On Constrained Nonconvex Stochastic Optimization: A Case Study for Generalized Eigenvalue Decomposition

Zhehui Chen[1] **Xingguo Li**[2] **Lin Yang**[2] **Jarvis Haupt**[3] **Tuo Zhao**[1]
Georgia Tech[1] Princeton University[2] University of Minnesota[3]

Abstract

We study constrained nonconvex optimization problems in machine learning and signal processing. It is well-known that these problems can be rewritten to a min-max problem in a Lagrangian form. However, due to the lack of convexity, their landscape is not well understood and how to find the stable equilibria of the Lagrangian function is still unknown. To bridge the gap, we study the landscape of the Lagrangian function. Further, we define a special class of Lagrangian functions. They enjoy the following two properties: 1.Equilibria are either stable or unstable (Formal definition in Section 2); 2.Stable equilibria correspond to the global optima of the original problem. We show that a generalized eigenvalue (GEV) problem, including canonical correlation analysis and other problems as special examples, belongs to the class. Specifically, we characterize its stable and unstable equilibria by leveraging an invariant group and symmetric property (more details in Section 3). Motivated by these neat geometric structures, we propose a simple, efficient, and stochastic primal-dual algorithm solving the online GEV problem. Theoretically, under sufficient conditions, we establish an asymptotic rate of convergence and obtain the first sample complexity result for the online GEV problem by diffusion approximations, which are widely used in applied probability. Numerical results are also provided to support our theory.

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89. Copyright 2019 by the author(s).

1 Introduction

We often encounter the following optimization problem in machine learning and signal processing:

$$\min_X f(X) \quad \text{subject to} \quad X \in \Omega, \tag{1}$$

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 $\Omega = \{X \in \mathbb{R}^d : \mathcal{G}(X) = 0\}$. Principal component analysis (PCA), canonical correlation analysis (CCA), matrix factorization/sensing/completion, phase retrieval, and many other problems (Friedman et al., 2001; Sun et al., 2016; Bhojanapalli et al., 2016; Li et al., 2016b; Ge et al., 2016b; Chen et al., 2017; Zhu et al., 2017) can be viewed as special examples of (1). Many algorithms have been proposed to solve (1). For the unconstrained ($\Omega = \mathbb{R}^d$) or a simple constraint $\mathcal{G}(X)$, e.g., the spherical constraint, $\mathcal{G}(X) := ||X||_2 - 1$, we can apply simple first order algorithms such as the projected gradient descent algorithm (Luenberger et al., 1984).

However, when $\mathcal{G}(X)$ is complicated, the aforementioned algorithms are often not applicable or inefficient. This is because the projection to Ω does not admit a closed form expression and can be computationally expensive in each iteration. To address this issue, we convert (1) to a min-max problem using the Lagrangian multiplier method. Specifically, instead of solving (1), we solve the following:

$$\min_{X \in \mathbb{R}^d} \max_{Y \in \mathbb{R}^m} \mathcal{L}(X, Y) := f(X) + Y^\top \mathcal{G}(X), \tag{2}$$

where $Y \in \mathbb{R}^m$ is the Lagrangian multiplier. $\mathcal{L}(X, Y)$ is often referred as the Lagrangian function in literature (Boyd and Vandenberghe, 2004). The literature on optimization also refers to X as

the primal variable and Y as the dual variable. Accordingly, (1) is called the primal problem. From the perspective of game theory, it can be viewed as two players competing with each other and eventually achieving some equilibrium. When $f(X)$ is convex and Ω is convex or the boundary of a convex set, the optimization landscape of (2) is essentially convex-concave, i.e., for any fixed Y, $\mathcal{L}(X,Y)$ is convex in X, and for any fixed X, $\mathcal{L}(X,Y)$ is concave in Y. Such a landscape further implies that the equilibrium of (2) is a saddle point, whose primal variable is equivalent to the global optimum of (1) under strong duality conditions. To solve (2), we resort to primal-dual algorithms, which iterate over both X and Y (usually in an alternating manner). The global convergence rates to the equilibrium are also established accordingly for these algorithms (Lan et al., 2011; Chen et al., 2014; Iouditski and Nesterov, 2014).

When $f(X)$ and Ω are nonconvex, both (1) and (2) become much more computationally challenging, NP-Hard in general. Significant progress has been made toward solving primal problem (1). For example, Ge et al. (2015) show that when certain tensor factorization problem satisfies the so-called strict saddle properties, one can apply some first order algorithms, e.g., the projected gradient algorithm, and the global convergence in polynomial time can be guaranteed. Their results further motivate many follow-up works, proving that many problems can be formulated as strict saddle optimization problems, including PCA, multiview learning, phase retrieval, and matrix factorization/sensing/completion (Sun et al., 2016; Bhojanapalli et al., 2016; Li et al., 2016b; Ge et al., 2016b; Chen et al., 2017; Zhu et al., 2017; Liu et al., 2018). Note that these strict saddle optimization problems are either unconstrained or with a simple constraint. However, for many other nonconvex optimization problems, Ω can be much more complicated. To the best of our knowledge, when Ω is nonconvex and complicated, the applicable algorithms and convergence guarantees are still largely unknown.

To handle the complicated Ω, this paper proposes to investigate min-max problem (2). Specifically, we first define a special class of Lagrangian functions, where the landscape of $\mathcal{L}(X,Y)$ enjoys the following good properties: (1) *There exist only two types of equilibria – stable and unstable equilibria. At an unstable one, $\mathcal{L}(X,Y)$ has negative curvature with respect to the primal variable X. More details in Section 2*; (2) *All stable equilibria correspond to the global optima of primal problem* (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 descent direction. On the other hand, the second property ensures that we do not get spurious local optima of (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, Mika et al. (1999)), and sufficient dimension reduction (SDR, Cook and Ni (2005)) as special examples. Specifically, GEV solves

$$\begin{aligned} X^* = \operatorname*{argmin}_{X\in\mathbb{R}^{d\times r}} \ & f(X) := -\operatorname{tr}(X^\top AX) \\ \text{s.t.} \ \ & X \in \mathcal{T}_B := \{X \in \mathbb{R}^{d\times r} : X^\top BX = I_r\}, \end{aligned} \quad (3)$$

where $A, B \in \mathbb{R}^{d\times d}$ are symmetric, B is positive semidefinite. We rewrite (3) as a min-max problem,

$$\min_X \max_Y \mathcal{L}(X,Y) = -\operatorname{tr}(X^\top AX) + \langle Y, X^\top BX - I_r\rangle, \quad (4)$$

where $Y \in \mathbb{R}^{r\times r}$ is the Lagrangian multiplier. Theoretically, we show that the Lagrangian function in (4) exactly belongs to our previously defined class. Motivated by our defined landscape structures, we then solve an online version of (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, Gorrell (2006)). Theoretically, we establish its asymptotic rate of convergence to stable equilibria for our stochastic GHA (SGHA) based on the diffusion approximations (Kushner and Yin, 2003). 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 (Harold et al., 1997; Li et al., 2016a; Chen et al., 2017). 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 also presented to justify our theory.

Our work is closely related to several recent results on solving GEV problems. For example, Ge et al.

(2016a) propose a multistage semi-stochastic optimization algorithm for solving the GEV problem 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 forbidden in our setting. Similar matrix inversion approaches are also adopted by a few other recently proposed algorithms for solving the GEV problem (Allen-Zhu and Li, 2016; Arora et al., 2017). 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 and Generative Adversarial Network (Sutton et al., 2000; Shapiro et al., 2009; Goodfellow et al., 2014). Many primal-dual algorithms have been proposed to solve these problems. However, most of these algorithms are 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 (equal or more difficult than our online GEV). This paper 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.

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, $X_{:,\mathcal{I}}$ ($X_{\mathcal{I},:}$) as the column (row) submatrix of X indexed by $\mathcal{I}$, $\mathrm{vec}(X) \in \mathbb{R}^{dr}$ as the vectorization of X, Col/Null(X) as the column/null space of X. Given a symmetric matrix $X \in \mathbb{R}^{d\times d}$, we denote $\lambda_{\min/\max}(X)$ as its smallest/largest singular value, and denote the 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.

2 Characterization of Equilibria

We start with characterizing the equilibria of (2). By KKT conditions, an equilibrium (X, Y) satisfies

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

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 1. *Given the Lagrangian function $\mathcal{L}(X, Y)$ in (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 1 has a similar strict saddle property over a manifold in Ge et al. (2015). The motivation behind Definition 1 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 strong convexity, which relates to several conditions, e.g., Polyak Łojasiewicz conditions (Polyak, 1963), 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 (Luo and Tseng, 1993). With this property, we cannot decrease $\mathcal{L}(X, Y)$ along any direction with respect to X. Definition 1 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 equilibrium, 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 paper. They enjoy the following properties: (1) *All equilibria are either stable or unstable (i.e., no high order unstable equilibria)*; (2) *All stable equilibria correspond to the global optima of the primal problem.*
As mentioned earlier, the first property ensures that second order information can identify the type of

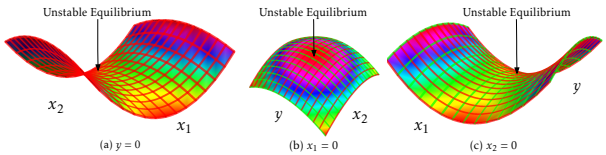


Figure 1: An illustration of an unstable equilibrium: $\min_{x_1,x_2}\max_y \mathcal{L}(x_1,x_2,y) = x_1^2 - x_2^2 - y^2$. Notice that $(0,0,0)$ is an unstable equilibrium. For visualization, we show three views: (a) $\mathcal{L}(x_1,x_2,0)$; (b) $\mathcal{L}(0,x_2,y)$; (c) $\mathcal{L}(x_1,0,y)$. Red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

equilibria. The second property guarantees that we do not get spurious optima for (1) as long as an algorithm attains a stable equilibrium. Several machine learning problems belong to this class, such as generalized eigenvalue problem.

3 Generalized Eigenvalue Problem

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):

$$\min_{X\in\mathbb{R}^{d\times r}} \max_{Y\in\mathbb{R}^{r\times r}} \mathcal{L}(X,Y),$$

where $\mathcal{L}(X,Y) = -\operatorname{tr}(X^\top AX) + \langle Y, X^\top BX - I_r\rangle$.

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

Assumption 1. *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 $\lambda_1^{\widetilde{A}},\ldots,\lambda_d^{\widetilde{A}}$, 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 1 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.

To characterize all equilibria of GEV, we leverage the idea of an invariant group. Li et al. (2016b) use similar techniques for an unconstrained matrix factorization problem. However, it does not work for the Lagrangian function due to the more complicated 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 (3). Therefore, GEV belongs to the class we defined earlier.

3.1 Invariant Group and Symmetric Property

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) has the same landscape as $\mathcal{L}(X\Psi,\Psi^\top Y\Psi)$. 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 (Dummit and Foote, 2004).

Definition 2. *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: *$\phi(\mathbf{1},x) = x \;\; \forall x \in \mathcal{X}$, where $\mathbf{1}$ denotes the identity element of $\mathcal{H}$.*
Compatibility: *$\phi(gh,x) = \phi(g,\phi(h,x)) \forall g,h \in \mathcal{H}, x \in \mathcal{X}$.*

Definition 3. *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*

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

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

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

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

$$(X_1,Y_1) = (X_2\Psi,\Psi^{-1}Y_2\Psi) = (X_2\Psi,\Psi^\top Y_2\Psi). \quad (5)$$

To find all equilibria of GEV, we examine the KKT conditions of (4):

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

Given the eigendecomposition $B = O^B\Lambda^B O^{B\top}$, we denote $\widetilde{A} = (\Lambda^B)^{-\frac{1}{2}}O^{B\top}AO^B(\Lambda^B)^{-\frac{1}{2}}$ and $\widetilde{X} =$

$(\Lambda^B)^{\frac{1}{2}} O^{B\top} X$. We then consider eigendecomposition $\widetilde{A} = O^{\widetilde{A}} \Lambda^{\widetilde{A}} O^{\widetilde{A}\top}$. 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 index set to determine a column submatrix.

Theorem 4 (Symmetric Property)**.** *Suppose Assumption 1 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 is provided in Appendix A.1. Theorem 4 implies that under the equivalence relation given in (5), there are $\binom{d}{r}$ equilibria of $\mathcal{L}(X, Y)$. Each corresponds to an $O^{\widetilde{A}}_{:,\mathcal{I}}$, where $\mathcal{I} \in \mathcal{X}_d^r$ is an index set. Then whole equilibria set is generated by $O^{\widetilde{A}}_{:,\mathcal{I}}$'s with the transformation matrix $O^B (\Lambda^B)^{-\frac{1}{2}}$ and the invariant group action induced by $\mathcal{G}$.

3.2 Unstable and Stable Equilibria

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 1, $(X, \mathcal{D}(X))$ 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 5. *Suppose Assumption 1 holds, and $(X, \mathcal{D}(X))$ is an equilibrium in (4). By Theorem 4, 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 $\mathcal{I} \in \mathcal{X}_d^r$. Then, if $\mathcal{I} \neq [r]$, equilibrium $(X, \mathcal{D}(X))$ is unstable 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 $\lambda^{\widetilde{A}}_{\max / \min \mathcal{I}} = \max / \min_{i \in \mathcal{I}} \lambda^{\widetilde{A}}_i$, and $\lambda^{\widetilde{A}}_i$ is the i-th leading eigenvalue of $\widetilde{A}$;
Otherwise, we have $H_X \succeq 0$ and $\operatorname{rank}(H_X) = dr - r(r-1)/2$. Moreover, $(X, \mathcal{D}(X))$ is a stable equilibrium of problem (4).

The proof of Theorem 5 is provided in Appendix A.2. Theorem 5 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 $\operatorname{Null}(H_X)$ 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 (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 (3) and other equilibria are unstable, which further indicates that GEV belongs to the class we defined earlier.

4 Stochastic Optimization for GEV

For GEV, we propose a fully stochastic primal-dual algorithm to solve (4), which only requires access to the stochastic approximations (SA) of A and B matrices. This is very different from other existing semi-stochastic algorithms that require to access the exact B matrix (Ge et al., 2016a). Specifically, we propose a stochastic variant of the generalized Hebbian algorithm (GHA), also referred as Sanger's rule in existing literature (Sanger, 1989), to solve (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:

Primal Update:

$$X^{(k+1)} \leftarrow X^{(k)} - \eta \underbrace{\left(B^{(k)} X^{(k)} Y^{(k)} - A^{(k)} X^{(k)} \right)}_{\text{Stoc. Approx. of } \nabla_X \mathcal{L}(X^{(k)}, Y^{(k)})}, \quad (6)$$

Dual Update:

$$Y^{(k+1)} \leftarrow \underbrace{X^{(k)\top} A^{(k)} X^{(k)}}_{\text{Stoc. Approx. of } X^{(k)\top} A X^{(k)}}, \quad (7)$$

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). SGHA is simple and easy to implement. The constraint is handled by the dual update. Further, motivated by the landscape of GEV, we analyze the algorithm by diffusion approximations and obtain the asymptotical sample complexity.

4.1 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 $A = U \cdot \mathrm{diag}(1, 1, 1, 0.1, ..., 0.1) \cdot U^\top$ and $B = U \cdot \mathrm{diag}(2, 2, 2, 1, ..., 1) \cdot U^\top$;
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 $A = U \cdot \mathrm{diag}(1, 1, 1, 0.1, ..., 0.1) \cdot U^\top$ and $B = V \cdot \mathrm{diag}(2, 2, 2, 1, ..., 1) \cdot V^\top$.
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 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 and vertical axises correspond to the number of iterations and the optimization error

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

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

4.2 Analysis for Commutative A and B

As a special case, we first prove the convergence of SGHA for GEV with $r = 1$, and commutative A and B. 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 2. *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 $\Lambda^B = \mathrm{diag}(\mu_1, ..., \mu_d)$ and $\Lambda^A = \mathrm{diag}(\lambda_1, ..., \lambda_d)$ 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 Assumption 2 are mild, but (b) is stringent. For convenience of analysis, we combine (6) and (7) as

$$X^{(k+1)} \leftarrow X^{(k)} - \eta\big(B^{(k)}X^{(k)}X^{(k)\top} - I_d\big)A^{(k)}X^{(k)}. \quad (8)$$

We remark that (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

$$\Lambda = (\Lambda^B)^{-\frac{1}{2}}\Lambda^A(\Lambda^B)^{-\frac{1}{2}} : \mathrm{diag}(\beta_1, \cdots, \beta_d).$$

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

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

Moreover, we denote $W^{(k)} = (\Lambda^B)^{\frac{1}{2}}OX^{(k)}$. One can verify that (8) can be rewritten as follows:

$$W^{(k+1)} \leftarrow W^{(k)} - \eta\Big((\Lambda^B)^{\frac{1}{2}}\widehat{\Lambda}_B^{(k)}(\Lambda^B)^{-\frac{1}{2}} \cdot W^{(k)}W^{(k)\top} - \Lambda^B\Big) \cdot \widetilde{\Lambda}^{(k)}W^{(k)}, \quad (9)$$

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 $W^* = [1, 0, 0, ..., 0]^\top$ corresponds to the optimal solution of (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.

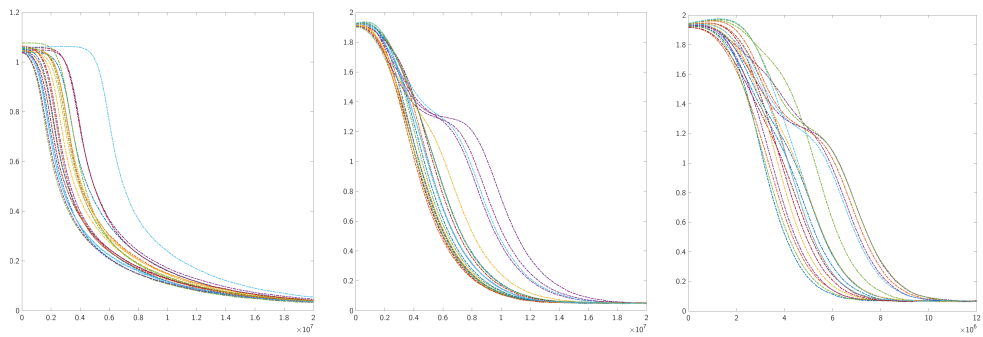
Overall, we have an asymptotic sample complexity.
ODE Characterization: To demonstrate an ODE characterization for the trajectory of our algorithm, we introduce a continuous time random process

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

where $k = \lfloor \frac{t}{\eta} \rfloor$ and η is the step size in (8). For notational simplicity, we drop (t) when it is clear from the context. Instead of directly showing a global convergence of $w^{(\eta)}$, we construct a new quantity as

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

where $w_i^{(\eta)}$ is the i-th component (coordinate) of $w^{(\eta)}$. We then show that $v_{i,j}^{(\eta)}$ converges to an exponential decay function.



(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.

Lemma 6. *Suppose Assumption 2 holds and the initial is away from saddle points, i.e., given constants, $\tau > 0$ and $\delta < \frac{1}{2}$, there exist i, j such that*

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

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

$$dx_{k,j} = x_{k,j} \cdot (\mu_j \mu_k (\beta_k - \beta_j))\, dt \quad \forall k \neq j. \tag{10}$$

The proof of Lemma 6 is provided in Appendix C.1. Lemma 6 essentially implies the global convergence of SGHA. Specifically, the solution of (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 in the order of $\mathcal{O}(\eta)$ 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 the order of $\mathcal{O}(1)$. The following lemma characterizes how the algorithm escapes from the saddle, i.e., $w^{(\eta)}(0) \approx e_i$, for $i \neq 1$, in Phase I.

Lemma 7. *Suppose Assumption 2 holds and the initial is close to a saddle point, i.e., given constants $\delta < \frac{1}{2}$ and D, there exists an $i \in [d] \backslash \{1\}$ such that*

$$|w_i^{(\eta)} - 1| \leq D\eta^{\frac{1}{2}+\delta} \quad \text{and} \quad |w_j^{(\eta)}| \leq D\eta^{\frac{1}{2}+\delta} \quad \forall j \neq i.$$

As $\eta \to 0$, then for $\forall j \neq i$ $z_j^{(\eta)}$ weakly converges to the solution of the following SDE:

$$dz_j(t) = (-\beta_j \mu_i \cdot z_i + \lambda_i z_i)\, dt + \sqrt{G_{j,i}} dB(t), \tag{11}$$

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.

The proof of Lemma 7 is provided in Appendix C.2. Note that (11) is a Fokker-Plank equation, whose solution is an Ornstein-Uhlenbeck (O-U) process (Doob, 1942) as follows:

$$z_j(t) = \exp\left[-\mu_j \left(\beta_i - \beta_j\right) t\right] \cdot \underbrace{\left[z_j(0) + \sqrt{G_{j,i}} \int_0^t \exp\left[\mu_j \left(\beta_i - \beta_j\right) s\right] dB(s)\right]}_{Q_1}. \tag{12}$$

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 8. *Suppose that Assumption 2 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 for $\forall i \neq 1$, $z_i^{(\eta)}$ weakly converges to the solution of the following SDE:*

$$dz_i(t) = (-\beta_1 \cdot \mu_i z_i + \lambda_i z_i)\, dt + \sqrt{G_{i,1}} dB(t), \tag{13}$$

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 8 is provided in Appendix C.3. The solution of (13) is as follows:

$$z_i(t) = \sqrt{G_{i,1}} \int_0^t \exp\left[\mu_i \left(\beta_1 - \beta_i\right)(s-t)\right] dB(s) + z_i(0) \cdot \exp[-\mu_i \left(\beta_1 - \beta_i\right) t]. \tag{14}$$

Note the second term of the right hand side in (14) 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.

By Lemma 6, 7, and 8, we have the next theorem.

Theorem 9. *Suppose Assumption 2 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

$$T \asymp \frac{\mu_{\max}/\mu_{\min}}{\mu_1 \cdot \mathsf{gap}} \log\left(\eta^{-1}\right) \tag{15}$$

such that with probability at least $\frac{5}{8}$, $||w(T) - W^*||_2^2 \leq \epsilon$, *where* W^* *is the optima of* (3).

The proof of Theorem 9 is provided in Appendix C.4. Theorem 9 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 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, problem (3) is more difficult to solve.

4.3 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) 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 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 under the commutative assumption. We suspect 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 Discussion

Here we briefly discuss a few related works:

- Li et al. (2016b) propose a framework for characterizing the stationary points in the unconstrained nonconvex matrix factorization problem, while our GEV problem is constrained. Different from their analysis, we analyze the optimization landscape of the corresponding Lagrangian function. When characterize the stationary points, we need to take both primal and dual variables into consideration, which is technically more challenging.

- Ge et al. (2016a) consider the (off-line) generalized eigenvalue problem in a finite sum form. Unlike online setting, they access exact A and B in each iteration. Specifically, they need to access exact A and B to compute an approximate inverse of B to find the descent direction. Meanwhile, they also need a modified Gram Schmidt process, which also requires accessing exact B, to maintain the solution on the generalized Stiefel manifold (defined by $X^\top B X = I_r$ via exact B, Mishra and Sepulchre (2016)). Our proposed stochastic search, however, is a full stochastic primal-dual algorithm, which neither require accessing exact A and B, nor enforcing the the primal variables to stay on the manifold.

Acknowledgements

This research was partially supported by DARPA Young Faculty Award N66001-14-1-4047.

References

Allen-Zhu, Z. and Li, Y. (2016). Doubly accelerated methods for faster CCA and generalized eigendecomposition. *arXiv preprint arXiv:1607.06017* .

Arora, R., Marinov, T. V., Mianjy, P. and Srebro, N. (2017). Stochastic approximation for canonical correlation analysis. In *Advances in Neural Information Processing Systems*.

Bhojanapalli, S., Neyshabur, B. and Srebro, N. (2016). Global optimality of local search for low rank matrix recovery. *arXiv preprint arXiv:1605.07221* .

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Chen, Y., Lan, G. and Ouyang, Y. (2014). Optimal primal-dual methods for a class of saddle point problems. *SIAM Journal on Optimization* **24** 1779–1814.

Chen, Z., Yang, F. L., Li, C. J. and Zhao, T. (2017). Online multiview representation learning: Dropping convexity for better efficiency. *arXiv preprint arXiv:1702.08134* .

Cook, R. D. and Ni, L. (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *Journal of the American Statistical Association* **100** 410–428.

Doob, J. L. (1942). The brownian movement and stochastic equations. *Annals of Mathematics* 351–369.

Dummit, D. S. and Foote, R. M. (2004). *Abstract algebra*, vol. 3. Wiley Hoboken.

Ethier, S. N. and Kurtz, T. G. (2009). *Markov processes: characterization and convergence*, vol. 282. John Wiley & Sons.

Friedman, J., Hastie, T. and Tibshirani, R. (2001). *The elements of statistical learning*, vol. 1. Springer series in statistics New York.

Ge, R., Huang, F., Jin, C. and Yuan, Y. (2015). Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*.

Ge, R., Jin, C., Netrapalli, P., Sidford, A. et al. (2016a). Efficient algorithms for large-scale generalized eigenvector computation and canonical correlation analysis. In *International Conference on Machine Learning*.

Ge, R., Lee, J. D. and Ma, T. (2016b). Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*.

Gorrell, G. (2006). Generalized hebbian algorithm for incremental singular value decomposition in natural language processing. In *EACL*, vol. 6. Citeseer.

Harold, J., Kushner, G. and Yin, G. (1997). Stochastic approximation and recursive algorithm and applications. *Application of Mathematics* **35**.

Iouditski, A. and Nesterov, Y. (2014). Primal-dual subgradient methods for minimizing uniformly convex functions. *arXiv preprint arXiv:1401.1792* .

Kushner, H. and Yin, G. G. (2003). *Stochastic approximation and recursive algorithms and applications*, vol. 35. Springer Science & Business Media.

Lan, G., Lu, Z. and Monteiro, R. D. (2011). Primal-dual first-order methods with {\mathcal {O}(1/\epsilon)} iteration-complexity for cone programming. *Mathematical Programming* **126** 1–29.

Li, C. J., Wang, M., Liu, H. and Zhang, T. (2016a). Near-optimal stochastic approximation for online principal component estimation. *arXiv preprint arXiv:1603.05305* .

Li, X., Wang, Z., Lu, J., Arora, R., Haupt, J., Liu, H. and Zhao, T. (2016b). Symmetry, saddle points, and global geometry of nonconvex matrix factorization. *arXiv preprint arXiv:1612.09296* .

Liu, T., Chen, Z., Zhou, E. and Zhao, T. (2018). Toward deeper understanding of nonconvex stochastic optimization with momentum using diffusion approximations. *arXiv preprint arXiv:1802.05155* .

Luenberger, D. G., Ye, Y. et al. (1984). *Linear and nonlinear programming*, vol. 2. Springer.

Luo, Z.-Q. and Tseng, P. (1993). Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research* **46** 157–178.

Mika, S., Ratsch, G., Weston, J., Scholkopf, B. and Mullers, K.-R. (1999). Fisher discriminant analysis with kernels. In *Neural networks for signal processing IX, 1999. Proceedings of the 1999 IEEE signal processing society workshop*. Ieee.

Mishra, B. and Sepulchre, R. (2016). Riemannian preconditioning. *SIAM Journal on Optimization* **26** 635–660.

Polyak, B. T. (1963). Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki* **3** 643–653.

Sanger, T. D. (1989). Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural networks* **2** 459–473.

Shapiro, A., Dentcheva, D. and Ruszczyński, A. (2009). *Lectures on stochastic programming: modeling and theory*. SIAM.

Sun, J., Qu, Q. and Wright, J. (2016). A geometric analysis of phase retrieval. In *Information Theory (ISIT), 2016 IEEE International Symposium on*. IEEE.

Sutton, R. S., McAllester, D. A., Singh, S. P. and Mansour, Y. (2000). Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*.

Zhu, Z., Li, Q., Tang, G. and Wakin, M. B. (2017). The global optimization geometry of nonsymmetric matrix factorization and sensing. *arXiv preprint arXiv:1703.01256* .

A Proofs for Determining Stationary Points

A.1 Proof of Theorem 4

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

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

Then, 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 $\Psi \in \mathcal{G}$, one can verify 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 $\phi = c_1 O^{\widetilde{A}}_{:,i} + c_2 O^{\widetilde{A}}_{:,j}$ 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 follows 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 inequality holds because of $c_1, c_2 \neq 0$, $c_1^2 + c_2^2 = 1$, $\lambda^{\widetilde{A}}_j \neq \lambda^{\widetilde{A}}_j$ for $i \neq j$. □

A.2 Proof of Theorem 5

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

$$H_X = 2\,\mathrm{sym}\left(I_r \otimes ((BXX^\top - I_d)A) + (X^\top AX) \otimes B + (AX) \boxtimes (BX)\right), \tag{16}$$

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

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

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 holds at X.

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

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

where $(*)$ is from $c_{j_i} = 0$ 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

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

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

□

B Singular case for B

When B **is Singular**, we assume $\operatorname{rank}(B) = m < d$ and $\operatorname{rank}(A) = d$. Note that we require $m \geq r$; Otherwise, the feasible region of (3) becomes $\mathcal{T}_B = \emptyset$.

Before we proceed with our analysis, we first exclude an ill-defined case, where the objective function of (3) is unbounded from above. The following proposition shows the sufficient and necessary condition of the existence of the global optima of (3).

Proposition 10. *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 (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 $u_i \in \mathrm{Col}(B)$ and each column of $X_{B^\perp} = [v_1, ..., v_r]$ with $v_i \in \mathrm{Null}(B)$. Note such a decomposition is unique. Then (3) becomes

$$\min -\sum_{i=1}^{r}(u_i^\top Au_i) - 2\sum_{i=1}^{r}(u_i^\top Av_i) - \sum_{i=1}^{r}(v_i^\top Av_i) \quad \text{s.t.} \quad X_B^\top BX_B = I_r. \tag{17}$$

If (17) has an optimal solution, we have $v^\top Av \le 0$, for all $v \in \mathrm{Null}(B)$; otherwise, fixing the feasible X_B, we use $X_B = [\lambda v, ..., \lambda v]$ and increase λ, then there is no lower bound of the objective function. Further, given a vector $v \in \mathrm{Null}(B)$ with $v^\top Av = 0$, $u^\top Av = 0$ must hold for all $u \in \mathrm{Col}(B)$; otherwise, W.L.O.G, we assume that $u_1 \in \mathrm{Col}(B)$ $u_1^\top Av > 0$, we can construct a feasible $X_B = \mu[u_1, ..., u_r]$, where μ is a normalization constant such that $\mu^2 u_1^\top Bu_1 = 1$. Then constructing $X_{B^\perp} = \lambda[v, 0, ...0]$., if we increase λ, there is no lower bound of the objective function. Therefore, for a vector $v \in \mathrm{Null}(B)$, 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 $\Lambda_{11}^B = \mathrm{diag}(\lambda_1, ..., \lambda_m)$ with $\lambda_1 \ge \cdots \ge \lambda_m > 0$. We then left multiply $O^{B\top}$ and right multiply O^B to A:

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

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

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

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

$$AX = BX \cdot X^\top AX \quad \text{and} \quad X^\top BX = I_r. \tag{18}$$

For notational simplicity, we define

$$V(P) := \left(\Lambda^{B\dagger}(P)\right)^{-\frac{1}{2}} O^{B\top}\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} V_1 \\ V_2(P) \end{bmatrix}, \tag{19}$$

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 (19), we have

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = O^B \left(\Lambda^{B\dagger}(P)\right)^{\frac{1}{2}} \begin{bmatrix} V_1 \\ V_2(P) \end{bmatrix}. \tag{20}$$

Combining (20) and (18), we get the following equation system:

$$\begin{cases} \widetilde{A}(P)V(P) = \begin{bmatrix} V_1 \\ 0 \end{bmatrix} V(P)^\top \widetilde{A}(P)V(P), & \text{(21a)} \\ V(P)^\top \mathrm{diag}(I_m, 0)V(P) = I_r, & \text{(21b)} \end{cases}$$

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 (18) is equivalent to doing the transformation (20) to the solution of (21). We then denote

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

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 11. *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 $\operatorname{rank}(B) = m < d$, satisfying the well-defined condition in Proposition 10,$(X, \mathcal{D}(X))$ is an equilibrium of $\mathcal{L}(X, Y)$ if and only if X can be represented as*

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

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

Proof. By definition, we have

$$\begin{cases} AX = BX \cdot Y \\ X^\top BX = I_r \end{cases} \Longrightarrow \begin{cases} AX = BX \cdot X^\top AX \\ X^\top BX = I_r \end{cases}. \tag{22}$$

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 (22) 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

$$\begin{cases} \widetilde{A}(P) V(P) = \begin{bmatrix} V_1 \\ 0 \end{bmatrix} V(P)^\top \widetilde{A}(P) V(P) & \text{(23a)} \\ V(P)^\top \operatorname{diag}(I_m, 0) V(P) = I_r, & \text{(23b)} \end{cases}$$

where $\widetilde{A}(P) = (\Lambda^{B\dagger}(P))^{\frac{1}{2}} W (\Lambda^{B\dagger}(P))^{\frac{1}{2}}$. Simplifying (23a), we obtain

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

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 (23), we obtain the following equations:

$$\begin{cases} \widehat{A} V_1 = V_1 V_1^\top \widehat{A} V_1 & \text{(24a)} \\ V_1^\top V_1 = I_r. & \text{(24b)} \end{cases}$$

Note (24) are the KKT conditions of the following problem:

$$V_1^* = \underset{V_1 \in \mathbb{R}^{m\times r}}{\operatorname{argmin}} - \operatorname{tr}(V_1^\top \widehat{A} V_1) \quad \text{s.t.} \quad V_1^\top V_1 = I_r. \tag{25}$$

Because (25) is not a degenerate case, Theorem 4 can be directly applied to (25). Then, we get the stable equilibria and unstable equilibria of (25). 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 (24) can be represented as $V_1 = O_{:,\mathcal{I}}^{\widehat{A}} \cdot \Psi$, where $\mathcal{I} \in \Big\{\{i_1, ..., i_r\} : \{i_1, ..., i_r\} \subseteq [m]\Big\}$ and $\Psi \in \mathcal{G}$. Then, we know the primal variable X at an equilibrium of $\mathcal{L}(X, Y)$ satisfies

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

where $O_{:,\mathcal{I}}^{\widehat{A}}$ is an equilibrium for the Lagrangian function of (25). □

Theorem 11 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.

C Proofs for the Convergence Rate of Algorithm.

C.1 Proof of Lemma 6

Proof. Denote $k = \lfloor \frac{t}{\eta} \rfloor$, and $\Delta(t) = w^{(\eta)}(t+\eta) - w^{(\eta)}(t)$ 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

$$\begin{aligned}
\frac{1}{\eta}\mathbb{E}\left(\Delta(t)\Big|w^{(\eta)}(t)\right) &= \frac{1}{\eta}\mathbb{E}\left(W^{(k+1)} - W^{(k)}\big|W^{(k)}\right) \\
&= \frac{1}{\eta}\mathbb{E}\left[\eta\left(\Lambda^B - (\Lambda^B)^{\frac{1}{2}}\widehat{\Lambda}_B^{(k)}(\Lambda^B)^{-\frac{1}{2}}W^{(k)}W^{(k)\top}\right)\cdot\widetilde{\Lambda}^{(k)}W^{(k)}\big|W^{(k)}\right] \\
&= \Lambda^A w^{(\eta)}(t) - \left(w^{(\eta)}(t)\right)^{\top}\left(\Lambda^B\right)^{-\frac{1}{2}}\Lambda^A\left(\Lambda^B\right)^{-\frac{1}{2}}w^{(\eta)}(t)\Lambda^B w^{(\eta)}(t). \qquad (26)
\end{aligned}$$

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{\left(w_i^{(\eta)}\right)^{\mu_1}}{\left(w_1^{(\eta)}\right)^{\mu_i}}(t+\eta) - \frac{\left(w_i^{(\eta)}\right)^{\mu_1}}{\left(w_1^{(\eta)}\right)^{\mu_i}}(t)\Bigg|\frac{\left(w_i^{(\eta)}\right)^{\mu_1}}{\left(w_1^{(\eta)}\right)^{\mu_i}}(t)\right) \\
=&\frac{1}{\eta}\mathbb{E}\left(\frac{\left(w_i^{(\eta)}(t)+\Delta_i\right)^{\mu_1}}{\left(w_1^{(\eta)}(t)+\Delta_1\right)^{\mu_i}} - \frac{\left(w_i^{(\eta)}(t)\right)^{\mu_1}}{\left(w_1^{(\eta)}(t)\right)^{\mu_i}}\Bigg|\frac{\left(w_i^{(\eta)}(t)\right)^{\mu_1}}{\left(w_1^{(\eta)}(t)\right)^{\mu_i}}\right) \\
=&\frac{1}{\eta}\mathbb{E}\left(\frac{\left(w_i^{(\eta)}(t)\right)^{\mu_1}}{\left(w_1^{(\eta)}(t)\right)^{\mu_i}}\cdot\frac{(1+\frac{\Delta_i}{w_i^{(\eta)}(t)})^{\mu_1}}{(1+\frac{\Delta_1}{w_1^{(\eta)}(t)})^{\mu_i}} - \frac{\left(w_i^{(\eta)}(t)\right)^{\mu_1}}{\left(w_1^{(\eta)}(t)\right)^{\mu_i}}\Bigg|\frac{\left(w_i^{(\eta)}(t)\right)^{\mu_1}}{\left(w_1^{(\eta)}(t)\right)^{\mu_i}}\right) \\
=&\frac{1}{\eta}\frac{\left(w_i^{(\eta)}(t)\right)^{\mu_1}}{\left(w_1^{(\eta)}(t)\right)^{\mu_i}}\mathbb{E}\left([1+\mu_1\frac{\Delta_i}{w_i^{(\eta)}}+\mathcal{O}(\eta^2)]\cdot[1-\mu_i\frac{\Delta_1}{w_1^{(\eta)}}+\mathcal{O}(\eta^2)]-1\Big|w^{(\eta)}(t)\right) \\
=&\frac{1}{\eta}\frac{\left(w_i^{(\eta)}(t)\right)^{\mu_1}}{\left(w_1^{(\eta)}(t)\right)^{\mu_i}}\left(\frac{\mu_1}{w_i^{(\eta)}}\mathbb{E}(\Delta_i\big|w^{(\eta)}(t)) - \frac{\mu_i}{w_1^{(\eta)}}\mathbb{E}(\Delta_1\big|w^{(\eta)}(t))\right)+\mathcal{O}(\eta) \\
=&\frac{\left(w_i^{(\eta)}(t)\right)^{\mu_1}}{\left(w_1^{(\eta)}(t)\right)^{\mu_i}}\Big[\frac{\mu_1}{w_i^{(\eta)}}\left(-\sum_{k=1}^{d}\frac{\lambda_k}{\mu_k}\left(w_k^{(\eta)}\right)^2\mu_i w_i^{(\eta)}+\lambda_i w_i^{(\eta)}\right) - \\
&\qquad\frac{\mu_i}{w_1^{(\eta)}}\left(-\sum_{k=1}^{d}\frac{\lambda_k}{\mu_k}\left(w_k^{(\eta)}\right)^2\mu_1 w_1^{(\eta)}+\lambda_1 w_1^{(\eta)}\right)\Big]+\mathcal{O}(\eta) \\
=&\frac{\left(w_i^{(\eta)}(t)\right)^{\mu_1}}{\left(w_1^{(\eta)}(t)\right)^{\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 (26). Then, we calculate the infinitesimal conditional variance. From the update of W in (9), if $t \in [0,\ T]$ with a finite T, then $w^{(\eta)}(t)$ is bounded with probability 1. Denote

$||w^{(\eta)}(t)||_2^2 \leq D < \infty$. Then we have

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

where the second inequality holds because of the mean inequality and the last inequality is from the independence of $A^{(k)}$ and $B^{(k)}$, $(w^{(\eta)})^\top\widetilde{\Lambda}w^{(\eta)} \leq ||\widetilde{\Lambda}||_2(w^{(\eta)})^\top w^{(\eta)} \leq \frac{||A^{(k)}||_2}{\mu_{\min}}D$ (since $\widetilde{\Lambda}$ is symmetric), and $C = \frac{\left(w_i^{(\eta)}(t)\right)^{2\mu_1-2}}{\left(w_1^{(\eta)}(t)\right)^{2\mu_i}}$. By Section 4 of Chapter 7 in Ethier and Kurtz (2009), 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 (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 that 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:

$$
\frac{dw_i}{dt} = \mu_i(\beta_i - \sum_{j=1}^d \beta_j w_j^2)w_i. \tag{27}
$$

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 that our algorithm converges to the neighbor of the global optima. □

C.2 Proof of Lemma 7

Proof. Suppose the initial is near the saddle point. Without loss of generality, we assume that $|w_i^{(\eta)}(0) - 1| \leq C\eta^{\frac{1}{2}+\delta}$ and $|w_i^{(\eta)}(0)_j| \leq C\eta^{\frac{1}{2}+\delta}$ for all $j \neq i$. We calculate the infinitesimal conditional expectation for $j \neq i$

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

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

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

By Section 4 of Chapter 7 in Ethier and Kurtz (2009), we obtain that the algorithm converges to the solution of (13) if it is already near the saddle point e_i. □

C.3 Proof of Lemma 8

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 (9), we show $w^{(\eta)\top} w^{(\eta)}(t)$ weakly converges to the following ODE by a similar proof in Lemma 6:

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

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

$$x(t) = \begin{cases} \frac{1}{1-\exp(-\lambda_1 t + C)} & \text{if } x(0) > 1 \\ \frac{1}{1+\exp(-\lambda_1 t + C)} & \text{if } x(0) < 1 \\ 1 & \text{if } x(0) = 1 \end{cases}.$$

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

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

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

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

By Section 4 of Chapter 7 in Ethier and Kurtz (2009), we have that the algorithm converges to the solution of (13) if it is already near our optimal solution. □

C.4 Proof of Theorem 9

Proof. Assume the initial is near a saddle point, e_i, where $i \neq 1$. According to (11) and Lemma 7, we obtain the closed form solution of (11) 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 (32) be larger than 95%. Then with a sufficiently small η, we need

$$T_1 \asymp \frac{1}{\mu_1(\beta_1-\beta_i)}\log\left(\frac{200(\beta_1-\beta_i)}{\mu_1 G_{1,i}}+1\right) \tag{28}$$

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 6 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

$$T_2 \asymp \frac{\mu_{\max}}{\mu_1\mu_{\min}\cdot \mathsf{gap}}\log(\eta^{-1}) \tag{29}$$

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 6. After restarting the counter of time, we have

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

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 6. Therefore, given $\sum_{i=2}^d w_i^2 \leq \kappa\eta^{1+2\delta}$ and a sufficiently small η, we need

$$T_2' \asymp \frac{\mu_{\max}}{\mu_1\mu_{\min}\cdot \mathsf{gap}}\log(\eta^{-1}) \tag{30}$$

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 8 and (13), we obtain the closed form solution of (13) 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 Ito-Integral, we have

$$\mathbb{E}\left(z_i(t)\right)^2 = \left(z_i(0)\right)^2 e^{-2\mu_i(\beta_1-\beta_i)t} + \frac{G_{i,1}}{2\mu_i\left(\beta_1-\beta_i\right)}\left[1-e^{-2\mu_i(\beta_1-\beta_i)t}\right]. \tag{31}$$

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

$$\begin{aligned}
&\mathbb{P}(w_1^2 \le 1-\epsilon)\\
=&\mathbb{P}\left(\sum_{i=2}^d w_i^2 \ge \epsilon\right) \le \frac{\mathbb{E}\left(\sum_{i=2}^d w_i^2\right)}{\epsilon} = \frac{\mathbb{E}\left(\sum_{i=2}^d z_i^2\right)}{\eta^{-1}\epsilon}\\
=&\frac{1}{\eta^{-1}\epsilon}\left(\sum_{i=2}^d \left(z_i(0)\right)^2 e^{-2\mu_i(\beta_1-\beta_i)t} + \frac{G_i}{2\mu_i\left(\beta_1-\beta_i\right)}\left[1-e^{-2\mu_i(\beta_1-\beta_i)t}\right]\right)\\
\le&\frac{1}{\eta^{-1}\epsilon}\left(\eta^{-1}\delta^2 e^{-2\mu_{\min}\cdot\mathsf{gap}\cdot\mathsf{t}} + \frac{\phi}{2\mu_{\min}\cdot\mathsf{gap}}\right).
\end{aligned} \tag{32}$$

Let the right hand side of (32) be no larger than $\frac{1}{16}$.

$$\begin{aligned}
&\frac{1}{\eta^{-1}\epsilon}\left(\eta^{-1}\delta^2 e^{-2\mu_{\min}\cdot\mathsf{gap}\cdot t} + \frac{\phi}{2\mu_{\min}\cdot\mathsf{gap}}\right) \le \frac{1}{16}\\
\implies\ & e^{2\mu_{\min}\cdot\mathsf{gap}\cdot t} \ge \frac{16\cdot\mu_{\min}\cdot\mathsf{gap}\cdot\delta^2}{\epsilon\cdot\mu_{\min}\cdot\mathrm{gap} - 16\cdot\eta\cdot\phi}.
\end{aligned}$$

Then after restarting the counter of time, we need

$$T_3 \asymp \frac{1}{\mu_{\min}\cdot\mathsf{gap}}\cdot\log\left(\frac{\mu_{\min}\cdot\mathsf{gap}\cdot\delta^2}{\epsilon\cdot\mu_{\min}\cdot\mathrm{gap} - 16\cdot\eta\cdot\phi}\right) \tag{33}$$

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

Combining (28), (29), (30), (33), if our algorithm starts from a saddle point, then with probability at least $\frac{5}{8}$, we need

$$T = T_1 + T_2 + T_2' + T_3 \asymp \frac{\mu_{\max}/\mu_{\min}}{\mu_1\cdot\mathsf{gap}}\log\left(\eta^{-1}\right) \tag{34}$$

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

Moreover, we choose

$$\eta \asymp \frac{\epsilon\cdot\mu_{\min}\cdot\mathsf{gap}}{\phi}. \tag{35}$$

Combining (34) and (35) together, we get the asymptotic sample complexity

$$N \asymp \frac{T}{\eta} \asymp \frac{\phi\cdot\mu_{\max}/\mu_{\min}}{\epsilon\cdot\mu_1\cdot\mu_{\min}\cdot\mathsf{gap}^2}\log\left(\frac{\phi}{\epsilon\cdot\mu_{\min}\cdot\mathsf{gap}}\right) \tag{36}$$

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