{bmatrix}\mathbf{A}_\lambda^{-1}\mathbf{B} - \begin{bmatrix} \mathbf{x}\mathbf{x}^\top & \\ & \mathbf{y}\mathbf{y}^\top \end{bmatrix}\right)\mathbf{w}_t.$$

Define the shorthands $\mathbf{D} = \mathbf{B}^{-\frac{1}{2}} \begin{bmatrix} \lambda \mathbf{x}\mathbf{x}^\top & -\mathbf{x}\mathbf{y}^\top \\ -\mathbf{y}\mathbf{x}^\top & \lambda \mathbf{y}\mathbf{y}^\top \end{bmatrix} \mathbf{B}^{-\frac{1}{2}}$ and $\mathbf{E} = \mathbf{B}^{-\frac{1}{2}} \begin{bmatrix} \mathbf{x}\mathbf{x}^\top & \\ & \mathbf{y}\mathbf{y}^\top \end{bmatrix} \mathbf{B}^{-\frac{1}{2}}$.

Then we have

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{2}\left\|\nabla\phi(\mathbf{w}_{t+1}^*)\right\|^2_{\left(\nabla^2 f(\mathbf{w}_{t+1}^*)\right)^{-1}}\right] &= \mathbb{E}\left[\frac{1}{2}\left\|\nabla\phi(\mathbf{w}_{t+1}^*)\right\|^2_{\mathbf{A}_\lambda^{-1}}\right] \\
&= \mathbb{E}\left[\frac{1}{2}\left\|\mathbf{B}^{-\frac{1}{2}}\nabla\phi(\mathbf{w}_{t+1}^*)\right\|^2_{\mathbf{B}^{\frac{1}{2}}\mathbf{A}_\lambda^{-1}\mathbf{B}^{\frac{1}{2}}}\right] \\
&= \frac{1}{2}\mathbb{E}\left[\left(\mathbf{B}^{\frac{1}{2}}\mathbf{w}_t\right)^\top (\mathbf{M}_\lambda\mathbf{D} - \mathbf{E})\cdot\mathbf{M}_\lambda\cdot(\mathbf{D}\mathbf{M}_\lambda - \mathbf{E})\left(\mathbf{B}^{\frac{1}{2}}\mathbf{w}_t\right)\right] \\
&= \frac{1}{2}\mathbb{E}\left[\mathbf{r}_t^\top(\mathbf{M}_\lambda\mathbf{D} - \mathbf{E})\cdot\mathbf{M}_\lambda\cdot(\mathbf{D}\mathbf{M}_\lambda - \mathbf{E})\mathbf{r}_t\right]. \qquad (32)
\end{aligned}$$

Bounded case For the bounded case where $\sup\left(\|\mathbf{x}\|^2, \|\mathbf{y}\|^2\right) \le 1$, the derivation is relatively simple. We can bound $\|\mathbf{D}\| \le \frac{3}{\gamma}$ and $\|\mathbf{E}\| \le \frac{1}{\gamma}$, and thus

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{2}\left\|\nabla\phi(\mathbf{w}_{t+1}^*)\right\|^2_{\left(\nabla^2 f(\mathbf{w}_{t+1}^*)\right)^{-1}}\right] &\le \frac{\|\mathbf{M}_\lambda\|\|\mathbf{r}_t\|^2}{2}\mathbb{E}\|\mathbf{D}\mathbf{M}_\lambda - \mathbf{E}\|^2 \\
&\le \beta_1\|\mathbf{r}_t\|^2\left(\mathbb{E}\|\mathbf{D}\mathbf{M}_\lambda\|^2 + \mathbb{E}\|\mathbf{E}\|^2\right) \\
&= \mathcal{O}\left(\frac{\beta_1^3\|\mathbf{r}_t\|^2}{\gamma^2}\right)
\end{aligned}$$

where we have used the triangle inequality and the fact that $(x+y)^2 \le 2x^2 + 2y^2$ in the second inequality.

Sub-Gaussian/regular polynomial tail cases We now omit the subscript λ in $\mathbf{M}_\lambda$ and t from iterates for convenience. Using the fact that $\|\mathbf{x}+\mathbf{y}\|^2 \le 2\|\mathbf{x}\|^2 + 2\|\mathbf{y}\|^2$ with $\mathbf{x} = \mathbf{M}^{\frac{1}{2}}\mathbf{D}\mathbf{M}\mathbf{r}$ and $\mathbf{y} = \mathbf{M}^{\frac{1}{2}}\mathbf{E}\mathbf{r}$, we continue from (32) and obtain

$$\mathbb{E}\left[\frac{1}{2}\left\|\nabla\phi(\mathbf{w}_{t+1}^*)\right\|^2_{\left(\nabla^2 f(\mathbf{w}_{t+1}^*)\right)^{-1}}\right] \le \mathbb{E}\left[\mathbf{r}^\top\mathbf{M}\mathbf{D}\mathbf{M}\mathbf{D}\mathbf{M}\mathbf{r}\right] + \mathbb{E}\left[\mathbf{r}^\top\mathbf{E}\mathbf{M}\mathbf{E}\mathbf{r}\right].$$

Introduce the notation $\mathbf{u} = \mathbf{E}_{xx}^{-\frac{1}{2}}\mathbf{x}$, $\mathbf{v} = \mathbf{E}_{yy}^{-\frac{1}{2}}\mathbf{y}$, and parition $\mathbf{r}$ and $\mathbf{M}$ according to $(\mathbf{x}, \mathbf{y})$:

$$\mathbf{r} = \begin{bmatrix}\mathbf{r}_x \\ \mathbf{r}_y\end{bmatrix}, \qquad \mathbf{M} = \begin{bmatrix}\mathbf{M}_{xx} & \mathbf{M}_{xy} \\ \mathbf{M}_{yx} & \mathbf{M}_{yy}\end{bmatrix}.$$

In view of Lemma 23, we can assume $\max\left(\mathbb{E}\|\mathbf{u}\|^4, \mathbb{E}\|\mathbf{v}\|^4\right) \le Cd^2$. From now on, we use C for a generic constant whose specific value may change between appearances.

We have

$$\mathbf{E}\mathbf{M}\mathbf{E} = \begin{bmatrix}\mathbf{u}\mathbf{u}^\top\mathbf{M}_{xx}\mathbf{u}\mathbf{u}^\top & \mathbf{u}\mathbf{u}^\top\mathbf{M}_{xy}\mathbf{v}\mathbf{v}^\top \\ \mathbf{v}\mathbf{v}^\top\mathbf{M}_{yx}\mathbf{u}\mathbf{u}^\top & \mathbf{v}\mathbf{v}^\top\mathbf{M}_{yy}\mathbf{v}\mathbf{v}^\top\end{bmatrix}.$$

and thus

$$\mathbf{r}^\top\mathbf{E}\mathbf{M}\mathbf{E}\mathbf{r} = \mathbf{r}_x^\top\mathbf{u}\mathbf{u}^\top\mathbf{M}_{xx}\mathbf{u}\mathbf{u}^\top\mathbf{r}_x + \mathbf{r}_y^\top\mathbf{v}\mathbf{v}^\top\mathbf{M}_{yy}\mathbf{v}\mathbf{v}^\top\mathbf{r}_y + 2\mathbf{r}_x^\top\mathbf{u}\mathbf{u}^\top\mathbf{M}_{xy}\mathbf{v}\mathbf{v}^\top\mathbf{r}_y.$$

We have for the first term that

$$\begin{aligned}\mathbb{E}\left[\mathbf{r}_x^\top \mathbf{u}\mathbf{u}^\top \mathbf{M}_{xx}\mathbf{u}\mathbf{u}^\top \mathbf{r}_x\right] &= \mathbb{E}\left[\left|\mathbf{r}_x^\top \mathbf{u}\right|^2 \mathbf{u}^\top \mathbf{M}_{xx}\mathbf{u}\right] \\ &\le \sqrt{\mathbb{E}\left|\mathbf{r}_x^\top \mathbf{u}\right|^4}\sqrt{\mathbb{E}\left|\mathbf{u}^\top \mathbf{M}_{xx}\mathbf{u}\right|^2} \\ &\le C\left\|\mathbf{r}_x\right\|^2 \sqrt{\mathbb{E}\left|\mathbf{u}^\top \mathbf{M}_{xx}\mathbf{u}\right|^2} \\ &\le C\left\|\mathbf{M}_{xx}\right\|\left\|\mathbf{r}_x\right\|^2 \sqrt{\mathbb{E}\left\|\mathbf{u}\right\|^4} \\ &\le C\left\|\mathbf{M}\right\|\left\|\mathbf{r}_x\right\|^2 d.\end{aligned}$$

Similar arguments also lead to

$$\mathbb{E}\left[\mathbf{r}_y^\top \mathbf{v}\mathbf{v}^\top \mathbf{M}_{yy}\mathbf{v}\mathbf{v}^\top \mathbf{r}_y\right] \le C\left\|\mathbf{M}\right\|\left\|\mathbf{r}_y\right\|^2 d.$$

For the third term, we have

$$\begin{aligned}\mathbb{E}\left[\mathbf{r}_x^\top \mathbf{u}\mathbf{u}^\top \mathbf{M}_{xy}\mathbf{v}\mathbf{v}^\top \mathbf{r}_y\right] &\le \sqrt{\mathbb{E}\left|\mathbf{r}_x^\top \mathbf{u}\right|^2\left|\mathbf{r}_y^\top \mathbf{v}\right|^2}\sqrt{\mathbb{E}\left|\mathbf{u}^\top \mathbf{M}_{xy}\mathbf{v}\right|^2} \\ &\le \left\|\mathbf{M}\right\|\left(\mathbb{E}\left|\mathbf{r}_x^\top \mathbf{u}\right|^4\right)^{\frac{1}{4}}\left(\mathbb{E}\left|\mathbf{r}_y^\top \mathbf{v}\right|^4\right)^{\frac{1}{4}}\left(\mathbb{E}\left\|\mathbf{u}\right\|^4\right)^{\frac{1}{4}}\left(\mathbb{E}\left\|\mathbf{v}\right\|^4\right)^{\frac{1}{4}} \\ &\le C\left\|\mathbf{M}\right\|\left\|\mathbf{r}_x\right\|\left\|\mathbf{r}_y\right\| d.\end{aligned}$$

Therefore,

$$\mathbb{E}\left[\mathbf{r}^\top \mathbf{E}\mathbf{M}\mathbf{E}\mathbf{r}\right] \le C\left\|\mathbf{M}\right\|\left\|\mathbf{r}\right\|^2 d.$$

Now we need to bound $\mathbb{E}\left[\mathbf{r}^\top \mathbf{M}\mathbf{D}\mathbf{M}\mathbf{D}\mathbf{M}\mathbf{r}\right]$. Using the fact that $\left\|\mathbf{x}+\mathbf{y}\right\|^2 \le 2\left\|\mathbf{x}\right\|^2 + 2\left\|\mathbf{y}\right\|^2$ with $\mathbf{x} = \mathbf{M}^{\frac{1}{2}}\mathbf{D}_1\mathbf{M}\mathbf{r}$ and $\mathbf{y} = \mathbf{M}^{\frac{1}{2}}\mathbf{D}_2\mathbf{M}\mathbf{r}$, this can be bounded by two terms:

$$\mathbb{E}\left[\mathbf{r}^\top \mathbf{M}\mathbf{D}\mathbf{M}\mathbf{D}\mathbf{M}\mathbf{r}\right] \le 2\mathbb{E}\left[\mathbf{r}^\top \mathbf{M}\mathbf{D}_1\mathbf{M}\mathbf{D}_1\mathbf{M}\mathbf{r}\right] + 2\mathbb{E}\left[\mathbf{r}^\top \mathbf{M}\mathbf{D}_2\mathbf{M}\mathbf{D}_2\mathbf{M}\mathbf{r}\right]$$

where

$$\mathbf{D}_1 = \lambda\begin{bmatrix}\mathbf{u}\mathbf{u}^\top & 0 \\ 0 & \mathbf{v}\mathbf{v}^\top\end{bmatrix}, \qquad \mathbf{D}_2 = -\begin{bmatrix}0 & \mathbf{u}\mathbf{v}^\top \\ \mathbf{v}\mathbf{u}^\top & 0\end{bmatrix}.$$

The bound for $\mathbb{E}\left[\mathbf{r}^\top \mathbf{M}\mathbf{D}_1\mathbf{M}\mathbf{D}_1\mathbf{M}\mathbf{r}\right]$ can be derived using the same argument that bounds $\mathbb{E}\left[\mathbf{r}^\top \mathbf{E}\mathbf{M}\mathbf{E}\mathbf{r}\right]$ (now $\mathbf{M}\mathbf{r}$ plays the role of $\mathbf{r}$ in bounding $\mathbb{E}\left[\mathbf{r}^\top \mathbf{E}\mathbf{M}\mathbf{E}\mathbf{r}\right]$), and thus we have

$$\mathbb{E}\left[\mathbf{r}^\top \mathbf{M}\mathbf{D}_1\mathbf{M}\mathbf{D}_1\mathbf{M}\mathbf{r}\right] \le C\lambda^2\left\|\mathbf{M}\right\|\left\|\mathbf{M}\mathbf{r}\right\|^2 d \le C\left\|\mathbf{M}\right\|^3\left\|\mathbf{r}\right\|^2\lambda^2 d.$$

Finally, we bound $\mathbb{E}\left[\mathbf{r}^\top \mathbf{M}\mathbf{D}_2\mathbf{M}\mathbf{D}_2\mathbf{M}\mathbf{r}\right]$. Note that

$$-\mathbf{D}_2\mathbf{M}\mathbf{D}_2 = \begin{bmatrix}\mathbf{u}\mathbf{v}^\top \mathbf{M}_{yy}\mathbf{v}\mathbf{u}^\top & \mathbf{u}\mathbf{v}^\top \mathbf{M}_{yx}\mathbf{u}\mathbf{v} \\ \mathbf{v}\mathbf{u}^\top \mathbf{M}_{xy}\mathbf{v}\mathbf{u}^\top & \mathbf{v}\mathbf{u}^\top \mathbf{M}_{xx}\mathbf{u}\mathbf{v}^\top\end{bmatrix}.$$

Let

$$\mathbf{Mr} = \begin{bmatrix} \mathbf{m}_x \\ \mathbf{m}_y \end{bmatrix},$$

and then

$$-\mathbf{r}^\top \mathbf{M}\mathbf{D}_2\mathbf{M}\mathbf{D}_2\mathbf{M}\mathbf{r} = \mathbf{m}_x^\top \mathbf{u}\mathbf{v}^\top \mathbf{M}_{yy}\mathbf{v}\mathbf{u}^\top \mathbf{m}_x + \mathbf{m}_y^\top \mathbf{v}\mathbf{u}^\top \mathbf{M}_{xx}\mathbf{u}\mathbf{v}^\top \mathbf{m}_y + 2\mathbf{m}_x^\top \mathbf{u}\mathbf{v}^\top \mathbf{M}_{yx}\mathbf{u}\mathbf{v}^\top \mathbf{m}_y.$$

Similarly to what we have done above,

$$\begin{aligned} \mathbb{E}\left|\mathbf{m}_x^\top \mathbf{u}\mathbf{v}^\top \mathbf{M}_{yy}\mathbf{v}\mathbf{u}^\top \mathbf{m}_x\right| &\leq \sqrt{\mathbb{E}\left|\mathbf{m}_x^\top \mathbf{u}\right|^4}\sqrt{\mathbb{E}\left|\mathbf{v}^\top \mathbf{M}_{yy}\mathbf{v}\right|^2} \\ &\leq C\left\|\mathbf{M}\right\|\left\|\mathbf{m}_x\right\|^2 d \\ &\leq C\left\|\mathbf{M}\right\|^3\left\|\mathbf{r}\right\|^2 d. \end{aligned}$$

The same bound also holds for $\mathbb{E}\left|\mathbf{m}_y^\top \mathbf{v}\mathbf{u}^\top \mathbf{M}_{xx}\mathbf{u}\mathbf{v}^\top \mathbf{m}_y\right|$ with the same argument. For the term $\mathbb{E}\left|\mathbf{m}_x^\top \mathbf{u}\mathbf{v}^\top \mathbf{M}_{yx}\mathbf{u}\mathbf{v}^\top \mathbf{m}_y\right|$, we have

$$\begin{aligned} \mathbb{E}\left|\mathbf{m}_x^\top \mathbf{u}\mathbf{v}^\top \mathbf{M}_{yx}\mathbf{u}\mathbf{v}^\top \mathbf{m}_y\right| &\leq \left\|\mathbf{M}\right\|\left(\mathbb{E}\left|\mathbf{m}_x\mathbf{u}\right|^4\right)^{\frac{1}{4}}\left(\mathbb{E}\left|\mathbf{m}_y\mathbf{v}\right|^4\right)^{\frac{1}{4}}\left(\mathbb{E}\left\|\mathbf{u}\right\|^4\right)^{\frac{1}{4}}\left(\mathbb{E}\left\|\mathbf{v}\right\|^4\right)^{\frac{1}{4}} \\ &\leq C\left\|\mathbf{M}\right\|\left\|\mathbf{m}_x\right\|\left\|\mathbf{m}_y\right\| d \\ &\leq C\left\|\mathbf{M}\right\|^3\left\|\mathbf{r}\right\|^2 d. \end{aligned}$$

Combining all the terms, and noting that $\lambda \leq 2$, we have shown that

$$\mathbb{E}\left[\mathbf{r}^\top \mathbf{M}\mathbf{D}\mathbf{M}\mathbf{D}\mathbf{M}\mathbf{r}\right] \leq C\left\|\mathbf{M}\right\|^3\left\|\mathbf{r}\right\|^2 d.$$

And the final bound is

$$\mathbb{E}\left[\mathbf{r}^\top(\mathbf{M}\mathbf{D}-\mathbf{E})\mathbf{M}(\mathbf{D}\mathbf{M}-\mathbf{E})\mathbf{r}\right] \leq C\left[\left\|\mathbf{M}\right\|^3 + \left\|\mathbf{M}\right\|\right]\left\|\mathbf{r}\right\|^2 d = \mathcal{O}\left(\beta_1^3\left\|\mathbf{r}\right\|^2 d\right).$$

■

E.3. Proof of Lemma 17

Proof For notational simplicity, we omit the subscript $t+1$ below.

According to Frostig et al. (2015, Theorem 4.1), we have that for iteration τ of Algorithm 1

$$\begin{aligned} \mathbb{E}\left[f(\mathbf{w}^\tau) - f^*\right] \leq \frac{1}{1-4s}\Bigg[\left(\frac{S}{\mu m_\tau s} + 4s\right)\mathbb{E}\left[f(\mathbf{w}^{\tau-1}) - f^*\right] \\ + \frac{1+2s}{k_\tau}\left(\sqrt{\frac{S}{\mu}\mathbb{E}\left[f(\mathbf{w}^{\tau-1}) - f^*\right]} + \sigma\right)^2\Bigg]. \end{aligned} \tag{33}$$

Using the inequality $(x+y)^2 \le 2(x^2+y^2)$, it holds that

$$\left(\sqrt{\frac{S}{\mu}\mathbb{E}\left[f(\mathbf{w}^{\tau-1})-f^*\right]}+\sigma\right)^2 \le \frac{2S}{\mu}\mathbb{E}\left[f(\mathbf{w}^{\tau-1})-f^*\right]+2\sigma^2.$$

Now, set for this iteration $s=\frac{c_2}{8}$, $m_\tau=\lceil\frac{S}{\mu c_2^2}\rceil$, and $k_\tau=\max\left(\lceil\frac{S}{\mu c_2}\rceil,\lceil\frac{\sigma^2}{\beta_1\|\mathbf{r}_t\|^2 c_3}\rceil\right)$, for some $c_2,c_3\in(0,1)$. We continue from (33) and have

$$\begin{aligned}\mathbb{E}\left[f(\mathbf{w}^\tau)-f^*\right] &\le \frac{1}{1-4s}\left[\left(\frac{S}{\mu m_\tau s}+4s+\frac{2+4s}{k_\tau}\frac{S}{\mu}\right)\mathbb{E}\left[f(\mathbf{w}^{\tau-1})-f^*\right]+\frac{2+4s}{k_\tau}\sigma^2\right]\\ &\le \frac{1}{1-c_2/2}\left[\left(8c_2+\frac{c_2}{2}+\frac{2+4c_2}{2}c_2\right)\mathbb{E}\left[f(\mathbf{w}^{\tau-1})-f^*\right]+\frac{4+c_2}{2k_\tau}\sigma^2\right]\\ &\le 22c_2\cdot\mathbb{E}\left[f(\mathbf{w}^{\tau-1})-f^*\right]+10c_3\cdot\frac{\beta_1\left\|\mathbf{r}_t\right\|^2}{2}.\end{aligned}$$

We can now calculate the number of samples used in this iteration, which is

$$k_\tau+m_\tau=\mathcal{O}\left(\frac{d\beta_1^2}{c_3}+\frac{d\beta_1^2}{\gamma^2c_2^2}\right)=\mathcal{O}\left(\frac{d}{\Delta^2c_3}+\frac{d}{\Delta^2\gamma^2c_2^2}\right) \tag{34}$$

for sub-Gaussian/regular polynomial-tail inputs, and

$$k_\tau+m_\tau=\mathcal{O}\left(\frac{\beta_1^2}{\gamma^2c_3}+\frac{\beta_1^2}{\gamma^2c_2^2}\right)=\mathcal{O}\left(\frac{1}{\Delta^2\gamma^2c_3}+\frac{1}{\Delta^2\gamma^2c_2^2}\right) \tag{35}$$

for bounded inputs.

Let us fix $c_2=\frac{1}{44}$ for $\tau=1,\ldots,\Gamma$. In view of our initialization strategy (16), setting $c_3=\frac{1}{20}$ for $\tau=1$ gives $\mathbb{E}\left[f(\mathbf{w}^1)-f^*\right]\le\frac{\beta_1\|\mathbf{r}_t\|^2}{2}$. Afterwards, we halve c_3 at each outer loop $\tau=2,\ldots$, and this makes sure that $\mathbb{E}\left[f(\mathbf{w}^\tau)-f^*\right]\le\frac{\beta_1\|\mathbf{r}_t\|^2}{2^\tau}$.

To achieve the desired accuracy, we need $\Gamma=\mathcal{O}\left(\log\frac{1}{\eta_t}\right)$ outer iterations. Summing (34) and (34) over $\tau=1,\ldots,\Gamma$, and noting $\sum_{\tau=1}^{\Gamma}2^{\tau-1}=\mathcal{O}\left(\frac{1}{\eta_t}\right)$, the total sample complexity is

$$\mathcal{O}\left(\frac{d}{\Delta^2}\cdot 20\sum_{\tau=1}^{\Gamma}2^{\tau-1}+\frac{44^2d}{\Delta^2\gamma^2}\cdot\log\frac{1}{\eta_t}\right)=\mathcal{O}\left(\frac{d}{\Delta^2\eta_t}+\frac{d}{\Delta^2\gamma^2}\log\frac{1}{\eta_t}\right)$$

for sub-Gaussian/regular polynomial-tail inputs, and

$$\mathcal{O}\left(\frac{1}{\Delta^2\gamma^2}\left(20\sum_{\tau=1}^{\Gamma}2^{\tau-1}+44^2\cdot\log\frac{1}{\eta_t}\right)\right)=\mathcal{O}\left(\frac{1}{\Delta^2\gamma^2\eta_t}\right)$$

for bounded inputs (we have dropped the second term since $\log\frac{1}{\eta_t}$ is of lower order compared with $\frac{1}{\eta_t}$). ■

E.4. Proof of Theorem 18

Proof Recall that our streaming CCA algorithm performs shift-and-invert power iterations on the population matrices directly. Following the same argument in the ERM case in Corollary 13, as long as each least squares objective is solved to sufficient accuracy, i.e.,

$$\frac{f_{t+1}(\mathbf{w}_{t+1}) - f_{t+1}^*}{\mathbf{w}_t^\top \widehat{\mathbf{B}} \mathbf{w}_t} \leq \min\left(\sum_{i=2}^{d} \xi_{ti}^2/\beta_i,\ \xi_{t1}^2/\beta_1\right) \cdot \frac{(\beta_1 - \beta_2)^2}{32}, \tag{36}$$

the algorithm converges linearly, and therefore we only need to solve $T = \mathcal{O}\left(\log \frac{1}{\epsilon}\right)$ linear systems. But due to the zero initialization we use in the online setting, the ratio between initial error and final error for each f_{t+1} is different from the offline setting.

When $G(\mathbf{r}_t) > 1$, we are in the regime where $\sum_{i=2}^{d} \xi_{ti}^2/\beta_i \geq \xi_{t1}^2/\beta_1$, and we can ensure the sufficient accuracy in (36) by setting the ratio between the initial and the final error to be

$$\eta_t = \frac{(\beta_1 - \beta_2)^2 \left(\xi_{t1}^2\right)}{16\beta_1^2}$$

in Lemma 17. Since $\frac{\beta_1^2}{(\beta_1-\beta_2)^2} \leq 4$, this implies that

$$\frac{1}{\eta_t} \leq \frac{64}{\cos^2\theta_t} = 64(1+\tan^2\theta_t) \leq 64\left(1 + \frac{\beta_2}{\beta_1} G^2(\mathbf{r}_t)\right) \leq 64\left(1 + G^2(\mathbf{r}_t)\right) \leq 64\left(1 + G^2(\mathbf{r}_0)\right).$$

Note that the sample complexity of this phase does not depend on the final accuracy in alignment.

When $G(\mathbf{r}_t) \leq 1$, indicating that we are in the converging regime where $\sum_{i=2}^{d} \xi_{ti}^2/\beta_i \leq \xi_{t1}^2/\beta_1$, we can ensure the sufficient accuracy in (36) by setting

$$\eta_t = \frac{(\beta_1 - \beta_2)^2 \left(\sum_{i=2}^{d} \xi_{ti}^2\right)}{16\beta_1^2}$$

in Lemma 17. This implies that

$$\frac{1}{\eta_t} \leq \frac{64}{\sin^2\theta_t}.$$

Our goal is to have $\sin^2\theta_T \leq \frac{\epsilon}{4}$, as this implies $\cos\theta_T = \sqrt{1-\sin^2\theta_T} \geq 1 - \sin^2\theta_T \geq 1 - \frac{\epsilon}{4}$, and by Lemma 15 this further implies $\operatorname{align}((\mathbf{u}_T, \mathbf{v}_T); (\mathbf{u}^*, \mathbf{v}^*)) \geq 1 - \epsilon$ as desired. Since $\sin^2\theta_t \leq G^2(\mathbf{r}_t)$, and we have shown that $G^2(\mathbf{r}_t)$ decreases at a geometric rate, we can bound $\frac{1}{\sin^2\theta_t}$ by a geometrically increasing series where the last term is $\frac{4}{\epsilon}$, and the sum of the truncated series up to time T is of the same order of the last term, i.e., $\sum_{t=1}^{T} \frac{1}{\eta_t} = \mathcal{O}\left(\frac{1}{\epsilon}\right)$.

And the theorem follows from Lemma 17, by summing the sample complexity of least squares problems over the outer shift-and-invert iterations.

We remark that to achieve the result with probability $1-\delta$, we require each least squares problem to be solved to the desired accuracy with failure probability $\delta/\log(1/\epsilon)$ (using the Markov inequality) and finally apply the union bound. This would only cause additional $\log(1/\epsilon)$ factors in the total sample complexity. ■

Algorithm 2 Non-uniform sampling SVRG for optimizing finite-sum of nonconvex components $F(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n f_i(\mathbf{w})$.

Input: Stepsize s.
  Initialize $\mathbf{w}_0 \in \mathbb{R}^d$.
  for $j = 1, 2, \ldots, M$ **do**
    $\tilde{\mathbf{u}} \leftarrow \mathbf{w}_{j-1}$
    Evaluate the batch gradient $\nabla F(\tilde{\mathbf{u}}) = \frac{1}{n}\sum_{i=1}^n \nabla f_i(\tilde{\mathbf{u}})$
    $\mathbf{u}_0 \leftarrow \mathbf{w}_{j-1}$
    for $t = 1, 2, \ldots, m$ **do**
      Randomly pick i_t from $\{1, \ldots, n\}$ with probability $\{p_i\}_{i=1}^n$.
      $\mathbf{u}_t \leftarrow \mathbf{u}_{t-1} - s\left(\frac{\nabla f_{i_t}(\mathbf{u}_{t-1}) - \nabla f_{i_t}(\tilde{\mathbf{u}})}{p_{i_t} n} + \nabla F(\tilde{\mathbf{u}})\right)$
    end for
    $\mathbf{w}_j \leftarrow \frac{1}{n}\sum_{t=1}^m \mathbf{w}_t$
  end for
Output: $\mathbf{w}_M$ is the approximate solution.

Appendix F. SVRG with non-uniform sampling for finite-sum of nonconvex components

In this section, we show that for optimizing a convex objective that is the finite-sum of nonconvex components, sampling each components with probability proportional to its smoothness parameter, as shown in Algorithm 2, leads to improved convergence rate. In particular, the final time complexity depends on average smoothness parameter rather than the maximum smoothness.

Lemma 30 *Let $F(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n f_i(\mathbf{w})$, where $F(\mathbf{w})$ is μ-strongly convex, and each component $f_i(\mathbf{w})$ is L_i-smooth. Let $\mathbf{w}^* = \arg\min_{\mathbf{w}} F(\mathbf{w})$. In the inner loop of Algorithm 2, sample i_t using weighted sampling probability $\{p_i\}_{i=1}^n$ from $\{1, \ldots, n\}$ where $p_i = \frac{L_i^2}{\sum_{j=1}^n L_j^2}$, and set $s = \frac{2\mu n}{11\sum_{i=1}^n L_i^2}$, $m = \frac{121\sum_{i=1}^n L_i^2}{8n\mu^2}$. Then the iteration complexity (number of vector operations) to reach ϵ-suboptimality is*

$$\mathcal{O}\left(\left(n + \frac{\sum_{i=1}^n L_i^2}{n\mu^2}\right)\log\frac{(\frac{1}{n}\sum_{i=1}^n L_i)\cdot(F(\mathbf{w}_0) - F(\mathbf{w}^*))}{\mu\varepsilon}\right).$$

Proof For the inner loop of Algorithm 2, we are performing updates of the following form:

$$\mathbf{u}_t \leftarrow \mathbf{u}_{t-1} - s\mathbf{v}_t,$$

where

$$\mathbf{v}_t = \frac{\nabla f_{i_t}(\mathbf{u}_{t-1}) - \nabla f_{i_t}(\tilde{\mathbf{u}})}{p_{i_t} n} + \nabla F(\tilde{\mathbf{u}}).$$

Taking expectation over the random choice of component i_t, we have

$$\mathbb{E}_t[\mathbf{v}_t] = \nabla F(\mathbf{u}_{t-1})$$

We now upper bound the variance of $\mathbf{v}_t$:

$$\begin{aligned}
\mathbb{E}_t \left\|\mathbf{v}_t - \nabla F(\mathbf{u}_{t-1})\right\|^2 =& \mathbb{E}_t \left[\frac{\nabla f_{i_t}(\mathbf{u}_{t-1}) - \nabla f_{i_t}(\tilde{\mathbf{u}})}{p_{i_t} n} + \nabla F(\tilde{\mathbf{u}}) - \nabla F(\mathbf{u}_{t-1})\right]^2 \\
=& \mathbb{E}_t \left[\frac{1}{(p_{i_t} n)^2} \left\|\nabla f_{i_t}(\mathbf{u}_{t-1}) - \nabla f_{i_t}(\tilde{\mathbf{u}})\right\|^2\right] - \left\|\nabla F(\mathbf{u}_{t-1}) - \nabla F(\tilde{\mathbf{u}})\right\|^2 \\
\leq& \frac{1}{n} \sum_{i=1}^n \frac{1}{p_i n} \left\|\nabla f_i(\mathbf{u}_{t-1}) - \nabla f_i(\tilde{\mathbf{u}})\right\|^2 \\
\leq& \frac{2}{n} \sum_{i=1}^n \frac{1}{p_i n} \left(\left\|\nabla f_i(\mathbf{u}_{t-1}) - \nabla f_i(\mathbf{w}^*)\right\|^2 + \left\|\nabla f_i(\tilde{\mathbf{u}}) - \nabla f_i(\mathbf{w}^*)\right\|^2\right) \\
\leq& \frac{2}{n} \sum_{i=1}^n \frac{L_i^2}{p_i n} \left(\left\|\mathbf{u}_{t-1} - \mathbf{w}^*\right\|^2 + \left\|\tilde{\mathbf{u}} - \mathbf{w}^*\right\|^2\right)
\end{aligned}$$

where we have used the fact that $\mathbb{E}\left\|\mathbf{x} - \mathbb{E}[\mathbf{x}]\right\|^2 = \mathbb{E}\left\|\mathbf{x}\right\|^2 - (\mathbb{E}[\mathbf{x}])^2$ for a random vector $\mathbf{x}$ in the second equality, and that $\left\|\mathbf{x}+\mathbf{y}\right\|^2 \leq 2\left\|\mathbf{x}\right\|^2 + 2\left\|\mathbf{y}\right\|^2$ in the second inequality.

By choosing

$$p_i = \frac{L_i^2}{\sum_{i=1}^n L_i^2},$$

the above inequality becomes

$$\mathbb{E}_t \left\|\mathbf{v}_t - \nabla F(\mathbf{u}_{t-1})\right\|^2 \leq \frac{2\sum_{i=1}^n L_i^2}{n} \left(\left\|\mathbf{u}_{t-1} - \mathbf{w}^*\right\|^2 + \left\|\tilde{\mathbf{u}} - \mathbf{w}^*\right\|^2\right).$$

Define $\bar{L} = \frac{1}{n}\sum_{i=1}^n L_i$ which is an upper bound of the smoothness parameter of the average function $F(\mathbf{w})$ as

$$\begin{aligned}
F(\mathbf{a}) - F(\mathbf{b}) = \frac{1}{n}\sum_{i=1}^n f_i(\mathbf{a}) - f_i(\mathbf{b}) \leq& \frac{1}{n}\sum_{i=1}^n \left\langle \nabla f_i(\mathbf{b}), \mathbf{a} - \mathbf{b}\right\rangle + \frac{L_i}{2}\left\|\mathbf{a} - \mathbf{b}\right\|^2 \\
=& \left\langle \nabla F(\mathbf{b}), \mathbf{a} - \mathbf{b}\right\rangle + \frac{\frac{1}{n}\sum_{i=1}^n L_i}{2}\left\|\mathbf{a} - \mathbf{b}\right\|^2,
\end{aligned}$$

and define $\hat{L} = \sqrt{\frac{1}{n}\sum_{i=1}^n L_i^2}$. We then bound the distance from each iterate to the optimum:

$$\begin{aligned}
&\mathbb{E}_t \left\|\mathbf{u}_t - \mathbf{w}^*\right\|^2 = \left\|\mathbf{u}_{t-1} - \mathbf{w}^*\right\|^2 - 2s\left\langle \mathbf{u}_{t-1} - \mathbf{w}^*, \mathbb{E}_t\left[\mathbf{v}_t\right]\right\rangle + s^2 \mathbb{E}_t \left\|\mathbf{v}_t\right\|^2 \\
\leq& \left\|\mathbf{u}_{t-1} - \mathbf{w}^*\right\|^2 - 2s\left\langle \mathbf{u}_{t-1} - \mathbf{w}^*, \nabla F(\mathbf{u}_{t-1})\right\rangle + s^2 \left\|\nabla F(\mathbf{u}_{t-1})\right\|^2 + 2s^2\hat{L}^2\left(\left\|\mathbf{u}_{t-1} - \mathbf{w}^*\right\|^2 + \left\|\tilde{\mathbf{u}} - \mathbf{w}^*\right\|^2\right) \\
\leq& \left\|\mathbf{u}_{t-1} - \mathbf{w}^*\right\|^2 - 2s\mu\left\|\mathbf{u}_{t-1} - \mathbf{w}^*\right\|^2 + s^2\bar{L}^2\left\|\mathbf{u}_{t-1} - \mathbf{w}^*\right\|^2 + 2s^2\hat{L}^2\left(\left\|\mathbf{u}_{t-1} - \mathbf{w}^*\right\|^2 + \left\|\tilde{\mathbf{u}} - \mathbf{w}^*\right\|^2\right)
\end{aligned}$$

where we have used the fact that $\mathbb{E}\left\|\mathbf{x}\right\| = (\mathbb{E}[\mathbf{x}])^2 + \mathbb{E}\left\|\mathbf{x} - \mathbb{E}[\mathbf{x}]\right\|^2$ in the first inequality, and the smoothness and strong convexity of $F(\mathbf{w})$ in the second inequality.

By the Jensen's inequality, we have $\bar{L} \leq \hat{L}$. Therefore, we continue from above and obtain

$$\mathbb{E}_t\left[\left\|\mathbf{u}_t - \mathbf{w}^*\right\|^2\right] - \mathbb{E}\left[\left\|\mathbf{u}_{t-1} - \mathbf{w}^*\right\|^2\right] \leq \left(3s^2\hat{L}^2 - 2s\mu\right)\left\|\mathbf{u}_{t-1} - \mathbf{w}^*\right\|^2 + 2s^2\hat{L}^2\left\|\tilde{\mathbf{u}} - \mathbf{w}^*\right\|^2.$$

Summing the above inequality over the inner loop yields

$$\mathbb{E}\left\|\mathbf{u}_m - \mathbf{w}^*\right\|^2 - \mathbb{E}\left\|\mathbf{u}_0 - \mathbf{w}^*\right\|^2 \leq \left(3s^2\hat{L}^2 - 2s\mu\right)\sum_{t=1}^{m}\left\|\mathbf{u}_{t-1} - \mathbf{w}^*\right\|^2 + 2ms^2\hat{L}^2\left\|\tilde{\mathbf{u}} - \mathbf{w}^*\right\|^2$$

Using $\mathbf{u}_0 = \tilde{\mathbf{u}}$ and rearranging terms, we have

$$\left(2s\mu - 3s^2\hat{L}^2\right)\sum_{t=1}^{m}\left\|\mathbf{u}_{t-1} - \mathbf{w}^*\right\|^2 \leq \left(1 + 2ms^2\hat{L}^2\right)\left\|\tilde{\mathbf{u}} - \mathbf{w}^*\right\|^2$$

Using $\tilde{\mathbf{u}} = \mathbf{w}_{j-1}$ and $\mathbf{w}_j = \frac{1}{m}\sum_{t=0}^{m-1}\mathbf{u}_t$, we obtain

$$\mathbb{E}\left\|\mathbf{w}_j - \mathbf{w}^*\right\|^2 \leq \frac{1 + 2ms^2\hat{L}^2}{2ms\mu - 3ms^2\hat{L}^2}\mathbb{E}\left\|\mathbf{w}_{j-1} - \mathbf{w}^*\right\|^2.$$

Setting

$$s = \frac{2\mu}{11\hat{L}^2}, \qquad m = \frac{1}{2s^2\hat{L}^2} = \frac{121\hat{L}^2}{8\mu^2},$$

we obtain

$$\mathbb{E}\left\|\mathbf{w}_j - \mathbf{w}^*\right\|^2 \leq \frac{1}{2}\mathbb{E}\left\|\mathbf{w}_{j-1} - \mathbf{w}^*\right\|^2.$$

Therefore the squared distance to minimum decreases geometrically for each outer loop. After M iterations, we have

$$F(\mathbf{w}_M) - F(\mathbf{w}^*) \leq \frac{\bar{L}}{2}\left\|\mathbf{w}_M - \mathbf{w}^*\right\|^2 \leq \frac{\bar{L}}{2}\left(\frac{1}{2}\right)^M\left\|\mathbf{w}_0 - \mathbf{w}^*\right\|^2 \leq \frac{\bar{L}}{\mu}\left(\frac{1}{2}\right)^M(F(\mathbf{w}_0) - F(\mathbf{w}^*)).$$

Setting the right hand side to ε gives the number of outer iterations $M = \mathcal{O}\left(\log\frac{(\bar{L}/\mu)\cdot(F(\mathbf{w}_0)-F(\mathbf{w}^*))}{\varepsilon}\right)$. Finally, the total iteration complexity to reach ε-suboptimality is

$$\mathcal{O}\left((n+m)M\right) = \mathcal{O}\left(\left(n + \frac{\hat{L}^2}{\mu^2}\right)\log\frac{\bar{L}(F(\mathbf{w}_0) - F(\mathbf{w}^*))}{\mu\varepsilon}\right).$$

■