STOCHASTIC SYSTEMS
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 1946-5238 | EISSN 1946-5238 | 00 | 0000 | 0001

INFORMS
DOI 10.1287/xxxx.0000.0000

Online Companion to "A Diffusion Approximation Theory of Momentum SGD in Nonconvex Optimization"

Tianyi Liu, Zhehui Chen, Enlu Zhou, Tuo Zhao
School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30318, tianyiliu@gatech.edu, zchen451@gatech.edu, enlu.zhou@isye.gatech.edu, tourzhao@gatech.edu

1. Summary on Weak Convergence and Main Theorems

Here, we summarize the theory of weak convergence and theorems used in this paper. Recall that the continuous-time interpolation of the solution trajectory $V^\eta(\cdot)$ is defined as $V^\eta(t) = v_k^\eta$ on the time interval $[k\eta, k\eta+\eta)$. It has sample paths in the space of Càdlàg functions (right continuous and have left-hand limits) defined on $\mathbb{R}^d$, or *Skorokhod Space*, denoted by $D^d[0,\infty)$. Thus, the weak convergence we consider here is defined in this space $D^d[0,\infty)$ instead of $\mathbb{R}^d$. The special metric σ in $D^d[0,\infty)$ is called Skorokhod metric, and the topology generated by this metric is Skorokhod topology. Please refer to Sagitov (2013), Kushner and Yin (2003) for detailed explanations. The weak convergence in D^d is defined as follows:

Definition 1 (Weak Convergence in $D^d[0,\infty)$). Let $\mathcal{B}$ be the minimal σ-field induced by Skorokhod topology. Let $\{X_n, n<\infty\}$ and X be random variables on $D^d[0,\infty)$ defined on a probability space $(\Omega, P, \mathcal{F})$. Suppose that P_n and P_X are the probability measures on $(D^d, \mathcal{B})$ generated by X_n and X. We say P_n converges weakly to P ($P_n \Rightarrow P$), if for all bounded and continuous real-valued functions F on D^d, the following condition holds:

$$\mathbb{E}F(X_n) = \int F(x)dP_n(x) \to \mathbb{E}F(X) = \int F(x)dP(x). \tag{1}$$

With an abuse of terminology, we say X_n converges weakly to X and write $X_n \Rightarrow X$.

Another important definition we need is *tightness*:

DEFINITION 2. A set of D^d-valued random variables $\{X_n\}$ is said to be tight if for each $\delta > 0$, there is a compact set $B_\delta \in D^d$ such that:

$$\sup_n P\{X_n \notin B_\delta\} \leq \delta. \tag{2}$$

We care about tightness because it provides us a powerful way to prove weak convergence based on the following two theorems:

THEOREM 1 **(Prokhorov's Theorem)**. *Under Skorokhod topology, $\{X_n(\cdot)\}$ is tight in $D^d[0,\infty)$ if and only if it is relative compact which means each subsequence contains a further subsequence that converges weakly.*

THEOREM 2 **(Sagitov (2013), Theorem 3.8)**. *A necessary and sufficient condition for $P_n \Rightarrow P$ is each subsequence $P_{n'}$ contains a further subsequence $P_{n''}$ converging weakly to P.*

Thus, if we can prove $\{X_n(\cdot)\}$ is tight and all the further subsequences share the same weak limit X, then we have X_n converges weakly to X. However, (2) is hard to verified. We usually check another easier criteria. We first define the càdlàg modulus to characterize the discontinuity of any $f \in D^d[0,\infty]$.

DEFINITION 3 (NOWAKOWSKI (2013), DEFINITION 2.7). For $f \in D^d[0,\infty]$, $T > 0$ and $\epsilon > 0$, the modulus of continuity is defined by

$$\varpi'_T(f,\epsilon) := \inf_{\Pi_{T,\epsilon}} \max_{1\leq i\leq k} w(f,[t_{i-1},t_i)),$$

where $\Pi_{T,\epsilon} = \{0 = t_0 \leq t_1 \leq \cdots \leq t_k = T, \min_{1\leq i\leq k} t_i - t_{i-1} > \epsilon\}$ and

$$w(f,[t_{i-1},t_i)) := \sup_{s,t\in[t_{i-1},t_i)} |f(s) - f(t)|.$$

Next theorem provides an sufficient and necessary condition for the tightness of sequence X_n in $D^d[0,\infty)$.

THEOREM 3 **(Nowakowski (2013), Theorem 2.4)**. *Let $\{X_n(\cdot)\}$ be a sequence of processes that have paths in $D^d[0,\infty)$. Then $\{X_n(\cdot)\}$ is tight if and only if*
(i). For every $T > 0$, $\delta > 0$, there exists $n_0 > 0$ and $C > 0$ such that

$$\mathbb{P}\left(\sup_{t\in[0,T]} X_n(t) > C\right) \leq \delta, \ \forall n \geq n_0.$$

(ii). For every $T > 0$, $\delta > 0$, $\gamma > 0$, there exists $n_0 > 0$ and ϵ such that

$$\mathbb{P}\left(\varpi'_T(X_n,\epsilon) \geq \gamma\right) \leq \delta, \ \forall n \geq n_0.$$

Theorem 4 provides one sufficient condition for tightness. Let $\mathcal{F}_t^n$ be the σ-algebra generated by $\{X_n(s), s \le t\}$, and τ denotes a $\mathcal{F}_t^n$-stopping time.

THEOREM 4 **(Kushner and Yin (2003), Theorem 3.3, Chapter 7)**. *Let $\{X_n(\cdot)\}$ be a sequence of processes that have paths in $D^d[0,\infty)$. Suppose that for each $\delta > 0$ and each t in a dense set in $[0,\infty)$, there is a compact set $K_{\delta,t}$ in $\mathbb{R}$ such that*

$$\inf_n P\{X_n(t) \in K_{\delta,t}\} \ge 1 - \delta, \tag{3}$$

and for each positive T,

$$\lim_{\delta} \limsup_n \sup_{|\tau| \le T} \sup_{s \le \delta} \mathbb{E} \min[\|X_n(\tau + s) - X_n(\tau)\|, 1] = 0. \tag{4}$$

Then $\{X_n(\cdot)\}$ is tight in $D^d[0,\infty)$.

This theorem is used in Section 3 to prove tightness of the trajectory of Momentum SGD.

At last, we provide the theorem we use to prove the SDE approximation. Let's consider the following algorithm:

$$\theta_{n+1}^\eta = \theta_n^\eta + \eta Y_n^\eta, \tag{5}$$

where $Y_n^\eta = g_n^\eta(\theta_n^\eta, \xi_n^\eta) + M_n^\eta$, and M_n^η is a martingale difference sequence. Then the normalized process $U_n^\eta = (\theta_n^\eta - \bar{\theta})/\sqrt{\eta}$ satisfies:

$$U_{n+1}^\eta = U_n^\eta + \sqrt{\eta}(g_n^\eta(\theta_n^\eta, \xi_n^\eta) + M_n^\eta). \tag{6}$$

We further assume the fixed-state-chain exists as in Subsection 4.1 and use the same notation $\xi_i(\theta)$ to denote the fixed-θ-process. Then we have the following theorem:

THEOREM 5 **(Kushner and Yin (2003), Theorem 8.1, Chapter 10)**. *Assume the following conditions hold:*

C.1 For small $\rho > 0$, $\{|Y_n^\eta|^2 I_{|\theta_n^\eta - \bar{\theta}| \le \rho}\}$ is uniformly integrable.

C.2 There is a continuous function $\bar{g}(\cdot)$ such that for any sequence of integers $n_\eta \to 0$ satisfying $n_\eta \eta \to 0$ as $\eta \to 0$ and each compact set A,

$$\frac{1}{n_\eta} \sum_{i=jn_\eta}^{jn_\eta + n_\eta - 1} E_{jn_\eta}^\eta [g_i^\eta(\theta, \xi_i(\theta)) - \bar{g}(\theta)] I_{\{\xi_{jn_\eta}^\eta\}} \to 0,$$

in mean for each θ, as $j \to \infty$ and $\eta \to 0$.

C.3 Define

$$\Gamma_n^\eta(\theta) = \sum_{i=n}^{\infty} (1-\eta)^{i-n} E_n^\eta [g_i^\eta(\theta, \xi_i(\theta)) - \bar{g}(\theta)],$$

where when E_n^η is used, the initial condition is $\xi_n(\theta) = \xi_n^\eta$. For the initial conditions ξ_n^η confined to any compact set,

$$\{|\Gamma_n^\eta(\theta_n^\eta)|^2 I_{|\theta_n^\eta - \bar{\theta}| \le \rho}, |\Gamma_n^\eta(\bar{\theta})|^2; n, \eta\}$$

is uniformly integrable, and

$$E\left|E_n^\eta \Gamma_{n+1}^\eta(\theta_{n+1}^\eta) - \Gamma_{n+1}^\eta(\theta_n^\eta)\right|^2 I_{|\theta_n^\eta - \bar{\theta}| \le \rho} = O(\eta^2).$$

C.4 There is a Hurwitz matrix A such that

$$\bar{g}(\theta) = A(\theta - \bar{\theta}) + o(\theta - \bar{\theta}).$$

C.5 There is a matrix $\Sigma_0 = \{\sigma_{0,ij};\, i, j = i, ..., r\}$ such that as $n, m \to \infty$,

$$\frac{1}{m} \sum_{i=n}^{n+m-1} E_n^\eta [M_i^\eta (M_i^\eta)' - \Sigma_0] I_{|\theta_n^\eta - \bar{\theta}| \le \rho} \to 0,$$

in probability.

Then $\{U^\eta(\cdot)\}$ is tight. Given tightness, we further assumes the following assumptions hold.

C.6 There is a matrix $\bar{\Sigma}_0 = \{\bar{\sigma}_{0,ij};\, i, j = i, ..., r\}$ such that as $n, m \to \infty$,

$$\frac{1}{m} \sum_{i=n}^{n+m-1} E_n^\eta [g_i^\eta(\bar{\theta}, \xi_i(\bar{\theta}))(g_i^\eta(\bar{\theta}, \xi_i(\bar{\theta})))' - \bar{\Sigma}_0] \to 0,$$

in probability.

C.7 Define another function

$$G_n^{\eta,i}(\theta, \xi_n^\eta) = E_n^\eta \left[\Gamma_{n+1}^\eta(\theta_n^\eta)[Y_n^\eta]' I_{|\theta_n^\eta - \bar{\theta}| \le \rho} \middle| \theta_n^\eta = \theta\right].$$

It needs to be a continuous function in (θ, ξ_n^η), uniformly in n and η.

C.8 There is a matrix $\Sigma_1 = \{\sigma_{1,ij};\, i, j = i, ..., r\}$ such that as $n, m \to \infty$,

$$\frac{1}{m} \sum_{i=n}^{n+m-1} E_n^\eta [G_n^{\eta,i}(\bar{\theta}, \xi_i(\bar{\theta})) - \Sigma_1] \to 0$$

in probability.

Then there exists a Wiener process $W(\cdot)$ with covariance matrix $\Sigma = \Sigma_0 + \bar{\Sigma}_0 + \Sigma_1 + \Sigma_1'$ such that $\{U^\eta(\cdot)\}$ converges weakly to a stationary solution of

$$dU = AU\,dt + dW.$$

2. Proof of Theorem 1

The proof consists of two parts. In the first part, we show that $\{X^\eta(\cdot)\}$ is tight. Therefore, every sub-sequence has further one sub-sequence that weakly converges to some limit process. In the second part, we find the limit ODE and show that the solution to this ODE exists and is unique. Combining these two parts, we prove the result.

• **Tightness.** We first rewrite MSGD as follows:

$$x_{k+1}^\eta = x_1^\eta - \eta \sum_{j=1}^{k} \sum_{i=1}^{j} \mu^{j-i} f(x_i^\eta, \xi_i^\eta).$$

Under Assumption 1, we have

$$||x_k^\eta||_2 \leq ||x_1^\eta||_2 + \eta \sum_{j=1}^{k} \sum_{i=1}^{j} \mu^{j-i} C \leq ||x_1^\eta||_2 + \frac{Ck\eta}{1-\mu}.$$

Then the continuous interpolation $X^\eta(t)$ satisfies:

$$||X^\eta(t)||_2 \leq ||x_1^\eta||_2 + \frac{t}{\eta} \frac{C\eta}{1-\mu} = ||x_1^\eta||_2 + \frac{Ct}{1-\mu}.$$

We define $K_{\delta,t} = \left\{ x \middle| ||x||_2 \leq ||x_1^\eta||_2 + \frac{Ct}{1-\mu} \right\}$. Then for any $\delta > 0, t > 0$ we have

$$\inf_\eta \mathbb{P}\{X^\eta(t) \in K_{\delta,t}\} = 1 \geq 1 - \delta.$$

Moreover, $\forall \tau, s > 0$, we have

$$||X^\eta(\tau + s) - X^\eta(\tau)||_2 \leq \frac{Cs}{1-\mu}.$$

Therefore, for each positive T,

$$\lim_\delta \limsup_\eta \sup_{|\tau| \leq T} \sup_{s \leq \delta} \mathbb{E} \min[\|X^\eta(\tau + s) - X^\eta(\tau)\|_2, 1] = 0.$$

Then by Theorem 4, $\{X^\eta(\cdot)\}$ is tight.

• **Limit Process.** For simplicity, we define

$$\beta_k^\eta = \textstyle\sum_{i=0}^{k-1} \mu^{k-i} (\nabla f(x_i^\eta, \xi_i) - \nabla \mathcal{F}(x_i^\eta)) \quad \text{and} \quad \epsilon_k = \nabla f(x_k^\eta, \xi_k) - \nabla \mathcal{F}(x_k^\eta).$$

We then rewrite the algorithm as follows:

$$m_{k+1}^\eta = m_k^\eta + (1-\mu) \left[-m_k^\eta + \widetilde{M}(x_k^\eta) \right], \quad x_{k+1}^\eta = x_k^\eta + \eta(m_{k+1}^\eta + \beta_k^\eta + \epsilon_k^\eta),$$

where $\widetilde{M}(x_k^\eta) = -\frac{1}{1-\mu} \nabla \mathcal{F}(x_k^\eta)$ is the rescaled negative gradient and

$$m_{k+1} = -\sum_{i=0}^{k} \mu^i \nabla \mathcal{F}(x_i^\eta).$$

Define the sums

$$\mathcal{E}^\eta(t) = \eta \sum_{i=0}^{t/\eta-1} \epsilon_i^\eta, \quad B^\eta(t) = \eta \sum_{i=0}^{t/\eta-1} \beta_i^\eta,$$
$$\bar{G}^\eta(t) = \eta \sum_{i=0}^{t/\eta-1} \widetilde{M}(x_i^\eta), \quad \widetilde{G}^\eta(t) = \eta \sum_{i=0}^{t/\eta-1} [m_{i+1}^\eta - \widetilde{M}(x_i^\eta)].$$

Then the continuous-time interpolation of $X^\eta(t)$ can be decomposed as follows.

$$X^\eta(t) = x_0^\eta + \bar{G}^\eta(t) + \widetilde{G}^\eta(t) + B^\eta(t) + \mathcal{E}^\eta(t).$$

Define the process $W^\eta(t)$ by

$$W^\eta(t) = X^\eta(t) - x_0^\eta - \bar{G}^\eta(t) = \widetilde{G}^\eta(t) + B^\eta(t) + \mathcal{E}^\eta(t).$$

We have already shown that $\{X^\eta(\cdot))\}$ is tight in the first part of the proof. Specifically, there is a subsequence $\eta(k) \to 0$ and a process $X(\cdot)$ such that

$$X^{\eta(k)}(t) \Rightarrow X(t),$$

as $k \to \infty$. Under the bounded assumption of $\nabla f(x, \xi)$, one can show that

$$||x_{k+1}^\eta - x_k^\eta||_2 = \eta || \sum_{i=1}^{k} \mu^{j-i} f(x_i^\eta, \xi_i^\eta)||_2 \leq \frac{\eta}{1-\mu} C,$$

which further implies the uniform integrability of $\{x_k^\eta\}$. By Lemma 2.1 in Kushner and Vazquez-Abad (1996), we know that any weak sense limit $X(t)$ must have Lipschitz continuous path. For notational simplicity, we write $\eta(k)$ as η in the following proof.

For $t > 0$ and integer p, we take $s_i \leq t$, $i \leq p$, and $\tau > 0$. Let $g(\cdot)$ be a continuous, bounded and real-valued function. Then by definition of $W^\eta(t)$, we have

$$0 = Eg(X^\eta(s_i), i \leq p)[W^\eta(t+\tau) - W^\eta(t)] \tag{7}$$
$$- Eg(X^\eta(s_i), i \leq p)[\widetilde{G}^\eta(t+\tau) - \widetilde{G}^\eta(t)] \tag{8}$$
$$- Eg(X^\eta(s_i), i \leq p)[\mathcal{E}^\eta(t+\tau) - \mathcal{E}^\eta(t)] \tag{9}$$
$$- Eg(X^\eta(s_i), i \leq p)[B^\eta(t+\tau) - B^\eta(t)]. \tag{10}$$

Let $\mathcal{F}_n^\eta = \sigma\{x_i^\eta, \xi_{i-1}^\eta, i \leq n\}$, then $\mathcal{F}_{t/\eta}^\eta$ measures $\{\mathcal{E}^\eta(s), s \leq t\}$ by definition and the process $\mathcal{E}^\eta(\cdot)$ is actually an $\mathcal{F}_{t/\eta}^\eta$-martingale. By the tower property of the conditional expectation, we know term (9) equals to 0.

Next, we eliminate term (10). Note that for any $m, n > 0$, we have

$$\left\| \frac{1}{m} \sum_{i=n}^{n+m-1} \mathbb{E}[\beta_i^\eta | \mathcal{F}_n] \right\|_2 = \left\| \frac{1}{m} \sum_{i=n}^{n+m-1} \mu^{i-n} \beta_n^\eta \right\|_2 \leq \frac{1}{(1-\mu)m} \|\beta_n^\eta\|_2.$$

Since β_n^η is uniformly bounded in η, m and n, we have

$$\lim_{m,n,\eta} \frac{1}{m} \sum_{i=n}^{n+m-1} \mathbb{E}[\beta_i^\eta | \mathcal{F}_n] = 0$$

in $\mathcal{L}_2$, which also means

$$\lim_{\eta \to 0} \mathbb{E}[B^\eta(t+\tau) - B^\eta(t) | \mathcal{F}_{t/\eta}^\eta] = 0.$$

Together with the boundedness of f, by Dominated Convergence Theorem, we know that term (10) goes to 0, as $\eta \to 0$.

For term (8), we first bound $\|\widetilde{G}^\eta(t+\tau) - \widetilde{G}^\eta(t)\|_2$. Since $\frac{1}{1-\mu} = \sum_{i=0}^\infty \mu^i$, there exists $N(\eta) = \log_\mu(1-\mu)\eta$ such that $\sum_{i=N(\eta)}^\infty \mu^i < \eta$. When $k > N(\eta)$, write m_k^η and $\widetilde{M}(x_k^\eta)$ into summations:

$$m_{k+1}^\eta = -\sum_{i=0}^k \mu^i \nabla \mathcal{F}(x_i^\eta) = -\sum_{i=0}^{N(\eta)} \mu^i \nabla \mathcal{F}(x_i^\eta) - \sum_{i=N(\eta)+1}^k \mu^i \nabla \mathcal{F}(x_i^\eta),$$

and

$$\widetilde{M}(x_k^\eta) = -\frac{1}{1-\mu} \nabla \mathcal{F}(x_k^\eta) = -\sum_{i=0}^{N(\eta)} \mu^i \nabla \mathcal{F}(x_k^\eta) - \sum_{i=N(\eta)+1}^\infty \mu^i \nabla \mathcal{F}(x_k^\eta).$$

Note that $\|x_{k+1}^\eta - x_k^\eta\|_2 \le \frac{C}{1-\mu}\eta$. Then we have

$$\max_{i=0,1,\ldots,N(\eta)} \|x_{k-i}^\eta - x_k^\eta\|_2 \le \frac{C}{1-\mu} N(\eta)\eta \to 0,$$

as $\eta \to 0$. By the Lipschitz assumption, for $i = 0, 1, ..., N(\delta)$, we have

$$\|\nabla \mathcal{F}(x_k^\eta) - \nabla \mathcal{F}(x_{k-i}^\eta)\|_2 \le L \frac{C}{1-\mu} N(\eta)\eta.$$

Then

$$\left\| \sum_{i=0}^{N(\eta)} \mu^i \{\nabla \mathcal{F}(x_{k-i}^\eta) - \nabla \mathcal{F}(x_k^\eta)\} \right\|_2 \le \frac{LCN(\eta)\eta}{(1-\mu)^2}.$$

Since $\nabla \mathcal{F}(x_k^\eta)$ is bounded by C, both $\sum_{i=N(\eta)+1}^k \mu^i \nabla \mathcal{F}(x_{k-i}^\eta)$ and $\sum_{i=N(\eta)+1}^\infty \mu^i \nabla \mathcal{F}(x_k^\eta)$ are bounded by $C\eta$. Thus,

$$\|m_{k+1}^\eta - \widetilde{M}(x_k^\eta)\|_2 \le \frac{KCN(\eta)\eta}{1-\mu} + 2C\eta = O\left(\eta \log \frac{1}{\eta}\right).$$

For $k < N(\eta)$, following the same approach, we can bound $\|m_{k+1}^\eta - \widetilde{M}(m_k^\eta)\|_2$ by the same bound $O\left(\eta \log \frac{1}{\eta}\right)$. Therefore, we have the following bound for $\|\widetilde{G}^\eta(t+\tau) - \widetilde{G}^\eta(t)\|_2$.

$$\|\widetilde{G}^\eta(t+\tau) - \widetilde{G}^\eta(t)\|_2 \le \tau O\left(\eta \log \frac{1}{\eta}\right).$$

Thus, term (8) goes to 0 as $\eta \to 0$. Then we have

$$\lim_\eta E g(X^\eta(s_i), i \le p)[W^\eta(t+\tau) - W^\eta(t)] = 0.$$

Define

$$W(t) = X(t) - X(0) - \int_0^T \widetilde{M}(X(s))ds.$$

Then the weak convergence and the previous analysis together imply that

$$Eg(X^\eta(s_i), i \le p)[W(t+\tau) - W(t)] = 0.$$

Here, we need an important result in the martingale theory:

THEOREM 6 **(Kushner and Yin (2003), Theorem 4.1, Chapter 7)**. *Let $U(\cdot)$ be a random process with paths in $D^d[0,\infty)$, where $U(t)$ is measurable on the σ-algebra $\mathcal{F}_t^X$ determined by $\{X(s), s \le t\}$ for some given process $X(\cdot)$ and let $\mathbb{E}[U(t)] < \infty$ for each t. Suppose that for each real $t \ge 0$ and $\tau \ge 0$, each integer p and each set of real numbers $s_i \le t$, $i = 1, ..., p$, and each bounded and continuous real-valued function $h(\cdot)$,*

$$Eh(X^\eta(s_i), i \le p)[U(t+\tau) - U(t)] = 0,$$

then $U(t)$ is a $\mathcal{F}_t^X$-martingale.

By Theorem 6 , we know that $W(\cdot)$ is a martingale. It has locally Lipschitz continuous sample paths by the fact $X(\cdot)$ is Lipschitz. Since a Lipschitz continuous martingale must almost surely be a constant, we know $W(t) = W(0) = 0$ with probability 1. In other words, $X(t)$ is a solution to the following ODE

$$\dot{X} = -\frac{1}{1-\mu}\nabla\mathcal{F}(x), \quad x(0) = x_0. \tag{11}$$

Moreover, under Assumption 1 and by Theorem 12.70.B in Simmons (2016), we know that the above initial value problem has only one solution. Therefore, all sub-sequences of $\{X^\eta(\cdot)\}$ weakly converge to the same limit, which implies the weak convergence of the entire sequence. We prove the theorem. □

3. Detailed Proof in Section 4

3.1. Proof of Theorem 2

Proof. The proof follows from Theorem 10.8.1 in Kushner and Yin (2003) (Theorem 5). We need to check the Assumption C.1 to C.8 (in Appendix 1)

1. The uniform integrability in C.1 directly follows from the uniform boundedness assumption of $\nabla f(x, \xi)$.
2. C.2 can be easily got from the proof of ODE approximation.
3. To check condition C.4, we need use our isolated stationary point assumption, i.e, Assumption 2. At the local optimum x^*, the Hessian matrix must be positive definite. Then C.4 is obviously satisfied with the Hurwitz matrix $-\nabla^2\mathcal{F}(x^*)$.

The main challenge left is to calculate the variance of the Wiener process and check the other five assumptions.

For simplicity, $E_k^\eta[\cdot]$ means the conditional expectation for

$$\{\zeta_{k+j}, j \geq 0; \zeta_k(X) = \zeta_k^\eta\}.$$

From Equation (9), the variance can be decomposed into three parts. The first part is from the noise $\gamma_k^{\eta,i}$. Since we have assumed the weak convergence $x_k^\eta \Rightarrow x^*$, we have in distribution,

$$\lim_{\eta,k} E_k^\eta(\gamma_{k+j}^\eta(\gamma_{k+j}^\eta)^\top) = \Sigma.$$

Since the limit is a constant, the convergence also holds in probability. Thus, C.5 is satisfied. The second part comes from the fixed-state-chain:

$$\begin{aligned}
E_k^\eta(g(x^*, \zeta_{k+j}^\eta(x^*))g(x^*, \zeta_{k+j}^\eta(x^*))^\top) &= E_k^\eta(\zeta_{k+j}^\eta(x^*) - \nabla F(x^*))(\zeta_{k+j}^\eta(x^*) - \nabla F(x^*))^\top \\
&= E_k^\eta \zeta_{k+j}^\eta(x^*)(\zeta_{k+j}^\eta(x^*))^\top \\
&= \mu^{2j}(\zeta_k^\eta)(\zeta_k^\eta)^\top + \sum_{m=0}^{j-1} \mu^{2(j-m)} E_k^\eta[\nabla f(x^*, \xi_{k+m}) \nabla f(x^*, \xi_{k+m})^\top] \\
&\to \frac{\mu^2}{1-\mu^2}\Sigma,
\end{aligned}$$

in probability, as $k, j \to 0$. Thus, C.6 is satisfied.

The last part is from the term $g(\zeta_k^\eta, x_k^\eta) - g(\zeta_k(x_k^\eta), x_k^\eta)$. Define the discounted sequence

$$\Gamma_k^\eta(x) = \sum_{j=0}^{\infty} (1-\eta)^j E_k^\eta[g(x, \zeta_{k+j}^\eta(x)) - \widetilde{M}(x)].$$

Note that

$$\begin{aligned}
E_k^\eta[\zeta_{k+j}^\eta(x)] &= E_k^\eta[\mu^j \zeta_k^\eta - \sum_{m=0}^{j-1} \mu^{j-m} \nabla f(x, \xi_{k+m})] \\
&= \mu^j \zeta_k^\eta - \sum_{m=0}^{j-1} \mu^{j-m} \nabla \mathcal{F}(x).
\end{aligned}$$

Thus, we have

$$E_k^\eta[g(x, \zeta_{k+j}^\eta(x)) - \widetilde{M}(x)] = \mu^j \zeta_k^\eta + \frac{\mu^{j+1}}{1-\mu} \nabla \mathcal{F}(x).$$

Then

$$\Gamma_k^\eta(x) = \sum_{j=0}^{\infty} (1-\eta)^j \left\{ \mu^j \zeta_k^\eta + \frac{\mu^{j+1}}{1-\mu} \nabla \mathcal{F}(x) \right\} = \frac{1}{1-(1-\eta)\mu} \left(\zeta_k^\eta - \frac{\mu}{1-\mu} \widetilde{M}(x) \right).$$

Since M is locally Lipschitz, and $\|x_{k+1}^\eta - x_k^\eta\|_2 = O(\eta)$, the following result holds:

$$\begin{aligned}
\|E_k^\eta[\Gamma_{k+1}^\eta(x_{k+1}^\eta) - \Gamma_{k+1}^\eta(x_k^\eta)]\|_2^2 &= \left\| \frac{\mu}{(1-(1-\eta)\mu)(1-\mu)} \left\{ E_k^\eta[\widetilde{M}(x_{k+1}^\eta) - \widetilde{M}(x_k^\eta)] \right\} \right\|_2^2 \\
&= O(\eta^2).
\end{aligned}$$

Then, Assumption C.3 holds.

Define another function

$$G_k^\eta(x,\zeta_k^\eta) = E_k^\eta\left[\Gamma_{k+1}^{\eta,i}(x_k^\eta)(Z_k^\eta)^\top | x_k^\eta = x\right].$$

It is easy to check this is a continuous function in (x,ζ_k^η), uniformly in k and η (Assumption C.7). Moreover,

$$\begin{aligned}\Gamma_{k+1}^\eta(x_k^\eta)(Z_k^\eta)^\top &= \frac{1}{1-(1-\eta)\mu}\left(\zeta_{k+1}^\eta - \frac{\mu}{1-\mu}\widetilde{M}(x_k^\eta)\right)\frac{1}{\mu}(\zeta_{k+1}^\eta)^\top \\ &= \frac{1}{1-(1-\eta)\mu}\left(\frac{1}{\mu}\zeta_{k+1}^\eta(\zeta_{k+1}^\eta)^\top - \frac{1}{1-\mu}\widetilde{M}(x_k^\eta)(\zeta_{k+1}^\eta)^\top\right).\end{aligned}$$

Then we have

$$\begin{aligned}E_k^\eta[\zeta_{k+1}^\eta(\zeta_{k+1}^\eta)^\top | x_k^\eta = x^*] &= E_k^\eta\left[\left(\mu\zeta_k^\eta - \mu\nabla f(x_k^\eta,\xi_k)\right)\left(\mu\zeta_k^\eta - \mu\nabla f(x_k^\eta,\xi_k)\right)^\top \middle| x_k^\eta = x^*\right] \\ &= \mu^2\zeta_k^\eta(\zeta_k^\eta)^\top + \mu^2\Sigma,\end{aligned}$$

and

$$E_k^\eta[\widetilde{M}(x_k^\eta)\zeta_{k+1}^\eta | x_k^\eta = x^*] = 0.$$

Those imply that

$$\begin{aligned}E_k^\eta G_{k+j}^\eta(x^*,\zeta_{k+j}^\eta(x^*)) &= \frac{\mu}{1-(1-\eta)\mu}(E_k^\eta\zeta_{k+j}^\eta(x^*)(\zeta_{k+j}^\eta(x^*))^\top + \Sigma) \\ &\to \frac{1}{1-\mu^2}\frac{\mu}{1-\mu}\Sigma,\end{aligned}$$

in probability. Thus, C.8 is satisfied. We have proved all the assumptions of Theorem 5 are satisfied. As a result, there exists a Wiener Process W, such that any subsequence of $\{U^{\eta,i}\}$ converges weakly to a stationary solution of

$$dU = -\frac{1}{1-\mu}\nabla^2\mathcal{F}(x^*)U dt + dW,$$

where the variance of W is $[1+\frac{\mu^2}{1-\mu^2}+2\frac{1}{1-\mu^2}\frac{\mu}{1-\mu}]\Sigma = \frac{1}{(1-\mu)^2}\Sigma$.

Lastly, we show that the above SDE has one unique solution given any initial. In fact, one can verify that both the drift term and the diffusion term are Lipschitz continuous. By Theorem 5.2.5 in Karatzas and Shreve (1998), we know that the solution exists and is unique.

Therefore, $\{U^{\eta,i}\}$ converges weakly to the unique stationary solution of

$$dU = -\frac{1}{1-\mu}\nabla^2\mathcal{F}(x^*)U dt + dW,$$

We finish the proof. □

3.2. Proof of Theorem 3

Proof of Theorem 3 Since we restart our record time, we assume here the algorithm is initialized around one local optimum x^*. Thus, we have $||U^\eta(0)||_2^2 = \eta^{-1}\delta^2 < \infty$. Note that $U^\eta(t)$ converges to $U(t)$ in this neighborhood, and the second moment of $U(t)$ is:

$$\begin{aligned}
\mathbb{E}\left(||U(t)||_2^2\right) &= \mathbb{E}\left[\mathrm{tr}(U(t)U(t)^\top)\right] = \mathrm{tr}\left[\mathbb{E}U(t)U(t)^\top\right] \\
&= \mathrm{tr}\left[\exp\left(-\frac{t}{1-\mu}\nabla^2\mathcal{F}(x^*)\right)[U(0)U(0)^\top]\exp\left(-\frac{t}{1-\mu}\nabla^2\mathcal{F}(x^*)\right)\right] \\
&\quad + \mathrm{tr}\left[\int_0^t \exp\left(-\frac{1}{1-\mu}\nabla^2\mathcal{F}(x^*)s\right)\frac{1}{(1-\mu)^2}\Sigma\exp\left(-\frac{1}{1-\mu}\nabla^2\mathcal{F}(x^*)s\right)ds\right] \\
&= \mathrm{tr}\left[\exp\left(-\frac{t}{1-\mu}\nabla^2\mathcal{F}(x^*)\right)[U(0)U(0)^\top]\exp\left(-\frac{t}{1-\mu}\nabla^2\mathcal{F}(x^*)\right)\right] \\
&\quad + \frac{1}{(1-\mu)^2}\int_0^t \mathrm{tr}\left(\exp\left(-\frac{1}{1-\mu}\nabla^2\mathcal{F}(x^*)s\right)\Sigma\exp\left(-\frac{1}{1-\mu}\nabla^2\mathcal{F}(x^*)s\right)\right)ds \\
&= \frac{1}{(1-\mu)^2}\int_0^t \left\|\exp\left(-\frac{1}{1-\mu}\nabla^2\mathcal{F}(x^*)s\right)\Sigma^{\frac{1}{2}}\right\|_{\mathrm{F}}^2 ds \\
&\quad + \left\|\exp\left(-\frac{1}{1-\mu}\nabla^2\mathcal{F}(x^*)t\right)U(0)\right\|_{\mathrm{F}}^2 \\
&= \sum_{i=1}^d\left\{\left\|e_ie_i^\top U(0)\right\|_{\mathrm{F}}^2\exp\left(-\frac{2\lambda_i}{1-\mu}t\right) + \int_0^t\frac{1}{(1-\mu)^2}\left\|e_ie_i^\top\Sigma^{\frac{1}{2}}\right\|_{\mathrm{F}}^2\exp\left(-\frac{2\lambda_i}{1-\mu}s\right)ds\right\} \\
&= \sum_{i=1}^d\left\|e_ie_i^\top U(0)\right\|_{\mathrm{F}}^2\exp\left(-\frac{2\lambda_i}{1-\mu}t\right) + \frac{1}{(1-\mu)}\frac{1-\exp(-\frac{2\lambda_i}{1-\mu}t)}{2\lambda_i}\left\|e_ie_i^\top\Sigma^{\frac{1}{2}}\right\|_{\mathrm{F}}^2 \\
&= \sum_{i=1}^d(U(0)^\top e_i)^2\exp\left(-\frac{2\lambda_i}{1-\mu}t\right) + \frac{1-\exp(-\frac{2\lambda_i}{1-\mu}t)}{2(1-\mu)\lambda_i}e_i^\top\Sigma e_i,
\end{aligned}$$

By Markov inequality, we have:

$$\begin{aligned}
\eta^{-1}\epsilon\mathbb{P}\left(\|X^\eta(T_3)-x^*\|_2^2>\epsilon\right) &\le \eta^{-1}\mathbb{E}\left(\|X^\eta(T_3)-x^*\|_2^2\right) = \mathbb{E}\left(||U^\eta(T_3)||_2^2\right) \\
&\to \sum_{i=1}^d(U(0)^\top e_i)^2\exp\left(-\frac{2\lambda_i}{1-\mu}T_3\right) + \frac{1-\exp(-\frac{2\lambda_i}{1-\mu}T_3)}{2(1-\mu)\lambda_i}e_i^\top\Sigma e_i, \quad \text{as } \eta\to 0.
\end{aligned}$$

Thus, for a sufficiently small η, we have

$$\begin{aligned}
\mathbb{P}\left(\|X^\eta(T_3)-x^*\|_2^2>\epsilon\right) &\le \frac{2}{\eta^{-1}\epsilon}\sum_{i=1}^d(U(0)^\top e_i)^2\exp\left(-\frac{2\lambda_i}{1-\mu}T_3\right) + \frac{1-\exp(-\frac{2\lambda_i}{1-\mu}T_3)}{2(1-\mu)\lambda_i}e_i^\top\Sigma e_i \\
&\le \frac{2}{\eta^{-1}\epsilon}\Big(\eta^{-1}\delta^2\exp\left[-2\frac{\lambda_dT_3}{1-\mu}\right] + \frac{\phi}{2(1-\mu)\lambda_d}\Big(1-\exp\big(-2\frac{\lambda_1T_3}{1-\mu}\big)\Big)\Big) \\
&\le \frac{2}{\eta^{-1}\epsilon}\Big(\eta^{-1}\delta^2\exp\left[-2\frac{\lambda_dT_3}{1-\mu}\right] + \frac{\phi}{2(1-\mu)\lambda_d}\Big),
\end{aligned}$$

where $\phi = \sum_{i=1}^d e_i^\top\Sigma e_i$. The above inequality actually implies that the desired probability is asymptotically upper bounded by the term on the right hand. Thus, to guarantee

$$\mathbb{P}\left((\|X^\eta(T_3)-x^*\|_2^2>\epsilon\right) \le \frac{1}{4}$$

when η is sufficiently small, we need

$$\frac{2}{\eta^{-1}\epsilon}\Big(\eta^{-1}\delta^2\exp\left[-2\frac{\lambda_d T_3}{1-\mu}\right]+\frac{\phi}{2(1-\mu)\lambda_d}\Big)\leq\frac{1}{4}.$$

The above inequality has a solution only when:

$$(1-\mu)\lambda_d\epsilon-4\eta\phi>0.$$

Moreover, when the above inequality holds, we have:

$$T_3=\frac{1-\mu}{2\lambda_d}\log\left(\frac{8(1-\mu)\lambda_d\delta^2}{(1-\mu)\lambda_d\epsilon-4\eta\phi}\right).$$

We finish the proof. □

3.3. Proof of Theorem 5

Proof of Theorem 5. Recall that Theorem 4 holds when $u_k^\eta=(x_k^\eta-\widehat{x})/\sqrt{\eta}$ is bounded. Thus, if $||X^\eta(T_1)||_2^2\geq\delta^2$ holds at some time T_1, the algorithm has successfully escaped from the saddle point. We approximate $U^\eta(t)$ by the limiting process approximation, which is Gaussian distributed at time t. As $\eta\to0$, by simple manipulation, we have

$$\mathbb{P}\left(||X^\eta(T_1)||_2^2\geq\delta^2\right)=\mathbb{P}\left(||U^\eta(T_1)||_2^2\geq\eta^{-1}\delta^2\right).$$

We then prove $\mathbb{P}\left(||U^\eta(T_1)||_2^2\geq\eta^{-1}\delta^2\right)\geq1-\nu$. At time t, $U^\eta(t)$ converges to a Gaussian distribution with mean 0 and covariance matrix

$$\int_0^{T_1}\exp\left(-\frac{1}{1-\mu}\nabla^2\mathcal{F}(\widehat{x})s\right)\frac{1}{(1-\mu)^2}\Sigma\exp\left(-\frac{1}{1-\mu}\nabla^2\mathcal{F}(\widehat{x})s\right)ds.$$

Let $\nabla^2\mathcal{F}(\widehat{x})=P\Lambda P^\top$ where $\Lambda=\mathrm{diag}(\lambda_1,...,\lambda_d)$ and $\lambda_1\geq\lambda_2\geq...\geq\lambda_d$ and $\lambda_d<0$. Since P is orthogonal, we have $||P^\top U^\eta(T_1)||_2=||U^\eta(T_1)||_2$, and $P^\top U^\eta$ converges to a Gaussian distribution with mean 0 and covariance matrix

$$\begin{aligned}&\int_0^{T_1}P^\top\exp\left(-\frac{1}{1-\mu}\nabla^2\mathcal{F}(\widehat{x})s\right)\frac{1}{(1-\mu)^2}\Sigma\exp\left(-\frac{1}{1-\mu}\nabla^2\mathcal{F}(\widehat{x})s\right)Pds\\=&\int_0^{T_1}P^\top P\exp\left(-\frac{1}{1-\mu}\Lambda s\right)\frac{1}{(1-\mu)^2}P^\top\Sigma P\exp\left(-\frac{1}{1-\mu}\Lambda s\right)P^\top Pds\\=&\int_0^{T_1}\exp\left(-\frac{1}{1-\mu}\Lambda s\right)\frac{1}{(1-\mu)^2}P^\top\Sigma P\exp\left(-\frac{1}{1-\mu}\Lambda s\right)ds.\end{aligned}$$

Moreover , $(P^\top U^\eta)^{(d)}$ converge to normal distribution with mean 0 and variance

$$\int_0^{T_1}\exp\left(-\frac{2\lambda_d}{1-\mu}s\right)\frac{1}{(1-\mu)^2}(P^\top\Sigma P)_{d,d}ds=\frac{(P^\top\Sigma P)_{d,d}}{2\lambda_d(1-\mu)}\left(1-\exp\left(-\frac{2\lambda_d}{1-\mu}s\right)\right).$$

Therefore, let $\Phi(x)$ be the CDF of $N(0,1)$, we have

$$\mathbb{P}\left(\frac{\left|(P^\top U^\eta(T_1))^{(d)}\right|}{\sqrt{\frac{(P^\top \Sigma P)_{d,d}}{2\lambda_d(1-\mu)}\left(1-\exp\left(-\frac{2\lambda_d}{1-\mu}s\right)\right)}} \geq \Phi^{-1}\left(\frac{1+\nu/2}{2}\right)\right) \to 1-\nu/2, \quad \text{as } \eta \to 0.$$

When the following inequality holds,

$$\eta^{-\frac{1}{2}}\delta \leq \Phi^{-1}\left(\frac{1+\nu/2}{2}\right)\cdot\sqrt{\frac{(P^\top \Sigma P)_{d,d}}{2\lambda_d(1-\mu)}\left(1-\exp\left(-\frac{2\lambda_d}{1-\mu}s\right)\right)},$$

we get

$$T_1 = \frac{(1-\mu)}{2|\lambda_d|}\log\left(\frac{2\eta^{-1}\delta^2(1-\mu)|\lambda_d|}{\Phi^{-1}\left(\frac{1+\nu/2}{2}\right)^2 (P^\top \Sigma P)_{d,d}}+1\right).$$

Thus, for a sufficiently small ϵ, we have

$$\begin{aligned}\mathbb{P}\left(||U^\eta(T_1)||_2^2 \geq \eta^{-1}\delta^2\right) &= \mathbb{P}\left(||P^\top U^\eta(T_1)||_2^2 \geq \eta^{-1}\delta^2\right)\\ &\geq \mathbb{P}\left(\left|\left(P^\top U^\eta(T_1)\right)^{(d)}\right| \geq \eta^{-1/2}\delta\right)\\ &\geq 1-\nu.\end{aligned}$$

Take $\nu = \frac{1}{4}$, and we prove the theorem. □

4. Detailed Proof in Section 5

4.1. Derivation of Momentum Stochastic Generalized Hebbian Algorithm

SGHA is essentially a primal-dual algorithm. Specifically, we consider the Lagrangian function of (17):

$$L(v,\lambda) = v^\top \mathbb{E}_{X\sim\mathcal{D}}[XX^\top]v - \lambda(v^\top v - 1),$$

where λ is the Lagrangian multiplier. We then check the optimal KKT conditions:

$$\mathbb{E}_{X\sim\mathcal{D}}[XX^\top]v - \lambda v = 0 \text{ and } v^\top v = 1,$$

which implies $\lambda = v^\top \mathbb{E}_{X\sim\mathcal{D}}[XX^\top]v$. At the k-th iteration, SGHA takes the following primal-dual update:

- Dual Update: $\lambda_k = v_k^\top \Sigma_k v_k$,
- Primal Update: $v_{k+1} = v_k + \eta(\Sigma_k v_k - \lambda_k v_k)$,

where $\Sigma_k = X_k X_k^\top$ and $\mu(v_k - v_{k-1})$ is the momentum with a parameter $\mu \in [0,1)$. Combine the primal and dual updates together, we obtain a dual free update:

$$v_{k+1} = v_k + \eta(\Sigma_k v_k - v_k^\top \Sigma_k v_k v_k) = v_k + \eta(I - v_k v_k^\top)\Sigma_k v_k.$$

Adding the additional momentum term $\mu(v_k - v_{k-1})$, we get update (18).

4.2. Proof of Corollary 1

To apply Theorem 1 to prove the ODE approximation for algorithm (18), we only need to check whether Assumptions 1 and 2 hold. From our landscape analysis in Section 5, we know that Assumption 2 holds naturally for streaming PCA. We only need to verify the uniform boundedness and Lipschitz continuity.

The next lemma shows that the algorithm trajectory of (18) is bounded and thus the boundedness and Lipschitz continuity in Assumption 1 holds for (18).

LEMMA 1. *Under Assumption* (4)*, given* $v_0 \in \mathbb{S}$*, for any* $k \le O(1/\eta)$*, we have*

$$\|v_k\|^2 \le 1 + O((1-\mu)^{-3}\eta) \quad \text{and} \quad \|v_{k+1} - v_k\| \le \frac{2C_d\eta}{1-\mu}.$$

Proof. First, if we assume $\{v_k\}$ is uniformly bounded by 2, by formulation (18), we then have

$$\begin{aligned} & v_{k+1} - v_k = \mu(v_k - v_{k-1}) + \eta\{\Sigma_k v_k - v_k^\top \Sigma_k v_k v_k\}, \\ \Longrightarrow & v_{k+1} - v_k = \sum_{i=0}^{k} \mu^{k-i}\eta\{\Sigma_i v_i - v_i^\top \Sigma_i v_i v_i\}, \\ \Longrightarrow & \|v_{k+1} - v_k\|_2 \le C_\delta \frac{\eta}{1-\mu}, \end{aligned}$$

where $C_\delta = \sup_{\|v\|_2 \le 2, \|X\|_2 \le C_d} \|XX^Tv - v^TXX^Tvv\|_2 \le 2C_d$. Next, we show the boundedness assumption on v can be taken off. In fact, with an initialization on $\mathbb{S}$ (the sphere of the unit ball), the algorithm is bounded in a much smaller ball of radius $1 + O(\eta)$.

Recall $\delta_{k+1} = v_{k+1} - v_k$. Let's consider the difference between the norm of two iterates,

$$\begin{aligned} \Delta_k &= \|v_{k+1}\|_2^2 - \|v_k\|_2^2 = \|\delta_{k+1}\|_2^2 + 2v_k^\top \delta_{k+1} \\ \Delta_{k+1} - \Delta_k &= \|\delta_{k+2}\|_2^2 + 2v_{k+1}^\top \delta_{k+2} - \|\delta_{k+1}\|_2^2 - 2v_k^\top \delta_{k+1} \\ &= \|\delta_{k+2}\|_2^2 - \|\delta_{k+1}\|_2^2 + 2\mu v_{k+1}^\top \delta_{k+1} + 2\eta v_{k+1}^\top \Sigma_{k+1} v_{k+1}(1 - v_{k+1}^\top v_{k+1}) - 2v_k^\top \delta_{k+1} \\ &= \|\delta_{k+2}\|_2^2 - \|\delta_{k+1}\|_2^2 + 2\mu v_k^\top \delta_{k+1} + 2\mu\|\delta_{k+1}\|_2^2 + 2\eta v_{k+1}^\top \Sigma_{k+1} v_{k+1}(1 - v_{k+1}^\top v_{k+1}) - 2v_k^\top \delta_{k+1} \\ &= |\delta_{k+2}\|_2^2 + \mu\|\delta_{k+1}\|_2^2 - (1-\mu)(\|\delta_{k+1}\|_2^2 + 2v_k^\top \delta_{k+1}) + 2\eta v_{k+1}^\top \Sigma_{k+1} v_{k+1}(1 - v_{k+1}^\top v_{k+1}) \\ &= \|\delta_{k+2}\|_2^2 + \mu\|\delta_{k+1}\|_2^2 - (1-\mu)\Delta_k + 2\eta v_{k+1}^\top \Sigma_{k+1} v_{k+1}(1 - v_{k+1}^\top v_{k+1}) \\ &\le \|\delta_{k+2}\|_2^2 + \mu\|\delta_{k+1}\|_2^2 - (1-\mu)\Delta_k. \end{aligned}$$

The last inequality holds when $\|v_{k+1}\|_2 \ge 1$. Let $\kappa = \inf\{i : \|v_{i+1}\|_2 > 1\}$, then

$$\Delta_{\kappa+1} \le (1+\mu)\left(\frac{C_\delta}{1-\mu}\right)^2 \eta^2 + \mu\Delta_\kappa.$$

Moreover, if $1 < \|v_{\kappa+i}\|_2 \le 2$ holds for $i = 1, ..., n < \frac{t}{\eta}$, we have

$$\begin{aligned}\Delta_{\kappa+i} &\le (1+\mu)\left(\frac{C_\delta}{1-\mu}\right)^2 \eta^2 + \mu\Delta_{\kappa+i-1} \\ &\le \frac{1+\mu}{1-\mu}\left(\frac{C_\delta}{1-\mu}\right)^2 \eta^2 + \mu^i \Delta_\kappa.\end{aligned}$$

Thus,

$$\begin{aligned}\|v_{\kappa+n+1}\|_2^2 &= \|v_\kappa\|_2^2 + \sum_{i=0}^{n} \Delta_{\kappa+i} \\ &\le 1 + \frac{1}{1-\mu}\Delta_k + \frac{t}{\eta}\frac{1+\mu}{1-\mu}\left(\frac{C_\delta}{1-\mu}\right)^2 \eta^2 \\ &\le 1 + O\left(\frac{\eta}{(1-\mu)^3}\right).\end{aligned}$$

In other words, when η is very small, we cannot go far from $\mathbb{S}$ and the assumption that $\|v\|_2 \le 2$ can be removed. □

Therefore all the assumptions for Theorem 1 holds and we know that $V^\eta(\cdot) \Rightarrow V(\cdot)$ in the weak sense as $\eta \to 0$ in the space $D^d[0,\infty)$, where $V(\cdot)$ is the unique solution to the following ODE:

$$\dot{V} = \frac{1}{1-\mu}(\Sigma V - V^\top \Sigma V V), \quad V(0) = v_0.$$

To solve ODE (19), we rotate the coordinate to decouple each dimension. Under Assumption 3, there exists an orthogonal matrix Q such that: $\Sigma = Q\Lambda Q^\top$, where $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, ..., \lambda_d)$. Let $H(t) = Q^\top V(t)$, or equivalently $V(t) = QH(t)$. Substitute $V(t)$ with $QH(t)$ in ODE (6), then we can obtain the following ODE.

$$\dot{H} = \frac{1}{1-\mu}[\Lambda H - H^\top \Lambda H H]. \tag{12}$$

ODE (12) is different from (4.6) in Chen et al. (2017) by a constant $\frac{1}{1-\mu}$, and has an explicit form solution. Then we have the initial value problem (19) has a solution $V(t) = QH(t)$, where

$$H^{(i)}(t) = \Big(\sum_{i=1}^{d}[H^{(i)}(0)\exp\Big(\frac{\lambda_i t}{1-\mu}\Big)]^2\Big)^{-\frac{1}{2}} H^{(i)}(0)\exp\left(\frac{\lambda_i t}{1-\mu}\right), \quad i = 1, ..., d. \tag{13}$$

where $H(0) = Q^\top v_0, v_0 \in \mathbb{S}$. Moreover, suppose $v_0 \neq \pm v^i,\ \forall i = 2, ..., d$, as $t \to \infty$, one can easily verify that $V(t)$ converges to v^1, which is the global maximum to (17).

Last, we show the uniqueness of the above solution. Define $f(t,v) = \frac{1}{1-\mu}[\Sigma v - v^\top \Sigma v v]$ and a domain $\mathcal{R} = \big\{(t,v)\big| t \ge 0, ||v||_2 \le 1\big\}$. Since $f(t,v)$ is continuously differentiable with respect to (t,v), $f(t,v)$ satisfies Lipschitz continuous condition in $\mathcal{R}$ with respect to v and uniformly in t. By Theorem 1.2.1 in Hu and Li (2004), we know the solution is unique. □

4.3. Proof of Corollary 2

Proof of Corollary 2 Phase I and III are a directly application of Theorems 3 5. Here we only consider Phase II.

After Phase I, we restart our record time, i.e., $H^{\eta,1}(0) \geq \delta$ and we obtain

$$\mathbb{P}(\left\|V^{\eta}(T_2) - v^1\right\|_2^2 \leq \delta^2) \to \mathbb{P}(\left\|V(T_2) - v^1\right\|_2^2 \leq \delta^2) = \mathbb{P}(\left\|H(T_2) - e^1\right\|_2^2 \leq \delta^2),$$

where H is defined in (13). Since H is deterministic and

$$\begin{aligned}\left(H^{(1)}(T_2)\right)^2 &= \left(\sum_{j=1}^{d}\left(\left(H^{(j)}(0)\right)^2 \exp\left(2\frac{\lambda_j}{1-\mu}T_2\right)\right)\right)^{-1}\left(H^{(1)}(0)\right)^2 \exp\left(2\frac{\lambda_1}{1-\mu}T_2\right) \\ &\geq \left(\delta^2 \exp\left(2\frac{\lambda_1}{1-\mu}T_2\right) + (1-\delta^2)\exp\left(2\frac{\lambda_2}{1-\mu}T_2\right)\right)^{-1} \delta^2 \exp\left(2\frac{\lambda_2}{1-\mu}T_2\right), \qquad (14)\end{aligned}$$

Thus, when the term (14) satisfies

$$\left(\delta^2 \exp\left(2\frac{\lambda_1}{1-\mu}T_2\right) + (1-\delta^2)\exp\left(2\frac{\lambda_2}{1-\mu}T_2\right)\right)^{-1} \delta^2 \exp\left(2\frac{\lambda_1}{1-\mu}T_2\right) \geq 1 - \delta^2/2, \qquad (15)$$

we have

$$\mathbb{P}(\left(H^{(1)}(T_2)\right)^2 \geq 1 - \delta^2/2) = 1.$$

Then for sufficiently small η, we have

$$\mathbb{P}(\left(H^{\eta,1}(T_2)\right)^2 \geq 1 - \delta^2/2) \geq \frac{3}{4}.$$

Note that when $\left(H^{(1)}(T_2)\right)^2 \geq 1 - \delta^2/2$, we have

$$\|H^{\eta}(T_2) - e^1\|_2^2 \leq 2 - 2\sqrt{1 - \delta^2/2} \leq \delta^2.$$

Therefore,

$$\mathbb{P}(\|V^{\eta}(T_2) - v^1\|_2^2 \leq \delta^2) = \mathbb{P}(\|H^{\eta}(T_2) - v^1\|_2^2 \leq \delta^2) \geq \frac{3}{4}.$$

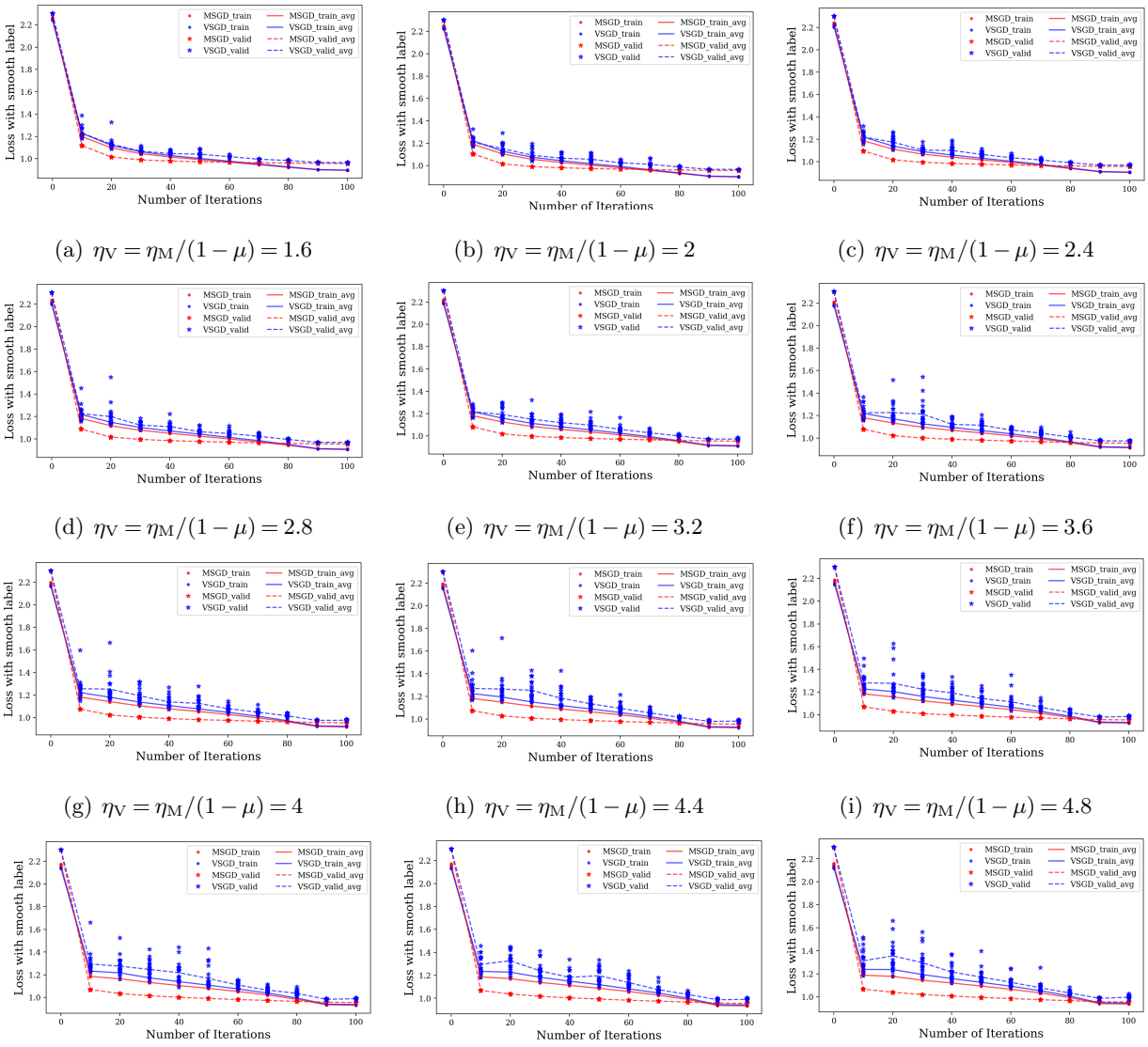
Solving the above inequality (15), we get

$$T_2 = \frac{1-\mu}{2(\lambda_1 - \lambda_2)} \log \frac{2-\delta^2}{\delta^2} \ .$$

We finish the proof. □

5. Deep Neural Networks Experiments

Figure 1 Experimental Results of ResNet-9 on CIFAR-10. VSGD uses the Equivalent Step Sizes of MSGD.



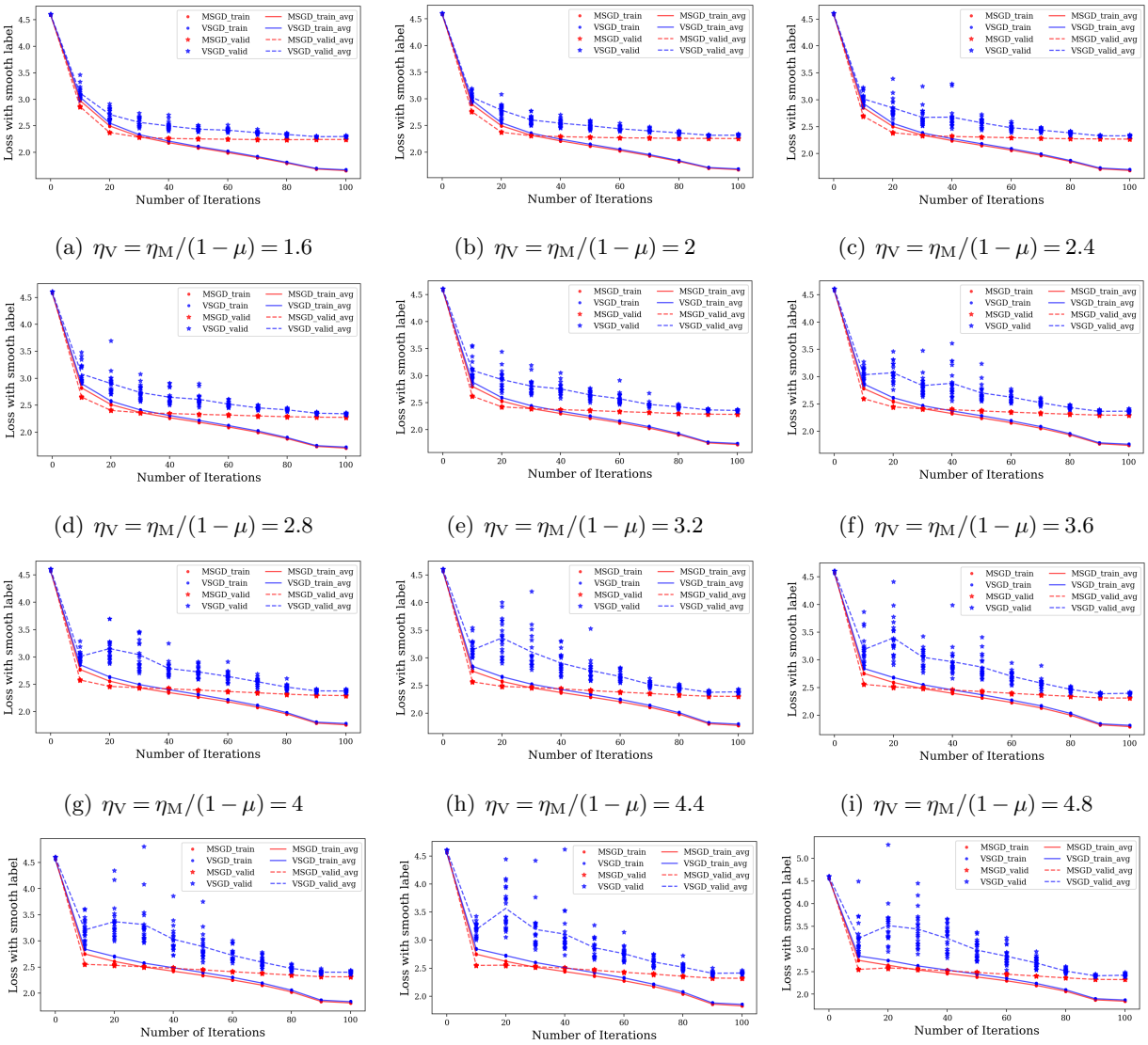
(a) $\eta_{\mathrm{V}} = \eta_{\mathrm{M}}/(1-\mu) = 1.6$ (b) $\eta_{\mathrm{V}} = \eta_{\mathrm{M}}/(1-\mu) = 2$ (c) $\eta_{\mathrm{V}} = \eta_{\mathrm{M}}/(1-\mu) = 2.4$

(d) $\eta_{\mathrm{V}} = \eta_{\mathrm{M}}/(1-\mu) = 2.8$ (e) $\eta_{\mathrm{V}} = \eta_{\mathrm{M}}/(1-\mu) = 3.2$ (f) $\eta_{\mathrm{V}} = \eta_{\mathrm{M}}/(1-\mu) = 3.6$

(g) $\eta_{\mathrm{V}} = \eta_{\mathrm{M}}/(1-\mu) = 4$ (h) $\eta_{\mathrm{V}} = \eta_{\mathrm{M}}/(1-\mu) = 4.4$ (i) $\eta_{\mathrm{V}} = \eta_{\mathrm{M}}/(1-\mu) = 4.8$

(j) $\eta_{\mathrm{V}} = \eta_{\mathrm{M}}/(1-\mu) = 5.2$ (k) $\eta_{\mathrm{V}} = \eta_{\mathrm{M}}/(1-\mu) = 5.6$ (l) $\eta_{\mathrm{V}} = \eta_{\mathrm{M}}/(1-\mu) = 6$

Figure 2 Experimental Results of ResNet-9 on CIFAR-100. VSGD uses the Equivalent Step Sizes of MSGD.



(a) $\eta_{\mathrm{V}} = \eta_{\mathrm{M}}/(1-\mu) = 1.6$ (b) $\eta_{\mathrm{V}} = \eta_{\mathrm{M}}/(1-\mu) = 2$ (c) $\eta_{\mathrm{V}} = \eta_{\mathrm{M}}/(1-\mu) = 2.4$

(d) $\eta_{\mathrm{V}} = \eta_{\mathrm{M}}/(1-\mu) = 2.8$ (e) $\eta_{\mathrm{V}} = \eta_{\mathrm{M}}/(1-\mu) = 3.2$ (f) $\eta_{\mathrm{V}} = \eta_{\mathrm{M}}/(1-\mu) = 3.6$

(g) $\eta_{\mathrm{V}} = \eta_{\mathrm{M}}/(1-\mu) = 4$ (h) $\eta_{\mathrm{V}} = \eta_{\mathrm{M}}/(1-\mu) = 4.4$ (i) $\eta_{\mathrm{V}} = \eta_{\mathrm{M}}/(1-\mu) = 4.8$

(j) $\eta_{\mathrm{V}} = \eta_{\mathrm{M}}/(1-\mu) = 5.2$ (k) $\eta_{\mathrm{V}} = \eta_{\mathrm{M}}/(1-\mu) = 5.6$ (l) $\eta_{\mathrm{V}} = \eta_{\mathrm{M}}/(1-\mu) = 6$

References

Chen Z, Yang FL, Li CJ, Zhao T (2017) Online multiview representation learning: Dropping convexity for better efficiency. *arXiv preprint arXiv:1702.08134* .

Hu J, Li WP (2004) Theory of ordinary differential equations: Existence, uniqueness and stability.

Karatzas I, Shreve SE (1998) Brownian motion. *Brownian Motion and Stochastic Calculus*, 47–127 (Springer).

Kushner HJ, Vazquez-Abad FJ (1996) Stochastic approximation methods for systems over an infinite horizon. *SIAM Journal on Control and Optimization* 34(2):712–756.

Kushner HJ, Yin GG (2003) *Stochastic Approximation and Recursive Algorithms and Applications*, volume 35 (Springer-Verlag, New York, NY).

Nowakowski BD (2013) *On Multi-parameter Semimartingales, Their Integrals and Weak Convergence.*

Sagitov S (2013) Weak convergence of probability measures. *Chalmers University of Technology and Gothenburg University* .

Simmons GF (2016) *Differential equations with applications and historical notes* (CRC Press).