Fast Rate Analysis of Some Stochastic Optimization Algorithms

Chao Qu A0117143@U.NUS.EDU
Department of Mechanical Engineering, National University of Singapore

Huan Xu ISEXUH@NUS.EDU.SG
Department of Industrial and Systems Engineering, National University of Singapore

Chong Jin Ong MPEONGCJ@NUS.EDU.SG
Department of Mechanical Engineering , National University of Singapore

Abstract

In this paper, we revisit three fundamental and popular stochastic optimization algorithms (namely, Online Proximal Gradient, Regularized Dual Averaging method and ADMM with online proximal gradient) and analyze their convergence speed under conditions weaker than those in literature. In particular, previous works showed that these algorithms converge at a rate of $O(\ln T/T)$ when the loss function is strongly convex, and $O(1/\sqrt{T})$ in the weakly convex case. In contrast, we relax the strong convexity assumption of the loss function, and show that the algorithms converge at a rate $O(\ln T/T)$ if the *expectation* of the loss function is *locally* strongly convex. This is a much weaker assumption and is satisfied by many practical formulations including Lasso and Logistic Regression. Our analysis thus extends the applicability of these three methods, as well as provides a general recipe for improving analysis of convergence rate for stochastic and online optimization algorithms.

1. Introduction

The last decade has witnessed the surge of attention in big data: learning and decision tasks involving datasets with unprecedented size – e.g., data from computational biology, video, social networks – are becoming ubiquitous. Big data brings in severe challenges: the memory cannot fit the size of data, the computation time can be prohibitively long, etc. A popular and powerful tool to overcome these challenges is stochastic and online optimization methods, as they draw one data point at a time (hence mitigate the storage issue), and update the variable to optimize with low complexity at each iteration (Shalev-Shwartz, 2011; Zhang, 2004). Yet, stochastic optimization methods may suffer from slow convergence. Take the (arguably simplest) stochastic subgradient method (SGD) as an example. SGD converges at a rate of $O(1/\sqrt{T})$, and hence requires a significant number of iterations if an accurate solution is in need (Kushner & Yin, 2003).

To solve problems in the high-dimensional (i.e., $p \ll n$ setting), various formulations have been proposed based on the idea of exploiting the lower dimensional structure such as sparsity via regularization. For example, ℓ_1 norm regularization is widely used to obtain sparse solutions. SGD cannot efficiently exploit the structure of such regularized formulations. Fortunately, several algorithms have been developed to successfully address this setting. Duchi's Forward-Backward-Splitting algorithm (Duchi et al., 2010; Singer & Duchi, 2009), also termed online proximal gradient (OPG), is an online version of the celebrated proximal gradient method (Combettes & Wajs, 2005). Xiao's Regularized Dual Averaging method (RDA) (Xiao, 2009) extends Nesterov's work (Nesterov, 2009) into the online and regularized version. Based upon these two fundamental algorithms, several variants have been developed. Notably, Suzuki considers new variants of online ADMM with online proximal gradient descent type method (OPG-ADMM) and regularized dual averaging type method (RDA-ADMM) (Suzuki, 2013). OPG-ADMM is also independently developed by Ouyang et al (Ouyang et al., 2013). It can solve problems with structured sparsity regularization such as overlapped group lasso (Jacob et al., 2009) or low rank tensor estimation(Signoretto et al., 2010).

This paper studies the convergence speed of OPG, RDA and OPG-ADMM. All these three methods are known to achieve a convergence rate of $O(1/\sqrt{T})$ when the function to optimize is (weakly) convex (Duchi et al., 2010; Xiao,

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

2009; Suzuki, 2013). In contrast, when the loss function or the regularization is strongly convex, they achieve a fast convergence rate of $O(\ln T/T)$. Yet, it has been observed in practice that in many weakly convex problems these algorithms perform better than what the theory predicts, indicating room of improvement for analysis under this case. We revisit these three methods in this paper and present some new results about their convergence speed in the weakly convex case. In particular, we show that the convergence rate of $O(\ln T/T)$ is achievable if the *expectation* of the loss function is *locally strongly convex* around the optimal solution. This is a much weaker assumption than the standard assumption that the loss function is strongly convex, and is satisfied by many practical formulations. Some examples *par excellence* are the renowned Lasso and logistic regression, where loss functions are $\|y - x^T\theta\|_2^2$ and $\ln(1 + \exp(-y(x^T\theta)))$ respectively, which are obviously not strongly convex. However, under mild conditions, its expectation is indeed strongly convex. We remark that our proof technique is very general: it applies to all three methods we study, and we believe can be easily adapted to analyzing other online and stochastic optimization methods.

Before concluding this section, we discuss some relevant literature. Recently, some approaches without strongly convex assumption on loss function to achieve convergence rate $O(1/T)$ has been proposed (Rakhlin et al., 2012; Bach & Moulines, 2013; Zhong & Kwok, 2013; Bach, 2014). Rakhlin et al. (2012) analyze SGD when the expectation of loss function is strongly convex, while our analysis is on more complex and general algorithms (OPG, RDA and OPG-ADMM). Bach & Moulines (2013) propose a novel stochastic gradient method to solve (unregularized) least-squares regression and logistic regression, under a smoothness assumption of the loss function. Bach (2014) exploits local strong convexity of the objective function, however it needs the objective is three-times differentiable. Zhong & Kwok (2013) develop an ADMM type method using historical gradients and hence require extra memory to store gradients. These works are under different conditions from ours.

2. Problem setup and notations

We consider the following stochastic learning problem.

$$\min_{w\in\Omega} G(w) := E_z f(w, z) + r(w), \tag{1}$$

where w is the variable to optimize, and z is an input-output pair which is generated from an unknown distribution. The loss function $f(w, z)$ is convex with respect to w for any z. As an example, one commonly used loss function is the least squares $f(w, (x, y)) = (y - w^T x)^2$,where $z = (x, y)$. The set Ω is a compact convex set, and $r(w)$ is a convex regularization function. Notice that we make no assumption on $f(\cdot, z)$ and $r(\cdot)$ beyond being convex. We further assume that $F(w) := E_z f(w, z)$ is strongly convex around $w^* := \arg\min_{w\in\Omega} G(w)$, i.e., there exists $\beta > 0$ such that $F(w) - F(w^*) \geq \frac{\beta}{2}\|w - w^*\|_2^2 + \langle \nabla F(w^*), w - w^*\rangle$. We will see below that this condition is indeed implied by the condition that $F(\cdot)$ is strongly convex in a neighborhood of w^*.

As the distribution of z is unknown, a common approach to solve the learning problem is to approximate the expectation using a finite set of observation and to minimize the empirical loss

$$\min_{w\in\Omega} \frac{1}{T}\sum_{t=1}^{T} f(w, z_t) + r(w), \tag{2}$$

where $f(w, z_t)$ is a convex loss function associated with a data point z_t, and $\{z_t\}_{t=1}^T$ are drawn from the underlying distribution.

In the traditional batch learning, we have to access the whole data set, e.g., computing the gradient of the objective function in (2), which is impossible in the big data setting. In contrast, the stochastic optimization algorithm is a promising approach. It sequentially draw the data and optimize the decision variable based upon the last observation. Throughout the paper, we use the subscript to denote the variable at a certain time step, e.g., w_t, z_{t+1}. Given a function $f(\cdot, z)$, $\partial f(w, z)$ denotes its subdifferential set evaluated at w, and $f'(w, z)$ denotes a particular subgradient in this set.

We first present formally the assumptions needed for our results.

Assumption 1.

- Both f and r are convex, and Ω is a convex compact set with radius R.
- The subgradient of f is bounded, i.e., there exists a constant L, such that $\|f'(w_t, z_t)\|_2 \leq L$.
- F is *Locally Strongly Convex*: there exists $\beta > 0$ such that $F(w) - F(w^*) \geq \frac{\beta}{2}\|w - w^*\|_2^2 + \langle \nabla F(w^*), w - w^*\rangle$, where $w^* := \arg\min_{w\in\Omega} G(w)$.

Some remarks on the assumption are in order.

1. Local strong convexity of F is clearly a much weaker assumption than $f(w, z)$ being strongly convex – the latter indeed immediately implies the former.
2. Our condition of local strong convexity is a strictly weaker condition than F being strongly convex in a neighborhood of w^* (say with radius r): it is indeed implied by the latter. To see this, by strong convexity in the neighborhood, we have for all $\hat{w}$ in the neighborhood,

$$F(\hat{w}) - F(w^*) \geq \gamma\|\hat{w} - w^*\|_2^2 + \langle \nabla F(w^*), \hat{w} - w^*\rangle.$$

Now notice w is in the compact set Ω with radius R. Let $\hat{w}$ be the furthest point in the neighborhood of w^* on the line segment between w and w^*, we have

$$F(w)-F(w^*) \geq (r^2/R^2)\gamma\|w-w^*\|_2^2+\langle\nabla F(w^*), w-w^*\rangle. \tag{3}$$

Notice (3) is indeed our definition of local strong convexity.

The rest of the paper focuses on analyzing the convergence of stochastic optimization algorithms under Assumption 1. In specific, we will review OPG, RDA, and OPG-ADMM and establish new convergence results with the locally strong convexity of $F(w)$.

3. Online Proximal Gradient

This section is devoted to the study of Online Proximal Gradient algorithm (Duchi et al., 2010). We first briefly review OPG, and then present our new result that the algorithm converges at a rate $O(\ln T/T)$ under locally strong convexity of $F(w)$. Finally, we provide a roadmap of the proof and discuss the general insight to obtain this new result.

OPG iteratively solves the following problem

$$w_{t+1} = \arg\min_{w\in\Omega} \frac{1}{2}\|w_t-w\|_2^2+\eta_t\langle f'(w_t,z_t), w\rangle+\eta_t r(w),$$

where η_t is a step size parameter. Note that here we use $\frac{1}{2}\|w_t-w\|_2^2$ for simplicity. This can be replaced by a more general term, i.e., Bregman divergence, and the analysis is identical. The output of the algorithm is $\bar{w}_T$, where $\bar{w}_T = \frac{1}{T}\sum_{t=1}^T w_t$.

To establish the convergence rate of OPG, standard technique (Duchi et al., 2010; Xiao, 2009; Suzuki, 2013) first defines the regret

$$Reg_T := \sum_{t=1}^T (f(w_t,z_t)+r(w_t))-\sum_{t=1}^T (f(w^*,z_t)+r(w^*))$$

and

$$\begin{aligned} Diff_T :&= \sum_{t=1}^T (G(w_t)-G(w^*)) \\ &= \sum_{t=1}^T (F(w_t)+r(w_t)) - T(F(w^*)+r(w^*)), \end{aligned}$$

where $G(w) := F(w)+r(w)$. Recall that $F(w) = E_z f(w,z)$, w^* is the optimal solution of $G(w)$, and the goal is to bound $Diff_T$. Duchi et al. (2010) shows when $f(w,z)$ is weakly convex, Reg_T can be upper bounded by $O(\sqrt{T})$. This is in sharp contrast to the strongly convex case where the upper bound is $O(\ln T)$. Then by standard technique, one can convert this bound of the regret to the convergence rate (Cesa-Bianchi et al., 2004; Kakade & Tewari, 2009). We now show that by a refined analysis, we can relax the strong convexity assumption of $f(w,z)$ yet still achieve $O(\ln T/T)$ convergence rate. We present our main theorem in the following subsection.

3.1. Stochastic Convergence Result of OPG

We now present the main result of this section, namely the convergence of OPG to solve (1) under weaker assumption. As standard, the convergence rate is for the average value of w_t generated by OPG.

Theorem 1. *If $f(w,z)$ is bounded by B, $F(w)$ is a locally strongly convex function with parameter β around w^*, assume there is a constant L such that $\|f'(w_t,z_t)\|_2 \leq L$. We set $\eta_t = \frac{2}{\beta t}$, let $\{w_t\}_{t=1}^T$ be the sequence generated by the algorithm then*

$$\begin{aligned} G(\bar{w}_T) - G(w^*) \leq \frac{C_1L^2\ln T}{\beta T} + \frac{C_2L^2}{\beta T}\sqrt{ln(1/\delta)}\sqrt{\ln T} \\ + 2\max(\frac{16L^2}{\beta}, 6B)\frac{\ln(1/\delta)}{T} \end{aligned} \tag{4}$$

with probability at least $1-4\ln(T)\delta$, where $\bar{w}_T = \frac{1}{T}\sum_{t=1}^T w_t$, and C_1, C_2 are some universal constants.

Theorem 1 essentially states that $G(\bar{w}_T) - G(w^*)$ is bounded by $O(\ln T/T)$ with high probability. Notice that while the strong convexity parameter β does not affect the order of convergence rate, the speed of convergence is proportional to the $1/\beta$. $F(w)$ with larger convexity parameter would converge faster.

3.2. Roadmap of the proof

We now outline the proof of Theorem 1. The main innovation is that instead of analyzing Reg_T and $Diff_T$ separately as in the traditional approach, which is hard to exploit locally strong convexity of $F(w)$, we analyze $Diff_T - Reg_T$ directly. Define

$$\begin{aligned} \xi_t = (F(w_t)+r(w_t)-F(w^*)-r(w^*)) \\ -(f(w_t,z_t)+r(w_t)-f(w^*,z_t)-r(w^*)). \end{aligned} \tag{5}$$

Notice we have the following relation

$$\begin{aligned} \frac{1}{2}Diff_T &\leq Diff_T - \sum_{t=1}^T (\frac{\beta}{4}\|w^*-w_t\|_2^2) \\ &= Reg_T - \sum_{t=1}^T (\frac{\beta}{4}\|w^*-w_t\|_2^2) + \sum_{t=1}^T \xi_t, \end{aligned}$$

where the first inequality holds using the fact that $G(w_t) - G(w^*) \geq \frac{\beta}{2}\|w_t-w^*\|_2^2$.

We then establish the following lemma, which is similar to Lemma 1 of Duchi et al. (2010). The latter concerns bound on Reg_T and hence requires strong convexity of $f(\cdot, z)$. Instead, we bound $Reg_T - \sum_{t=1}^T (\frac{\beta}{4}\|w_t - w^*\|_2^2)$, and thus relax this strong convexity assumption.

Lemma 1. *If we set $\eta_t = \frac{2}{\beta t}$, let $\{w_t\}_{t=1}^T$ be the sequence generated by the algorithm. Assume there is a constant L such that $\|f'(w_t, z_t)\|_2 \leq L$, then for $w^* \in \Omega$,*

$$\begin{aligned} & Reg_T - \sum_{t=1}^T (\frac{\beta}{4}\|w_t - w^*\|_2^2) \\ & \leq \sum_{t=1}^T \frac{\eta_t}{2}\|f'(w_t, z_t)\|_2^2 - \frac{1}{2\eta_T}\|w_{T+1} - w^*\|_2^2 + r(w_1) \\ & \leq \frac{CL^2}{\beta}\ln(T), \end{aligned}$$

where C is some universal constant, β is the locally strongly convexity parameter of $F(\cdot)$ around w^*. Recall our assumption $F(w) - F(w^*) \geq \frac{\beta}{2}\|w - w^*\|_2^2 + \langle \nabla F(w^*), w - w^* \rangle$.

Lemma 1 asserts that $Reg_T - \sum_{t=1}^T (\frac{\beta}{4}\|w_t - w^*\|_2^2)$ is bounded by $O(\ln T)$, even when $f(w)$ is weakly convex.

Proof. The first step is same with the proof of Lemma 1 in (Duchi et al., 2010), we present here for completeness.

$$\begin{aligned} & \eta_t(f(w_t, z_t) + r(w_{t+1})) - \eta_t(f(w^*, z_t) + r(w^*)) \\ & \leq \eta_t\langle f'(w_t, z_t), w_t - w^* \rangle + \eta_t\langle w_{t+1} - w^*, r'(w_{t+1}) \rangle \\ & = \eta_t\langle w^* - w_{t+1}, w_t - w_{t+1} - \eta_t f'(w_t, z_t) - \eta_t r'(w_{t+1}) \\ & + \langle w^* - w_{t+1}, w_{t+1} - w_t \rangle + \eta_t\langle w_t - w_{t+1}, f'(w_t, z_t) \rangle \\ & \leq \frac{1}{2}\|w^* - w_t\|_2^2 - \frac{1}{2}\|w^* - w_{t+1}\|_2^2 + \frac{\eta_t^2}{2}\|f'(w_t, z_t)\|_2^2, \end{aligned} \tag{6}$$

where the second inequality holds from the optimality of w_{t+1}, i.e., $\langle w_{t+1} - w_t + \eta_t f'(w_t, z_t) + \eta_t r'(w_{t+1}), w - w_{t+1} \rangle \geq 0$ for all w, and the fact that $\eta_t\langle w_t - w_{t+1}, f'(w_t, z_t) \rangle \leq \frac{\eta_t^2}{2}\|f'(w_t, z_t)\|_2^2 + \frac{1}{2}\|w_t - w_{t+1}\|_2^2$.

Now we divide η_t at both sides of (6) , we have

$$\begin{aligned} & (f(w_t, z_t) + r(w_{t+1})) - (f(w^*, z_t) + r(w^*)) \\ & \leq \frac{1}{2\eta_t}\|w^* - w_t\|_2^2 - \frac{1}{2\eta_t}\|w^* - w_{t+1}\|_2^2 + \frac{\eta_t}{2}\|f'(w_t, z_t)\|_2^2. \end{aligned}$$

Subtract $\frac{\beta}{4}\|w_t - w^*\|_2^2$ at both sides and sum over both side, we get

$$\begin{aligned} & \sum_{t=1}^T [f(w_t, z_t) + r(w_{t+1}) - f(w^*, z_t) - r(w^*) \\ & - \frac{\beta}{4}\|w_t - w^*\|_2^2] \\ & \leq \sum_{t=1}^T \frac{\eta_t}{2}\|f'(w_t, z_t)\|_2^2 + \frac{1}{2\eta_1}\|w^* - w_1\|_2^2 \\ & - \frac{1}{2\eta_T}\|w_{T+1} - w^*\|_2^2 \\ & + \sum_{t=1}^{T-1} [\|w_{t+1} - w^*\|_2^2 (\frac{1}{2\eta_{t+1}} - \frac{1}{2\eta_t}) - \frac{\beta}{4}\|w_{t+1} - w^*\|_2^2] \\ & - \frac{\beta}{4}\|w_1 - w^*\|_2^2. \end{aligned}$$

Choose $\eta_t = \frac{2}{\beta t}$ and use the assumption $\|f'(w_t, z_t)\|_2 \leq L$, we get

$$\begin{aligned} & \sum_{t=1}^T [f(w_t, z_t) + r(w_{t+1}) - f(w^*, z_t) - r(w^*) \\ & - \frac{\beta}{4}\|w_t - w^*\|_2^2] \\ & \leq \sum_{t=1}^T \frac{\eta_t}{2}\|f'(w_t, z_t)\|_2^2 + C_2 \\ & \leq C_1 \frac{L^2}{\beta}\ln T + C_2 \end{aligned}$$

□

Now we have

$$\frac{1}{2}Diff_T \leq C\ln T + \sum_{t=1}^T \xi_t.$$

What remains is to convert this to the rate of convergence. For that, we show $\{\xi_t\}$ is a martingale sequence.

Lemma 2. *let $\{w_t\}_{t=1}^T$ be the sequence generated by the algorithm and $F(w)$ is locally strongly convex. Assume there is a constant L such that $\|f'(w_t, z_t)\|_2 \leq L$. ξ_t is a martingale difference, i.e. $E_{t-1}\xi_t = 0$, where E_{t-1} means the conditional expectation given $z_1, z_2, ..., z_{t-1}$. Furthermore define the conditional variance $\mathrm{Var}_{t-1}\xi_t := E_{t-1}\xi_t^2$ and we have $\mathrm{Var}_{t-1}\xi_t \leq \frac{2L^2}{\beta}(G(w_t) - G(w^*))$.*

Proof. $\xi_t = F(w_t) - F(w^*) - (f(w_t, z_t) - f(w^*, z_t))$. Notice w_t just depends on $z_1, z_2, ..., z_{t-1}$, so we have

$$\begin{aligned} E_{t-1}\xi_t &= F(w_t) - F(w^*) - E_{t-1}(f(w_t, z_t) - f(w^*, z_t)) \\ &= F(w_t) - F(w^*) - (F(w_t) - F(w^*)) = 0, \end{aligned} \tag{7}$$

which shows it is a martingale difference. To bound the variance,

$$\begin{aligned}
&E_{t-1}\xi_t^2 \\
&= (F(w_t) - F(w^*))^2 + E_{t-1}(f(w_t, z_t) - f(w^*, z_t))^2 \\
&- 2(F(w_t) - F(w^*))^2 \\
&= E_{t-1}(f(w_t, z_t) - f(w^*, z_t))^2 - (F(w_t) - F(w^*))^2 \\
&\leq E_{t-1}(f(w_t, z_t) - f(w^*, z_t))^2 \\
&\leq E_{t-1}(L^2\|w_t - w^*\|_2^2) \\
&\leq \frac{2L^2}{\beta}(G(w_t) - G(w^*)),
\end{aligned}$$

where the last step holds from the locally strongly convexity of G. G is locally strongly convex from the locally strong convexity of F and definition. □

Lemma 2 is similar in spirit to Lemma 1 of (Kakade & Tewari, 2009), except that we only use the locally strong convexity of $F(w)$. Then we use the martingale inequality developed in (Kakade & Tewari, 2009) which connects the regret to generalization to prove our main theorem.

Proof of Theorem 1.

$$\begin{aligned}
Diff_T - Reg_T &= \sum_{t=1}^{T}(G(w_t) - G(w^*) - \frac{\beta}{4}\|w_t - w^*\|_2^2) \\
&- \sum_{t=1}^{T}(f(w_t, z_t) + r(w_t) - f(w^*, z_t) - r(w^*) \\
&- \frac{\beta}{4}\|w_t - w^*\|_2^2).
\end{aligned} \tag{8}$$

Define $\tilde{Diff}_T = \sum_{t=1}^{T}(G(w_t) - G(w^*) - \frac{\beta}{4}\|w_t - w^*\|_2^2)$ and

$$\begin{aligned}
\tilde{Reg}_T &= \sum_{t=1}^{T}(f(w_t, z_t) + r(w_t) - f(w^*, z_t) - r(w^*) \\
&- \frac{\beta}{4}\|w_t - w^*\|_2^2).
\end{aligned}$$

Using Lemma 1, we have $\tilde{Reg}_T \leq \frac{CL^2}{\beta}\ln T$. Notice we have $\tilde{Diff}_T \geq \frac{1}{2}Diff_T$ using the the fact that $G(w_t) - G(w^*) \geq \frac{\beta}{2}\|w_t - w^*\|_2^2$, thus

$$\frac{1}{2}Diff_T - \tilde{Reg} \leq Diff_T - Reg \leq \sum_{t=1}^{T}\xi_t,$$

which implies $\frac{1}{2}Diff_T - \frac{CL^2\ln(T)}{\beta} \leq \sum_{t=1}^{T}\xi_t$.

Using Lemma 2 and Lemma A.1 in the supplementary material, we have a upper bound of $\sum_{t=1}^{T}\xi_t$. Particularly, the following relation holds with probability at least $1-4\delta\ln T$

$$\sum_{t=1}^{T}\xi_t \leq \max\{2\sqrt{\sum_{t=1}^{T}Var_{t-1}\xi_t}, 6B\sqrt{\ln 1/\delta}\}\sqrt{\ln 1/\delta}.$$

Now we relate RHS to $Diff_T$ using Lemma 2.

$$\sum_{t=1}^{T}\xi_t \leq \max(2\sqrt{\frac{2L^2}{\beta}Diff_T}, 6B\sqrt{\ln(1/\delta)})\sqrt{\ln(1/\delta)}$$

with probability at least $1 - 4\ln(T)\delta$.
Now we have a inequality of $Diff_T$,

$$\begin{aligned}
&\frac{1}{2}Diff_T - \frac{CL^2\ln(T)}{\beta} \\
&\leq \max(2\sqrt{\frac{2L^2}{\beta}Diff_T}, 6B\sqrt{\ln(1/\delta)})\sqrt{\ln(1/\delta)}.
\end{aligned} \tag{9}$$

Solving the inequality using Lemma A.2 in the supplementary material , we get

$$\begin{aligned}
\frac{1}{2}Diff_T &\leq \frac{C_1L^2\ln T}{\beta} + C_2\frac{4L^2}{\beta}\sqrt{ln(1/\delta)}\sqrt{\ln T} \\
&+ \max(\frac{16L^2}{\beta}, 6B)\ln(1/\delta).
\end{aligned} \tag{10}$$

Using the convexity of $G(w)$, we have with high probability

$$G(\bar{w}_T) - G(w^*) \leq Diff_T/T \leq O(\ln T/T).$$

□

This proof idea can be extended to other algorithms. Indeed, our proofs of RDA and OPG-ADMM follow a similar roadmap except for some technical details.

4. Regularized Dual Averaging method

We briefly explain the RDA method. In RDA, an auxiliary function $h(w)$ is needed, which is a strongly convex function with convex parameter 1 (e.g., $\frac{1}{2}\|w\|_2^2$) on domain of $r(w)$, and also satisfies $\arg\min h(w) = \arg\min r(w)$. Define w_0 as $w_0 = \arg\min h(w)$. The Regularized Dual Averaging method iteratively generates a sequence $\{w_\tau\}_{\tau=1}^t$ in the following way,

$$w_{t+1} := \arg\min_w(\langle\sum_{\tau=1}^{t}f'(w_\tau, z_\tau), w\rangle + tr(w) + \beta_t h(w)),$$

with initialization $w_1 = w_0$. Notice we only need to store the summation of subgradients rather than each individual one. It achieves $O(1/\sqrt{T})$ convergence rate when f

is (weakly) convex and $O(\ln T/T)$ in the strongly convex case. For more details on choices of $h(w)$, examples of applications of RDA and proofs, we refer the readers to Xiao (2009).

To achieve the $O(\ln T/T)$ convergence rate under our assumption, we propose a modified RDA method slightly different from the original one. It updates w_t in the following way.

$$w_{t+1} = \arg\min_w (\langle \sum_{\tau=1}^{t} f'(w_\tau, z_\tau), w\rangle + tr(w) + \beta_t h(w) + \gamma \sum_{\tau=1}^{t} \|w_\tau - w\|_2^2). \tag{11}$$

Notice we add an additional strongly convex term. It is easy to see that solving w_{t+1} only requires knowing the *sum* of w_t rather than each individual one, so the memory consumption is almost same as the original algorithm. This additional term help us to bound $Reg_t - \gamma \sum_{\tau=1}^{t} \|w_\tau - w\|_2^2$ in the following lemma. Similarly to the proof of OPG, we first define Reg_t.

$$Reg_t = \sum_{\tau=1}^{t} (f(w_\tau, z_\tau) + r(w_\tau)) - \sum_{\tau=1}^{t} (f(w, z_\tau) + r(w)).$$

The following lemma upper bounds the regret subtracted by a strongly convex term.

Lemma 3. *If we set $\beta_t = \gamma(1 + \ln t)$ for $t \geq 1$, $\beta_0 = \beta_1 = \gamma$, let $\{w_\tau\}_{\tau=1}^t$ be the sequence generated by the algorithm (11). Assume there is a constant L such that $\|f'(w_\tau, z_\tau)\|_2 \leq L$, then we have*

$$Reg_t - \gamma \sum_{\tau=1}^{t} \|w_\tau - w\|_2^2 \leq C_1 \gamma D^2 + C_2 \frac{L^2}{\gamma}(1 + \ln(t)),$$

for any $w \in \mathcal{F}_D$, where $\mathcal{F}_D = \{w | h(w) \leq D^2\}$, C_1, C_2 are some universal constants.

This lemma states $Reg_t - \gamma \sum_{\tau=1}^{t} \|w_\tau - w\|_2^2$ is upper bounded by $O(\ln t)$. Notice $\gamma \sum_{\tau=1}^{t} \|w_\tau - w\|_2^2$ is indeed the new term introduced in (11).

Following a same roadmap of derivation as that of OPG, we can apply the martingale inequality on the regret bound given in the Lemma 3. This leads to the following theorem that establishes an improved convergence rate of RDA to solve the learning problem 1. The detailed proof is deferred to the appendix.

Theorem 2. *If we set $\beta_t = \gamma(1 + \ln t)$, $\gamma = \frac{\beta}{4}$. Let $\{w_\tau\}_{\tau=1}^t$ be the sequence generated by the algorithm. Assume there is a constant L such that $\|f'(w_\tau, z_\tau)\|_2 \leq L$, If $f(w, z)$ is bounded by B, $F(w)$ is a locally strongly convex function with parameter β around w^*, then*

$$G(\bar{w}_T) - G(w^*) \leq \frac{C_1 + C_2 \ln T}{T} + 4\sqrt{\ln(1/\delta)} \frac{\sqrt{C_1 + C_2 \ln T}}{T} + 2\max(\frac{16L^2}{\beta}, 6B) \frac{\ln(1/\delta)}{T}, \tag{12}$$

with probability at least $1 - 4\ln(T)\delta$, where $\bar{w}_T = \frac{1}{T}\sum_{t=1}^{T} w_t$, C_1, C_2 are some constants depending on D, L and β.

5. OPG-ADMM

ADMM (Gabay & Mercier, 1976) is a framework for optimization of a composite function and has wide range of applications (Candès et al., 2011; Aguiar et al., 2011; Shen et al., 2012). It has gained lots of attentions in the machine learning society recently (Boyd et al., 2011; Ouyang et al., 2013; Suzuki, 2013). ADMM considers the following optimization problem:

$$\min_{x \in \mathcal{X}, y \in \mathcal{Y}} \quad \frac{1}{T}\sum_{t=1}^{T} f(x, z_t) + \psi(y) \quad s.t. \quad Ax + By - b = 0, \tag{13}$$

where $\mathcal{X}$ and $\mathcal{Y}$ are some convex compact sets with radius D, i.e., $\|x - x'\|_2 \leq D$ for all $x, x' \in \mathcal{X}$ and similarly $\|y - y'\|_2 \leq D$ for all $y, y' \in \mathcal{Y}$. B is a squared matrix and B^{-1} exists. One example is $B = I$ and $b = 0$, i.e., $Ax = -y$. ADMM splits the optimization with respect to x and y using the augmented Lagrangian technique. It has $O(1/T)$ convergence rate in the weakly convex case and linear convergence rate in the strongly convex case (He & Yuan, 2012; Deng & Yin, 2012).

However, ADMM is basically a batch method which needs to store the whole data in memory. To resolve this issue, several online variants of ADMM are proposed (Ouyang et al., 2013; Suzuki, 2013; Wang & Banerjee, 2012). The algorithm of OPG-ADMM (Suzuki, 2013) is given by the following update rules.

$$\begin{aligned} x_{t+1} &= \arg\min_{x \in \mathcal{X}} g_t^T x - \lambda_t^T (Ax + By_t - b) \\ &\quad + \frac{\rho}{2}\|Ax + By_t - b\|_2^2 + \frac{1}{2\eta_t}\|x - x_t\|_{G_t}^2, \\ y_{t+1} &= \arg\min_{y \in \mathcal{Y}} \psi(y) - \lambda_t^T (Ax_{t+1} + By - b) \\ &\quad + \frac{\rho}{2}\|Ax_{t+1} + By - b\|_2^2, \\ \lambda_{t+1} &= \lambda_t - \rho(Ax_{t+1} + By_{t+1} - b), \end{aligned} \tag{14}$$

where g_t stands for $f'(w_t, z_t)$, $G_t = \gamma I - \eta_t \rho A^T A$, and $\|x\|_{G_t}^2$ denotes $x^T G_t x$. γ, ρ, η_t are chosen such that $G_t \succeq$

I for simplicity. We initialize $x_1 = 0, \lambda_1 = 0$ and $By_1 = b$. Moreover we define $\tilde{\lambda}_t = \lambda_t - \rho(Ax_{t+1} + By_t - b)$. We assume that it is easy to compute the proximal operation corresponding to ψ, i.e., the update rule for y_t is computationally easy. It is known to achieve a convergence rate $O(1/\sqrt{T})$ in the weakly convex case and $O(\ln T/T)$ in the strongly convex case.

To establish a fast rate for OPG-ADMM, similarly to OPG, we need to establish an upper bound of the $Reg_T - \sum_{t=1}^T \frac{\beta}{4}\|x_t - \hat{x}\|_2^2$.

Lemma 4. *Let $\{x_t\}_{t=1}^T$, $\{y_t\}_{t=1}^T$ and $\{\lambda_t\}_{t=1}^T$ be the sequence generated by algorithm (14). Assume f is weakly convex, then for all $\hat{x} \in \mathcal{X}$, $\hat{y} \in \mathcal{Y}$ and $\hat{\lambda}$, we have*

$$\begin{aligned}
&\sum_{t=1}^T (f(x_t, z_t) + \psi(y_t)) - \sum_{t=1}^T (f(\hat{x}, z_t) + \psi(\hat{y})) \\
&+ \sum_{t=1}^T \begin{pmatrix} -A^T\tilde{\lambda}_t \\ -B^T\tilde{\lambda}_t \\ Ax_t + By_t - b \end{pmatrix}^T \begin{pmatrix} x_t - \hat{x} \\ y_t - \hat{y} \\ \tilde{\lambda}_t - \hat{\lambda} \end{pmatrix} \\
&+ \sum_{t=1}^T \frac{\|\lambda_t - \lambda_{t+1}\|_2^2}{2\rho} + \frac{\|\lambda_{T+1} - \hat{\lambda}\|_2^2}{2\rho} - \sum_{t=1}^T \frac{\beta}{4}\|x_t - \hat{x}\|_2^2 \\
&\le \frac{\|\hat{x}\|_{G_1}^2}{2\eta_1} + \sum_{t=2}^T (\frac{\gamma}{2\eta_t} - \frac{\gamma}{2\eta_{t-1}})\|x_t - \hat{x}\|_2^2 + \sum_{t=1}^T \frac{\eta_t}{2}\|g_t\|_{G_t^{-1}}^2 \\
&+ \frac{\rho}{2}\|b - B\hat{y}\|_2^2 + \frac{\|\hat{\lambda}\|_2^2}{2\rho} + \langle Ax_{T+1}, \hat{\lambda}\rangle \\
&+ \langle B(\hat{y} - y_{T+1}), \lambda_{T+1} - \hat{\lambda}\rangle - \langle B\hat{y} - b, \hat{\lambda}\rangle \\
&- \sum_{t=1}^T \frac{\beta}{4}\|x_t - \hat{x}\|_2^2,
\end{aligned} \tag{15}$$

where g_t denotes $f'(x_t, z_t)$ for short. Here we assume $\|g_t\|_2$ is bounded by L, $f(x, w)$ is bounded by B, the sub gradient of ψ is bounded by L_ψ, and X and Y are two convex compact set with radius D.

Notice that in the RHS of (15), the term $\langle B(\hat{y} - y_{T+1}), \lambda_{T+1} - \hat{\lambda}\rangle \le \langle \hat{y} - y_{T+1}, \nabla\psi(y_{T+1}) - B^T\hat{\lambda}\rangle$ by the optimality of y_{t+1}. The RHS can also be bounded by a constant since y_{T+1} and $\nabla\psi(y_{T+1})$ are bounded by constants using assumptions. Furthermore, We can choose a special $\hat{\lambda} = 0$ to make the third term in the LHS in (15) vanishing. Thus, we can show the RHS is bounded by $C_1 \ln T + C_2$ if we choose $\eta_t = \frac{2\gamma}{\beta t}$. We defer the details to the appendix.

We remark that (x_t, y_t) generated by the algorithm may not satisfy the constraint $Ax_t + By_t - b = 0$. Same as Suzuki (2013), we use $y_t' := B^{-1}(b - Ax_t)$ to replace y_t.

Similarly as the previous two methods, based on the regret bound we establish the improved stochastic convergence rate of OPG-ADMM.

Theorem 3. *Set $\eta_t = \frac{2\gamma}{\beta t}$. Let $\{x_t\}_{t=1}^T$, $\{y_t\}_{t=1}^T$ and $\{\lambda_t\}_{t=1}^T$ be the sequence generated by the algorithm. Assume X and Y are two convex compact set with radius D, $f(x, z)$ is bounded by B, $\|g_t\|_2$ is bounded by L, the sub gradient of ψ is bounded by L_ψ, and $F(x)$ is locally strongly convex function around x^* with parameter β, then with high probability*

$$G(\bar{x}_T, \bar{y}_T') - G(x^*, y^*) \le C\frac{\ln T}{T}$$

where $\bar{x}_T = \frac{1}{T}\sum_{t=1}^T x_t$, $\bar{y}_T' = \frac{1}{T}\sum_{t=1}^T y_t'$, C is some constant depending on β, B, L, L_ψ. $G(x, y) = F(x) + \psi(y), F(x) = E_z f(x, z)$ and (x^*, y^*) is the optimal solution of $G(x, y)$ with constraint $Ax + By - b = 0$.

6. Simulation results

In this section, we perform numerical simulations to illustrate and validate our results. We emphasize that this paper concerns providing tighter theoretical analysis of three well-known algorithms, rather than proposing new algorithms. As such, the goal of the simulation is not about showing the superiority of the algorithms (which has been validated in many real applications (Ouyang et al., 2013; Suzuki, 2013; Xiao, 2009), but to show that their empirical performance matches our theorem. In light of this, synthesized data is more preferable in our experiments, as we are able to compute the global optimum (and hence the optimality gap) based on the underlying mechanism that generates the data, which is typically impossible for real data sets.

We use LASSO as the problem to optimize, i.e. $f(w, (x, y)) + r(w) = \frac{1}{2}\|y - w^T x\|_2^2 + \lambda\|w\|_1$, as this is a widely used formulation in various areas such as sparse coding, compressive sensing, and high dimensional statistics. We generate the input and output pair (x, y) in the following way: $w^\dagger$ is a k sparse vector of dimension d where the no-zeros entries are generated from the standard normal distribution, x is a d dimension vector drawn from the standard normal distribution, and $y = (w^\dagger)^T x + \xi$, where ξ is a Gaussian noise term with variance $\sigma^2 = 0.25$. Observe this satisfies our setting where $f(w, (x, y)) + r(w)$ is a weakly convex function, but $E(f(w, (x, y)))$ is strongly convex. In Lasso, given all conditions above, we can calculate $G(w)$ analytically as follows. $G(w) = \frac{1}{2}\|w - w^\dagger\|_2^2 + \lambda\|w\|_1 + \frac{\sigma^2}{2}$. We set $\lambda = 0.1$. The optimal solution of $G(w)$ can be calculated by the standard soft thresholding operation of $w^\dagger$. The simulation results are reported in Figure 1 and 2. The Y axis is the optimality gap $G(\bar{w_T}) - G(w^*)$ and the X axis is the number of steps. All results are averaged over 10 trials and drawn in the log-log scale.

We observe the following: all three algorithms converge slowly at beginning. After certain time steps T_0, they de-

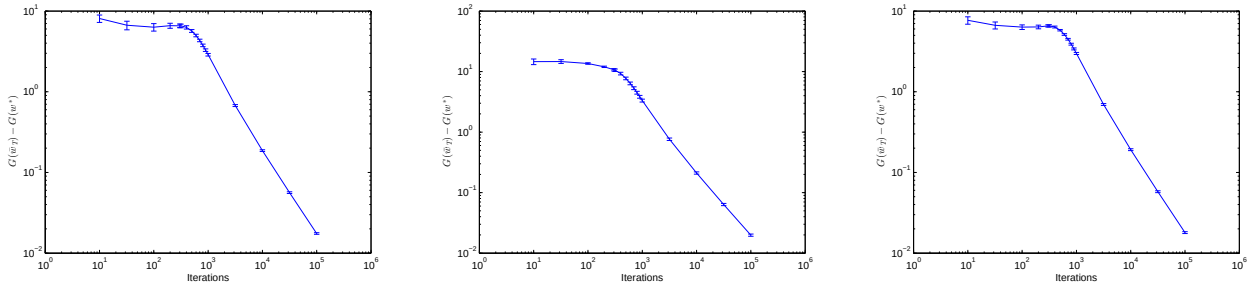

Figure 1. Stochastic convergence rates of OPG, RDA and OPG-ADMM in LASSO with dimension d=300 and sparsity k=10. The experimental results are averaged with 10 trials.

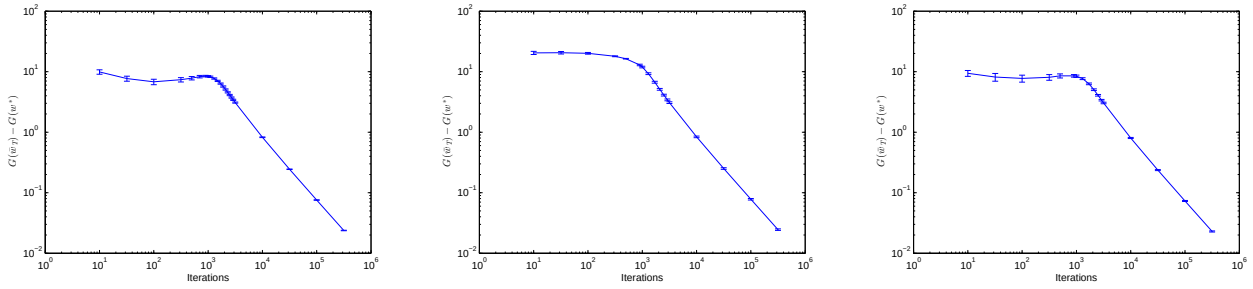

Figure 2. Stochastic convergence rates of OPG, RDA and OPG-ADMM in LASSO with dimension d=1000 and sparsity k=10. The experimental results are averaged over 10 trials.

crease with ratio -1 on log-log scale, i.e., the convergence rate is proportional to $1/T$ for large T, which validates our theoretical results. In terms of the convergence speed at the initial stages of the algorithms, we suspect that this may be due to fact that we measure the population error $G(\bar{w}_T) - G(w^*)$ rather than the empirical error. Thus, when relatively few samples are given, overfitting may happen. Interestingly, we find that the value T_0 is closely related the dimension of the problem. In the case d=300, T_0 is around 300, whereas T_0 is around 1000 when $d = 1000$, which seems to support our conjecture.

7. Conclusion

In this paper, we analyzed three widely used stochastic optimization algorithms, namely OPG, RDA, and OPG-ADMM, and established an $O(\ln T/T)$ upper bound of their convergence speed without the strong convexity assumption on the loss function. Instead, we only require the expectation of the loss function to be locally strongly convex, a much weaker assumption that is easily satisfied in many cases. This closed a gap between known theoretic results and empirical performance of these algorithms on widely used formulations such as Lasso and logistic regression. The key novelty of our analysis is that we analyze the $Diff_T - Reg_T$ directly instead of themselves separately, which makes it possible to utilize the strong convexity of the expectation of the loss function. We believe this is a technique that can be easily adapted to studying other algorithms, and hence provides a general recipe to obtain improved convergence rate for stochastic optimization algorithms.

Acknowledgements

The research is partially supported by Agency for Science, Technology and Research (A*STAR) of Singapore through SERC PSF Grant R266000101305.

References

Aguiar, Pedro, Xing, Eric P, Figueiredo, Mário, Smith, Noah A, and Martins, André. An augmented lagrangian approach to constrained map inference. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 169–176, 2011.

Bach, Francis. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *The Journal of Machine Learning Research*, 15(1):595–627, 2014.

Bach, Francis and Moulines, Eric. Non-strongly-convex smooth stochastic approximation with convergence rate

o (1/n). In *Advances in Neural Information Processing Systems*, pp. 773–781, 2013.

Boyd, Stephen, Parikh, Neal, Chu, Eric, Peleato, Borja, and Eckstein, Jonathan. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

Candès, Emmanuel J, Li, Xiaodong, Ma, Yi, and Wright, John. Robust principal component analysis? *Journal of the ACM*, 58(3):11, 2011.

Cesa-Bianchi, Nicolo, Conconi, Alex, and Gentile, Claudio. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50 (9):2050–2057, 2004.

Combettes, Patrick L and Wajs, Valérie R. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling & Simulation*, 4(4):1168–1200, 2005.

Deng, Wei and Yin, Wotao. On the global and linear convergence of the generalized alternating direction method of multipliers. Technical report, Rice University CAAM TR12-14, 2012.

Duchi, John, Shalev-Shwartz, Shai, Singer, Yoram, and Tewari, Ambuj. Composite objective mirror descent. In *Proceedings of the Annual Conference on Computational Learning Theory*, 2010.

Gabay, Daniel and Mercier, Bertrand. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications*, 2(1):17–40, 1976.

He, Bingsheng and Yuan, Xiaoming. On the o(1/n) convergence rate of the douglas-rachford alternating direction method. *SIAM Journal on Numerical Analysis*, 50(2): 700–709, 2012.

Jacob, Laurent, Obozinski, Guillaume, and Vert, Jean-Philippe. Group lasso with overlap and graph lasso. In *Proceedings of 26th International Conference on Machine Learning*, pp. 433–440, 2009.

Kakade, Sham M and Tewari, Ambuj. On the generalization ability of online strongly convex programming algorithms. In *Advances in Neural Information Processing Systems*, pp. 801–808, 2009.

Kushner, Harold J and Yin, George. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

Nesterov, Yurii. Primal-dual subgradient methods for convex problems. *Mathematical programming*, 120(1):221–259, 2009.

Ouyang, Hua, He, Niao, Tran, Long, and Gray, Alexander. Stochastic alternating direction method of multipliers. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 80–88, 2013.

Rakhlin, Alexander, Shamir, Ohad, and Sridharan, Karthik. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Conference on Machine Learning*, pp. 449–456, 2012.

Shalev-Shwartz, Shai. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011.

Shen, Chao, Chang, Tsung-Hui, Wang, Kun-Yu, Qiu, Zhengding, and Chi, Chong-Yung. Distributed robust multicell coordinated beamforming with imperfect csi: an admm approach. *IEEE Transactions on Signal Processing*, 60(6):2988–3003, 2012.

Signoretto, Marco, De Lathauwer, Lieven, and Suykens, Johan AK. Nuclear norms for tensors and their use for convex multilinear estimation. *Submitted to Linear Algebra and Its Applications*, 43, 2010.

Singer, Yoram and Duchi, John C. Efficient learning using forward-backward splitting. In *Advances in Neural Information Processing Systems*, pp. 495–503, 2009.

Sridharan, Karthik, Shalev-Shwartz, Shai, and Srebro, Nathan. Fast rates for regularized objectives. In *Advances in Neural Information Processing Systems*, pp. 1545–1552, 2009.

Suzuki, Taiji. Dual averaging and proximal gradient descent for online alternating direction multiplier method. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 392–400, 2013.

Wang, Huahua and Banerjee, Arindam. Online alternating direction method. In *Proceedings of the 29th International Conference on Machine Learning*, pp. 1119–1126, 2012.

Xiao, Lin. Dual averaging method for regularized stochastic learning and online optimization. In *Advances in Neural Information Processing Systems*, pp. 2116–2124, 2009.

Zhang, Tong. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the 21st International Conference on Machine learning*, 2004.

Zhong, Leon Wenliang and Kwok, James T. Fast stochastic alternating direction method of multipliers. In *Proceedings of the 30th International Conference on Machine Learning*, 2013.