Online Generalized Eigenvalue Decomposition: Primal Dual Geometry and Inverse-Free Stochastic Optimization

Xingguo Li[*] lixx1661@umn.edu
Zhehui Chen[†] zhchen@gatech.edu
Lin Yang[◇] lin.yang@princeton.edu
Jarvis Haupt[*] jdhaupt@umn.edu
Tuo Zhao[†] tourzhao@gatech.edu
[*]*University of Minnesota* [◇]*Princeton University* [†]*Georgia Institute of Technology*

Abstract

The generalized eigenvalue decomposition (GEV) can be rewritten as a constrained minimization problem with a nonconvex objective and constraint. Using the Lagrangian multiplier method, we recast GEV into an unconstrained min-max problem. By exploiting the underlying symmetry of the min-max problem, we investigate the mechanism that generates unstable stationary points. We then show that all stationary points are unstable except the convex-concave saddle points, which correspond to the global optima of the original constrained problem. We apply a stochastic generalized Hebbian algorithm (SGHA) to solve the GEV problem without any sophisticated (approximate) matrix version operation. By applying a diffusion approximation analysis, we obtain a global rate of convergence for the limiting process of SGHA under a simultaneously orthogonal diagonalizable condition. Numerical results are provided to support our theory.

1 Introduction

We consider the generalized eigenvalue decomposition (GEV) [23] as follows:

$$X^* = \operatorname{argmin}_{X\in\mathbb{R}^{d\times r}} \ \mathcal{F}(X) = -\operatorname{tr}(X^\top AX) \text{ subject to } X^\top BX = I_r, \tag{1}$$

where $A, B \in \mathbb{R}^{d\times d}$ are symmetric, B is positive semidefinite, and $I_r \in \mathbb{R}^{r\times r}$ is the identity matrix. As a generalization of the ordinary eigenvalue problem [15], GEV is closely related to popular methods for classification, dimension reduction, and feature extraction [5, 17, 21, 22] in practice.

Significant efforts have been made on designing efficient solvers for GEV. Although (1) is nonconvex, there are many algorithms that can obtain global optima in polynomial time under the batch or finite sums settings [2, 3, 6, 13, 19], where $A = \frac{1}{n}\sum_{k=1}^n A^{(k)}$ and $B = \frac{1}{n}\sum_{k=1}^n B^{(k)}$. All these algorithms, however, require intensively computing the approximate inverse of B, which make them not applicable to the online setting, which is our particular interest here. Specifically, at each iteration, we obtain independent stochastic approximations $A^{(k)}$'s and $B^{(k)}$ in a streaming fashion with $\mathbb{E}A^{(k)} = A$ and $\mathbb{E}B^{(k)} = B$. At the $(k+1)$-th iteration, $A^{(k)}$'s and $B^{(k)}$ are discarded. Thus, it is impossible to get good approximate inverses of B in such settings.

To overcome this drawback, we recast the GEV problem (1) as an unconstrained min-max problem. Specifically, using the Lagrangian multiplier method with the dual variable $Y \in \mathbb{R}^{r\times r}$, we solve

$$\min_X \max_Y \ \mathcal{L}(X,Y), \quad \text{where } \mathcal{L}(X,Y) = -\operatorname{tr}(X^\top AX) + \langle Y, X^\top BX - I_r\rangle, \tag{2}$$

where X is the primal variable. Such a primal-dual formulation allows us to develop new stochastic algorithms without (approximate) inversion of B. Since (1) is nonconvex, the associated min-max problem (2) does not have a nice convex-concave structure (convex in X and concave in Y). Existing theory on convex-concave saddle point problems cannot be applied for analyzing the convergence of stochastic algorithms for solving (2) directly. To the best of our knowledge, existing literature only studies the properties of eigenvalues and eigenspaces for special type of A and B [4, 9]. There is no clear characterization of the generic geometry for GEV. Especially, it is unclear yet what are the complete set of stationary points and the intrinsic mechanism that generates these stationary points.

Our first goal is to answer the questions above. Recent progress has been made to characterize the geometry for nonconvex problems [8, 20, 24]. However, these are either unconstrained or with simple spherical

OPTML 2017: 10th NIPS Workshop on *Optimization for Machine Learning* (NIPS 2017).

constraint that limits their applicability to (1). Here, we leverage the symmetry property and invariant group, also discussed in [18] for unconstrained problems, to characterize all stationary points (including unstable ones and global optima) of our constrained problem. Such a characterization is also closely related to the topology of $\mathcal{F}(X)$ on the generalized Stiefel manifold [1]. With clear geometry, our second goal is to propose a stochastic variant of the generalized Hebbian algorithm [10], called SGHA, for the GEV problem with a global convergence under a simultaneously orthogonal diagonalizable condition. Our convergence analysis is motivated by characterizing the associated diffusion process, where the discrete trajectory of the stochastic update is approximated by solution of a stochastic differential equation under the asymptotic setting, leveraging the idea in [12, 16]. Though online algorithms are popular and scalable [7, 11, 14], there is no iteration complexity analysis for solving the GEV problem. Guarantees exists only for simple orthogonal constraints, e.g., for tensor decomposition [8]. Ours is the first iteration complexity analysis of a simple online algorithm for solving the GEV problem without inversion.

2 Characterization of Stationary and Saddle Points

Recall that we consider the min-max problem $\min_X \max_Y \mathcal{L}(X,Y)$ in (2). By KKT conditions of the primal and dual variables, X and Y at a stationary point satisfy $\nabla_X \mathcal{L}(X,Y)=0$ and $\nabla_Y \mathcal{L}(X,Y)=0$. This indicates $Y=\mathcal{D}(X)\triangleq X^\top AX$ at a stationary point. Denote the gradient by $\nabla\mathcal{L}\triangleq\begin{bmatrix}\nabla_X\mathcal{L}(X,Y)\\ \nabla_Y\mathcal{L}(X,Y)\end{bmatrix}=\begin{bmatrix}2BXY-2AX\\ X^\top BX-I_r\end{bmatrix}$. Our aim is to find the set of stationary points of $\mathcal{L}(X,Y)$ and further distinguish unstable ones and convex-concave saddle points defined as follows:

Definition 1. *Given the Lagrangian function $\mathcal{L}(X,Y)$, (X,Y) is called:* **(1)** *A **stationary point** of $\mathcal{L}(X,Y)$, if $\nabla\mathcal{L}=0$;* **(2)** *An **unstable stationary point** of $\mathcal{L}(X,Y)$, if (X,Y) is a stationary point and for any neighborhood $\mathcal{B}\subseteq\mathbb{R}^{d\times r}$ of X, there exist $X_1,X_2\in\mathcal{B}$ such that $\mathcal{L}(X_1,Y)|_{Y=\mathcal{D}(X_1)}\le\mathcal{L}(X,Y)|_{Y=\mathcal{D}(X)}\le\mathcal{L}(X_2,Y)|_{Y=\mathcal{D}(X_2)}$ and $\lambda_{\min}(\nabla_X^2\mathcal{L}(X,Y)|_{Y=\mathcal{D}(X)})<0$;* **(3)** *A **convex-concave saddle point**, or a **minimax point** of $\mathcal{L}(X,Y)$, if (X,Y) is a stationary point and (X,Y) is a global optimum to (1), i.e. $(X,Y)=\arg\min_{\widetilde{X}}\max_{\widetilde{Y}}\mathcal{L}(\widetilde{X},\widetilde{Y})$.*

In general, it requires to solve an expensive large system $\nabla\mathcal{L}=0$ for finding stationary points. Here, we apply a symmetry property for functions with an invariant group to facilitate a more efficient characterization. We focus on nonsingular B, and the extension to singular B is analogous.

2.1 Invariant Group and Symmetry Property

It is straightforward that $\mathcal{L}(X,Y)$ has an invariant group $\mathcal{G}=\{\Psi\in\mathbb{R}^{r\times r}:\Psi\Psi^\top=\Psi^\top\Psi=I_r\}$. Further consider orthogonal decomposition $\mathbb{R}^{d\times r}=\mathcal{U}\oplus\mathcal{V}$, where $\oplus$ is the direct sum, $\mathcal{U}=\{U\in\mathbb{R}^{d\times r_1}\}$ and $\mathcal{V}=\{V\in\mathbb{R}^{d\times(r-r_1)}:V^\top U=0\text{ for all }U\in\mathcal{U}\}$. Given $V\in\mathcal{V}$, a subgroup of $\mathcal{G}$ is induced as $\mathcal{G}_{\mathcal{U}}(V)=\{g_{\mathcal{U}}:g_{\mathcal{U}}(V)=g(U\oplus V),g\in\mathcal{G},U\in\mathcal{U}\}$. Denote the eigendecomposition of $B^{-1/2}AB^{-1/2}=O^\dagger\Lambda^\dagger(O^\dagger)^\top$, where B^{-1} is the inverse of B and $\Lambda^\dagger$ is a diagonal matrix with eigenvalues $\lambda_1^\dagger\ge\cdots\ge\lambda_d^\dagger$. Then we characterize the stationary point of $\mathcal{L}(X,Y)$ as follows.

Theorem 2 (Symmetry Property)**.** *Suppose A and B are symmetric, B is nonsingular with the subspace pair $(\mathcal{U}_{\mathcal{S}},\mathcal{V}_{\widetilde{\mathcal{S}}})$ defined as $\mathcal{U}_{\mathcal{S}}=\{U\in\mathbb{R}^{d\times s}:U=O^\dagger_{:,\mathcal{S}},\mathcal{S}\subseteq[r]\text{ with }|\mathcal{S}|=s\le r\}$ and $\mathcal{V}_{\widetilde{\mathcal{S}}}=\{V\in\mathbb{R}^{d\times(r-s)}:V=O^\dagger_{:,\widetilde{\mathcal{S}}},\widetilde{\mathcal{S}}\subseteq[d]\backslash[r]\text{ with }|\widetilde{\mathcal{S}}|=r-s,|\mathcal{S}|=s\le r\}$. Then X is a stationary point, i.e., $\nabla\mathcal{L}=0$, if and only if $X=B^{-1/2}\widetilde{X}$ for any $\widetilde{X}\in\mathcal{G}_{\mathcal{U}_{\mathcal{S}}}(V)$ with any $V\in\mathcal{V}_{\widetilde{\mathcal{S}}}$.*

From Theorem 2, given a subset of eigenvectors corresponding to top r eigenvalues of $B^{-1/2}AB^{-1/2}$, a stationary point is formed by taking its direct sum with a subset of eigenvectors corresponding to bottom $d-r$ eigenvalues, followed by the invariant group operation via $\mathcal{G}$. Such a symmetry property allows us to find all stationary points in a recursive fashion. The symmetry property is also discussed in [18], but they consider an unconstrained problem and use a fixed point of the invariant group. Due to the min-max structure of our problem, a fixed point of the invariant group does not help, which motivates us to consider a more general symmetry property.

2.2 Unstable Stationary vs. Saddle Point

We then characterize whether a stationary point is unstable or a convex-concave saddle point. Specifically, denote the Hessian matrix $H_X\triangleq\nabla_X^2\mathcal{L}(X,Y)|_{Y=\mathcal{D}(X)}$. Then we analyze the eigenspace of H_X, where X is an unstable stationary point if H_X has both positive and negative eigenvalues or positive semi-definite; or X corresponds to the min-max global optimum if H_X is negative semi-definite. Since B is nonsingular, we can reparameterize (1) as

$$\widetilde{X}^*=\operatorname{argmin}_{\widetilde{X}\in\mathbb{R}^{d\times r}}\ -\operatorname{tr}(\widetilde{X}^\top\widetilde{A}\widetilde{X})\quad\text{s.t.}\quad\widetilde{X}^\top\widetilde{X}=I_r,\tag{3}$$

where $\widetilde{X} = B^{1/2}X$ and $\widetilde{A} = B^{-1/2}AB^{-1/2}$. Given a stationary point $\widetilde{X}$ of (3), we can obtain a stationary point of (1) by $X = B^{-1/2}\widetilde{X}$. The following lemma distinguishes the saddle point of the min-max problem, i.e., the global optimum of (1), and unstable stationary points, respectively.

Lemma 3. *Let $X = B^{-1/2}\widetilde{X}$ for any $\widetilde{X} \in \mathcal{G}_{\mathcal{U}_\mathcal{S}}(V)$ and any $V \in \mathcal{V}_{\widetilde{\mathcal{S}}}$ with $\mathcal{S} \subseteq [r]$. If $\mathcal{S} = [r]$ and $\widetilde{\mathcal{S}} = \emptyset$, then X is a saddle point of the min-max problem. Otherwise, if $\mathcal{S} \subset [r]$ and $\widetilde{\mathcal{S}} \neq \emptyset$, then X is an unstable stationary point with $\lambda_{\min}(H_X) \leq \frac{2(\lambda^\dagger_{\max \mathcal{S}\cup\widetilde{\mathcal{S}}} - \lambda^\dagger_{\min \mathcal{S}^\perp\cap\widetilde{\mathcal{S}}^\perp})}{\|X_{:,\min \mathcal{S}^\perp\cap\widetilde{\mathcal{S}}^\perp}\|_2^2}$ and $\lambda_{\max}(H_X) \geq \frac{4\lambda^\dagger_{\min \mathcal{S}\cup\widetilde{\mathcal{S}}}}{\|X_{:,\min \mathcal{S}\cup\widetilde{\mathcal{S}}}\|_2^2}$, where $\lambda^\dagger_{\max \mathcal{S}}$ ($\lambda^\dagger_{\min \mathcal{S}}$) is the smallest (largest) eigenvalue indexed by a set $\mathcal{S}$ and $\mathcal{S}^\perp = [d]\backslash\mathcal{S}$.*

We have from Lemma 3 that when $\mathcal{U}_\mathcal{S}$ contains eigenvectors corresponding to top r eigenvalue of $B^{-1/2}AB^{-1/2}$, then $X = B^{-1/2}\widetilde{X}$ is the global optimum of (1). Otherwise, any other stationary point is unstable since $\lambda^\dagger_{\max \mathcal{S}^\perp\cap\widetilde{\mathcal{S}}^\perp} - \lambda^\dagger_{\min \mathcal{S}\cup\widetilde{\mathcal{S}}} \leq 0$. Moreover, when $\lambda^\dagger_{\min \mathcal{S}\cup\widetilde{\mathcal{S}}} < 0$, $\mathcal{L}(X, Y)$ may be concave at the corresponding stationary point, which is also unstable. Using the manifold terminology, there are $\binom{d}{r}$ smooth curves corresponding to stationary points, where one of them corresponds to the global optima, and the rest are unstable. It is also important to note that there is no spurious local optimum for the min-max problem (2), i.e., all local optima are global optima.

3 Stochastic Optimization Algorithms without Matrix Inversion

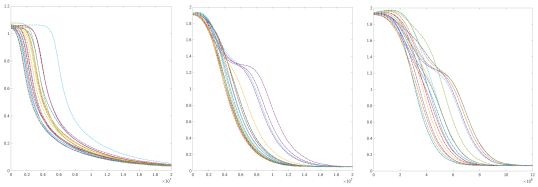
Motivated by the geometric structure, we apply a stochastic variant of the generalized Hebbian algorithm (SGHA) [10]. SGHA is an intuitive primal-dual stochastic algorithm with primal update $X^{(k+1)} \leftarrow X^{(k)} - \eta(B^{(k)}X^{(k)}Y^{(k)} - A^{(k)}X^{(k)})$ and dual update $Y^{(k+1)} \leftarrow X^{(k)\top}A^{(k)}X^{(k)}$, where $\eta > 0$ is the step size parameter. Combining two updates, we have a dual-free update

$$X^{(k+1)} \leftarrow X^{(k)} - \eta\left(B^{(k)}X^{(k)}X^{(k)\top} - I_d\right)A^{(k)}X^{(k)}. \tag{4}$$

The constraint is naturally handled by the dual update. Thus, we do not need to perform any (approximate) matrix inversion or projection onto the constraint set at each iteration. The initial solution $X^{(0)} \in \mathbb{R}^{d\times r}$ only needs to be chosen as a random matrix with orthonormal columns. We then provide numerical to illustrate the efficiency of SGHA and a preliminary convergence analysis.

3.1 Numerical Simulations

We set $d = 500$ with three different settings: (1) Set $A_{ii} = 1/100$ for all $i \in [d]$, and $A_{ij} = 0.5/100$ for all $i \neq j$, and $B_{ij} = 0.5^{|i-j|}/3$ for all $i, j \in [d]$; (2) Randomly generate a orthonormal matrix $U \in \mathbb{R}^{d\times d}$, set $A = U\text{diag}(1, 1, 1, 0.1, ..., 0.1)U^\top$ and $B = U\text{diag}(2, 2, 2, 1, ..., 1)U^\top$; (3) Randomly generate orthonormal matrices $U, V \in \mathbb{R}^{d\times d}$, set $A = U\text{diag}(1, 1, 1, 0.1, ..., 0.1)U^\top$ and $B = V\text{diag}(2, 2, 2, 1, ..., 1)V^\top$. At each iteration of SGHA, we independently sample 40 vectors from $N(0, A)$ and $N(0, B)$, and compute the sample covariance $A^{(k)}$ and $B^{(k)}$ respectively. We repeat the numerical simulations under each setting for 20 times, and present the results in Figure 1. The horizontal axis corresponds to the number of iterations, and the vertical axis corresponds to the optimization error defined as $||B^{1/2}X^{(t)}X^{(t)\top}B^{1/2} - B^{1/2}X^*X^{*\top}B^{1/2}||_\text{F}$. Our results indicate that SGHA convergences to a global optimum in all settings.



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

Figure 1: Plot of errors on: (a) $\eta = 1 \times 10^{-4}$, $r = 1$; (b) $\eta = 5 \times 10^{-5}$, $r = 3$; (c) $\eta = 2.5 \times 10^{-5}$, $r = 3$.

3.2 Convergence Analysis

Before we proceed with our convergence analysis, we first introduce an important assumption.

Assumption 1. *(a) $A^{(k)}$'s and $B^{(k)}$'s are independently sampled from $\mathcal{D}_A$ and $\mathcal{D}_B$ with $\mathbb{E}A^{(k)} = A$ and $\mathbb{E}B^{(k)} = B \succ 0$; (b) A and B are simultaneously orthogonal diagonalizable, i.e., there exists an orthonormal matrix O such that $A = O\Lambda^A O^\top$ and $B = O\Lambda^B O^\top$, where $\Lambda^A = \text{diag}(\lambda_1, ..., \lambda_d)$ with $\lambda_j \neq 0$ and $\Lambda^B = \text{diag}(\mu_1, ..., \mu_d)$; (c) $A^{(k)}$ and $B^{(k)}$ satisfy the moment conditions with bounded spectral norms $\mathbb{E}||A^{(k)}||_2^2 \leq C_0$ and $\mathbb{E}||B^{(k)}||_2^2 \leq C_1$ for generic constants C_0 and C_1.*

Note that Assumption 1.(b) is very strong, which is likely an artifact of our proof technique. To ease the analysis, we consider $r = 1$. Even with these restrictions, the analysis is highly involved. For notational convenience, we define $W^{(k)} = O^\top B^{\frac{1}{2}}X^{(k)}$, then (1) can be rewritten as

$$W^* = \text{argmax}_W \; W^\top\Lambda W \quad \text{subject to} \quad W^\top W = 1, \tag{5}$$

where $\Lambda = (\Lambda^B)^{-\frac{1}{2}}\Lambda^A(\Lambda^B)^{-\frac{1}{2}} = \text{diag}\left(\frac{\lambda_1}{\mu_1}, ..., \frac{\lambda_d}{\mu_d}\right)$ with $\frac{\lambda_1}{\mu_1} > \frac{\lambda_2}{\mu_2} \geq \cdots \geq \frac{\lambda_d}{\mu_d}$. μ_i and λ_i are not necessarily monotonic. We also define $\mathsf{gap} = \frac{\lambda_1}{\mu_1} - \frac{\lambda_2}{\mu_2}$, $\mu_{\min} = \min_{i=2,...,d}\mu_i$, and $\mu_{\max} = \max_{i=2,...d}\mu_i$. One can verify that we can rewrite (4) in the SGHA algorithm as follow:

$$W^{(k+1)} \leftarrow W^{(k)} - \eta\left((\Lambda^B)^{\frac{1}{2}}\widehat{\Lambda}_B^{(k)}(\Lambda^B)^{-\frac{1}{2}}W^{(k)}W^{(k)\top} - (\Lambda^B)\right)\cdot\widetilde{\Lambda}^{(k)}W^{(k)}, \tag{6}$$

where $\widehat{\Lambda}_B^{(k)} = O^\top B^{(k)}O$ and $\widetilde{\Lambda}^{(k)} = O^\top B^{-\frac{1}{2}}A^{(k)}B^{-\frac{1}{2}}O$. Consider the random process defined as $w^{(\eta)}(t) := W^{(\lfloor\frac{t}{\eta}\rfloor)}$.

Theorem 4. *Suppose Assumption 1 holds. Given a sufficiently small pre-specified error $\epsilon > 0$, if we choose a step size $\eta \asymp \frac{\epsilon\cdot\mathsf{gap}}{d\cdot\left(\frac{1}{\mu_1}C_0\cdot C_1 + \mu_{\max}C_1\right)}$, then with probability at least $\frac{3}{4}$, we need a very short time $T = O\left[\frac{1}{\mathsf{gap}\cdot\mu_{\min}}\log\left(\frac{d^{1+\mu_{\max}/\mu_1}}{\epsilon\cdot\mathsf{gap}}\right)\right]$ such that $||w^\eta(T) - W^*||_2^2 \leq \epsilon$ as $\eta \to 0$.*

Note that our analysis implies that the sample complexity not only depends on the eigengap $\frac{\lambda_1}{\mu_1} - \frac{\lambda_2}{\mu_2}$, but also on $\frac{\mu_{\max}}{\mu_{\min}}$, which ratio is analogous to the condition number of B (but with μ_1 excluded).

Our proof contains two major parts: (1) Given a random initialization, we show that the trajectory of the limiting process of our algorithm can be approximated by an ODE; (2) To analyze the sample complexity, we first show that the norm of the solution converges to a constant. Then after proper rescaling of time, the limiting process can be characterized by an SDE.

ODE Characterization: We use $w^{(\eta)}(t)$ to demonstrate the ODE characterization for the trajectory of the limiting process. For notational simplicity, we drop (t) when the context is clear. Instead of showing a global convergence of $w^{(\eta)}$ directly, we show the quantity $\frac{(w_i^{(\eta)})^{\mu_1}}{(w_1^{(\eta)})^{\mu_i}}$ converges to an exponential decay function, where $w_i^{(\eta)}$ is the i-th component of $w^{(\eta)}$.

Lemma 5. *Suppose Assumption 1 holds and initial point is away from saddle points, that is, given pre-specified constants τ and $\delta < \frac{1}{2}$, $||w_1^{(\eta)}||_1 > \tau$ and $||w_i^{(\eta)}||_2 > \eta^{\frac{1}{2}-\delta}$. As the stepsize $\eta \to 0$, quantities $\frac{(w_i^{(\eta)})^{\mu_1}}{(w_1^{(\eta)})^{\mu_i}}$ $\forall i = 2, ..., d$ weakly converge to the solution of the ODE $dx = x\cdot\left(\mu_1\mu_i(\frac{\lambda_i}{\mu_i} - \frac{\lambda_1}{\mu_1})\right)dt$.*

Lemma 5 implies the global convergence of the algorithm. Specifically, the solution of ODE is $x(t) = x(0)\cdot\exp\left(\mu_1\mu_i\left(\frac{\lambda_i}{\mu_i} - \frac{\lambda_1}{\mu_1}\right)t\right)$, $\forall i \in \{2, 3, ..., d\}$, where $x(0)$ is the initial value of $\frac{(w_i^{(\eta)})^{\mu_1}}{(w_1^{(\eta)})^{\mu_i}}$. This implies as $t \to \infty$, the dominated component of w is w_1.

SDE Characterization: ODE approximation of the limiting process implies that after large enough time t, i.e., for any large enough iterations, the algorithm solution can be arbitrarily close to the optimal. Nevertheless, to obtain the "convergence rate", we need to study the variance of the trajectory at time t. We notice that such a variance is of order $O(\eta)$ and is vanishing under the limit of $\eta \to 0$. To characterize this variance, we rescale the updates by a factor of $\eta^{-\frac{1}{2}}$. After rescaling, the variance is of order $O(1)$. Specifically, we define $z^{(\eta)} = \eta^{-\frac{1}{2}}w^{(\eta)}$ to highlight the dependence of η.

Lemma 6. *Suppose Assumption 1 holds and initial point is near the optimal point, that is, given pre-specified constants κ and $\delta < \frac{1}{2}$, $\frac{|w_1^{(\eta)}|^2}{||w^{(\eta)}||_2^2} > 1 - \kappa\eta^{\frac{1}{2}-\delta}$. As stepsize $\eta \to 0$, $||w^{(\eta)}(t)||_2 \xrightarrow{t\to\infty} 1$ and i-th component of $z^{(\eta)}$, i.e., $z_i^{(\eta)}$, $i = 2, ..., d$, weakly converges to the following SDE:*

$$dz_i(t) = \left(-\frac{\lambda_1}{\mu_1}\cdot\mu_i z_i + \lambda_i z_i\right)dt + \sqrt{G_i}dB(t), \tag{7}$$

where $B(t)$ is a standard brownian motion, $G_i = \mathbb{E}\left(\left(\widehat{\Lambda}_B\right)_{i,1}\left(\frac{\mu_i}{\mu_1}\right)^{\frac{1}{2}}\widetilde{\Lambda}_{1,1} - \mu_i\Lambda_{i,1}\right)^2$, and $M_{i,j}$ is the i-th row and j-th column entry of a matrix M.

Notice that (7) is a Fokker-Plank equation, whose solution is an Ornstein-Uhlenbeck process as $z_i(t) = z_i(0)\cdot\exp\left[-\left(\frac{\lambda_1}{\mu_1}\mu_i - \lambda_i\right)t\right] + \sqrt{G_i}\int_0^t\exp\left[\left(\frac{\lambda_1}{\mu_1}\mu_i - \lambda_i\right)(s-t)\right]dB(s)$ with the first term on right hand side goes to 0 as time $t \to \infty$. The remaining part is a pure random walk. Thus, the fluctuation of $z_i(t)$ is the error fluctuation of the limiting process after sufficiently many iterations.

Acknowledgment: The authors graciously acknowledge support from the DARPA YFA, Grant N66001-14-1-4047.

References

[1] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

[2] R. Arora, A. Cotter, K. Livescu, and N. Srebro. Stochastic optimization for PCA and PLS. In *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*, pages 861–868. IEEE, 2012.

[3] A. Balsubramani, S. Dasgupta, and Y. Freund. The fast convergence of incremental pca. In *Advances in Neural Information Processing Systems*, pages 3174–3182, 2013.

[4] M. Benzi, G. H. Golub, and J. Liesen. Numerical solution of saddle point problems. *Acta numerica*, 14:1–137, 2005.

[5] C. M. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[6] J. Chen, T. Yang, and S. Zhu. Efficient low-rank stochastic gradient descent methods for solving semidefinite programs. In *Artificial Intelligence and Statistics*, pages 122–130, 2014.

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

[8] R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping from saddle points-online stochastic gradient for tensor decomposition. In *COLT*, pages 797–842, 2015.

[9] G. H. Golub and C. F. Van Loan. *Matrix computations*, volume 3. JHU Press, 2012.

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

[11] M. Hardt and E. Price. The noisy power method: A meta algorithm with applications. In *Advances in Neural Information Processing Systems*, pages 2861–2869, 2014.

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

[13] P. Jain, C. Jin, S. Kakade, and P. Netrapalli. Global convergence of non-convex gradient descent for computing matrix squareroot. In *Artificial Intelligence and Statistics*, pages 479–488, 2017.

[14] P. Jain, C. Jin, S. M. Kakade, P. Netrapalli, and A. 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.

[15] I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

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

[17] K.-C. Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

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

[19] Z. Ma, Y. Lu, and D. Foster. Finding linear structure in large datasets with scalable canonical correlation analysis. In *International Conference on Machine Learning*, pages 169–178, 2015.

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

[21] B. Thompson. *Canonical correlation analysis: Uses and interpretation*. Number 47. Sage, 1984.

[22] M. Welling. Fisher linear discriminant analysis. *Department of Computer Science, University of Toronto*, 3:1–4, 2005.

[23] J. H. Wilkinson and J. H. Wilkinson. *The algebraic eigenvalue problem*, volume 87. Clarendon Press Oxford, 1965.

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