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

Kobi Cohen[‡], Angelia Nedić[§], R. Srikant[§]
[‡] Department of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel
yakovsec@bgu.ac.il
[§] Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, IL 61801
{angelia, rsrikant}@illinois.edu

***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 constraint set term 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 $\rho \leq 4$ under some mild conditions for all $d \geq 1$. We further improve the upper bound by showing that $\rho \leq 4/3$ for the case of $d = 1$ and unbounded parameter set. 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 [1].

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:

$$\begin{array}{l} \min\limits_{\boldsymbol{\omega}} E\left[||\boldsymbol{x}^T\boldsymbol{\omega} - y||^2\right] \\ \text{subject to } \ \boldsymbol{\omega} \in \Omega \subseteq \mathbb{R}^d \end{array} \quad (1)$$

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 propose and 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 $c > 0$, $\gamma \geq 1$ 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

This work was supported by AFOSR MURI FA 9550–10–1–0573, ONR Grant N00014–13–1–003, NSF Grant CNS–1161404.

[1]It should be noted that previous studies on PSGD algorithms with weighted averaging (see [2], [3]) considered only a fixed form of the step size without tuning parameters.

(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 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) (which is the minimizer in the absence of computational constraints) as the number of iterations approaches infinity. We show 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.

B. Related Work

Accelerating SGD-based algorithms using averaging techniques has been studied in past and more recent years in [2]–[22]. In [12], 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 [9], for which the first fast algorithm was originally constructed by Nesterov [5] for unconstrained problems, and was extended recently by Beck and Teboulle in [15] to a larger class of problems. Ghadimi and Lan used averaging in [20] 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. [11] considered a mirror-descent algorithm with averaging to construct aggregate estimators with the best achievable learning rate. 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 [21]. Other related works are concerned with iterate-averaging for best achievable rate of stochastic subgradients methods [17], [19], as well as gradient-averaging [8], [10], [13], [14], [16], [18] and a sort of momentum [23], [24], in which the algorithm uses a sort of weighting over previous gradients (instead of the iterate minimizer) in the construction of the algorithm.

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 [4] for convex-concave saddle-point problems, by Polyak and Juditsky [6] for stochastic gradient approximations and by Polyak [7] for convex feasibility problems. In [6], an asymptotically optimal performance has been achieved. However, a finite sample analysis remained open. More recently, Lacoste-Julien et al. [2] used this averaging approach for a projected stochastic subgradient method to achieve $1/k$ convergence rate for strongly convex functions. Nedić and Lee [3] 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.

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 [22], [25]–[27]. We focus on the strongly convex case, in which $O(1/k)$ is the best attainable convergence rate [26]. 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 [26], 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 [25], 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 [22], 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 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 does not require the sample batch size to grow geometrically occasionally as in [26] or setting small step-sizes in the beginning of the algorithm as in [22], [26]. 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.

C. 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 $\boldsymbol{R}_x = E\left[\boldsymbol{x}\boldsymbol{x}^T\right]$ the correlation matrix of $\boldsymbol{x}$.

It is desired to minimize the expected least squares loss in (1) from the samples stream $(\boldsymbol{x}_k, y_k)$ at times $k = 1, 2, ...$ It is assumed that $\boldsymbol{R}_x$ is invertible (i.e., strongly convex case). We denote by μ the smallest eigenvalue of $\boldsymbol{R}_x$, so that $\mu > 0$. We denote the optimal solution of (1) 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 $\boldsymbol{\omega}$, 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 $\nabla \boldsymbol{f}(\boldsymbol{\omega}) = E\left[2\boldsymbol{x}\left(\boldsymbol{x}^T\boldsymbol{\omega} - y\right)\right]$, where $\boldsymbol{g}_k(\boldsymbol{\omega}) \triangleq 2\boldsymbol{x}_k\left(\boldsymbol{x}_k^T\boldsymbol{\omega} - y_k\right)$ is defined as 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 estimation error at the k^{th} iteration is defined by $\boldsymbol{e}_k \triangleq \boldsymbol{\omega}_k - \boldsymbol{\omega}^*$.

III. Projected Stochastic Gradient descent algorithm with Weighted Averaging

We propose a Projected Stochastic Gradient descent algorithm with Weighted Averaging (PSGD-WA) for solving (1). 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 a decreasing step size with k. Let $\boldsymbol{\omega}_0 \in \Omega$ be an initial estimate of $\boldsymbol{\omega}$ (possibly random). At iteration $k = 1$ we compute the projected estimate of $\boldsymbol{\omega}^*$ based on the random measurements $(\boldsymbol{x}_0, y_0)$ and the initial estimate $\boldsymbol{\omega}_0$:

$$\boldsymbol{\omega}_1 = \arg\min_{\boldsymbol{\omega}\in\Omega}\left\{\lambda_0\boldsymbol{g}_0^T\cdot(\boldsymbol{\omega} - \boldsymbol{\omega}_0) + \frac{1}{2}||\boldsymbol{\omega} - \boldsymbol{\omega}_0||^2\right\}, \quad (2)$$

and we update this estimate iteratively. In general, 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 $\boldsymbol{\omega}_k$:

$$\boldsymbol{\omega}_{k+1} = \arg\min_{\boldsymbol{\omega}\in\Omega}\left\{\lambda_k\boldsymbol{g}_k^T\cdot(\boldsymbol{\omega} - \boldsymbol{\omega}_k) + \frac{1}{2}||\boldsymbol{\omega} - \boldsymbol{\omega}_k||^2\right\} \quad \forall k \geq 0. \quad (3)$$

It can be verified that $\boldsymbol{\omega}_{k+1}$ projects the unconstrained gradient descent iterate into Ω. Motivated by previous studies on SGD with iterate-averaging (e.g., [2], [3]), in addition to the estimate $\boldsymbol{\omega}_{k+1}$, we propose to compute the weighted average estimate:

$$\bar{\boldsymbol{\omega}}_{k+1} = \sum_{i=0}^{k+1}\beta_{k+1,i}\boldsymbol{\omega}_i, \quad (4)$$

where $\beta_{k,0}, \beta_{k,1}, ..., \beta_{k,k}$ are nonnegative scalars with the sum equals 1, and 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 $\bar{\boldsymbol{\omega}}_k$ can be computed recursively (see (5) in Section III-A). In Section IV we will analyze the convergence rate of $\bar{\boldsymbol{\omega}}_k$ to the solution of (1).

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

A. Implementation

The PSGD-WA algorithm is computationally efficient as compared to existing methods. At iteration k, the algorithm requires to store $\boldsymbol{\omega}_k$, the weighted average $\bar{\boldsymbol{\omega}}_{k-1}$ and the normalization term $S_{k-1} = \sum_{r=0}^{k-1} 1/\alpha_r$, where $\alpha_r = \gamma/(\gamma + r)$ (see (6) below). The weighted average $\bar{\boldsymbol{\omega}}_k$ can be updated recursively by computing $S_k = S_{k-1} + 1/\alpha_k$ and by setting:

$$\bar{\boldsymbol{\omega}}_k = \frac{S_{k-1}}{S_k}\bar{\boldsymbol{\omega}}_{k-1} + \left(1 - \frac{S_{k-1}}{S_k}\right)\boldsymbol{\omega}_k. \quad (5)$$

Note that PSGD-WA does not require the sample batch size to grow as in [26]. The storage memory required by PSGD-WA is very similar to that required by the SGD with averaging and constant step size algorithm proposed in [22].

IV. Performance Analysis

In this section we analyze the algorithm's performance. Consider first the case where 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 $\boldsymbol{\omega}^*$.

Theorem 1: Assume that PSGD-WA is implemented, with

$$\begin{aligned}\lambda_k &= \frac{1}{2\mu}\alpha_k = \frac{1}{2\mu}\frac{\gamma}{\gamma + k} \\ \beta_{k,i} &= \frac{1/\alpha_i}{\sum_{r=0}^{k} 1/\alpha_r},\end{aligned} \quad (6)$$

where $\gamma \geq 2$. Then, for all $k \geq 0$ we have:

$$\begin{aligned}&E\left[f\left(\bar{\boldsymbol{\omega}}_k\right)\right] - f\left(\boldsymbol{\omega}^*\right) \\ &\leq \frac{\left(\log(k+1) + 1\right)\gamma^2 E\left[||\boldsymbol{x}\boldsymbol{x}^T||^2\right]C^2}{\mu^2(\gamma + k)^2} \\ &\quad + \frac{(k+1)\gamma^2 E\left[||\boldsymbol{x}||^2\right]\sigma^2}{\mu(\gamma + k)^2}, \quad \forall\gamma \geq 2\ \forall k \geq 0,\end{aligned} \quad (7)$$

where

$$C^2 \triangleq 4e_{max}dE\left[||\boldsymbol{x}\boldsymbol{x}^T||^2\right] + 4\sigma^2 E\left[||\boldsymbol{x}||^2\right]. \quad (8)$$

The proof is given in the extended version of this paper [28].

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 (7)) and the size of the constraint set (first term on the RHS of (7)). 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 $\log(k)/k^2$. The best asymptotic (as k increases) bound is obtained by setting $\gamma = 2$.

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 $\lim_{k\to\infty} k\left(E\left[f\left(\bar{\boldsymbol{\omega}}_k\right)\right] - f\left(\boldsymbol{\omega}^*\right)\right) \leq 4d\sigma^2$, where $\lim_{k\to\infty} k\left(E\left[f\left(\boldsymbol{\omega}_k^{ERM}\right)\right] - f\left(\boldsymbol{\omega}^*\right)\right) \sim d\sigma^2$ under the ERM scheme. Hence, the asymptotic ratio ρ between the convergence rate of our scheme and the ERM scheme is upper bounded by $\rho \leq 4$ as the number of iterations approaches infinity.

Remark 3: The streaming SVRG algorithm proposed in [26] 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 [22] 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. Theorem 1, however, shows that PSGD-WA achieves $\rho \leq 4$, where the step sizes can be large in the beginning of the algorithm.

For purposes of analysis whether further improvement in the resulting error can be expected, we analyze the following case. Assume: A1) Ω is unbounded (thus, the algorithm applies SGD-WA without projection); A2) $d = 1$; A3) the step size satisfies[3] $\lambda_k = \frac{1}{2x_k^2}\frac{\gamma}{\gamma+k}$; and A4) $\beta_{k,i}$ is set as in (6). The following theorem provides a better bound on the error for an unbounded constraint set.

Theorem 2: Assume that SGD-WA is implemented and Assumptions $A1 - A4$ hold. Then,
a) for all $k \geq 0$ we have:

$$E\left[f\left(\bar{\omega}_k\right)\right] - f\left(\omega^*\right) \leq \frac{4\gamma^2\sigma^2 E[x^2]E[1/x^2]}{3k} + O\left(k^{-2}\right). \tag{9}$$

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

$$\lim_{k\to\infty} k\left(E\left[f\left(\bar{\omega}_k\right)\right] - f\left(\omega^*\right)\right) \leq \frac{4}{3}\sigma^2. \tag{10}$$

The proof is given in the extended version of this paper [28].

Remark 4: Note that when the conditions in Theorem 2.b hold, then the asymptotic ratio ρ between the convergence rate of SGD-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 set the following parameters (very similar to the experiment setup in [22]): $d = 25$, $\boldsymbol{x}_k \in \mathbb{R}^{25}$ are i.i.d r.v. drawn from a normal distribution with covariance matrix I_d. $y_k = \boldsymbol{x}_k^T\omega^* + v_k$, where $v_k \sim N(0,1)$ is an additive Gaussian noise. $\boldsymbol{\omega}^* = [1\ 2\ ...\ 25]^T$ 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 and tuned their parameters: i) a standard Projected SGD with decreasing step size $10/(10+k)$, referred to as PSGD; ii) a Projected SGD using a constant step size 0.002 with Averaging, referred to as PSGD-A (i.e., a projected version of the algorithm proposed in [22]); iii) the proposed Projected SGD algorithm with decreasing step size $10/(10+k)$ and Weighted Averaging (PSGD-WA). As a benchmark, we computed the empirical risk minimizer (ERM), which solves (1) directly by using the *entire data* at each iteration. The performance of the algorithms are presented in Fig. 1. It can be seen that the proposed PSGD-WA algorithm performs the best among the streaming algorithms and obtains performance close to the ERM algorithm for all tested k. The ratio between the errors under PSGD-WA and the ERM schemes was less than 1.335 for all $k > 2 \cdot 10^4$ and equals 1.31 for $k = 10^5$. These results coincide with the upper bound $\rho \leq 4/3$ obtained in Theorem 2 under $d = 1$. However, showing $\rho \leq 4/3$ theoretically for $d > 1$ remains open. It can also be seen that PSGD-A performs the worst for $k \leq 6 \cdot 10^4$ iterations but outperforms PSGD for $k > 6 \cdot 10^4$. These results confirm the advantages of the proposed PSGD-WA in the finite sample regime. It should be noted that similar results have been observed under many different scenarios.

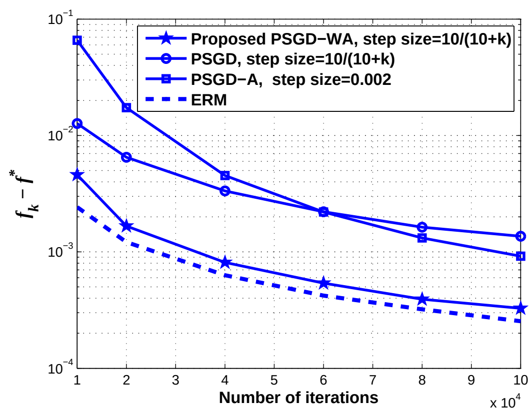

Fig. 1. The error as a function of the number of iterations.

VI. Conclusion

We considered a least squares regression of a d-dimensional unknown parameter. We proposed and analyzed a stochastic gradient descent algorithms with weighted iterate-averaging that uses a single pass over the data. When the constraint set of the unknown parameter is bounded, we provided an explicit $O(1/k)$ upper bound on the convergence rate, showing 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. Under some mild conditions, we showed that $\rho \leq 4$ for all $d \geq 1$. We further improved the upper bound by showing that $\rho \leq 4/3$ for the case of $d = 1$ and when the parameter set is unbounded.

It should be noted that SGD with a constant step size does not converge to the global optimum in general [29], [30]. Thus, it is desirable to analyze the proposed PSGD-WA algorithm with decreasing step size under other loss functions (e.g., logistic regression) as a future research direction.

[3]Note that when $x_k = x$ for all k, then $\mu = x^2$, and $\lambda_k = \frac{1}{2x^2}\frac{\gamma}{\gamma+k}$, which is a special case of the step size in (6) when $d = 1$.

References

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

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

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

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

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

[17] 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.

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

[19] 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.

[20] 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.

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

[22] 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.

[23] 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.

[24] 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.

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

[26] R. Frostig, R. Ge, S. M. Kakade, and A. Sidford, "Competing with the empirical risk minimizer in a single pass," *arXiv preprint arXiv:1412.6606*, 2014.

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

[28] K. Cohen, A. Nedić, and R. Srikant, "On projected stochastic gradient descent algorithm with weighted averaging for least squares regression," *technical report, available at http://www.ee.bgu.ac.il/~yakovsec/*.

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

[30] 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.