This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2017.2705559, IEEE Transactions on Automatic Control

On Projected Stochastic Gradient Descent Algorithm with Weighted Averaging for Least Squares Regression

Kobi Cohen, Angelia Nedić and R. Srikant

***Abstract*— The problem of least squares regression of a d-dimensional unknown parameter is considered. A stochastic gradient descent based algorithm with weighted iterate-averaging that uses a single pass over the data is studied and its convergence rate is analyzed. We first consider a bounded constraint set of the unknown parameter. Under some standard regularity assumptions, we provide an explicit $O(1/k)$ upper bound on the convergence rate, depending on the variance (due to the additive noise in the measurements) and the size of the constraint set. We show that the variance term dominates the error and decreases with rate $1/k$, while the term which is related to the size of the constraint set decreases with rate $\log k/k^2$. We then compare the asymptotic ratio ρ between the convergence rate of the proposed scheme and the empirical risk minimizer (ERM) as the number of iterations approaches infinity. We show that $\rho \leq 4$ for all $d \geq 1$ when the random entries of the sensing vector are uncorrelated and identically distributed. We further improve the upper bound by showing that $\rho \leq 4/3$ for the case of $d = 1$ and unbounded parameter set when the random sensing entries are equal across time. Simulation results demonstrate strong performance of the algorithm as compared to existing methods, and coincide with $\rho \leq 4/3$ even for large d in practice.**

***Index Terms*—** Convex optimization, projected stochastic gradient descent, weighted averaging, empirical risk minimizer.

I. Introduction

For large-scale optimization problems, it is often desirable to minimize an unknown objective under computational constraints. Stochastic Gradient Descent (SGD) is a popular optimization method in a variety of machine learning tasks when dealing with very large data or with data streams. Specifically, instead of computing the true gradient (which is often computationally expensive) as in a standard gradient descent algorithm, in SGD-based methods the gradient is approximated by a single (or few) sample at each iteration. Using stochastic approximation analysis, it has been shown that SGD converges almost surely to a global minimum when the objective function is convex (otherwise it converges to a local minimum) under an appropriate learning rate and some regularity conditions [2].

In this paper, we consider the problem of least mean squares regression, in which a d-dimensional unknown parameter is desired to be estimated from streaming noisy measurements. Specifically, let $\boldsymbol{x}$, y be random variables with values in $\mathbb{R}^d$, and $\mathbb{R}$, respectively, and let $\Omega \subseteq \mathbb{R}^d$ be a compact convex constraint set for the unknown parameter. It is desired to minimize the expected least squares loss:

$$\min_{\boldsymbol{\omega}} E\left[||\boldsymbol{x}^T\boldsymbol{\omega} - y||^2\right] \quad \text{subject to } \boldsymbol{\omega} \in \Omega \subseteq \mathbb{R}^d \qquad (1)$$

from the samples stream $(\boldsymbol{x}_k, y_k)$ at times $k = 1, 2, ...$ Motivated by recent studies on accelerated methods of SGD-based algorithms, we focus on a projected SGD method with weighted iterate-averaging to solve (1).

A. Main Results

Solving (1) directly is computationally inefficient since it requires high storage memory for the entire data and high computational complexity. Thus, our goal is to solve (1) efficiently so that the running time and space usage are small. Motivated by recent studies showing that using averaging of the estimated parameter accelerates the convergence of SGD-based algorithms, we analyze a Projected SGD with Weighted Averaging (PSGD-WA) algorithm for solving (1). Specifically, a projected SGD iterates are computed at each time k, where averaged iterates are computed as byproducts of the algorithm (but not used in the construction of the PSGD iterates). The averaging weights are specified in terms of the step-sizes that the algorithm uses such that recent measurements are given higher weights (see Section III for details). Our main results are as follows: i) We consider a bounded constraint set of the unknown parameter and propose a PSGD-WA algorithm that requires a single pass over the data. The proposed step size has a general form[1] of $c\frac{\gamma}{k+\gamma}$, where $c, \gamma > 0$ are tunable parameters; ii) in contrast to previous studies on PSGD algorithms with weighted averaging showing a general order $O(1/k)$ of the error rate, we provide an explicit finite sample upper bound on the error obtained by the proposed PSGD-WA algorithm, depending on the variance (due to the additive noise in the measurements) and the size of the constraint set. Our bounds hold for all k (i.e., finite-sample regime), which differentiate our work from many existing asymptotic results (as k approaches infinity) on SGD-based methods as well as unprojected SGD methods, and unweighted averaging methods (see Sec. I-B for a detailed discussion on related works). Specifically, we show that the variance term dominates the error and decreases with rate $1/k$, while the term which is related to the diameter of the constraint set decreases with rate $\log k/k^2$; iii) we compare the asymptotic ratio ρ between the convergence rate of the proposed PSGD-WA and the empirical risk minimizer (ERM) as the number of iterations approaches infinity (by contrast, a related analysis has been established recently by different streaming algorithms under different settings of unprojected SGD iterates and a constant step size. See Sec. I-B for details). The ERM algorithm computes the minimizer in the absence of computational constraints, by using the entire sampled data observed up to the current iteration. The theoretical performance bounds that we establish guarantee that $\rho \leq 4$ for all $d \geq 1$ when the random components of $\boldsymbol{x}$ are identically distributed and uncorrelated. We further improve the upper bound by showing that $\rho \leq 4/3$ for the case of $d = 1$ and $x_k = x$ for all k. Simulation results demonstrate strong performance of the algorithm as compared to existing methods, and coincide with $\rho \leq 4/3$ even for large d in practice.

Kobi Cohen is with the Department of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel, email: yakovsec@bgu.ac.il

Angelia Nedić is with the School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ 85287-5706. email: angelia.nedich@asu.edu

R. Srikant is with the Department of Electrical and Computer Engineering, and Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, IL 61801. email: rsrikant@illinois.edu

Part of this work was presented at the 41st IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2016 [1].

This work was supported by AFOSR MURI FA 9550–10–1–0573, ONR Grant N00014–13–1–003, NSF Grant CNS–1161404, and Cyber Security Research Center at Ben-Gurion University of the Negev.

[1] It should be noted that previous studies on PSGD algorithms with weighted averaging (see [3]–[5]) considered only a fixed form of the step size without tuning parameters.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2017.2705559, IEEE Transactions on Automatic Control

B. Related Work

SGD is a computationally efficient method for solving large-scale optimization problems when dealing with very large data or with data streams. Accelerating SGD-based algorithms using averaging techniques has been studied in past and more recent years in [3]–[29]. A number of studies have shown convergence rate of $1/k^2$ for a noiseless case (i.e., $\sigma^2 = 0$). In [3], Tseng has developed an accelerated SGD-based algorithm with iterate-averaging that achieves convergence rate of $1/k^2$ for problems where the objective function has Lipschitz continuous gradients. This rate is known to be the best in the class of convex functions with Lipshitz gradients [14], for which the first fast algorithm was originally constructed by Nesterov [7] for unconstrained problems, and was extended recently by Beck and Teboulle in [19] to a larger class of problems. Ghadimi and Lan used averaging in [24] to develop an algorithm that has the rate $1/k^2$ when the objective function has Lipschitz continuous gradients, and rate $1/k$ when the objective function is strongly convex. Juditsky et al. [16] considered a mirror-descent algorithm with averaging to construct aggregate estimators with the best achievable learning rate. Averaging techniques for non-smooth optimization have been investigated by Shamir and Zhang in [26]. Averaging techniques for the mirror-descent algorithm for stochastic problems involving the sum of a smooth objective and a nonsmooth objective function have been studied by Lan in [25]. Other related works are concerned with iterate-averaging for best achievable rate of stochastic subgradients methods [21], [23], as well as gradient-averaging [13], [15], [17], [18], [20], [22], [27], [30] and a sort of momentum [31], [32], in which the algorithm uses a sort of weighting over previous gradients (instead of the iterate minimizer) in the construction of the algorithm. In [10], [33], asymptotic analysis has been established for averaging both states and observations. Averaging methods with feedback have been analyzed in [34].

The averaged iterates considered in this paper are not used in the construction of the PSGD iterates, but only computed as byproducts of them (see Section III for details). Such methods have been studied by Nemirovski and Yudin [6] for convex-concave saddle-point problems, by Polyak and Juditsky [9] for stochastic gradient approximations and by Polyak [12] for convex feasibility problems. In [9], an asymptotically optimal performance has been achieved. However, a finite sample analysis remained open. More recently, Lacoste-Julien et al. [4] used this averaging approach for a projected stochastic subgradient method to achieve $1/k$ convergence rate for strongly convex functions. Nedić and Lee [5] used a similar form of this scheme for a more general projected stochastic subgradient method using Bregman distances, which achieves $1/k$ convergence rate for strongly convex functions, and $1/\sqrt{k}$ convergence rate for general convex functions. Here we use a similar scheme with a more general form of step-sizes for least squares regression from noisy measurements and establish a fine-grained finite-sample analysis (see details in subsequent paragraph and Section I-A).

In this paper we focus on the testing error (i.e., the expected error on unseen data) of regression from noisy measurements, in which the convergence rate deteriorates (varies from $1/k$ to $1/\sqrt{k}$ per-iterate). While accelerating methods cannot be made faster, they have ability to produce estimates with low-variance, which attracted much interest in recent years [29], [35]–[37]. We focus on the strongly convex case, in which $O(1/k)$ is the best attainable convergence rate [36]. However, this convergence rate is only optimal in the limit of large samples, and in practice other non-dominant terms may come into play in the finite sample regime. In [36], Frostig et al. have developed a Streaming Stochastic Variance Reduced Gradient (Streaming SVRG) algorithm using a constant step size, inspired by the SVRG algorithm developed by Johnson and Zhang [35], and provided a finite sample analysis for a general strongly convex regression problems. They showed that the asymptotic ratio ρ between the convergence rate of the Streaming SVRG and the ERM algorithm approaches $\rho = 1$ as the number of iterations approaches infinity. However, achieving $\rho = 1$ requires the sample batch size to grow geometrically occasionally for gradient-computing, as well as setting the constant step size close to zero (which deteriorates performance in the finite sample regime). In [29], Defossez and Bach have developed a SGD algorithm using a constant step size with averaging for least mean squares regression, and provided a finite sample analysis. They showed that $\rho = 1$ as the constant step size is set close to zero, which deteriorates performance in the finite sample regime. In this paper, however, the proposed PSGD-WA algorithm uses decreasing step-sizes which can be large in the beginning of the algorithm and decrease as the number of iterations increases. The proposed PSGD-WA algorithm uses a weighted averaging of the estimates, by letting higher weights to recent measurements. We provide a finite sample analysis as well as an asymptotic upper bound $\rho \leq 4$ when $d \geq 1$ and $\rho \leq 4/3$ when $d = 1$. Note that our results do not require the sample batch size to grow geometrically occasionally as in [36] or setting small step-sizes in the beginning of the algorithm as in [29], [36]. Thus, the proposed PSGD-WA algorithm is expected to perform well in the non-asymptotic case in addition to the nice asymptotic property as illustrated by simulation results provided in Section V.

Notations: Throughout the paper, small letters denote scalars, boldface small letters denote column vectors, and boldface capital letters denote matrices. All vectors are column vectors. The term $\mathbf{z}^T$ denotes the conjugate transpose of the vector $\mathbf{z}$, and $||\cdot||$ denotes the Euclidean norm. The subscript k associated with a r.v. denotes a realization at time k.

II. Problem Statement

Let $\boldsymbol{x}$, y be random variables with values in $\mathbb{R}^d$, and $\mathbb{R}$, respectively. At each time k, we observe i.i.d. samples across time $(\boldsymbol{x}_k, y_k)$. We assume that $E\left[\boldsymbol{x}^T\boldsymbol{x}\right]$ is finite and we denote by $R_x = E\left[\boldsymbol{x}\boldsymbol{x}^T\right]$ the correlation matrix of $\boldsymbol{x}$. It is desired to minimize the expected least squares loss:

$$\min_{\boldsymbol{\omega}} E\left[||\boldsymbol{x}^T\boldsymbol{\omega} - y||^2\right] \quad \text{subject to } \boldsymbol{\omega} \in \Omega \subseteq \mathbb{R}^d \tag{2}$$

from the samples stream $(\boldsymbol{x}_k, y_k)$ at times $k = 1, 2, ...$ It is assumed that R_x is invertible (i.e., strongly convex case). We denote by μ the smallest eigenvalue of R_x, so that $\mu > 0$.

We denote the optimal solution of (2) by $\boldsymbol{\omega}^* \in \mathbb{R}^d$, and it is assumed that a decision maker knows that $\boldsymbol{\omega}^*$ lies in the interior of a convex constraint set $\Omega \subseteq \mathbb{R}^d$.

Let $f(\boldsymbol{\omega}) \triangleq E\left[||\boldsymbol{x}^T\boldsymbol{\omega} - y||^2\right]$ be the mean squares loss as a function of $\boldsymbol{\omega}$ (i.e., conditioned on $\boldsymbol{\omega}$, $f(\boldsymbol{\omega}) = E\left[||\boldsymbol{x}^T\boldsymbol{\omega} - y||^2|\boldsymbol{\omega}\right]$), and $f^* = f(\boldsymbol{\omega}^*) \in \mathbb{R}$ be the value at the minimum. The term $v_k = \boldsymbol{x}_k^T\boldsymbol{\omega}^* - y_k$ denotes the zero-mean additive noise with variance σ^2. The gradient of f at $\boldsymbol{\omega}$ is defined by $\nabla\boldsymbol{f}(\boldsymbol{\omega}) = E\left[2\boldsymbol{x}\left(\boldsymbol{x}^T\boldsymbol{\omega} - y\right)\right] = E\left[\boldsymbol{g}_k(\boldsymbol{\omega})\right]$, where $\boldsymbol{g}_k(\boldsymbol{\omega}) \triangleq 2\boldsymbol{x}_k\left(\boldsymbol{x}_k^T\boldsymbol{\omega} - y_k\right)$ is the estimate of the gradient at $\boldsymbol{\omega}$ based on a single sample at iteration[2] k. For convenience, we write $\nabla\boldsymbol{f}_k \triangleq \nabla\boldsymbol{f}(\boldsymbol{\omega}_k)$ and $\boldsymbol{g}_k \triangleq \boldsymbol{g}_k(\boldsymbol{\omega}_k)$ when referring to the gradients at $\boldsymbol{\omega}_k$, where $\boldsymbol{\omega}_k$ is the estimate of $\boldsymbol{\omega}$ at iteration k obtained by an iterative algorithm (see the next section for details). The error at the k^{th}

[2] When a few samples are available per iteration we estimate the gradient by averaging.

iteration is defined by $\boldsymbol{e}_k \triangleq \boldsymbol{\omega}_k - \boldsymbol{\omega}^*$. Note that

$$\begin{aligned}\boldsymbol{g}_k = 2\boldsymbol{x}_k\left(\boldsymbol{x}_k^T\boldsymbol{\omega}_k - y_k\right) &= 2\boldsymbol{x}_k\left(\boldsymbol{x}_k^T\boldsymbol{\omega}_k - \boldsymbol{x}_k^T\boldsymbol{\omega}^* + v_k\right)\\ &= 2\boldsymbol{x}_k\left(\boldsymbol{x}_k^T\boldsymbol{e}_k + v_k\right).\end{aligned} \qquad (3)$$

III. Projected Stochastic Gradient descent algorithm with Weighted Averaging

We investigate a Projected Stochastic Gradient descent algorithm with Weighted Averaging (PSGD-WA). According to PSGD-WA, we hold two estimates of $\boldsymbol{\omega}^*$ at each iteration, denoted by $\boldsymbol{\omega}_k, \bar{\boldsymbol{\omega}}_k$. The estimate $\boldsymbol{\omega}_k$ is computed at each iteration (say k), and $\bar{\boldsymbol{\omega}}_k$ is the weighted average estimate based on all estimates up to time k. Let λ_k be the step-size at time k, and assume that it diminishes with k. At iteration $k+1$ we compute the projected estimate of $\boldsymbol{\omega}^*$ based on the random measurements $(\boldsymbol{x}_k, y_k)$ and the last estimate $\boldsymbol{\omega}_k$:

$$\boldsymbol{\omega}_{k+1} = \arg\min_{\boldsymbol{\omega}\in\Omega}\left\{\lambda_k\boldsymbol{g}_k^T\cdot(\boldsymbol{\omega}-\boldsymbol{\omega}_k) + \frac{1}{2}||\boldsymbol{\omega}-\boldsymbol{\omega}_k||^2\right\} \;\forall k \geq 0, \qquad (4)$$

where $\boldsymbol{\omega}_0 \in \Omega$ is an initial estimate of $\boldsymbol{\omega}$ (possibly random). It can be verified that $\boldsymbol{\omega}_{k+1}$ projects the unconstrained gradient descent iterate $\boldsymbol{\omega}_k - \lambda_k\boldsymbol{g}_k$ into Ω. Similar to [4], [5], in addition to the estimate $\boldsymbol{\omega}_{k+1}$, we compute a weighted average estimate:

$$\bar{\boldsymbol{\omega}}_{k+1} = \sum_{i=0}^{k+1}\beta_{k+1,i}\boldsymbol{\omega}_i, \qquad (5)$$

where $\beta_{k,0}, \beta_{k,1}, ..., \beta_{k,k}$ are nonnegative scalars with the sum equal to 1. The weighted average estimate $\bar{\boldsymbol{\omega}}_k$ is computed based on the first k iterations. These convex weights will be defined in terms of the step size values $\lambda_0, \lambda_1, ..., \lambda_k$, and $\bar{\boldsymbol{\omega}}_k$ will be computed recursively (see (7) in Section III-A). In Section IV we will analyze the convergence rate of $\bar{\boldsymbol{\omega}}_k$ to the solution of (2).

A. Implementation and Complexity Discussion

The PSGD-WA algorithm is simple for implementation as compared to existing methods. Let

$$\beta_{k,i} = \frac{1/\alpha_i}{\sum_{r=0}^{k}1/\alpha_r}, \qquad (6)$$

where α_i is a decreasing sequence with i. At iteration k, the algorithm requires to store $\boldsymbol{\omega}_k$, the weighted average $\bar{\boldsymbol{\omega}}_{k-1}$ and the normalization term $S_{k-1} = \sum_{r=0}^{k-1}1/\alpha_r$. The weighted average $\bar{\boldsymbol{\omega}}_k$ can be updated recursively by computing $S_k = S_{k-1} + 1/\alpha_k$ and then by setting:

$$\bar{\boldsymbol{\omega}}_k = \frac{S_{k-1}}{S_k}\bar{\boldsymbol{\omega}}_{k-1} + \left(1 - \frac{S_{k-1}}{S_k}\right)\boldsymbol{\omega}_k. \qquad (7)$$

As a result, only $O(1)$ computations are required per iteration as needed by the classic SGD algorithm. Note that PSGD-WA does not require the sample batch size to grow as in [36]. The storage memory required by PSGD-WA is similar to that required by the average SGD with constant step size algorithm proposed in [29].

IV. Performance Analysis

In this section we analyze the performance of the algorithm when the constraint set Ω is bounded. Let $e_{max} = \sup_{\boldsymbol{\omega}\in\Omega}\left\{||\boldsymbol{\omega}-\boldsymbol{\omega}^*||^2\right\}$ be the maximal square error of any projected estimate of $\boldsymbol{\omega}^*$. Let $\mathcal{F}_{k-1} = \sigma\left\{\boldsymbol{\omega}_0, \boldsymbol{x}_0, y_0, \boldsymbol{x}_1, y_1, ..., \boldsymbol{x}_{k-1}, y_{k-1}\right\}$ be the filtration generated by the history of the algorithm starting at time 0 up to time $k-1$. Note that $\boldsymbol{\omega}_0, \boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_k$ are known once $\mathcal{F}_{k-1}$ is given.

Lemma 1: Assume that (4) is implemented. Then, for all $\boldsymbol{\omega} \in \Omega$ and $k \geq 0$, we have:

$$\begin{aligned}&\frac{1}{2}E\left[||\boldsymbol{\omega}_{k+1}-\boldsymbol{\omega}||^2|\mathcal{F}_{k-1}\right] + \lambda_k\nabla\boldsymbol{f}_k^T\cdot(\boldsymbol{\omega}_k-\boldsymbol{\omega})\\ &\leq \frac{1}{2}||\boldsymbol{\omega}_k-\boldsymbol{\omega}||^2 + 2\lambda_k^2E\left[||\boldsymbol{x}_k\boldsymbol{x}_k^T\boldsymbol{e}_k||^2|\mathcal{F}_{k-1}\right] + 2\lambda_k^2\sigma^2E\left[||\boldsymbol{x}_k||^2\right].\end{aligned} \qquad (8)$$

Proof: We first upper bound the term $\lambda_k\boldsymbol{g}_k^T\cdot(\boldsymbol{\omega}_{k+1}-\boldsymbol{\omega})$. Since $\boldsymbol{\omega}_{k+1}$ solves (4), we have:

$$\begin{aligned}&\nabla_{\boldsymbol{\omega}}q(\boldsymbol{\omega}_{k+1})^T(\boldsymbol{\omega}-\boldsymbol{\omega}_{k+1})\\ &= (\lambda_k\boldsymbol{g}_k + \boldsymbol{\omega}_{k+1} - \boldsymbol{\omega}_k)^T(\boldsymbol{\omega}-\boldsymbol{\omega}_{k+1}) \geq 0 \;\; \forall\boldsymbol{\omega}\in\Omega,\end{aligned} \qquad (9)$$

where $q(\boldsymbol{\omega}) = \lambda_k\boldsymbol{g}_k^T\cdot(\boldsymbol{\omega}-\boldsymbol{\omega}_k) + \frac{1}{2}||\boldsymbol{\omega}-\boldsymbol{\omega}_k||^2$ is the objective function in (4). Arranging terms yields:

$$\begin{aligned}&\lambda_k\boldsymbol{g}_k^T(\boldsymbol{\omega}_{k+1}-\boldsymbol{\omega}) \leq (\boldsymbol{\omega}_{k+1}-\boldsymbol{\omega}_k)^T(\boldsymbol{\omega}-\boldsymbol{\omega}_{k+1})\\ &= \frac{1}{2}||\boldsymbol{\omega}_k-\boldsymbol{\omega}||^2 - \frac{1}{2}||\boldsymbol{\omega}_{k+1}-\boldsymbol{\omega}||^2 - \frac{1}{2}||\boldsymbol{\omega}_k-\boldsymbol{\omega}_{k+1}||^2.\end{aligned} \qquad (10)$$

Next, we lower bound the term $\lambda_k\boldsymbol{g}_k^T\cdot(\boldsymbol{\omega}_{k+1}-\boldsymbol{\omega})$.

$$\begin{aligned}&\lambda_k\boldsymbol{g}_k^T(\boldsymbol{\omega}_{k+1}-\boldsymbol{\omega}) = \lambda_k\boldsymbol{g}_k^T(\boldsymbol{\omega}_{k+1}-\boldsymbol{\omega}_k) + \lambda_k\boldsymbol{g}_k^T(\boldsymbol{\omega}_k-\boldsymbol{\omega})\\ &\geq -\frac{\lambda_k^2}{2}||\boldsymbol{g}_k||^2 - \frac{1}{2}||\boldsymbol{\omega}_{k+1}-\boldsymbol{\omega}_k||^2 + \lambda_k\boldsymbol{g}_k^T(\boldsymbol{\omega}_k-\boldsymbol{\omega}).\end{aligned} \qquad (11)$$

Finally, combining the lower and upper bounds on $\lambda_k\boldsymbol{g}_k^T\cdot(\boldsymbol{\omega}_{k+1}-\boldsymbol{\omega})$ yields:

$$\lambda_k\boldsymbol{g}_k^T\cdot(\boldsymbol{\omega}_k-\boldsymbol{\omega}) \leq \frac{1}{2}||\boldsymbol{\omega}_k-\boldsymbol{\omega}||^2 - \frac{1}{2}||\boldsymbol{\omega}_{k+1}-\boldsymbol{\omega}||^2 + \frac{\lambda_k^2}{2}||\boldsymbol{g}_k||^2. \qquad (12)$$

Taking expectation conditioned on $\mathcal{F}_{k-1}$ yields:

$$\begin{aligned}\lambda_k\nabla\boldsymbol{f}_k^T\cdot(\boldsymbol{\omega}_k-\boldsymbol{\omega}) &\leq \frac{1}{2}||\boldsymbol{\omega}_k-\boldsymbol{\omega}||^2 - \frac{1}{2}E\left[||\boldsymbol{\omega}_{k+1}-\boldsymbol{\omega}||^2|\mathcal{F}_{k-1}\right]\\ &\quad + \frac{\lambda_k^2}{2}E\left[||\boldsymbol{g}_k||^2|\mathcal{F}_{k-1}\right],\end{aligned} \qquad (13)$$

where we used the fact that $E\left[\boldsymbol{g}_k|\mathcal{F}_{k-1}\right] = \nabla\boldsymbol{f}_k$, and $\boldsymbol{\omega}_k$ is deterministic conditioned on $\mathcal{F}_{k-1}$. Finally, using (3) we have $E\left[||\boldsymbol{g}_k||^2|\mathcal{F}_{k-1}\right] \leq 4E\left[||\boldsymbol{x}_k\boldsymbol{x}_k^T\boldsymbol{e}_k||^2|\mathcal{F}_{k-1}\right] + 4\sigma^2E\left[||\boldsymbol{x}_k||^2\right]$. Thus, (8) follows. ■

Next, consider a sequence

$$\alpha_k = \nu\frac{\gamma}{\gamma+k}, \quad k = 0, 1, ... \qquad (14)$$

where ν, γ are positive constants.

Lemma 2: The sequence α_k, with $\nu \geq 1$, and $\gamma \geq 2/\nu$ satisfies:

$$\alpha_k^2 \geq \frac{1}{\sum_{r=0}^{k}1/\alpha_r} \quad \forall k \geq 0. \qquad (15)$$

Proof: Note that it suffices to show that the step size satisfies:

$$\frac{1}{\alpha_{r+1}^2} - \frac{1}{\alpha_r^2} \leq \frac{1}{\alpha_{r+1}} \qquad (16)$$

for $r = 0, 1, ...,$ since summing (16) over $r = 0, 1, k-1$ yields: $\frac{1}{\alpha_k^2} - \frac{1}{\alpha_0^2} \leq \sum_{r=1}^{k}\frac{1}{\alpha_r}$. Since $\nu \geq 1$ we have $\alpha_0^2 \geq \alpha_0$, resulting in $\frac{1}{\alpha_k^2} \leq \frac{1}{\alpha_0^2} + \sum_{r=1}^{k}\frac{1}{\alpha_r} \leq \frac{1}{\alpha_0} + \sum_{r=1}^{k}\frac{1}{\alpha_r}$, which yields (15).

Next, we show that the step size with $\gamma \geq 2/\nu$ satisfies (16) for $r \geq 0$. Note that (16) can be written as $\frac{1-\alpha_{r+1}}{\alpha_{r+1}^2} \leq \frac{1}{\alpha_r^2}$, where substituting $\alpha_r = \frac{\nu\gamma}{\gamma+r}$ in the last inequality yields:

$$\frac{1-\frac{\nu\gamma}{\gamma+r+1}}{\left[\frac{\nu\gamma}{\gamma+r+1}\right]^2} \leq \frac{1}{\left[\frac{\nu\gamma}{\gamma+r}\right]^2}. \qquad (17)$$

After some algebraic manipulations we obtain the following quadratic inequality: $\gamma^2 + \frac{\nu(r+1)-2}{\nu}\gamma + \frac{-2r-1}{\nu} \geq 0$, where the solution yields

$$\gamma \geq \gamma(r) \triangleq \frac{-\nu(r+1)+2+\sqrt{\nu^2(r+1)^2+4\nu r+4}}{2\nu} \quad \forall r \geq 0. \qquad (18)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2017.2705559, IEEE Transactions on Automatic Control

Thus, setting $\alpha_r = \frac{\tilde{\gamma}(r)}{\tilde{\gamma}(r)+r}$ with $\tilde{\gamma}(r) \geq \gamma(r)$ satisfies (15) for all $r \geq 0$. Next, it can be verified that $\gamma(r)$ is monotonically increasing for all $r \geq 0$ and has limit $\lim_{r\to\infty} \gamma(r) = 2/\nu$. Thus, $\gamma(r) \leq 2/\nu$ for all $r \geq 0$. Hence, setting $\gamma \geq 2/\nu$ with $\nu \geq 1$ is sufficient to satisfy (15) for all $r \geq 0$. ■

Next, we present the performance bound on the error rate for all k obtained by PSGD-WA. We assume the following conditions (referred to as Assumption A). (A1) The samples $(\boldsymbol{x}_k, y_k)$ are i.i.d. across time. (A2) The constraint set Ω is bounded. (A3) The term

$$C^2 \triangleq 4e_{max} d E\left[||\boldsymbol{x}_k\boldsymbol{x}_k^T||^2\right] + 4\sigma^2 E\left[||\boldsymbol{x}_k||^2\right] \tag{19}$$

is bounded, where $||\boldsymbol{x}_k\boldsymbol{x}_k^T||$ is the spectral norm of $\boldsymbol{x}_k\boldsymbol{x}_k^T$.

Theorem 1: Assume that PSGD-WA is implemented with

$$\lambda_k = \frac{1}{2\mu}\alpha_k, \tag{20}$$

where $\alpha_k = \dfrac{\nu\gamma}{\gamma + k}$, with $\nu \geq 1$, $\gamma \geq 2/\nu$ such that $\gamma \neq 1$, and Assumptions A1-A3 hold. Then, for all $k \geq 0$ we have:

$$\begin{aligned} &E\left[f\left(\bar{\boldsymbol{\omega}}_k\right)\right] - f\left(\boldsymbol{\omega}^*\right) \\ &\leq \frac{\nu^4\gamma^4 C^2}{4\mu^3(\gamma+k)^2}\left[\log\left(\frac{\gamma+k+1}{\gamma}\right) + \frac{1+\gamma}{\gamma^2} + 1\right] E\left[||\boldsymbol{x}_k\boldsymbol{x}_k^T||^2\right] \\ &\qquad + \frac{(k+1)\nu^2\gamma^2\sigma^2}{\mu(\gamma+k)^2} E\left[||\boldsymbol{x}_k||^2\right]. \end{aligned} \tag{21}$$

Proof: By Lemma 1, setting $\boldsymbol{\omega} = \boldsymbol{\omega}^*$ in (8) yields:

$$\begin{aligned} &\frac{1}{2}E\left[||\boldsymbol{\omega}_{k+1} - \boldsymbol{\omega}^*||^2 | \mathcal{F}_{k-1}\right] + \frac{\alpha_k}{2\mu}\nabla \boldsymbol{f}_k^T \cdot (\boldsymbol{\omega}_k - \boldsymbol{\omega}^*) \\ &\quad\leq \frac{1}{2}||\boldsymbol{\omega}_k - \boldsymbol{\omega}^*||^2 + \frac{\alpha_k^2}{2\mu^2}E\left[||\boldsymbol{x}_k\boldsymbol{x}_k^T\boldsymbol{e}_k||^2 | \mathcal{F}_{k-1}\right] \\ &\qquad + \frac{\alpha_k^2}{2\mu^2}\sigma^2 E\left[||\boldsymbol{x}_k||^2\right]. \end{aligned} \tag{22}$$

Note that 2μ-strong convexity of f implies:

$$\nabla \boldsymbol{f}_k^T \cdot (\boldsymbol{\omega}_k - \boldsymbol{\omega}^*) \geq f(\boldsymbol{\omega}_k) - f(\boldsymbol{\omega}^*) + \mu||\boldsymbol{\omega}_k - \boldsymbol{\omega}^*||^2. \tag{23}$$

Substituting (23) in (22) and smoothing with respect to the filtration $\mathcal{F}_{k-1}$ yields:

$$\begin{aligned} &\frac{1}{2}E\left[||\boldsymbol{\omega}_{k+1} - \boldsymbol{\omega}^*||^2\right] + \frac{\alpha_k}{2\mu}\left(E\left[f\left(\boldsymbol{\omega}_k\right)\right] - f\left(\boldsymbol{\omega}^*\right)\right) \\ &\overset{(a)}{\leq} \frac{1}{2}E\left[||\boldsymbol{\omega}_{k+1} - \boldsymbol{\omega}^*||^2\right] + \frac{\alpha_k}{2\mu}E\left[\nabla \boldsymbol{f}_k^T \cdot (\boldsymbol{\omega}_k - \boldsymbol{\omega}^*)\right] \\ &\qquad - \frac{\alpha_k}{2}E\left[||\boldsymbol{\omega}_k - \boldsymbol{\omega}^*||^2\right] \\ &\overset{(b)}{\leq} \frac{1}{2}E\left[||\boldsymbol{\omega}_k - \boldsymbol{\omega}^*||^2\right] + \frac{\alpha_k^2}{2\mu^2}E\left[||\boldsymbol{x}_k\boldsymbol{x}_k^T\boldsymbol{e}_k||^2\right] \\ &\quad + \frac{\alpha_k^2}{2\mu^2}\sigma^2 E\left[||\boldsymbol{x}_k||^2\right] - \frac{\alpha_k}{2}E\left[||\boldsymbol{\omega}_k - \boldsymbol{\omega}^*||^2\right] \\ &= \frac{1-\alpha_k}{2}E\left[||\boldsymbol{\omega}_k - \boldsymbol{\omega}^*||^2\right] + \frac{\alpha_k^2}{2\mu^2}E\left[||\boldsymbol{x}_k\boldsymbol{x}_k^T\boldsymbol{e}_k||^2\right] \\ &\qquad + \frac{\alpha_k^2}{2\mu^2}\sigma^2 E\left[||\boldsymbol{x}_k||^2\right] \\ &\leq \frac{1-\alpha_k}{2}E\left[||\boldsymbol{\omega}_k - \boldsymbol{\omega}^*||^2\right] + \frac{\alpha_k^2}{2\mu^2}E\left[||\boldsymbol{x}_k\boldsymbol{x}_k^T||^2\right]E\left[||\boldsymbol{e}_k||^2\right] \\ &\qquad + \frac{\alpha_k^2}{2\mu^2}\sigma^2 E\left[||\boldsymbol{x}_k||^2\right]. \end{aligned} \tag{24}$$

Inequality (a) follows from (23), and inequality (b) follows from (22). Next, we upper bound $E\left[||\boldsymbol{e}_k||^2\right]$. Note that $E\left[||\boldsymbol{g}_k||^2\right]$ is bounded by $E\left[||\boldsymbol{g}_k||^2\right] = E\left[||2\boldsymbol{x}_k\left(\boldsymbol{x}_k^T\boldsymbol{e}_k + v_k\right)||^2\right] \leq C^2$, where $||\boldsymbol{g}_k||$ is the Euclidean norm of $\boldsymbol{g}_k$ and f is 2μ-strongly convex. As a result, using a similar argument as in [5, Lemma 3] we have: $E\left[||\boldsymbol{e}_k||^2\right] \leq \frac{k\alpha_{k-1}^2 C^2}{4\mu^2}$. Substituting $\alpha_{k-1} = \frac{\nu\gamma}{\gamma+k-1}$ yields:

$$E\left[||\boldsymbol{e}_k||^2\right] \leq \frac{k\nu^2\gamma^2C^2}{4\mu^2(\gamma+k-1)^2}. \tag{25}$$

As a result, substituting (25) in (24) yields:

$$\begin{aligned} &\frac{1}{2}E\left[||\boldsymbol{\omega}_{k+1} - \boldsymbol{\omega}^*||^2\right] + \frac{\alpha_k}{2\mu}\left(E\left[f\left(\boldsymbol{\omega}_k\right)\right] - f\left(\boldsymbol{\omega}^*\right)\right) \\ &\leq \frac{1-\alpha_k}{2}E\left[||\boldsymbol{\omega}_k - \boldsymbol{\omega}^*||^2\right] + \alpha_k^2\widetilde{C}^2\frac{k}{(\gamma+k-1)^2}E\left[||\boldsymbol{x}_k\boldsymbol{x}_k^T||^2\right] \\ &\qquad + \frac{\alpha_k^2\sigma^2}{2\mu^2}E\left[||\boldsymbol{x}_k||^2\right], \end{aligned} \tag{26}$$

where $\widetilde{C} \triangleq \frac{\nu\gamma}{\sqrt{8}\mu^2}C$. Next, by dividing both sides of the inequality by α_k^2 and using $(1-\alpha_k)/\alpha_k^2 \leq 1/a_{k-1}^2$ for $k \geq 1$ (see (16) in the proof of Lemma 2) we obtain:

$$\begin{aligned} &\frac{1}{2\alpha_k^2}E\left[||\boldsymbol{\omega}_{k+1} - \boldsymbol{\omega}^*||^2\right] + \frac{1}{2\alpha_k\mu}\left(E\left[f\left(\boldsymbol{\omega}_k\right)\right] - f\left(\boldsymbol{\omega}^*\right)\right) \\ &\leq \frac{1}{2\alpha_{k-1}^2}E\left[||\boldsymbol{\omega}_k - \boldsymbol{\omega}^*||^2\right] + \widetilde{C}^2\frac{k}{(\gamma+k-1)^2}E\left[||\boldsymbol{x}_k\boldsymbol{x}_k^T||^2\right] \\ &\qquad + \frac{\sigma^2}{2\mu^2}E\left[||\boldsymbol{x}_k||^2\right]. \end{aligned} \tag{27}$$

Next, summing (27) over $1, 2, ..., k$ yields:

$$\begin{aligned} &\frac{1}{2\alpha_k^2}E\left[||\boldsymbol{\omega}_{k+1} - \boldsymbol{\omega}^*||^2\right] + \frac{1}{2\mu}\sum_{r=1}^{k}\frac{1}{\alpha_r}\left(E\left[f\left(\boldsymbol{\omega}_k\right)\right] - f\left(\boldsymbol{\omega}^*\right)\right) \\ &\leq \frac{1}{2\nu^2}E\left[||\boldsymbol{\omega}_1 - \boldsymbol{\omega}^*||^2\right] + \frac{k\sigma^2}{2\mu^2}E\left[||\boldsymbol{x}_k||^2\right] + E\left[||\boldsymbol{x}_k\boldsymbol{x}_k^T||^2\right]\widetilde{C}^2 \times \\ &\qquad \left[\log\left(\frac{\gamma+k-1}{\gamma}\right) + \frac{1-\gamma^2+\gamma}{\gamma^2} + \frac{\gamma-1}{\gamma+k-1}\right], \end{aligned} \tag{28}$$

where the first term on the RHS of (28) follows by the fact that $\alpha_0 = \nu$. Note that $\frac{k}{(\gamma+k-1)^2}$ decreases with k for all $k \geq 1, \gamma > 0$. Thus, the term $\sum_{r=1}^{k}\frac{r}{(\gamma+r-1)^2} = 1/\gamma^2 + \sum_{r=2}^{k}\frac{r}{(\gamma+r-1)^2}$ is upper bounded by $1/\gamma^2 + \int_1^k \frac{rdr}{(\gamma+r-1)^2}$ which leads to the last term on the RHS of (28).

The term for $k = 0$ is obtained by dividing (26) by α_0^2 (note that $\alpha_0 = \nu \geq 1$) and substituting $k = 0$ (note that $\gamma \neq 1$):

$$\begin{aligned} &\frac{1}{2\nu^2}E\left[||\boldsymbol{\omega}_1 - \boldsymbol{\omega}^*||^2\right] + \frac{1}{2\mu\alpha_0}\left(E\left[f\left(\boldsymbol{\omega}_0\right)\right] - f\left(\boldsymbol{\omega}^*\right)\right) \\ &\leq \frac{\sigma^2}{2\mu^2}E\left[||\boldsymbol{x}_k||^2\right]. \end{aligned} \tag{29}$$

As a result, by combining (28) and (29) we obtain for all $k \geq 0$:

$$\begin{aligned} &\frac{1}{2\alpha_k^2}E\left[||\boldsymbol{\omega}_{k+1} - \boldsymbol{\omega}^*||^2\right] + \frac{1}{2\mu}\sum_{r=0}^{k}\frac{1}{\alpha_r}\left(E\left[f\left(\boldsymbol{\omega}_k\right)\right] - f\left(\boldsymbol{\omega}^*\right)\right) \\ &\leq \widetilde{C}^2\left[\log\left(\frac{\gamma+k+1}{\gamma}\right) + \frac{1+\gamma}{\gamma^2} + 1\right]E\left[||\boldsymbol{x}_k\boldsymbol{x}_k^T||^2\right] \\ &\qquad + \frac{(k+1)\sigma^2}{2\mu^2}E\left[||\boldsymbol{x}_k||^2\right]. \end{aligned} \tag{30}$$

Multiplying by $2\mu\alpha_k^2$ and rearranging terms yield:

$$\begin{aligned}&\alpha_k^2 \sum_{r=0}^{k} \frac{1}{\alpha_r}\left(E\left[f\left(\boldsymbol{\omega}_k\right)\right] - f\left(\boldsymbol{\omega}^*\right)\right)\\ &\le 2\mu\alpha_k^2 \widetilde{C}^2 \left[\log\left(\frac{\gamma+k+1}{\gamma}\right) + \frac{1+\gamma}{\gamma^2} + 1\right] E\left[||\boldsymbol{x}_k\boldsymbol{x}_k^T||^2\right]\\ &\quad + \frac{(k+1)\alpha_k^2\sigma^2}{\mu} E\left[||\boldsymbol{x}_k||^2\right] - \mu E\left[||\boldsymbol{\omega}_{k+1} - \boldsymbol{\omega}^*||^2\right]\\ &\le 2\mu\alpha_k^2 \widetilde{C}^2 \left[\log\left(\frac{\gamma+k+1}{\gamma}\right) + \frac{1+\gamma}{\gamma^2} + 1\right] E\left[||\boldsymbol{x}_k\boldsymbol{x}_k^T||^2\right]\\ &\qquad + \frac{(k+1)\alpha_k^2\sigma^2}{\mu} E\left[||\boldsymbol{x}_k||^2\right].\end{aligned} \tag{31}$$

Next, we use (15) in Lemma 2 to get:

$$\begin{aligned}&\frac{1}{\sum_{r=0}^{k}\frac{1}{\alpha_r}} \sum_{r=0}^{k} \frac{1}{\alpha_r}\left(E\left[f\left(\boldsymbol{\omega}_k\right)\right] - f\left(\boldsymbol{\omega}^*\right)\right)\\ &\le 2\mu\alpha_k^2 \widetilde{C}^2 \left[\log\left(\frac{\gamma+k+1}{\gamma}\right) + \frac{1+\gamma}{\gamma^2} + 1\right] E\left[||\boldsymbol{x}_k\boldsymbol{x}_k^T||^2\right]\\ &\qquad + \frac{(k+1)\alpha_k^2\sigma^2}{\mu} E\left[||\boldsymbol{x}_k||^2\right].\\ &= \frac{\nu^4\gamma^4 C^2}{4\mu^3(\gamma+k)^2} \left[\log\left(\frac{\gamma+k+1}{\gamma}\right) + \frac{1+\gamma}{\gamma^2} + 1\right] E\left[||\boldsymbol{x}_k\boldsymbol{x}_k^T||^2\right]\\ &\qquad + \frac{(k+1)\nu^2\gamma^2\sigma^2}{\mu(\gamma+k)^2} E\left[||\boldsymbol{x}_k||^2\right],\end{aligned} \tag{32}$$

where the last equality holds by substituting $\alpha_k = \frac{\nu\gamma}{\gamma+k}$, $\widetilde{C} = \frac{\nu\gamma}{\sqrt{8}\mu^2}C$. Next, using the convexity of f we have:

$$\begin{aligned}&E\left[f\left(\bar{\boldsymbol{\omega}}_k\right)\right] - f\left(\boldsymbol{\omega}^*\right)\\ &\le \frac{\nu^4\gamma^4 C^2}{4\mu^3(\gamma+k)^2} \left[\log\left(\frac{\gamma+k+1}{\gamma}\right) + \frac{1+\gamma}{\gamma^2} + 1\right] E\left[||\boldsymbol{x}_k\boldsymbol{x}_k^T||^2\right]\\ &+ \frac{(k+1)\nu^2\gamma^2\sigma^2}{\mu(\gamma+k)^2} E\left[||\boldsymbol{x}_k||^2\right],\ \forall \nu \ge 1,\ \gamma \ge 2/\nu,\ \gamma \ne 1,\ k \ge 0.\end{aligned} \tag{33}$$

■

Remark 1: From Theorem 1, we obtain an explicit $O(1/k)$ upper bound on the convergence rate, depending on the noise variance (second term on the RHS of (21)) and the size of the constraint set (first term on the RHS of (21)). The variance term dominates the error and decreases with rate $1/k$, while the other term (which is related to the diameter e_{max} of the constraint set) decreases faster at rate $\log(k)/k^2$. Note that any step size of the form of $\lambda_k = c\frac{\gamma}{\gamma+k}$ can be written as $\lambda_k = \frac{1}{2\mu}\frac{\nu\gamma}{\gamma+k}$, where $\nu = 2\mu c$. Thus, it suffices to know a lower bound on μ (say $\mu \ge \mu_L$) to choose $c \ge \frac{1}{2\mu_L} \ge \frac{1}{2\mu}$ so that $\nu \ge 1$, and choose $\gamma \ge 1/(\mu_L c) \ge 1/(\mu c) = 2/\nu$ so that (21) holds. The best asymptotic (as k increases) bound is obtained by setting $\nu\gamma = 2$.

Remark 2: Note that when the random components of $\boldsymbol{x}$ are identically distributed and uncorrelated (thus, the correlation matrix of $\boldsymbol{x}_k$ can be written as $E\left[\boldsymbol{x}\boldsymbol{x}^T\right] = \mu I_d$, where I_d is the identity matrix and its minimal eigenvalue is μ) we obtain: $E\left[||\boldsymbol{x}||^2\right] = d\mu$. As a result, we have $\lim_{k\to\infty} k\left(E\left[f\left(\bar{\boldsymbol{\omega}}_k\right)\right] - f\left(\boldsymbol{\omega}^*\right)\right) \le 4d\sigma^2$, where $\lim_{k\to\infty} k\left(E\left[f\left(\boldsymbol{\omega}_k^{ERM}\right)\right] - f\left(\boldsymbol{\omega}^*\right)\right) = d\sigma^2$ under the ERM scheme [36]. Hence, the asymptotic ratio ρ between the convergence rate of our scheme and the ERM scheme is upper bounded by $\rho \le 4$ as the number of iterations approaches infinity.

Remark 3: The streaming SVRG algorithm proposed in [36] for a general strongly convex regression problem achieves $\rho = 1$ asymptotically with the price of geometrically increasing batch sample size occasionally and setting the constant step size close to zero, which deteriorates performance in the finite regime. The SGD with averaging and constant step size scheme proposed in [29] for a linear least squares regression problem requires a fixed batch sample size as required by PSGD-WA. However, obtaining $\rho = 1$ asymptotically requires to set the constant step size close to zero, which deteriorates performance in the finite regime (due to a term that depends on $1/\zeta^2$ and blows up as the constant step size ζ approaches zero [29]). Controlling the decay step sizes, however, as suggested by PSGD-WA avoids that blowing up term. Theorem 1 shows that PSGD-WA achieves $\rho \le 4$, where the step sizes can be large in the beginning of the algorithm and approach zero only asymptotically. This insight is demonstrated by numerical experiments in Section V, where significant performance gain is demonstrated by PSGD-WA in the finite regime, while unweighted averaging is expected to perform well as the number of iterations becomes very large.

A. A case of $d = 1$

For purposes of analysis whether further improvement in the resulting error can be expected, we provide a better bound for the error when $d = 1$, and Ω is unbounded.

Let

$$\lambda_k = \frac{1}{2x_k^2}\alpha_k, \tag{34}$$

where $\alpha_k = \dfrac{\gamma}{\gamma+k}$. Note that when $x_k = x$ for all k, then $\mu = x^2$, and $\lambda_k = \frac{1}{2x^2}\alpha_k = \frac{1}{2x^2}\frac{\gamma}{\gamma+k}$, which is a special case of the step size in (20) when $d = 1$.

Since Ω is unbounded, the proposed PSGD-WA algorithm updates the estimate ω_{k+1} using a SGD update and computes a weighted average over iterates as byproduct of the algorithm. Specifically, at iteration $k+1$ we compute the estimate of ω^* as follows:

$$\omega_{k+1} = \omega_k - \alpha_k\left(\omega_k - y_k/x_k\right), \tag{35}$$

where $\alpha_k = \frac{\gamma}{\gamma+k}$. In addition to the estimate ω_{k+1}, we compute the weighted average estimate as in (5).

Let

$$\eta_k \triangleq \omega_k - \omega^* \;\;,\;\; \bar{\eta}_k \triangleq \bar{\omega}_k - \omega^* = \sum_{i=0}^{k}\beta_{k,i}\eta_i, \tag{36}$$

where the last equality holds since $\sum_{i=0}^{k}\beta_{k,i} = 1$ for all $k \ge 0$.

We define:

$$\widetilde{M}_{i,j} \triangleq \prod_{r=i+1}^{j}\left(1 - \alpha_{r-1}\right) \quad , \quad \text{and} \quad M_{i,j} \triangleq \beta_j\widetilde{M}_{i,j}. \tag{37}$$

For the ease of presentation we also set[3]:

$$\begin{aligned}&\beta_0 = 1\ ,\ \beta_k = \frac{1}{\alpha_{k-1}} = \frac{\gamma+k-1}{\gamma}\ \forall k \ge 1\ ,\\ &\beta_{k,i} = \frac{\beta_i}{\sum_{r=0}^{k}\beta_r}\ \forall k \ge 0,\end{aligned} \tag{38}$$

where $\gamma \ge 1$.

Lemma 3: Assume that (38) holds. Then, for all $i < k$,

$$\sum_{j=i+1}^{k} M_{i+1,j} \le \frac{(i+\gamma)(k-i)}{\gamma}. \tag{39}$$

Proof: Since $1 - \alpha_i = 1 - \frac{\gamma}{\gamma+i} = \frac{i}{\gamma+i}$, we can rewrite $M_{i+1,j}$ as:

$$M_{i+1,j} = \frac{\gamma+j-1}{\gamma}\prod_{r=i+2}^{j}\frac{r-1}{\gamma+r-1}\ \forall j \ge i+1. \tag{40}$$

[3] It should be noted that a similar asymptotic result in this section is obtained by setting $\beta_{k,i}$ as in (20)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2017.2705559, IEEE Transactions on Automatic Control

As a result, we obtain:

$$\begin{aligned} M_{i+1,j} = &\frac{\gamma+j-1}{\gamma}\times \\ &\left[\frac{i+1}{\gamma+i+1}\cdot\frac{i+2}{\gamma+i+2}\cdots\frac{j-2}{\gamma+j-2}\cdot\frac{j-1}{\gamma+j-1}\right] \\ &\leq \frac{\gamma+j-1}{\gamma}\times \\ &\left[(i+1)\cdots(i+\lfloor\gamma\rfloor)\frac{i+\lfloor\gamma\rfloor+1}{\gamma+i+1}\cdots\frac{j-1}{j-1+\gamma-\lfloor\gamma\rfloor}\times\right. \\ &\left.\frac{1}{j+\gamma-\lfloor\gamma\rfloor}\cdots\frac{1}{j+\gamma-1}\right] \\ &\leq \frac{\gamma+j-1}{\gamma}\left[\frac{i+1}{j+\gamma-\lfloor\gamma\rfloor}\cdots\frac{i+\lfloor\gamma\rfloor}{j+\gamma-1}\right] \\ &\leq \frac{\gamma+j-1}{\gamma}\left[\frac{i+\lfloor\gamma\rfloor}{j+\gamma-1}\right]\leq\frac{i+\gamma}{\gamma}. \end{aligned} \tag{41}$$

Thus, summing over j yields (39). ■

Next, we present the performance bound on the error obtained by PSGD-WA. We assume the following conditions (referred to as Assumption B). (B1) The samples $(\boldsymbol{x}_k, y_k)$ are i.i.d. across time. (B2) $d=1$.

Theorem 2: Assume that PSGD-WA is implemented, where the parameters satisfies (38). Then,
a) If Assumptions B1-B2 hold, then for all $k\geq 0$ we have:

$$E\left[f\left(\bar{\omega}_k\right)\right]-f\left(\omega^*\right)\leq\frac{4\gamma^2\sigma^2E[x^2]E[1/x^2]}{3k}+O\left(k^{-2}\right). \tag{42}$$

b) In addition, if $x_k=x$ for all k and $\gamma=1$, we have:

$$\lim_{k\to\infty} k\left(E\left[f\left(\bar{\omega}_k\right)\right]-f\left(\omega^*\right)\right)\leq\frac{4}{3}\sigma^2. \tag{43}$$

Proof: Since we consider a least squares loss, we have:

$$E\left[f\left(\bar{\omega}_k\right)\right]-f\left(\omega^*\right)=E[x^2]E\left[\bar{\eta}_k^2\right], \tag{44}$$

Next, we compte $\bar{\eta}_k$. Note that η_i can be written recursively as follows:

$$\begin{aligned}\eta_i=\omega_i-\omega^*&=\omega_{i-1}-\alpha_{i-1}(\omega_{i-1}-y_{i-1}/x_{i-1})-\omega^* \\ &=(1-\alpha_{i-1})\eta_{i-1}+\alpha_{i-1}v_{i-1}/x_{i-1},\end{aligned} \tag{45}$$

and by iterating over η_i we obtain:

$$\eta_k=\sum_{i=0}^{k-1}\widetilde{M}_{i+1,k}\alpha_i v_i/x_i. \tag{46}$$

Hence,

$$\begin{aligned}\bar{\eta}_k&=\frac{1}{\sum_{r=0}^{k}\beta_r}\sum_{j=0}^{k}\sum_{i=0}^{j-1}\beta_j\widetilde{M}_{i+1,j}\alpha_i v_i/x_i \\ &=\frac{1}{\sum_{r=0}^{k}\beta_r}\sum_{i=0}^{k-1}\left(\sum_{j=i+1}^{k}M_{i+1,j}\right)\alpha_i v_i/x_i.\end{aligned} \tag{47}$$

Next, we compute $E\left[\bar{\eta}_k^2\right]$. Note that:

$$\begin{aligned}E\left[\bar{\eta}_k^2\right]=&\frac{1}{\left(\sum_{r=0}^{k}\beta_r\right)^2}\times \\ &E\left[\sum_{i=0}^{k-1}\left(\sum_{j=i+1}^{k}M_{i+1,j}\right)\alpha_i\frac{v_i}{x_i}\sum_{\ell=0}^{k-1}\left(\sum_{p=\ell+1}^{k}M_{\ell+1,p}\right)\alpha_\ell\frac{v_\ell}{x_\ell}\right].\end{aligned} \tag{48}$$

Since cross terms are canceled (due to independence across time), we obtain:

$$E\left[\bar{\eta}_k^2\right]=\frac{\sigma^2E[1/x^2]}{\left(\sum_{r=0}^{k}\beta_r\right)^2}\sum_{i=0}^{k-1}\left(\sum_{j=i+1}^{k}M_{i+1,j}\right)^2\alpha_i^2. \tag{49}$$

Setting β_r according to (38) yields:

$$\left(\sum_{r=0}^{k}\beta_r\right)^2=\frac{k^4}{4}+O(k^3). \tag{50}$$

Next, applying Lemma 3 and setting $\alpha_i=\frac{\gamma}{\gamma+i}$ yields:

$$E\left[\bar{\eta}_k^2\right]\leq\frac{\gamma^2\sigma^2E[1/x^2]}{k^4/4+O(k^3)}\sum_{i=0}^{k-1}\left(k-i\right)^2. \tag{51}$$

Finally, since $\sum_{i=0}^{k-1}(k-i)^2=k^3/3+O(k^2)$, (42) follows. Setting $\gamma=1$, $x_k=x$ for all k, and letting $k\to\infty$ yields (43). ■

Remark 4: Note that when the conditions in Theorem 2.b hold, then the asymptotic ratio ρ between the convergence rate of PSGD-WA and the ERM scheme is upper bounded by $\rho\leq 4/3$ as the number of iterations approaches infinity. Thus, the upper bound on the error is better then $\rho\leq 4$ obtained in Theorem 1. Simulation results demonstrate $\rho\leq 4/3$ even for large d in practice.

V. Numerical Examples

In this section, we provide numerical examples to illustrate the performance of the algorithms. We have performed experiments on synthetic as well as real date set.

A. Experiments Over Synthetic Data

In this section we numerically study the performance of the algorithms using synthetic data. We set the following parameters: the streaming data $\boldsymbol{x}_k\in\mathbb{R}^d$ are i.i.d r.v. drawn from a normal distribution with covariance matrix I_d, and $y_k=\boldsymbol{x}_k^T\omega^*+v_k$, where $v_k\sim N(0,1)$ is an additive Gaussian noise. $\boldsymbol{\omega}^*=[1\ 2\ \ldots\ d]^T$ is the unknown parameter. The constraint set for the projected SGD iterates was set to $\boldsymbol{\omega}^*\pm 100$. We compared three streaming algorithms that require a very similar computational complexity: i) a standard Projected SGD, referred to as PSGD; ii) a Projected SGD using a constant step size with a standard arithmetic (unweighted) Averaging, referred to as PSGD-A (i.e., a projected version of the algorithm proposed in [29]); iii) the proposed Projected SGD algorithm with decreasing step sizes and Weighted Averaging (PSGD-WA). We performed 1000 Monte-Carlo experiments to compute the average performance. The parameters of the stepsizes were tuned by performing an extensive grid search process.

First, we set $d=100$. We investigated the proposed PSGD-WA algorithm with stepsizes $5/(5+k)$, $0.2/(1+k)^{3/4}$, referred to as PSGD-WA 1, and PSGD-WA 2, respectively. Note that the theoretical analysis provided in this paper holds for PSGD-WA 1. We examined the PSGD-A algorithm with stepsizes 0.012 and 0.001, referred to as PSGD-A 1 and PSGD-A 2, respectively. The performance of the algorithms are presented in Fig. 1. As a benchmark, we computed the empirical risk minimizer (ERM), which solves (1) directly by using the *entire data* at each iteration (using a constrained ordinary least squares solver). It can be seen that the proposed PSGD-WA 1 algorithm performs the best among the streaming algorithms and obtains performance close to the ERM algorithm for a large range of tested k. The ratio between the errors under PSGD-WA algorithm and the ERM schemes is less than 1.31 for all $k>8\cdot 10^5$. These results coincide with the upper bound $\rho\leq 4/3$ obtained in Theorem 2 under $d=1$. However, showing that $\rho\leq 4/3$ theoretically for $d>1$ remains open. It can be seen that PSGD-WA 2 algorithm performs well as k increases. Note that when using PSGD-A, in which the step size is fixed, one can always improve the asymptotic performance by decreasing the step size (with the price of deteriorating performance in the finite regime), as was shown in [29]. Indeed, it can be seen that PSGD-A 2 has a better error decrease rate than PSGD-A 1. However, the price in terms of deteriorating performance in the finite regime is high, whereas PSGD-A 2 performs poorly for for all tested k. These results confirm the advantages of the proposed PSGD-WA algorithm in the finite sample regime, and demonstrate its nice asymptotic property (up to a constant ratio between the asymptotic errors under the PSGD-WA algorithm and the ERM scheme).

Next, we set $d = 1000$. We investigated the proposed PSGD-WA algorithm with stepsizes $5/(5+k)$, $1/(1+k)^{3/4}$, referred to as PSGD-WA 1, and PSGD-WA 2, respectively. The standard Projected SGD, referred to as PSGD was implemented with stepsize sequence $5/(5+k)$. We examined the PSGD-A algorithm with stepsize 0.0006. The performance of the algorithms are presented in Fig. 2. It can be seen that the proposed PSGD-WA algorithm performs the best among the streaming algorithms for a large range of tested k. It can also be seen that PSGD-A outperforms the standard PSGD for $k > 3 \cdot 10^6$ and is expected to perform well for very large k. Improving the error decrease rate obtained by PSGD-A can be done by decreasing the step size with the price of deteriorating performance in the finite regime. Again, these results confirm the advantages of the proposed PSGD-WA algorithm in the finite sample regime and demonstrate its nice asymptotic property.

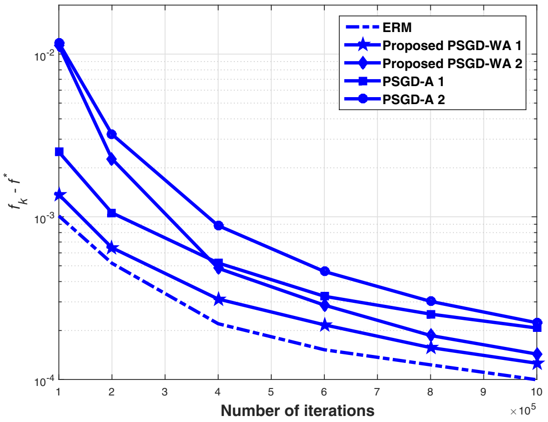

Fig. 1. The error as a function of the number of iterations under various PSGD algorithms with problem dimension $d = 100$ as described in Sec. V-A.

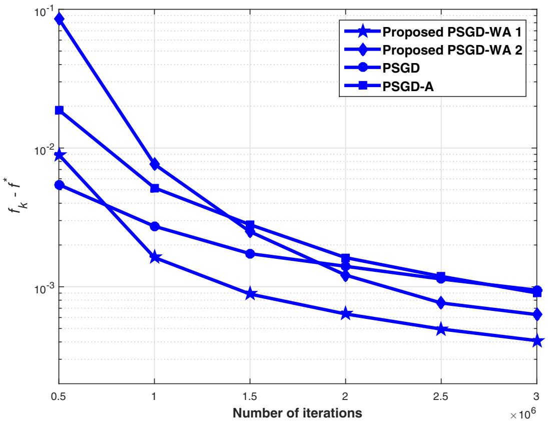

Fig. 2. The error as a function of the number of iterations under various PSGD algorithms with problem dimension $d = 1000$ as described in Sec. V-A.

B. Experiments Over the Million Song Dataset

In this section we examine the performance of the algorithms for prediction of a release year of a song from audio features. We used the dataset available by UCI Machine Learning Repository [38], extracted from the Million Song Dataset collaborative project between The Echo Nest and LabROSA [39]. The Million Song Dataset contains songs which are mostly western, commercial tracks ranging from 1922 to 2011. Each song is associated with a released year (i.e., y in our model that we aim to estimate), and 90 audio attributes (i.e., $\boldsymbol{x}$ in our model). We compared three streaming algorithms as described in Sec. V-A. Here, we did not assume prior knowledge on the constraint set of the parameters. Thus, the projected update degenerates to an unconstrained update: $\boldsymbol{\omega}_{k+1} = \boldsymbol{\omega}_{k+1} - \lambda_k \boldsymbol{g}_k$.

It should be noted that the theoretical performance analysis established in Theorem 1 does not apply to the unconstrained optimization in this simulations. Nevertheless, simulation results demonstrate very good performance of the proposed algorithm in this case as well. In Fig. 3, we present the average prediction error of the released year of a song $|\hat{y}_k - y_k|$ as a function of the number of iterations. It can be seen that the proposed PSGD-WA algorithm (implemented with step size sequence $0.01/(100+k)$) performs the best among the streaming algorithms for all tested k. It can also be seen that the standard PSGD algorithm (implemented with step size sequence $0.01/(100+k)$) performs the worst for all tested k. It should be noted that simulation results demonstrate that PSGD-A (implemented with step size 0.00015) has high decrease rate, thus, expected to perform well as k becomes large. These results confirm the advantages of the proposed PSGD-WA algorithm in the finite sample regime and provide important design principles when implementing PSGD-based algorithms for regression tasks.

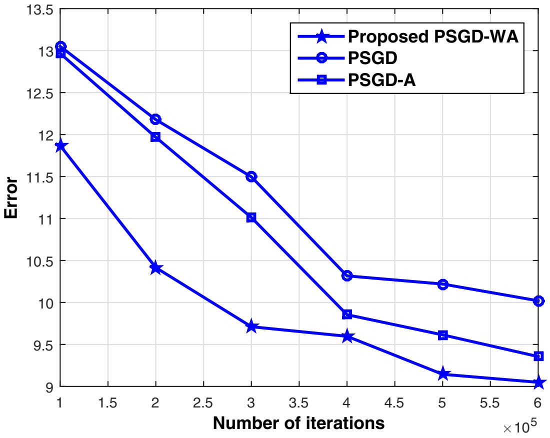

Fig. 3. The average prediction error of the released year of a song $|\hat{y}_k - y_k|$ as a function of the number of iterations under various PSGD algorithms as described in Sec. V-B.

VI. Conclusion

We considered a least squares regression of a d-dimensional unknown parameter. We proposed and analyzed a projected stochastic gradient descent algorithms with weighted iterate-averaging. We established an explicit $O(1/k)$ upper bound on the convergence rate when the constraint set of the unknown parameter is bounded. We showed that the variance term dominates the error and decreases with rate $1/k$, while the term which is related to the size of the constraint set decreases with rate $\log k/k^2$. We then compared the asymptotic ratio ρ between the convergence rate of the proposed scheme and the empirical risk minimizer (ERM) as the number of iterations approaches infinity. We showed that $\rho \leq 4$ for all $d \geq 1$ when the random entries of the sensing vector are uncorrelated and identically distributed. We further improved the upper bound by showing that $\rho \leq 4/3$ for the case of $d = 1$ and unbounded parameter set when the random sensing entries are equal across time. Simulation results demonstrated strong performance of the algorithm, and coincide with $\rho \leq 4/3$ even for large d in practice.

Future research directions are the following: (i) In addition to the analysis of the proposed PSGD-WA algorithm with a step size that decays with rate $1/k$, it is desirable to analyze the proposed PSGD-WA algorithm with a step size that decays with rate $1/k^{\alpha}$, where $1/2 < \alpha < 1$. (ii) In addition to the finite-sample upper bounds on the convergence rate established in this paper, a future research direction will be to analyze the asymptotic covariance of the algorithm and compare it to the optimal one. (iii) Another possible method is to use a projected weighted averaging of unprojected SGD iterates. Thus, it is desirable to analyze this scheme as well. (iv) This work has focused on regression from noisy observations with a constant variance. It is desirable to establish a fine-grained analysis for the case in which the variance is not constant, but is a function of the sensing vector (i.e., data with heteroscedasticity). (v) It should be noted that SGD with a constant step size does not converge to the global optimum under general smooth loss functions [40], [41]. Thus, it is desirable to analyze the proposed PSGD-WA algorithm with a decreasing step size under other loss functions (e.g., logistic regression).

References

[1] K. Cohen, A. Nedic, and R. Srikant, "On projected stochastic gradient descent algorithm with weighted averaging for least squares regression," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2314–2318, 2016.

[2] L. Bottou, "Online learning and stochastic approximations," *On-line learning in neural networks*, vol. 17, no. 9, p. 25, 1998.

[3] P. Tseng, "On accelerated proximal gradient methods for convex-concave optimization," *submitted to SIAM Journal on Optimization*, 2008.

[4] S. Lacoste-Julien, M. Schmidt, and F. Bach, "A simpler approach to obtaining an $O(1/t)$ convergence rate for the projected stochastic subgradient method," *arXiv preprint arXiv:1212.2002*, 2012.

[5] A. Nedić and S. Lee, "On stochastic subgradient mirror-descent algorithm with weighted averaging," *SIAM Journal on Optimization*, vol. 24, no. 1, pp. 84–107, 2014.

[6] A. Nemirovskii and D. Yudin, "Cezare convergence of gradient method approximation of saddle points for convex-concave functions," *Doklady Akademii Nauk SSSR*, vol. 239, pp. 1056–1059, 1978.

[7] Y. Nesterov, "A method of solving a convex programming problem with convergence rate $O(1/k^2)$," in *Soviet Mathematics Doklady*, vol. 27, pp. 372–376, 1983.

[8] X.-R. Cao, "Convergence of parameter sensitivity estimates in a stochastic experiment," *IEEE Transactions on Automatic Control*, vol. 30, no. 9, pp. 845–853, 1985.

[9] B. T. Polyak and A. B. Juditsky, "Acceleration of stochastic approximation by averaging," *SIAM Journal on Control and Optimization*, vol. 30, no. 4, pp. 838–855, 1992.

[10] R. Schwabe, "Stability results for smoothed stochastic approximation procedures," *Z. Angew. Math. Mech.*, vol. 73, pp. 639–643, 1993.

[11] M. Pelletier, "Asymptotic almost sure efficiency of averaged stochastic algorithms," *SIAM Journal on Control and Optimization*, vol. 39, no. 1, pp. 49–72, 2000.

[12] B. Polyak, "Random algorithms for solving convex inequalities," *Studies in Computational Mathematics*, vol. 8, pp. 409–422, 2001.

[13] W.-P. Ang and B. Farhang-Boroujeny, "A new class of gradient adaptive step-size LMS algorithms," *IEEE transactions on signal processing*, vol. 49, no. 4, pp. 805–810, 2001.

[14] Y. Nesterov, "Introductory lectures on convex optimization: a basic course," 2004.

[15] H.-C. Shin, A. H. Sayed, and W.-J. Song, "Variable step-size NLMS and affine projection algorithms," *IEEE signal processing letters*, vol. 11, no. 2, pp. 132–135, 2004.

[16] A. Juditsky, P. Rigollet, A. B. Tsybakov, *et al.*, "Learning by mirror averaging," *The Annals of Statistics*, vol. 36, no. 5, pp. 2183–2206, 2008.

[17] Y. Zhang, N. Li, J. A. Chambers, and Y. Hao, "New gradient-based variable step size LMS algorithms," *EURASIP Journal on Advances in Signal Processing*, vol. 2008, p. 105, 2008.

[18] J.-K. Hwang and Y.-P. Li, "Variable step-size LMS algorithm with a gradient-based weighted average," *IEEE Signal Processing Letters*, vol. 12, no. 16, pp. 1043–1046, 2009.

[19] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM journal on imaging sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[20] Y. Nesterov, "Primal-dual subgradient methods for convex problems," *Mathematical programming*, vol. 120, no. 1, pp. 221–259, 2009.

[21] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, "Robust stochastic approximation approach to stochastic programming," *SIAM Journal on Optimization*, vol. 19, no. 4, pp. 1574–1609, 2009.

[22] L. Xiao, "Dual averaging method for regularized stochastic learning and online optimization," in *Advances in Neural Information Processing Systems*, pp. 2116–2124, 2009.

[23] A. Rakhlin, O. Shamir, and K. Sridharan, "Making gradient descent optimal for strongly convex stochastic optimization," *In The 29th International Conference on Machine Learning (ICML)*, 2012.

[24] S. Ghadimi and G. Lan, "Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization I: A generic algorithmic framework," *SIAM Journal on Optimization*, vol. 22, no. 4, pp. 1469–1492, 2012.

[25] G. Lan, "An optimal method for stochastic composite optimization," *Mathematical Programming*, vol. 133, no. 1-2, pp. 365–397, 2012.

[26] O. Shamir and T. Zhang, "Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes.," in *International Conference on Machine Learning (ICML)*, pp. 71–79, 2013.

[27] M. Schmidt, N. L. Roux, and F. Bach, "Minimizing finite sums with the stochastic average gradient," *arXiv preprint arXiv:1309.2388*, 2013.

[28] S. Bonnabel, "Stochastic gradient descent on Riemannian manifolds," *IEEE Transactions on Automatic Control*, vol. 58, no. 9, pp. 2217–2229, 2013.

[29] A. Défossez and F. Bach, "Averaged least-mean-squares: Bias-variance trade-offs and optimal sampling distributions," in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pp. 205–213, 2015.

[30] D. Blatt, A. O. Hero, and H. Gauchman, "A convergent incremental gradient method with a constant step size," *SIAM Journal on Optimization*, vol. 18, no. 1, pp. 29–51, 2007.

[31] P. Tseng, "An incremental gradient (-projection) method with momentum term and adaptive stepsize rule," *SIAM Journal on Optimization*, vol. 8, no. 2, pp. 506–531, 1998.

[32] N. L. Roux, M. Schmidt, and F. Bach, "A stochastic gradient method with an exponential convergence rate for finite training sets," in *Advances in Neural Information Processing Systems*, pp. 2663–2671, 2012.

[33] R. Schwabe and H. Walk, "On a stochastic approximation procedure based on averaging," *Metrika*, vol. 44, no. 1, pp. 165–180, 1996.

[34] H. J. Kushner and J. Yang, "Stochastic approximation with averaging and feedback: rapidly convergent "on-line" algorithms," *IEEE Transactions on Automatic Control*, vol. 40, no. 1, pp. 24–34, 1995.

[35] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," in *Advances in Neural Information Processing Systems*, pp. 315–323, 2013.

[36] R. Frostig, R. Ge, S. M. Kakade, and A. Sidford, "Competing with the empirical risk minimizer in a single pass.," in *Proc. Conference on Learning Theory (COLT) (also available at arXiv: 1412.6606)*, pp. 728–763, 2015.

[37] S. Shalev-Shwartz and T. Zhang, "Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization," *Mathematical Programming*, pp. 1–41, 2014.

[38] M. Lichman, "UCI machine learning repository," in *University of California, Irvine, School of Information and Computer Sciences, http://archive.ics.uci.edu/ml*, 2013.

[39] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere, "The million song dataset," in *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.

[40] A. Nedić and D. Bertsekas, "Convergence rate of incremental subgradient algorithms," in *Stochastic optimization: algorithms and applications*, pp. 223–264, Springer, 2001.

[41] F. Bach and E. Moulines, "Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$," in *Advances in Neural Information Processing Systems*, pp. 773–781, 2013.