On Accelerated Gradient Approximation for Least Square Regression with L_1-regularization

Yongquan Zhang
Department of Information and Mathematics Sciences
China Jiliang University
Hangzhou 310018, ZheJiang, China
Email: zyqmath@163.com

Jianyong Sun
School of Computer Science and Electrical Engineering
University of Essex
Wivenhoe Park, Colchester, CO4 3SQ, U.K.
Email: jysun@essex.ac.uk

***Abstract*—In this paper, we consider an online least square regression problem where the objective function is composed of a quadratic loss function and an L_1 regularization on model parameter. For each training sample, we propose to approximate the L_1 regularization by a convex function. This results in an overall convex approximation to the original objective function. We apply an efficient accelerated stochastic approximation algorithm to solve the approximation. The developed algorithm does not need to store previous samples, thus can reduce the space complexity. We further prove that the developed algorithm is guaranteed to converge to the global optimum with a convergence rate $O(\ln n/\sqrt{n})$ where n is the number of training samples. The proof is based on a weaker assumption than those applied in similar research work.**

I. Introduction

In supervised learning settings with many input features, over-fitting usually occurs when there are no ample training data. Regularization is a well-known solution to avoid over-fitting when there is only a small number of training examples, and/or when there are a large number of parameters to be learned. The L_1 regularization is often used for sparse representation, and has been shown to have good generalization capability (e.g. see [3]). The L_1-regularization proposed by Tibshirani [4] is especially useful because it selects variables according to the amount of penalization on the L_1-norm of the coefficients, in a manner that is less greedy than forward selection and backward deletion. Since then, the L_1 regularizer and its variants including SCAD [5], Adaptive Lasso [6], Elastic net [7], Stage-wise Lasso [8] and Dantzig selector [9], have become the dominantly-used tools for data analysis.

In the online learning context, some stochastic gradient methods has been successfully applied with L_1 regularization (e.g., Bottou and LeCunn [11]; Shalev-Shwartz et al. [12], [13], and Hu et al. [1]) recently. But they have shown that the classical stochastic gradient descent method cannot produce sparse solutions. Further, the algorithms become slow because of the introduction of the regularization. Therefore, it is appealing to develop fast online stochastic learning algorithms that can achieve sparsity.

Literature work to address the aforementioned problems mostly fall into the following two categories. First, Duchi et al. [16] suggested to projecting the L_1 regularization term onto a simplex, and then use projected sub-gradient method for convex optimization. Second, Langford et al. [17] proposed to use truncated gradient method, where a less aggressive strategy is adopted to remove non-zero parameters but with small weights periodically to prompt sparsity. Recently, Xiao [10] proposed a regularized dual averaging (RDA) method for stochastic online learning, which is an extension of the simple dual averaging scheme of Nesterov [21]. In the RDA, an auxiliary strongly convex function is introduced in the objective function of the regularized stochastic learning problem which includes a loss-function and a regularization term. He claimed that the RDA can exploit the structure of the regularized online learning problem more effectively. He further proved that the RDA with L_1 regularization can achieve a convergence rate $O(1/\sqrt{n})$ where n is the number of samples. However, in the RDA, at each iteration, previous solutions are to be stored for updating present solution.

In this paper, we develop a new stochastic online learning algorithms for regression with L_1 regularization, called the accelerated stochastic gradient descent methods (AC-SGD), which is able to address the sparsity problem and decrease the space complexity. The developed algorithm is based on the so-called accelerated stochastic approximation algorithm. A short literature reivew can be summarised as follows.

For a class of convex programming (CP) problems, Nesterov presented an accelerated gradient (AG) method in his work [18]. The accelerated gradient method has also been generalized by Beck and Teboulle [19], Tseng [20], and Nesterov [21], [22] for an emerging class of composite CP problems. In 2012, Lan [23] showed that the AG method is optimal for solving not only the smooth CP problems, but also general non-smooth and stochastic CP problems. The accelerated stochastic approximation (AC-SA) algorithm was proposed by Ghadimi and Lan [24], in which a properly modified Nesterov's optimal method for smooth CP is applied. Recently, they developed a generic accelerated stochastic approximation algorithmic framework, which can be specialized to yield optima or nearly optimal methods for strongly convex stochastic composite optimization problems [24], [25], [26], [27].

In this paper, we propose to approximate the L_1 regularization by a convex function. With a convex loss function, this

978-1-4799-7560-0/15/$31 ©2015 IEEE

978-1-4799-7560-0/15 $31.00 © 2015 IEEE
DOI 10.1109/SSCI.2015.221

will result in a convex approximation to the original objective function of the online regression learning. Thus the considered problem is also a convex programming problem. We then propose to apply the accelerated stochastic approximation algorithm to the considered problem, motivated by those mentioned work. Further, we prove, with weaker assumptions than a similar work in [2], that the developed algorithm guarantees the convergence to global optimum with a convergence rate $O(\ln n/\sqrt{n})$.

The rest of the paper is organized as follows. In Section II, we first give a brief introduction of the stochastic accelerated gradient algorithm, and present the analysis results on the convergence rate for the online least square regression with L_1 regularization. In Section 3, we present the comparison between the work with related work in the literature. Section 4 concludes the paper.

II. The Algorithm

In this section, we present the accelerated stochastic gradient algorithm for L_1 regularization least square regression. The objective function we considered in the paper is of the following form:

$$f(\theta) = \frac{1}{2}\mathbb{E}[(y - \langle\theta, x\rangle)^2] + \|\theta\|_1 \qquad (1)$$

where (x, y) is an input-output pair of data drawn from an (unknown) underlying distribution and θ is the model parameters, where $x \in \mathcal{F}$ and $y \in \mathbb{R}$. $\langle\theta, x\rangle$ denotes the inner product of the parameter θ and the decision variable x.

Before presenting the stochastic accelerated gradient algorithm for the regression problem, we make the following assumptions:

(a) $\mathcal{F}$ is a $d-$dimension Euclidean space, with $d \geq 1$.
(b) Let (X, d) be a compact metric space and let $Y = \mathbb{R}$. Let ρ be a probability distribution on $Z = \mathcal{F} \times Y$ and $(\mathcal{X}, \mathcal{Y})$ be corresponding random variable. Denote by $\mathbf{z} = \{z_i\}_{i=1}^k = \{(x_i, y_i)\}_{i=1}^k \in Z$ a set of random samples, which is independently drawn according to ρ.
(c) $\mathbb{E}\|x_k\|^2$ is finite, i.e., $\mathbb{E}\|x_k\|^2 \leq M$ for any $k \geq 1$.
(d) The global minimum of $f(\theta)$ is attained at a certain $\theta^* \in \mathcal{F}$.

The assumptions (a-d) are standard in stochastic approximation. In Bach et al. [2], they addressed the same problem as presented in Eq. (1). In their work, they made assumption on the covariance operator $\mathcal{H} = \mathbb{E}(x_k \bigotimes x_k)$ to be invertible for any $k \geq 1$, and that the operator $\mathbb{E}(x_k \bigotimes x_k)$ satisfies $\mathbb{E}[\xi_i \bigotimes \xi_i] \preceq \sigma^2\mathcal{H}$ and $\mathbb{E}(\|x_i\|^2 x_k \bigotimes x_k) \preceq R^2\mathcal{H}$ for a positive number. We do not require such rather strong assumptions in the analysis.

A. The accelerated gradient algorithm for regression learning

In the sequel, we let $\xi_k = (y_k - \langle\theta^*, x_k\rangle)\, x_k$ denote the residual. For any $k \geq 1$, we have $\mathbb{E}\xi_k = 0$. We also assume that $\mathbb{E}\xi_k^2 \leq \sigma^2$ for every k and $\overline{\xi}_k = \frac{1}{k}\sum_{i=1}^k \xi_i$.

Since $\|\theta\|_1$ is a non-differentiable function, we propose to approximate it by a smooth function for $\delta > 0$ defined as follows:

$$\begin{aligned} h(\theta, \delta) &= \frac{1}{(2\delta)^d}\int_{\theta_1-\delta}^{\theta_1+\delta}\cdots\int_{\theta_d-\delta}^{\theta_d+\delta}\|\mathbf{t}\|_1 dt_1\cdots dt_d \\ &= \frac{1}{2\delta}\int_{\theta_1-\delta}^{\theta_1+\delta}|t_1|dt_1 + \cdots + \frac{1}{2\delta}\int_{\theta_d-\delta}^{\theta_d+\delta}|t_d|dt_d \end{aligned}$$

where $\mathbf{t} \in \mathbb{R}^d$. It can be proved that Eq. (1) is convex. We have the following theorem 2.1 (Please see Appendix A for the proof).

Theorem 2.1: For $\theta \in \mathbb{R}^d$ and any $\delta > 0$, $h(\theta, \delta)$ is convex. Moreover, it can be easily proven that

$$\begin{aligned} \Big|\|\theta\|_1 - h(\theta,\delta)\Big| &\leq \left||\theta_1| - \frac{1}{2\delta}\int_{\theta_1-\delta}^{\theta_1+\delta}|t_1|dt_1\right| + \cdots \\ &\quad + \left||\theta_d| - \frac{1}{2\delta}\int_{\theta_d-\delta}^{\theta_d+\delta}|t_d|dt_d\right| \\ &\leq \omega(|\theta_1|, \delta) + \cdots + \omega(|\theta_d|, \delta), \end{aligned} \qquad (2)$$

where $\omega(\|\theta\|_1, \delta)$ denotes smoothness model of $\|\theta\|_1$. Properties of the smoothness model tell us that

$$\omega(|\theta_i|, \delta) \leq \delta, \qquad \text{for any } i = 1, 2, \ldots, d. \qquad (3)$$

From Eqs. (2) and (3), we have

$$f(\theta) \leq \frac{1}{2}\mathbb{E}[(y_k - \langle\theta, x_k\rangle)^2] + h(\theta, \delta) + d\delta. \qquad (4)$$

Since $h(\theta, \delta)$ is a convex function, it is easy to obtain the gradient of $h(\theta, \delta)$ with respect to θ:

$$\nabla h(\theta, \delta) = \left(\frac{|\theta_1 + \delta| - |\theta_1 - \delta|}{2\delta}, \ldots, \frac{|\theta_d + \delta| - |\theta_d - \delta|}{2\delta}\right)^\intercal$$

Since $\nabla h(\theta, \delta)$ satisfies

$$\|\nabla h(\theta, \delta) - \nabla h(\vartheta, \delta)\|_1 \leq \frac{\sum_i |\theta_i - \vartheta_i|}{\delta} = \frac{\|\theta - \vartheta\|_1}{\delta}$$

This implies that $\nabla h(\theta, \delta)$ is Lipschitz continuous with constant $\frac{1}{\delta}$. If we let

$$g(\theta) = \frac{1}{2}\mathbb{E}[(y_k - \langle\theta, x_k\rangle)^2]$$

then $g(\theta)$ is a convex function and its gradient w.r.t. θ is $\nabla g(\theta) = \mathbb{E}\left(\langle\theta, x_k\rangle x_k - y_k x_k\right)$. From Eq. (4), we know that

$$f(\theta) \leq g(\theta) + h(\theta, \delta) + d\delta,$$

Therefore, it can be seen that both $\nabla g(\theta)$ and $\nabla h(\theta, \delta)$ are Lipschitz continuous with constant

$$\frac{1}{\delta} + \mathbb{E}\|x_k\|^2 \leq M + \frac{1}{\delta} = L$$

In the sequel, we denote $G_L^{1,1}$ the class of convex functions on convex set X whose gradient is Lipschitz-continuous with constant L. It is well known that functions belonging to this class satisfy for any $\theta, \vartheta \in X$ and $\delta > 0$

$$\begin{aligned} g(\theta) + h(\theta, \delta) &\geq g(\vartheta) + h(\vartheta, \delta) + \\ &\quad \langle \nabla g(\vartheta) + \nabla h(\vartheta, \delta), \theta - \vartheta \rangle \quad (5) \\ g(\theta) + h(\theta, \delta) &\leq g(\vartheta) + h(\vartheta) + \frac{M\delta + 1}{2\delta}\|\theta - \vartheta\|_2^2 + \\ &\quad \langle \nabla g(\vartheta) + \nabla h(\vartheta), \theta - \vartheta \rangle \quad (6) \end{aligned}$$

From Eqs. (2), (3), (5) and (6), we know that

$$\begin{aligned} f(\theta) &\geq g(\vartheta) + h(\vartheta) + \\ &\quad \langle \nabla g(\vartheta) + \nabla h(\vartheta), \theta - \vartheta \rangle - d\delta \quad (7) \\ f(\theta) &\leq f(\vartheta) + \langle \nabla g(\vartheta) + \nabla h(\vartheta), \theta - \vartheta \rangle \\ &\quad + \frac{M\delta + 1}{2\delta}\|\theta - \vartheta\|_2^2 + 2d\delta. \quad (8) \end{aligned}$$

In the following, we let $\theta_0 \in \mathcal{F}$, $\{\alpha_k\}$ satisfy $\alpha_1 = 1$ and $\{\alpha_k > 0\}$ for any $k \geq 2$, $\{\beta_k > 0\}$, and $\{\lambda_k > 0\}$. Based on Eqs. (7)(8), the accelerated gradient algorithm for regression learning can be summarised in Alg. 1.

Algorithm 1 The accelerated stochastic gradient algorithm.

1: Set the initial $\theta_0^{ag} = \theta_0$ and $k = 1$

2: Update auxiliary variable θ_k^{md} by a convex combination between θ^{ag} and θ_{k-1} as

$$\theta_k^{md} = (1 - \alpha_k)\theta_{k-1}^{ag} + \alpha_k \theta_{k-1}. \quad (9)$$

3: Update θ_k as

$$\theta_k = \theta_{k-1} - \lambda_k \left(\nabla g(\theta_k^{md}) + \nabla h(\theta_k^{md})\right), \quad (10)$$

4: Update the auxlilary variable θ^{ag} as

$$\theta_k^{ag} = \theta_k^{md} - \beta_k \left(\nabla g(\theta_k^{md}) + \nabla h(\theta_k^{md}) + \bar{\xi}_k\right), \quad (11)$$

5: Set $k \leftarrow k + 1$ and go to step 2.

In Alg. 1, we introduce an auxiliary value θ_k^{md} at each step, which is updated as a linear combination between another auxiliary value θ^{ag} and previous estimation of the model parameter denoted as θ_{k-1} in Step 2. The model parameter is then updated in Step 3 with a parameter λ_k. In step 4, another auxiliary value θ_k^{ag} is introduced with the parameter β_k. The settings of the parameters α_k, β_k and λ_k are defined in the following section.

B. The Convergence Rate

To establish the convergence rate of the stochastic accelerated gradient algorithm, we need the following Lemma (see Lemma 1 in [25]).

Lemma 2.2: Let α_k be the step sizes in the accelerated gradient algorithm and the sequence $\{\eta_k\}$ satisfies

$$\eta_k = (1 - \alpha_k)\eta_{k-1} + \tau_k, \quad k = 1, 2, \ldots,$$

where

$$\Gamma_k = \begin{cases} 1, & k = 1, \\ (1 - \alpha_k)\Gamma_{k-1}, & k \geq 2. \end{cases} \quad (12)$$

Then we have

$$\eta_k \leq \Gamma_k \sum_{i=1}^{k} \frac{\tau_i}{\Gamma_i} \text{ for any } \ k \geq 1$$

Based on lemma 2.2, we present Theorem 2.3 which describes the convergence property of the accelerated gradient algorithm for the least-square regression with L_1 regularization. The proof of the theorem can be found in Appendix B.

Theorem 2.3: Let $\{\theta_k^{md}, \theta_k^{ag}\}$ be computed by the accelerated gradient algorithm and Γ_k be defined in (5) and assumptions (a-d) hold. If $\{\alpha_k\}$, $\{\beta_k\}$ and $\{\lambda_k\}$ are chosen such that

$$\alpha_k \lambda_k \leq \beta_k \leq \frac{\delta}{M\delta + 1}, \text{ and } \frac{\alpha_1}{\lambda_1 \Gamma_1} \geq \frac{\alpha_2}{\lambda_2 \Gamma_2} \geq \cdots,$$

then for any $n \geq 1$, we have

$$\begin{aligned} \mathbb{E}\Big[f\left(\theta_n^{ag}\right) - f(\theta^*)\Big] &\leq \frac{\Gamma_n}{2\lambda_1}\|\theta_0 - \theta^*\|^2 + \\ &\Gamma_n \sigma^2 \sum_{k=1}^{n} \frac{M\delta + 1}{2\delta \Gamma_k}\beta_k^2 + 4d\Gamma_n \sum_{k=1}^{n} \frac{\delta}{\Gamma_k}. \quad (13) \end{aligned}$$

In the following, we specialize the results of Theorem 2.3 for some particular selections of $\{\alpha_k\}, \{\beta_k\}$ and $\{\lambda_k\}$. Proof is available in Appendix C.

Corollary 2.4: Suppose that α_k, λ_k and β_k in the accelerated gradient algorithm for regression learning are set to

$$\begin{aligned} \alpha_k &= \frac{3}{2(k+1)}, \lambda_k = \frac{1}{M(k+1)\Gamma_k}, \\ \beta_k &= \frac{3}{2M(k+1)^2\Gamma_k}, \forall k \geq 1, \quad (14) \end{aligned}$$

for any $n \geq 1$, we have

$$\begin{aligned} \mathbb{E}\left[f\left(\theta_n^{ag}\right) - f(\theta^*)\right] &\leq \frac{e\|\theta_0 - \theta^*\|^2}{2M\sqrt{n}} + \\ &\frac{7e^2\sigma^2\left(2.5e^2 + \ln n + 1\right) + 64de^3\ln(n+1)}{14M\sqrt{n}}. \quad (15) \end{aligned}$$

With such corollary, we can conclude that the developed algorithm converges to the global optimum with a convergence rate $O(\ln n/\sqrt{n})$.

III. Comparisons with related work

In Section 2, we have discussed the accelerated stochastic approximation algorithms for least-square regression with L_1 regularization. We have derived the convergence rate $O\left(\ln n/\sqrt{n}\right)$ of accelerated stochastic approximation learning algorithms by using the convexity of the aim function. This rate is similar to some related work, but with weaker assumptions. For examples, in [10] and [27], the authors considered similar problems, while they obtained a convergence rate $O(1/\sqrt{n})$.

Our convergence analysis of accelerate stochastic learning algorithms with L_1 regularization is based on a similar analysis for stochastic composite optimization by Ghadimi and Lan in [27]. Beside the convergence analysis results, the other difference between our work and that of Ghadimi and Lan is that at each iteration, the parameters β_k, λ_k of the developed algorithm in [27] depends on the maximum iteration N. In our algorithm, we do not need this assumption.

The work that is most closely related to ours is that of Xiao [10], who consider regularized stochastic learning and online optimization problems. In their work, the objective function is considered as the sum of two convex terms: one is the loss function of the learning task, and the other is a simple regularization term such as L_1-norm for promoting sparsity as follows:

$$\min \{\mathbf{E}_z f(\omega, z) + \|\omega\|_1\}$$

where $\omega \in \mathbb{R}^d$ is the optimization variable, the sample $z = (x, y)$ is an input-output pair of data drawn from an (unknown) probability distribution. Both Xiao and we studied the convergence performance of regularized stochastic learning problems when the aim function f is strong-convexity function. The difference is that in the algorithm developed in [10], the average of previous solutions is used for the update while we only used the current solution.

Appendix

A. Proof of Theorem 2.1

Proof For any $\theta_1, \theta_2 \in X, \iota \in (0,1)$, let $\omega = \iota\theta + (1-\iota)\vartheta$, we have

$$\begin{aligned}
h(\omega,\delta) &= \frac{1}{2\delta}\int_{\omega-\delta}^{\omega+\delta}|t_1|dt_1 + \cdots + \frac{1}{2\delta}\int_{\omega-\delta}^{\omega+\delta}|t_d|dt_d \\
&= \frac{1}{2\delta}\int_{-\delta}^{\delta}|t_1-\omega|dt_1 + \cdots + \frac{1}{2\delta}\int_{-\delta}^{\delta}|t_d-\omega|dt_d \\
&\le \frac{\iota}{2\delta}\int_{-\delta}^{\delta}|t_1-\theta_1|dt_1 + \cdots + \frac{\iota}{2\delta}\int_{-\delta}^{\delta}|t_d-\theta_d|dt_d + \\
&\quad \frac{1-\iota}{2\delta}\int_{-\delta}^{\delta}|t_1-\vartheta_1|dt_1 + \cdots + \frac{1-\iota}{2\delta}\int_{-\delta}^{\delta}|t_d-\vartheta_d|dt_d \\
&= \frac{\iota}{2\delta}\int_{\theta_1-\delta}^{\theta_1+\delta}|t_1|dt_1 + \cdots + \frac{\iota}{2\delta}\int_{\theta_d-\delta}^{\theta_d+\delta}|t_d|dt_d + \\
&\quad \frac{1-\iota}{2\delta}\int_{\vartheta_1-\delta}^{\vartheta_1+\delta}|t_1|dt_1 + \cdots + \frac{1-\iota}{2\delta}\int_{\vartheta_d-\delta}^{\vartheta_d+\delta}|t_d|dt_d \\
&= \iota h(\theta) + (1-\iota)h(\vartheta).
\end{aligned}$$

This completes the proof. ∎

B. Proof of Theorem 2.3

Proof Let $F(\theta) = f(\theta) + g(\theta)$. From Eqs. (7), we have

$$\begin{aligned}
f(\theta_k^{ag}) &\le f(\theta_k^{md}) + \left\langle \nabla F(\theta_k^{md}), \theta_k^{ag} - \theta_k^{md}\right\rangle + \\
&\quad \frac{M\delta+1}{2\delta}\|\theta_k^{ag} - \theta_k^{md}\|_2^2 + 2d\delta \\
&= f(\theta_k^{md}) - \beta_k\|\nabla F(\theta_k^{md})\|^2 - \beta_k\langle \nabla F(\theta_k^{md}), \xi_k\rangle + \\
&\quad \frac{M\delta+1}{2\delta}\beta_k^2\|\nabla F(\theta_k^{md}) + \xi_k\|^2 + 2d\delta.
\end{aligned}$$

From Eqs. (7),(8), we have

$$\begin{aligned}
& f(\theta_k^{md}) - [(1-\alpha_k)f(\theta_{k-1}^{ag}) + \alpha_k f(\theta)] \\
= \; & \alpha_k[f(\theta_k^{md}) - f(\theta)] + (1-\alpha_k)[f(\theta_k^{md}) - f(\theta_{k-1}^{ag})] \\
\le \; & \alpha_k\langle \nabla F(\theta_k^{md}), \theta_k^{md} - \theta\rangle + 2d\delta + \\
& (1-\alpha_k)\langle \nabla F(\theta_k^{md}), \theta_k^{md} - \theta_{k-1}^{ag}\rangle \\
= \; & \langle \nabla F(\theta_k^{md}), \alpha_k(\theta_k^{md} - \theta) + (1-\alpha_k)(\theta_k^{md} - \theta_{k-1}^{ag})\rangle + 2d\delta \\
= \; & \alpha_k\langle \nabla F(\theta_k^{md}), \theta_{k-1} - \theta\rangle + 2d\delta.
\end{aligned}$$

Thus we can obtain

$$\begin{aligned}
f(\theta_k^{ag}) &\le (1-\alpha_k)f(\theta_{k-1}^{ag}) + \alpha_k f(\theta) + 4d\delta + \\
&\quad \alpha_k\langle \nabla F(\theta_k^{md}), \theta_{k-1} - \theta\rangle - \beta_k\|\nabla F(\theta_k^{md})\|^2 \\
&\quad - \beta_k\langle \nabla F(\theta_k^{md}), \xi_k\rangle + \frac{M\delta+1}{2\delta}\beta_k^2\|\nabla F(\theta_k^{md}) + \xi_k\|^2.
\end{aligned}$$

It follows from Eq. (10) that

$$\begin{aligned}
\|\theta_k - \theta\|^2 &= \left\|\theta_{k-1} - \lambda_k \nabla F\left(\theta_k^{md}\right) - \theta\right\|^2 \\
&= -2\lambda_k\left\langle \nabla F(\theta_k^{md}), \theta_{k-1} - \theta\right\rangle + \\
&\quad \|\theta_{k-1} - \theta\|^2 + \lambda_k^2\left\|\nabla F(\theta_k^{md})\right\|^2.
\end{aligned}$$

Then we have

$$\begin{aligned}
\left\langle \nabla F(\theta_k^{md}), \theta_{k-1} - \theta\right\rangle &= \frac{\lambda_k}{2}\left\|\nabla F(\theta_k^{md})\right\|^2 \\
&\quad + \frac{1}{2\lambda_k}\Big[\|\theta_{k-1} - \theta\|^2 - \|\theta_k - \theta\|^2\Big].
\end{aligned}$$

While

$$\begin{aligned}
\left\|\nabla F(\theta_k^{md}) + \xi_k\right\|^2 &= \left\|\nabla F(\theta_k^{md})\right\|^2 + \|\xi_k\|^2 + \\
&\quad 2\left\langle \nabla F(\theta_k^{md}), \xi_k\right\rangle.
\end{aligned}$$

Combining the above two inequalities, we obtain

$$\begin{aligned}
f(\theta_k^{ag}) &\le (1-\alpha_k)f(\theta_{k-1}^{ag}) + \alpha_k f(\theta) + \\
&\quad \frac{\alpha_k}{2\lambda_k}\left[\|\theta_{k-1} - \theta\|^2 - \|\theta_k - \theta\|^2\right] \\
&\quad - \beta_k\left(1 - \frac{\lambda_k\alpha_k}{2\beta_k} - \frac{M\delta+1}{2\delta}\beta_k\right)\|\nabla F(\theta_k^{md})\|^2 + \\
&\quad \frac{M\delta+1}{2\delta}\beta_k^2\|\xi_k\|^2 + 4d\delta \\
&\quad + \left\langle \xi_k, \left(\frac{M\delta+1}{\delta}\beta_k^2 - \beta_k\right)\nabla F(\theta_k^{md})\right\rangle.
\end{aligned}$$

The above inequality is equal to

$$\begin{aligned}
f(\theta_k^{ag}) - f(\theta) &\le (1-\alpha_k)[f(\theta_{k-1}^{ag}) - f(\theta)] + \\
&\quad \frac{\alpha_k}{2\lambda_k}\left[\|\theta_{k-1} - \theta\|^2 - \|\theta_k - \theta\|^2\right] \\
&\quad - \beta_k\left(1 - \frac{\lambda_k\alpha_k}{2\beta_k} - \frac{M\delta+1}{2\delta}\beta_k\right)\|\nabla F(\theta_k^{md})\|^2 + \\
&\quad \frac{M\delta+1}{2\delta}\beta_k^2\|\xi_k\|^2 + 4d\delta \\
&\quad + \left\langle \xi_k, \left(\frac{M\delta+1}{\delta}\beta_k^2 - \beta_k\right)\nabla F(\theta_k^{md})\right\rangle.
\end{aligned}$$

Using Lemma 1, we have

$$f(\theta_n^{ag}) - f(\theta) \le \Gamma_n \sum_{k=1}^{n} \frac{\alpha_k}{2\lambda_k \Gamma_k} \left[\|\theta_{k-1} - \theta\|^2 - \|\theta_k - \theta\|^2 \right]$$
$$- \Gamma_n \sum_{k=1}^{n} \frac{\beta_k}{\Gamma_k} \left(1 - \frac{\lambda_k \alpha_k}{2\beta_k} - \frac{M\delta + 1}{2\delta} \beta_k \right) \|\nabla F(\theta_k^{md})\|^2$$
$$+ \Gamma_n \sum_{k=1}^{n} \frac{M\delta+1}{2\delta\Gamma_k} \beta_k^2 \|\xi_k\|^2 + 4d\Gamma_n \sum_{k=1}^{n} \frac{\delta}{\Gamma_k} +$$
$$\Gamma_n \sum_{k=1}^{n} \frac{1}{\Gamma_k} \left\langle \xi_k, \left(\frac{M\delta+1}{\delta} \beta_k^2 - \beta_k \right) \nabla F(\theta_k^{md}) \right\rangle.$$

Since

$$\frac{\alpha_1}{\lambda_1 \Gamma_1} = \frac{\alpha_2}{\lambda_2 \Gamma_2} = \cdots, \alpha_1 = \Gamma_1 = 1$$

then

$$\sum_{k=1}^{n} \frac{\alpha_k}{2\lambda_k \Gamma_k} \left[\|\theta_{k-1} - \theta\|^2 - \|\theta_k - \theta\|^2 \right] \le$$
$$\frac{\alpha_1}{2\lambda_1 \Gamma_1} \left[\|\theta_0 - \theta\|^2 - \|\theta_n - \theta\|^2 \right] \le \frac{1}{2\lambda_1} \|\theta_0 - \theta\|^2.$$

So we obtain

$$f(\theta_n^{ag}) - f(\theta) \le \Gamma_n \sum_{k=1}^{n} \frac{M\delta+1}{2\delta\Gamma_k} \beta_k^2 \|\xi_k\|^2 +$$
$$\frac{\Gamma_n}{2\lambda_1} \|\theta_0 - \theta\|^2 + 4d\Gamma_n \sum_{k=1}^{n} \frac{\delta}{\Gamma_k}$$
$$+ \Gamma_n \sum_{k=1}^{n} \frac{1}{\Gamma_k} \left\langle \xi_k, \left(\frac{M\delta+1}{\delta} \beta_k^2 - \beta_k \right) \nabla F(\theta_k^{md}) \right\rangle,$$

where the inequality follows from the assumption

$$\alpha_k \lambda_k \le \beta_k \le \frac{\delta}{M\delta + 1} \le \delta.$$

Under the assumption (d), we have $\mathbb{E}\xi_k = 0, \quad \mathbb{E}\xi_k^2 = \sigma^2$. Taking expectation on both sides of the inequality above with respect to (x_i, y_i), we obtain for $\theta \in \mathbb{R}^d$,

$$\mathbb{E}\left[f(\theta_n^{ag}) - f(\theta) \right] \le \Gamma_n \sigma^2 \sum_{k=1}^{n} \frac{M\delta+1}{2\delta\Gamma_k} \beta_k^2 +$$
$$\frac{\Gamma_n}{2\lambda_1} \|\theta_0 - \theta\|^2 + 4d\Gamma_n \sum_{k=1}^{n} \frac{\delta}{\Gamma_k}.$$

Now, fixing $\theta = \theta^*$, we have

$$\mathbb{E}\left[f(\theta_n^{ag}) - f(\theta^*) \right] \le \Gamma_n \sigma^2 \sum_{k=1}^{n} \frac{M\delta+1}{2\delta k \Gamma_k} \beta_k^2 +$$
$$\frac{\Gamma_n}{2\lambda_1} \|\theta_0 - \theta^*\|^2 + 4d\Gamma_n \sum_{k=1}^{n} \frac{\delta}{\Gamma_k}.$$

This completes the proof. ∎

C. *Proof of Corollary 2.4*

Proof In Eqs. (12) and (14), we have for $k \ge 2$

$$\Gamma_k = (1 - \alpha_k)\Gamma_{k-1}$$
$$= \frac{2k+1}{2(k+1)} \frac{2(k-1)+1}{2k} \times \cdots \times \frac{1}{2}\Gamma_1$$
$$= \frac{1}{2(k+1)} \times \left(1 + \frac{1}{2k}\right) \times \cdots \times \left(1 + \frac{1}{2}\right).$$

It is now to estimate

$$\left(1 + \frac{1}{2k}\right) \times \left(1 + \frac{1}{2(k-1)}\right) \times \cdots \times \left(1 + \frac{1}{2}\right)$$
$$= \exp\left\{ \ln\left(1 + \frac{1}{2k}\right) + \cdots + \ln\left(1 + \frac{1}{2}\right) \right\}.$$

While

$$\int_1^{k+1} \ln\left(1 + \frac{1}{2x}\right) dx =$$
$$\int_k^{k+1} \ln\left(1 + \frac{1}{2x}\right) dx + \cdots \int_1^2 \ln\left(1 + \frac{1}{2x}\right) dx$$
$$\le \ln\left(1 + \frac{1}{2k}\right) + \cdots + \ln\left(1 + \frac{1}{2}\right)$$
$$\le \int_1^k \ln\left(1 + \frac{1}{2x}\right) dx + \ln\left(1 + \frac{1}{2}\right)$$
$$\le \int_1^k \ln\left(1 + \frac{1}{2x}\right) dx + 1.$$

Taylor expansion tells us that

$$\ln(1+t) = t - \frac{t^2}{2} + \frac{t^3}{3} + \cdots + (-1)^{n-1}\frac{t^n}{n} + \cdots$$

For any $0 < t \le 1$, we have

$$\ln(1+t) = t - \sum_{n=1}^{\infty} \left(\frac{t^{2n}}{2n} - \frac{t^{2n+1}}{2n+1} \right) \quad \le \quad t$$
$$\ln(1+t) = t - \frac{t^2}{2} + \sum_{n=1}^{\infty} \left(\frac{t^{2n+1}}{2n+1} - \frac{t^{2n+2}}{2n+2} \right) \quad \ge \quad t - \frac{t^2}{2}$$

Then we have

$$\begin{aligned}
\frac{1}{2}\ln(k+1) - 1 \quad &\le \quad \frac{1}{2}\ln(k+1) - \frac{1}{4} \\
&\le \quad \int_1^{k+1} \left(\frac{1}{2x} - \frac{1}{8x^2} \right) dx \\
&\le \quad \int_1^{k+1} \ln\left(1 + \frac{1}{2x}\right) dx \\
&\le \quad \ln\left(1 + \frac{1}{2k}\right) + \cdots + \ln\left(1 + \frac{1}{2}\right) \\
&\le \quad \int_1^k \ln\left(1 + \frac{1}{2x}\right) dx + 1 \\
&\le \quad \int_1^k \frac{1}{2x} dx + 1 = \frac{1}{2}\ln k + 1.
\end{aligned}$$

So we have

$$\begin{aligned} e^{-1}(k+1)^{\frac{1}{2}} &= e^{\frac{1}{2}\ln(k+1)-1} \\ &\leq \left(1+\frac{1}{2k}\right)\times\cdots\times\left(1+\frac{1}{2}\right) \\ &= \exp\left\{\ln\left(1+\frac{1}{2k}\right)+\cdots+\ln\left(1+\frac{1}{2}\right)\right\} \\ &\leq e^{\frac{1}{2}\ln k+1} = ek^{\frac{1}{2}}. \end{aligned}$$

We obtain

$$\frac{1}{2e(k+1)^{\frac{1}{2}}} \leq \Gamma_k \leq \frac{ek^{\frac{1}{2}}}{2(k+1)} \leq \frac{e}{2k^{\frac{1}{2}}}.$$

It is easy to check

$$\alpha_k\lambda_k = \frac{1}{2M(k+1)^2\Gamma_k} \leq \beta_k \leq \frac{1}{M}$$

and

$$\frac{\alpha_1}{\lambda_1\Gamma_1} = \frac{\alpha_2}{\lambda_2\Gamma_2} = \cdots = \frac{M}{2}.$$

Then we obtain

$$\begin{aligned} \Gamma_n\sigma^2\sum_{k=1}^{n}\frac{M\delta_k+1}{2\delta_k\Gamma_k}\beta_k^2 &\leq \frac{e\sigma^2}{4\sqrt{n}}\sum_{k=1}^{n}\frac{M\beta_k^2+\beta_k}{\Gamma_k} \\ &\leq \frac{e\sigma^2}{4\sqrt{n}}\sum_{k=1}^{n}\left(\frac{8Me^3(k+1)^{3/2}}{4M^2(k+1)^4}+\frac{4e(k+1)}{2M(k+1)^2}\right) \\ &\leq \frac{e^4\sigma^2}{2M\sqrt{n}}\sum_{k=1}^{n}\frac{1}{(k+1)^{5/2}}+\frac{e^2\sigma^2}{2M\sqrt{n}}\sum_{k=1}^{n}\frac{1}{k+1} \\ &\leq \frac{e^4\sigma^2}{2M\sqrt{n}}\left(\int_1^n\frac{1}{x^{5/2}}dx+1\right)+\frac{e^2\sigma^2}{2M\sqrt{n}}\left(\int_1^n\frac{1}{x}dx+1\right) \\ &\leq \frac{e^4\sigma^2}{2M\sqrt{n}}\left(\frac{5}{2}-\frac{2}{3}n^{-3/2}\right)+\frac{e^2\sigma^2}{2M\sqrt{n}}(\ln n+1) \\ &\leq \frac{e^2\sigma^2\left(\frac{5}{2}e^2+\ln n+1\right)}{2M\sqrt{n}}. \end{aligned}$$

We also obtain

$$\begin{aligned} \Gamma_n\sum_{k=1}^{n}\frac{\delta_k}{\Gamma_k} &\leq \frac{e}{2\sqrt{n}}\sum_{k=1}^{n}\frac{2e\sqrt{k+1}\beta_k}{1-M\beta_k} \\ &\leq \frac{e^3}{M\sqrt{n}}\sum_{k=1}^{n}\frac{1}{k+1}\times\frac{1}{1-\frac{1}{2(k+1)^2}} \\ &\leq \frac{8e^3}{7M\sqrt{n}}\sum_{k=1}^{n}\frac{1}{k+1} \\ &\leq \frac{8e^3}{7M\sqrt{n}}\int_0^n\frac{1}{x+1}dx \leq \frac{8e^3\ln(n+1)}{7M\sqrt{n}} \end{aligned}$$

From the result of Theorem 2.3, we have

$$\begin{aligned} \mathbb{E}\left[f\left(\theta_n^{ag}\right)-f(\theta^*)\right] &\leq \frac{e\|\theta_0-\theta^*\|^2}{2\lambda_1\sqrt{n}}+\frac{32de^3\ln(n+1)}{7M\sqrt{n}} \\ &+\frac{\sigma^2e^2(2.5e^2+\ln n+1)}{2M\sqrt{n}} \\ &= \frac{e\|\theta_0-\theta^*\|^2}{2\lambda_1\sqrt{n}}+ \\ &\frac{7e^2\sigma^2(2.5e^2+\ln n+1)+64de^3\ln(n+1)}{14M\sqrt{n}} \end{aligned}$$

The proof of Corollary 1 is completed. ∎

ACKNOWLEDGEMENT

YZ was supported by the Natural Science Foundation of Zhejiang Province of China (No. Q12A010096) and the National Natural Science Foundation of China (No. 11301494 and 61573326). JS was supported by the National Natural Science Foundation of China (No. 61273313 and 61175063), and the Key Science and Technology Project of Wuhan under Grant No. 2014010202010108.

REFERENCES

[1] Hu, C. and James, T.K. and Pan, W. Accelerated Gradient Methods for Stochastic Optimization and Online Learning, Advances in Neural Information Processing Systems (NIPS), 2009.
[2] Bach, F. and Moulines, E. Non-strongly-convex smooth stochastic approximation with convergence rate O(1/n), Advances in Neural Information Processing Systems (NIPS), 2013.
[3] Ng, A. Y. Feature selection, l_1 vs. l_2 regularization, and rotational invariance. In International Conference on Machine Learning. 2004.
[4] Tibshirani R. Regression shrinkage and selection via the Lasso. J Royal Statist Soc B, 58: 267-288, 1996.
[5] Fan J., Heng P. Nonconcave penalty likelihood with a diverging number of parameters. Ann Statist, 32: 928-961, 2004.
[6] Zou H. The adaptive Lasso and its oracle properties. J Amer Statist Assoc, 101: 1418-1429, 2006.
[7] Zou H. and Hastie T. Regularization and variable selection via the elastic net. J Royal Statist Soc B, 67: 301-320, 2005.
[8] Zhao P. and Yu B. Stagewise Lasso. Journal of Machine Learning Research, 8: 2701-2726, 2007.
[9] Candes E. and Tao T. The Dantzig selector: Statistical estimation when p is much larger than n. Ann Statist, 35: 231-2351, 2007.
[10] Xiao, L. Dual averaging method for regularized stochastic learning and online optimization. Journal of Machine Learning Research, 11: 2543-2596, 2010.
[11] Bottou L. and LeCunn Y. On-line learning for very large datasets. Applied Stochastic Models in Business and Industry, 21(2): 137-151, 2005.
[12] Shalev-Shwartz, S. Online Learning: Theory, Algorithms, and Applications. PhD thesis, The Hebrew University, 2007.
[13] Shalev-Shwartz, S. and Tewari, A. Stochastic Methods for L_1-regularized Loss Minimization. Journal of Machine Learning Research, 12: 1865-1892, 2011.
[14] Bottou, L. and Bousquet, O. The tradeoffs of large scale learning. In Advances in Neural Information Processing Systems 20, pages 161-168, 2008.
[15] Shalev-Shwartz, S. and Srebro, N. SVM optimization: Inverse dependence on training set size. In International Conference on Machine Learning, pages 928-935, 2008.
[16] Duchi J., Shalev-Shwartz, S., Singer, Y. and Chandra, T.. Efficient projections onto the L_1-ball for learning in high dimensions. In International Conference on Machine Learning, pages 272-279, 2008.
[17] Langford, J. Li, L. and Zhang, T. Sparse online learning via truncated gradient. In Advances in Neural Information Processing Systems 21, pages 905-912, 2009.

[18] Nesterov, Y. E. A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. Doklady AN SSSR, 269: 543-547, 1983.
[19] Beck, A. and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sciences, 2: 183-202, 2009.
[20] Tseng, P. On accelerated proximal gradient methods for convex-concave optimization. Manuscript, University of Washington, Seattle, May 2008.
[21] Nesterov, Y. E. Smooth minimization of nonsmooth functions. Mathematical Programming, 103: 127-152, 2005.
[22] Nesterov, Y. E. Gradient methods for minimizing composite objective functions. Technical report, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain, September 2007.
[23] Lan, G. An optimal method for stochastic composite optimization. Mathematical Programming, 133(1): 365-397, 2012.
[24] Ghadimi, S. and Lan, G. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, I: a generic algorithmic framework. SIAM Journal on Optimization, 22(4): 1469-1492, 2012.
[25] Ghadimi, S. and Lan, G. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. SIAM Journal on Optimization, 23(4): 2341-2368, 2013.
[26] Ghadimi, S., Lan, G. and Zhang, H. Mini-batch stochastic approximation methods for constrained nonconvex stochastic programming. Manuscript, Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611, USA, August 2013.
[27] Ghadimi, S. and Lan, G. Accelerated Gradient Methods for Nonconvex Nonlinear and Stochastic Programming. Math. Program., Ser. A., 2015