COMMUN. MATH. SCI.
Vol. 16, No. 3, pp. 777–789

SEMIGROUPS OF STOCHASTIC GRADIENT DESCENT AND ONLINE PRINCIPAL COMPONENT ANALYSIS: PROPERTIES AND DIFFUSION APPROXIMATIONS*

YUANYUAN FENG†, LEI LI‡, AND JIAN-GUO LIU§

Abstract. We study the Markov semigroups for two important algorithms from machine learning: stochastic gradient descent (SGD) and online principal component analysis (PCA). We investigate the effects of small jumps on the properties of the semigroups. Properties including regularity preserving, L^∞ contraction are discussed. These semigroups are the dual of the semigroups for evolution of probability, while the latter are L^1 contracting and positivity preserving. Using these properties, we show that stochastic differential equations (SDEs) in $\mathbb{R}^d$ (on the sphere $\mathbb{S}^{d-1}$) can be used to approximate SGD (online PCA) weakly. These SDEs may be used to provide some insights of the behaviors of these algorithms.

Keywords. semigroup; Markov chain; stochastic gradient descent; online principle component analysis; stochastic differential equations.

AMS subject classifications. 60J20.

1. Introduction

Stochastic gradient descent (SGD) is a stochastic approximation of the gradient descent optimization method for minimizing an objective function. It is widely used in support vector machines, logistic regression, graphical models and artificial neural networks, which shows amazing performance for large-scale learning due to its computational and statistical efficiency [2–4]. Principal component analysis (PCA) is a dimension reduction method which preserves most of the information in the large data set [8]. Online PCA updates the current PCA each time new data are observed without recomputing it from scratch [10]. SGD and online PCA are both popular algorithms in machine learning. Computational efficiency and convegence behavior in the context of large-scale learning [1, 6] of these two algorithms are studied tremendously.

In this paper, we focus on the properties of discrete semigroups for SGD and online PCA in Section 2 and Section 3 respectively in the small jump regimes. Properties including regularity preserving, L^∞ contraction are discussed. These semigroups are the dual of the semigroups for evolution of probability, while the latter are L^1 contracting and positivity preserving. Based on these properties, we show that SGD and online PCA can be approximated in the weak sense by continuous-time stochastic differential equations (SDEs) in $\mathbb{R}^d$ or on the sphere $\mathbb{S}^{d-1}$ respectively. These will help us understand the discrete algorithms in the viewpoint of diffusion approximation and randomly perturbed dynamical system. Other related works regarding diffusion approximation for SGD can be found in [9, 11], while diffusion approximation using SDEs on the sphere for online PCA seems new.

*Received: September 3, 2017; accepted (in revised form): January 29, 2018. Communicated by Shi Jin.

†Department of Mathematics, Carnegie Mellon University, Pittsburgh, PA 15213, USA (yuanyuaf@andrew.cmu.edu).

‡Department of Mathematics, Duke University, Durham, NC 27708, USA (leili@math.duke.edu).

§Departments of Mathematics and Physics, Duke University, Durham, NC 27708, USA (jliu@phy.duke.edu).

2. The semigroups from SGD

In machine learning, one optimization problem that appears frequently is

$$\min_{x\in\mathbb{R}^d} f(x), \tag{2.1}$$

where $f(x)$ is the loss function associated with a certain training set, and d is the dimension for the parameter x. Usually, the training set is large and one instead considers the stochastic loss functions $f(x;\xi)$ such that

$$f(x)=\mathbb{E}f(x;\xi), \tag{2.2}$$

and $f(x;\xi)$ is often much simpler (e.g. the loss function for a few randomly chosen samples) and thus much easier to handle. Here, $\xi\sim\nu$ is a random vector and ν is some probability distribution. The stochastic gradient descent (SGD) is then to consider

$$x_{n+1}=x_n-\eta\nabla f(x_n;\xi_n), \tag{2.3}$$

where $\eta>0$ is the learning rate and $\xi_n\sim\nu$ are i.i.d so that ξ_n is independent of x_n, with the hope that $\{x_n\}$ can lead to some approximation (if not exact) solution to the optimization problem (2.1). Our goal in this section is to study the time homogeneous Markov chains formed by the SGD (2.3).

REMARK 2.1. To make $\{x_n\}$ a time homogeneous Markov chain, the i.i.d assumption of ξ_n's might be relaxed (for example, one may assume ξ_n to be independent of the past conditioning on x_n and the law of ξ_n depends on the value of x_n only). Though ξ_n's are assumed to be i.i.d, the noises $\epsilon_n:=\eta\nabla f(x_n)-\eta\nabla f(x_n;\xi_n)$ are not i.i.d and their laws can even be different. See the example in Section 2.3 for how ξ_n is involved.

We introduce the following set of smooth functions

$$C_b^m(\mathbb{R}^d)=\left\{f\in C^m(\mathbb{R}^d)\ \middle|\ \|f\|_{C^m}:=\sum_{|\alpha|\le m}|D^\alpha f|_\infty<\infty\right\}. \tag{2.4}$$

2.1. The semigroup and the properties. Let $\mathbb{E}_{x_0}$ denote the expectation under the distribution of this Markov chain starting from x_0 and $\mu^n(\cdot;x_0)$ be the law of x_n. Let $\mu(y,\cdot)$ be the transition probability. Then, for any Borel set E, by the Markov property:

$$\mu^{n+1}(E;x_0)=\int_{\mathbb{R}^d}\mu(y,E)\mu^n(dy;x_0)=\int_{\mathbb{R}^d}\mu^n(E;z)\mu(x_0,dz). \tag{2.5}$$

For a fixed test function $\varphi\in L^\infty(\mathbb{R}^d)$, we define

$$u^n(x_0)=\mathbb{E}_{x_0}\varphi(x_n)=\int_{\mathbb{R}^d}\varphi(y)\mu^n(dy;x_0). \tag{2.6}$$

The Markov property implies that

$$u^{n+1}(x_0)=\mathbb{E}_{x_0}(\mathbb{E}_{x_0}(\varphi(x_{n+1})|x_1))=\mathbb{E}_{x_0}u^n(x_1)=\int_{\mathbb{R}^d}\mu(x_0,dx_1)\int_{\mathbb{R}^d}\varphi(y)\mu^n(dy;x_1), \tag{2.7}$$

which is consistent with (2.5). Given the SGD (2.3), we find explicitly that

$$u^{n+1}(x)=\mathbb{E}(u^n(x-\eta\nabla f(x;\xi)))=:Su^n(x). \tag{2.8}$$

Then, $u^0=\varphi$ and $\{S^n\}_{n\geq 0}$ forms a semigroup for the Markov chain.

For the convenience of discussion, let us introduce $\mathrm{dist}(A,B)$ to mean the distance of two sets $A,B\subset\mathbb{R}^d$:

$$\mathrm{dist}(A,B)=\inf_{x\in A,y\in B}|x-y|.$$

Let X and Y be two Banach spaces. We introduce $\|O\|_{X\to Y}$ to mean the operator norm of a linear operator $O:X\to Y$. We have the following claims regarding the effects of small jumps (small $\eta\nabla f(x,\xi)$) on the properties of semigroups:

THEOREM 2.1. *We fix time $T>0$. Consider SGD (2.3). Let u^n and S be defined by (2.6) and (2.8) respectively. Then:*

(i) (Regularity.) *Suppose that for some $k\in\mathbb{N}$, $\varphi\in C^k(\mathbb{R}^d)$ and $\sup_\xi\|f(\cdot;\xi)\|_{C^{k+1}}<\infty$. Then there exists $\eta_0>0$, such that*

$$\|u^n\|_{C^k}\leq C(k,T,\eta_0)\|\varphi\|_{C^k},\ \forall\ \eta\leq\eta_0,\ n\eta\leq T.$$

(ii) (L^∞ contraction.) $\|S\|_{L^\infty\to L^\infty}\leq 1$.

(iii) (Finite speed.) *If $\mathrm{supp}\,\varphi\subset K$ and $\sup_\xi\|f(\cdot;\xi)\|_{C^1}<\infty$, then for any $n\geq 0$, $n\eta\leq T$, we have that $\mathrm{dist}(\mathrm{supp}\,u^n,K)\leq CT$, where C is a constant depending only on f.*

(iv) (Mass confinement.) *Suppose $\sup_\xi\|f(\cdot;\xi)\|_{C^2}<\infty$ and that there exist $R>0,\delta>0$ such that whenever $|x|\geq R$, $\frac{x}{|x|}\cdot\nabla f(x;\xi)\geq\delta$ for any ξ. If $|x_0|\leq R$, then for any $n\geq 0$, $\eta<\frac{2\delta R}{C^2}$, it holds that*

$$\mathrm{supp}\,\mu^n\subset B(0,R+C\eta).$$

Proof. (i). Since $u^{n+1}(x_0)=\mathbb{E}(u^n(x_0-\eta\nabla f(x_0;\xi)))$, for any $1\leq i\leq d$,

$$\partial_i u^{n+1}(x_0)=\mathbb{E}\left(\sum_{j=1}^d\partial_j u^n(x_0-\eta\nabla f(x_0;\xi)(\delta_{ij}-\eta\partial_i\partial_j f(x_0,\xi))\right).$$

By $\sup_\xi\|f(\cdot;\xi)\|_{C^{k+1}}<\infty$, we have $\|u^{n+1}\|_{C^1}\leq(1+C\eta)\|u^n\|_{C^1}$. Similar calculation reveals that for any k, there exits $\eta_0>0$ such that for any $\eta\leq\eta_0$ there exists $C(k,\eta_0)$ satisfying

$$\|u^{n+1}\|_{C^k}\leq(1+C(k,\eta_0)\eta)\|u^n\|_{C^k}.$$

Since $n\eta\leq T$, we have

$$\|u^n\|_{C^k}\leq(1+C(k,\eta_0)\eta)^n\|\varphi\|_{C^k}\leq e^{C(k,\eta_0)n\eta}\|\varphi\|_{C^k}\leq e^{C(k,\eta_0)T}\|\varphi\|_{C^k}.$$

(ii). That $\|S\|_{L^\infty\to L^\infty}\leq 1$ is clear by (2.8).

(iii). By equation (2.8), $x\in\mathrm{supp}\,u^{n+1}\Leftrightarrow x-\eta\nabla f(x;\xi)\in\mathrm{supp}\,u^n$. Hence, with the assumption $\sup_\xi\|f(\cdot;\xi)\|_{C^1}<\infty$,

$$\mathrm{dist}(\mathrm{supp}\,u^{n+1},\mathrm{supp}\,u^n)\leq C\eta,$$

which implies the claimed result.

(iv). If $|x_n| \le R$, then $|x_{n+1}| \le |x_n| + \eta|\nabla f(x_n, \xi_n)| \le R + C\eta$. If $|x_n| > R$,

$$|x_{n+1}|^2 = |x_n|^2 - 2\eta x_n \cdot \nabla f(x_n;\xi_n) + \eta^2 |\nabla f(x_n;\xi_n)|^2.$$

By assumption, $|x_{n+1}|^2 \le |x_n|^2 - 2\eta\delta|x_n| + C^2\eta^2$. If we take $\eta < \frac{2\delta R}{C^2}$, we obtain that $|x_{n+1}|^2 \le |x_n|^2$. Hence we conclude that $|x_{n+1}| \le R + C\eta$ for any $n \ge 0$. ∎

PROPOSITION 2.1. *Assume* $\sup_\xi \|f(\cdot;\xi)\|_{C^2} < \infty$. *If* η *is sufficiently small, then there exists* $S^* : L^1(\mathbb{R}^d) \to L^1(\mathbb{R})$ *such that* S *is the dual of* S^*, *and* S^* *is given by*

$$S^*\rho = \mathbb{E}\left(\frac{\rho}{|\det(I - \eta\nabla^2 f)|} \circ h(\cdot;\xi) \right), \tag{2.9}$$

where $h(\cdot;\xi)$ *is the inverse mapping of* $x \mapsto x - \eta\nabla f(x;\xi)$ *and* $'\circ'$ *means function composition. Further,* S^* *satisfies:*

(i) $\int_{\mathbb{R}^d} S^*\rho dx = \int_{\mathbb{R}^d} \rho dx$.

(ii) If $\rho \in L^1$ *is nonnegative, then* $S^*\rho \ge 0$ *and* $\|S^*\rho\|_1 = \|\rho\|_1$. S^* *is* L^1 *contraction.*

Proof. From the expression, we see directly that

$$|S^*\rho| \le \mathbb{E}\left| \left(\frac{\rho}{|\det(I - \eta\nabla^2 f)|} \circ h(\cdot;\xi) \right) \right| = \mathbb{E}\left(\frac{|\rho|}{\det(I - \eta\nabla^2 f)} \circ h(\cdot;\xi) \right).$$

This inequality implies that if $\rho \in L^1$, then $S^*\rho \in L^1$.

Take $\rho \in L^1$. Since $Su(x) = \mathbb{E}(u(x - \eta\nabla f(x;\xi)))$ for $u \in L^\infty$, we find that

$$\langle Su(x), \rho \rangle = \int_{\mathbb{R}^d} \mathbb{E}(u(x - \eta\nabla f(x;\xi)))\rho(x)\,dx = \mathbb{E}\int_{\mathbb{R}^d} u(x - \eta\nabla f(x;\xi))\rho(x)\,dx.$$

When η is small, $x - \eta\nabla f(x;\xi)$ is bijective for each ξ because $\|f(\cdot;\xi)\|_{C^2}$ is uniformly bounded. Denote $h(y;\xi)$ the inverse mapping of $y = x - \eta\nabla f(x;\xi)$. It follows that

$$\langle Su(x), \rho \rangle = \mathbb{E}\int_{\mathbb{R}^d} u(y) \frac{\rho}{|\det(I - \eta\nabla^2 f)|} \circ h(y;\xi) dy.$$

Using the fact that $(L^1)' = L^\infty$, we conclude that S is the dual operator of S^*.

(i) Take $u \equiv 1$, we have $Su \equiv 1$. In this case, $\int_{\mathbb{R}^d} S^*\rho dx = \int_{\mathbb{R}^d} \rho Su dx = \int_{\mathbb{R}^d} \rho dx$.

(ii) That $\rho \ge 0$ implies that $S^*\rho \ge 0$ is obvious by the expression. Using Crandall–Tartar lemma [5, Proposition 1], we get $\|S^*\|_{L^1 \to L^1} \le 1$.

REMARK 2.2. We remark that (2.9) is consistent with the first equality in (2.5). To see this, assume μ^n is absolutely continuous to Lebesgue measure so that $\rho^n(\cdot;x_0) = \frac{d\mu^n(\cdot;x_0)}{dx}$. Then, (2.5) implies that

$$\rho^{n+1}(x;x_0) = \mathbb{E}\int_{\mathbb{R}^d} \rho^n(y;x_0)\delta(x - (y - \eta\nabla f(y;\xi)))\,dy = S^*\rho^n(x;x_0).$$

∎

2.2. The diffusion approximation. The discrete semigroups are close to the continuous semigroups generated by certain SDEs in the weak sense. Consider the SDE in Itô sense [14] given by

$$dX = b(X)\,dt + \sqrt{\eta\Sigma}\,dW, \tag{2.10}$$

where b and $\sqrt{\Sigma}$ are assumed to be bounded functions in this paper. We use $e^{tL}\varphi$ to represent the solution to the backward Kolmogrov equation

$$\partial_t u = Lu := b(x)\cdot\nabla u + \frac{1}{2}\eta\Sigma : \nabla^2 u,\ u(\cdot,0) = \varphi.$$

It is well-known that [14]

$$u(x,t) = e^{tL}\varphi = \mathbb{E}_x\varphi(X(t)). \tag{2.11}$$

Similarly, we use $e^{tL^*}\rho_0$ to represent the solution of the forward Kolmogrov equation (Fokker–Planck equation) at time t:

$$\partial_t\rho = L^*\rho := -\nabla\cdot(b\rho) + \frac{1}{2}\eta\sum_{i,j}\partial_{ij}(\Sigma_{ij}\rho),\ \rho(\cdot,0) = \rho_0.$$

Then, $\{e^{tL}\}$ and $\{e^{tL^*}\}$ form two semigroups. If ρ_0 is the initial distribution of $X(t)$, then $e^{tL^*}\rho_0$ is the probability distribution at time t.

LEMMA 2.1.

(i) If $\rho_0 \in L^1(\mathbb{R}^d)$ and $\rho_0 \geq 0$, then $e^{tL^*}\rho_0 \geq 0$ and $\|e^{tL^*}\rho_0\|_1 = \|\rho_0\|_1$. Consequently, e^{tL} is L^1-contraction, i.e.,

$$\|e^{tL^*}\|_{L^1\to L^1} \leq 1,\ \forall t \geq 0.$$

(ii) $e^{tL} : L^\infty(\mathbb{R}^d) \to L^\infty(\mathbb{R}^d)$ is a contraction.

Proof. (i). Since e^{tL^*} is the solution operator to Fokker–Planck equations, it is well-known that it preserves positivity and probability [15], and for general initial data $\rho_0 \in L^1$,

$$\int_{\mathbb{R}^d} e^{tL^*}\rho_0\,dx = \int_{\mathbb{R}^d}\rho_0\,dx.$$

Then, $\forall u, v \in L^1$ and $u \leq v$, it holds $e^{tL^*}u \leq e^{tL^*}v$, and consequently

$$\|e^{tL^*}\|_{L^1\to L^1} \leq 1$$

by Crandall-Tartar lemma [5, Proposition 1].

(ii). Fix $\varphi \in L^\infty$. We take $\rho \in L^1$, and have

$$\langle e^{tL}\varphi, \rho\rangle = \langle \varphi, e^{tL^*}\rho\rangle \leq \|\varphi\|_{L^\infty}\|e^{tL^*}\rho\|_{L^1} \leq \|\varphi\|_{L^\infty}\|\rho\|_{L^1},$$

which yields that $\|e^{tL}\|_{L^\infty\to L^\infty} \leq 1$. ∎

We now show that the discrete semigroups can be approximated by the continuous semigroups in the weak sense (this is the standard terminology in SDE analysis while in functional analysis a more appropriate term might be 'weak-star sense'):

THEOREM 2.2. *Fix time $T>0$. Assume that $\sup_\xi \|f(\cdot,\xi)\|_{C^5}<\infty$. Consider the SDE (2.10) in Itô sense. $u^n(x)$ and $u(x,t)$ are given in (2.6) and (2.11) respectively. If we choose $b(x)=-\nabla f(x)$ and $\Sigma\in C_b^2(\mathbb{R}^d)$ be positive semi-definite, then for all $\varphi\in C_b^4(\mathbb{R}^d)$, there exist $\eta_0>0$ and $C(T,\|\varphi\|_{C^4},\eta_0)>0$ such that*

$$\sup_{n:n\eta\le T}\|u^n-u(\cdot,n\eta)\|_\infty\le C(T,\|\varphi\|_{C^4},\eta_0)\eta,\ \forall\eta\le\eta_0.$$

If instead $\sup_\xi\|f(\cdot,\xi)\|_{C^7}<\infty$ and we choose $b(x)=-\nabla f(x)-\frac{1}{4}\eta\nabla|\nabla f(x)|^2$, $\Sigma=\mathrm{var}(\nabla f(x;\xi))$, then for all $\varphi\in C_b^6(\mathbb{R}^d)$, there exist $\eta_0>0$ and $C(T,\|\varphi\|_{C^6},\eta_0)>0$ such that

$$\sup_{n:n\eta\le T}\|u^n-u(\cdot,n\eta)\|_\infty\le C(T,\|\varphi\|_{C^6},\eta_0)\eta^2,\ \forall\eta\le\eta_0.$$

Proof. First of all, we have

$$u(x,(n+1)\eta)=e^{\eta L}u(x,n\eta),\ \forall n\ge 0. \tag{2.12}$$

Since $\sup_\xi\|f(\cdot,\xi)\|_{C^5}<\infty$, for any $\varphi\in C_b^4(\mathbb{R}^d)$, there exists $\eta_0>0$ such that we have the semigroup expansion by Theorem 2.1,

$$|e^{\eta L}u^n(x)-u^n(x)-\eta Lu^n(x)|\le C(\|u^n\|_{C^4})\eta^2\le C(T,\|\varphi\|_{C^4},\eta_0)\eta^2,\ \forall\eta\le\eta_0.$$

We therefore have

$$\left|e^{\eta L}u^n(x)-u^n(x)-\eta b\cdot\nabla u^n(x)-\frac{1}{2}\eta^2\Sigma:\nabla^2u^n(x)\right|\le C(T,\|\varphi\|_{C^4},\eta_0)\eta^2. \tag{2.13}$$

If $\sup_\xi\|f(\cdot,\xi)\|_{C^7}<\infty$ and we take $\varphi\in C_b^6(\mathbb{R}^d)$, then there exists $\eta_0>0$ and we have for $\eta\le\eta_0$ by Theorem 2.1:

$$|e^{\eta L}u^n(x)-u^n(x)-\eta Lu^n(x)-\frac{1}{2}\eta^2L^2u^n(x)|\le C(\|u^n\|_{C^6},\eta_0)\eta^3\le C(T,\|\varphi\|_{C^6},\eta_0)\eta^3.$$

Therefore, we have

$$\left|e^{\eta L}u^n(x)-u^n(x)-\eta b\cdot\nabla u^n(x)-\frac{1}{2}\eta^2(\Sigma+bb^T):\nabla^2u^n(x)\right.$$
$$\left.-\frac{1}{4}\eta^2\nabla|b|^2\cdot\nabla u^n(x)\right|\le C(T,\|\varphi\|_{C^6},\eta_0)\eta^3. \tag{2.14}$$

On the other hand, we apply Taylor expansion to (2.8),

$$|u^{n+1}(x)-u^n(x)+\eta\nabla f(x)\cdot\nabla u^n(x)-\frac{1}{2}\eta^2\mathbb{E}(\nabla f(x;\xi)\nabla f(x;\xi)^T):\nabla^2u^n(x)|\le C\eta^3. \tag{2.15}$$

If we choose $b(x)=-\nabla f(x)$, (2.13) and (2.15) imply that

$$R^n:=\|u^{n+1}-e^{\eta L}u^n(x)\|_\infty\le\frac{1}{4}\eta^2\|\nabla|b|^2\nabla u^n\|_\infty+C(T,\|\varphi\|_{C^4},\eta_0)\eta^2\le C\eta^2. \tag{2.16}$$

If instead we choose $b(x)=-\nabla f(x)-\frac{1}{4}\eta\nabla|\nabla f(x)|^2$ and $\Sigma=\mathrm{var}(\nabla f(x;\xi))$, then (2.14) and (2.15) imply that

$$R^n:=\|u^{n+1}-e^{\eta L}u^n(x)\|_\infty\le C(T,\|\varphi\|_{C^6},\eta_0)\eta^3. \tag{2.17}$$

We define $E^n := \|u^n - u(\cdot, n\eta)\|_{L^\infty}$. (2.12) and the definition of R^n yield that

$$E^{n+1} \le \left\| e^{\eta L}\Big(u(\cdot,n\eta) - u^n\Big)\right\|_\infty + R^n \le \|u(\cdot,n\eta) - u^n\|_{L^\infty} + R^n = E^n + R^n.$$

The second inequality holds because e^{tL} is L^∞ contraction. The result then follows from $E^n \le \sum_{m=0}^{n-1} R^m$ and $n\eta \le T$. ∎

REMARK 2.3. To get the $O(\eta)$ weak approximation, we can even take $\Sigma = 0$. Indeed, with $\Sigma = 0$, $Z = X(n\eta) - x_n$ is a noise with magnitude $O(\sqrt{\eta})$. The weak order is $O(\eta)$ because $\mathbb{E}Z = O(\eta)$. Choosing $\Sigma = \mathrm{var}(\nabla f(x;\xi))$ (with $b(x) = -\nabla f(x)$) does not improve the weak order from $O(\eta)$ to $O(\eta^2)$ but we believe it characterizes the leading order fluctuation, which is left for future study.

2.3. A specific example. In learning a deep neural network with $N \gg 1$ training samples, the loss function is often given by

$$f(x) = \frac{1}{N}\sum_{k=1}^{N} f_k(x).$$

To train a neural network, the back propagation algorithm is often applied to compute $\nabla f_k(x)$, which is usually not trivial, making computing $\nabla f(x)$ expensive. One strategy is to pick $\xi \in \{1,\dots,N\}$ uniformly so that

$$f(x;\xi) = f_\xi(x).$$

The following SGD is applied

$$x_{n+1} = x_n - \eta \nabla f_\xi(x_n).$$

Such SGD algorithm and its SDE approximation were studied in [11]. The results in Theorem 2.2 can be viewed as a slight generalization of the results in [11], but our proof is performed in a clear way based on the semi-groups. The operators S and Σ for this specific example are given respectively by

$$Su(x) = \frac{1}{N}\sum_{k=1}^{N} u(x - \eta\nabla f_k(x)), \quad \Sigma = \frac{1}{N}\sum_{k=1}^{N} (\nabla f(x) - \nabla f_k(x)) \otimes (\nabla f(x) - \nabla f_k(x)).$$

Now that we have the connection between the SGD and SDEs, the well-known results in SDEs can be borrowed to understand the behaviors of SGD. For example, this gives us the intuition how the batch size in a general SGD can help to escape from sharp minimizers and saddle points of the loss function by affecting the diffusion (see [9]).

3. The semigroups from online PCA

Assume some data have d coordinates and they can be represented by points in $\mathbb{R}^d$. Let $\xi \in \mathbb{R}^d$ be a random vector sampled from the distribution of the data with mean centered to zero (if not, we use $\xi - \mathbb{E}\xi$) and the covariance matrix to be

$$\Sigma = \mathbb{E}(\xi\xi^T).$$

Principal component analysis (PCA) is a procedure to find k $(k < d)$ linear combinations of these coordinates: $w_1^T\xi, \dots, w_k^T\xi$ such that for all $i = 1, 2, \dots, k$, the optimization problem is solved

$$\max \mathbb{E}(w_i^T\xi)^2, \text{ subject to } w_i^T w_j = \delta_{ij},\ j \le i. \tag{3.1}$$

It is well-known that $w_1,\dots,w_k$ are the eigenvectors of Σ corresponding to the first k eigenvalues.

In the online PCA procedure, an adaptive system receives a stream of data $\xi(n)\in\mathbb{R}^d$ and tries to compute the estimates of $w_1,\dots,w_k$ [12, 13]. The algorithm in [12] in the case $k=1$ can be summarized as

$$w^n=\mathbb{Q}\left(w^{n-1}+\eta\nabla f(w^{n-1};\xi(n))\right)=\frac{w^{n-1}+\eta(n)\xi(n)\xi(n)^Tw^{n-1}}{|w^{n-1}+\eta(n)\xi(n)\xi(n)^Tw^{n-1}|},\tag{3.2}$$

where $f(w;\xi(n))=(w^T\xi(n))^2$ and

$$\mathbb{Q}v:=v/|v|.$$

This algorithm is also called the stochastic gradient ascent (SGA) algorithm.

Suppose we choose $\eta(n)=\eta$ to be a constant and take

$$\xi(n)\sim\nu,\ i.i.d.,$$

where ν is some probability distribution in $\mathbb{R}^d$. Then, $\{w^n\}$ forms a time-homogeneous Markov chain on $\mathbb{S}^{d-1}$. Our goal in this section is to study this Markov chain and its semigroups.

For the convenience of following discussion, we assume:

ASSUMPTION 3.1. *There exists $C>0$ such that $\forall\xi\sim\nu$,*

$$|\xi|\le C.$$

3.1. The semigroups and properties. Similarly as in Section 2, fix a test function $\varphi\in L^\infty(\mathbb{S}^{d-1})$, and define

$$u^n(w^0)=\mathbb{E}_{w^0}\varphi(w^n)=\int_{\mathbb{S}^{d-1}}\varphi(y)\mu^n(dS;w^0).\tag{3.3}$$

Again by the Markov property, for the SGA algorithm, we have

$$u^{n+1}(w)=\mathbb{E}u^n(\mathbb{Q}(w+\eta\xi\xi^Tw))=:Su^n(w),\tag{3.4}$$

with $u^0(w)=\varphi(w)$. Clearly, $\{S^n\}_{n\ge0}$ form a semigroup.

For discussing the dynamics on sphere, we find it is convenient to extend w into a neighborhood of the sphere as

$$w=\frac{x}{|x|},\ x\in\mathbb{R}^d,$$

and introduce the projection:

$$\mathbb{P}:=I-w\otimes w.\tag{3.5}$$

This extension allows us to perform the computation on sphere by performing computation in $\mathbb{R}^d$. For example, if $\psi\in C^1(\mathbb{S}^{d-1})$, then we extend ψ into a neighborhood as well and the gradient operator on the sphere can be written in terms of this extension as:

$$\nabla_S\psi(w)=\mathbb{P}\nabla\psi(x)|_{x=w}=(I-w\otimes w)\cdot\nabla\psi(w),\ w\in\mathbb{S}^{d-1}.\tag{3.6}$$

Clearly, $\nabla_S\psi$ only depends on the values of ψ on $\mathbb{S}^{d-1}$, not on the extension. The extension is introduced for the convenience of computation. We use $C^k(\mathbb{S}^{d-1})$ to denote the space of functions that are k-th order continuously differentiable on the sphere with respect to ∇_S, with the norm:

$$\|\psi\|_{C^k(\mathbb{S}^{d-1})} := \sum_{|\alpha|\le k} \sup_{w\in\mathbb{S}^{d-1}} |\nabla_S^\alpha \psi(w)|. \tag{3.7}$$

Here $\alpha=(\alpha_1,\dots,\alpha_m)$, $m\le k$ is a multi-index so that for a vector $v=(v^1,v^2,\dots,v^d)$, $v^\alpha=\prod_{j=1}^m v^{\alpha_j}$.

Now we study the semigroups for the online PCA:

PROPOSITION 3.1.

(i) $S: L^\infty(\mathbb{S}^{d-1})\to L^\infty(\mathbb{S}^{d-1})$ *is a contraction.*

(ii) Fix time $T>0$. *For any* $\varphi\in C^k(\mathbb{S}^{d-1})$, *there exists* $\eta_0>0$ *and* $C(k,\eta_0,T)$ *such that*

$$\|u^n\|_{C^k(\mathbb{S}^{d-1})}\le C(k,T)\|\varphi\|_{C^k(\mathbb{S}^{d-1})},\ \forall \eta\le\eta_0,\ n\eta\le T.$$

(iii) There exists $S^*: L^1(\mathbb{S}^{d-1})\to L^1(\mathbb{S}^{d-1})$ *such that* S *is the dual of* S^*. S^* *is a contraction on* L^1. *Further, for any* $\rho\in L^1$ *and* $\rho\ge 0$, *we have* $S^*\rho\ge 0$ *and* $\|S^*\rho\|_1=\|\rho\|_1$.

Proof. (i). That S is an L^∞ contraction is clear from (3.4).

(ii). For each k, there exists $\eta_0>0$ and $C(k,T,\eta_0)>0$ such that

$$\|\mathbb{Q}(w+\eta\xi\xi^T w)\|_{C^k(\mathbb{S}^{d-1};\mathbb{S}^{d-1})}\le 1+C\eta,\ \eta\le\eta_0.$$

By (3.4) and chain rule for derivatives on sphere, we have

$$\|u^{n+1}\|_{C^k(\mathbb{S}^{d-1})}\le(1+C_1\eta)\|u^n\|_{C^k(\mathbb{S}^{d-1})},\ \eta\le\eta_0, \tag{3.8}$$

and the second claim follows.

(iii). Similar to Proposition 2.1, we find that S^* is given by

$$S^*\rho(v)=\mathbb{E}(\rho\big(h(v;\xi)\big)|J_h(v;\xi))|),$$

where $h(v;\xi)=\frac{(I+\eta\xi\xi^T)^{-1}v}{|(I+\eta\xi\xi^T)^{-1}v|}$ is the inverse mapping of $\mathbb{Q}(w+\eta\xi\xi^T w)$, and $|J_h|$ accounts for the volume change on $\mathbb{S}^{d-1}$ under h. The properties are then similarly proved as in Proposition 2.1. ∎

3.2. The diffusion approximation. Now, we move onto the SDE approximation to the online PCA, i.e. seeking a semigroup generated by diffusion processes on sphere to approximate the discrete semigroups.

Before the discussion, let's consider the matrix $\nabla_S^2 u$

$$\nabla_S^2 u=\nabla_S(\nabla_S u). \tag{3.9}$$

Note that $\nabla_S^2 u$ is not the Hessian of u on the sphere as a Riemannian sub-manifold. The Hessian has the matrix representation in $\mathbb{R}^d$ as

$$H(u)=\nabla_S^2 u\cdot\mathbb{P}. \tag{3.10}$$

Recall that for a vector field $\phi \in C^2(\mathbb{S}^{d-1};\mathbb{R}^d)$, the divergence of ϕ is defined as

$$\int_{\mathbb{S}^{d-1}} \operatorname{div}(\phi)\varphi dS = -\int_{\mathbb{S}^{d-1}} \phi \cdot \nabla_S \varphi dS.$$

Again we extend the vector field into a neighborhood of the sphere so that we can use formulas in $\mathbb{R}^d$ to compute. Using the formula $\int_{\mathbb{S}^{d-1}} \nabla_S \cdot \phi dS = \int_{\mathbb{S}^{d-1}} (\nabla \cdot w) w \cdot \phi dS = (d-1)\int_{\mathbb{S}^{d-1}} w \cdot \phi dS$, one can derive that

$$\operatorname{div}(\phi) = \nabla_S \cdot (\mathbb{P}\phi). \tag{3.11}$$

Indeed, the covariant derivative of $\mathbb{P}\phi$ along a tangent vector Y is $\mathbb{P}(Y \cdot \nabla(\mathbb{P}\phi))$ and the divergence of $\mathbb{P}\phi$ is $\operatorname{tr}(\nabla_S(\mathbb{P}\phi)\cdot\mathbb{P}) = \nabla_S \cdot (\mathbb{P}\phi)$. It follows that

$$\operatorname{tr}(\nabla_S^2 u) = \nabla_S \cdot (\nabla_S u) = \operatorname{div}(\nabla_S u) = \operatorname{tr}(H(u)) =: \Delta_S u, \tag{3.12}$$

because $\mathbb{P}\nabla_S u = \nabla_S u$. Δ_S is called the Laplace–Beltrami operator on $\mathbb{S}^{d-1}$.

Next we give the Taylor expansion and the proof can be found in Appendix A.

LEMMA 3.1. *Let $w \in \mathbb{S}^{d-1}$, and $v \in \mathbb{R}^d$. Suppose $u \in C^3(\mathbb{S}^{d-1})$, then we have*

$$u\left(\frac{w+\eta v}{|w+\eta v|}\right) = u(w) + \eta v \cdot \nabla_S u + \frac{1}{2}\eta^2 (vv : H(u) - 2w \cdot v \otimes v \cdot \nabla_S u) + R(\eta)\eta^3,$$

where $R(\eta)$ is a bounded function.

Now, we introduce $f(w;\xi) = \frac{1}{2} w^T \xi \xi^T w$ and thus

$$f(w) = \mathbb{E} f(w;\xi) = \frac{1}{2} w^T \Sigma w. \tag{3.13}$$

Define the following second moment:

$$M(w) := \mathbb{E}\nabla f(w;\xi) \nabla f(w,\xi)^T = w \otimes w : \mathbb{E}(\xi \otimes \xi \otimes \xi \otimes \xi). \tag{3.14}$$

Recall (3.4). By Lemma 3.1, we then have:

$$u^{n+1}(w) = u^n(w) + \eta \nabla_S f(w) \cdot \nabla_S u^n + \frac{1}{2}\eta^2 (M(w) : H(u^n) - 2w \cdot M(w) \cdot \nabla_S u^n) + R(\eta)\eta^3. \tag{3.15}$$

We now construct SDEs on $\mathbb{S}^{d-1}$ whose semigroups approximate the semigroup $\{S^n\}_{n\geq 0}$ based on (3.15). Consider the following general SDE in Stratonovich sense in $\mathbb{R}^d$:

$$dX = \mathbb{P}b(X)\,dt + \mathbb{P}\sigma(X) \circ dW, \tag{3.16}$$

where W is the standard Wiener process (Brownian motion) in $\mathbb{R}^d$ while $'\circ'$ here represents Stratonovich stochastic integrals, convenient for SDEs on manifold [7].

LEMMA 3.2. *X stays on $\mathbb{S}^{d-1}$ if $X(0) \in \mathbb{S}^{d-1}$. In other words, $|X(t)| = 1$.*

To see this, we only need to apply Itô's formula (for Stratonovich integrals) with the test function $f(x) = |x|^2$ and noting $\mathbb{P}\nabla f = \nabla_S f = 0$.

Let $\varphi \in C^3(\mathbb{S}^{d-1})$ and we extend it to the ambient space of $\mathbb{S}^{d-1}$. Consider

$$u(x,t) = \mathbb{E}_x \varphi(X). \tag{3.17}$$

We find using Itô's formula (for Stratonovich integrals) that (the explicit expression of ∇_S^2 in Appendix A is needed to derive this)

$$\partial_t u = (\mathbb{P}b + \mathbb{P}b_1(\sigma)) \cdot \nabla_S u + \frac{1}{2} \mathbb{P}\sigma\sigma^T \mathbb{P} : H(u) =: L_S u, \tag{3.18}$$

where

$$(\mathbb{P}b_1(\sigma))_i = \frac{1}{2} (\mathbb{P}\sigma)_{kj} (\mathbb{P}\partial_k \sigma)_{ij}. \tag{3.19}$$

The operator L_S is elliptic on sphere.

REMARK 3.1. Clearly, if we take $\sigma = I$, then we have

$$\partial_t u = \frac{1}{2} \Delta_S u.$$

This means that $dX = P \circ dW$ gives the spherical Brownian motion, which is the Stroock's representation of spherical Brownian motion [7, Section 3.3].

With (3.15) and (3.18), we conclude that the discrete semigroups for online PCA can be approximated in the weak sense by the stochastic processes on the sphere:

THEOREM 3.1. *Assume Assumption 3.1. Consider the SDE (3.16) with* $\sigma =: \sqrt{\eta} S$*:*

$$dw = \mathbb{P}b(w)\, dt + \sqrt{\eta} \mathbb{P} S(w) \circ dW. \tag{3.20}$$

u^n and $u(w,t)$ are given as in (3.3) and (3.17) respectively. If we take $\mathbb{P}b(w) = \nabla_S f(w) = (I - w \otimes w) \cdot \Sigma w$ and $S(\cdot) \in C_b^2(\mathbb{S}^{d-1})$ to be positive semi-definite, then for all $\varphi \in C_b^4(\mathbb{S}^{d-1})$, there exist $\eta_0 > 0$ and $C(T, \|\varphi\|_{C^4}, \eta_0) > 0$ such that $\forall \eta \le \eta_0$,

$$\sup_{n:n\eta \le T} \|u^n - u(\cdot, n\eta)\|_{L^\infty(\mathbb{S}^{d-1})} \le C(T, \|\varphi\|_{C^4}, \eta_0)\eta.$$

If instead we take

$$S(w) = \sqrt{\mathrm{var}(\nabla f(w;\xi))} = \sqrt{M(w) - \nabla_S f(w) \nabla_S f(w)^T},$$
$$\mathbb{P}b(w) = \nabla_S f(w) - \eta \mathbb{P}(S(w)^2 \cdot w) - \eta \mathbb{P} b_1(S)(w) - \frac{1}{2} \eta \nabla_S f(w) \cdot H(f)(w).$$

where $Pb_1(\cdot)$ is given in (3.19), then for all $\varphi \in C_b^6(\mathbb{S}^{d-1})$, there exist $\eta_0 > 0$ and $C(T, \|\varphi\|_{C^6}, \eta_0) > 0$ such that $\forall \eta \le \eta_0$,

$$\sup_{n:n\eta \le T} \|u^n - u(\cdot, n\eta)\|_{L^\infty(\mathbb{S}^{d-1})} \le C(T, \|\varphi\|_{C^6}, \eta_0)\eta^2.$$

The proof is similar as that for Theorem 2.2 and we omit it here for brevity.

Acknowledgements. The work of J.-G Liu is partially supported by KI-Net NSF RNMS11-07444, NSF DMS-1514826, and NSF DMS-1812573. Y. Feng is supported by NSF DMS-1252912.

Appendix A. Proof of Lemma 3.1.

Proof. **(Proof of Lemma 3.1.)** We set

$$g(\eta) = u\left(\frac{w+\eta v}{|w+\eta v|}\right).$$

By the fact that $|w| = 1$ and direct computation, we have

$$g'(0) = \nabla u \cdot (I - w \otimes w) \cdot v = v \cdot \nabla_S u,$$

and that

$$g''(0) = (v \cdot (I - ww)) \cdot \nabla^2 u \cdot ((I - ww) \cdot v) + \nabla u \cdot (-2v(w \cdot v) - w(v \cdot v) + 3w(w \cdot v)^2).$$

Since $w = x/|x|$, we have for $x = w \in \mathbb{S}^{d-1}$ that

$$\nabla w = I - w \otimes w.$$

It follows that

$$\nabla_S^2 u(w) = (I - w \otimes w) \cdot \nabla^2 u(w) \cdot (I - w \otimes w) - [(w \cdot \nabla u)(I - w \otimes w) + ((I - w \otimes w) \cdot \nabla u) \otimes w].$$

Hence, we find that

$$g''(0) = v \otimes v : \nabla_S^2 u + \nabla u \cdot (w(w \cdot v)^2 - v(w \cdot v)) = v \otimes v : H(u) - 2w \cdot v \otimes v \cdot \nabla_S u.$$

Moreover, $g'''(\eta)$ is bounded by the assumption. Hence, the claim follows from Taylor expansion. ∎

REFERENCES

[1] Z. Allen-Zhu and Y. Li, *First efficient convergence for streaming k-pca: A global, gap-free, and near-optimal rate*, 2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS), Berkeley, California, USA, 487–492:2017.

[2] L. Bottou, *Large-scale machine learning with stochastic gradient descent*, in Proceedings of COMPSTAT'2010, Springer, 177–186, 2010.

[3] L. Bottou, F.E. Curtis, and J. Nocedal, *Optimization methods for large-scale machine learning*, SIAM Rev., 60(2):223C311, 2018.

[4] S. Bubeck, *Convex optimization: Algorithms and complexity*, Foundations and Trends® in Machine Learning, 8(3-4):231–357, 2015.

[5] M.G. Crandall and L. Tartar, *Some relations between nonexpansive and order preserving mappings*, Proceedings of the American Mathematical Society, 78(3):385–390, 1980.

[6] S. Ghadimi and G. Lan, *Stochastic first-and zeroth-order methods for nonconvex stochastic programming*, SIAM Journal on Optimization, 23(4):2341–2368, 2013.

[7] E.P. Hsu, *Stochastic Analysis on Manifolds*, American Math. Soc., 38, 2002.

[8] J. Josse, J. Pagès, and F. Husson, *Multiple imputation in principal component analysis*, Advances in Data Analysis and Classification, 5(3):231–246, 2011.

[9] W. Hu, C. J. Li, L. Li, and J. Liu, *On the diffusion approximation of nonconvex stochastic gradient descent*, Annals of Mathematical Sciences and applications, to appear. arXiv:1705.07562, 2017.

[10] C.J. Li, M. Wang, H. Liu, and T. Zhang, *Near-optimal stochastic approximation for online principal component estimation*, Mathematical Programming, 167(1):75–97, 2018.

[11] Q. Li, C. Tai, and W.E., *Stochastic modified equations and adaptive stochastic gradient algorithms*, International Conference on Machine Learning, 2101–2110, 2017.

[12] E. Oja, *Simplified neuron model as a principal component analyzer*, J. Math. Biology, 15(3):267–273, 1982.
[13] E. Oja, *Principal components, minor components, and linear neural networks*, Neural Networks, 5(6):927–935, 1992.
[14] B. Øksendal, *Stochastic Differential Equations: An Introduction with Applications*, Springer, Berlin, Heidelberg, Sixth Edition, 2003.
[15] C. Soize, *The Fokker–Planck equation for stochastic dynamical systems and its explicit steady state solutions*, World Scientific, 17, 1994.