Online ICA: Understanding Global Dynamics of Nonconvex Optimization via Diffusion Processes

Chris Junchi Li **Zhaoran Wang** **Han Liu**
Department of Operations Research and Financial Engineering, Princeton University
{junchil, zhaoran, hanliu}@princeton.edu

Abstract

Solving statistical learning problems often involves nonconvex optimization. Despite the empirical success of nonconvex statistical optimization methods, their global dynamics, especially convergence to the desirable local minima, remain less well understood in theory. In this paper, we propose a new analytic paradigm based on diffusion processes to characterize the global dynamics of nonconvex statistical optimization. As a concrete example, we study stochastic gradient descent (SGD) for the tensor decomposition formulation of independent component analysis. In particular, we cast different phases of SGD into diffusion processes, i.e., solutions to stochastic differential equations. Initialized from an unstable equilibrium, the global dynamics of SGD transit over three consecutive phases: (i) an unstable Ornstein-Uhlenbeck process slowly departing from the initialization, (ii) the solution to an ordinary differential equation, which quickly evolves towards the desirable local minimum, and (iii) a stable Ornstein-Uhlenbeck process oscillating around the desirable local minimum. Our proof techniques are based upon Stroock and Varadhan's weak convergence of Markov chains to diffusion processes, which are of independent interest.

1 Introduction

For solving a broad range of large-scale statistical learning problems, e.g., deep learning, nonconvex optimization methods often exhibit favorable computational and statistical efficiency empirically. However, there is still a lack of theoretical understanding of the global dynamics of these nonconvex optimization methods. In specific, it remains largely unexplored why simple optimization algorithms, e.g., stochastic gradient descent (SGD), often exhibit fast convergence towards local minima with desirable statistical accuracy. In this paper, we aim to develop a new analytic framework to theoretically understand this phenomenon.

The dynamics of nonconvex statistical optimization are of central interest to a recent line of work. Specifically, by exploring the local convexity within the basins of attraction, [1, 5–8, 10–13, 20–22, 24–26, 31, 35, 36, 39, 46–58] establish local fast rates of convergence towards the desirable local minima for a variety statistical problems. Most of these characterizations of local dynamics are based on two decoupled ingredients from statistics and optimization: (i) the local (approximately) convex geometry of the objective functions, which is induced by the underlying statistical models, and (ii) adaptation of classical optimization analysis [19, 34] by incorporating the perturbations induced by nonconvex geometry as well as random noise. To achieve global convergence guarantees, they rely on various problem-specific approaches to obtain initializations that provably fall into the basins of attraction. Meanwhile, for some learning problems, such as phase retrieval and tensor decomposition for latent variable models, it is empirically observed that good initializations within the basins of attraction are not essential to the desirable convergence. However, it remains highly challenging to characterize the global dynamics, especially within the highly nonconvex regions outside the local basins of attraction.

In this paper, we address this problem with a new analytic framework based on diffusion processes. In particular, we focus on the concrete example of SGD applied on the tensor decomposition formula-

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

tion of independent component analysis (ICA). Instead of adapting classical optimization analysis accordingly to local nonconvex geometry, we cast SGD in different phases as diffusion processes, i.e., solutions to stochastic differential equations (SDE), by analyzing the weak convergence from discrete Markov chains to their continuous-time limits [17, 40]. The SDE automatically incorporates the geometry and randomness induced by the statistical model, which allows us to establish the exact dynamics of SGD. In contrast, classical optimization analysis only yields upper bounds on the optimization error, which are unlikely to be tight in the presence of highly nonconvex geometry, especially around the stationary points that have negative curvatures along certain directions. In particular, we identify three consecutive phases of the global dynamics of SGD, which is illustrated in Figure 1.

(i) We consider the most challenging initialization at a stationary point with negative curvatures, which can be cast as an unstable equilibrium of the SDE. Within the first phase, the dynamics of SGD are characterized by an unstable Ornstein-Uhlenbeck process [2, 37], which departs from the initialization at a relatively slow rate and enters the second phase.

(ii) Within the second phase, the dynamics of SGD are characterized by the exact solution to an ordinary differential equation. This solution evolves towards the desirable local minimum at a relatively fast rate until it approaches a small basin around the local minimum.

(iii) Within the third phase, the dynamics of SGD are captured by a stable Ornstein-Uhlenbeck process [2, 37], which oscillates within a small basin around the local minimum.

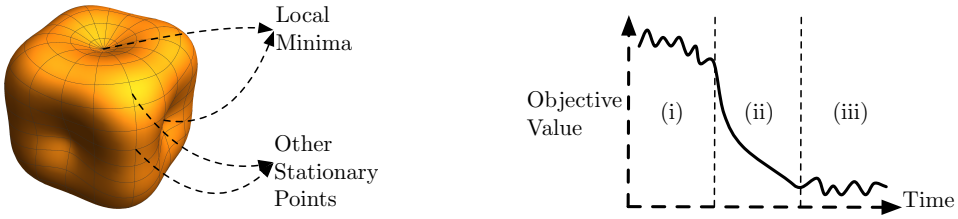

Figure 1: Left: an illustration of the objective function for the tensor decomposition formulation of ICA. Note that here we use the spherical coordinate system and add a global offset of 2 to the objective function for better illustration. Right: An illustration of the three phases of diffusion processes.

More related work. Our results are connected with a very recent line of work [3, 18, 27, 29, 38, 42–45] on the global dynamics of nonconvex statistical optimization. In detail, they characterize the global geometry of nonconvex objective functions, especially around their saddle points or local maxima. Based on the geometry, they prove that specific optimization algorithms, e.g., SGD with artificial noise injection, gradient descent with random initialization, and second-order methods, avoid the saddle points or local maxima, and globally converge to the desirable local minima. Among these results, our results are most related to [18], which considers SGD with noise injection on ICA. Compared with this line of work, our analysis takes a completely different approach based on diffusion processes, which is also related to another line of work [14, 15, 30, 32, 33, 41].

Without characterizing the global geometry, we establish the global exact dynamics of SGD, which illustrate that, even starting from the most challenging stationary point, it may be unnecessary to use additional techniques such as noise injection, random initialization, and second-order information to ensure the desirable convergence. In other words, the unstable Ornstein-Uhlenbeck process within the first phase itself is powerful enough to escape from stationary points with negative curvatures. This phenomenon is not captured by the previous upper bound-based analysis, since previous upper bounds are relatively coarse-grained compared with the exact dynamics, which naturally give a sharp characterization simultaneously from upper and lower bounds. Furthermore, in Section 5 we will show that our sharp diffusion process-based characterization provides understanding on different phases of dynamics of our online/SGD algorithm for ICA.

A recent work [29] analyzes an online principal component analysis algorithm based on the intuition gained from diffusion approximation. In this paper, we consider a different statistical problem with a rigorous characterization of the diffusion approximations in three separate phases.

Our contribution. In summary, we propose a new analytic paradigm based on diffusion processes for characterizing the global dynamics of nonconvex statistical optimization. For SGD on ICA, we identify the aforementioned three phases for the first time. Our analysis is based on Stroock and Varadhan's weak convergence of Markov chains to diffusion processes, which are of independent interest.

2 Background

In this section we formally introduce a special model of *independent component analysis* (ICA) and the associated SGD algorithm. Let $\{\boldsymbol{X}^{(i)}\}_{i=1}^n$ be the data sample identically distributed as $\boldsymbol{X} \in \mathbb{R}^d$. We make assumptions for the distribution of $\boldsymbol{X}$ as follows. Let $\|\cdot\|$ be the ℓ_2-norm of a vector.

Assumption 1. There is an orthonormal matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ such that $\boldsymbol{X} = \mathbf{A}\boldsymbol{Y}$, where $\boldsymbol{Y} \in \mathbb{R}^d$ is a random vector that has independent entries satisfying the following conditions:

(i) The distribution of each Y_i is symmetric about 0;
(ii) There is a constant B such that $\|\boldsymbol{Y}\|^2 \le B$;
(iii) The $Y_1, \dots, Y_d$ are independent with identical m moments for $m \le 8$, denoted by $\psi_m \equiv \mathbb{E}Y_1^m$;
(iv) The $\psi_1 = \mathbb{E}Y_i = 0$, $\psi_2 = \mathbb{E}Y_i^2 = 1$, $\psi \equiv \psi_4 \neq 3$.

Assumption 1(iii) above is a generalization of i.i.d. tensor components. Let $\mathbf{A} = (\mathbf{a}_1, \dots, \mathbf{a}_d)$ whose columns form an orthonormal basis. Our goal is to estimate the orthonormal basis $\mathbf{a}_i$ from online data $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$. We first establish a preliminary lemma.

Lemma 1. Let $\mathbf{T} = \mathbb{E}(\boldsymbol{X}^{\otimes 4})$ be the 4th-order tensor whose (i,j,k,l)-entry is $\mathbb{E}(X_iX_jX_kX_l)$. Under Assumption 1, we have

$$\mathbf{T}(\mathbf{u},\mathbf{u},\mathbf{u},\mathbf{u}) \equiv \mathbb{E}\left(\mathbf{u}^\top \boldsymbol{X}\right)^4 = 3 + (\psi - 3)\sum_{i=1}^d (\mathbf{a}_i^\top \mathbf{u})^4. \tag{2.1}$$

Lemma 1 implies that finding $\mathbf{a}_i$'s can be cast into the solution to the following population optimization problem

$$\operatorname{argmin} -\operatorname{sign}(\psi-3)\cdot \mathbb{E}\left(\mathbf{u}^\top \boldsymbol{X}\right)^4 = \operatorname{argmin} \sum_{i=1}^d -(\mathbf{a}_i^\top \mathbf{u})^4 \quad \text{subject to } \|\mathbf{u}\| = 1. \tag{2.2}$$

It is straightforward to conclude that all stable equilibria of (2.2) are $\pm\mathbf{a}_i$ whose number linearly grows with d. Meanwhile, by analyzing the Hessian matrices the set of unstable equilibria of (2.2) includes (but not limited to) all $\mathbf{v}^* = d^{-1/2}(\pm 1, \cdots, \pm 1)$, whose number grows exponentially as d increases [18, 44].

Now we introduce the SGD algorithm for solving (2.2) with finite samples. Let $\mathcal{S}^{d-1} = \{\mathbf{u} : \|\mathbf{u}\| = 1\}$ be the unit sphere in $\mathbb{R}^d$, and denote $\Pi\mathbf{u} = \mathbf{u}/\|\mathbf{u}\|$ for $\mathbf{u} \neq 0$ the projection operator onto $\mathcal{S}^{d-1}$. With appropriate initialization, the SGD for tensor method iteratively updates the estimator via the following Eq. (2.3):

$$\mathbf{u}^{(n)} = \Pi\left\{\mathbf{u}^{(n-1)} + \operatorname{sign}(\psi-3)\cdot\beta\left(\mathbf{u}^{(n-1)\,\top}\boldsymbol{X}^{(n)}\right)^3 \boldsymbol{X}^{(n)}\right\}. \tag{2.3}$$

The SGD algorithms that performs stochastic approximation using single online data sample in each update has the advantage of less temporal and spatial complexity, especially when d is high [18, 29]. An essential issue of this nonconvex optimization problem is how the algorithm escape from unstable equilibria. [18] provides a method of adding artificial noises to the samples, where the noise variables are uniformly sampled from $\mathcal{S}^{d-1}$. In our work, we demonstrate that under some reasonable distributional assumptions, the online data provide sufficient noise for the algorithm to escape from the unstable equilibria.

By symmetry, our algorithm in Eq. (2.3) converges to a uniformly random tensor component from d components. In order to solve the problem completely, one can repeatedly run the algorithm using different set of online samples until all tensor components are found. In the case where d is high, the well-known coupon collector problem [16] implies that it takes $\approx d\log d$ runs of SGD algorithm to obtain all d tensor components.

Remark. From Eq. (2.2) we see the tensor structure in Eq. (2.1) is unidentifiable in the case of $\psi = 3$, see more discussion in [4, 18]. Therefore in Assumption 1 we rule out the value $\psi = 3$ and call the value $|\psi - 3|$ the *tensor gap*. The reader will see later that, analogous to eigengap in SGD algorithm for principal component analysis (PCA) [29], tensor gap plays a vital role in the time complexity in the algorithm analysis.

3 Markov Processes and Differential Equation Approximation

To work on the approximation we first conclude the following proposition.

Proposition 1. The iteration $\mathbf{u}^{(n)}, n = 0, 1, \ldots$ generated by Eq. (2.3) forms a discrete-time, time-homogeneous Markov process that takes values on $\mathcal{S}^{d-1}$. Furthermore, $\mathbf{u}^{(n)}$ holds strong Markov property.

For convenience of analysis we use the transformed iteration $\mathbf{v}^{(n)} \equiv \mathbf{A}^\top \mathbf{u}^{(n)}$ in the rest of this paper. The update equation in Eq. (2.3) is equivalently written as

$$\begin{aligned} \mathbf{v}^{(n)} = \mathbf{A}^\top \mathbf{u}^{(n)} &= \Pi \left\{ \mathbf{A}^\top \mathbf{u}^{(n-1)} \pm \beta \left(\mathbf{u}^{(n-1)\top} \mathbf{A}\mathbf{A}^\top \boldsymbol{X}^{(n)} \right)^3 \mathbf{A}^\top \boldsymbol{X}^{(n)} \right\} \\ &= \Pi \left\{ \mathbf{v}^{(n-1)} \pm \beta \left(\mathbf{v}^{(n-1)\top} \boldsymbol{Y}^{(n)} \right)^3 \boldsymbol{Y}^{(n)} \right\}. \end{aligned} \tag{3.1}$$

Here $\pm\beta$ has the same sign with $\psi - 3$. It is obvious from Proposition 1 that the (strong) Markov property applies to $\mathbf{v}^{(n)}$, and one can analyze the iterates $\mathbf{v}^{(n)}$ generated by Eq. (3.1) from a perspective of Markov processes.

Our next step is to conclude that as the stepsize $\beta \to 0^+$, the iterates generated by Eq. (2.3), under the time scaling that speeds up the algorithm by a factor β^{-1}, can be globally approximated by the solution to the following ODE system. To characterize such approximation we use theory of weak convergence to diffusions [17, 40] via computing the infinitesimal mean and variance for SGD for the tensor method. We remind the readers of the definition of weak convergence $Z^\beta \Rightarrow Z$ in stochastic processes: for any $0 \le t_1 < t_2 < \cdots < t_n$ the following convergence in distribution occurs as $\beta \to 0^+$

$$\left(Z^\beta(t_1), Z^\beta(t_2), \ldots, Z^\beta(t_n) \right) \xrightarrow{d} \left(Z(t_1), Z(t_2), \ldots, Z(t_n) \right).$$

To highlight the dependence on β we add it in the superscipts of iterates $\mathbf{v}^{\beta,(n)} = \mathbf{v}^{(n)}$. Recall that $\lfloor t\beta^{-1} \rfloor$ is the integer part of the real number $t\beta^{-1}$.

Theorem 1. If for each $k = 1, \ldots, d$, as $\beta \to 0^+$ $v_k^{\beta,(0)}$ converges weakly to some constant scalar V_k^o then the Markov process $v_k^{\beta,(\lfloor t\beta^{-1} \rfloor)}$ converges weakly to the solution of the ODE system

$$\frac{\mathrm{d}V_k}{\mathrm{d}t} = |\psi - 3|\, V_k \left(V_k^2 - \sum_{i=1}^d V_i^4 \right), \qquad k = 1, \ldots, d, \tag{3.2}$$

with initial values $V_k(0) = V_k^o$.

To understand the complex ODE system in Eq. (3.2) we first investigate into the case of $d = 2$. Consider a change of variable $V_1^2(t)$ we have by chain rule in calculus and $V_2^2 = 1 - V_1^2$ the following derivation:

$$\begin{aligned} \frac{\mathrm{d}V_1^2}{\mathrm{d}t} &= 2V_1 \cdot \frac{\mathrm{d}V_1}{\mathrm{d}t} = 2V_1 \cdot |\psi - 3|\, V_1 \left(V_1^2 - V_1^4 - V_2^4 \right) \\ &= 2\,|\psi - 3|\, V_1^2 \left(V_1^2 - V_1^4 - (1 - V_1^2)^2 \right) = -2\,|\psi - 3|\, V_1^2 \left(V_1^2 - \frac{1}{2} \right) (V_1^2 - 1). \end{aligned} \tag{3.3}$$

Eq. (3.3) is an autonomous, first-order ODE for V_1^2. Although this equation is complex, a closed-form solution is available:

$$V_1^2(t) = 0.5 \pm 0.5(1 + C \exp\left(-|\psi - 3|t\right))^{-0.5},$$

and $V_2^2(t) = 1 - V_1^2(t)$, where the choices of $\pm$ and C depend on the initial value. The above solution allows us to conclude that if the initial vector $(V_1^o)^2 < (V_2^o)^2$ (resp. $(V_1^o)^2 > (V_2^o)^2$), then it approaches to 1 (resp. 0) as $t \to \infty$. This intuition can be generalized to the case of higher d that the ODE system in Eq. (3.2) converges to the coordinate direction $\pm\mathbf{e}_k$ if $(V_k^o)^2$ is strictly maximal among $(V_1^o)^2, \ldots, (V_d^o)^2$ in the initial vector. To estimate the time of traverse we establish the following Proposition 2.

Proposition 2. Fix $\delta \in (0, 1/2)$ and the initial value $V_k(0) = V_k^o$ that satisfies $(V_{k_0}^o)^2 \ge 2(V_k^o)^2$ for all $1 \le k \le d, k \ne k_0$, then there is a constant (called traverse time) T that depends only on d, δ such that $V_{k_0}^2(T) \ge 1 - \delta$. Furthermore T has the following upper bound: let $y(t)$ solution to the following auxillary ODE

$$\frac{\mathrm{d}y}{\mathrm{d}t} = y^2 (1 - y), \tag{3.4}$$

with $y(0) = 2/(d+1)$. Let T_0 be the time that $y(T_0) = 1 - \delta$. Then

$$T \le |\psi - 3|^{-1} T_0 \le |\psi - 3|^{-1} \left(d - 3 + 4 \log(2\delta)^{-1} \right). \tag{3.5}$$

Proposition 2 concludes that, by admitting a gap of 2 between the largest $(V_{k_0}^o)^2$ and second largest $(V_k^o)^2$, $k \neq k_0$ the estimate on traverse time can be given, which is tight enough for our purposes in Section 5.

Remark. In an earlier paper [29] which focuses on the SGD algorithm for PCA, when the stepsize is small, the algorithm iteration is approximated by the solution to ODE system after appropriate time rescaling. The approximate ODE system for SGD for PCA is

$$\frac{\mathrm{d}V_k}{\mathrm{d}t} = -2V_k \sum_{i=1}^{d} (\lambda_k - \lambda_i) V_i^2, \qquad k = 1, \ldots, d. \tag{3.6}$$

The analysis there also involves computation of infinitesimal mean and variance for each coordinate as the stepsize $\beta \to 0^+$ and theory of convergence to diffusions [17, 40]. A closed-form solution to Eq. (3.6) is obtained in [29], called the *generalized logistic curves*. In contrast, to our best knowledge a closed-form solution to Eq. (3.2) is generally *not* available.

4 Local Approximation via Stochastic Differential Equation

The ODE approximation in Section 3 is very informative: it characterizes globally the trajectory of our algorithm for ICA or tensor method in Eq. (2.3) with $\mathcal{O}(1)$ approximation errors. However it fails to characterize the behavior near equilibria where the gradients in our ODE system are close to zero. For instance, if the SGD algorithm starts from $\mathbf{v}^*$, on a microscopic magnitude of $\mathcal{O}(\beta^{1/2})$ the noises generated by online samples help escaping from a neighborhood of $\mathbf{v}^*$.

Our main goal in this section is to demonstrate that under appropriate spatial and temporal scalings, the algorithm iteration converges locally to the solution to certain stochastic differential equations (SDE). We provide the SDE approximations in two scenarios, separately near an arbitrary tensor component (Subsection 4.1) which indicates that our SGD for tensor method converges to a local minimum at a desirable rate, and a special local maximum (Subsection 4.2) which implies that the stochastic nature of our SGD algorithm for tensor method helps escaping from unstable equilibria. Note that in the algorithm iterates, the escaping from stationary points occurs first, followed by the ODE and then by the phase of convergence to local minimum. We discuss this further in Section 5.

4.1 Neighborhood of Local Minimizers

To analyze the behavior of SGD for tensor method we first consider the case where the iterates enter a neighborhood of one local minimizer, i.e. the tensor component. Since the tensor decomposition in Eq. (2.2) is full-rank and symmetric, we consider without loss of generality the neighborhood near $\mathbf{e}_1$ the first tensor component. The following Theorem 2 indicates that under appropriate spatial and temporal scalings, the process admits an approximation by Ornstein-Uhlenbeck process. Such approximation is characterized rigorously using weak convergence theory of Markov processes [17, 40]. The readers are referred to [37] for fundamental topics on SDE.

Theorem 2. If for each $k = 2, \ldots, d$, $\beta^{-1/2} v_k^{\beta,(0)}$ converges weakly to $U_k^o \in (0, \infty)$ as $\beta \to 0^+$ then the stochastic process $\beta^{-1/2} v_k^{\beta,(\lfloor t\beta^{-1} \rfloor)}$ converges weakly to the solution of the stochastic differential equation

$$\mathrm{d}U_k(t) = -|\psi - 3| \, U_k(t)\mathrm{d}t + \psi_6^{1/2} \mathrm{d}B_k(t), \tag{4.1}$$

with initial values $U_k(0) = U_k^o$. Here $B_k(t)$ is a standard one-dimensional Brownian motion.

We identify the solution to Eq. (4.1) as an Ornstein-Uhlenbeck process which can be expressed in terms of a Itô integral, with

$$U_k(t) = U_k^o \exp\left(-|\psi - 3|t\right) + \psi_6^{1/2} \int_0^t \exp\left(-|\psi - 3|(t-s)\right) \mathrm{d}B_k(s). \tag{4.2}$$

Itô isometry along with mean-zero property of Itô integral gives

$$\begin{aligned}
\mathbb{E}(U_k(t))^2 &= (U_k^o)^2 \exp\left(-2|\psi - 3|t\right) + \psi_6 \int_0^t \exp\left(-2|\psi - 3|(t-s)\right) \mathrm{d}s \\
&= \frac{\psi_6}{2|\psi - 3|} + \left((U_k^o)^2 - \frac{\psi_6}{2|\psi - 3|} \right) \exp\left(-2|\psi - 3|t\right),
\end{aligned}$$

which, by taking the limit $t \to \infty$, approaches $\psi_6/(2|\psi - 3|)$. From the above analysis we conclude that the Ornstein-Uhlenbeck process has the *mean-reverting* property that its mean decays exponentially towards 0 with persistent fluctuations at equilibrium.

4.2 Escape from Unstable Equilibria

In this subsection we consider SGD for tensor method that starts from a sufficiently small neighborhood of a special unstable equilibrium. We show that after appropriate rescalings of both time and space, the SGD for tensor iteration can be approximated by the solution to a second SDE. Analyzing the approximate SDE suggests that our SGD algorithm iterations can get rid of the unstable equilibria (including local maxima and stationary points with negative curvatures) whereas the traditional gradient descent (GD) method gets stuck. In other words, under weak distributional assumptions the stochastic gradient plays a vital role that helps the escape. As a illustrative example, we consider the special stationary points $\mathbf{v}^* = d^{-1/2}(\pm 1, \ldots, \pm 1)$. Consider a submanifold $\mathcal{S}_F \subseteq \mathcal{S}^{d-1}$ where

$$\mathcal{S}_F = \left\{ \mathbf{v} \in \mathcal{S}^{d-1} : \text{there exists } 1 \le k < k' \le d \text{ such that } v_k^2 = v_{k'}^2 = \max_{1 \le i \le d} v_i^2 \right\}.$$

In words, $\mathcal{S}_F$ consists of all $\mathbf{v} \in \mathcal{S}^{d-1}$ where the maximum of v_k^2 is *not* unique. In the case of $d = 3$, it is illustrated by Figure 1 that $\mathcal{S}_F$ is the frame of a 3-dimenisional box, and hence we call $\mathcal{S}_F$ the *frame*. Let

$$W_{kk'}^{\beta}(t) = \beta^{-1/2} \log\left(v_k^{\beta,(\lfloor t\beta^{-1} \rfloor)}\right)^2 - \beta^{-1/2} \log\left(v_{k'}^{\beta,(\lfloor t\beta^{-1} \rfloor)}\right)^2, \tag{4.3}$$

The reason we study $W_{kk'}^{\beta}(t)$ is that these $d(d-1)$ functions of $\mathbf{v} \in \mathcal{S}^{d-1}$ form a local coordinate map around $\mathbf{v}^*$ and further characterize the distance between $\mathbf{v}$ and $\mathcal{S}_F$ on a spatial scale of $\beta^{1/2}$. We define the positive constant $\Lambda_{d,\psi}$ as

$$\begin{aligned} \Lambda_{d,\psi}^2 = 8d^{-2} \big(& \psi_8 + (16d - 28)\psi_6 + 15d\psi_4^2 \\ & -5(72d^2 - 228d + 175)\psi_4 + 15(2d-7)(d-2)(d-3) \big). \end{aligned} \tag{4.4}$$

We have our second SDE approximation result as follows.

Theorem 3. Let $W_{kk'}^{\beta}(t)$ be defined as in Eq. (4.3), and let $\Lambda_{d,\psi}$ be as in Eq. (4.4). If for each distinct $k, k' = 1, \ldots, d$, $W_{kk'}^{\beta}(0)$ converges weakly to $W_{kk'}^{o} \in (0, \infty)$ as $\beta \to 0^+$ then the stochastic process $W_{kk'}^{\beta}(t)$ converges weakly to the solution of the stochastic differential equation

$$\mathrm{d}W_{kk'}(t) = \frac{2\,|\psi - 3|}{d} W_{kk'}(t)\mathrm{d}t + \Lambda_{d,\psi}\mathrm{d}B_{kk'}(t) \tag{4.5}$$

with initial values $W_{kk'}(0) = W_{kk'}^{o}$. Here $B_{kk'}(t)$ is a standard one-dimensional Brownian motion. We can solve Eq. (4.5) and obtain an unstable Ornstein-Uhlenbeck process as

$$W_{kk'}(t) = \left(W_{kk'}^{o} + \Lambda_{d,\psi} \int_0^t \exp\left(-\frac{2\,|\psi - 3|}{d} s \right) \mathrm{d}B_{kk'}(s) \right) \exp\left(\frac{2\,|\psi - 3|}{d} t \right). \tag{4.6}$$

Let $C_{kk'}$ be defined as

$$C_{kk'} \equiv W_{kk'}^{o} + \Lambda_{d,\psi} \int_0^{\infty} \exp\left(-\frac{4\,|\psi - 3|}{d} s \right) \mathrm{d}B_{kk'}(s). \tag{4.7}$$

We conclude that the following holds.

(i) $C_{kk'}$ is a normal variable with mean $W_{kk'}^{o}$ and variance $d\Lambda_{d,\psi}^2 / (4\,|\psi - 3|)$;

(ii) When t is large $W_{kk'}(t)$ has the following approximation

$$W_{kk'}(t) \approx C_{kk'} \exp\left(\frac{2\,|\psi - 3|}{d} t \right). \tag{4.8}$$

To verify (i) above we have the Itô integral in Eq. (4.6)

$$\mathbb{E}\left(\Lambda_{d,\psi} \int_0^{\infty} \exp\left(-\frac{2\,|\psi - 3|}{d} s \right) \mathrm{d}B_{kk'}(s) \right) = 0,$$

and by using Itô isometry

$$\begin{aligned} \mathbb{E}\left(\Lambda_{d,\psi} \int_0^{\infty} \exp\left(-\frac{2\,|\psi - 3|}{d} s \right) \mathrm{d}B_{kk'}(s) \right)^2 &= \Lambda_{d,\psi}^2 \int_0^t \exp\left(-\frac{4\,|\psi - 3|}{d} s \right) \mathrm{d}s \\ &\approx \Lambda_{d,\psi}^2 \int_0^{\infty} \exp\left(-\frac{4\,|\psi - 3|}{d} s \right) \mathrm{d}s = \frac{d\Lambda_{d,\psi}^2}{4\,|\psi - 3|}. \end{aligned}$$

The analysis above on the unstable Ornstein-Uhlenbeck process indicates that the process has the *momentum* nature that when t is large, it can be regarded as at a normally distributed location centered at 0 and grows exponentially. In Section 5 we will see how the result in Theorem 3 provides explanation on the escape from unstable equilibria.

5 Phase Analysis

In this section, we utilize the weak convergence results in Sections 3 and 4 to understand the dynamics of online ICA in different phases. For purposes of illustration and brevity, we restrict ourselves to the case of starting point $\mathbf{v}^*$, a local maxima that has negative curvatures in every direction. In below we denote by $Z^\beta \asymp W^\beta$ as $\beta \to 0^+$ when the limit of ratio $Z^\beta / W^\beta \to 1$.

Phase I (Escape from unstable equilibria). Assume we start from $\mathbf{v}^*$, then $W^o_{kk'} = 0$ for all $k \neq k'$. We have from Eqs. (4.6) and (4.7) that

$$\log\left(\frac{v_k^{(n)}}{v_{k'}^{(n)}}\right)^2 = \beta^{1/2} W^\beta_{kk'}(n\beta) \approx \left(\beta \frac{d\Lambda^2_{d,\psi}}{4\,|\psi-3|}\right)^{1/2} \chi_{kk'} \exp\left(\frac{2\,|\psi-3|}{d}\cdot \beta n\right). \tag{5.1}$$

Suppose k_1 is the index that maximizes $\left(v_k^{(N_1^\beta)}\right)^2$ and k_2 maximizes $\left(v_k^{(N_1^\beta)}\right)^2, k \neq k_1$. Then by Eq. (5.1) we know $\chi_{k_1 k_2}$ is positive. By setting

$$\log\left(v_{k_1}^{(N_1^\beta)}\right)^2 - \log\left(v_{k_2}^{(N_1^\beta)}\right)^2 = \log 2,$$

we have from the construction in the proof of Theorem 3 that as $\beta \to 0^+$

$$N_1^\beta = \frac{1}{2}\,|\psi-3|^{-1}\, d\beta^{-1} \log\left(\left(\beta \frac{d\Lambda^2_{d,\psi}}{4\,|\psi-3|}\right)^{-1/2} \chi^{-1}_{k_1 k_2} \log 2\right) \asymp \frac{1}{4}\,|\psi-3|^{-1}\, d\beta^{-1} \log\left(\beta^{-1}\right).$$

Phase II (Deterministic traverse). By (strong) Markov property we can restart the counter of iteration, we have the max and second max

$$\left(v_{k_1}^{(0)}\right)^2 = 2\left(v_{k_2}^{(0)}\right)^2,$$

Proposition 2 implies that it takes time

$$T \leq |\psi-3|^{-1}\left(d - 3 + 4\log(2\delta)^{-1}\right),$$

for the ODE to traverse from $V_1^2 = 2/(d+1) = 2V_k^2$ for $k > 1$. Converting to the timescale of the SGD, the second phase has the following relations as $\beta \to 0^+$

$$N_2^\beta \asymp T\beta^{-1} \leq |\psi-3|^{-1}\left(d - 3 + 4\log(2\delta)^{-1}\right)\beta^{-1}.$$

Phase III (Convergence to stable equilibria). Again restart our counter. We have from the approximation in Theorem 3 and Eq. (4.2) that

$$\begin{aligned}\mathbb{E}(v_k^{(n)})^2 &= (v_k^{(0)})^2 \exp\left(-2|\psi-3|\beta n\right) + \beta\psi_6 \int_0^{\beta n} \exp\left(-2|\psi-3|(t-s)\right) \mathrm{d}s \\ &= \frac{\beta\psi_6}{2|\psi-3|} + \left((v_k^{(0)})^2 - \frac{\beta\psi_6}{2|\psi-3|}\right)\exp\left(-2\beta|\psi-3|n\right).\end{aligned}$$

In terms of the iterations $\mathbf{v}^{(n)}$, note the relationship $\mathbb{E}\sin^2\angle(\mathbf{v}, \mathbf{e}_1) = \sum_{k=2}^d v_k^2 = 1 - v_1^2$. The end of ODE phase implies that $\mathbb{E}\sin^2\angle(\mathbf{v}^{(0)}, \mathbf{e}_1) = \delta$, and hence

$$\mathbb{E}\sin^2\angle(\mathbf{v}^{(n)}, \mathbf{e}_1) = \frac{\beta(d-1)\psi_6}{2|\psi-3|} + \left(\delta - \frac{\beta(d-1)\psi_6}{2|\psi-3|}\right)\exp\left(-2\beta|\psi-3|n\right).$$

By setting

$$\mathbb{E}\sin^2\angle(\mathbf{v}^{(N_3^\beta)}, \mathbf{e}_1) = (C_0+1)\cdot\frac{\beta(d-1)\psi_6}{2|\psi-3|},$$

we conclude that as $\beta \to 0^+$

$$N_3^\beta = \frac{1}{2\beta|\psi-3|}\log\left(\beta^{-1}\cdot\frac{2|\psi-3|\delta - \beta(d-1)\psi_6}{C_0(d-1)\psi_6}\right) \asymp \frac{1}{2}|\psi-3|^{-1}\beta^{-1}\log\left(\beta^{-1}\right).$$

6 Summary and discussions

In this paper, we take online ICA as a first step towards understanding the global dynamics of stochastic gradient descent. For general nonconvex optimization problems such as training deep networks, phase-retrieval, dictionary learning and PCA, we expect similar multiple-phase phenomenon. It is believed

that the flavor of asymptotic analysis above can help identify a class of stochastic algorithms for nonconvex optimization with statistical structure.

Our continuous-time analysis also reflects the dynamics of the algorithm in discrete time. This is substantiated by Theorems 1, 2 and 3 which rigorously characterize the convergence of iterates to ODE or SDE by shifting to different temporal and spatial scales. In detail, our results imply when $\beta \to 0^+$:

Phase I takes iteration number $N_1^\beta \asymp (1/4)|\psi - 3|^{-1} d \cdot \beta^{-1} \log(\beta^{-1})$;
Phase II takes iteration number $N_2^\beta \asymp |\psi - 3|^{-1} d \cdot \beta^{-1}$;
Phase III takes iteration number $N_3^\beta \asymp (1/2)|\psi - 3|^{-1} \cdot \beta^{-1} \log(\beta^{-1})$.

After the three phases, the iteration reaches a point that is $C \cdot \left(\psi_6 |\psi - 3|^{-1} \cdot d\beta\right)^{1/2}$ distant on average to one local minimizer. As $\beta \to 0^+$ we have $N_2^\beta / N_1^\beta \to 0$. This implies that the algorithm demonstrates the *cutoff* phenomenon which frequently occur in discrete-time Markov processes [28, Chap. 18]. In words, the Phase II where the objective value in Eq. (2.2) drops from $1 - \varepsilon$ to ε is a short-time phase compared to Phases I and III, so the convergence curve illustrated in the right figure in Figure 1 instead of an exponentially decaying curve. As $\beta \to 0^+$ we have $N_1^\beta / N_3^\beta \asymp d/2$, which suggests that Phase I of escaping from unstable equlibria dominates Phase III by a factor of $d/2$.

References

[1] Agarwal, A., Anandkumar, A., Jain, P. and Netrapalli, P. (2013). Learning sparsely used overcomplete dictionaries via alternating minimization. *arXiv preprint arXiv:1310.7991*.

[2] Aldous, D. (1989). Probability approximations via the Poisson clumping heuristic. *Applied Mathematical Sciences*, **77**.

[3] Anandkumar, A. and Ge, R. (2016). Efficient approaches for escaping higher order saddle points in non-convex optimization. *arXiv preprint arXiv:1602.05908*.

[4] Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M. and Telgarsky, M. (2014). Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, **15** 2773–2832.

[5] Anandkumar, A., Ge, R. and Janzamin, M. (2014). Analyzing tensor power method dynamics in overcomplete regime. *arXiv preprint arXiv:1411.1488*.

[6] Arora, S., Ge, R., Ma, T. and Moitra, A. (2015). Simple, efficient, and neural algorithms for sparse coding. *arXiv preprint arXiv:1503.00778*.

[7] Balakrishnan, S., Wainwright, M. J. and Yu, B. (2014). Statistical guarantees for the EM algorithm: From population to sample-based analysis. *arXiv preprint arXiv:1408.2156*.

[8] Bhojanapalli, S., Kyrillidis, A. and Sanghavi, S. (2015). Dropping convexity for faster semi-definite optimization. *arXiv preprint arXiv:1509.03917*.

[9] Bronshtein, I. N. and Semendyayev, K. A. (1998). *Handbook of mathematics*. Springer.

[10] Cai, T. T., Li, X. and Ma, Z. (2015). Optimal rates of convergence for noisy sparse phase retrieval via thresholded Wirtinger flow. *arXiv preprint arXiv:1506.03382*.

[11] Candès, E., Li, X. and Soltanolkotabi, M. (2014). Phase retrieval via Wirtinger flow: Theory and algorithms. *arXiv preprint arXiv:1407.1065*.

[12] Chen, Y. and Candès, E. (2015). Solving random quadratic systems of equations is nearly as easy as solving linear systems. In *Advances in Neural Information Processing Systems*.

[13] Chen, Y. and Wainwright, M. J. (2015). Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*.

[14] Darken, C. and Moody, J. (1991). Towards faster stochastic gradient search. In *Advances in Neural Information Processing Systems*.

[15] De Sa, C., Olukotun, K. and Ré, C. (2014). Global convergence of stochastic gradient descent for some non-convex matrix problems. *arXiv preprint arXiv:1411.1134*.

[16] Durrett, R. (2010). *Probability: Theory and examples*. Cambridge University Press.

[17] Ethier, S. N. and Kurtz, T. G. (1985). *Markov processes: Characterization and convergence*, vol. 282. John Wiley & Sons.

[18] Ge, R., Huang, F., Jin, C. and Yuan, Y. (2015). Escaping from saddle points — online stochastic gradient for tensor decomposition. *arXiv preprint arXiv:1503.02101*.

[19] Golub, G. H. and Van Loan, C. F. (2012). *Matrix computations*. JHU Press.

[20] Gu, Q., Wang, Z. and Liu, H. (2014). Sparse PCA with oracle property. In *Advances in neural information processing systems*.

[21] Gu, Q., Wang, Z. and Liu, H. (2016). Low-rank and sparse structure pursuit via alternating minimization. In *International Conference on Artificial Intelligence and Statistics*.

[22] Hardt, M. (2014). Understanding alternating minimization for matrix completion. In *Foundations of Computer Science*.

[23] Hirsch, M. W., Smale, S. and Devaney, R. L. (2012). *Differential equations, dynamical systems, and an introduction to chaos*. Academic Press.

[24] Jain, P., Jin, C., Kakade, S. M. and Netrapalli, P. (2015). Computing matrix squareroot via non convex local search. *arXiv preprint arXiv:1507.05854*.

[25] Jain, P. and Netrapalli, P. (2014). Fast exact matrix completion with finite samples. *arXiv preprint arXiv:1411.1087*.

[26] Jain, P., Netrapalli, P. and Sanghavi, S. (2013). Low-rank matrix completion using alternating minimization. In *Symposium on Theory of Computing*.

[27] Lee, J. D., Simchowitz, M., Jordan, M. I. and Recht, B. (2016). Gradient descent converges to minimizers. *arXiv preprint arXiv:1602.04915*.

[28] Levin, D. A., Peres, Y. and Wilmer, E. L. (2009). *Markov chains and mixing times*. American Mathematical Society.

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

[30] Li, Q., Tai, C. et al. (2015). Dynamics of stochastic gradient algorithms. *arXiv preprint arXiv:1511.06251*.

[31] Loh, P.-L. and Wainwright, M. J. (2015). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, **16** 559–616.

[32] Mandt, S., Hoffman, M. D. and Blei, D. M. (2016). A variational analysis of stochastic gradient algorithms. *arXiv preprint arXiv:1602.02666*.

[33] Mobahi, H. (2016). Training recurrent neural networks by diffusion. *arXiv preprint arXiv:1601.04114*.

[34] Nesterov, Y. (2004). *Introductory lectures on convex optimization: A basic course*, vol. 87. Springer.

[35] Netrapalli, P., Jain, P. and Sanghavi, S. (2013). Phase retrieval using alternating minimization. In *Advances in Neural Information Processing Systems*.

[36] Netrapalli, P., Niranjan, U., Sanghavi, S., Anandkumar, A. and Jain, P. (2014). Non-convex robust pca. In *Advances in Neural Information Processing Systems*.

[37] Oksendal, B. (2003). *Stochastic differential equations*. Springer.

[38] Panageas, I. and Piliouras, G. (2016). Gradient descent converges to minimizers: The case of non-isolated critical points. *arXiv preprint arXiv:1605.00405*.

[39] Qu, Q., Sun, J. and Wright, J. (2014). Finding a sparse vector in a subspace: Linear sparsity using alternating directions. In *Advances in Neural Information Processing Systems*.

[40] Stroock, D. W. and Varadhan, S. S. (1979). *Multidimensional diffusion processes*, vol. 233. Springer.

[41] Su, W., Boyd, S. and Candès, E. (2014). A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. In *Advances in Neural Information Processing Systems*.

[42] Sun, J., Qu, Q. and Wright, J. (2015). Complete dictionary recovery over the sphere i: Overview and the geometric picture. *arXiv preprint arXiv:1511.03607*.

[43] Sun, J., Qu, Q. and Wright, J. (2015). Complete dictionary recovery over the sphere ii: Recovery by Riemannian trust-region method. *arXiv preprint arXiv:1511.04777*.

[44] Sun, J., Qu, Q. and Wright, J. (2015). When are nonconvex problems not scary? *arXiv preprint arXiv:1510.06096*.

[45] Sun, J., Qu, Q. and Wright, J. (2016). A geometric analysis of phase retrieval. *arXiv preprint arXiv:1602.06664*.

[46] Sun, R. and Luo, Z.-Q. (2015). Guaranteed matrix completion via nonconvex factorization. In *Foundations of Computer Science*.

[47] Sun, W., Lu, J., Liu, H. and Cheng, G. (2015). Provable sparse tensor decomposition. *arXiv preprint arXiv:1502.01425*.

[48] Sun, W., Wang, Z., Liu, H. and Cheng, G. (2015). Non-convex statistical optimization for sparse tensor graphical model. In *Advances in Neural Information Processing Systems 28*.

[49] Tan, K. M., Wang, Z., Liu, H. and Zhang, T. (2016). Sparse generalized eigenvalue problem: Optimal statistical rates via truncated rayleigh flow. *arXiv preprint arXiv:1604.08697*.

[50] Tu, S., Boczar, R., Soltanolkotabi, M. and Recht, B. (2015). Low-rank solutions of linear matrix equations via procrustes flow. *arXiv preprint arXiv:1507.03566*.

[51] Wang, Z., Gu, Q., Ning, Y. and Liu, H. (2015). High dimensional EM algorithm: Statistical optimization and asymptotic normality. In *Advances in Neural Information Processing Systems*.

[52] Wang, Z., Liu, H. and Zhang, T. (2014). Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *Annals of statistics*, **42** 2164.

[53] Wang, Z., Lu, H. and Liu, H. (2014). Nonconvex statistical optimization: Minimax-optimal sparse PCA in polynomial time. *arXiv preprint arXiv:1408.5352*.

[54] White, C. D., Sanghavi, S. and Ward, R. (2015). The local convexity of solving systems of quadratic equations. *arXiv preprint arXiv:1506.07868*.

[55] Yang, Z., Wang, Z., Liu, H., Eldar, Y. C. and Zhang, T. (2015). Sparse nonlinear regression: Parameter estimation and asymptotic inference under nonconvexity. *arXiv preprint arXiv:1511.04514*.

[56] Zhang, Y., Chen, X., Zhou, D. and Jordan, M. I. (2014). Spectral methods meet em: A provably optimal algorithm for crowdsourcing. In *Advances in neural information processing systems*.

[57] Zhao, T., Wang, Z. and Liu, H. (2015). A nonconvex optimization framework for low rank matrix estimation. In *Advances in Neural Information Processing Systems*.

[58] Zheng, Q. and Lafferty, J. (2015). A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements. *arXiv preprint arXiv:1506.06081*.