Stochastic Primal-Dual Approaches for Efficient Nonconvex Optimization in Multiview Learning

Chris Junchi Li[◇]

Department of Electrical Engineering and Computer Sciences[◇]
University of California, Berkeley

October 7, 2024

Abstract

In this paper, we propose a novel framework for efficient nonconvex optimization in scalable multiview learning. Leveraging stochastic gradient-based methods, we address computational challenges in Partial Least Squares (PLS) and Canonical Correlation Analysis (CCA), two essential techniques for extracting shared latent structures from high-dimensional, multimodal data. By exploring the Lagrangian landscape, we extend our analysis to constrained nonconvex optimization problems, introducing a stochastic primal-dual algorithm that guarantees convergence to global optima. Our algorithm operates within an online setting, tackling the generalized eigenvalue problem (GEV) with stochastic approximations of the problem's parameters. The theoretical foundations of our method are reinforced by diffusion approximations, showing significant improvements in sample complexity and computational efficiency. Finally, we present numerical evaluations to validate the empirical performance of our approach.

Keywords: Nonconvex optimization, multiview learning, stochastic gradient descent, partial least squares, canonical correlation analysis, Lagrangian dynamics, stochastic primal-dual algorithms

1 Introduction

The rapid growth of data across various fields such as computer vision, natural language processing, and bioinformatics has led to an increasing demand for efficient multiview learning techniques. Multiview learning focuses on analyzing datasets where multiple sets of features (or "views") describe the same underlying phenomena, such as images paired with captions or audio recordings paired with transcriptions. A central challenge in this domain is extracting shared latent structures that capture correlations between views, while maintaining scalability for high-dimensional and multimodal datasets.

Traditional approaches to multiview learning have primarily relied on convex optimization techniques like Sample Average Approximation (SAA). While effective for smaller datasets, these methods often struggle with scalability in big data settings due to their reliance on offline, batch processing. In response, nonconvex optimization techniques—particularly those based on stochastic approximation—have emerged as promising alternatives, offering improved computational efficiency and superior empirical performance in large-scale scenarios.

In this work, we propose a novel nonconvex optimization framework for solving multiview learning problems, with a focus on Partial Least Squares (PLS) and Canonical Correlation Analysis (CCA). These are two prominent methods for extracting latent structures that maximize covariance between multiple views. By formulating these problems as stochastic optimization tasks, we develop a stochastic primal-dual algorithm that leverages Lagrangian-based methods to efficiently handle

constrained nonconvex optimization. This approach is particularly well-suited for large-scale data, where batch processing is impractical and online updates are preferred.

Constrained nonconvex optimization is a common problem in machine learning, signal processing, and control, where minimizing a loss function subject to complex constraints is necessary. While traditional methods like projected gradient descent work efficiently with simple constraints, they become computationally expensive or inapplicable in more complex feasible regions. To address these limitations, we introduce a Lagrangian-based method that transforms the original problem into a min-max formulation, enabling the use of primal-dual algorithms.

At the heart of our approach is the exploration of the Lagrangian landscape, where we analyze the properties of equilibria—points where the primal and dual variables converge. We demonstrate that the landscape exhibits two types of equilibria: stable and unstable. At unstable equilibria, the Lagrangian function shows negative curvature, allowing the algorithm to escape toward more promising regions. Importantly, all stable equilibria correspond to global optima, ensuring that spurious local minima are avoided.

We apply this framework to solve an online version of the generalized eigenvalue problem (GEV), which has significant applications in tasks such as CCA and sufficient dimension reduction (SDR). Our algorithm, a variant of the Generalized Hebbian Algorithm (GHA), operates under stochastic approximations and provides strong theoretical guarantees for convergence. Using diffusion approximations, we analyze the sample complexity and convergence properties of our approach, with empirical results demonstrating scalability and accuracy in both synthetic and real-world data settings.

1.1 Contributions

The contributions of this paper are threefold:

- We introduce a novel stochastic primal-dual algorithm for solving the Partial Least Squares (PLS) and Canonical Correlation Analysis (CCA) problems in multiview learning.
- We provide a comprehensive theoretical analysis of the algorithm's convergence, including sample complexity and asymptotic behavior based on diffusion approximations.
- We demonstrate the empirical performance of the proposed algorithm on both synthetic and real-world datasets, highlighting its scalability and efficiency compared to traditional approaches.

1.2 Organization

The remainder of this paper is organized as follows. Section 2 reviews the related work on multiview learning and stochastic optimization. Section 3 presents the proposed nonconvex optimization framework for scalable multiview learning, as a stochastic approach to partial least squares and canonical correlation analysis. Section 4 explores the primal-dual dynamics in constrained nonconvex optimization, with stochastic search and Lagrangian landscape insights. Section 5 concludes the paper with discussions on future work.

1.3 Notations

Given an integer d, we denote I_d as a $d \times d$ identity matrix, $[d] = \{1, 2, \ldots, d\}$. Given an index set $\mathcal{I} \subseteq [d]$ and a matrix $X \in \mathbb{R}^{d \times r}$, we denote $\mathcal{I}^\perp = [d] \backslash \mathcal{I}$ as the complement set of $\mathcal{I}$, $X_{:,i}$ ($X_{i,:}$) as

the i-th column (row) of X, $X_{i,j}$ as the (i,j)-th entry of X, and $X_{:,\mathcal{I}}$ ($X_{\mathcal{I},:}$) as the column (row) submatrix of X indexed by $\mathcal{I}$, $\mathrm{vec}(X) \in \mathbb{R}^{dr}$ as the vectorization of X, $\mathrm{Col}(X)$ as the column space of X, and $\mathrm{Null}(X)$ as the null space of X. Given a symmetric matrix $X \in \mathbb{R}^{d\times d}$, we denotethe eigenvalue decomposition of X as $X = O\Lambda O^\top$, where $\Lambda = \mathrm{diag}(\lambda_1, ...\lambda_d)$ with $\lambda_1 \geq \cdots \geq \lambda_d$, denote $\|X\|_2$ as the spectral norm of X. Given two matrices X and Y, $X \otimes Y$ as the Kronecker product of X, Y. Given a vector $v = (v^{(1)}, \ldots, v^{(d)})^\top \in \mathbb{R}^d$, we define vector norms: $\|v\|_1 = \sum_j |v^{(j)}|$, $\|v\|_2^2 = \sum_j (v^{(j)})^2$, and $\|v\|_\infty = \max_j |v^{(j)}|$. Given a matrix $A \in \mathbb{R}^{d\times d}$, we use $A_j = (A_{1j}, ..., A_{dj})^\top$ to denote the j-th column of A and define the matrix norms $\|A\|_{\mathrm{F}}^2 = \sum_j \|A_j\|_2^2$ and $\|A\|_2$ as the largest singular value of A.

2 Related Works

Multiview data analysis has broad applications, particularly in scenarios such as computer vision, natural language processing, and acoustic recognition, where multiple sets of features, like pixels and captions, or audio signals and transcriptions, represent the same underlying phenomena [VSTC02, BALHJ12, AL12, CKLS09, KSE05, SFF10, HSST04, DFU11]. A core challenge in multiview learning is extracting latent representations that capture correlations between views, while maintaining scalability for high-dimensional data. Although many multiview learning tasks are unsupervised, they rely on discovering intrinsic associations between views to reveal low-dimensional structures.

Traditional approaches, such as the Sample Average Approximation (SAA) and Principal Component Analysis (PCA), have been widely used in multiview and matrix factorization problems [FHT01, SQW16]. However, these methods often struggle in large-scale or high-dimensional settings due to scalability issues with offline, batch processing. In response, nonconvex optimization methods have gained traction, offering both computational and empirical advantages, particularly in stochastic approximation frameworks. For instance, algorithms like matrix sensing, factorization, and phase retrieval [BNS16, GLM16, ZLTW17] share similarities with multiview learning and can benefit from nonconvex approaches.

One popular strategy is to relax the convexity assumption in optimization, allowing heuristic nonconvex methods to solve complex problems with more efficiency than traditional convex approaches [ZWL15, CLS15, GHJY15]. These methods have shown better empirical performance, as they avoid costly projection steps required by algorithms like the Matrix Stochastic Gradient (MSG) algorithm [AMM16], which can become inefficient due to high computational and memory demands. Furthermore, recent works have demonstrated that first-order algorithms, like projected gradient descent, can converge to global optima under certain conditions, such as strict saddle properties [GHJY15].

For more complex constraints, such as those in generalized eigenvalue problems, traditional algorithms can struggle to compute projections efficiently [BV04, LY$^+$84]. To address these challenges, Lagrangian-based primal-dual methods have emerged, reformulating constrained problems as min-max problems [LLM11, CLO14, IN14]. These approaches have been particularly effective when the optimization landscape exhibits convex-concave properties, enabling saddle-point convergence under strong duality conditions. However, in nonconvex settings where the feasible region is more complicated, solving the min-max problem becomes significantly more challenging and often NP-hard [GHJY15, BNS16]. While substantial progress has been made in solving nonconvex problems with simple constraints, many challenges remain when both the objective and feasible region are nonconvex, leaving open questions regarding the applicability and convergence guarantees of

these algorithms in more complex scenarios.

3 Efficient Nonconvex Optimization for Scalable Multiview Learning

One ubiquitous approach is partial least square (PLS) for multiview representation learning. Specifically, given a data set of n samples of two sets of random variables (views), $X \in \mathbb{R}^m$ and $Y \in \mathbb{R}^d$, PLS aims to find an r-dimensional subspace ($r \ll \min(m, d)$) that preserves most of the covariance between two views. Existing literature has shown that such a subspace is spanned by the leading r components of the singular value decomposition (SVD) of $\Sigma_{XY} = \mathbb{E}_{(X,Y)\sim\mathcal{D}}\left[XY^\top\right]$ [ACLS12], where we sample (X, Y) from some unknown distribution $\mathcal{D}$. Throughout the rest of the paper, if not clear specified, we denote $\mathbb{E}_{(X,Y)\sim\mathcal{D}}$ by $\mathbb{E}$ for notational simplicity.

A straightforward approach for PLS is "Sample Average Approximation" (SAA, [Abd03, AZ05]), where we run an offline (batch) SVD algorithm on the empirical covariance matrix after seeing sufficient data samples. However, in the "big data" regime, this approach requires unfeasible amount of storage and computation time. Therefore, it is much more practical to consider the multiview learning problem in a "data laden" setting, where we draw independent samples from an underlying distribution $\mathcal{D}$ over $\mathbb{R}^m \times \mathbb{R}^d$, one at a time. This further enables us to formulate PLS as a stochastic (online) optimization problem. Here we only consider the rank-1 case ($r = 1$) for simplicity, and solve

$$(\widehat{u},\ \widehat{v}) = \underset{u\in\mathbb{R}^m, v\in\mathbb{R}^d}{\operatorname{argmax}}\ \mathbb{E}\left(v^\top Y X^\top u\right) \quad \text{subject to} \quad u^\top u = 1, v^\top v = 1 \tag{3.1}$$

We will explain more details on the rank-r case in the later section.

Several nonconvex stochastic approximation (SA) algorithms have been proposed in [ACLS12]. These algorithms work great in practice, but lack theoretic justifications, since the nonconvex nature of (3.1) makes the theoretical analysis very challenging. To overcome this obstacle, [AMM16] propose a convex relaxation of (3.1). Specifically, by a reparametrization $M = uv^\top$ (Recall that we are interested in the rank-1 PLS), they rewrite (3.1) as[1]

$$\widehat{M} = \underset{M}{\operatorname{argmax}}\ \langle M, \Sigma_{XY}\rangle \quad \text{subject to} \quad \|M\|_* \le 1 \text{ and } \|M\|_2 \le 1 \tag{3.2}$$

where $\Sigma_{XY} = \mathbb{E}XY^\top$, and $\|M\|_2$ and $\|M\|_*$ are the spectral (i.e., the largest singular value of M) and nuclear (i.e., the sum of all singular values of M) norms of M respectively. By examining the KKT conditions of (3.2), one can verify that $\widehat{M} = \widehat{u}\widehat{v}^\top$ is the optimal solution, where $\widehat{u}, \widehat{v}$ are the leading left and right singular vectors of Σ_{XY}, i.e., a pair of global optimal solutions to (3.1) for $r = 1$. Accordingly, they propose a projected stochastic gradient-type algorithm to solve (3.2), which is often referred to the Matrix Stochastic Gradient (MSG) algorithm. Particularly, at the $(k+1)$-th iteration, MSG takes

$$M_{k+1} = \Pi_{\text{Fantope}}(M_k + \eta X_k Y_k^\top)$$

where X_k and Y_k are independently sampled from $\mathcal{D}$, and $\Pi_{\text{Fantope}}(\cdot)$ is a projection operator to the feasible set of (3.2). They further prove that given a pre-specified accuracy ϵ, MSG requires $N = \mathcal{O}(\epsilon^{-2}\log(1/\epsilon))$ iterations such that $\langle \widehat{M}, \mathbb{E}xy^\top\rangle - \langle M_N, \mathbb{E}xy^\top\rangle \le \epsilon$ with high probability.

[1] For $r > 1$ case, we replace $\|M\|_* \le 1$ with $\|M\|_* \le r$

Main Results We aims to bridge the gap between theory and practice for solving multiview representation learning problems by nonconvex approaches. Specifically, we first illustrate the nonoconvex geometry of (3.1), we analyze the convergence properties of a simple stochastic optimization algorithm for solving (3.1) based on diffusion processes. Our analysis takes advantage of the strong Markov properties of the stochastic optimization algorithm updates and casts the trajectories of the algorithms as a diffusion processes [EK09, LWL16a]. By leveraging the weak convergence from discrete Markov chains to their continuous time limits, we demonstrate that the trajectories are essentially the solutions to stochastic differential equations. Such an SDE-type analysis automatically incorporates the geometry of the objective and the randomness of the algorithm, and eventually demonstrates three phases of convergence.

(i) Starting from an unstable equilibrium with negative curvature, the dynamics of the algorithm can be described by an Ornstein-Uhlenbeck process with a steady driven force pointing away from the initial.

(ii) When the algorithm is sufficiently distant from the initial unstable equilibrium, the dynamics can be characterized by a deterministic ordinary differential equation (ODE). The trajectory of this phase is evolving directly toward the desired global maximum until it reaches a small basin around the global maximum.

(iii) In this phase, the trajectory can be also described by an Ornstein-Uhlenbeck process oscillating around the global maximum. The process has a drifting term that gradually dies out and eventually becomes a nearly unbiased random walk centered at the maximum.

The sharp characterization in these three phases eventually allows us to establish strong convergence guarantees. Particularly, we show that the nonconvex stochastic gradient algorithm guarantees an ϵ-optimal solution in $\mathcal{O}(\epsilon^{-1}\log(\epsilon^{-1}))$ iterations with high probability, which is a significant improvement over convex MSG by a factor of ϵ^{-1}. Our theoretical analysis reveals the power of the nonconvex optimization in PLS. The simple heuristic algorithms drop the convexity, but achieve much better efficiency.

Our convergence analysis also has important implications on stochastic optimization algorithm for Canonical Correlation Analysis (CCA). Specifically, CCA considers a similar setting to PLS, and solves

$$(\widehat{u},\widehat{v}) = \underset{u,v}{\operatorname{argmax}}\ u^\top \mathbb{E}XY^\top v \quad \text{subject to} \quad \mathbb{E}(X^\top u)^2 = 1,\ \mathbb{E}(Y^\top v)^2 = 1 \tag{3.3}$$

From an optimization perspective, CCA is equivalent to PLS under some linear transformation, but more challenging. We will explain more details on CCA in our later discussions.

3.1 Stochastic Nonconvex Optimization

Recall that we solve (3.1)

$$(\widehat{u},\widehat{v}) = \underset{u,v}{\operatorname{argmax}}\ u^\top \mathbb{E}XY^\top v \quad \text{subject to} \quad \|u\|_2^2 = 1,\ \|v\|_2^2 = 1 \tag{3.4}$$

where (X, Y) follows some unknown distribution $\mathcal{D}$. Due to the symmetrical structure of (3.4), $(-\widehat{u}, -\widehat{v})$ is also a pair of global optimum. Our analysis holds for both optima. Throughout the rest of the paper, if not clearly specified, we consider $(\widehat{u}, \widehat{v})$ as the global optimum for simplicity.

We apply the stochastic approximation (SA) of the generalized Hebbian algorithm (GHA) to solve (3.4). GHA, which is also referred as Sanger's rule [San89], is essentially a primal-dual algorithm. Specifically, we consider the Lagrangian function of (3.4):

$$L(u, v, \mu, \sigma) = u^\top \mathbb{E}XY^\top v - \mu(u^\top u - 1) - \sigma(v^\top v - 1) \tag{3.5}$$

where μ and σ are Lagrangian multipliers. We then check the optimal KKT conditions,

$$\mathbb{E}XY^\top v - 2\mu u = 0, \quad \mathbb{E}YX^\top u - 2\sigma v = 0, \quad u^\top u = 1 \quad \text{and} \quad v^\top v = 1 \tag{3.6}$$

which further imply

$$u^\top \mathbb{E}XY^\top v - 2\mu u^\top u = u^\top \mathbb{E}XY^\top v - 2\mu = 0,$$
$$v^\top \mathbb{E}YX^\top u - 2\sigma v^\top v = v^\top \mathbb{E}YX^\top u - 2\sigma = 0$$

Solving the above equations, we obtain the optimal Lagrangian multipliers as

$$\mu = \sigma = \frac{1}{2} u^\top \mathbb{E}XY^\top v \tag{3.7}$$

GHA is inspired by (3.6) and (3.7). At k-th iteration GHA takes

$$\text{Dual Update}: \quad \mu_k = \sigma_k = \frac{1}{2} \underbrace{u_k^\top X_k Y_k^\top v_k}_{\text{SA (stochastic approximation) of } u_k^\top \Sigma v_k} \tag{3.8}$$

$$\text{Primal Update}: \quad u_{k+1} = u_k + \eta \underbrace{\left(X_k Y_k^\top v_k - 2\mu_k u_k\right)}_{\text{SA of } \nabla_u L(u,v,\mu,\sigma)}, \quad v_{k+1} = v_k + \eta \underbrace{\left(Y_k X_k^\top u_k - 2\sigma_k v_k\right)}_{\text{SA of } \nabla_v L(u,v,\mu,\sigma)} \tag{3.9}$$

where $\eta > 0$ is the step size. Combining (3.8) and (3.9), we obtain a dual-free update as follow:

$$u_{k+1} = u_k + \eta \left(X_k Y_k^\top v_k - u_k^\top X_k Y_k^\top v_k u_k\right) \quad \text{and} \quad v_{k+1} = v_k + \eta \left(Y_k X_k^\top u_k - u_k^\top X_k Y_k^\top v_k v_k\right) \tag{3.10}$$

Different from the projected SGD algorithm, which is a primal algorithm, Stochastic GHA does not need projection at each iteration.

3.2 Optimization Landscape

We illustrate the nonconvex optimization landscape of (3.1), which helps us understand the intuition behind the algorithmic convergence. We first study its stationary points based the Lagrangian function (3.5). By the KKT conditions (3.6), we define the stationary point of (3.5) as follows.

Definition 1. *Given* (3.1) *and* (3.5)*, we define:*

(i) A quadruplet of (u, v, μ, σ) *is called a stationary point of* (3.5)*, if it satisfies* (3.6)*.*

(ii) A pair of (u, v) *is called a stable stationary point of* (3.1)*, if* (u, v, μ, σ) *is a stationary point of* (3.5)*, and* $\nabla^2_{u,v} L(u, v, \mu, \sigma)$ *is negative semi-definite.*

(iii) A pair of (u, v) is called an unstable stationary point of (3.1), if (u, v, μ, σ) is a stationary point of (3.5), and $\nabla^2_{u,v} L(u, v, \mu, \sigma)$ has a positive eigenvalue.

We then obtain all stationary points by solving (3.6). For notational simplicity, we denote $\Sigma_{XY} = \mathbb{E}XY^\top$. Before we proceed with our analysis, we introduce the following assumption.

Assumption 1. *Suppose $d \leq m$ and rank$(\Sigma_{XY}) = r$. We have $\lambda_1 > \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_r > 0$, where λ_i's are the i-th singular values of Σ_{XY}.*

We impose such an eigengap assumption ($\lambda_1 > \lambda_2$) to ensure the identifiability of the leading pair of singular vectors. Thus, the leading pair of singular vectors are uniquely determined only up to sign change. Let $O_1 \in \mathbb{R}^{m\times m}$ and $O_2 \in \mathbb{R}^{d\times d}$ be any pair of left and right singular matrices[2]. Let $\overline{u}_i$ and $\overline{v}_j$ denote the i-th column of O_1 and j-th column of O_2, respectively. The next proposition reveals the connection between stationary points and singular vectors.

Proposition 1. *Suppose Assumption 1 holds. A quadruplet of (u, v, μ, σ) is the stationary point of (3.5), if either of the following condition holds:*

(i) (u, v) are a pair of singular vectors associated with the same nonzero singular value;

(ii) u and v belong to the row and column null spaces of Σ_{XY} respectively: $\Sigma_{XY} v = 0$, $\Sigma^\top_{XY} u = 0$.

The proof of Proposition 1 is presented in Appendix A.1.1. We then determine the types of these obtained stationary points. The next proposition characterizes the maximum eigenvalues of $\nabla^2_{u,v} L(u, v, \mu, \sigma)$ at these stationary points of (3.5).

Proposition 2. *Suppose Assumption 1 holds. All pairs of singular vectors associated with the leading singular value are global optima of (3.1), i.e., also the saddle points of (3.5), and they are stable stationary points. All other stationary points of (3.5) are all unstable with*

$$\lambda_{\max}(\nabla^2_{u,v} L(u, v, \mu, \sigma)) \geq \lambda_1 - \lambda_2$$

The proof of Proposition 2 is presented in Appendix A.1.2. Proposition 2 essentially characterizes the geometry of (3.1) at all stationary points, and the unstableness allows the stochastic gradient algorithm to escape, as will be shown in the next sections.

3.3 Global Convergence by ODE

Before we proceed with our analysis, we first impose some mild assumptions on the problem.

Assumption 2. *X_k, Y_k, $k = 1, 2, ...N$ are data samples identically independently distributed as $X \in \mathbb{R}^d$, $Y \in \mathbb{R}^d$ respectively satisfying the following conditions:*

(i) For any $\Delta > 0$, $\max\{\mathbb{E}\|X\|_2^{4+\Delta}, \mathbb{E}\|Y\|_2^{4+\Delta}\} < \infty$ and $\max\{\mathbb{E}\|X\|_2^2, \mathbb{E}\|Y\|_2^2\} \leq Bd$ for a constant B;[3]

(ii) $\lambda_1 > \lambda_2 \geq \lambda_3 \geq ... \geq \lambda_d > 0$, where λ_i's are the singular values of $\Sigma_{XY} = \mathbb{E}XY^\top$.

[2]Since all singular values are not necessarily distinct, some pairs of singular vectors are not unique, e.g., when $\lambda_i = \lambda_j$, $(\overline{u}_i, \overline{v}_i)$ and $(\overline{u}_j, \overline{v}_j)$ are uniquely determined up to rotation. Note that our analysis works for all possible combinations of O_1 and O_2. See more details in [GVL12].

[3]We only need $(4 + \Delta)$-th moments of $\|X\|_2$ and $\|Y\|_2$ to be bounded, while the preliminary results require both $\|X\|_2$ and $\|Y\|_2$ to be bounded random variables.

Here we assume X and Y are of the same dimensions (i.e., $m = d$) and Σ_{XY} is full rank for convenience of analysis. The extension to $m \neq d$ in a rank deficient setting is straightforward, but more involved (See more details in Section 3.4.4). Moreover, for a multiview learning problem, it is also natural to impose the following additional assumptions.

Assumption 3. *Given the observed random variables X and Y, there exist two orthogonal matrices $O_X \in \mathbb{R}^{d\times d}$, $O_Y \in \mathbb{R}^{d\times d}$ such that $X = O_X\overline{X}$, $Y = O_Y\overline{Y}$, where $\overline{X} = (\overline{X}^{(1)}, ..., \overline{X}^{(d)})^\top \in \mathbb{R}^d$ and $\overline{Y} = (\overline{Y}^{(1)}, ..., \overline{Y}^{(d)})^\top \in \mathbb{R}^d$ are the latent variables satisfying:*

(i) $\overline{X}^{(i)}$ and $\overline{Y}^{(j)}$ are uncorrelated if $i \neq j$, so that O_X and O_Y are the left and right singular matrices of Σ_{XY} respectively;

(ii) $\mathrm{Var}(\overline{X}^{(i)}) = \gamma_i$, $\mathrm{Var}(\overline{Y}^{(i)}) = \omega_i$, $\mathbb{E}\left(\overline{X}^{(i)}\overline{Y}^{(i)}\overline{X}^{(j)}\overline{Y}^{(j)}\right) = \alpha_{ij}$, where γ_i, α_{ij}, and ω_i are constants.

The next proposition characterizes the strong Markov property of our algorithm.

Proposition 3. *Using (3.10), we get a sequence of (u_k, v_k), $k = 1, 2, ..., N$. They form a discrete-time Markov process.*

With Proposition 3, we can construct a continuous time process to derive an ordinary differential equation to analyze the algorithmic convergence. Specifically, as the fixed step size $\eta \to 0^+$, two processes $U_\eta(t) = u_{\lfloor \eta^{-1}t \rfloor}$, $V_\eta(t) = v_{\lfloor \eta^{-1}t \rfloor}$ based on the sequence generated by (3.10), weakly converge to the solution of the following ODE system in probability (see more details in [EK09]),

$$\frac{dU}{dt} = \left(\Sigma_{XY}V - U^\top\Sigma_{XY}VU\right), \quad \frac{dV}{dt} = \left(\Sigma_{XY}^\top U - V^\top\Sigma_{XY}^\top UV\right) \tag{3.11}$$

where $U(0) = u_0$ and $V(0) = v_0$. To highlight the sequence generated by (3.10) depending on η, we redefine $u_{\eta,k} = u_k$, $v_{\eta,k} = v_k$.

Theorem 3.1. *As $\eta \to 0^+$, the processes $u_{\eta,k}$, $v_{\eta,k}$ weakly converge to the solution of the ODE system in (3.11) with sphere initial $U(0) = u_0$, $V(0) = v_0$, i.e., $\|u_0\|_2 = \|v_0\|_2 = 1$.*

The proof of Theorem 3.1 is presented in Appendix A.2.1. Under Assumption 2, the above ODE system admits a closed form solution. Specifically, we solve U and V simultaneously, since they are coupled together in (3.11). To simplify (3.11), we define $W = \frac{1}{\sqrt{2}}\begin{pmatrix} U^\top & V^\top \end{pmatrix}^\top$ and $w_k = \frac{1}{\sqrt{2}}\begin{pmatrix} u_k^\top & v_k^\top \end{pmatrix}^\top$. We then rewrite (3.11)as

$$\frac{dW}{dt} = QW - W^\top QWW \tag{3.12}$$

where $Q = \begin{pmatrix} 0 & \Sigma_{XY} \\ \Sigma_{XY}^\top & 0 \end{pmatrix}$. By Assumption 3, O_X and O_Y are the left and right singular matrices of Σ_{XY} respectively, i.e., $\Sigma_{XY} = \mathbb{E}XY^\top = O_X\mathbb{E}\overline{XY}^\top O_Y^\top$, where $\mathbb{E}\overline{XY}^\top$ is diagonal. For notational simplicity, we define $D = \mathrm{diag}(\lambda_1, \lambda_2, ..., \lambda_d)$ such that $\Sigma_{XY} = O_X D O_Y^\top$. One can verify $Q = P\Lambda P^\top$, where

$$P = \frac{1}{\sqrt{2}}\begin{pmatrix} O_X & O_X \\ O_Y & -O_Y \end{pmatrix} \quad \text{and} \quad \Lambda = \begin{pmatrix} D & 0 \\ 0 & -D \end{pmatrix} \tag{3.13}$$

By left multiplying $P^\top$ both sides of (3.12), we obtain

$$H(t) = P^\top W(t) \text{ with } \frac{dH}{dt} = \Lambda H - H^\top \Lambda H H \tag{3.14}$$

which is a coordinate separable ODE system. Accordingly, we define $h_k^{(i)}$'s as:

$$h_k = P^\top w_k \quad \text{and} \quad h_k^{(i)} = P_i^\top w_k \tag{3.15}$$

Thus, we can obtain a closed form solution to (3.14) based on the following theorem.

Theorem 3.2. *Given* (3.14)*, we write the ODE in each component* $H^{(i)}$*,*

$$\frac{d}{dt}H^{(i)} = H^{(i)} \sum_{j=1}^{2d} (\lambda_i - \lambda_j) (H^{(j)})^2 \tag{3.16}$$

where $\lambda_i = -\lambda_{i-d}$ *when* $i > d$*. This ODE System has a closed form solution as follows:*

$$H^{(i)}(t) = \big(C(t)\big)^{-\frac{1}{2}} H^{(i)}(0) \exp(\lambda_i t) \tag{3.17}$$

for $i = 1, 2, ..., 2d$*, where*

$$C(t) = \sum_{j=1}^{2d} \left(\left(H^{(j)}(0) \right)^2 \exp(2\lambda_j t) \right)$$

is a normalization function such that $\|H(t)\|_2 = 1$*.*

The proof of Theorem 3.2 is presented in Appendix A.2.2. Without loss of generality, we assume $H^{(1)}(0) > 0$. As can be seen, $H_1(t) \to 1$, as $t \to \infty$. We have successfully characterized the global convergence performance of our algorithm with an approximate error $o(1)$. The solution to the ODE system in (3.17), however, does not fully reveal the algorithmic behavior (more precisely, the rate of convergence) near the equilibria of the ODE system. This further motivates us to exploit the stochastic differential equation approach to characterize the dynamics of the algorithm.

3.4 Global Dynamics by SDE

We analyze the dynamics of the algorithm near the equilibria based on stochastic differential equation by rescaling analysis. Specifically, we characterize three stages for the trajectories of solutions: [a] Neighborhood around unstable equilibria — minimizers and saddle points of (3.4), [b] Neighborhood around stable equilibria — maximizers of (3.4), and [c] deterministic traverses between equilibria. Moreover, we provide the approximate the number of iterations in each phase until convergence.

3.4.1 Phase I: Escaping from Unstable Equilibria

Suppose that the algorithm starts to iterate around a unstable equilibrium, (e.g. saddle point). Different from our previous analysis, we rescale two aforementioned processes $U_\eta(t)$ and $V_\eta(t)$ rescaled by a factor of $\eta^{-1/2}$. This eventually allows us to capture the uncertainty of the algorithm updates by stochastic differential equations. Roughly speaking, the ODE approximation is essentially

a variant of law of large number for Markov process, while the SDE approximation serves as a variant of central limit theorem accordingly.

Recall that P is an orthonormal matrix for diagonalizing Q, and H is defined in (3.14). Let $Z_\eta^{(i)}$ and $z_{\eta,k}^{(i)}$ denote the i-th coordinates of $Z_\eta = \eta^{-1/2} H_\eta$ and $z_{\eta,k} = \eta^{-1/2} h_{\eta,k}$ respectively. The following theorem characterizes the dynamics of the algorithm around the unstable equilibrium.

Theorem 3.3. *Suppose $z_{\eta,0}$ is initialized around some saddle point or minimizer (e.g. j-th column of P with $j \neq 1$), i.e., $Z^{(j)}(0) \approx \eta^{-\frac{1}{2}}$ and $Z^{(i)}(0) \approx 0$ for $i \neq j$. Then for any $C > 0$, there exist $\tau > 0$ and $\eta' > 0$ such that*

$$\sup_{\eta < \eta'} \mathbb{P}(\sup_t |Z_\eta^{(i)}(t)| \leq C) \leq 1 - \tau \tag{3.18}$$

Here we provide the proof sketch and leave the whole proof of Theorem 3.3 in Appendix A.3.1.

Proof Sketch. We prove this argument by contradiction. Assume the conclusion does not hold, that is there exists a constant $C > 0$, such that for any $\eta' > 0$ we have

$$\sup_{\eta \leq \eta'} \mathbb{P}(\sup_t |Z_\eta^{(i)}(t)| \leq C) = 1$$

That implies there exists a sequence $\{\eta_n\}_{n=1}^\infty$ converging to 0 such that

$$\lim_{n \to \infty} \mathbb{P}(\sup_t |Z_{\eta_n}^{(i)}(t)| \leq C) = 1 \tag{3.19}$$

Then we show $\{Z_{\eta_n}^{(i)}(\cdot)\}_n$ is tight and thus converges weakly. Furthermore, $\{Z_{\eta_n}^{(i)}(\cdot)\}_n$ weakly converges to a stochastic differential equation,

$$dZ^{(i)}(t) = -(\lambda_j - \lambda_i) Z^{(i)}(t) dt + \beta_{ij} dB(t) \tag{3.20}$$

We compute the solution of this stochastic differential equation and then show (3.18) does not hold. □

Theorem 3.3 implies that for $i > j$, with a constant probability τ, escapes from the saddle points at some time T_1, i.e., $(H^{(j)}(T_1))^2$ is smaller than $1 - \delta^2$, where $(\delta = O(\sqrt{\eta}))$. Note that (3.20) is a Fokker-Planck equation, which admits a closed form solution as follows,

$$\begin{aligned} Z^{(i)}(t) &= Z^{(i)}(0) \exp\left[-(\lambda_j - \lambda_i)t\right] + \beta_{ij} \int_0^t \exp\left[(\lambda_j - \lambda_i)(s-t)\right] dB(s) \\ &= \Bigg[\underbrace{Z^{(i)}(0) + \beta_{ij} \int_0^t \exp\left[(\lambda_j - \lambda_i)s\right] dB(s)}_{T_1} \Bigg] \underbrace{\exp\left[(\lambda_i - \lambda_j)t\right]}_{T_2} \quad \text{for } i \neq j \end{aligned} \tag{3.21}$$

Such a solution is well known as the Ornstein-Uhlenbeck process [Øks03], and also implies that the distribution of $z_{\eta,k}^{(i)}$ can be well approximated by the normal distribution of $Z^{(i)}(t)$ for a sufficiently small step size. This continuous approximation further has the following implications:

[a] For $\lambda_i > \lambda_j$, $T_1 = \beta_{ij} \int_0^t \exp\left[(\lambda_j - \lambda_i)s\right] dB(s) + Z^{(i)}(0)$ is essentially a random variable with mean $Z^{(i)}(0)$ and variance smaller than $\frac{\beta_{ij}^2}{2(\lambda_i - \lambda_j)}$. The larger t is, the closer its variance gets to this upper bound. While $T_2 = \exp\left[(\lambda_i - \lambda_j)t\right]$ essentially amplifies T_1 by a factor exponentially increasing in t. This tremendous amplification forces $Z^{(i)}(t)$ to quickly get away from 0, as t increases.

[b] For $\lambda_i < \lambda_j$, we have

$$\mathbb{E}[Z^{(i)}(t)] = Z^{(i)}(0)\exp\left[-(\lambda_j - \lambda_i)t\right] \quad \text{and} \quad \mathrm{Var}[Z^{(i)}(t)] = \frac{\beta_{ij}^2}{2(\lambda_j - \lambda_i)}\left[1 - \exp\left[-2(\lambda_j - \lambda_i)t\right]\right]$$

As has been shown in **[a]** that t does not need to be large for $Z^{(i)}(t)$ to get away from 0. Here we only consider relatively small t. Since the initial drift for $Z^{(i)}(0) \approx 0$ is very small, $Z^{(i)}$ tends to stay at 0. As t increases, the exponential decay term makes the drift quickly become negligible. Moreover, by mean value theorem, we know that the variance is bounded, and increases far slower than the variance in **[a]**. Thus, roughly speaking, $Z^{(i)}(t)$ oscillates near 0.

[c] For $\lambda_j = \lambda_i$, we have $\mathbb{E}[Z^{(i)}(t)] = Z^{(i)}(0)$ and $\mathrm{Var}[Z^{(i)}(t)] = \beta_{ij}^2$. This implies that $Z^{(i)}(t)$ also tends to oscillate around 0, as t increases.

Overall speaking, **[a]** is dominative so that it is the major driving force for the algorithm to escape from this unstable equilibrium. More precisely, let us consider one special case for Phase I, that is we start from the second maximum singular value, with $h_{\eta,k}^{(2)}(0) = 1$. We then approximately calculate the number of iterations to escape Phase I using the algorithmic behavior of $h_{\eta,k}^{(1)} = \eta^{1/2} z_{\eta,k}^{(1)} \approx \eta^{1/2} Z_\eta^{(1)}(t)$ with $t = k\eta$ by the following proposition.

Proposition 4. *Given pre-specified $\nu > 0$ and sufficiently small η, there exists some $\delta \asymp \eta^\mu$, where $\mu \in (0, 0.5)$ is a generic constant, such that the following result holds: We need at most*

$$N_1 = \frac{\eta^{-1}}{2(\lambda_1 - \lambda_2)} \log\left(\frac{2\eta^{-1}\delta^2(\lambda_1 - \lambda_2)}{\Phi^{-1}\left(\frac{1+\nu}{2}\right)^2 \beta_{12}^2} + 1\right)$$

iterations such that $(h_{\eta,N_1}^{(2)})^2 \le 1 - \delta^2$ with probability at least $1 - \nu$, where $\Phi(x)$ is the CDF of standard normal distribution.

The proof of Proposition 4 is provided in Appendix A.3.2. Proposition 4 suggests that SGD can escape from unstable equilibria within a few iterations. After escaping from the saddle, SGD gets into the next phase, which is a deterministic traverse between equilibria.

3.4.2 Phase II: Traverse between Equilibria

When the algorithm is close to neither the saddle points nor the optima, the algorithm's performance is nearly deterministic. Since $Z(t)$ is a rescaled version of $H(t)$, their trajectories are similar. Like before, we have the following proposition to calculate the approximate iterations, N_2, following our results in Section 3.3. We restart the counter of iteration by Proposition 3.

Proposition 5. *After restarting counter of iteration, given sufficiently small η and δ defined in Proposition 4, we need at most*

$$N_2 = \frac{\eta^{-1}}{2(\lambda_1 - \lambda_2)} \log\frac{1 - \delta^2}{\delta^2}$$

iterations such that $\left(h_{\eta,N_2}^{(1)}\right)^2 \ge 1 - \delta^2$.

The proof of Proposition 5 is provided in Appendix A.3.3. Combining Propositions 4 and 5, we know that after $N_1 + N_2$ iteration numbers, SGD is close to the optimum with high probability, and gets into its third phase, i.e., convergence to stable equilibria.

3.4.3 Phase III: Convergence to Stable Equilibria

Again, we restart the counter of iteration by the strong Markov property. The trajectory and analysis are similar to Phase I, since we also characterize the convergence using an Ornstein-Uhlenbeck process. The following theorem characterizes the dynamics of the algorithm around the stable equilibrium.

Theorem 3.4. *Suppose $z_{\eta,0}$ is initialized around some maximizer (the first column of P), i.e., $Z^{(1)}(0) \approx \eta^{-\frac{1}{2}}$ and $Z^{(i)}(0) \approx 0$ for $i \neq 1$. Then as $\eta \to 0^+$, for all $i \neq 1$, $z_{\eta,k}^{(i)}$ weakly converges to a diffusion process $Z^{(i)}(t)$ satisfying the following SDE for $i \neq 1$,*

$$dZ^{(i)}(t) = -(\lambda_1 - \lambda_i)Z^{(i)}(t)dt + \beta_{i1}dB(t) \tag{3.22}$$

where $B(t)$ is a brownian motion, and

$$\beta_{i1} = \begin{cases} \frac{1}{2}\sqrt{\gamma_i\omega_1 + \gamma_1\omega_i + 2\alpha_{i1}} & \text{if } 1 \leq i \leq d, \\ \frac{1}{2}\sqrt{\gamma_i\omega_1 + \gamma_1\omega_i - 2\alpha_{i1}} & \text{otherwise} \end{cases}$$

The proof of Theorem 3.4 is provided in Appendix A.3.4. Similar to (3.21), the closed form solution to (3.22) for $i \neq 1$ is as follow:

$$Z^{(i)}(t) = Z^{(i)}(0)\exp\left[-(\lambda_1 - \lambda_i)t\right] + \beta_{i1}\int_0^t \exp\left[(\lambda_1 - \lambda_i)(s-t)\right]dB(s) \tag{3.23}$$

By the property of the O-U process, we characterize the expectation and variance of $Z^{(i)}(t)$ for $i \neq 1$.

$$\mathbb{E}Z^{(i)}(t) = Z^{(i)}(0)\exp\left[-(\lambda_1 - \lambda_i)t\right],$$
$$\mathbb{E}\left(Z^{(i)}(t)\right)^2 = \frac{\beta_{i1}^2}{2(\lambda_1 - \lambda_i)} + \left[\left(Z^{(i)}(0)\right)^2 - \frac{\beta_{i1}^2}{2(\lambda_1 - \lambda_i)}\right]\exp\left[-2(\lambda_1 - \lambda_i)t\right]$$

Recall that the distribution of $z_{\eta,k}^{(i)}$ can be well approximated by the normal distribution of $Z^{(i)}(t)$ for a sufficiently small step size. This further implies that after sufficiently many iterations, SGD enforces $z_{\eta,k}^{(i)} \to 0$ except $i = 1$. Meanwhile, SGD behaves like a biased random walk towards the optimum, when it iterates within a small neighborhood the optimum. But unlike Phase I, the variance gradually becomes a constant.

Based on theorem 3.4, we further establish an iteration complexity bound for SGD in following proposition.

Proposition 6. *Given a pre-specified $\epsilon > 0$, a sufficiently small η, and δ defined in Proposition 4, after restarting counter of iteration, we need at most*

$$N_3 = \frac{\eta^{-1}}{2(\lambda_1 - \lambda_2)}\log\left(\frac{4(\lambda_1 - \lambda_2)\delta^2}{(\lambda_1 - \lambda_2)\epsilon\eta^{-1} - 4d\max\limits_{1\leq i\leq d}\beta_{i1}^2}\right)$$

iterations such that $\sum_{i=2}^{2d}\left(h_{\eta,N_3}^{(i)}\right)^2 \leq \epsilon$ with probability at least $3/4$.

The proof of Proposition 6 is provided in Appendix A.3.5. Combining Propositions 4, 5, and 6, we obtain a more refined result in the following corollary.

Corollary 1. *Given a sufficiently small pre-specified $\epsilon > 0$, we choose*

$$\eta \asymp \frac{\epsilon(\lambda_1 - \lambda_2)}{d \max_{1 \le i \le d} \beta_{i1}^2}$$

We need at most

$$N = O\left[\frac{d}{\epsilon(\lambda_1 - \lambda_2)^2} \log\left(\frac{d}{\epsilon}\right)\right]$$

iterations such that we have $\|u_{\eta,n} - \widehat{u}\|_2^2 + \|v_{\eta,n} - \widehat{v}\|_2^2 \le 3\epsilon$ with probability at least $\frac{3}{4}$.

The proof of Corollary 1 is provided in Appendix A.3.6. We can further improve the probability to $1 - \nu$ for some $\nu > 0$ by repeating $\mathcal{O}(\log 1/\nu)$ replicates of SGD. We then compute the geometric median of all output solutions. See more details in [CLM+16].

3.4.4 Extension to $m \neq d$

Our analysis can further extend to the case where X and Y have different dimensions, i.e., $m \neq d$. Specifically, we consider an alternative way to construct P defined in (3.13). We follow the same notations to Assumption 3, and use O_X and O_Y to denote the transition matrix between the observed data and latent variables. The dimensions of O_X and O_Y, however, are different now, i.e., $O_X \in \mathbb{R}^{m \times m}$ and $O_Y \in \mathbb{R}^{d \times d}$. Without loss of generality, we assume $m > d$ and $O_X = (\widetilde{O}_X \; O_X^0)$, where $\widetilde{O}_X \in \mathbb{R}^{m \times d}$ and $O_X^0 \in \mathbb{R}^{m \times (m-d)}$, and O_Y are the transform matrix of X and Y, respectively. Then we have the singular value decomposition as follows,

$$O_X^\top \Sigma_{XY} O_Y = D, \quad \text{where } D = \begin{pmatrix} \widetilde{D} \\ 0 \end{pmatrix} \text{ and } \widetilde{D} = \mathrm{diag}(\lambda_1, \lambda_2, ..., \lambda_d) \tag{3.24}$$

Thus, we have $\widetilde{O}_X^\top \Sigma_{XY} O_Y = \widetilde{D}$ and $(O_X^0)^\top \Sigma_{XY} O_Y = 0$. Now we design the orthogonal transform matrix P.

$$P = \begin{pmatrix} \frac{1}{\sqrt{2}} \widetilde{O}_X & O_X^0 & \frac{1}{\sqrt{2}} \widetilde{O}_X \\ \frac{1}{\sqrt{2}} O_Y & 0 & -\frac{1}{\sqrt{2}} O_Y \end{pmatrix} \tag{3.25}$$

One can check that

$$\begin{pmatrix} 0 & \Sigma_{XY} \\ \Sigma_{XY}^\top & 0 \end{pmatrix} = P \begin{pmatrix} D & 0 \\ 0 & -D^\top \end{pmatrix} P^\top = P \begin{pmatrix} \widetilde{D} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -\widetilde{D} \end{pmatrix} P^\top \tag{3.26}$$

Then our previous analysis using ODE and SDE still holds.

Note that for $d = m$, any column vector of P in (3.13) is a stationary solution. Here the square matrix P in (3.25) contains $m + d$ column vectors, but only the first d and last d column vectors are stationary solutions. This is because the remaining $m - d$ column vectors are even not feasible solutions, and violate the constraint $v^\top v = 1$. Thus, given a feasible initial, the algorithm will not be trapped in the subspace spanned by th remaining $m - d$ column vectors .

3.4.5 Extension to Missing Values

Our methodology and theory can tolerate missing values. For simplicity, we assume the entries of X and Y misses independently with probability $1-p$ in each iteration, where $p \in (0,1)$. We then set all missing entries as 0 values. We denote such imputed vectors by $\widetilde{X}_k$ and $\widetilde{Y}_k$. One can verify $\frac{1}{p^2}\widetilde{X}_k \cdot \widetilde{Y}_k^\top$ is an unbiased estimator of $\Sigma_{XY} = \mathbb{E}X_k Y_k^\top$. Note that $1/p^2$ can be further absorbed into the step size η, denoted by η_p. Then (3.10) becomes:

$$u_{k+1} = u_k + \eta_p \left(\widetilde{X_k}\widetilde{Y_k}^\top v_k - u_k^\top \widetilde{X_k}\widetilde{Y_k}^\top v_k u_k \right) \quad \text{and} \quad v_{k+1} = v_k + \eta_p \left(\widetilde{Y_k}\widetilde{X_k}^\top u_k - u_k^\top \widetilde{X_k}\widetilde{Y_k}^\top v_k v_k \right) \tag{3.27}$$

The convergence analysis is very similar to the standard setting with a different choice of η_p, and therefore is omitted.

3.5 Discussions

We establish the convergence rate of stochastic gradient descent (SGD) algorithms for solving online partial least square (PLS) problems based on diffusion process approximation. Our analysis indicates that for PLS, dropping convexity actually improves efficiency and scalability. Our convergence results are tighter than existing convex relaxation based method by a factor of $O(1/\epsilon)$, where ϵ is a pre-specified error. We believe the following directions should be of wide interests:

(i) Our current results hold only for the top pair of left and right singular vectors, i.e., $r = 1$. For $r > 1$, we need to solve

$$(\widehat{U},\ \widehat{V}) = \underset{U \in \mathbb{R}^{m\times r}, V \in \mathbb{R}^{d\times r}}{\operatorname{argmax}} \ \mathbb{E}\ \operatorname{tr}(V^\top Y X^\top U) \quad \text{subject to} \quad U^\top U = I_r, \quad V^\top V = I_r \tag{3.28}$$

Our approximations using ODE and SDE, however, do not admit unique solution due to rotation or permutation. Thus, extension of our analysis to $r > 1$ is a challenging, but also an important future direction.

(ii) Our current results are only applicable to a fixed step size $\eta \asymp \epsilon(\lambda_1 - \lambda_2)d^{-1}$. Our results suggest that the diminishing step size $\eta_k \asymp k^{-1}(\lambda_1 - \lambda_2)^{-1}\log d$, k from 1 to N, where N is the sample complexity from theory, achieves a better empirical performance. One possible probability tool is Stein's method [R+11].

(iii) Our current results rely on the classical central limit theorem-type analysis by taking $\eta \to 0^+$. Note the analysis of $\|u\|_2 = \|v\|_2 = 1$ is an asymptotic result, and in result, when η is small, u and v exactly stay on the sphere. But to get a more general result, connecting our analysis to discrete algorithmic proofs such as [JJK+16, Sha15, LWLZ16] should be an important direction [BC05]. One possible probability tool for addressing this issue is Stein's method [R+11].

Moreover, our proposed SGD algorithm for PLS is also closely related to Canonical Correlation Analysis. Specifically, CCA solves a similar problem

$$(\widehat{u}, \widehat{v}) = \underset{u,v}{\operatorname{argmax}}\ u^\top \mathbb{E}XY^\top v \quad \text{subject to} \quad \mathbb{E}(X^\top u)^2 = 1,\ \mathbb{E}(Y^\top v)^2 = 1 \tag{3.29}$$

For notational simplicity, we denote $\Sigma_{XY} = \mathbb{E}XY^\top$, $\Sigma_{XX} = \mathbb{E}XX^\top$, and $\Sigma_{YY} = \mathbb{E}YY^\top$. Since computing $\mathbb{E}XX^\top$ and $\mathbb{E}YY^\top$ is not affordable, the projected stochastic gradient algorithms are not applicable. Thus we consider an alternative approach to avoid the projection operation. We consider the Lagrangian function of (3.29) as

$$L(u, v, \mu, \sigma) = u^\top \Sigma_{XY} v - \mu(u^\top \Sigma_{XX} u - 1) - \sigma(v^\top \Sigma_{YY} v - 1) \tag{3.30}$$

where μ and σ are Lagrangian multipliers. We then check the optimal KKT conditions,

$$\Sigma_{XY} v - 2\Sigma_{XX}\mu u = 0, \quad \Sigma_{XY} u - 2\Sigma_{YY}\sigma v = 0, \quad u^\top \Sigma_{XX} u = 1 \quad \text{and} \quad v^\top \Sigma_{YY} v = 1$$

which further imply

$$u^\top \Sigma_{XY} v - 2\mu u^\top \Sigma_{XX} u = u^\top \Sigma_{XY} v - 2\mu = 0 \quad \text{and} \quad v^\top \Sigma_{XY} u - 2\sigma v^\top \Sigma_{YY} v = v^\top \mathbb{E}YX^\top u - 2\sigma = 0$$

Solving the above equations, we obtain the optimal Lagrangian multipliers as

$$\mu = \sigma = \frac{1}{2} u^\top \mathbb{E}XY^\top v \tag{3.31}$$

Similarly, we then apply the dual free stochastic gradient method to solve (3.29). Specifically, at the k-th iteration, we independently sample (X_k, Y_k) and $(\widetilde{X}_k, \widetilde{Y}_k)$ from $\mathcal{D}$. Then we obtain

$$u_{k+1} = u_k + \eta\Big(\widetilde{X}_k \widetilde{Y}_k^\top v_k - u_k^\top X_k Y_k^\top v_k \cdot \widetilde{X}_k \widetilde{X}_k^\top u_k\Big), \quad v_{k+1} = v_k + \eta\Big(\widetilde{Y}_k \widetilde{X}_k^\top u_k - v_k^\top Y_k X_k^\top u_k \cdot \widetilde{Y}_k \widetilde{Y}_k^\top v_k\Big) \tag{3.32}$$

Here we sample two pairs of X and Y to ensure the unbiasedness of the stochastic gradient.

Then we can convert (3.32) to ordinary differential equations by taking $\eta \to 0^+$, we get

$$\frac{dU}{dt} = \Sigma_{XY} V - U^\top \Sigma_{XY} V \cdot \Sigma_{XX} U, \quad \frac{dV}{dt} = \Sigma_{XY}^\top U - V^\top \Sigma_{XY}^\top U \cdot \Sigma_{YY} V$$

Different from PLS, the above ordinary differential equations do not admit a closed form solution, which makes our ODE/SDE-type convergence analysis not applicable in a straightforward manner. A possible alternative approach is to establish the lower bounds for $|\widehat{u}^\top U(t)|$ and $|\widehat{v}^\top V(t)|$, and further prove that as $t \to \infty$, we have $U(t) \to \widehat{u}$ and $V(t) \to \widehat{v}$. We will leave this option for further investigation.

4 Exploring Primal-Dual Dynamics in Constrained Nonconvex Optimization

To formulate, we often encounter the following optimization problem in machine learning, signal processing, and stochastic control:

$$\min_X f(X) \quad \text{subject to} \quad X \in \Omega \tag{4.1}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is a loss function, $\Omega :\triangleq \{X \in \mathbb{R}^d : g_i(X) = 0, i = 1, 2, ..., m\}$ denotes a feasible set, m is the number of constraints, and $g_i : \mathbb{R}^d \to \mathbb{R}$'s are the differentiable functions that impose constraints into model parameters. For notational simplicity, we define $\mathcal{G}(X) = [g_1(X), ..., g_m(X)]^\top$ and $\Omega = \{X \in \mathbb{R}^d : \mathcal{G}(X) = 0\}$.

Main Results To handle the complicated Ω, this section proposes to investigate the min-max problem:

$$\min_{X\in\mathbb{R}^d}\max_{Y\in\mathbb{R}^m}\ \mathcal{L}(X,Y) := f(X) + Y^\top \mathcal{G}(X) \tag{4.2}$$

where $Y \in \mathbb{R}^m$ is the Lagrangian multiplier. $\mathcal{L}(X,Y)$ is often referred as the Lagrangian function in existing literature [BV04]. Specifically, we first define a special class of Lagrangian functions, where the landscape of $\mathcal{L}(X,Y)$ enjoys the following good properties:

- *There exist only two types of equilibria – stable and unstable equilibria. At an unstable equilibrium,* $\mathcal{L}(X,Y)$ *has negative curvature with respect to the primal variable* X*. More details in Section 4.1.*
- *All stable equilibria correspond to the global optima of the primal problem* (4.1).

Both properties are intuitive. On the one hand, the negative curvature in the first property enables the primal variable to escape from the unstable equilibria along some decent direction. On the other hand, the second property ensures that we do not get spurious local optima of (4.1), that is all local minima must also be global optima.

We then study a generalized eigenvalue (GEV) problem, which includes CCA, Fisher discriminant analysis (FDA, [MRW+99]), sufficient dimension reduction (SDR, [CN05]) as special examples. Specifically, GEV solves

$$X^* = \operatorname*{argmin}_{X\in\mathbb{R}^{d\times r}}\ f(X) := -\operatorname{tr}(X^\top AX)\ \text{ s.t. }\ X\in\mathcal{T}_B := \{X\in\mathbb{R}^{d\times r} : X^\top BX = I_r\} \tag{4.3}$$

where $A, B \in \mathbb{R}^{d\times d}$ are symmetric, B is positive semidefinite. We rewrite (4.3) as a min-max problem,

$$\min_X \max_Y \mathcal{L}(X,Y) = -\operatorname{tr}(X^\top AX) + \langle Y, X^\top BX - I_r\rangle \tag{4.4}$$

where $Y \in \mathbb{R}^{r\times r}$ is the Lagrangian multiplier. Theoretically, we show that the Lagrangian function in (4.4) exactly belongs to our previously defined class. Motivated by our defined landscape structures, we then solve an online version of (4.4), where we can only access independent unbiased stochastic approximations of A, B and directly accessing A and B is prohibited. Specifically, at the k-th iteration, we only obtain independent $A^{(k)}$ and $B^{(k)}$ satisfying

$$\mathbb{E}A^{(k)} = A \quad \text{and} \quad \mathbb{E}B^{(k)} = B$$

Computationally, we propose a simple stochastic primal-dual algorithm, which is a stochastic variant of the generalized Hebbian algorithm (GHA, [Gor06]). Theoretically, we establish its asymptotic rate of convergence to stable equilibria for our stochastic GHA (SGHA) based on the diffusion approximations [KY03]. Specifically, we show that, asymptotically, the solution trajectory of SGHA weakly converges to the solutions of stochastic differential equations (SDEs). By studying the analytical solutions of these SDEs, we further establish the asymptotic sample/iteration complexity of SGHA under certain regularity conditions [HKY97, LWLZ16]. To the best of our knowledge, this is the first asymptotic sample/iteration complexity analysis of a stochastic optimization algorithm for solving the online version of GEV problem. Numerical experiments are presented to justify our theory.

Our work is closely related to several recent results on solving GEV problems. For example, [GJN+16] propose a multistage semi-stochastic optimization algorithm for solving GEV problems with a finite sum structure. At each optimization stage, their algorithm needs to access the exact

B matrix, and compute the approximate inverse of B by solving a quadratic program, which is not allowed in our setting. Similar matrix inversion approaches are also adopted by a few other recently proposed algorithms for solving GEV problem [AZL16, AMMS17]. In contrast, our proposed SGHA is a fully stochastic algorithm, which does not require any matrix inversion.

Moreover, our work is also related to several more complicated min-max problems, such as Markov Decision Process with function approximation, Generative Adversarial Network, multistage stochastic programming and control [SMSM00, SDR09, GPAM$^+$14]. Many primal-dual algorithms have been proposed to solve these problems. However, most of these algorithms are even not guaranteed to converge. As mentioned earlier, when the convex-concave structure is missing, the min-max problems go far beyond the existing theories. Moreover, both primal and dual iterations involve sophisticated stochastic approximations (equally or more difficult than our online version of GEV). This section makes the attempt on understanding the optimization landscape of these challenging min-max problems. Taking our results as an initial start, we expect more sophisticated and stronger follow-up works that apply to these min-max problems.

The remainder of the section is organized as follows: Section 2 characterizes the equilibria of the Lagrangian function and introduces the theoretical properties of stable and unstable equilibria. Section 3 presents the generalized eigenvalue problem and formulates it as a min-max problem. Section 4 describes the proposed stochastic primal-dual algorithm and provides theoretical convergence analysis. Section 5 presents numerical results on synthetic data. Finally, Section 6 concludes the section and discusses potential future work.

4.1 Characterization of Equilibria

Recall the Lagrangian function in (4.2). Then we start with characterizing its equilibria. By KKT conditions, an equilibrium (X, Y) satisfies

$$\nabla_X \mathcal{L}(X,Y) = \nabla_X f(X) + Y^\top \nabla_X \mathcal{G}(X) = 0 \quad \text{and} \quad \nabla_Y \mathcal{L}(X,Y) = \mathcal{G}(X) = 0$$

which only contains the first order information of $\mathcal{L}(X, Y)$. To further distinguish the difference among the equilibria, we define two types of equilibria by the second order information.

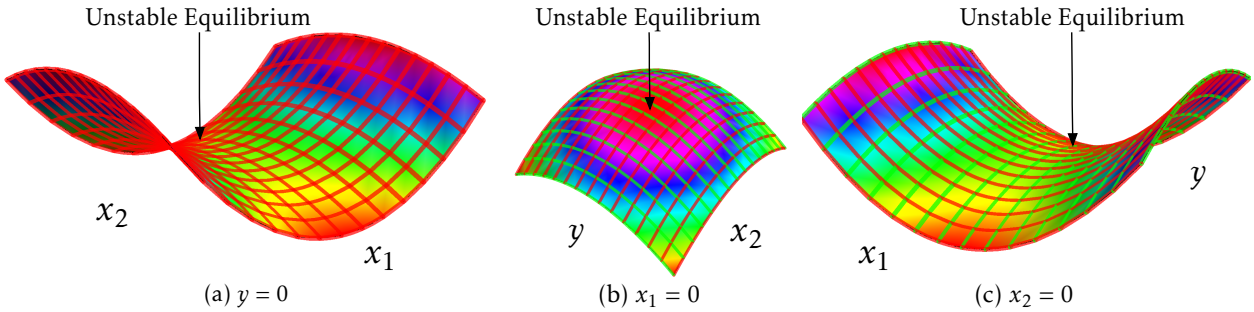
Definition 2. *Given the Lagrangian function $\mathcal{L}(X, Y)$ in (4.2), a point (X, Y) is called:*

- **(1)** *An **equilibrium** of $\mathcal{L}(X, Y)$, if*

$$\nabla \mathcal{L}(X,Y) = \begin{bmatrix} \nabla_X \mathcal{L}(X,Y) \\ \nabla_Y \mathcal{L}(X,Y) \end{bmatrix} = 0$$

- **(2)** *An **equilibrium** (X, Y) is **unstable**, if (X, Y) is an equilibrium and $\lambda_{\min}\left(\nabla_X^2 \mathcal{L}(X,Y)\right) < 0$.*
- **(3)** *An **equilibrium** (X, Y) is **stable**, if (X, Y) is an equilibrium, $\nabla_X^2 \mathcal{L}(X,Y) \succeq 0$, and $\mathcal{L}(X, Y)$ is strongly convex over a restricted domain.*

Note that (2) in Definition 2 has a similar strict saddle property over a manifold in [GHJY15]. The motivation behind Definition 2 is intuitive. When $\mathcal{L}(X, Y)$ has negative curvature with respect to the primal variable X at an equilibrium, we can find a direction in X to further decrease $\mathcal{L}(X, Y)$. Therefore, a tiny perturbation can break this unstable equilibrium. An illustrative example is presented in Figure 1. Moreover, at a stable equilibrium (X^*, Y^*), there is restricted



(a) $y = 0$ (b) $x_1 = 0$ (c) $x_2 = 0$

Figure 1. An illustration of an unstable equilibrium: $\min_{x_1,x_2} \max_y \mathcal{L}(x_1, x_2, y) = x_1^2 - x_2^2 - y^2$. Notice that $(0, 0, 0)$ is an equilibrium but unstable. For visualization, we show three views: (a) $\mathcal{L}(x_1, x_2, 0)$; (b) $\mathcal{L}(0, x_2, y)$; (c) $\mathcal{L}(x_1, 0, y)$. The red lines correspond to x_1 and x_2, and the green one corresponds to the y.

strong convexity, which relates to several conditions, e.g., Polyak Łojasiewicz conditions [Pol63], i.e.

$$\|\nabla_X \mathcal{L}(X, Y^*)\|^2 \geq \mu(\mathcal{L}(X, Y^*) - \mathcal{L}(X^*, Y^*))$$

for X belonging to a small region near X^* and $\mu > 0$ is a constant, or Error Bound conditions [LT93]. With this property, we cannot decrease $\mathcal{L}(X, Y)$ along any direction with respect to X. Definition 2 excludes the high order unstable equilibrium, which may exist due to the degeneracy of $\nabla_X^2 \mathcal{L}(X, Y)$. Specifically, such a high order unstable equilibrium cannot be identified by the second order information, e.g.

$$\mathcal{L}(x_1, x_2, y) = x_1^3 + x_2^2 + y \cdot (x_1 - x_2)$$

$(0, 0, 0)$ is an equilibrium with a positive semidefinite Hessian matrix. However, it is an unstable equilibria, since a small perturbation to x_1 can break this equilibrium. Such an equilibrium makes the landscape highly more complicated. Overall, we consider a specific class of Lagrangian functions throughout the rest of this section. They enjoy the following properties:

- All equilibria are either stable or unstable (i.e., no high order unstable equilibria);
- All stable equilibria correspond to the global optima of the primal problem.

As mentioned earlier, the first property ensures that the second order information can identify the type of equilibria. The second property guarantees that we do not get spurious optima for (4.1) as long as an algorithm attains a stable equilibrium. Several machine learning problems belong to this class, such as the generalized eigenvalue decomposition problem.

4.2 Generalized Eigenvalue Decomposition

We consider the generalized eigenvalue (GEV) problem as a motivating example, which includes CCA, FDA, SDR, etc. as special examples. Recall its min-max formulation (4.4):

$$\min_{X \in \mathbb{R}^{d \times r}} \max_{Y \in \mathbb{R}^{r \times r}} \mathcal{L}(X, Y) = -\operatorname{tr}(X^\top A X) + \langle Y, X^\top B X - I_r \rangle$$

Before we proceed, we impose the following assumption on the problem.

Assumption 4. *Given a symmetric matrix $A \in \mathbb{R}^{d\times d}$ and a positive definite matrix $B \in \mathbb{R}^{d\times d}$, the eigenvalues of $\widetilde{A} = B^{-\frac{1}{2}}AB^{-\frac{1}{2}}$, denoted by $\lambda_1^{\widetilde{A}}, ..., \lambda_d^{\widetilde{A}}$, satisfy*

$$\lambda_1^{\widetilde{A}} \geq \cdots \geq \lambda_r^{\widetilde{A}} > \lambda_{r+1}^{\widetilde{A}} \geq \cdots \geq \lambda_d^{\widetilde{A}}$$

Such an eigengap assumption avoids the identifiability issue. The full rank assumption on B in Assumption 4 ensures that the original constrained optimization problem is bounded. This assumption can be further relaxed but require more involved analysis. We will discuss this in Appendix B.2.

To characterize all equilibria of GEV, we leverage the idea of an invariant group. [LWL+16b] use similar techniques for an unconstrained matrix factorization problem. However, it does not work for the Lagrangian function due to the more complicate landscape. Therefore, we consider a more general invariant group. Moreover, by analyzing the Hessian matrix of $\mathcal{L}(X, Y)$ at the equilibria, we demonstrate that each equilibrium is either unstable or stable and the stable equilibria correspond to the global optima of the primal problem (4.3). Therefore, GEV belongs to the class we defined earlier.

4.2.1 Invariant Group and Symmetric Property

We first denote the orthogonal group in dimension r as

$$O(r, \mathbb{R}) = \left\{\Psi \in \mathbb{R}^{r\times r} : \Psi\Psi^\top = \Psi^\top\Psi = I_r\right\}$$

Notice that for any $\Psi \in O(r, \mathbb{R})$, $\mathcal{L}(X, Y)$ in (4.4) has the same landscape with $\mathcal{L}(X\Psi, \Psi^\top Y\Psi)$. This further indicates that given an equilibrium (X, Y), $(X\Psi, \Psi^\top Y\Psi)$ is also an equilibrium. This symmetric property motivates us to characterize the equilibria of $\mathcal{L}(X, Y)$ with an invariant group.

We introduce several important definitions in group theory [DF04].

Definition 3. *Given a group $\mathcal{H}$ and a set $\mathcal{X}$, a map $\phi(\cdot, \cdot)$ from $\mathcal{H} \times \mathcal{X}$ to $\mathcal{X}$ is called the* **group action** *of $\mathcal{H}$ on $\mathcal{X}$ if ϕ satisfies the following two properties:*

Identity: $\phi(\mathbf{1}, x) = x \quad \forall x \in \mathcal{X}$, *where* $\mathbf{1}$ *denotes the identity element of* $\mathcal{H}$.

Compatibility: $\phi(gh, x) = \phi(g, \phi(h, x)) \quad \forall g, h \in \mathcal{H}, \quad x \in \mathcal{X}$.

Definition 4. *Given a function $f(x, y) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, a group $\mathcal{H}$ is a* ***stationary invariant group*** *of f with respect to two group actions of $\mathcal{H}$, ϕ_1 on $\mathcal{X}$ and ϕ_2 on $\mathcal{Y}$, if $\mathcal{H}$ satisfies*

$$f(x, y) = f(\phi_1(g, x), \phi_2(g, y)) \quad \forall x \in \mathcal{X}, \quad y \in \mathcal{Y}, \quad \text{and} \quad g \in \mathcal{H}$$

For notational simplicity, we denote $\mathcal{G} = O(r, \mathbb{R})$. Given the group $\mathcal{G}$, two sets $\mathbb{R}^{d\times r}$ and $\mathbb{R}^{r\times r}$, we define a group action with ϕ_1 of $\mathcal{G}$ on $\mathbb{R}^{d\times r}$ and a group action ϕ_2 of $\mathcal{G}$ on $\mathbb{R}^{r\times r}$ as

$$\phi_1(\Psi, X) = X\Psi \quad \forall \Psi \in \mathcal{G}, \ X \in \mathbb{R}^{d\times r} \quad \text{and} \quad \phi_2(g, Y) = \Psi^{-1}Y\Psi \quad \forall \Psi \in \mathcal{G}, \ Y \in \mathbb{R}^{r\times r}$$

One can check that the orthogonal group $\mathcal{G}$ is a stationary invariant group of $\mathcal{L}(X, Y)$ with respect to two group actions of $\mathcal{G}$, ϕ_1 on $\mathbb{R}^{d\times r}$ and ϕ_2 on $\mathbb{R}^{r\times r}$. By this invariant group, we define the equivalence relation between (X_1, Y_1) and (X_2, Y_2), if there exists a $\Psi \in \mathcal{G}$ such that

$$(X_1, Y_1) = (X_2\Psi, \Psi^{-1}Y_2\Psi) = (X_2\Psi, \Psi^\top Y_2\Psi) \tag{4.5}$$

To find all equilibria of GEV, we examine the KKT conditions of (4.4):

$$2BXY - 2AX = 0 \;\text{ and }\; X^\top BX - I_r = 0 \Longrightarrow Y = X^\top AX =: \mathcal{D}(X)$$

Given the eigenvalue decomposition $B = O^B \Lambda^B O^{B\top}$, we denote

$$\widetilde{A} = (\Lambda^B)^{-\frac{1}{2}} O^{B\top} A O^B (\Lambda^B)^{-\frac{1}{2}} \quad \text{and} \quad \widetilde{X} = (\Lambda^B)^{\frac{1}{2}} O^{B\top} X$$

We then consider the eigenvalue decomposition $\widetilde{A} = O^{\widetilde{A}} \Lambda^{\widetilde{A}} O^{\widetilde{A}\top}$. The following theorem shows the connection between the equilibrium of $\mathcal{L}(X, Y)$ and the column submatrix of $O^{\widetilde{A}}$, denoted as $O^{\widetilde{A}}_{:,\mathcal{I}}$, where

$$\mathcal{I} \in \mathcal{X}_d^r := \Big\{ \{i_1, ..., i_r\} : \{i_1, ..., i_r\} \subseteq [d] \Big\}$$

is the column index set to determine a column submatrix.

Theorem 4.1 (Symmetric Property)**.** *Suppose Assumption 4 holds. Then $(X, \mathcal{D}(X))$ is an equilibrium of $\mathcal{L}(X, Y)$, if and only if X can be written as*

$$X = (O^B (\Lambda^B)^{-\frac{1}{2}} O^{\widetilde{A}}_{:,\mathcal{I}}) \cdot \Psi$$

where index $\mathcal{I} \in \mathcal{X}_d^r$ and $\Psi \in \mathcal{G}$.

The proof of Theorem 4.1 is provided in Appendix B.1.1. Theorem 4.1 implies that there are $\binom{d}{r}$ equilibria of $\mathcal{L}(X, Y)$ under the equivalence relation given in (4.5). Each of them corresponds to an $O^{\widetilde{A}}_{:,\mathcal{I}}$, where $\mathcal{I} \in \mathcal{X}_d^r$ is the index set. Then whole equilibria set is generated by these $O^{\widetilde{A}}_{:,\mathcal{I}}$ with the transformation matrix $O^B (\Lambda^B)^{-\frac{1}{2}}$ and the invariant group action induced by $\mathcal{G}$.

4.2.2 Unstable Equilibrium vs. Stable Equilibrium

We further identify the stable and unstable equilibria. Specifically, given (X, Y) as an equilibrium of $\mathcal{L}(X, Y)$, we denote the Hessian matrix of $\mathcal{L}(X, Y)$ with respect to the primal variable X as

$$H_X \triangleq \nabla_X^2 \mathcal{L}(X, Y)|_{Y=\mathcal{D}(X)} \in \mathbb{R}^{dr \times dr}$$

Then we calculate the eigenvalues of H_X. By Definition 2, $(X, \mathcal{D}(X))$ is unstable if H_X has a negative eigenvalue; Otherwise, we analyze the local landscape at $(X, \mathcal{D}(X))$ to determine whether it is stable or not. The following theorem shows that all equilibria are either stable or unstable and demonstrates how the choice of index set $\mathcal{I}$ corresponds to the unstable and stable equilibria of $\mathcal{L}(X, Y)$.

Theorem 4.2. *Suppose Assumption 4 holds, and $(X, \mathcal{D}(X))$ is an equilibrium in (4.4). By Theorem 4.1, X can be represented as $X = (O^B (\Lambda^B)^{-\frac{1}{2}} O^{\widetilde{A}}_{:,\mathcal{I}}) \cdot \Psi$ for some $\Psi \in \mathcal{G}$ and $\mathcal{I} \in \mathcal{X}_d^r$.*

If $\mathcal{I} \neq [r]$, then $(X, \mathcal{D}(X))$ is an unstable equilibrium with

$$\lambda_{\min}(H_X) \leq \frac{2(\lambda^{\widetilde{A}}_{\max \mathcal{I}} - \lambda^{\widetilde{A}}_{\min \mathcal{I}^\perp})}{\|X_{:,\min \mathcal{I}^\perp}\|_2^2} < 0$$

where $\lambda^{\widetilde{A}}_{\max \mathcal{I}} = \max_{i \in \mathcal{I}} \lambda_i^{\widetilde{A}}$, and $\lambda^{\widetilde{A}}_{\min \mathcal{I}} = \min_{i \in \mathcal{I}} \lambda_i^{\widetilde{A}}$, $\lambda_i^{\widetilde{A}}$ is the i-th leading eigenvalue of $\widetilde{A}$.

Otherwise, we have $H_X \succeq 0$ and $\operatorname{rank}(H_X) = d \times r - r(r-1)/2$. Moreover, $(X, \mathcal{D}(X))$ is a stable equilibrium of min-max problem (4.4).

The proof of Theorem 4.2 is provided in Appendix B.1.2. Theorem 4.2 indicates that when $\widetilde{X} = O^{\widetilde{A}}_{:,[r]}$, that is, the eigenvectors of $\widetilde{A}$ corresponding to the r largest eigenvalues, $(X, \mathcal{D}(X))$ is a stable equilibrium of $\mathcal{L}(X, Y)$, where $X = (O^B(\Lambda^B)^{-\frac{1}{2}} O^{\widetilde{A}}_{:,\mathcal{I}})) \cdot \Psi$ for some $\Psi \in \mathcal{G}$. Although H_X is degenerate at this equilibrium, all directions in Null(H_X) essentially point to the primal variables of other stable equilibria. Excluding these directions, the rest all have positive curvature, which implies that this equilibrium is stable. Moreover, such an X corresponds to the optima of (4.3). When $\mathcal{I} \neq [r]$, due to the negative curvature, these equilibria are unstable. Therefore, all stable equilibria of $\mathcal{L}(X, Y)$ correspond to the global optima in (4.3) and other equilibria are unstable, which further indicates that GEV belongs to the class we defined earlier.

4.3 Stochastic Search for Online GEV

For GEV, we propose a fully stochastic primal-dual algorithm to solve (4.4), which only requires access to the stochastic approximations of A and B matrices. This is very different from other existing semi-stochastic algorithms that require to access the exact B matrix [GJN$^+$16]. Specifically, we propose a stochastic variant of the generalized Hebbian algorithm (GHA), also referred as Sanger's rule in existing literature [San89], to solve (4.4). For online setting, accessing the exact A and B is prohibitive and we only get $A^{(k)} \in \mathbb{R}^{d\times d}$ and $B^{(k)} \in \mathbb{R}^{d\times d}$ that are independently sampled from the distribution associated with A and B at the k-th iteration. Our proposed SGHA updates primal and dual variables as follows:

$$\text{Primal Update: } X^{(k+1)} \leftarrow X^{(k)} - \eta \cdot \underbrace{\left(B^{(k)}X^{(k)}Y^{(k)} - A^{(k)}X^{(k)}\right)}_{\text{Stochastic Approximation of } \nabla_X \mathcal{L}(X^{(k)}, Y^{(k)})} \tag{4.6}$$

$$\text{Dual Update: } Y^{(k+1)} \leftarrow \underbrace{X^{(k)\top}A^{(k)}X^{(k)}}_{\text{Stochastic Approximation of } X^{(k)\top}AX^{(k)}} \tag{4.7}$$

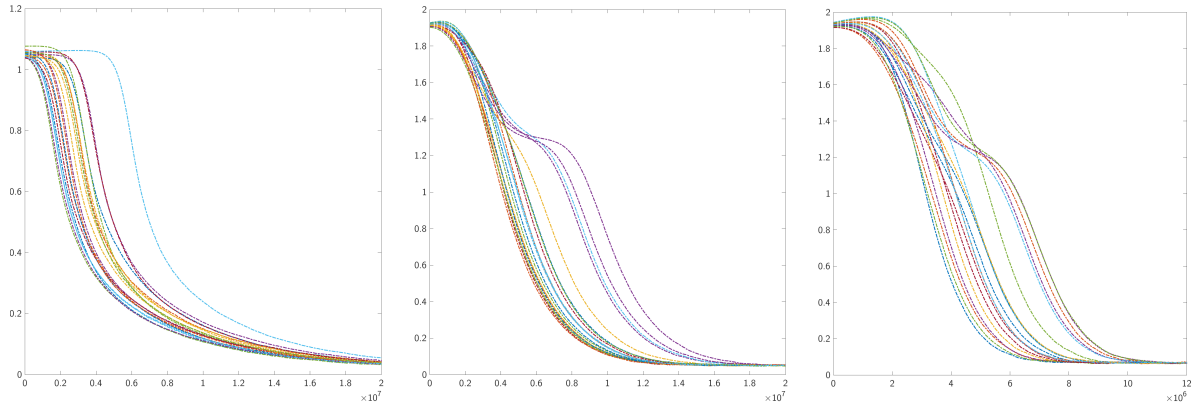
where $\eta > 0$ is a step size parameter. Note that the primal update is a stochastic gradient descent step, while the dual update is motivated by the KKT conditions of (4.4). SGHA is simple and easy to implement. The constraint is naturally handled by the dual update. Further, motivated by the the landscape of GEV, we analyze the algorithm by diffusion approximations and obtain the asymptotical sample complexity.

4.3.1 Numerical Evaluations

We first provide numerical evaluations to illustrate the effectiveness of SGHA, and then provide an asymptotic convergence analysis of SGHA. We choose $d = 500$ and select three different settings:

- **Setting(1) :** $\eta = 10^{-4}$, $r = 1$, $A_{ii} = 1/100 \ \ \forall i \in [d]$, $A_{ij} = 0.5/10$ and $B_{ij} = 0.5^{|i-j|}/3 \ \ \forall i \neq j$;
- **Setting(2) :** $\eta = 5 \times 10^{-5}$, $r = 3$, and randomly generate an orthogonal matrix $U \in \mathbb{R}^{d\times d}$ such that $A = U \cdot \text{diag}(1, 1, 1, 0.1, ..., 0.1) \cdot U^\top$ and $B = U \cdot \text{diag}(2, 2, 2, 1, ..., 1) \cdot U^\top$;
- **Setting(3) :** $\eta = 2.5 \times 10^{-5}$, $r = 3$, and randomly generate two orthogonal matrices $U, V \in \mathbb{R}^{d\times d}$ such that $A = U \cdot \text{diag}(1, 1, 1, 0.1, ..., 0.1) \cdot U^\top$ and $B = V \cdot \text{diag}(2, 2, 2, 1, ..., 1) \cdot V^\top$.

At the k-th iteration of SGHA, we independently sample 40 random vectors from $N(0, A)$ and $N(0, B)$ respectively. Accordingly, we compute the sample covariance matrices $A^{(k)}$ and $B^{(k)}$ as



(a) Setting (1) (b) Setting (2) (c) Setting (3)

Figure 2. Plots of the optimization error $\|B^{1/2}X^{(t)}X^{(t)\top}B^{1/2} - B^{1/2}X^*X^{*\top}B^{1/2}\|_{\mathrm{F}}$ over SGHA iterations on synthetic data of 20 random data generations under different settings of parameters.

the approximations of A and B. We repeat numerical simulations under each setting for 20 times using random data generations, and present all results in Figure 2. The horizontal axis corresponds to the number of iterations, and the vertical axis corresponds to the optimization error

$$\|B^{1/2}X^{(t)}X^{(t)\top}B^{1/2} - B^{1/2}X^*X^{*\top}B^{1/2}\|_{\mathrm{F}}$$

Our experiments indicate that SGHA converges to a global optimum in all settings.

4.3.2 Convergence Analysis for Commutative A and B

As a special case, we first prove the convergence of SGHA for GEV with $r = 1$, and A and B are commutative. We will discuss more on noncommutative cases and $r > 1$ in the next section. Before we proceed, we introduce our assumptions on the problem.

Assumption 5. *We assume that the following conditions hold:*

- **(a):** *$A^{(k)}$'s and $B^{(k)}$'s are independently sampled from two different distributions $\mathcal{D}_A$ and $\mathcal{D}_B$ respectively, where $\mathbb{E}A^{(k)} = A$ and $\mathbb{E}B^{(k)} = B \succ 0$;*
- **(b):** *A and B are commutative, i.e., there exists an orthogonal matrix O such that $A = O\Lambda^A O^\top$ and $B = O\Lambda^B O^\top$, where $\Lambda^A = \mathrm{diag}(\lambda_1, ..., \lambda_d)$ and $\Lambda^B = \mathrm{diag}(\mu_1, ..., \mu_d)$ are diagonal matrices with $\lambda_j \neq 0$;*
- **(c):** *$A^{(k)}$ and $B^{(k)}$ satisfy the moment conditions, that is, for some generic constants C_0 and C_1, $\mathbb{E}\|A^{(k)}\|_2^2 \leq C_0$ and $\mathbb{E}\|B^{(k)}\|_2^2 \leq C_1$.*

Note that (a) and (c) in (5) are mild, but (b) is stringent. For convenience of analysis, we combine (4.6) and (4.7) as

$$X^{(k+1)} \leftarrow X^{(k)} - \eta\big(B^{(k)}X^{(k)}X^{(k)\top} - I_d\big)A^{(k)}X^{(k)} \tag{4.8}$$

We remark that (4.8) is very different from existing optimization algorithms over the generalized Stiefel manifold. Specifically, computing the gradient over the generalized Stiefel manifold requires

B^{-1}, which is not allowed in our setting. For notational convenience, we further denote

$$\Lambda = (\Lambda^B)^{-\frac{1}{2}}\Lambda^A(\Lambda^B)^{-\frac{1}{2}} = \mathrm{diag}\left(\frac{\lambda_1}{\mu_1}, ..., \frac{\lambda_d}{\mu_d}\right) =: \mathrm{diag}(\beta_1, \cdots, \beta_d)$$

Without loss of generality, we assume $\beta_1 > \beta_2 \geq \beta_3 \geq \cdots \geq \beta_d$, and $\beta_i \neq 0 \ \ \forall i \in [d]$. Note that μ_i and λ_i, however, are not necessarily to be monotonic. We denote

$$\mu_{\min} = \min_{i\neq 1} \mu_i, \quad \mu_{\max} = \max_{i\neq 1} \mu_i, \quad \text{and} \quad \mathsf{gap} = \beta_1 - \beta_2$$

Denote $W^{(k)} = (\Lambda^B)^{\frac{1}{2}}OX^{(k)}$. One can verify that (4.8) can be rewritten as follows:

$$W^{(k+1)} \leftarrow W^{(k)} - \eta\Big((\Lambda^B)^{\frac{1}{2}}\widehat{\Lambda}_B^{(k)}(\Lambda^B)^{-\frac{1}{2}} \cdot W^{(k)}W^{(k)\top} - \Lambda^B\Big) \cdot \widetilde{\Lambda}^{(k)}W^{(k)} \tag{4.9}$$

where $\widehat{\Lambda}_B^{(k)} = O^\top B^{(k)}O$ and $\widetilde{\Lambda}^{(k)} = O^\top B^{-\frac{1}{2}}A^{(k)}B^{-\frac{1}{2}}O$. Note that $W^* = (1, \underbrace{0, 0, ..., 0}_{(d-1)})^\top$ corresponds to the optimal solution of (4.3).

By diffusion approximation, we show that our algorithm converges through three Phases:

- **Phase I:** Given an initial near a saddle point, we show that after rescaling of time properly, the algorithm can be characterized by a stochastic differential equation (SDE). Such an SDE further implies our algorithm can escape from the saddle fast;
- **Phase II:** We show that away from the saddle, the trajectory of our algorithm can be approximated by an ordinary differential equation (ODE);
- **Phase III:** We first show that after Phase II, the norm of solution converges to a constant. Then, the algorithm can be characterized by an SDE, like Phase I. By the SDE, we analyze the error fluctuation when the solution is within a small neighborhood of the global optimum.

Overall, we obtain an asymptotic sample complexity.

ODE Characterization: To demonstrate an ODE characterization for the trajectory of our algorithm, we introduce a continuous time random process

$$w^{(\eta)}(t) := W^{(k)}$$

where $k = \lfloor \frac{t}{\eta} \rfloor$ and η is the step size in (4.8). For notational simplicity, we drop (t) when it is clear from the context. Instead of showing a global convergence of $w^{(\eta)}$, we show that the quantity

$$v_{i,j}^{(\eta)} = \frac{(w_i^{(\eta)})^{\mu_j}}{(w_j^{(\eta)})^{\mu_i}}$$

converges to an exponential decay function, where $v_i^{(\eta)}$ is the i-th component (coordinate) of $w^{(\eta)}$.

Lemma 1. *Suppose that Assumption 5 holds and the initial solution is away from any saddle point, i.e., given pre-specified constants, $\tau > 0$ and $\delta < \frac{1}{2}$, there exist i, j such that*

$$i \neq j, \quad |w_j^{(\eta)}| > \tau, \quad \textit{and} \quad |w_i^{(\eta)}| > \eta^{\frac{1}{2}+\delta}$$

As $\eta \to 0$, $v_{k,j}^{(\eta)}$ weakly converges to the solution of the following ODE:

$$dx_{k,j} = x_{k,j} \cdot (\mu_j\mu_k(\beta_k - \beta_j))\, dt \ \ \forall k \neq j \tag{4.10}$$

The proof of Lemma 1 is provided in Appendix B.3.1. Lemma 1 essentially implies the global convergence of SGHA. Specifically, the solution of (4.10) is

$$x_{k,j}(t) = x_{k,j}(0) \cdot \exp\left(\mu_j \mu_k \left(\beta_k - \beta_j\right) t\right) \quad \forall k \neq j$$

where $x_{k,j}(0)$ is the initial value of $v_{k,j}^{(\eta)}$. In particular, we consider $j = 1$. Then, as $t \to \infty$, the dominating component of w will be w_1.

The ODE approximation of the algorithm implies that after long enough time, i.e., t is large enough, the solution of the algorithm can be arbitrarily close to a global optimum. Nevertheless, to obtain the asymptotic "convergence rate", we need to study the variance of the trajectory at time t. Thus, we resort to the following SDE-based approach for a more precise characterization.

SDE Characterization: We notice that such a variance with order $\mathcal{O}(\eta)$ vanishes as $\eta \to 0$. To characterize this variance, we rescale the updates by a factor of $\eta^{-\frac{1}{2}}$, i.e., by defining a new process as $z^{(\eta)} = \eta^{-\frac{1}{2}} w^{(\eta)}$. After rescaling, the variance of $z^{(\eta)}$ is of order $\mathcal{O}(1)$. The following lemma characterizes how the algorithm escapes from the saddle, i.e., $w^{(\eta)}(0) \approx e_i$, where $i \neq 1$, in Phase I.

Lemma 2. *Suppose Assumption 5 holds and the initial is close to a saddle point , i.e., $z_j^{(\eta)}(0) \approx \eta^{-\frac{1}{2}}$ and $z_i^{(\eta)}(0) \approx 0$ for $i \neq j$. Then for any $C > 0$, there exist $\tau > 0$ and $\eta' > 0$ such that*

$$\sup_{\eta < \eta'} \mathbb{P}(\sup_t |z_i^{(\eta)}(t)| \leq C) \leq 1 - \tau \tag{4.11}$$

Here we provide the proof sketch and leave the whole proof of Lemma 2 in Appendix B.3.2.

Proof Sketch. We prove this argument by contradiction. Assume the conclusion does not hold, that is there exists a constant $C > 0$, such that for any $\eta' > 0$ we have

$$\sup_{\eta \leq \eta'} \mathbb{P}(\sup_t |z_i^{(\eta)}(t)| \leq C) = 1$$

That implies there exists a sequence $\{\eta_n\}_{n=1}^{\infty}$ converging to 0 such that

$$\lim_{n \to \infty} \mathbb{P}(\sup_t |z_i^{(\eta_n)}(t)| \leq C) = 1 \tag{4.12}$$

Then we show $\{z_i^{(\eta_n)}(\cdot)\}_n$ is tight and thus converges weakly. Furthermore, $\{z_i^{(\eta_n)}(\cdot)\}_n$ weakly converges to a stochastic differential equation,

$$dz_j(t) = \left(-\beta_j \mu_i \cdot z_i + \lambda_i z_i\right) dt + \sqrt{G_{j,i}} dB(t) \quad \text{for } j \in [d] \backslash \{i\}, \tag{4.13}$$

where $G_{j,i} = \mathbb{E}\Big(\big(\widehat{\Lambda}_B^{(k)}\big)_{j,i} \cdot \sqrt{\mu_j/\mu_i} \cdot \widetilde{\Lambda}_{i,i} - \mu_j \widetilde{\Lambda}_{j,i}\Big)^2$ and $B(t)$ is a standard Brownian motion. We compute the solution of this stochastic differential equation and then show (4.11) holds. □

Note that (4.13) is a Fokker-Plank equation, whose solution is an Ornstein-Uhlenbeck (O-U) process [Doo42] as follows:

$$z_j(t) = \underbrace{\Big[z_j(0) + \sqrt{G_{j,i}} \int_0^t \exp\left[\mu_j \left(\beta_i - \beta_j\right) s\right] dB(s) \Big]}_{Q_1} \cdot \exp\left[-\mu_j \left(\beta_i - \beta_j\right) t\right] \tag{4.14}$$

We consider $j = 1$. Note that Q_1 is essentially a random variable with mean $z_j(0)$ and variance smaller than $\frac{G_{1,i}\mu_1}{2(\beta_1-\beta_i)}$. However, the larger t is, the closer its variance gets to this upper bound. Moreover, the term $\exp\left[\mu_1(\beta_1 - \beta_i)t\right]$ essentially amplifies Q_1 by a factor exponentially increasing in t. This tremendous amplification forces $z_1(t)$ to quickly get away from 0, as t increases, which indicates that the algorithm will escape from the saddle. Further, the following lemma characterizes the local behavior of the algorithm near the optimal.

Lemma 3. *Suppose that Assumption 5 holds and the initial solution is close to an optimal solution, that is, given pre-specified constants κ and $\delta < \frac{1}{2}$, we have $\frac{|w_1^{(\eta)}|^2}{\|w^{(\eta)}\|_2^2} > 1 - \kappa\eta^{1+2\delta}$. As $\eta \to 0$, then we have $\|w^{(\eta)}(t)\|_2 \xrightarrow{t\to\infty} 1$ and $z_i^{(\eta)}$ weakly converges to the solution of the following SDE:*

$$dz_i(t) = (-\beta_1 \cdot \mu_i z_i + \lambda_i z_i)\, dt + \sqrt{G_{i,1}} dB(t) \quad \text{for } i \neq 1, \tag{4.15}$$

where $G_{i,1} = \mathbb{E}\big((\widehat{\Lambda}_B)_{i,1} \cdot \sqrt{\mu_i/\mu_1} \cdot \widetilde{\Lambda}_{1,1} - \mu_i\Lambda_{i,1}\big)^2$, and $B(t)$ is a standard Brownian motion.

The proof of Lemma 3 is provided in Appendix B.3.3. The solution of (4.15) is as follows:

$$z_i(t) = \sqrt{G_{i,1}} \int_0^t \exp\left[\mu_i\left(\beta_1 - \beta_i\right)(s-t)\right] dB(s) + z_i(0) \cdot \exp\left[-\mu_i\left(\beta_1 - \beta_i\right)t\right] \tag{4.16}$$

Note the second term of the right hand side in (4.16) decays to 0, as time $t \to \infty$. The rest is a pure random walk. Thus, the fluctuation of $z_i(t)$ is essentially the error fluctuation of the algorithm after sufficiently long time.

Combining Lemma 1, 2, and 3, we obtain the following theorem.

Theorem 4.3. *Suppose Assumption 5 holds. Given a sufficiently small error $\epsilon > 0$, $\phi = \sum_{i=1}^d G_{i,1}$, and*

$$\eta \asymp \frac{\epsilon \cdot \mu_{\min} \cdot \mathsf{gap}}{\phi}$$

we need

$$T \asymp \frac{\mu_{\max}/\mu_{\min}}{\mu_1 \cdot \mathsf{gap}} \log\left(\eta^{-1}\right) \tag{4.17}$$

such that with probability at least $\frac{5}{8}$, $\|w(T) - W^*\|_2^2 \leq \epsilon$, where W^* is the optima of (4.3).

The proof of Theorem 4.3 is provided in Appendix B.3.4. Theorem 4.3 implies that asymptotically, our algorithm yields an iterations of complexity:

$$N \asymp \frac{T}{\eta} \asymp \frac{\phi \cdot \mu_{\max}/\mu_{\min}}{\epsilon \cdot \mu_1 \cdot \mu_{\min} \cdot \mathsf{gap}^2} \log\left(\frac{\phi}{\epsilon \cdot \mu_{\min} \cdot \mathsf{gap}}\right)$$

which not only depends on the gap, i.e., $\beta_1 - \beta_2$, but also depends on $\frac{\mu_{\max}}{\mu_{\min}}$, which is the condition number of B in the worst case. As can be seen, for an ill-conditioned B, the problem (4.3) is more difficult to solve.

When A and B are Noncommutative? Unfortunately, when A and B are noncommutative, the analysis is more difficult, even for $r = 1$. Recall that the optimization landscape of the Lagrangian function in (4.4) enjoys a nice geometric property: At an unstable equilibrium, the negative curvature with respect to the primal variable encourages the algorithm to escape. Specifically, suppose the algorithm is initialized at an unstable equilibrium $(X^{(0)}, Y^{(0)})$, the descent direction for $X^{(0)}$ is determined by the eigenvectors of

$$H_{X^{(0)}} = A + Y^{(0)}B$$

associated with the negative eigenvalues. After one iteration, we obtain $(X^{(1)}, Y^{(1)})$. The Hessian matrix becomes

$$H_{X^{(1)}} = A + Y^{(1)}B$$

Since $Y^{(1)} = X^{(0)\top}A^{(0)}X^{(0)}$ is a stochastic approximation, the random noise can make $Y^{(1)}$ significantly different from $Y^{(0)}$. Thus, the eigenvectors of $H_{X^{(1)}}$ associated with the negative eigenvalues can be also very different from those of $H_{X^{(0)}}$. This phenomenon can seriously confuse the algorithm about the descent direction of the primal variable. We remark that such an issue does not appear if we assume A and B are commutative. We suspect that this is very likely an artifact of our proof technique, since our numerical experiments have provided some empirical evidences of the convergence of SGHA.

5 Conclusion

Multiview representation learning plays a crucial role in latent factor analysis across fields like machine learning, data analysis, and information retrieval, modeling dependent structures among diverse data sources. Traditionally, convex optimization frameworks have been favored due to their theoretical guarantees of global optimality. However, empirical evidence shows that nonconvex approaches, especially those using stochastic algorithms, can achieve comparable or even superior performance, despite lacking formal guarantees. Motivated by this gap, we introduced a nonconvex formulation for multiview learning that leverages stochastic gradient descent (SGD).

Our work rigorously established the convergence rates of SGD for solving online Partial Least Squares (PLS), using diffusion process approximation to characterize the nonconvex optimization landscape. By dropping convexity, we demonstrated significant improvements in efficiency and scalability over convex methods. Our analysis tightens existing results by a factor of $O(1/\epsilon)$, where ϵ is a pre-specified error, and we provided theoretical grounding for the global convergence of the algorithm.

Moreover, we extended our framework to Canonical Correlation Analysis (CCA), demonstrating its broad applicability in multiview learning tasks. Numerical experiments confirmed that our nonconvex approach not only accelerates convergence but also reduces computational and memory requirements compared to traditional methods. We believe future work should explore extensions to more complex multiview scenarios, higher-rank solutions, and adaptive step-size schemes, as these represent promising directions for both theory and practice.

In this paper, we explored the landscape of Lagrangian functions for constrained nonconvex optimization problems, frequently encountered in machine learning, signal processing, and stochastic control. These problems are typically reformulated as minimax problems using Lagrangian functions, but the nonconvex nature of the feasible set poses significant challenges in understanding the landscape and identifying stable equilibria.

To address this, we analyzed the structure of the Lagrangian landscape and defined a special class of Lagrangian functions. These functions exhibit two key properties: (1) equilibria are either stable or unstable, and (2) stable equilibria correspond to global optima of the original problem. This classification fills a crucial gap in understanding nonconvex optimization landscapes.

We applied our framework to the generalized eigenvalue problem (GEV), a class of problems that includes canonical correlation analysis (CCA) and matrix factorization. By leveraging symmetry and invariant group properties, we characterized both stable and unstable equilibria. Based on these insights, we developed a fully stochastic primal-dual algorithm to solve the online GEV problem, which is efficient both computationally and theoretically.

Theoretically, we derived conditions for asymptotic convergence and provided the first sample complexity results for solving the online GEV problem using diffusion approximations, a widely adopted technique in stochastic control. Empirically, we demonstrated the efficiency of our approach through numerical experiments, confirming the algorithm's convergence to global optima and its practical viability.

Our work is closely related to several recent advances in the field:

- [LWL+16b] propose a framework for characterizing the stationary points in unconstrained nonconvex matrix factorization problems. However, our focus is on the constrained generalized eigenvalue problem (GEV). Unlike their approach, we analyze the optimization landscape by considering both primal and dual variables, making the analysis more challenging.

- [GJN+16] also study the generalized eigenvalue problem, but in an offline setting with a finite sum structure. Their method requires access to exact matrices A and B in each iteration, along with a modified Gram Schmidt process to enforce the solution on the generalized Stiefel manifold. In contrast, our proposed stochastic primal-dual algorithm operates in an online setting, without requiring access to exact matrices or enforcing strict manifold constraints.

This work not only advances the theoretical understanding of constrained nonconvex optimization but also provides practical algorithms with strong convergence guarantees, opening the door for future applications and refinements in this area.

References

[Abd03] Hervé Abdi. Partial least square regression (pls regression). *Encyclopedia for research methods for the social sciences*, pages 792–795, 2003.

[ACLS12] Raman Arora, Andrew Cotter, Karen Livescu, and Nathan Srebro. Stochastic optimization for PCA and PLS. In *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*, pages 861–868. IEEE, 2012.

[AL12] Raman Arora and Karen Livescu. Kernel cca for multi-view learning of acoustic features using articulatory measurements. In *MLSLP*, pages 34–37. Citeseer, 2012.

[AMM16] Raman Arora, Poorya Mianjy, and Teodor Marinov. Stochastic optimization for multiview representation learning using partial least squares. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 1786–1794, 2016.

[AMMS17] Raman Arora, Teodor Vanislavov Marinov, Poorya Mianjy, and Nati Srebro. Stochastic approximation for canonical correlation analysis. In *Advances in Neural Information Processing Systems*, pages 4778–4787, 2017.

[AZ05] Rie Kubota Ando and Tong Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6(Nov):1817–1853, 2005.

[AZL16] Zeyuan Allen-Zhu and Yuanzhi Li. Doubly accelerated methods for faster CCA and generalized eigendecomposition. *arXiv preprint arXiv:1607.06017*, 2016.

[BALHJ12] Sujeeth Bharadwaj, Raman Arora, Karen Livescu, and Mark Hasegawa-Johnson. Multiview acoustic feature learning using articulatory measurements. In *Intl. Workshop on Stat. Machine Learning for Speech Recognition*. Citeseer, 2012.

[BC05] Andrew D Barbour and Louis Hsiao Yun Chen. *An introduction to Stein's method*, volume 4. World Scientific, 2005.

[BNS16] Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Global optimality of local search for low rank matrix recovery. *arXiv preprint arXiv:1605.07221*, 2016.

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[CKLS09] Kamalika Chaudhuri, Sham M Kakade, Karen Livescu, and Karthik Sridharan. Multi-view clustering via canonical correlation analysis. In *Proceedings of the 26th annual international conference on machine learning*, pages 129–136. ACM, 2009.

[CLM+16] Michael B Cohen, Yin Tat Lee, Gary Miller, Jakub Pachocki, and Aaron Sidford. Geometric median in nearly linear time. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing*, pages 9–21. ACM, 2016.

[CLO14] Yunmei Chen, Guanghui Lan, and Yuyuan Ouyang. Optimal primal-dual methods for a class of saddle point problems. *SIAM Journal on Optimization*, 24(4):1779–1814, 2014.

[CLS15] Emmanuel J Candes, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

[CN05] R Dennis Cook and Liqiang Ni. Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *Journal of the American Statistical Association*, 100(470):410–428, 2005.

[DF04] David Steven Dummit and Richard M Foote. *Abstract algebra*, volume 3. Wiley Hoboken, 2004.

[DFU11] Paramveer Dhillon, Dean P Foster, and Lyle H. Ungar. Multi-view learning of word embeddings via cca. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 199–207. Curran Associates, Inc., 2011.

[Doo42] Joseph L Doob. The brownian movement and stochastic equations. *Annals of Mathematics*, pages 351–369, 1942.

[EK09] Stewart N Ethier and Thomas G Kurtz. *Markov processes: characterization and convergence*, volume 282. John Wiley & Sons, 2009.

[Eva88] WD Evans. Partial differential equations, 1988.

[FHT01] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[GHJY15] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.

[GJN+16] Rong Ge, Chi Jin, Praneeth Netrapalli, Aaron Sidford, et al. Efficient algorithms for large-scale generalized eigenvector computation and canonical correlation analysis. In *International Conference on Machine Learning*, pages 2741–2750, 2016.

[GLM16] Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981, 2016.

[Gor06] Genevieve Gorrell. Generalized hebbian algorithm for incremental singular value decomposition in natural language processing. In *EACL*, volume 6, pages 97–104. Citeseer, 2006.

[GPAM+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[GVL12] Gene H Golub and Charles F Van Loan. *Matrix computations*, volume 3. JHU Press, 2012.

[HKY97] J Harold, G Kushner, and George Yin. Stochastic approximation and recursive algorithm and applications. *Application of Mathematics*, 35, 1997.

[HSST04] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.

[IN14] Anatoli Iouditski and Yuri Nesterov. Primal-dual subgradient methods for minimizing uniformly convex functions. *arXiv preprint arXiv:1401.1792*, 2014.

[JJK+16] Prateek Jain, Chi Jin, Sham M Kakade, Praneeth Netrapalli, and Aaron Sidford. Streaming PCA: Matching matrix Bernstein and near-optimal finite sample guarantees for Oja's algorithm. In *29th Annual Conference on Learning Theory*, pages 1147–1164, 2016.

[KSE05] Einat Kidron, Yoav Y Schechner, and Michael Elad. Pixels that sound. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 88–95. IEEE, 2005.

[KY03] Harold Kushner and G George Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

[LLM11] Guanghui Lan, Zhaosong Lu, and Renato DC Monteiro. Primal-dual first-order methods with {\mathcal {O}(1/\epsilon)} iteration-complexity for cone programming. *Mathematical Programming*, 126(1):1–29, 2011.

[LT93] Zhi-Quan Luo and Paul Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46(1):157–178, 1993.

[LWL16a] Chris Junchi Li, Zhaoran Wang, and Han Liu. Online ica: Understanding global dynamics of nonconvex optimization via diffusion processes. In *Advances in Neural Information Processing Systems*, pages 4967–4975, 2016.

[LWL+16b] Xingguo Li, Zhaoran Wang, Junwei Lu, Raman Arora, Jarvis Haupt, Han Liu, and Tuo Zhao. Symmetry, saddle points, and global geometry of nonconvex matrix factorization. *arXiv preprint arXiv:1612.09296*, 2016.

[LWLZ16] Chris J Li, Mengdi Wang, Han Liu, and Tong Zhang. Near-optimal stochastic approximation for online principal component estimation. *arXiv preprint arXiv:1603.05305*, 2016.

[LY+84] David G Luenberger, Yinyu Ye, et al. *Linear and nonlinear programming*, volume 2. Springer, 1984.

[MRW+99] Sebastian Mika, Gunnar Ratsch, Jason Weston, Bernhard Scholkopf, and Klaus-Robert Mullers. Fisher discriminant analysis with kernels. In *Neural networks for signal processing IX, 1999. Proceedings of the 1999 IEEE signal processing society workshop.*, pages 41–48. Ieee, 1999.

[Now13] Brian David Nowakowski. On multi-parameter semimartingales, their integrals and weak convergence. 2013.

[Øks03] Bernt Øksendal. Stochastic differential equations. In *Stochastic differential equations*, pages 65–84. Springer, 2003.

[Pol63] Boris Teodorovich Polyak. Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 3(4):643–653, 1963.

[R+11] Nathan Ross et al. Fundamentals of stein's method. *Probab. Surv*, 8:210–293, 2011.

[San89] Terence D Sanger. Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural networks*, 2(6):459–473, 1989.

[SDR09] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, 2009.

[SFF10] Richard Socher and Li Fei-Fei. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 966–973. IEEE, 2010.

[Sha15] Ohad Shamir. Fast stochastic algorithms for svd and pca: Convergence properties and convexity. *arXiv preprint arXiv:1507.08788*, 2015.

[SMSM00] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pages 1057–1063, 2000.

[SQW16] Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retrieval. In *Information Theory (ISIT), 2016 IEEE International Symposium on*, pages 2379–2383. IEEE, 2016.

[VSTC02] Alexei Vinokourov, John Shawe-Taylor, and Nello Cristianini. Inferring a semantic representation of text via cross-language correlation analysis. In *NIPS*, volume 1, page 4, 2002.

[ZLTW17] Zhihui Zhu, Qiuwei Li, Gongguo Tang, and Michael B Wakin. The global optimization geometry of nonsymmetric matrix factorization and sensing. *arXiv preprint arXiv:1703.01256*, 2017.

[ZWL15] Tuo Zhao, Zhaoran Wang, and Han Liu. A nonconvex optimization framework for low rank matrix estimation. In *Advances in Neural Information Processing Systems*, pages 559–567, 2015.

A Appendix for Section 3

A.1 Proof Detailed Proofs in Section 3.2

A.1.1 Proof of Proposition 1

Proof. We consider a compact singular value decomposition of Σ_{XY} as follow:

$$\Sigma_{XY} = \sum_{i=1}^{r} \lambda_i \overline{u}_i \overline{v}_i^\top$$

where $\lambda_1 > \lambda_2 \geq ... \geq \lambda_r > 0$ are nonzero singular values, and $(\overline{u}_i,\ \overline{v}_i)$'s are a pair of singular vectors associated with λ_i. Plugging (3.7) into (3.6), we have

$$\Sigma_{XY} v - (u^\top \Sigma_{XY} v)u = 0 \quad \text{and} \quad \Sigma_{XY}^\top u - (u^\top \Sigma_{XY} v)v = 0 \tag{A.1}$$

Since every vector $u \in \mathbb{R}^m$ and $v \in \mathbb{R}^d$ can be expanded as

$$u = \sum_{i=1}^{r} c_i \overline{u}_i + \sum_{j=r+1}^{m} c_j \overline{u}_j \quad \text{and} \quad v = \sum_{i=1}^{r} l_i \overline{v}_i + \sum_{j=r+1}^{d} l_j \overline{v}_j \tag{A.2}$$

where $\overline{u}_j$ for $j = r+1, ..., m$ and $\overline{v}_j$ for $j = r+1, ..., d$ are orthonormal basis vectors, and complementary to $\overline{u}_i$'s and $\overline{v}_i$'s for $i = 1, ..., r$ in $\mathbb{R}^m$ and $\mathbb{R}^d$ respectively, and c_i's and l_i's are the coefficients. Plugging (A.2) into the first equation of (A.1), we get

$$\begin{aligned}
0 &= \sum_{i=1}^{r} \lambda_i \overline{u}_i \overline{v}_i^\top \cdot \sum_{i=1}^{d} c_i \overline{v}_i - \sum_{i=1}^{m} l_i \overline{u}_i \cdot \sum_{i=1}^{r} \lambda_i \overline{u}_i \overline{v}_i^\top \cdot \sum_{i=1}^{d} c_i \overline{v}_i \cdot \sum_{i=1}^{m} l_i \overline{u}_i \\
&= \sum_{i=1}^{r} c_i \lambda_i \overline{u}_i - \sum_{i=1}^{m} \big(\sum_{k=1}^{r} l_k \lambda_k c_k \big) \cdot l_i \overline{u}_i \\
&= \sum_{i=1}^{r} \left(c_i \lambda_i - \big(\sum_{k=1}^{r} l_k \lambda_k c_k \big) \cdot l_i \right) \overline{u}_i - \sum_{i=r+1}^{m} \big(\sum_{k=1}^{r} l_k \lambda_k c_k \big) \cdot l_i \overline{u}_i
\end{aligned} \tag{A.3}$$

The second equality holds because $\overline{u}_i$ and $\overline{v}_j$ are the columns of the orthogonal matrices. Since $\overline{u}_i$'s are the basis vectors of $\mathbb{R}^m$, by (A.3), we know the coefficients of all $\overline{u}_i$'s should be 0. Therefore we consider two scenarios:

(i) If $\sum_{i=1}^{r} l_k \lambda_k c_k = 0$, then we have $c_i = 0$, $i = 1, 2, ..., r$. Similarly, plugging (A.2) into the second equation of (A.1), we have $l_i = 0$, $i = 1, 2, ..., r$. Thus, u and v are in the row and column null space of Σ_{XY} respectively.

(ii) If $\sum_{i=1}^{r} l_k \lambda_k c_k \neq 0$, then we have $l_i = 0$, $i = r+1, ..., m$, which further leads to:

$$c_i \lambda_i = \big(\sum_{k=1}^{r} l_k \lambda_k c_k \big) \cdot l_i \quad \text{and} \quad l_i \lambda_i = \big(\sum_{k=1}^{r} l_k \lambda_k c_k \big) \cdot c_i \quad \text{for } i = 1, 2, ..., r \tag{A.4}$$

Note that (A.4) holds if and only if there exists only one $i \in \{1, 2, ..., r\}$. $c_j = l_j = \pm \delta_{ij},\ j = 1, 2, ..., r$, where δ_{ij} is the Kronecker delta, i.e., $\delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$.

The verification of the above points satisfying (A.1) is straightforward, and therefore omitted. □

A.1.2 Proof of Proposition 2

Proof. For notation simplicity, we denote $\nabla^2_{u,v}L(u,v)$ as $\nabla^2_{u,v}L(u,v,\mu,\sigma)\Big|_{\mu=\sigma=\frac{1}{2}u^\top Av}$

$$\nabla^2_{u,v}L(u,v) = \begin{pmatrix} -u^\top\Sigma_{XY}v\cdot I_m & \Sigma_{XY} \\ \Sigma_{XY}^\top & -u^\top\Sigma_{XY}v\cdot I_d \end{pmatrix}$$

a. If u and v are in the row and column null space of Σ_{XY} respectively, then

$$\nabla^2_{u,v}L(u,v) = \begin{pmatrix} 0 & \Sigma_{XY} \\ \Sigma_{XY}^\top & 0 \end{pmatrix} \quad \text{and} \quad \lambda_{\max}(\nabla^2_{u,v}L(u,v)) = \lambda_1$$

Therefore, it is an unstable stationary point because of the positive curvature.

b. If (u,v) is a pair of singular vector of λ_i, then by simple linear algebra, we know that

$$\nabla^2_{u,v}L(u,v) \sim \begin{pmatrix} -u^\top\Sigma_{XY}v\cdot I_m & 0 \\ 0 & \frac{1}{u^\top\Sigma_{XY}v}\Sigma_{XY}^\top\Sigma_{XY} - u^\top\Sigma_{XY}v\cdot I_d \end{pmatrix}$$

One can verify

$$\lambda_{\max}(\nabla^2_{u,v}L(u,v)) = \frac{\lambda_1^2-\lambda_i^2}{\lambda_i} \geq \lambda_1 - \lambda_2$$

Therefore, the Hessian matrix is negative semi-definite if and only if $u^\top\Sigma_{XY}v = \lambda_1$, i.e., (u,v) is the optimum of (3.1). The Hessian has a positive eigenvalue.

Thus, only the optima of (3.5) are stable stationary points. All the others are unstable. □

A.2 Proof Detailed Proofs in Section 3.3

A.2.1 Proof of Theorem 3.1

Proof. First, we calculate the infinitesimal conditional expectation. Since the optimization problem is symmetric about u and v, we only prove the claim for u,

$$\begin{aligned}\frac{d}{dt}\mathbb{E}\left(U_\eta(t)-U_\eta(0)\right)\big|_{t=0} &= \eta^{-1}\mathbb{E}\left(U_\eta(\eta)-U_\eta(0)\big|U_\eta(0),V_\eta(0)\right) \\ &= \Sigma_{XY}V(0) - U(0)^\top\Sigma_{XY}V(0)U(0)\end{aligned}$$

Next, we show that if the initial is on the sphere, then with probability 1, all iterations are on the sphere as $\eta\to 0^+$. Given $\|u_k\|_2 = \|v_k\|_2 = 1$, we have

$$\begin{aligned}\|u_{k+1}\|_2^2 &= \Big(u_k+\eta\cdot(X_kY_k^\top v_k - u_k^\top X_kY_k^\top v_k u_k)\Big)^\top\cdot\Big(u_k+\eta\cdot(X_kY_k^\top v_k - u_k^\top X_kY_k^\top v_k u_k)\Big) \\ &= u_k^\top u_k + 2\eta(u_k^\top X_kY_k^\top v_k - u_k^\top X_kY_k^\top v_k u_k^\top u_k) + \eta^2\|X_kY_k^\top v_k - u_k^\top X_kY_k^\top v_k u_k\|_2^2 \\ &= 1+\eta^2\|X_kY_k^\top v_k - u_k^\top X_kY_k^\top v_k u_k\|_2^2\end{aligned}$$

Therefore, we get

$$\mathbb{P}\Big(\lim_{\eta\to 0^+}\|u_{k+1}\|_2 = 1\Big|\|u_k\|_2=1\Big) = \mathbb{P}(|X_k^\top Y_k|<\infty) = 1$$

The last equality holds, since $\mathbb{E}|X_k^\top Y_k|$ is finite:

$$\mathbb{E}|X_k^\top Y_k| \le \sqrt{\mathbb{E}\|X_k\|_2^2 \cdot \mathbb{E}\|Y_k\|_2^2} \le B^2 d$$

Finally, we bound the infinitesimal conditional variance.

$$\begin{aligned}
&\frac{d}{dt}\mathbb{E}\left(U_\eta^{(j)}(t) - U_\eta^{(j)}(0)\right)^2\Big|_{t=0} \\
&\le \eta^{-1}\cdot \mathrm{tr}\left(\mathbb{E}\left[\left(U_\eta(\eta) - U_\eta(0)\right)\left(U_\eta(\eta) - U_\eta(0)\right)^\top\right)\Big| U_\eta(0) = u_k,\ V_\eta(0) = v_k\right] \\
&= \eta^{-1}\cdot \mathbb{E}\left[\eta\left(X_kY_k^\top u_k - u_k^\top X_kY_k^\top v_k u_k\right)^\top \cdot \eta\left(X_kY_k^\top u_k - u_k^\top X_kY_k^\top v_k u_k\right)\right] \\
&= \eta\cdot \mathbb{E}\left(u_k^\top Y_kX_k^\top X_kY_k^\top u_k - 2u_k^\top Y_kX_k^\top u_k u_k^\top X_kY_k^\top v_k + u_k^\top u_k (u_k^\top X_kY_k^\top v_k)^2\right) \\
&\le \eta\cdot\left(\sqrt{\mathbb{E}\|X_k\|_2^4\mathbb{E}\|Y_k\|_2^4} + 2\sqrt{\mathbb{E}(u_k^\top Y_kX_k^\top u_k)^2\mathbb{E}(u_k^\top Y_kX_k^\top v_k)^2} + \mathbb{E}(u_k^\top X_kY_k^\top v_k)^2\right) \\
&\le \eta\cdot\left(\sqrt{\mathbb{E}\|X_k\|_2^4\mathbb{E}\|Y_k\|_2^4} + 3\mathbb{E}(|Y_k^\top||X_k|)^2\right) \\
&= O(\eta)
\end{aligned}$$

Last equality holds by the Assumption 2.

Therefore, by Section 4 of Chapter 7 in [EK09], we know that, as $\eta \to 0^+$, $U_\eta(t)$ and $V_\eta(t)$ weakly converge to the solution of (3.11) with the same initial. By definition of $U_\eta(t)$ and $V_\eta(t)$, we complete the proof. □

A.2.2 Proof of Theorem 3.2

Proof. Since P is an orthonormal matrix, $\|H_j\|_2 = \|W_j\|_2 = 1$ for all $j = 1, ..., d$. Thus, we have

$$\begin{aligned}
\frac{d}{dt}H^{(i)} &= \lambda_i H^{(i)} - \sum_{j=1}^{2d}\lambda_j (H^{(j)})^2 H^{(i)} \\
&= \lambda_i \sum_{j=1}^{2d}(H^{(j)})^2 H^{(i)} - \sum_{j=1}^{2d}\lambda_j (H^{(j)})^2 H^{(i)} \\
&= H^{(i)}\sum_{j=1}^{2d}\left(\lambda_i - \lambda_j\right)(H^{(j)})^2
\end{aligned}$$

We then verify (3.17) satisfies (3.16). By [Eva88], we know that since $H_j(t)$ is continuously differentiable in t, the solution to the ODE is unique. For notational simplicity, we denote

$$S^{(j)}(t) = H^{(j)}(0)\exp(\lambda_j t)$$

Then we have

$$H^{(i)}(t) = \frac{S^{(i)}(t)}{\sqrt{\sum_{j=1}^{2d}\left(S^{(j)}(t)\right)^2}}$$

Now we only need to verify

$$\begin{aligned}\frac{d}{dt}H^{(i)}(t) &= \frac{\left(\lambda_i S^{(i)}(t)\right)\sqrt{\sum_{j=1}^{2d}\left(S^{(j)}(t)\right)^2} - \frac{\left(2\sum_{j=1}^{2d}\lambda_j\left(S^{(j)}(t)\right)^2\right)S^{(i)}(t)}{2\sqrt{\sum_{j=1}^{2d}\left(S^{(j)}(t)\right)^2}}}{\sum_{j=1}^{2d}\left(S^{(j)}(t)\right)^2} \\ &= \lambda_i\frac{S^{(i)}(t)}{\sqrt{\sum_{j=1}^{2d}\left(S^{(j)}(t)\right)^2}} - \sum_{j=1}^{2d}\lambda_j\frac{\left(S^{(j)}(t)\right)^2}{\sum_{j=1}^{2d}\left(S^{(j)}(t)\right)^2}\frac{S^{(i)}(t)}{\sqrt{\sum_{j=1}^{2d}\left(S^{(j)}(t)\right)^2}} \\ &= \lambda_i H^{(i)}(t) - \sum_{j=1}^{2d}\lambda_j\left(H^{(j)}(t)\right)^2 H^{(i)}(t)\end{aligned}$$

which completes the proof. □

A.3 Proof Detailed Proofs in Section 3.4

A.3.1 Proof of Theorem 3.3

Proof. We prove this by contradiction. Assume the conclusion does not hold, that is there exists a constant $C > 0$, such that for any $\eta' > 0$ we have

$$\sup_{\eta\le\eta'}\mathbb{P}(\sup_t|Z_\eta^{(i)}(t)| \le C) = 1$$

That implies there exists a sequence $\{\eta_n\}_{n=1}^\infty$ converging to 0 such that

$$\lim_{n\to\infty}\mathbb{P}(\sup_t|Z_{\eta_n}^{(i)}(t)| \le C) = 1 \tag{A.5}$$

Thus, condition (i) in Theorem 2.4 [Now13] holds. We next check the second condition. When $\sup_t|Z_{\eta_n}^{(i)}(t)| \le C$ holds, Assumption 2 yields that $z_{\eta_n,k+1}^{(i)} - z_{\eta_n,k}^{(i)} = C'\eta_n$, where C' is some constant. Thus, for any $t, \epsilon > 0$, we have

$$|Z_{\eta_n}^{(i)}(t) - Z_{\eta_n}^{(i)}(t+\epsilon)| = \frac{\epsilon}{\eta}C'\eta = C'\epsilon$$

Thus, condition (ii) in Theorem Theorem 2.4 [Now13] holds. Then we have $\{Z_{\eta_n}^{(i)}(\cdot)\}_n$ is tight and thus converges weakly.

We then calculate the infinitesimal conditional expectation and variance for $Z_{\eta_n}^{(i)},\ i \ne j$.

$$\begin{aligned}\frac{d}{dt}\mathbb{E}Z_{\eta_n}^{(i)}(t)\big|_{t=0} &= \eta_n^{-1}\mathbb{E}\left[Z_{\eta_n}^{(i)}(\eta_n) - Z_{\eta_n}^{(i)}(0)\big|H_{\eta_n}(0) = h\right] \\ &= \eta_n^{-1}\mathbb{E}\left[\eta_n^{-1/2}\left(H_{\eta_n}^{(i)}(\eta_n) - H_{\eta_n}^{(i)}(0)\right)\big|H_{\eta_n}(0) = h\right] \\ &= \eta_n^{-1/2}h^{(i)}\sum_{l=1}^{2d}\left(\lambda_i - \lambda_l\right)(h^{(l)})^2 = Z_{\eta_n}^{(i)}\left(\lambda_i - \lambda_j\right) + o(1),\end{aligned} \tag{A.6}$$

where the last equality comes from the assumption that the algorithm starts near j^{th} column of $P,\ j \ne 1$, i.e., $h \approx e_j$. To compute variance, we first compute $\widehat{\Lambda}$,

$$\widehat{\Lambda} = P^\top QP = \frac{1}{2}\begin{pmatrix} \overline{Y}\,\overline{X}^\top + \overline{X}\,\overline{Y}^\top & \overline{Y}\,\overline{X}^\top - \overline{X}\,\overline{Y}^\top \\ -\overline{Y}\,\overline{X}^\top + \overline{X}\,\overline{Y}^\top & -\overline{Y}\,\overline{X}^\top - \overline{X}\,\overline{Y}^\top \end{pmatrix}$$

where Q is defined in (3.12). Then we analyze $e_i^\top \widehat{\Lambda} e_j$ by cases:

$$e_i^\top \widehat{\Lambda} e_j = \begin{cases} \frac{1}{2}\left(\overline{X}^{(i)}\overline{Y}^{(j)} + \overline{X}^{(j)}\overline{Y}^{(i)}\right) & \text{if } \max(i,j) \le d, \\ \frac{1}{2}\left(-\overline{X}^{(j)}\overline{Y}^{(i-d)} + \overline{X}^{(i-d)}\overline{Y}^{(j)}\right) & \text{if } j \le d < i, \\ \frac{1}{2}\left(\overline{X}^{(j-d)}\overline{Y}^{(i)} - \overline{X}^{(i)}\overline{Y}^{(j-d)}\right) & \text{if } i \le d < j, \\ \frac{1}{2}\left(-\overline{X}^{(i-d)}\overline{Y}^{(j-d)} - \overline{X}^{(j-d)}\overline{Y}^{(i-d)}\right) & \text{if } \min(i,j) > d \end{cases}$$

which further implies

$$\begin{aligned} \frac{d}{dt}\mathbb{E}(Z_{\eta_n}^{(i)}(t) - Z_{\eta_n}^{(i)}(0))^2\big|_{t=0} &= \eta_n^{-1}\mathbb{E}\big[\big(Z_{\eta_n}^{(i)}(\eta) - Z_{\eta_n}^{(i)}(0)\big)^2 \big| H_{\eta_n}(0) = h\big] \\ &= \eta_n^{-2}\mathbb{E}[\eta_n^2(\widehat{\Lambda}h - h^\top\widehat{\Lambda}hh)(\widehat{\Lambda}h - h^\top\widehat{\Lambda}hh)^\top]_{i,i} \\ &= \mathbb{E}(e_i^\top\widehat{\Lambda}e_j e_j^\top\widehat{\Lambda}^\top e_i) + o(1) \\ &= \frac{1}{4}\big(\gamma_i\omega_j + \gamma_j\omega_i + 2\,\mathrm{sign}(i-d-1/2)\cdot\mathrm{sign}(j-1/2-d)\cdot\alpha_{ij}\big) \quad \text{(A.7)} \end{aligned}$$

By (A.6) and (A.7), we get the limit stochastic differential equation,

$$dZ^{(i)}(t) = -(\lambda_j - \lambda_i)Z^{(i)}(t)dt + \beta_{ij}dB(t)$$

Therefore, $\{Z_{\eta_n}^{(i)}(\cdot)\}$ converges weakly to a solution of The process defined by the equation above is an unstable O-U process with mean 0 and exploding variance. Thus, for any τ, there exist a time t', such that

$$\mathbb{P}(|Z^{(i)}(t')| \ge C) \ge 2\tau$$

Since $\{Z_{\eta_n}^{(i)}\}_n$ converges weakly to Z^i, thus $\{Z_{\eta_n}^{(i)}(t')\}_n$ converges in distribution to $Z^i(t')$. This implies that there exists an $N > 0$, such that for any $n > N$

$$|\mathbb{P}(|Z^i(T)| \ge C) - \mathbb{P}(|Z_{\eta_n}^{(i)}(T)| \ge C)| \le \tau$$

Then we find a t' such that

$$\mathbb{P}(|Z_{\eta_n}^{(i)}(t')| \ge C) \ge \tau, \forall n > N,$$

or equivalently

$$\mathbb{P}(|Z_{\eta_n}^{(i)}(t')| \le C) < 1 - \tau, \forall n > N$$

Since $\left\{\omega \middle| \sup_t |Z_{\eta_n}^{(i)}(t)(\omega)| \le C\right\} \subset \left\{\omega \middle| |Z_{\eta_n}^{(i)}(\tau')(\omega) < C\right\}$, we have

$$\mathbb{P}(\sup_t |H^{\eta_n,i}(t)| \le C\sqrt{\eta_n}) = \mathbb{P}(\sup_t |Z_{\eta_n}^{(i)}(t)| \le C) \le 1 - \delta, \forall n > N,$$

which leads to a contradiction with A.5. Our assumption does not hold.

□

A.3.2 Proof of Proposition 4

Proof. Our analysis is based on approximating $z_{\eta,k}^{(1)}$ by its continuous approximation $Z_{\eta}^{(1)}(t)$, which is normal distributed at time t. By simple manipulation, we have

$$\mathbb{P}\left((h_{\eta,N_1}^{(2)})^2 \leq 1-\delta^2\right) = \mathbb{P}\left((z_{\eta,N_1}^{(2)})^2 \leq \eta^{-1}(1-\delta^2)\right) \geq \mathbb{P}(|z_{\eta,N_1}^{(1)}| \geq \eta^{-\frac{1}{2}}\delta).$$

We then prove $P\left(\left|z_{\eta,N_1}^{(1)}\right| \geq \eta^{-\frac{1}{2}}\delta\right) \geq 1-\nu$. At time t, $z_{\eta,k}^{(1)}$ approximates to a normal distribution with mean 0 and variance $\frac{\beta_{12}^2}{2(\lambda_1-\lambda_2)}\left[\exp\left(2(\lambda_1-\lambda_2)\eta N_1\right)-1\right]$. Therefore, let $\Phi(x)$ be the CDF of $N(0,1)$, we have

$$\mathbb{P}\left(\frac{\left|z_{\eta,N_1}^{(1)}\right|}{\sqrt{\frac{\beta_{12}^2}{2(\lambda_1-\lambda_2)}\cdot\left[\exp\left(2(\lambda_1-\lambda_2)\eta N_1\right)-1\right]}} \geq \Phi^{-1}\left(\frac{1+\nu}{2}\right)\right) \approx 1-\nu$$

which requires

$$\eta^{-\frac{1}{2}}\delta \leq \Phi^{-1}\left(\frac{1+\nu}{2}\right)\cdot\sqrt{\frac{\beta_{12}^2}{2(\lambda_1-\lambda_2)}\cdot\left[\exp\left(2(\lambda_1-\lambda_2)\eta N_1\right)-1\right]}$$

Solving the above inequality, we get

$$N_1 = \frac{\eta^{-1}}{2(\lambda_1-\lambda_2)}\log\left(\frac{2\eta^{-1}\delta^2(\lambda_1-\lambda_2)}{\Phi^{-1}\left(\frac{1+\nu}{2}\right)^2\beta_{12}^2}+1\right)$$

□

A.3.3 Proof of Proposition 5

Proof. After Phase I, we restart our counter, i.e., $h_{\eta,0}^{(1)} = \delta$. By (3.17) and $h_{\eta,N_2}^{(1)}$ approximating to the process $H^{(1)}(\eta N_2)$, we obtain

$$\begin{aligned}\left(h_{\eta,N_2}^{(1)}(t)\right)^2 = \left(H^{(1)}(\eta N_2)\right)^2 &= \left(\sum_{j=1}^{2d}\left(\left(H^{(j)}(0)\right)^2\exp\left(2\lambda_j\eta N_2\right)\right)\right)^{-1}\left(H^{(1)}(0)\right)^2\exp\left(2\lambda_1\eta N_2\right)\\ &\geq \left(\delta^2\exp(2\lambda_1\eta N_2)+(1-\delta^2)\exp(2\lambda_2\eta N_2)\right)^{-1}\delta^2\exp(2\lambda_1\eta N_2)\end{aligned}$$

which requires

$$\left(\delta^2\exp(2\lambda_1\eta N_2)+(1-\delta^2)\exp(2\lambda_2\eta N_2)\right)^{-1}\delta^2\exp(2\lambda_1\eta N_2) \geq \eta^{-1}(1-\delta^2)$$

Solving the above inequality, we get

$$N_2 = \frac{\eta^{-1}}{2(\lambda_1-\lambda_2)}\log\frac{1-\delta^2}{\delta^2}$$

□

A.3.4 Proof of Theorem 3.4

Proof. For $i = 2, ..., 2d$, we compute the infinitesimal conditional expectation and variance,

$$\begin{aligned}
\frac{d}{dt}\mathbb{E}Z^{(i)}_{\eta_n}(t)\big|_{t=t_0} &= \eta^{-1}\mathbb{E}\left[Z^{(i)}_{\eta_n}(t_0+\eta) - Z^{(i)}_{\eta_n}(t_0)\big|H^\eta(t_0)=h\right] \\
&= \eta^{-1/2}h_i\sum_{j=1}^{2d}(\lambda_i-\lambda_j)\,h_j^2 + O(\eta) = Z^{(i)}\,(\lambda_i-\lambda_1) + o(1), \\
\frac{d}{dt}\mathbb{E}\left(Z^{(i)}_{\eta_n}(t) - Z^{(i)}_{\eta_n}(t_0)\right)^2\big|_{t=t_0} &= \eta^{-1}\mathbb{E}\left[\left(Z^{(i)}_{\eta_n}(t_0+\eta) - Z^{(i)}_{\eta_n}(t_0)\right)^2\big|H^\eta(t_0)=h\right] \\
&= \eta^{-2}\mathbb{E}\left[\eta^2(\widehat{\Lambda}h - h^\top\widehat{\Lambda}hh)(\widehat{\Lambda}h - h^\top\widehat{\Lambda}hh)^\top\right]_{i,i} + O(\eta) \\
&= \mathbb{E}(e_i^\top\widehat{\Lambda}e_1e_1^\top\widehat{\Lambda}^\top e_i) + o(1) = \frac{1}{4}\left(\gamma_i\omega_1 + \gamma_1\omega_i - 2\,\mathrm{sign}(i-d-1/2)\alpha_{i1}\right) + o(1).
\end{aligned}$$

Following similar lines to the proof of Theorem 3.3, by Section 4 of Chapter 7 in [EK09], we have for each $k = 2, ..., 2d$, if $Z^{(i)}(0) = \eta^{-1/2}h^{(i)}_{\eta,0}$ as $\eta \to 0^+$, then the stochastic process $\eta^{-1/2}h^{(k)}_{\eta,\lfloor t\eta^{-1}\rfloor}$ weakly converges to the solution of the stochastic differential equation (3.22). □

A.3.5 Proof of Proposition 6

Proof. Since we restart our counter, we have $\sum_{i=2}^{2d}(z^{(i)}_{\eta,0})^2 = \eta^{-1}\delta^2$. Since $z^{(i)}_{\eta,k}$ approximates to $Z^{(i)}(\eta k)$ and its second moment:

$$\mathbb{E}\left(Z^{(i)}(t)\right)^2 = \frac{\beta_{i1}^2}{2(\lambda_1-\lambda_i)} + \left(\left(Z^{(i)}(0)\right)^2 - \frac{\beta_{i1}^2}{2(\lambda_1-\lambda_i)}\right)\exp\left[-2(\lambda_1-\lambda_i)t\right], \quad \text{for } i \neq 1$$

we use the Markov inequality:

$$\begin{aligned}
\mathbb{P}\left(\sum_{i=2}^{2d}\left(h^{(i)}_{\eta,N_3}\right)^2 > \epsilon\right) &\le \frac{\mathbb{E}\left(\sum_{i=2}^{2d}\left(h^{(i)}_{\eta,N_3}\right)^2\right)}{\epsilon} = \frac{\mathbb{E}\left(\sum_{i=2}^{2d}\left(z^{(i)}_{\eta,N_3}\right)^2\right)}{\eta^{-1}\epsilon} \\
&= \frac{1}{\eta^{-1}\epsilon}\sum_{i=2}^{2d}\frac{\beta_{i1}^2}{2(\lambda_1-\lambda_i)}\Big(1-\exp\big(-2(\lambda_1-\lambda_i)\eta N_3\big)\Big) + \left(z^{(i)}_{\eta,0}\right)^2\exp\left[-2(\lambda_1-\lambda_i)\eta N_3\right] \\
&\le \frac{1}{\eta^{-1}\epsilon}\Big(\frac{d\max_{2\le i\le d}(\beta_{i1}^2)}{2(\lambda_1-\lambda_2)}\Big(1-\exp\big(-2(\lambda_1-\lambda_d)\eta N_3\big)\Big) \\
&\qquad + \frac{d\max_{d+1\le i\le 2d}(\beta_{i1}^2)}{2(\lambda_1+\lambda_d)}\Big(1-\exp\big(-4\lambda_1\eta N_3\big)\Big) + \delta^2\exp\left[-2(\lambda_1-\lambda_2)\eta N_3\right]\Big) \\
&\le \frac{1}{\eta^{-1}\epsilon}\left(\frac{d\max_{1\le i\le d}(\beta_{i1}^2)}{(\lambda_1-\lambda_2)} + \delta^2\exp\left[-2(\lambda_1-\lambda_2)\eta N_3\right]\right)
\end{aligned}$$

To guarantee $\frac{1}{\eta^{-1}\epsilon}\left(\frac{d\max_{1\le i\le d}(\beta_{i1}^2)}{(\lambda_1-\lambda_2)} + \delta^2\exp\left[-2(\lambda_1-\lambda_2)\eta N_3\right]\right) \le \frac{1}{4}$, we get:

$$N_3 \ge \frac{\eta^{-1}}{2(\lambda_1-\lambda_2)}\log\left(\frac{4(\lambda_1-\lambda_2)\delta^2}{(\lambda_1-\lambda_2)\epsilon\eta^{-1} - 4d\max_{1\le i\le d}\beta_{i1}^2}\right)$$

□

A.3.6 Proof of Corollary 1

Proof. First, we prove that $\|u_{\eta,k}-\widehat{u}\|_2^2+\|v_{\eta,k}-\widehat{v}\|_2^2$ can be bounded by $3\sum_{i=2}^{2d}\left(h_{\eta,k}^{(i)}\right)^2$, when it is near the optima. Recall that $h_{\eta,k}=\frac{1}{\sqrt{2}}P^\top(u_{\eta,k}^\top\ v_{\eta,k}^\top)^\top$ and $e_1=\widehat{h}=\frac{1}{\sqrt{2}}\mathbb{P}(\widehat{u}^\top\ \widehat{v}^\top)^\top$. Our analysis has shown that when k is large enough, the SGD iterates near the optima. Then we have

$$\begin{aligned}\|u_{\eta,k}-\widehat{u}\|_2^2+\|v_{\eta,k}-\widehat{v}\|_2^2 &= 4-2\langle u_{\eta,k},\widehat{u}\rangle-2\langle v_{\eta,k},\widehat{v}\rangle = 4-4h_{\eta,k}^1\\ &= 4-4\sqrt{1-\sum_{i=2}^{2d}\left(h_{\eta,k}^{(i)}\right)^2}=\frac{16\sum_{i=2}^{2d}\left(h_{\eta,k}^{(i)}\right)^2}{4+4\sqrt{1-\sum_{i=2}^{2d}\left(h_{\eta,k}^{(i)}\right)^2}}\le 3\sum_{i=2}^{2d}\left(h_{\eta,k}^{(i)}\right)^2 \end{aligned}\tag{A.8}$$

where the last inequality holds since k is large enough such that $\sum_{i=2}^{2d}\left(h_{\eta,k}^{(i)}\right)^2$ is sufficiently small. By Propositions 4, 5, and 6, the total iteration number is

$$N=N_1+N_2+N_3 \tag{A.9}$$

To explicitily bound N in (A.9) in terms of sample size n, we consider

$$N_1=\frac{\eta^{-1}}{2(\lambda_1-\lambda_2)}\log\left(\frac{2\eta^{-1}\delta^2(\lambda_1-\lambda_2)}{\Phi^{-1}\left(\frac{1+\nu}{2}\right)^2\beta_{12}^2}+1\right), \tag{A.10}$$

$$N_2=\frac{\eta^{-1}}{2(\lambda_1-\lambda_2)}\log\frac{1-\delta^2}{\delta^2}, \tag{A.11}$$

$$N_3=\frac{\eta^{-1}}{2(\lambda_1-\lambda_2)}\log\left(\frac{4(\lambda_1-\lambda_2)\delta^2}{(\lambda_1-\lambda_2)\epsilon\eta^{-1}-4d\max\limits_{1\le i\le d}\beta_{i1}^2}\right) \tag{A.12}$$

Given a small enough ϵ, we choose η as follow:

$$\eta\asymp\frac{\epsilon(\lambda_1-\lambda_2)}{d\max_{1\le i\le d}\beta_{i1}^2} \tag{A.13}$$

Combining the above sample complexities (A.10), (A.11), (A.12), and (A.13), we get

$$N=O\left[\frac{d}{\epsilon(\lambda_1-\lambda_2)^2}\log\left(\frac{d}{\epsilon}\right)\right] \tag{A.14}$$

By Proposition 6 with (A.8), after at most N iterations, we have

$$\|u_{\eta,n}-\widehat{u}\|_2^2+\|v_{\eta,n}-\widehat{v}\|_2^2\le 3\|h_{\eta,n}-\widehat{h}\|_2^2\le 3\epsilon$$

with probability at least $\frac{3}{4}$. □

B Appendix for Section 4

B.1 Proofs for Determining Stationary Points

B.1.1 Proof of Theorem 4.1

Proof. Remind that the eigendecomposition of $\widetilde{A}$ is $(\Lambda^B)^{-\frac{1}{2}}O^{B\top}AO^B(\Lambda^B)^{-\frac{1}{2}} = O^{\widetilde{A}}\Lambda^{\widetilde{A}}(O^{\widetilde{A}})^\top$. Given the eigendecomposition of B is $B = O^B\Lambda^B(O^B)^\top$, we can write B^{-1} as

$$B^{-1} = O^B(\Lambda^B)^{-1}(O^B)^\top$$

We denote $\widetilde{X}$ as $\widetilde{X} = O^{\widetilde{A}}_{:,\mathcal{I}}$ for some $\mathcal{I} \subseteq [d]$ with $|\mathcal{I}| = r$. For $X = (B^{-1/2}O^{\widetilde{A}}_{:,\mathcal{I}}) \cdot \Psi$, where $\Psi \in \mathcal{G}$. It is easy to see that $\nabla_Y\mathcal{L}(X,Y) = 0$. Ignore the constant 2 in the gradient $\nabla_X\mathcal{L}(X,Y)$ for convenience, we have,

$$\begin{aligned}
\nabla_X\mathcal{L}(X,Y) &= -(I_d - BXX^\top)AX = -(I_d - BB^{-1/2}O^{\widetilde{A}}_{:,\mathcal{I}}(O^{\widetilde{A}}_{:,\mathcal{I}})^\top B^{-1/2})AB^{-1/2}O^{\widetilde{A}}_{:,\mathcal{I}} \\
&= -AB^{-1/2}O^{\widetilde{A}}_{:,\mathcal{I}} + B^{1/2}O^{\widetilde{A}}_{:,\mathcal{I}}(O^{\widetilde{A}}_{:,\mathcal{I}})^\top O^{\widetilde{A}}\Lambda^{\widetilde{A}}(O^{\widetilde{A}})^\top O^{\widetilde{A}}_{:,\mathcal{I}} \\
&= -B^{1/2}O^{\widetilde{A}}\Lambda^{\widetilde{A}}(O^{\widetilde{A}})^\top O^{\widetilde{A}}_{:,\mathcal{I}} + B^{1/2}O^{\widetilde{A}}_{:,\mathcal{I}}\Lambda^{\widetilde{A}}_{\mathcal{I},\mathcal{I}} \\
&= -B^{1/2}O^{\widetilde{A}}\Lambda^{\widetilde{A}}_{:\mathcal{I}} + B^{1/2}O^{\widetilde{A}}_{:,\mathcal{I}}\Lambda^{\widetilde{A}}_{\mathcal{I},\mathcal{I}} = 0
\end{aligned}$$

Next we show that if X is not as specified, then $\nabla_X\mathcal{L}(X,Y) \neq 0$. We only need to show that if $\widetilde{X} = [O^{\widetilde{A}}_{:,\mathcal{S}},\ \phi]\Psi$, where $\mathcal{S} \subseteq [d]$ with $|\mathcal{S}| = r-1$ and $\phi = c_1O^{\widetilde{A}}_{:,i} + c_2O^{\widetilde{A}}_{:,j}$ with $i,j \notin \mathcal{S}$, $i \neq j$, $c_1^2 + c_2^2 = 1$, and $c_1, c_2 \neq 0$, then we have $\nabla_X\mathcal{L}(X,Y) \neq 0$. The general scenario can be induced from this basic setting. It is easy to see that such an $X = B^{-1/2}\widetilde{X}$ satisfies the constraint,

$$X^\top BX = \Psi^\top[O^{\widetilde{A}}_{:,\mathcal{S}},\ \phi]^\top B^{-1/2}BB^{-1/2}[O^{\widetilde{A}}_{:,\mathcal{S}},\ \phi]\Psi = \Psi^\top\begin{bmatrix} I_{r-1} & 0_{(r-1)\times 1} \\ 0_{1\times(r-1)} & \phi^\top\phi \end{bmatrix}\Psi = I_r$$

where the last equality follow from $\phi^\top\phi = c_1^2 + c_2^2 = 1$.

Plugging such an X into the gradient, we have

$$\begin{aligned}
\nabla_X\mathcal{L}(X,Y) &= -(I_d - BXX^\top)AX = -(I_d - BB^{-1/2}[O^{\widetilde{A}}_{:,\mathcal{S}},\ \phi][O^{\widetilde{A}}_{:,\mathcal{S}},\ \phi]^\top B^{-1/2})AB^{-1/2}[O^{\widetilde{A}}_{:,\mathcal{S}},\ \phi]\Psi \\
&= -B^{1/2}(O^{\widetilde{A}}_{:,\mathcal{S}^\perp}(O^{\widetilde{A}}_{:,\mathcal{S}^\perp})^\top - \phi\phi^\top)O^{\widetilde{A}}\Lambda^{\widetilde{A}}[(I_d)_{\mathcal{S}},\ c_1e_i + c_2e_j]\Psi \\
&= -B^{1/2}[0_{d\times(r-1)},\ O^{\widetilde{A}}_{:,\mathcal{S}^\perp}\Lambda^{\widetilde{A}}_{\mathcal{S}^\perp,:}(c_1e_i + c_2e_j)]\Psi + [0_{d\times(r-1)},\ \phi(c_1^2\lambda^{\widetilde{A}}_i + c_2^2\lambda^{\widetilde{A}}_j)]\Psi \\
&= -B^{1/2}[0_{d\times(r-1)},\ c_1c_2^2(\lambda^{\widetilde{A}}_i + \lambda^{\widetilde{A}}_j)O^{\widetilde{A}}_{:,i} + c_2c_1^2(\lambda^{\widetilde{A}}_j - \lambda^{\widetilde{A}}_i)O^{\widetilde{A}}_{j,j}]\Psi \neq 0
\end{aligned}$$

where the last $\neq$ is from $c_1, c_2 \neq 0$, $c_1^2 + c_2^2 = 1$, $\lambda^{\widetilde{A}}_j \neq \lambda^{\widetilde{A}}_j$ for $i \neq j$. □

B.1.2 Proof of Theorem 4.2

Proof. We have the Hessian of $\mathcal{L}(X,Y)$ on X with $Y = \mathcal{D}(X)$ as

$$H_X = 2\,\mathrm{sym}\left(I_r \otimes ((BXX^\top - I_d)A) + (X^\top AX)\otimes B + (AX)\boxtimes(BX)\right) \tag{B.1}$$

where $\mathrm{sym}(M) = M + M^\top$, $\otimes$ is the Kronecker product, and for $U \in \mathbb{R}^{d\times r}$ and $V \in \mathbb{R}^{m\times k}$, $U \boxtimes V \in \mathbb{R}^{dk\times mr}$ is defined as

$$U \boxtimes V = \begin{bmatrix} U_{:,1}V_{:,1}^\top & U_{:,2}V_{:,1}^\top & \cdots & U_{:,r}V_{:,1}^\top \\ U_{:,1}V_{:,2}^\top & U_{:,2}V_{:,2}^\top & \cdots & U_{:,r}V_{:,2}^\top \\ \vdots & \vdots & \ddots & \vdots \\ U_{:,1}V_{:,k}^\top & U_{:,2}V_{:,k}^\top & \cdots & U_{:,r}V_{:,k}^\top \end{bmatrix}$$

To determine whether a stationary point is an unstable stationary or a minimax global optimum, we consider its Hessian. We start with checking that $\mathcal{S} = [r]$ corresponds to the global optimum, $X = B^{-1/2}O^{\widetilde{A}}_{:,[r]}\Psi$. Without loss of generality, we set $\Psi = I_r$. We only need to check that for any vector $v = [v_1^\top, \ldots, v_r^\top]^\top \in \mathbb{R}^{nr}$ with $v_i \in \mathbb{R}^n$ denoting the i-th block of v, which satisfies

$$v_i = c_{j_i}B^{-1/2}O^{\widetilde{A}}_{:,j_i} \text{ for any } j_i \in [d] \text{ and a real constant } c_j$$

such that $\|v\|_2 = 1$, then we have $v^\top H_X v \geq 0$. The general case is only a linear combination of such v's. Specifically, for $X = O^{\widetilde{A}}_{:,[r]}$, we have

$$\begin{aligned}
v^\top H_X v &= -v^\top \mathrm{sym}\left(I_r \otimes ((I_d - BXX^\top)A) - (X^\top AX)\otimes B - (AX)\boxtimes(BX)\right)v \\
&= -v^\top \mathrm{sym}\Big(I_r \otimes ((I_d - B^{1/2}O^{\widetilde{A}}_{:,[r]}O^{\widetilde{A}\top}_{:,[r]}B^{-1/2})A) - (O^{\widetilde{A}\top}_{:,[r]}B^{-1/2}AB^{-1/2}O^{\widetilde{A}}_{:,[r]})\otimes B \\
&\qquad\qquad - (AB^{-1/2}O^{\widetilde{A}}_{:,[r]})\boxtimes(B^{1/2}O^{\widetilde{A}}_{:,[r]})\Big)v \\
&= -v^\top \mathrm{sym}\left(I_r \otimes (B^{1/2}O^{\widetilde{A}}_{:,[d]\setminus[r]}O^{\widetilde{A}\top}_{:,[d]\setminus[r]}B^{-1/2}A) - \Lambda^{\widetilde{A}}_{:,[r]}\otimes B - (B^{1/2}O^{\widetilde{A}}\Lambda^{\widetilde{A}}_{:,[r]})\boxtimes(B^{1/2}O^{\widetilde{A}}_{:,[r]})\right)v \\
&= -2\sum_{i=1}^r c_{j_i}^2 O^{\widetilde{A}\top}_{:,j_i}O^{\widetilde{A}}_{:,[d]\setminus[r]}\Lambda^{\widetilde{A}}_{[d]\setminus[r],:}O^{\widetilde{A}\top}O^{\widetilde{A}}_{:,j_i} + 2\sum_{i=1}^r c_{j_i}^2\lambda_i^{\widetilde{A}} + 2\sum_{i=1}^r\sum_{k=1}^r c_{j_i}c_{j_k}e_{j_i}^\top\Lambda^{\widetilde{A}}_{:,k}O^{\widetilde{A}\top}_{:,i}O^{\widetilde{A}}_{j_k} \\
&\geq 0 + 2\sum_{i=1}^r c_{j_i}^2\lambda_i^{\widetilde{A}} + 2\sum_{i=1}^r\sum_{k=1}^r c_{j_i}c_{j_k}\lambda_{j_i}^{\widetilde{A}} = 0
\end{aligned}$$

where the last inequality is obtained by taking $j_k \in [r]$, $i = j_k$, and $k = j_i$ in the last term, and the last equality is obtained by setting $c_{j_k} = -c_{j_i}$ when $j_i = k$, which implies that the restricted strongly convex property at X holds.

For any other $\mathcal{I} \neq [r]$, we only need to show that the largest eigenvalue of $\nabla^2\mathcal{L}$ is positive and the smallest eigenvalue of $\nabla^2\mathcal{L}$ is negative, which implies that such a stationary point is unstable. Using the same construction as above, we have

$$\begin{aligned}
\lambda_{\min}(H_X) &\leq -v^\top \mathrm{sym}\left(I_r \otimes ((I_d - BXX^\top)A) - (X^\top AX)\otimes B - (AX)\boxtimes(BX)\right)v \\
&= -v^\top \mathrm{sym}\left(I_r \otimes (B^{1/2}O^{\widetilde{A}}_{:,\mathcal{I}^\perp}O^{\widetilde{A}\top}_{:,\mathcal{I}^\perp}B^{-1/2}A) - \Lambda^{\widetilde{A}}_{:,\mathcal{I}}\otimes B - (B^{1/2}O^{\widetilde{A}}\Lambda^{\widetilde{A}}_{:,\mathcal{I}})\boxtimes(B^{1/2}O^{\widetilde{A}}_{:,\mathcal{I}})\right)v \\
&= -2\sum_{i\in\mathcal{I}} c_{j_i}^2 O^{\widetilde{A}\top}_{:,j_i}O^{\widetilde{A}}_{:,\mathcal{I}^\perp}\Lambda^{\widetilde{A}}_{\mathcal{I}^\perp,:}O^{\widetilde{A}\top}O^{\widetilde{A}}_{:,j_i} + 2\sum_{i\in\mathcal{I}} c_{j_i}^2\lambda_i^{\widetilde{A}} + 2\sum_{i\in\mathcal{I}}\sum_{k\in\mathcal{I}} c_{j_i}c_{j_k}e_{j_i}^\top\Lambda^{\widetilde{A}}_{:,k}O^{\widetilde{A}\top}_{:,i}O^{\widetilde{A}}_{j_k} \\
&\overset{(i)}{=} 2c_{j_r}^2(\lambda^{\widetilde{A}}_{\max\mathcal{I}} - \lambda^{\widetilde{A}}_{\min\mathcal{I}^\perp})
\end{aligned}$$

where (i) is from setting $c_{j_i} = 0$ for all $j_i \in \mathcal{I}^\perp$ except j_r, and $c_{j_r} = 1/\|B^{-1/2}O^{\widetilde{A}}_{:,\min\mathcal{I}^\perp}\|_2$.

On the other hand, we have

$$\begin{aligned}
&\lambda_{\max}(H_X) \geq v^\top H_X v \\
&= -2\sum_{i\in\mathcal{I}} c_{j_i}^2 O_{:,j_i}^{\widetilde{A}\top} O_{:,\mathcal{I}^\perp}^{\widetilde{A}} \Lambda_{\mathcal{I}^\perp,:}^{\widetilde{A}} O^{\widetilde{A}\top} O_{:,j_i}^{\widetilde{A}} + 2\sum_{i\in\mathcal{I}} c_{j_i}^2 \lambda_i^{\widetilde{A}} + 2\sum_{i\in\mathcal{I}}\sum_{k\in\mathcal{I}} c_{j_i} c_{j_k} e_{j_i}^\top \Lambda_{:,k}^{\widetilde{A}} O_{:,i}^{\widetilde{A}\top} O_{j_k}^{\widetilde{A}} \\
&\overset{(i)}{=} 2c_{j_1}^2 \lambda_{\min\mathcal{I}}^{\widetilde{A}} + c_{j_1}^2 \lambda_{\min\mathcal{I}}^{\widetilde{A}} = 4c_{j_1}^2 \lambda_{\min\mathcal{I}}^{\widetilde{A}}
\end{aligned}$$

where (i) is from setting $c_{j_i} = 0$ for all $j_i \in \mathcal{I}$ except j_1, and $c_{j_1} = 1/\|B^{-1/2} O_{:,\min\mathcal{I}}^{\widetilde{A}}\|_2$.

□

B.2 Singular case for B

When B **is Singular**, we assume $\text{rank}(B) = m < d$ and $\text{rank}(A) = d$. Note that we require $m \geq r$; Otherwise, the feasible region of (4.3) becomes $\mathcal{T}_B = \emptyset$.

Before we proceed with our analysis, we first exclude an ill-defined case, where the objective function of (4.3) is unbounded from above. The following proposition shows the sufficient and necessary condition of the existence of the global optima of (4.3).

Proposition 7. *Given a full rank symmetric matrix $A \in \mathbb{R}^{d\times d}$ and a positive semidefinite matrix $B \in \mathbb{R}^{d\times d}$, the optimal solution of* (4.3) *exists if and only if for all $v \in$ Null(B), one of the following two condition holds: (1) $v^\top Av < 0$; (2) $v^\top Av = 0$ and $u^\top Av = 0$, $\forall u \in$ Col(B).*

Proof. We decompose $X = X_B + X_{B^\perp}$, where $X_B = [u_1, ..., u_r]$ with $u_i \in \text{Col}(B)$ and each column of $X_{B^\perp} = [v_1, ..., v_r]$ with $v_i \in \text{Null}(B)$. Note such decomposition is unique. Then (4.3) becomes

$$\min -\sum_{i=1}^r (u_i^\top A u_i) - 2\sum_{i=1}^r (u_i^\top A v_i) - \sum_{i=1}^r (v_i^\top A v_i) \quad \text{s.t.} \quad X_B^\top B X_B = I_r \tag{B.2}$$

If (B.2) has an optimal solution, we have $v^\top Av \leq 0$, for all $v \in \text{Null}(B)$; otherwise, fixing the feasible X_B, we use $X_B = [\lambda v, ..., \lambda v]$ and increase λ, then there is no lower bound of objective value. Further, given a vector $v \in \text{Null}(B)$ with $v^\top Av = 0$, $u^\top Av = 0$ must hold for all $u \in \text{Col}(B)$; otherwise, W.L.O.G, we assume that $u_1 \in \text{Col}(B)$ $u_1^\top Av > 0$, we can construct a feasible $X_B = \mu[u_1, ..., u_r]$, where μ is a normalization constant such that $\mu^2 u_1^\top B u_1 = 1$. Then constructing $X_{B^\perp} = \lambda[v, 0, ...0]$., if we increase λ, there is no lower bound the objective value. Therefore, for a vector $v \in \text{Null}(B)$, either $v^\top Av = 0$, or $u^\top Av = 0$ and $v^\top Av = 0$ hold. □

Throughout our following analysis, we exclude the ill-defined case.

The idea of characterizing all the equilibria is analogous to the nonsingular case, but much more involved. Since B is singular, we need to use general inverses. For notationally convenience, we use block matrices in our analysis. We consider the eigenvalue decomposition of B as follows:

$$B = \underbrace{\begin{bmatrix} O_{11}^B & O_{12}^B \\ O_{21}^B & O_{22}^B \end{bmatrix}}_{O^B} \underbrace{\begin{bmatrix} \Lambda_{11}^B & 0 \\ 0 & 0 \end{bmatrix}}_{\Lambda^B} \underbrace{\begin{bmatrix} O_{11}^{B\top} & O_{21}^{B\top} \\ O_{12}^{B\top} & O_{22}^{B\top} \end{bmatrix}}_{O^{B\top}}$$

where $O_{11}^B \in \mathbb{R}^{m\times m}$, $O_{22}^B \in \mathbb{R}^{(d-m)\times(d-m)}$, and $\Lambda_{11}^B = \text{diag}(\lambda_1, ..., \lambda_m)$ with $\lambda_1 \geq \cdots \geq \lambda_m > 0$. We then left multiply $O^{B\top}$ and right multiply O^B to A:

$$O^{B\top} A O^B =: W = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix}$$

where $W_{11} \in \mathbb{R}^{m\times m}, W_{22} \in \mathbb{R}^{(d-m)\times(d-m)}$. Here, we assume W_{22} is nonsingular (guaranteed in the well-defined case). Then we construct a general inverse of Λ^B. Specifically, given an arbitrary positive definite matrix $P \in \mathbb{R}^{(d-m)\times(d-m)}$, we define $\Lambda^{B\dagger}(P)$ as

$$\Lambda^{B\dagger}(P) := \begin{bmatrix} (\Lambda^B_{11})^{-1} & 0 \\ 0 & P \end{bmatrix}$$

Note $\Lambda^{B\dagger}(P)$ is invertible and depends on P. Recall the primal variable X at the equilibrium of $\mathcal{L}(X, Y)$ satisfies

$$AX = BX \cdot X^\top AX \quad \text{and} \quad X^\top BX = I_r \tag{B.3}$$

For notational simplicity, we define

$$V(P) := \left(\Lambda^{B\dagger}(P)\right)^{-\frac{1}{2}} O^{B\top} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} V_1 \\ V_2(P) \end{bmatrix} \tag{B.4}$$

where $V_1,\ X_1 \in \mathbb{R}^{m\times r}$, and $V_2(P),\ X_2 \in \mathbb{R}^{(d-m)\times r}$. Note that V_1 does not depend on P. From (B.4) we have

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = O^B \left(\Lambda^{B\dagger}(P)\right)^{\frac{1}{2}} \begin{bmatrix} V_1 \\ V_2(P) \end{bmatrix} \tag{B.5}$$

Combining (B.5) and (B.3) we get the following equation system:

$$\begin{cases} \widetilde{A}(P)V(P) = \begin{bmatrix} V_1 \\ 0 \end{bmatrix} V(P)^\top \widetilde{A}(P)V(P) & \text{(B.6a)} \\ V(P)^\top \mathrm{diag}(I_m, 0)V(P) = I_r & \text{(B.6b)} \end{cases}$$

where $\widetilde{A}(P) = (\Lambda^{B\dagger}(P))^{\frac{1}{2}} W (\Lambda^{B\dagger}(P))^{\frac{1}{2}}$. The invertibility of $\Lambda^{B\dagger}(P)$ ensures that solving (B.3) is equivalent to doing the transformation (B.5) to the solution of (B.6). We then denote

$$\widehat{A} = (\Lambda^B_{11})^{-\frac{1}{2}} \left(W_{11} - W_{12}W_{22}^{-1}W_{21}\right) (\Lambda^B_{11})^{-\frac{1}{2}}$$

and consider its eigenvalue decomposition as $\widehat{A} = O^{\widehat{A}}\Lambda^{\widehat{A}}O^{\widehat{A}\top}$. The following theorem characterizes all the equilibria of $\mathcal{L}(X, Y)$ with a singular B.

Theorem B.1. *Given a full rank symmetric matrix $A \in \mathbb{R}^{d\times d}$ and a positive semidefinite matrix $B \in \mathbb{R}^{d\times d}$ with $\mathrm{rank}(B) = m < d$, satisfying the well-defined condition in Proposition 7,$(X, \mathcal{D}(X))$ is an equilibrium of $\mathcal{L}(X, Y)$ if and only if X can be represented as*

$$X = O^B \begin{bmatrix} (\Lambda^B_{11})^{-\frac{1}{2}} \cdot O^{\widehat{A}}_{:,\mathcal{I}} \\ -W_{22}^{-1}W_{12}^\top (\Lambda^B_{11})^{-\frac{1}{2}} O^{\widehat{A}}_{:,\mathcal{I}} \end{bmatrix} \cdot \Psi$$

where $\Psi \in \mathcal{G}$ and $\mathcal{I} \in \mathcal{X}_m$ is the column index set.

Proof. By definition, we have

$$\begin{cases} AX = BX \cdot Y \\ X^\top BX = I_r \end{cases} \Longrightarrow \begin{cases} AX = BX \cdot X^\top AX \\ X^\top BX = I_r \end{cases} \tag{B.7}$$

We define $V(P) := \left(\Lambda^{B\dagger}(P)\right)^{-\frac{1}{2}} O^{B\top}\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} V_1 \\ V_2(P) \end{bmatrix}$, where $V_1,\ X_1 \in \mathbb{R}^{m\times r}$, and $V_2(P),\ X_2 \in \mathbb{R}^{(d-m)\times r}$. Note that V_1 does not depend on P. By (B.7) and replacing I_d with $O^B O^{B\top}$ and $\Lambda^{B\dagger}(P)^{\frac{1}{2}}\Lambda^{B\dagger}(P)^{-\frac{1}{2}}$, we have

$$\begin{cases} \widetilde{A}(P)V(P) = \begin{bmatrix} V_1 \\ 0 \end{bmatrix} V(P)^\top \widetilde{A}(P)V(P) & \text{(B.8a)} \\ V(P)^\top \mathrm{diag}(I_m, 0)V(P) = I_r & \text{(B.8b)} \end{cases}$$

where $\widetilde{A}(P) = (\Lambda^{B\dagger}(P))^{\frac{1}{2}} W (\Lambda^{B\dagger}(P))^{\frac{1}{2}}$. Simplifying (B.8a), we obtain

$$\begin{cases} W_{22}^{-1}W_{21}(\Lambda_{11}^B)^{-\frac{1}{2}}V_1 = P^{\frac{1}{2}}V_2(P) \\ V_1V_1^\top(\Lambda_{11}^B)^{-\frac{1}{2}}(W_{11} - W_{12}W_{22}^{-1}W_{21})(\Lambda_{11}^B)^{-\frac{1}{2}} \end{cases}$$

Let $\widehat{A} = (\Lambda_{11}^B)^{-\frac{1}{2}}\left(W_{11} - W_{12}W_{22}^{-1}W_{21}\right)(\Lambda_{11}^B)^{-\frac{1}{2}}$. Then, by (B.8), we obtain the following equations:

$$\begin{cases} \widehat{A}V_1 = V_1V_1^\top \widehat{A}V_1 & \text{(B.9a)} \\ V_1^\top V_1 = I_r & \text{(B.9b)} \end{cases}$$

Note (B.9) are the KKT conditions of the following problem:

$$V_1^* = \underset{V_1\in\mathbb{R}^{m\times r}}{\operatorname{argmin}} -\mathrm{tr}(V_1^\top \widehat{A}V_1) \quad \text{s.t.} \quad V_1^\top V_1 = I_r \tag{B.10}$$

Because (B.10) is not a degenerate case, Theorem 4.1 can be directly applied to (B.10). Then, we get the stable equilibria and unstable equilibria of (B.10). Specifically, denote the eigenvalue decomposition of $\widehat{A}$ as $\widehat{A} = O^{\widehat{A}}\Lambda^{\widehat{A}}O^{\widehat{A}\top}$. Then we know the equilibrium of (B.9) can be represented as $V_1 = O^{\widehat{A}}_{:,\mathcal{I}} \cdot \Psi$, where $\mathcal{I} \in \left\{\{i_1, ..., i_r\} : \{i_1, ..., i_r\} \subseteq [m]\right\}$ and $\Psi \in \mathcal{G}$. Then, we know the primal variable X at an equilibrium of $\mathcal{L}(X, Y)$ satisfies

$$X = O^B \begin{bmatrix} (\Lambda_{11}^B)^{-\frac{1}{2}} \cdot O^{\widehat{A}}_{:,\mathcal{I}} \\ -W_{22}^{-1}W_{12}^\top(\Lambda_{11}^B)^{-\frac{1}{2}}O^{\widehat{A}}_{:,\mathcal{I}} \end{bmatrix} \cdot \Psi$$

where $O^{\widehat{A}}_{:,\mathcal{I}}$ is an equilibrium for the Lagrangian function of (B.10). □

Theorem B.1 implies that for the well-defined degenerated case, there are only $\binom{m}{r}$ equilibria unique in the sense of invariant group, since B is rank deficient.

B.3 Proofs for the Convergence Rate of Algorithm.

B.3.1 Proof of Lemma 1

Proof. Denote $k = \lfloor \frac{t}{\eta} \rfloor$, $\Delta(t) = w^{(\eta)}(t+\eta) - w^{(\eta)}(t)$, Δ_i as the i-th component of Δ. For notational simplicity, we may drop (t) if it is clear from the context. By the definition of $w^\eta(t)$, we have

$$\begin{aligned} \frac{1}{\eta}\mathbb{E}\left(\Delta(t)\Big|w^{(\eta)}(t)\right) &= \frac{1}{\eta}\mathbb{E}\left(W^{(k+1)} - W^{(k)}\big|W^{(k)}\right) \\ &= \frac{1}{\eta}\mathbb{E}\left[\eta\left(\Lambda^B - (\Lambda^B)^{\frac{1}{2}}\widehat{\Lambda}_B^{(k)}(\Lambda^B)^{-\frac{1}{2}}W^{(k)}W^{(k)\top}\right)\cdot\widetilde{\Lambda}^{(k)}W^{(k)}\big|W^{(k)}\right] \\ &= \Lambda^A w^{(\eta)}(t) - (w^{(\eta)}(t))^\top(\Lambda^B)^{-\frac{1}{2}}\Lambda^A(\Lambda^B)^{-\frac{1}{2}}w^{(\eta)}(t)\Lambda^B w^{(\eta)}(t) \end{aligned} \tag{B.11}$$

Similarly, we calculate the infinitesimal conditional expectation of $v_{i,1} = \frac{(w_i^{(\eta)})^{\mu_1}}{(w_1^{(\eta)})^{\mu_i}}$ as

$$
\begin{aligned}
&\frac{1}{\eta}\mathbb{E}\left(\frac{(w_i^{(\eta)})^{\mu_1}}{(w_1^{(\eta)})^{\mu_i}}(t+\eta) - \frac{(w_i^{(\eta)})^{\mu_1}}{(w_1^{(\eta)})^{\mu_i}}(t)\middle|\frac{(w_i^{(\eta)})^{\mu_1}}{(w_1^{(\eta)})^{\mu_i}}(t)\right)\\
=&\frac{1}{\eta}\mathbb{E}\left(\frac{(w_i^{(\eta)}(t)+\Delta_i)^{\mu_1}}{(w_1^{(\eta)}(t)+\Delta_1)^{\mu_i}} - \frac{(w_i^{(\eta)}(t))^{\mu_1}}{(w_1^{(\eta)}(t))^{\mu_i}}\middle|\frac{(w_i^{(\eta)}(t))^{\mu_1}}{(w_1^{(\eta)}(t))^{\mu_i}}\right)\\
=&\frac{1}{\eta}\frac{(w_i^{(\eta)}(t))^{\mu_1}}{(w_1^{(\eta)}(t))^{\mu_i}}\mathbb{E}\left(\left[1+\mu_1\frac{\Delta_i}{w_i^{(\eta)}}+\mathcal{O}(\eta^2)\right]\cdot\left[1-\mu_i\frac{\Delta_1}{w_1^{(\eta)}}+\mathcal{O}(\eta^2)\right]-1\middle|w^{(\eta)}(t)\right)\\
=&\frac{1}{\eta}\frac{(w_i^{(\eta)}(t))^{\mu_1}}{(w_1^{(\eta)}(t))^{\mu_i}}\left(\frac{\mu_1}{w_i^{(\eta)}}\mathbb{E}(\Delta_i|w^{(\eta)}(t)) - \frac{\mu_i}{w_1^{(\eta)}}\mathbb{E}(\Delta_1|w^{(\eta)}(t))\right)+\mathcal{O}(\eta)\\
=&\frac{(w_i^{(\eta)}(t))^{\mu_1}}{(w_1^{(\eta)}(t))^{\mu_i}}\left[\frac{\mu_1}{w_i^{(\eta)}}\left(-\sum_{k=1}^{d}\frac{\lambda_k}{\mu_k}(w_k^{(\eta)})^2\mu_i w_i^{(\eta)}+\lambda_i w_i^{(\eta)}\right)-\frac{\mu_i}{w_1^{(\eta)}}\left(-\sum_{k=1}^{d}\frac{\lambda_k}{\mu_k}(w_k^{(\eta)})^2\mu_1 w_1^{(\eta)}+\lambda_1 w_1^{(\eta)}\right)\right]+\mathcal{O}(\eta)\\
=&\frac{(w_i^{(\eta)}(t))^{\mu_1}}{(w_1^{(\eta)}(t))^{\mu_i}}\mu_1\mu_i\left(\beta_i-\beta_1\right)+\mathcal{O}(\eta)
\end{aligned}
$$

where the third equality holds because of the Taylor expansion, the fourth holds for Δ is order of $\mathcal{O}(\eta)$ and the last equality holds due to (B.11). Then, we calculate the infinitesimal conditional variance. From the update of W in (4.9), if $t \in [0,\ T]$ with a finite T, then $w^{(\eta)}(t)$ is bounded with probability 1. Denote $\|w^{(\eta)}(t)\|_2^2 \le D < \infty$. Then we have

$$
\begin{aligned}
&\frac{1}{\eta}\mathbb{E}\left[\left(\frac{(w_i^{(\eta)})^{\mu_1}}{(w_1^{(\eta)})^{\mu_i}}(t+\eta) - \frac{(w_i^{(\eta)})^{\mu_1}}{(w_1^{(\eta)})^{\mu_i}}(t)\right)^2\middle|\frac{(w_i^{(\eta)})^{\mu_1}}{(w_1^{(\eta)})^{\mu_i}}(t)\right]\\
=&\frac{1}{\eta}\frac{(w_i^{(\eta)}(t))^{2\mu_1}}{(w_1^{(\eta)}(t))^{2\mu_i}}\mathbb{E}\left[\left(\mu_1\frac{\Delta_i}{w_i^{(\eta)}}-\mu_i\frac{\Delta_1}{w_1^{(\eta)}}\right)^2\middle|w^{(\eta)}(t)\right]+\mathcal{O}(\eta^2)\\
\le&\frac{2}{\eta}\frac{(w_i^{(\eta)}(t))^{2\mu_1}}{(w_1^{(\eta)}(t))^{2\mu_i}}\mathbb{E}\left[\left(\frac{\mu_1}{w_i^{(\eta)}}\right)^2\Delta_i^2+\left(\frac{\mu_i}{w_1^{(\eta)}}\right)^2\Delta_1^2\middle|w^{(\eta)}(t)\right]+\mathcal{O}(\eta^2)\\
\le&4\eta\frac{(w_i^{(\eta)}(t))^{2\mu_1}}{(w_1^{(\eta)}(t))^{2\mu_i}}\mathbb{E}\Big[\frac{\left((w^{(\eta)})^\top\widetilde{\Lambda}^{(k)}w^{(\eta)}\right)^2\left(e_i^\top(\widehat{\Lambda}_B^{(k)})(\Lambda^B)^{-\frac{1}{2}}w^{(\eta)}\right)^2+\mu_i\left(e_i\widetilde{\Lambda}^{(k)}w^{(\eta)}\right)^2}{(w_i^{(\eta)})^2}\mu_i\mu_1^2\\
&+\frac{\left((w^{(\eta)})^\top\widetilde{\Lambda}^{(k)}w^{(\eta)}\right)^2\left(e_1^\top(\widehat{\Lambda}_B^{(k)})(\Lambda^B)^{-\frac{1}{2}}w^{(\eta)}\right)^2+\mu_1\left(e_1\widetilde{\Lambda}^{(k)}w^{(\eta)}\right)^2}{(w_1^{(\eta)})^2}\mu_1\mu_i^2\Big|w^{(\eta)}(t)\Big]+\mathcal{O}(\eta^2)\\
\le&4\eta^{2\delta}\left(C\frac{C_0C_1}{\mu_{\min}^2}D^3\mu_i\mu_1^2+\mu_i^2\mu_1^2\frac{C_0}{\mu_{\min}}D\right)+\mathcal{O}(\eta)\\
=&\mathcal{O}(\eta^{2\delta})\xrightarrow{\eta\to 0}0
\end{aligned}
$$

where the second inequality holds because of the mean inequality and the last inequality is from the independence of $A^{(k)}$ and $B^{(k)}$, $(w^{(\eta)})^\top\widetilde{\Lambda}w^{(\eta)} \le \|\widetilde{\Lambda}\|_2(w^{(\eta)})^\top w^{(\eta)} \le \frac{\|A^{(k)}\|_2}{\mu_{\min}}D$, since $\widetilde{\Lambda}$ is symmetric,

and $C = \frac{(w_i^{(\eta)}(t))^{2\mu_1-2}}{(w_1^{(\eta)}(t))^{2\mu_i}}$. By Section 4 of Chapter 7 in [EK09], we have that when $t \in [0,T]$, as $\eta \to 0$, $\frac{(w_i^{(\eta)})^{\mu_1}}{(w_1^{(\eta)})^{\mu_i}}$ weakly converges to the solution of (4.10) if they have the same initial solutions. Then, let $T \to \infty$, we know the convergence of $\frac{(w_i^{(\eta)})^{\mu_1}}{(w_1^{(\eta)})^{\mu_i}}$ holds at any time t. Note we can replace 1 by j, where $j \neq i$, and the proof still holds.

Moreover, using the same techniques, we can show that for all $i \in [d]$, $w_i^{(\eta)}$ converges to the solution of the following equation:

$$\frac{dw_i}{dt} = \mu_i(\beta_i - \sum_{j=1}^{d} \beta_j w_j^2)w_i \tag{B.12}$$

Note that if any $w_i > 1$, $\mu_i(\beta_i - \sum_{j=1}^d \beta_j w_j^2)w_i < 0$, and if $\sum_{j=1}^d w_j^2 < 1$, $\mu_1(\beta_1 - \sum_{j=1}^d \beta_j w_j^2)w_1 > 0$, which means that w_1 will increase. This further indicates that w_1 converges to 1, while w_i converges to 0 for all $i \neq 1$. This shows our algorithm converges to the neighbor of the global optima. □

B.3.2 Proof of Lemma 2

Proof. We prove this by contradiction. Assume the conclusion does not hold, that is there exists a constant $C > 0$, such that for any $\eta' > 0$ we have

$$\sup_{\eta \leq \eta'} \mathbb{P}(\sup_t |z_i^{(\eta)}(t)| \leq C) = 1$$

That implies there exists a sequence $\{\eta_n\}_{n=1}^{\infty}$ converging to 0 such that

$$\lim_{n\to\infty} \mathbb{P}(\sup_t |z_i^{(\eta_n)}(t)| \leq C) = 1 \tag{B.13}$$

Thus, condition (i) in Theorem 2.4 [Now13] holds. We next check the second condition. When $\sup_t |z_i^{(\eta_n)}(t)| \leq C$ holds, Assumption 5 yields that $Z_i^{(\eta_n,k+1)} - z_i^{(\eta_n,k)} = C'\eta_n$, where C' is some constant. Thus, for any $t, \epsilon > 0$, we have

$$|z_i^{(\eta_n)}(t) - Z_i^{(\eta_n)}(t+\epsilon)| = \frac{\epsilon}{\eta} C'\eta = C'\epsilon$$

Thus, condition (ii) in Theorem Theorem 2.4 [Now13] holds. Then we have $\{Z_i^{(\eta_n)}(\cdot)\}_n$ is tight and thus converges weakly. We then calculate the infinitesimal conditional expectation

$$\begin{aligned}
\frac{d}{dt}\mathbb{E}(z_j^{(\eta)}(t)) &= \frac{1}{\eta}\mathbb{E}\left(z_j^{(\eta)}(t+\eta) - z_j^{(\eta)}(t) \middle| z_j^{(\eta)}(t)\right) = \eta^{-\frac{3}{2}}\mathbb{E}\left(w_j^{(\eta)}(t+\eta) - w_j^{(\eta)}(t) \middle| w_j^{(\eta)}(t)\right) \\
&= -\eta^{-\frac{1}{2}}\left[(w^{(\eta)}(t))^\top (\Lambda^B)^{-\frac{1}{2}}(\Lambda^A)(\Lambda^B)^{-\frac{1}{2}} w^{(\eta)}(t) \cdot (\Lambda^B) w^{(\eta)}(t) - (\Lambda^A) w^{(\eta)}(t)\right]_j \\
&= \lambda_i z_j - \beta_i \mu_j z_j + \mathcal{O}(\eta^{1-2\delta})
\end{aligned}$$

The last equality holds due to the fact that our initial point is near the saddle point $w_i^{(\eta)}(t) \approx e_i$ and $|w_j^{(\eta)}(t)| \leq C\eta^{\frac{1}{2}+\delta}$. Next, we turn to the infinitesimal conditional variance,

$$\begin{aligned}
&\frac{1}{\eta}\mathbb{E}\left[\left(z_j^{(\eta)}(t+\eta) - z_j^{(\eta)}(t)\right)^2 | z_j^{(\eta)}(t)\right] \\
=&\mathbb{E}\left[\left(e_j^\top \left((\Lambda^B)^{\frac{1}{2}}\widehat{\Lambda}_B^{(k)}(\Lambda^B)^{-\frac{1}{2}} w^{(k)} w^{(l)\top} - \Lambda^B\right)\cdot \widetilde{\Lambda} w^{(k)}\right)^2 | w^{(k)}\right] \\
=&\mathbb{E}\left[\left((\widehat{\Lambda}_B^{(k)})_{j,i}\cdot\sqrt{\mu_j/\mu_i}\cdot\widetilde{\Lambda}_{i,i} - \mu_j\widetilde{\Lambda}_{j,i}\right)^2\right] + \mathcal{O}(\eta^{3-6\delta}) \\
=&G_{j,i} + \mathcal{O}(\eta^{3-6\delta}) \leq 2\left(\frac{\mu_1}{\mu_j}\cdot C_0\cdot C_1 + \mu_i^2\cdot C_1\right)
\end{aligned}$$

Then, we get the limit stochastic differential equation,

$$dz_j(t) = (-\beta_j\mu_i\cdot z_i + \lambda_i z_i)\,dt + \sqrt{G_{j,i}}dB(t) \quad \text{for } j\in[d]\backslash\{i\}$$

Therefore, $\{Z_i^{(\eta_n)}(\cdot)\}$ converges weakly to a process defined by the equation above, which is an unstable O-U process with mean 0 and exploding variance. Thus, for any τ, there exist a time t', such that

$$\mathbb{P}(|z_i(t')| \geq C) \geq 2\tau$$

Since $\{z_i^{(\eta_n)}\}_n$ converges weakly to z_i, thus $\{z_i^{(\eta_n)}(t')\}_n$ converges in distribution to $Z_(t')$. This implies that there exists an $N>0$, such that for any $n>N$

$$|\mathbb{P}(|z_i(T)|\geq C) - \mathbb{P}(|z_i^{(\eta_n)}(T)|\geq C)| \leq \tau$$

Then we find a t' such that

$$\mathbb{P}(|z_i^{(\eta_n)}(t')|\geq C) \geq \tau, \forall n > N$$

or equivalently

$$\mathbb{P}(|z_i^{(\eta_n)}(t')|\leq C) < 1-\tau, \forall n > N$$

Since $\left\{\omega\middle|\sup_t|z_i^{(\eta_n)}(t)(\omega)|\leq C\right\} \subset \left\{\omega\middle||z_i^{(\eta_n)}(\tau')(\omega) < C\right\}$, we have

$$\mathbb{P}(\sup_t|w_i^{(\eta_n)}(t)|\leq C\sqrt{\eta_n}) = \mathbb{P}(\sup_t|z_i^{(\eta_n)}(t)|\leq C) \leq 1-\delta, \forall n > N$$

which leads to a contradiction with (B.13). Our assumption does not hold.

□

B.3.3 Proof of Lemma 3

Proof. Suppose the initial is near the stable equilibria, i.e., $|w_1^{(\eta)}(0)-1| \leq C\eta^{\frac{1}{2}+\delta}$ and $|w_j^{(\eta)}(0)| \leq C\eta^{\frac{1}{2}+\delta}$ for all $j\neq 1$.First we show that $\|w^{(\eta)}(t)\|_2 \to 1$ as $t\to\infty$. With update (4.9), we show $w^{(\eta)\top}w^{(\eta)}(t)$ weakly converges to the following ODE by a similar proof in Lemma 1:

$$\begin{aligned}
\frac{d}{dt}\mathbb{E}\left(w^{(\eta)\top}w^{(\eta)}(t)\right) &= -w^{(\eta)\top}(\Lambda^B)^{-\frac{1}{2}}(\Lambda^A)(\Lambda^B)^{-\frac{1}{2}}w^{(\eta)}\cdot w^{(\eta)\top}\Lambda^B w^{(\eta)} + w^{(\eta)\top}(\Lambda^A)w^{(\eta)} + \mathcal{O}(\eta) \\
&= -\lambda_1\left(\|w^{(\eta)}\|_2^4 - \|w^{(\eta)}\|_2^2\right) + \mathcal{O}(\eta^{1-2\delta})
\end{aligned}$$

Similarly, we can bound the infinitesimal conditional variance. Therefore, the norm of w weakly converges to the following ODE:

$$dx = -\lambda_1 \left(x^2 - x\right) dt$$

The solution of the above ODE is

$$x = \begin{cases} \frac{1}{1-\exp(-\lambda_1 t + C)} & \text{if } x > 1 \\ \frac{1}{1+\exp(-\lambda_1 t + C)} & \text{if } x < 1 \\ 1 & \text{if } x = 1 \end{cases}$$

This implies that $\|w^{(\eta)}(t)\|_2$ converges to 1 as $t \to \infty$. Then we calculate the infinitesimal conditional expectation for $i \neq 1$

$$\begin{aligned} \frac{d}{dt}\mathbb{E}(z_i^{(\eta)}(t)) &= \frac{1}{\eta}\mathbb{E}\left(z_i^{(\eta)}(t+\eta) - z_i^{(\eta)}(t) \middle| z_i^{(\eta)}(t)\right) = \eta^{-\frac{3}{2}}\mathbb{E}\left(w_i^{(\eta)}(t+\eta) - w_i^{(\eta)}(t) \middle| w_i^{(\eta)}(t)\right) \\ &= -\eta^{-\frac{1}{2}}\left[(w^{(\eta)}(t))^\top (\Lambda^B)^{-\frac{1}{2}}(\Lambda^A)(\Lambda^B)^{-\frac{1}{2}} w^{(\eta)}(t) \cdot (\Lambda^B) w^{(\eta)}(t) - (\Lambda^A) w^{(\eta)}(t)\right]_i \\ &= \lambda_i z_i - \beta_1 \mu_i z_i + \mathcal{O}(\eta^{1-2\delta}) \end{aligned}$$

The last equality is from the fact that our initial point is near an optimum. Next, we turn to the infinitesimal conditional variance,

$$\begin{aligned} &\frac{1}{\eta}\mathbb{E}\left[\left(z_i^{(\eta)}(t+\eta) - z_i^{(\eta)}(t)\right)^2 \left| z_i^{(\eta)}(t)\right.\right] = \mathbb{E}\left[\left(e_i^\top \left((\Lambda^B)^{\frac{1}{2}} \widehat{\Lambda}_B^{(k)} (\Lambda^B)^{-\frac{1}{2}} w^{(k)} w^{(l)\top} - \Lambda^B\right) \cdot \widetilde{\Lambda} w^{(k)}\right)^2 \left| w^{(k)}\right.\right] \\ &= \mathbb{E}\left[\left((\widehat{\Lambda}_B^{(k)})_{i,1} \cdot \sqrt{\mu_i/\mu_1} \cdot \widetilde{\Lambda}_{1,1} - \mu_i \widetilde{\Lambda}_{i,1}\right)^2\right] + \mathcal{O}(\eta^{3-6\delta}) \\ &= G_{i,1} + \mathcal{O}(\eta^{3-6\delta}) \leq 2\left(\frac{\mu_i}{\mu_1} \cdot C_0 \cdot C_1 + \mu_i^2 \cdot C_1\right) \end{aligned}$$

By Section 4 of Chapter 7 in [EK09], we have that the algorithm converges to the solution of (4.15) if it is already near our optimal solution. □

B.3.4 Proof of Theorem 4.3

Proof. Assume the initial is near a saddle point, e_i. According to Lemma 2 and (4.13), we obtain the closed form solution of (4.13) as follows:

$$\begin{aligned} z_j(t) &= z_j(0)\exp\left(-\mu_j\left(\beta_i - \beta_j\right)t\right) + \sqrt{G_{j,i}}\int_0^t \exp\left(\mu_j\left(\beta_i - \beta_j\right)(s-t)\right) dB(s) \\ &= \underbrace{\left(z_j(0) + \sqrt{G_{j,i}}\int_0^t \exp\left(\mu_j(\beta_i - \beta_j)s\right) dB(s)\right)}_{Q_1} \underbrace{\exp\left(-\mu_j\left(\beta_i - \beta_j\right)t\right)}_{Q_2} \end{aligned}$$

We consider $j = 1$. Note at time t, Q_1 essentially is a random variable with mean $z_1(0)$ and variance $\frac{G_{1,i}\mu_1}{2(\beta_1 - \beta_i)}\left(1 - \exp\left(-2\mu_1(\beta_1 - \beta_i)t\right)\right)$, which has an upper bound $\frac{G_{1,i}\mu_1}{2(\beta_1 - \beta_i)}$. Q_2, however, amplifies the magnitude of Q_1. Then it forces the algorithm escaping from the saddle point e_i. We consider the event $\{w_1(t)^2 > \eta\}$ and a random variable $v(t) \sim N\Big(0, \frac{G_{1,i}\mu_1}{2(\beta_1 - \beta_i)}\left(\exp\left(2\mu_1(\beta_1 - \beta_i)t\right) - 1\right)\Big)$. Because $z_j(0)$ might not be 0, we have

$$\mathbb{P}(w_1(t)^2 > \eta) \geq \mathbb{P}(v^2(t) > 1)$$

Let the right hand side of (B.18) larger than 95%. Then with a sufficiently small η, we need

$$T_1 \asymp \frac{1}{\mu_1(\beta_1-\beta_i)}\log(\frac{200(\beta_1-\beta_i)}{\mu_1 G_{1,i}}+1) \tag{B.14}$$

such that $\mathbb{P}(|w_1^{(\eta)}(T_1)|_2^2 > \eta) = 90\%$.

Now we consider the time required to converge under the ODE approximation.

By Lemma 1 with $j=1$, after restarting the counter of time, we have

$$\frac{w_1^{\mu_i}(t)}{w_i^{\mu_1}(t)} \geq \eta^{\mu_i/2}\exp(\mu_1\mu_i(\beta_1-\beta_i)t)$$

Let the right hand side equal to 1. Then with a sufficiently small η we need

$$T_2 \asymp \frac{\mu_{\max}}{\mu_1\mu_{\min}\cdot \mathsf{gap}}\log(\eta^{-1}) \tag{B.15}$$

such that $\mathbb{P}\left(\frac{w_i^{(\eta)\mu_1}(T_2)}{w_1^{(\eta)\mu_i}(T_2)} \leq 1\right) = \frac{5}{6}$.

Then let $i=1$ in Lemma 1. After restarting the counter of time, we have

$$\begin{aligned}\frac{w_i^{\mu_1}(t)}{w_1^{\mu_i}(t)} &\leq C\exp(\mu_{\max})\exp(\mu_1\mu_i(\beta_i-\beta_1)t)\\ \Longrightarrow w_i^2 &\leq (C\exp(\mu_{\max})\exp(\mu_1\mu_i(\beta_i-\beta_1)t))^{2/\mu_1}\end{aligned}$$

where $\exp(\mu_{\max})$ comes from the above stage and C is a constant containing $G_{1,i}$ and $G_{i,j}$. The second inequality holds due to the fact that $w_1 \leq 1$, mentioned in the proof of Lemma 1. Therefore, given $\sum_{i=2}^d w_i^2 \leq \kappa\eta^{1+2\delta}$ and a sufficiently small η, we need

$$T_2' \asymp \frac{\mu_{\max}}{\mu_1\mu_{\min}\cdot \mathsf{gap}}\log(\eta^{-1}) \tag{B.16}$$

such that $\mathbb{P}\left(\frac{|w_1^{(\eta)}(T_2')|^2}{\|w^{(\eta)}(T_2')\|_2^2} > 1-\kappa\eta^{1+2\delta}\right) = \frac{8}{9}$.

Then the algorithm goes into Phase III. According to Lemma 3 and (4.15), we obtain the closed form solution of (4.15) as follows:

$$z_i(t) = z_i(0)\exp\left(-\mu_i\left(\beta_1-\beta_i\right)t\right) + \sqrt{G_{i,1}}\int_0^t \exp\left(\mu_i\left(\beta_1-\beta_i\right)(s-t)\right)dB(s)$$

By the Ito isometry property of the Ito-Integral, we have

$$\mathbb{E}\left(z_i(t)\right)^2 = \left(z_i(0)\right)^2 e^{-2\mu_i(\beta_1-\beta_i)t} + \frac{G_{i,1}}{2\mu_i\left(\beta_1-\beta_i\right)}\left[1-e^{-2\mu_i(\beta_1-\beta_i)t}\right] \tag{B.17}$$

Then we consider the complement of the event $\{w_1^2 > 1-\epsilon\}$. By Markov inequality, we have

$$\begin{aligned}&\mathbb{P}(w_1^2 \leq 1-\epsilon)\\ =&\mathbb{P}\left(\sum_{i=2}^d w_i^2 \geq \epsilon\right) \leq \frac{\mathbb{E}\left(\sum_{i=2}^d w_i^2\right)}{\epsilon} = \frac{\mathbb{E}\left(\sum_{i=2}^d z_i^2\right)}{\eta^{-1}\epsilon}\\ =&\frac{1}{\eta^{-1}\epsilon}\left(\sum_{i=2}^d \left(z_i(0)\right)^2 e^{-2\mu_i(\beta_1-\beta_i)t} + \frac{G_i}{2\mu_i\left(\beta_1-\beta_i\right)}\left[1-e^{-2\mu_i(\beta_1-\beta_i)t}\right]\right)\\ \leq&\frac{1}{\eta^{-1}\epsilon}\left(\eta^{-1}\delta^2 e^{-2\mu_{\min}\cdot\mathsf{gap}\cdot\mathsf{t}} + \frac{\phi}{2\mu_{\min}\cdot\mathsf{gap}}\right)\end{aligned} \tag{B.18}$$

Let the right hand side of (B.18) be no larger than $\frac{1}{16}$.

$$\frac{1}{\eta^{-1}\epsilon}\left(\eta^{-1}\delta^2 e^{-2\mu_{\min}\cdot \mathsf{gap}\cdot t}+\frac{\phi}{2\mu_{\min}\cdot \mathsf{gap}}\right)\leq \frac{1}{16}$$
$$\implies e^{2\mu_{\min}\cdot \mathsf{gap}\cdot t}\geq \frac{16\cdot \mu_{\min}\cdot \mathsf{gap}\cdot \delta^2}{\epsilon\cdot \mu_{\min}\cdot \mathrm{gap}-16\cdot \eta\cdot \phi}$$

Then after restarting the counter of time, we need

$$T_3 \asymp \frac{1}{\mu_{\min}\cdot \mathsf{gap}}\cdot \log\left(\frac{\mu_{\min}\cdot \mathsf{gap}\cdot \delta^2}{\epsilon\cdot \mu_{\min}\cdot \mathrm{gap}-16\cdot \eta\cdot \phi}\right). \tag{B.19}$$

such that $\mathbb{P}(w_1^2(T_3)\geq 1-\epsilon)\geq \frac{15}{16}$.

Combining (B.14), (B.15), (B.16), (B.19), if our algorithm start from a saddle, then with probability at least $\frac{5}{8}$, we need

$$T=T_1+T_2+T_2'+T_3 \asymp \frac{\mu_{\max}/\mu_{\min}}{\mu_1\cdot \mathsf{gap}}\log\left(\eta^{-1}\right) \tag{B.20}$$

such that $w_1^2(T)>1-\epsilon$.

Moreover, we choose

$$\eta \asymp \frac{\epsilon\cdot \mu_{\min}\cdot \mathsf{gap}}{\phi} \tag{B.21}$$

Combining (B.20) and (B.21) together, we get the asymptotic sample complexity

$$N\asymp \frac{T}{\eta}\asymp \frac{\phi\cdot \mu_{\max}/\mu_{\min}}{\epsilon\cdot \mu_1\cdot \mu_{\min}\cdot \mathsf{gap}^2}\log\left(\frac{\phi}{\epsilon\cdot \mu_{\min}\cdot \mathsf{gap}}\right) \tag{B.22}$$

such that with probability at least $\frac{5}{8}$, we have $\|\widehat{W}-W^*\|_2^2\leq \epsilon$. □