COVER: A CLUSTER-BASED VARIANCE REDUCED METHOD FOR ONLINE LEARNING

Kun Yuan, Bicheng Ying*, and Ali H. Sayed†*

*Department of Electrical and Computer Engineering, University of California, Los Angeles
†School of Engineering, École Polytechnique Fédérale de Lausanne, Switzerland

ABSTRACT

In this paper, we develop a stochastic-gradient learning algorithm for situations involving streaming data that arise from an underlying clustered structure. In such settings, the variance of gradient noise can be decomposed into the in-cluster variance $\sigma_{\rm in}^2$ plus the between-cluster variance $\sigma_{\rm bet}^2$. We develop a cluster-based online variance-reduced method (COVER) to eliminate $\sigma_{\rm bet}^2$ and improve the MSD performance of stochastic-gradient descent (SGD) to the order of $O(\sigma_{\rm in}^2)$. We establish the convergence property of COVER and derive a tight closed-form mean-square deviation (MSD) performance expression. Our simulations illustrate the improved performance of COVER in terms of steady-state performance.

Index Terms— Online learning, Streaming data, Internal structure, Variance reduction, SGD, SAGA.

1. INTRODUCTION AND RELATED WORKS

Stochastic optimization focuses on the problem of optimizing the expectation of a loss function, written as

$$w^\star = \arg\min_{w\in\mathbb{R}^M} \; J(w) \; \triangleq \; \mathbb{E}[Q(w;\boldsymbol{x})], \tag{1}$$

where $J(w)$ is a risk function, $\boldsymbol{x}$ is a random variable representing the data, and the expectation is over the distribution of $\boldsymbol{x}$. Problems of this kind are common in many contexts, including in several adaptation and machine learning formulations [1, 2].

When $J(w)$ is differentiable, one of the most popular methods to solve (1) is stochastic gradient descent (SGD) [3, 4, 5, 6]:

$$\boldsymbol{w}_i = \boldsymbol{w}_{i-1} - \mu\nabla Q(\boldsymbol{w}_{i-1};\boldsymbol{x}_i), \quad i\geq 0, \tag{2}$$

where $\boldsymbol{x}_i$ is the realization of $\boldsymbol{x}$ at iteration i, and μ is the positive step-size parameter. Throughout the paper, we assume μ is constant to endow SGD with desirable adaptive and tracking abilities. When $J(w)$ is strongly convex and $\nabla J(w)$ is Lipschitz continuous, the steady-state mean-square-deviation (MSD) performance of SGD is established in [3, 6] to be on the order of μ:

$$\limsup_{i\to\infty} \mathbb{E}\|w^\star - \boldsymbol{w}_i\|^2 = O\left(\mu\sigma^2\right), \tag{3}$$

where σ^2 is the variance of gradient noise.

In this paper we consider a new setting in which the data $\boldsymbol{x}$ appearing in (1) has some internal structure. By "internal structure" we mean that the distribution of $\boldsymbol{x}$ can be expressed as a mixture of distributions, or equivalently, data $\boldsymbol{x}$ can be grouped into various clusters. Data with such internal structure are abundant in practice. For example, in social media, the profiles of users can be categorized by the their communities, nationalities, religions, and beyond. In images or videos, the pixels tend to be well clustered into multiple segments. In natural language processing, documents can be grouped into various topics. As a typical example, it is identified in [7] that the Covtype dataset can be grouped into 1145 clusters.

This paper seeks to exploit the internal structure to improve SGD performance. When data can be grouped into smaller clusters, it is shown in Section 2.1 that the variance of gradient noise in SGD can be decomposed into

$$\sigma^2 = \sigma_{\rm in}^2 + \sigma_{\rm bet}^2 \tag{4}$$

where $\sigma_{\rm in}^2$ is the averaged in-cluster variance and $\sigma_{\rm bet}^2$ is the between-cluster variance. By employing the variance-reduction technique [8, 9, 10], we reach a cluster-based online variance-reduced method (COVER) that eliminates the between-cluster variance and achieves an improved steady-state MSD performance as

$$\limsup_{i\to\infty} \mathbb{E}\|w^\star - \boldsymbol{w}_i\|^2 = O\left(\mu\sigma_{\rm in}^2\right). \tag{5}$$

Expression 5 implies that when $\sigma_{\rm in}^2 \ll \sigma^2$, the proposed COVER method converges much more accurately than SGD.

Contribution. There are three main contributions in this paper. We propose a new COVER algorithm that exploits the internal structure of the data and attains an improved steady-state MSD performance. We also examine the convergence property of COVER and establish its stability range on step-sizes. Furthermore, we derive a tight closed-form expression for the steady-state MSD performance when μ is sufficiently small. This MSD expression is important for two reasons. First, by comparing the MSD expressions of SGD and COVER, we will conclude that COVER will be more accurate than SGD when the same step-size is employed. Second, the derived expression defines confidence levels about how well the iterate $\boldsymbol{w}_i$ approaches the global minimum $w^\star$ and provides crucial clues on how to choose the constant step-size μ.

Related Works in the Literature. Variance reduction is an important technique to improve the convergence of stochastic algorithms. There is an extensive research on variance reduction stochastic methods such as SVRG[9], SAGA[10], SAG[8], Finito[11], AVRG [12], and beyond. Most of these methods are aimed at solving empirical risk minimization (ERM) problems with finite-size datasets. Different from these works, COVER is an *online* variance-reduced method to solve problem (1) in the presence of *streaming* data.

There exist recent works [13, 14] on variance-reduced techniques that can handle streaming data. However, these methods employ *decaying* step-sizes to guarantee convergence. When these step-sizes approach zero, the tracking ability of the algorithms is lost. In contrast, the proposed COVER method employs a constant step-size to enable continuous adaptation and tracking. In addition, while works [13, 14] do not derive MSD performance expressions, this article derives a tight MSD result in order to characterize ac-

Email:{kunyuan,ybc}@ucla.edu and ali.sayed@epfl.ch. This work is supported in part by NSF grant CCF-1524250.

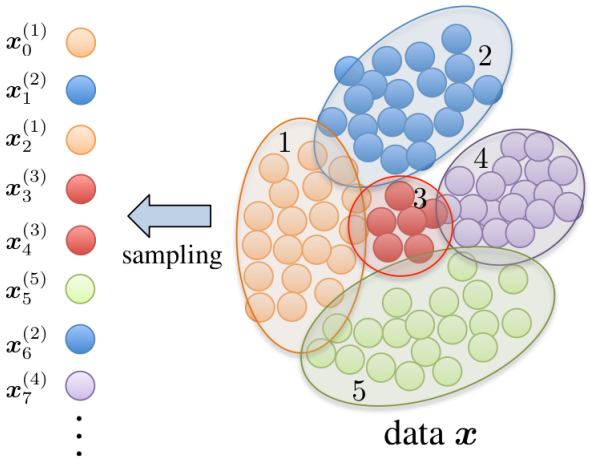

Fig. 1. An illustration of the internal structure in data $\boldsymbol{x}$ (the right part in the plot) and the way to generate data realization $\boldsymbol{x}_i^{(\boldsymbol{n}_i)}$ at iteration i. At iteration $i=4$, for example, the cluster index is first randomly generated (or sampled) and we assume $\boldsymbol{n}_i=3$. Then, a data realization $\boldsymbol{x}_4^{(3)}$ is generated according to distribution $f_3(x)$.

curately the dependence of MSD performance on the step-size parameter.

There exist also related works [15, 7] that exploit the internal structure of the data to improve the convergence rate of their respective learning algorithms. However, such as was the case with the earlier variance reduced methods, these techniques focus on ERM problems with *finite* data sets. One challenge in our contribution is to handle the internal structure for streaming data.

2. PROBLEM FORMULATION

In this section, we assume the probability distribution of the data variable $\boldsymbol{x}$ can be expressed as a mixture of N distributions, or equivalently, data $\boldsymbol{x}$ can be grouped into N smaller clusters. At iteration i, a data realization $\boldsymbol{x}_i$ is generated as follows. First, a random distribution index $\boldsymbol{n}_i=n$ is randomly generated with probability p_n (where $\sum_{n=1}^N p_n=1$). This index variable $\boldsymbol{n}_i$ indicates which distribution/cluster (we use distribution and cluster synonymously in this paper) that $\boldsymbol{x}_i$ arises from. Subsequently, the realization $\boldsymbol{x}_i$ is generated according to the n-th distribution $f_n(x)$. To emphasize that the data has internal structure, we rewrite variable $\boldsymbol{x}$ in problem (1) as $\boldsymbol{x}^{(\boldsymbol{n})}$ where the superscript $(\boldsymbol{n})$ indicates the cluster index $\boldsymbol{x}$ arises from. Also, we rewrite the i-th realization $\boldsymbol{x}_i$ as $\boldsymbol{x}_i^{(\boldsymbol{n}_i)}$. The process to generate $\boldsymbol{x}_i^{(\boldsymbol{n}_i)}$ is illustrated in Fig. 1. With such generation process, we express the probability distribution of $\boldsymbol{x}^{(\boldsymbol{n})}$ as

$$f(x)=\sum_{n=1}^N f_n(x)\mathbb{P}(\boldsymbol{n}=n)=\sum_{n=1}^N p_n f_n(x). \tag{6}$$

Since each $f_n(x)$ is generally unknown, the overall distribution $f(x)$ is also unknown. However, the mixture expression (i.e., the finite sum structure) of $f(x)$ in (6) enables the use of variance reduction techniques.

With (6), we rewrite the cost function in (1) as

$$J(w)=\mathbb{E}Q(w;\boldsymbol{x}^{(\boldsymbol{n})})=\int Q(w;x)f(x)dx\overset{(6)}{=}\sum_{n=1}^N p_n J_n(w) \tag{7}$$

where we define

$$J_n(w)\ \triangleq\ \mathbb{E}\,Q[w;\boldsymbol{x}^{(n)}]=\int Q(w;x)f_n(x)dx \tag{8}$$

as the n-th cluster risk function. With (7) and (8), problem (1) becomes

$$w^\star=\arg\min_{w\in\mathbb{R}^M}\sum_{n=1}^N p_n J_n(w). \tag{9}$$

If $p_1=p_2=\cdots=p_N$, problem (9) reduces to the problem formulation in [13, 14]. The finite-sum structure appearing in (9) makes it possible to employ variance reduction techniques. Throughout this paper, we introduce the following assumption:

Assumption 1 (CONDITIONS ON RISK FUNCTIONS)**.** *The loss function $Q(w;\boldsymbol{x}^{(\boldsymbol{n})})$ is δ-Lipschitz differentiable with respect to w, i.e., it holds for any w_1 and w_2 that*

$$\|\nabla Q(w_1;\boldsymbol{x}^{(\boldsymbol{n})})-\nabla Q(w_2;\boldsymbol{x}^{(\boldsymbol{n})})\|\le\delta\|w_1-w_2\|. \tag{10}$$

Moreover, we assume each cluster risk $J_n(w)$ is ν-strongly convex, i.e., it holds for any w_1 and w_2 that

$$\big(\nabla J_n(w_1)-\nabla J_n(w_2)\big)^{\mathsf{T}}(w_1-w_2)\ge\nu\|w_1-w_2\|^2. \tag{11}$$

The constants δ and ν are all positive. ■

2.1. The Variance of Gradient Noise in standard SGD

When data $\boldsymbol{x}_i^{(\boldsymbol{n}_i)}$ is observed, standard SGD takes the form

$$\boldsymbol{w}_i=\boldsymbol{w}_{i-1}-\mu\nabla Q(\boldsymbol{w}_{i-1};\boldsymbol{x}_i^{(\boldsymbol{n}_i)}). \tag{12}$$

where the gradient noise is defined as

$$\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\ \triangleq\ \nabla Q(\boldsymbol{w}_{i-1};\boldsymbol{x}_i^{(\boldsymbol{n}_i)})-\nabla J(\boldsymbol{w}_{i-1}). \tag{13}$$

This gradient noise is affected by two random variables: the random cluster variable $\boldsymbol{n}_i$ and the data realization $\boldsymbol{x}_i^{(n)}$ within a determined cluster n. To examine the variance of such gradient noise, we first introduce the filtration as

$$\boldsymbol{\mathcal{F}}_{i-1}\ \triangleq\ \{\boldsymbol{w}_{-1},\boldsymbol{w}_0,\boldsymbol{w}_1,\cdots,\boldsymbol{w}_{i-1}\}, \tag{14}$$

and the gradient noise within cluster n as

$$\boldsymbol{s}_i^{(n)}(\boldsymbol{w}_{i-1})\ \triangleq\ \nabla Q(\boldsymbol{w}_{i-1};\boldsymbol{x}_i^{(n)})-\nabla J_n(\boldsymbol{w}_{i-1}). \tag{15}$$

We then introduce the following assumption on $\boldsymbol{s}_i^{(n)}(\boldsymbol{w}_{i-1})$.

Assumption 2 (GRADIENT NOISE CONDITIONS)**.** *It is assumed for any $n=1,\cdots,N$ that*

$$\mathbb{E}[\boldsymbol{s}_i^{(n)}(\boldsymbol{w}_{i-1})|\boldsymbol{\mathcal{F}}_{i-1}]=0, \tag{16}$$

$$\mathbb{E}[\|\boldsymbol{s}_i^{(n)}(\boldsymbol{w}_{i-1})\|^2|\boldsymbol{\mathcal{F}}_{i-1}]\le\beta_n^2\|\widetilde{\boldsymbol{w}}_{i-1}\|^2+\sigma_n^2, \tag{17}$$

where $\widetilde{\boldsymbol{w}}_i\ \triangleq\ w^\star-\boldsymbol{w}_i$, constants β_n and σ_n are nonnegative, and σ_n^2 is referred to as the magnitude of gradient noise in cluster n.

With Assumption 2, and by following the analysis argument of Lemma 1 in [16], it is easy to reach the variance of SGD gradient noise $\boldsymbol{s}_i$ at $w^\star$ as

$$\mathbb{E}[\|\boldsymbol{s}_i(w^\star)\|^2|\boldsymbol{\mathcal{F}}_{i-1}]\le\sum_{n=1}^N p_n\sigma_n^2+\sum_{n=1}^N p_n\|\nabla J_n(w^\star)\|^2 \tag{18}$$

We define the first term as the averaged in-cluster variance $\sigma_{\rm in}^2$ and the second term as the between-cluster variance $\sigma_{\rm bet}^2$. By following the analysis in the proof of Lemma 3.1 in [6], we conclude that the steady-state MSD performance of the SGD recursion (12) is

$$\limsup_{i\to\infty}\mathbb{E}\|w^\star-\boldsymbol{w}_i\|^2=O\left(\mu(\sigma_{\rm in}^2+\sigma_{\rm bet}^2)\right), \tag{19}$$

3. VARIANCE-REDUCED ONLINE METHOD

In this section, we assume the data has an explicit internal structure.

Assumption 3 (EXPLICIT INTERNAL STRUCTURE)**.** *We assume that each time a realization $\boldsymbol{x}_i$ is observed, the value of its cluster index $\boldsymbol{n}_i$ is also observed. Furthermore, we assume $\mathbb{P}(\boldsymbol{n}=n)=p_n$ is also known in advance.* ■

Assumption 3 is not restrictive since we do not assume knowledge of the probability distribution of each cluster, $f_n(x)$. To correct the between-cluster variance, we propose to run SGD with a variance-reduced stochastic gradient. Inspired by the SAGA algorithm [10], we maintain N auxiliary variables $\{\boldsymbol{g}_i^{(n)}\}_{n=1}^N$ where each variable $\boldsymbol{g}_i^{(n)}$ approximates n-th cluster risk $\nabla J_n(\boldsymbol{w}_i)$. For each iteration, we propose the main recursion in COVER as

$$\boldsymbol{w}_i=\boldsymbol{w}_{i-1}-\mu\underbrace{\left(\nabla Q(\boldsymbol{w}_{i-1};\boldsymbol{x}_i^{(\boldsymbol{n}_i)})-\boldsymbol{g}_{i-1}^{(\boldsymbol{n}_i)}+\sum_{n=1}^N p_n\,\boldsymbol{g}_{i-1}^{(n)}\right)}_{\text{variance-reduced gradient}} \tag{20}$$

We establish in the future Lemma 1 that the improved gradient has less gradient noise than SGD. After recursion (20), only $\boldsymbol{g}_i^{(\boldsymbol{n}_i)}$ will be updated according to (22) and hence the averaged gradient $\bar{\boldsymbol{g}}_i=\sum_{n=1}^N p_n\boldsymbol{g}_i^{(n)}$ can be updated in a recursive manner. The complete COVER algorithm is listed in Algorithm 1. Variables α and α_n are relaxation coefficients, and quantity $p_{\min}=\min\{p_1,\cdots,p_N\}$.

Algorithm 1. The COVER method

Initialization: $\boldsymbol{w}_0=0,\ \bar{\boldsymbol{g}}_0=0,\ \alpha\in(0,p_{\min}),\ \boldsymbol{g}_0^{(n)}=0,\ \alpha_n=\alpha/p_n$ for any $n=1,\ldots,N$.

Repeat $i=1,2\ldots$, until convergence:

get the cluster index $\boldsymbol{n}_i$ in which data $\boldsymbol{x}_i$ arises from;

update $\boldsymbol{w}_i$, $\{\boldsymbol{g}_i^{(n)}\}_{n=1}^N$ and $\bar{\boldsymbol{g}}_i$ as follows:

$$\boldsymbol{w}_i=\boldsymbol{w}_{i-1}-\mu\Big(\nabla Q(\boldsymbol{w}_{i-1};\boldsymbol{x}_i^{(\boldsymbol{n}_i)})-\boldsymbol{g}_{i-1}^{(\boldsymbol{n}_i)}+\bar{\boldsymbol{g}}_{i-1}\Big) \tag{21}$$

$$\boldsymbol{g}_i^{(n)}=\begin{cases}(1-\alpha_n)\,\boldsymbol{g}_{i-1}^{(n)}+\alpha_n\nabla Q(\boldsymbol{w}_{i-1};\boldsymbol{x}_i^{(n)}) & \text{if } \boldsymbol{n}_i=n\\ \boldsymbol{g}_{i-1}^{(n)} & \text{if } \boldsymbol{n}_i\neq n\end{cases} \tag{22}$$

$$\bar{\boldsymbol{g}}_i=\bar{\boldsymbol{g}}_{i-1}-\alpha\big(\boldsymbol{g}_{i-1}^{(\boldsymbol{n}_i)}-\nabla Q(\boldsymbol{w}_{i-1};\boldsymbol{x}_i^{(\boldsymbol{n}_i)})\big) \tag{23}$$

End

4. CONVERGENCE PROPERTY

In this section we establish the convergence property of the proposed COVER algorithm. We first introduce

$$\boldsymbol{v}_i(\boldsymbol{w}_{i-1})=\nabla Q(\boldsymbol{w}_{i-1};\boldsymbol{x}_i^{(\boldsymbol{n}_i)})-\boldsymbol{g}_{i-1}^{(\boldsymbol{n}_i)}+\bar{\boldsymbol{g}}_{i-1}-\nabla J(\boldsymbol{w}_{i-1}) \tag{24}$$

as the gradient noise in COVER.

Lemma 1 (COVER GRADIENT NOISE PROPERTIES)**.** *Under Assumptions 1–3, it holds that*

$$\mathbb{E}[\boldsymbol{v}_i(\boldsymbol{w}_{i-1})|\boldsymbol{\mathcal{F}}_{i-1}]=0, \tag{25}$$

$$\mathbb{E}[\|\boldsymbol{v}_i(\boldsymbol{w}_{i-1})\|^2|\boldsymbol{\mathcal{F}}_{i-1}]\le(\bar\beta^2+2\delta^2)\|\widetilde{\boldsymbol{w}}_{i-1}\|^2+\bar\sigma^2+2\sum_{n=1}^N p_n\|\boldsymbol{g}_{i-1}^{(n)}-\nabla J_n(w^\star)\|^2, \tag{26}$$

where

$$\bar\beta^2\ \triangleq\ \sum_{n=1}^N p_n\beta_n^2,\qquad \bar\sigma^2\ \triangleq\ \sum_{n=1}^N p_n\sigma_n^2. \tag{27}$$

Proof. The proof is omitted due to space limitation. ■

By comparing (18) with (26), it is observed that COVER improves the term $\sum_{n=1}^N p_n\|\nabla J_n(w^\star)\|^2$ to $\sum_{n=1}^N p_n\|\boldsymbol{g}_{i-1}^{(n)}-\nabla J_n(w^\star)\|^2$. As iteration $i\to\infty$ and $\boldsymbol{g}_i^{(n)}\to\nabla J_n(w^\star)$, the term $\sum_{n=1}^N p_n\|\boldsymbol{g}_{i-1}^{(n)}-\nabla J_n(w^\star)\|^2$ vanishes to zero and hence the between-cluster variance is eliminated. This is the intuition why COVER converges more accurately than SGD.

Lemma 2. *Consider the COVER recursion in Algorithm 1. Under Assumptions 1–3, the following two inequalities hold.*

$$\mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^2\le\left(1-2\mu\nu+\mu^2(3\delta^2+\bar\beta^2)\right)\mathbb{E}\|\widetilde{\boldsymbol{w}}_{i-1}\|^2+2\mu^2\sum_{n=1}^N p_n\mathbb{E}\|\boldsymbol{g}_{i-1}^{(n)}-\nabla J_n(w^\star)\|^2+\mu^2\bar\sigma^2, \tag{28}$$

$$\sum_{n=1}^N p_n\mathbb{E}\|\boldsymbol{g}_i^{(n)}-\nabla J_n(w^\star)\|^2\le(1-\alpha)\sum_{n=1}^N p_n\mathbb{E}\|\boldsymbol{g}_{i-1}^{(n)}-\nabla J_n(w^\star)\|^2+\alpha\left(\delta^2+\bar\beta^2\right)\mathbb{E}\|\widetilde{\boldsymbol{w}}_{i-1}\|^2+\alpha\bar\sigma^2. \tag{29}$$

Proof. The proof is omitted due to space limitation. ■

Theorem 1 (STABILITY CONDITION)**.** *Consider $\boldsymbol{w}_i$ generated by the COVER algorithm. Under Assumptions 1–3, if the step-size μ satisfies*

$$\mu\le\min\left\{\frac{\nu}{6(\delta^2+\bar\beta^2)},\frac{\alpha}{6\nu}\right\} \tag{30}$$

where $\alpha\in(0,p_{\min})$, it holds that

$$\mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^2+\gamma\boldsymbol{G}_i\le(1-\mu\nu)\big(\mathbb{E}\|\widetilde{\boldsymbol{w}}_{i-1}\|^2+\gamma\boldsymbol{G}_{i-1}\big)+4\mu^2\bar\sigma^2 \tag{31}$$

where $\boldsymbol{G}_i=\sum_{n=1}^N p_n\mathbb{E}\|\boldsymbol{g}_i^{(n)}-\nabla J_n(w^\star)\|^2$ and $\gamma=3\mu^2/\alpha$.

Proof. With notation $\boldsymbol{G}_i$, recursion (29) can be simplified to

$$\boldsymbol{G}_i\le(1-\alpha)\boldsymbol{G}_{i-1}+\alpha\left(\delta^2+\bar\beta^2\right)\mathbb{E}\|\widetilde{\boldsymbol{w}}_{i-1}\|^2+\alpha\bar\sigma^2. \tag{32}$$

Combining (28) and (32), it holds that

$$\begin{aligned}
&\mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^2+\gamma\,\boldsymbol{G}_i\\
&\le\left[1-2\mu\nu+\left(3\mu^2+\gamma\alpha\right)\delta^2+\left(\mu^2+\gamma\alpha\right)\bar\beta^2\right]\mathbb{E}\|\widetilde{\boldsymbol{w}}_{i-1}\|^2\\
&\quad+\left(2\mu^2+\gamma(1-\alpha)\right)\boldsymbol{G}_{i-1}+\left(\mu^2+\gamma\alpha\right)\bar\sigma^2\\
&\le\left[1-2\mu\nu+\left(3\mu^2+\gamma\alpha\right)\left(\delta^2+\bar\beta^2\right)\right]\mathbb{E}\|\widetilde{\boldsymbol{w}}_{i-1}\|^2\\
&\quad+\left(2\mu^2+\gamma(1-\alpha)\right)\boldsymbol{G}_{i-1}+\left(\mu^2+\gamma\alpha\right)\bar\sigma^2\\
&=\left[1-2\mu\nu+\left(3\mu^2+\gamma\alpha\right)\left(\delta^2+\bar\beta^2\right)\right]\cdot\\
&\quad\left(\mathbb{E}\|\widetilde{\boldsymbol{w}}_{i-1}\|^2+\frac{\left(2\mu^2+\gamma(1-\alpha)\right)}{1-2\mu\nu+\left(3\mu^2+\gamma\alpha\right)\left(\delta^2+\bar\beta^2\right)}\boldsymbol{G}_{i-1}\right)\\
&\quad+\left(\mu^2+\gamma\alpha\right)\bar\sigma^2\\
&\le\left[1-2\mu\nu+\left(3\mu^2+\gamma\alpha\right)\left(\delta^2+\bar\beta^2\right)\right]\cdot\\
&\quad\left(\mathbb{E}\|\widetilde{\boldsymbol{w}}_{i-1}\|^2+\frac{2\mu^2+\gamma(1-\alpha)}{1-2\mu\nu}\boldsymbol{G}_{i-1}\right)+\left(\mu^2+\gamma\alpha\right)\bar\sigma^2
\end{aligned} \tag{33}$$

Since μ satisfies (30), we have $\mu \leq \alpha/6\nu$ which implies that $\alpha - 2\mu\nu > 0$. This fact, along with the definition of γ, implies that

$$\gamma = \frac{3\mu^2}{\alpha} \geq \frac{2\mu^2}{\alpha - 2\mu\nu} \Longleftrightarrow \gamma \geq \frac{2\mu^2 + \gamma(1-\alpha)}{1 - 2\mu\nu} \tag{34}$$

Also, since μ satisfies

$$\mu \leq \frac{\nu}{6(\delta^2 + \bar{\beta}^2)}, \tag{35}$$

we have

$$1 - 2\mu\nu + \left(3\mu^2 + \gamma\alpha\right)\left(\delta^2 + \bar{\beta}^2\right) \leq 1 - \mu\nu \tag{36}$$

With the relations (34) and (36), inequality (33) becomes

$$\begin{aligned}\mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^2 + \gamma\,\boldsymbol{G}_i \leq (1-\mu\nu)\left(\mathbb{E}\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 + \gamma\,\boldsymbol{G}_{i-1}\right) \\ + \left(\mu^2 + \gamma\alpha\right)\bar{\sigma}^2.\end{aligned} \tag{37}$$

■

With (31), it is easy to verify that

$$\begin{aligned}\mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^2 &\leq \mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^2 + \gamma\,\boldsymbol{G}_i \\ &\leq (1-\mu\nu)^i\left(\mathbb{E}\|\widetilde{\boldsymbol{w}}_0\|^2 + \gamma\,\boldsymbol{G}_0\right) + \frac{4\mu\bar{\sigma}^2}{\nu}.\end{aligned} \tag{38}$$

Recalling that $\widetilde{\boldsymbol{w}}_i = w^\star - \boldsymbol{w}_i$, inequality (38) implies that

$$\limsup_{i\to\infty} \mathbb{E}\|w^\star - \boldsymbol{w}_i\|^2 = O\left(\mu\bar{\sigma}^2\right) = O\left(\mu\sigma_{\text{in}}^2\right) \tag{39}$$

since $\sigma_{\text{in}}^2 = \bar{\sigma}^2 = \sum_{n=1}^N p_n\sigma_{\text{in}}^2$. Comparing (39) and (19), it is observed that COVER eliminates the between-cluster variance σ_{bet}^2 asymptotically. When $\sigma_{\text{in}}^2 \ll \sigma_{\text{in}}^2 + \sigma_{\text{bet}}^2$, we can expect COVER to have a much better performance than SGD.

5. STEADY-STATE PERFORMANCE

Relation (39) establishes a rough asymptotic upper bound for $\mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^2$. In this section we derive a closed-form expression to characterize the MSD performance for COVER more accurately. To this end, we introduce

$$R_s^{(n)} \triangleq \lim_{i\to\infty} \mathbb{E}[\boldsymbol{s}_i^{(n)}(w^\star)\boldsymbol{s}_i^{(n)}(w^\star)^{\mathsf{T}}] \tag{40}$$

as the limiting covariance matrix of the gradient noise in the n-th cluster. We also define $\bar{R}_s \triangleq \sum_{n=1}^N p_n R_s^{(n)}$ as the averaged limiting covariance matrix.

Assumption 4 (HESSIAN IS LIPSCHITZ CONTINUOUS)**.** *It is assumed that each cluster risk function $J_n(w)$, where $n = 1, \cdots, N$, is twice-differentiable and has a Lipschitz continuous Hessian matrix, i.e., there exists a constant η such that*

$$\|\nabla^2 J_n(w_1) - \nabla^2 J_n(w_2)\|^2 \leq \eta^2\|w_1 - w_2\|^2. \tag{41}$$

Theorem 2 (MSD EXPRESSION)**.** *Under Assumptions 1–4, when step-size is sufficiently small, the MSD expression for the COVER algorithm is*

$$\text{MSD}_{\text{cover}} = \limsup_{i\to\infty} \mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^2 = \frac{\mu}{2}\text{Tr}(H^{-1}\bar{R}_s), \tag{42}$$

where H is the Hessian of the cost function $J(w)$.

Proof. The proof is omitted due to space limitations. ■

For comparison purposes, recall in [16] that the MSD expression for SGD recursion (12) is

$$\text{MSD}_{\text{sgd}} = \frac{\mu}{2}\text{Tr}(H^{-1}(\bar{R}_s + R_b)). \tag{43}$$

where $R_b = \sum_{n=1}^N p_n \nabla J_n(w^\star)\nabla J_n(w^\star)^{\mathsf{T}}$. Comparing (42) and (43), we find it always holds that $\text{MSD}_{\text{cover}} < \text{MSD}_{\text{sgd}}$.

6. SIMULATION

We consider the application of the proposed COVER method in data augmentation. In practice, there are scenarios in which the amount of the available data is not enough for learning purposes. A common technique is to shift, rotate, or add random noise to the original data and augment the training dataset. Suppose there are N original data in the training set. Each time a data x_n is to be sampled, a perturbed realization of x_n, rather than the original x_n itself, will be the value that is used. By doing so, one can augment the dataset, reduce the overfitting, and hence get a more robust estimator. In this scenario, the whole augmented dataset can be infinitely large and the data has explicit internal structure – we regard each data realization arising from the same original data as one cluster. Furthermore, since we actively sample data from each cluster rather than receive data passively, the cluster index where data x belongs to is explicit.

In particular, we consider a binary classification task with digits 0 and 1 in MNIST dataset. We randomly pick up 1000 training images in the digit 0 and 1 classes and let them be the original dataset. Each image is vectorized into dimension $\mathbb{R}^{784}$ and its ℓ_2-norm is normalized to 1. Next, we add Gaussian noise with variance 0.1 to each original image and augment these 1000 training data to a total of 1000000. We then solve the problem with the following regularized logistic regression cost function

$$J(w) = \frac{\rho}{2}\|w\|^2 + \mathbb{E}\ln\left(1 + \exp(-\boldsymbol{\gamma}\boldsymbol{h}^{\mathsf{T}}w)\right). \tag{44}$$

We compare the proposed COVER method with SGD, S-MISO[13] and S-SAGA[14]. Note that both S-MISO and S-SAGA have been proposed in the literature with decaying step-size. To allow for a fair comparison, we adjust them to operate with a constant step-size. The step-sizes are carefully tuned to reach the best convergence performance. Also, we notice that S-SAGA is a special form of COVER when we let $\alpha = p_n = 1/N$ and $\alpha_n = 1$ in recursion (22). All these four algorithms are listed in Fig.2. By exploiting the internal structure, S-SAGA, S-MISO and COVER converge more accurately than SGD. It is observed that S-MISO is slightly better than S-SAGA, but both of them are worse than COVER. This indicates the importance of the relaxation step in (22). Furthermore, our derived MSD expression (42) shown with magenta dash line, matches very well with the steady-state MSD performance of the COVER method.

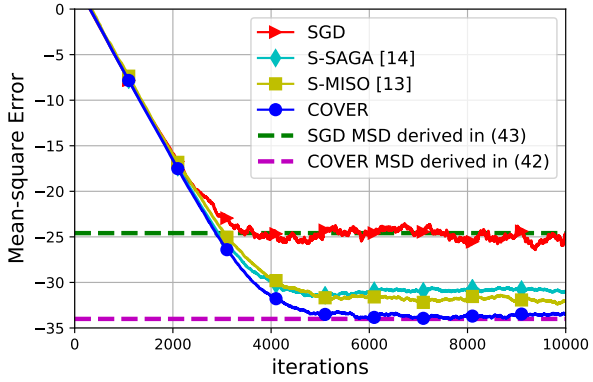

Fig. 2. Convergence comparison between SGD, S-SAGA, S-MISO and COVER.

7. REFERENCES

[1] A. H. Sayed, *Adaptive Filters*, Wiley, NY, 2008.

[2] S. Theodoridis, *Machine Learning: A Bayesian and Optimization Perspective*, Academic Press, NY, 2015.

[3] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proc. International Conference on Computational Statistics* (COMPSTAT), pp. 177–186. Springer, Paris, 2010.

[4] T. Zhang, "Solving large scale linear prediction problems using stochastic gradient descent algorithms," in *Proc. International Conference on Machine Learning* (ICML), Alberta, Canada, 2004, pp. 116 – 125.

[5] A. H. Sayed, "Adaptive networks," *Proceedings of the IEEE*, vol. 102, no. 4, pp. 460–497, Apr. 2014.

[6] A. H. Sayed, "Adaptation, learning, and optimization over networks," *Foundations and Trends in Machine Learning*, vol. 7, no. 4-5, pp. 311–801, Jul. 2014.

[7] Z. Allen-Zhu, Y. Yuan, and K. Sridharan, "Exploiting the structure: Stochastic gradient methods using raw clusters," in *Proc. Advances in Neural Information Processing Systems* (NIPS), Barcelona, Spain, 2016, pp. 1642–1650.

[8] M. Schmidt, N. Le Roux, and F. Bach, "Minimizing finite sums with the stochastic average gradient," *Mathematical Programming*, vol. 162, no. 1, pp. 83–112, Mar. 2017.

[9] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," in *Proc. Advances in Neural Information Processing Systems* (NIPS), Lake Tahoe, NV, 2013, pp. 315–323.

[10] A. Defazio, F. Bach, and S. Lacoste-Julien, "SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives," in *Proc. Advances in Neural Information Processing Systems* (NIPS), Montréal, Canada, 2014, pp. 1646–1654.

[11] A. Defazio, T. Caetano, and J. Domke, "Finito: A faster, permutable incremental gradient method for big data problems," in *Proc. International Conference on Machine Learning* (ICML), Beijing, China, 2014, pp. 1125–1133.

[12] B. Ying, K. Yuan, and A. H. Sayed, "Convergence of variance-reduced stochastic learning under random reshuffling," *submitted for publication.* Available as *arXiv:1708.01383*, Aug. 2017.

[13] A. Bietti and J. Mairal, "Stochastic optimization with variance reduction for infinite datasets with finite sum structure," in *Proc. Advances in Neural Information Processing Systems* (NIPS), Long Beach, California, 2017, pp. 1622–1632.

[14] S. Zheng and J. T. Kwok, "Lightweight stochastic optimization for minimizing finite sums with infinite data," in *Proc. International Conference on Machine Learning* (ICML), Stockholm, Sweden, 2018, pp. 1 – 8.

[15] T. Hofmann, A. Lucchi, S. Lacoste-Julien, and B. McWilliams, "Variance reduced stochastic gradient descent with neighbors," in *Proc. Advances in Neural Information Processing Systems* (NIPS), Montréal, Canada, 2015, pp. 2305–2313.

[16] K. Yuan, B. Ying, S. Vlaski, and A. H. Sayed, "Stochastic gradient descent with finite samples sizes," in *Proc. IEEE International Workshop on Machine Learning for Signal Processing* (MLSP), Salerno, Italy, 2016, pp. 1–6.