Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)

Fast-and-Light Stochastic ADMM

Shuai Zheng **James T. Kwok**
Department of Computer Science and Engineering
Hong Kong University of Science and Technology
Hong Kong
{szhengac, jamesk}@cse.ust.hk

Abstract

The alternating direction method of multipliers (ADMM) is a powerful optimization solver in machine learning. Recently, stochastic ADMM has been integrated with variance reduction methods for stochastic gradient, leading to SAG-ADMM and SDCA-ADMM that have fast convergence rates and low iteration complexities. However, their space requirements can still be high. In this paper, we propose an integration of ADMM with the method of stochastic variance reduced gradient (SVRG). Unlike another recent integration attempt called SCAS-ADMM, the proposed algorithm retains the fast convergence benefits of SAG-ADMM and SDCA-ADMM, but is more advantageous in that its storage requirement is very low, even independent of the sample size n. Experimental results demonstrate that it is as fast as SAG-ADMM and SDCA-ADMM, much faster than SCAS-ADMM, and can be used on much bigger data sets.

1 Introduction

In this big data era, tons of information are generated every day. Thus, efficient optimization tools are needed to solve the resultant large-scale machine learning problems. In particular, the well-known stochastic gradient descent (SGD) [Bottou, 2004] and its variants [Parikh and Boyd, 2014] have drawn a lot of interest. Instead of visiting all the training samples in each iteration, the gradient is computed by using one sample or a small mini-batch of samples. The per-iteration complexity is then reduced from $O(n)$, where n is the number of training samples, to $O(1)$. Despite its scalability, the stochastic gradient is much noisier than the batch gradient. Thus, the stepsize has to be decreased gradually as stochastic learning proceeds, leading to slower convergence.

Recently, a number of fast algorithms have been developed that try to reduce the variance of stochastic gradients [Defazio *et al.*, 2014; Johnson and Zhang, 2013; Roux *et al.*, 2012; Shalev-Shwartz and Zhang, 2013]. With the variance reduced, a larger constant stepsize can be used. Consequently, much faster convergence, even matching that of its batch counterpart, is attained. A prominent example is the stochastic average gradient (SAG) [Roux *et al.*, 2012], which reuses the old stochastic gradients computed in previous iterations. A related method is stochastic dual coordinate ascent (SDCA) [Shalev-Shwartz and Zhang, 2013], which performs stochastic coordinate ascent on the dual. However, a caveat of SAG is that storing the old gradients takes $O(nd)$ space, where d is the dimensionality of the model parameter. Similarly, SDCA requires storage of the dual variables, which scales as $O(n)$. Thus, they can be expensive in applications with large n (big sample size) and/or large d (high dimensionality).

Moreover, many machine learning problems, such as graph-guided fused lasso and overlapping group lasso, are too complicated for SGD-based methods. The alternating direction method of multipliers (ADMM) has been recently advocated as an efficient optimization tool for a wider variety of models [Boyd *et al.*, 2011]. Stochastic ADMM extensions have also been proposed [Ouyang *et al.*, 2013; Suzuki, 2013; Wang and Banerjee, 2012], though they only have suboptimal convergence rates. Recently, researchers have borrowed variance reduction techniques into ADMM. The resultant algorithms, SAG-ADMM [Zhong and Kwok, 2014] and SDCA-ADMM [Suzuki, 2014], have fast convergence rate as batch ADMM but are much more scalable. The downside is that they also inherit the drawbacks of SAG and SDCA. In particular, SAG-ADMM and SDCA-ADMM require $O(nd)$ and $O(n)$ space, respectively, to store the past gradients and weights or dual variables. This can be problematic in large multitask learning, where the space complexities is scaled by N, the number of tasks. For example, in one of our multitask learning experiments, SAG-ADMM needs 38.2TB for storing the weights, and SDCA-ADMM needs 9.6GB for the dual variables.

To alleviate this problem, one can integrate ADMM with another popular variance reduction method, namely, stochastic variance reduced gradient (SVRG) [Johnson and Zhang, 2013]. In particular, SVRG is advantageous in that no extra space for the intermediate gradients or dual variables is needed. However, this integration is not straightforward. A recent initial attempt is made in [Zhao *et al.*, 2015]. Essentially, their SCAS-ADMM algorithm uses SVRG as an inexact stochastic solver for one of the ADMM subproblems. The other ADMM variables are not updated until that subproblem has been approximately solved. Analogous to the difference between Jacobi iteration and Gauss-Seidel iteration, this slows down convergence. Indeed, on strongly con-

vex problems, SCAS-ADMM only has sublinear convergence while SDCA-ADMM has a linear rate. On general convex problems, SCAS-ADMM requires the stepsize to be gradually reduced. This defeats the original purpose of using SVRG-based algorithms, which aim at using a larger, constant learning rate to achieve fast convergence [Johnson and Zhang, 2013].

In this paper, we propose a tighter integration of SVRG and ADMM with a constant learning rate. The per-iteration computational cost of the resultant SVRG-ADMM algorithm is as low as existing stochastic ADMM methods, but yet it admits fast linear convergence on strongly convex problems. Among existing stochastic ADMM algorithms, a similar linear convergence result is only proved in SDCA-ADMM for a special ADMM setting. Besides, it is well-known that the penalty parameter in ADMM can significantly affect convergence [Nishihara *et al.*, 2015]. While its effect on the batch ADMM has been well-studied [Deng and Yin, 2015; Nishihara *et al.*, 2015], that on stochastic ADMM is still unclear. We show that its optimal setting is, interestingly, the same as that in the batch setting. Moreover, SVRG-ADMM does not need to store the gradients or dual variables throughout the iterations. This makes it particularly appealing when both the number of samples and label classes are large.

Notation: For a vector x, $\|x\|$ is its ℓ_2-norm, and $\|x\|_Q = \sqrt{x^T Q x}$. For a matrix X, $\|X\|$ is its spectral norm, $\sigma_{\max}(X)$ (resp. $\sigma_{\min}(X)$) is its largest (resp. smallest) eigenvalue, and $X^\dagger$ its pseudoinverse. For a function f, f' is a subgradient. When f is differentiable, we use ∇f as its gradient.

2 Related Work

Consider the regularized risk minimization problem: $\min_x \frac{1}{n}\sum_{i=1}^n f_i(x) + r(x)$, where x is the model parameter, n is the number of training samples, f_i is the loss due to sample i, and r is a regularizer. For many structured sparsity regularizers, $r(x)$ is of the form $g(Ax)$, where A is a matrix [Kim *et al.*, 2009; Jacob *et al.*, 2009]. By introducing an additional y, the problem can be rewritten as

$$\min_{x,y} \; f(x) + g(y) \;:\; Ax - y = 0, \tag{1}$$

where

$$f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x). \tag{2}$$

Problem (3) can be conveniently solved by the alternating direction method of multipliers (ADMM) [Boyd *et al.*, 2011]. In general, ADMM considers problems of the form

$$\min_{x,y} \; f(x) + g(y) \;:\; Ax + By = c, \tag{3}$$

where f, g are convex functions, and A, B (resp. c) are constant matrices (resp. vector). Let $\rho > 0$ be a penalty parameter, and u be the dual variable. At iteration t, ADMM performs the updates:

$$y_t = \arg\min_y g(y) + \frac{\rho}{2}\|Ax_{t-1} + By - c + u_{t-1}\|^2, \tag{4}$$

$$x_t = \arg\min_x f(x) + \frac{\rho}{2}\|Ax + By_t - c + u_{t-1}\|^2, \tag{5}$$

$$u_t = u_{t-1} + Ax_t + By_t - c. \tag{6}$$

With f in (2), solving (5) can be computationally expensive when the data set is large. Recently, a number of stochastic and online variants of ADMM have been developed [Wang and Banerjee, 2012; Ouyang *et al.*, 2013; Suzuki, 2013]. However, they converge much slower than the batch ADMM, namely, $O(1/\sqrt{T})$ vs $O(1/T)$ for convex problems, and $O(\log T/T)$ vs linear convergence for strongly convex problems.

For gradient descent, a similar gap in convergence rates between the stochastic and batch algorithms is well-known [Roux *et al.*, 2012]. As noted by [Johnson and Zhang, 2013], the underlying reason is that SGD has to control the gradient's variance by gradually reducing its stepsize η. Recently, by observing that the training set is always finite in practice, a number of variance reduction techniques have been developed that allow the use of a constant stepsize, and consequently faster convergence. In this paper, we focus on the SVRG [Johnson and Zhang, 2013], which is advantageous in that no extra space for the intermediate gradients or dual variables is needed. The algorithm proceeds in stages. At the beginning of each stage, the gradient $\tilde{z} = \frac{1}{n}\sum_{i=1}^n \nabla f_i(\tilde{x})$ is computed using a past parameter estimate $\tilde{x}$. For each subsequent iteration t in this stage, the approximate gradient

$$\hat{\nabla} f(x_{t-1}) = \frac{1}{b}\sum_{i_t \in \mathcal{I}_t} (\nabla f_{i_t}(x_{t-1}) - \nabla f_{i_t}(\tilde{x})) + \tilde{z} \tag{7}$$

is used, where $\mathcal{I}_t$ is a mini-batch of size b from $\{1, 2, \ldots, n\}$. Note that $\hat{\nabla} f(x_{t-1})$ is unbiased (i.e., $\mathbb{E}\hat{\nabla} f(x_{t-1}) = \nabla f(x_{t-1})$), and its (expected) variance goes to zero asymptotically.

Recently, variance reduction has also been incorporated into stochastic ADMM. For example, SAG-ADMM [Zhong and Kwok, 2014] is based on SAG [Roux *et al.*, 2012]; and SDCA-ADMM [Suzuki, 2014] is based on SDCA [Shalev-Shwartz and Zhang, 2013]. Both enjoy low iteration complexities and fast convergence. However, SAG-ADMM requires $O(nd)$ space for the old gradients and weights, where d is the dimensionality of x. As for SDCA-ADMM, even though its space requirement is lower, it is still proportional to N, the number of labels in a multiclass/multilabel/multitask learning problem. As N can easily be in the thousands or even millions (e.g., Flickr has more than 20 millions tags), SAG-ADMM and SDCA-ADMM can still be problematic.

3 Integrating SVRG with Stochastic ADMM

In this paper, we make the following assumptions on the f_i's in (2) and g in (3).

Assumption 1. *Each f_i is convex, continuously differentiable, and has L_i-Lipschitz-continuous gradient. In other words, there exists $L_i > 0$ such that $f_i(x_j) \le f_i(x_i) + \nabla f_i(x_i)^T(x_j - x_i) + \frac{L_i}{2}\|x_i - x_j\|^2$ for all x_i, x_j.*

Assumption 2. *g is convex, but can be nonsmooth.*

Let (x_*, y_*) be the optimal (primal) solution of (3), and u_* the corresponding dual solution. At optimality, we have

$$\nabla f(x_*) + \rho A^T u_* = 0, \;\; g'(y_*) + \rho B^T u_* = 0, \tag{8}$$

$$Ax_* + By_* = c. \tag{9}$$

3.1 Strongly Convex Problems

In this section, we consider the case where f is strongly convex. A popular example in machine learning is the square loss.

Assumption 3. *f is strongly convex, i.e., there exists $\lambda_f > 0$ such that $f(x_i) \geq f(x_j) + \nabla f(x_j)^T(x_i - x_j) + \frac{\lambda_f}{2}\|x_i - x_j\|^2$ for all x_i, x_j.*

Moreover, we assume that matrix A has full row rank. This assumption has been commonly used in the convergence analysis of ADMM algorithms [Deng and Yin, 2015; Ghadimi *et al.*, 2015; Giselsson and Boyd, 2014; Nishihara *et al.*, 2015].

Assumption 4. *Matrix A has full row rank.*

The proposed procedure is shown in Algorithm 1. Similar to SVRG, it is divided into stages, each with m iterations. The updates for y_t and u_t are the same as batch ADMM ((4) and (6)). The key change is on the more expensive x_t update. We first replace (5) by its first-order approximation $f(x_{t-1}) + \nabla f(x_{t-1})^T x$. As in SVRG, the full gradient $\nabla f(x_{t-1})$ is approximated by $\hat{\nabla} f(x_{t-1})$ in (7). Recall that $\hat{\nabla} f(x_{t-1})$ is unbiased and its (expected) variance goes to zero. In other words, $\hat{\nabla} f(x_{t-1}) \to \nabla f(x_*)$ when x_{t-1} and $\tilde{x}$ approach the optimal x_*, which allows the use of a constant stepsize. In contrast, traditional stochastic approximations such as OPG-ADMM [Suzuki, 2013] use $\frac{1}{b}\sum_{i_t \in \mathcal{I}_t} \nabla f_{i_t}(x_{t-1})$ to approximate the full gradient, and a decreasing step size is needed to ensure convergence.

Unlike SVRG, the optimization subproblem in Step 9 has the additional terms $\frac{\rho}{2}\|Ax + By_t - c + u_{t-1}\|^2$ (from subproblem (5)) and $\frac{1}{2\eta}\|x - x_{t-1}\|_G^2$ (to ensure that the next iterate is close to the current iterate x_{t-1}). A common setting for G is simply $G = I$ [Ouyang *et al.*, 2013]. Step 9 then reduces to

$$x_t = \left(\frac{1}{\eta} I + \rho A^T A\right)^{-1} \left(\frac{x_{t-1}}{\eta} - \hat{\nabla} f(x_{t-1}) + \rho A^T (By_t - c + u_{t-1})\right). \quad (10)$$

Note that $(\frac{1}{\eta} I + \rho A^T A)^{-1}$ above can be pre-computed. On the other hand, while some stochastic ADMM algorithms [Ouyang *et al.*, 2013; Zhong and Kwok, 2014] also need to compute a similar matrix inverse, their η's change with iterations and so cannot be pre-computed.

When $A^T A$ is large, storage of this matrix may still be problematic. To alleviate this, a common approach is *linearization* (also called the inexact Uzawa method) [Zhang *et al.*, 2011]. It sets $G = \gamma I - \eta\rho A^T A$ with

$$\gamma \geq \gamma_{\min} \equiv \eta\rho\|A^T A\| + 1 \quad (11)$$

to ensure that $G \succeq I$. The x_t update in (10) then simplifies to

$$x_t = x_{t-1} - \frac{\eta}{\gamma}\left(\hat{\nabla} f(x_{t-1}) + \rho A^T (Ax_{t-1} + By_t - c + u_{t-1})\right). \quad (12)$$

Note that steps 2 and 12 in Algorithm 1 involve the pseudo-inverse $A^\dagger$. As A is often sparse, this can be efficiently computed by the Lanczos algorithm [Golub and Van Loan, 2012].

Algorithm 1 SVRG-ADMM for strongly convex problems.

1: **Input:** $m, \eta, \rho > 0$.
2: initialize $\tilde{x}_0, \tilde{y}_0$ and $\tilde{u}_0 = -\frac{1}{\rho}(A^T)^\dagger \nabla f(\tilde{x}_0)$;
3: **for** $s = 1, 2, \ldots$ **do**
4: $\quad \tilde{x} = \tilde{x}_{s-1}$;
5: $\quad x_0 = \tilde{x}_{s-1}$; $y_0 = \tilde{y}_{s-1}$; $u_0 = \tilde{u}_{s-1}$;
6: $\quad \tilde{z} = \frac{1}{n}\sum_{i=1}^n \nabla f_i(\tilde{x})$;
7: $\quad$ **for** $t = 1, 2, \ldots, m$ **do**
8: $\quad\quad y_t \leftarrow \arg\min_y g(y) + \frac{\rho}{2}\|Ax_{t-1} + By - c + u_{t-1}\|^2$;
9: $\quad\quad x_t \leftarrow \arg\min_x \hat{\nabla} f(x_{t-1})^T x + \frac{\rho}{2}\|Ax + By_t - c + u_{t-1}\|^2 + \frac{\|x - x_{t-1}\|_G^2}{2\eta}$;
10: $\quad\quad u_t \leftarrow u_{t-1} + Ax_t + By_t - c$;
11: $\quad$ **end for**
12: $\quad \tilde{x}_s = \frac{1}{m}\sum_{t=1}^m x_t$; $\tilde{y}_s = \frac{1}{m}\sum_{t=1}^m y_t$; $\tilde{u}_s = -\frac{1}{\rho}(A^T)^\dagger \nabla f(\tilde{x}_s)$;
13: **end for**
14: **Output:** $\tilde{x}_s, \tilde{y}_s$;

In general, as in other stochastic algorithms, the stochastic gradient is computed based on a mini-batch of size b. The following Proposition shows that the variance can be progressively reduced. Note that this and other results in this section also hold for the batch mode, in which the whole data set is used in each iteration (i.e., $b = n$).

Proposition 1. *The variance of $\hat{\nabla} f(x_{t-1})$ is bounded by $\mathbb{E}\|\hat{\nabla} f(x_{t-1}) - \nabla f(x_{t-1})\|^2 \leq 4L_{\max}\beta(b)\left(J(x_{t-1}) - J(x_*) + J(\tilde{x}) - J(x_*)\right)$, where $L_{\max} \equiv \max_i L_i$, $\beta(b) = \frac{n-b}{b(n-1)}$, $J(x) = f(x) + \rho u_*^T Ax$, and $J(x_{t-1}) - J(x_*) + J(\tilde{x}) - J(x_*) \geq 0$.*

Using (8) and (9), $J(x) - J(x_*) = f(x) - f(x_*) - \nabla f(x_*)^T(x - x_*) = 0$ when $x \to x_*$, and thus the variance goes to zero. Moreover, as expected, the variance reduces when b increases, and goes to zero when $b = n$. However, a large b leads to a high per-iteration cost. Thus, there is a tradeoff between "high variance with cheap iterations" and "low variance with expensive iterations".

Convergence Analysis

In this section, we study the convergence w.r.t. $R(x, y) \equiv f(x) - f(x_*) - \nabla f(x_*)^T(x - x_*) + g(y) - g(y_*) - g'(y_*)^T(y - y_*)$. First, note that $R(x, y)$ is always non-negative.

Proposition 2. *$R(x, y) \geq 0$ for any x and y.*

Using the optimality conditions in (8) and (9), $R(x, y)$ can be rewritten as $f(x) + g(y) + \rho u_*^T(Ax + By - c) - (f(x_*) + g(y_*) + \rho u_*^T(Ax_* + By_* - c))$, which is the difference of the Lagrangians in (3) evaluated at (x, y, u_*) and (x_*, y_*, u_*). Moreover, $R(x, y) \geq 0$ is the same as the variational inequality used in [He and Yuan, 2012].

The following shows that Algorithm 1 converges linearly.

Theorem 1. *Let*

$$\kappa = \frac{\|G + \eta\rho A^T A\|}{\lambda_f \eta(1 - 4L_{\max}\eta\beta(b))m} + \frac{4L_{\max}\eta\beta(b)(m+1)}{(1 - 4L_{\max}\eta\beta(b))m} + \frac{L_f}{\rho(1 - 4L_{\max}\eta\beta(b))\sigma_{\min}(AA^T)m}. \quad (13)$$

Choose $0 < \eta < \min\left\{\frac{1}{L_f}, \frac{1}{4L_{\max}\beta(b)}\right\}$*, and the number of iterations* m *is sufficiently large such that* $\kappa < 1$*. Then,* $\mathbb{E}R(\tilde{x}_s, \tilde{y}_s) \leq \kappa^s R(\tilde{x}_0, \tilde{y}_0)$.

Theorem 1 is similar to the SVRG results in [Johnson and Zhang, 2013; Xiao and Zhang, 2014]. However, it is not a trivial extension because of the presence of the equality constraint and Lagrangian multipliers in the ADMM formulation. Moreover, for the existing stochastic ADMM algorithms, linear convergence is only proved in SDCA-ADMM for a special case ($B = -I$ and $c = 0$ in (3)). Here, we have linear convergence for a general B and any $G \succeq I$ (in step 9).

Corollary 1. *For a fixed* κ *and* $\epsilon > 0$*, the number of stages* s *required to ensure* $\mathbb{E}R(\tilde{x}_s, \tilde{y}_s) \leq \epsilon$ *is* $s \geq \log\left(\frac{R(\tilde{x}_0,\tilde{y}_0)}{\epsilon}\right) / \log\left(\frac{1}{\kappa}\right)$*. Moreover, for any* $\delta \in (0,1)$*, we have the high-probability bound:* $\text{Prob}(R(\tilde{x}_s, \tilde{y}_s) \leq \epsilon) \geq 1 - \delta$ *if* $s \geq \log\left(\frac{R(\tilde{x}_0,\tilde{y}_0)}{\epsilon\delta}\right) / \log\left(\frac{1}{\kappa}\right)$.

Optimal ADMM Parameter ρ

With linearization, the first term in (13) becomes $\|\gamma I\|/(\lambda_f\eta(1 - 4L_{\max}\eta\beta(b))m)$. Obviously, it is desirable to have a small convergence factor κ, and so we will always use $\gamma = \gamma_{\min}$ in (11). The following Proposition obtains the optimal ρ_*, which yields the smallest κ value and thus fastest convergence. Interestingly, this ρ_* is the same as that of its batch counterpart (Theorem 7 in [Nishihara *et al.*, 2015]). In other words, the optimal ρ_* is not affected by the stochastic approximation.

Proposition 3. *The smallest* κ *is obtained when* $\rho = \rho_* \equiv \sqrt{\frac{L_f\lambda_f}{\sigma_{\max}(AA^T)\sigma_{\min}(AA^T)}}$.

3.2 General Convex Problems

In this section, we consider (general) convex problems, and only Assumptions 1, 2 are needed. The procedure (Algorithm 2) differs slightly from Algorithm 1 in the initialization of each stage (steps 2, 5, 12) and the final output (step 14).

As expected, with a weaker form of convexity, the convergence rate of Algorithm 2 is no longer linear. Following [Ouyang *et al.*, 2013; Suzuki, 2013; Zhong and Kwok, 2014], we consider the convergence of $R(\bar{x}, \bar{y}) + \zeta\|A\bar{x} + B\bar{y} - c\|$, where $\zeta > 0$ and $\|A\bar{x} + B\bar{y} - c\|$ measures the feasibility of the ADMM solution. The following Theorem shows that Algorithm 2 has $O(1/s)$ convergence. Since both $R(\bar{x}, \bar{y})$ and $\|A\bar{x} + B\bar{y} - c\|$ are always nonnegative, obviously each term individually also has $O(1/s)$ convergence.

Theorem 2. *Choose* $0 < \eta < \min\left\{\frac{1}{L_f}, \frac{1}{8L_{\max}\beta(b)}\right\}$*. Then,*

$$\begin{aligned}
&\mathbb{E}(R(\bar{x}, \bar{y}) + \zeta\|A\bar{x} + B\bar{y} - c\|) \\
&\leq \frac{4L_{\max}\eta\beta(b)(m+1)}{(1-8L_{\max}\eta\beta(b))ms}\left(f(\hat{x}_0) - f(x_*) - \nabla f(x_*)^T(\hat{x}_0 - x_*)\right) \\
&\quad + \frac{\frac{1}{2\eta}\|\hat{x}_0 - x_*\|^2_{G+\eta\rho A^TA} + \rho\left(\|\hat{u}_0 - u_*\|^2 + \frac{\zeta^2}{\rho^2}\right)}{(1 - 8L_{\max}\eta\beta(b))ms}. \quad (14)
\end{aligned}$$

The following Corollary obtains a sublinear convergence rate for the batch case ($b = n$). This is similar to that of Remark 1 in [Ouyang *et al.*, 2013]. However, here we allow a general G while they require $G = I$.

Algorithm 2 SVRG-ADMM for general convex problems.

1: **Input:** $m, \eta, \rho > 0$.
2: initialize $\tilde{x}_0 = \hat{x}_0, \hat{y}_0$ and $\hat{u}_0$;
3: **for** $s = 1, 2, \ldots$ **do**
4: $\quad\tilde{x} = \tilde{x}_{s-1}$;
5: $\quad x_0 = \hat{x}_{s-1}$; $y_0 = \hat{y}_{s-1}$; $u_0 = \hat{u}_{s-1}$;
6: $\quad\tilde{z} = \frac{1}{n}\sum_{i=1}^n \nabla f_i(\tilde{x})$;
7: $\quad$**for** $t = 1, 2, \ldots, m$ **do**
8: $\quad\quad y_t \leftarrow \arg\min_y g(y) + \frac{\rho}{2}\|Ax_{t-1} + By - c + u_{t-1}\|^2$;
9: $\quad\quad x_t \leftarrow \arg\min_x \hat{\nabla} f(x_{t-1})^T x + \frac{\rho}{2}\|Ax + By_t - c + u_{t-1}\|^2 + \frac{\|x - x_{t-1}\|^2_G}{2\eta}$;
10: $\quad\quad u_t \leftarrow u_{t-1} + Ax_t + By_t - c$;
11: $\quad$**end for**
12: $\quad\tilde{x}_s = \frac{1}{m}\sum_{t=1}^m x_t$; $\tilde{y}_s = \frac{1}{m}\sum_{t=1}^m y_t$; $\hat{x}_s = x_m$; $\hat{y}_s = y_m$; $\hat{u}_s = u_m$;
13: **end for**
14: **Output:** $\bar{x} = \frac{1}{s}\sum_{i=1}^s \tilde{x}_i, \bar{y} = \frac{1}{s}\sum_{i=1}^s \tilde{y}_s$.

Corollary 2. *In batch learning,* $R(\bar{x}, \bar{y}) + \zeta\|A\bar{x} + b\bar{y} - c\| \leq \frac{1}{2\eta ms}\|\tilde{x}_0 - x_*\|^2_{G+\eta\rho A^TA} + \frac{\rho}{ms}\left(\|\tilde{u}_0 - u_*\|^2 + \frac{\zeta^2}{\rho^2}\right)$.

3.3 Comparison with SCAS-ADMM

The recently proposed SCAS-ADMM [Zhao *et al.*, 2015] is a more rudimentary integration of SVRG and ADMM. The main difference with our method is that SCAS-ADMM moves the updates of y and u outside the inner **for** loop. As such, the inner **for** loop focuses only on updating x, and is the same as using a one-stage SVRG to solve for an inexact x solution in (5). Variables y and u are not updated until the x subproblem has been approximately solved (after running m updates of x).

In contrast, we replace the x subproblem in (5) with its first-order stochastic approximation, and then update y and u in every iteration as x. This difference is analogous to that between the Jacobi iteration and Gauss-Seidel iteration. The use of first-order stochastic approximation has also shown clear speed advantage in other stochastic ADMM algorithms [Ouyang *et al.*, 2013; Suzuki, 2013; Zhong and Kwok, 2014; Suzuki, 2014], and is especially desirable on big data sets.

As a result, the convergence rates of SCAS-ADMM are inferior to those of SVRG-ADMM. On strongly convex problems, SVRG-ADMM attains a linear convergence rate, while SCAS-ADMM only has $O(1/s)$ convergence. On general convex problems, both SVRG-ADMM and SCAS-ADMM have a convergence rate of $O(1/s)$. However, SCAS-ADMM requires the stepsize to be gradually reduced as $O(1/s^\delta)$, where $\delta > 1$. This defeats the original purpose of using SVRG-based algorithms (e.g., SVRG-ADMM), which aims at using a constant learning rate for faster convergence [Johnson and Zhang, 2013]. Moreover, (14) shows that our rate consists of three components, which converge as $O(1/s)$, $O(1/(ms))$ and $O(1/(ms))$, respectively. On the other hand, while the sublinear convergence bound in SCAS-ADMM also has three similar components, they all converge as $O(1/s)$.

Table 1: Convergence rates and space requirements of various stochastic ADMM algorithms, including stochastic ADMM (STOC-ADMM) [Ouyang *et al.*, 2013], online proximal gradient descent ADMM (OPG-ADMM) [Suzuki, 2013], regularized dual averaging ADMM (RDA-ADMM) [Suzuki, 2013], stochastic averaged gradient ADMM (SAG-ADMM) [Zhong and Kwok, 2014], stochastic dual coordinate ascent ADMM (SDCA-ADMM) [Suzuki, 2014], scalable stochastic ADMM (SCAS-ADMM) [Zhao *et al.*, 2015], and the proposed SVRG-ADMM. Here, $d, \tilde{d}$ are dimensionalities of x and y in (3).

	general convex	strongly convex	space requirement
STOC-ADMM	$O(1/\sqrt{T})$	$O(\log T/T)$	$O(d\tilde{d} + d^2)$
OPG-ADMM	$O(1/\sqrt{T})$	$O(\log T/T)$	$O(d\tilde{d})$
RDA-ADMM	$O(1/\sqrt{T})$	$O(\log T/T)$	$O(d\tilde{d})$
SAG-ADMM	$O(1/T)$	unknown	$O(d\tilde{d} + nd)$
SDCA-ADMM	unknown	linear rate	$O(d\tilde{d} + n)$
SCAS-ADMM	$O(1/T)$	$O(1/T)$	$O(d\tilde{d})$
SVRG-ADMM	$O(1/T)$	linear rate	$O(d\tilde{d})$

To make the cost of full gradient computation less pronounced, a natural choice for m is $m = O(n)$ [Johnson and Zhang, 2013]. Hence, SCAS-ADMM can be much slower than SVRG-ADMM when n is large.

3.4 Space Requirement

The space requirements of Algorithms 1 and 2 mainly come from step 12. For simplicity, we consider $B = -I$ and $c = 0$, which are assumed in [Suzuki, 2013; 2014]. Moreover, we assume that the storage of the n old gradients can be reduced to the storage of n scalars, which is often the case in many machine learning models [Johnson and Zhang, 2013].

A summary of the space requirements and convergence rates for various stochastic ADMM algorithms is shown in Table 1. As can be seen, among those with variance reduction, the space requirements of SCAS-ADMM and SVRG-ADMM are independent of the sample size n. However, as discussed in the previous section, SVRG-ADMM has much faster convergence rates than SCAS-ADMM on both strongly convex and general convex problems.

4 Experiments

4.1 Graph-Guided Fused Lasso

We perform experiments on the generalized lasso model $\sum_{i=1}^{n} \ell_i(x) + \|Ax\|_1$, where ℓ_i is the logistic loss on sample i, and A is a matrix encoding the feature sparsity pattern. Here, we use graph-guided fused lasso [Kim *et al.*, 2009] and set $A = [G; I]$, where G is the sparsity pattern of the graph obtained by sparse inverse covariance estimation [Friedman *et al.*, 2008]. For the ADMM formulation, we introduce an additional variable y and the constraint $Ax = y$. Experiments are performed on four benchmark data sets[1] (Table 2). We use a mini-batch size of $b = 100$ on *protein* and *covertype*; and $b = 500$ on *mnist8m* and *dna*. Experiments are performed on a PC with Intel i7-3770 3.4GHz CPU and 32GB RAM,

[1]Downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/, http://osmot.cs.cornell.edu/kddcup/datasets.html, and http://largescale.ml.tu-berlin.de/instructions/.

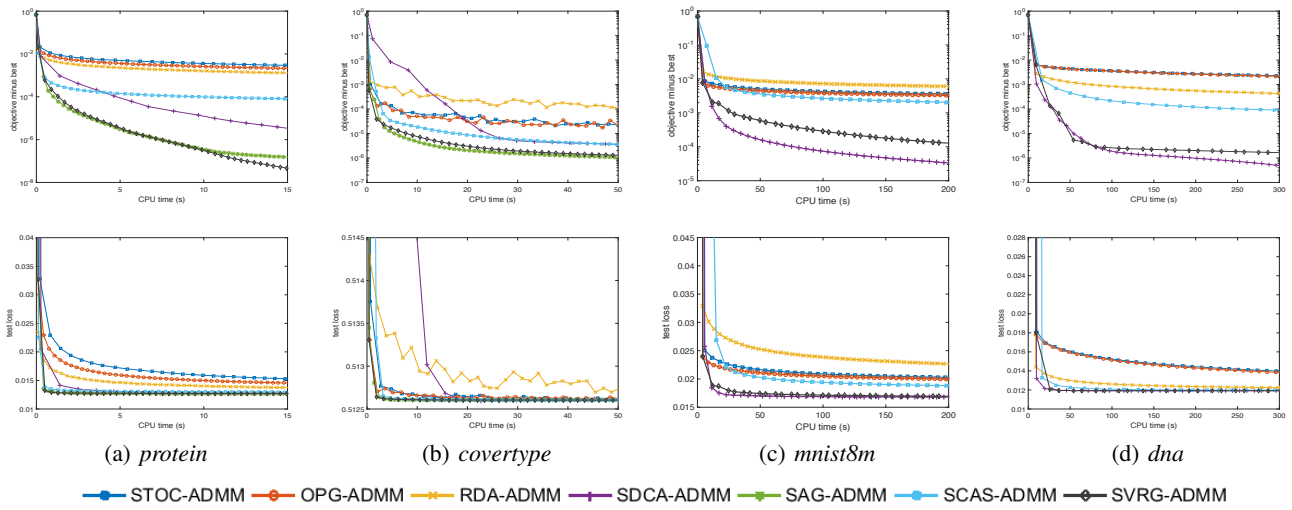
Table 2: Data sets for graph-guided fused lasso.

	#training	#test	dimensionality
protein	72,876	72,875	74
covertype	290,506	290,506	54
mnist8m	1,404,756	351,189	784
dna	2,400,000	600,000	800

All methods listed in Table 1 are compared and in Matlab. The proposed SVRG-ADMM uses the linearized update in (12) and $m = 2n/b$. For further speedup, we simply use the last iterates in each stage (x_m, y_m, u_m) as $\tilde{x}_s, \tilde{y}_s, \tilde{u}_s$ in step 12 of Algorithms 1 and 2. Both SAG-ADMM and SVRG-ADMM are initialized by running OPG-ADMM for n/b iterations.[2] For SVRG-ADMM, since the learning rate in (12) is effectively η/γ, we set $\gamma = 1$ and only tune η. All parameters are tuned as in [Zhong and Kwok, 2014]. Each stochastic algorithm is run on a small training subset for a few data passes (or stages). The parameter setting with the smallest training objective is then chosen. To ensure that the ADMM constraint is satisfied, we report the performance based on (x_t, Ax_t). Results are averaged over five repetitions.

Figure 1 shows the objective values and testing losses versus CPU time. SAG-ADMM cannot be run on *mnist8m* and *dna* because of its large memory requirement (storing the weights already takes 8.2GB for *mnist8m*, and 14.3GB for *dna*). As can be seen, stochastic ADMM methods with variance reduction (SVRG-ADMM, SAG-ADMM and SDCA-ADMM) have fast convergence, while those that do not use variance reduction are much slower. SVRG-ADMM, SAG-ADMM and SDCA-ADMM have comparable speeds, but SVRG-ADMM requires much less storage (see also Table 1). On the medium-sized *protein* and *covertype*, SCAS-ADMM has comparable performance with the other stochastic ADMM variants using variance reduction. However, it becomes much slower on the larger *minist8m* and *dna*, which is consistent with the analysis in Section 3.3.

[2]This extra CPU time is counted towards the first stages of SAG-ADMM and SVRG-ADMM.



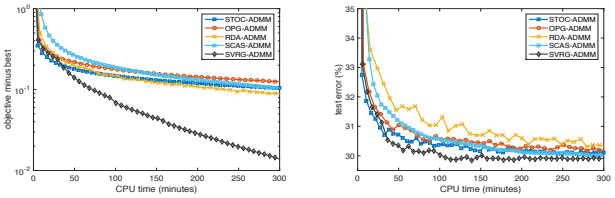
(a) *protein* (b) *covertype* (c) *mnist8m* (d) *dna*

Figure 1: Performance vs CPU time (in sec) on graph-guided fused lasso (Top: objective value; Bottom: testing loss).

4.2 Multitask Learning

When there are a large number of outputs, the much smaller space requirement of SVRG-ADMM is clearly advantageous. In this section, experiments are performed on an 1000-class ImageNet data set [Russakovsky *et al.*, 2015]. We use 1,281,167 images for training, and $50,000$ images for testing. 4096 features are extracted from the last fully connected layer of the convolutional net VGG-16 [Simonyan and Zisserman, 2014]. The multitask learning problem is formulated as: $\min_X \sum_{i=1}^N \ell_i(X) + \lambda_1\|X\|_1 + \lambda_2\|X\|_*$, where $X \in \mathbb{R}^{d\times N}$ is the parameter matrix, N is the number of tasks, d is the feature dimensionality, ℓ_i is the multinomial logistic loss on the ith task, and $\|\cdot\|_*$ is the nuclear norm. To solve this problem using ADMM, we introduce an additional variable X' with the constraint $X' = X$. On setting $A = [I; I]$, the regularizer is then $g(AX) = g([X; X']) = \lambda_1\|X\|_1 + \lambda_2\|X'\|_*$. We set $\lambda_1 = 10^{-5}$, $\lambda_2 = 10^{-4}$, and use a mini-batch size $b = 500$. SAG-ADMM requires 38.2TB for storing the weights, and SDCA-ADMM 9.6GB for the dual variables, while SVRG-ADMM requires 62.5MB for storing $\tilde{x}$ and the full gradient.

Figure 2 shows the objective value and testing error versus time. SVRG-ADMM converges rapidly to a good solution. The other non-variance-reduced stochastic ADMM algorithms are very aggressive initially, but quickly get much slower. SCAS-ADMM is again slow on this large data set.



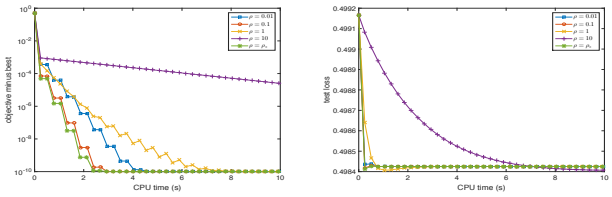
(a) objective. (b) testing error (%).

Figure 2: Performance vs CPU time (in min) on ImageNet.

4.3 Varying ρ

Finally, we perform experiments on total-variation (TV) regression [Boyd *et al.*, 2011] to demonstrate the effect of ρ. Samples z_i's are generated with i.i.d. components from the standard normal distribution. Each z_i is then normalized to $\|z_i\| = 1$. The parameter x is generated according to http://www.stanford.edu/~boyd/papers/admm/. The output o_i is obtained by adding standard Gaussian noise to $x^T z_i$. Given n samples $\{(z_1, o_1), \ldots, (z_n, o_n)\}$, TV regression is formulated as: $\min_x \frac{1}{2n}\sum_{i=1}^n \|o_i - x^T z_i\|^2 + \lambda\|Ax\|_1$, where $A_{ij} = 1$ if $i = j$; -1 if $j = i + 1$; and 0 otherwise.

We set $n = 100,000$, $d = 500$, $\lambda = 0.1/\sqrt{n}$, and a mini-batch size $b = 100$. Figure 3 shows the objective value and testing loss versus CPU time, with different ρ's. As can be seen, ρ_* in Proposition 3 outperforms the other choices of ρ.



(a) objective. (b) testing loss.

Figure 3: Performance of SVRG-ADMM at different ρ's.

5 Conclusion

This paper proposed a non-trivial integration of SVRG and ADMM. Its theoretical convergence rates are as fast as existing variance-reduced stochastic ADMM algorithms, but its storage requirement is much lower, even independent of the sample size. Experimental results demonstrate its benefits over other stochastic ADMM methods.

Acknowledgments

This research was supported in part by the Research Grants Council of the Hong Kong Special Administrative Region (Grant 614513).

References

[Bottou, 2004] L. Bottou. Stochastic learning. In *Advanced Lectures on Machine Learning*, pages 146–168. Springer Verlag, 2004.

[Boyd *et al.*, 2011] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

[Defazio *et al.*, 2014] A. Defazio, F. Bach, and S. Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, pages 2116–2124, 2014.

[Deng and Yin, 2015] W. Deng and W. Yin. On the global and linear convergence of the generalized alternating direction method of multipliers. *Journal of Scientific Computing*, pages 1–28, 2015.

[Friedman *et al.*, 2008] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[Ghadimi *et al.*, 2015] E. Ghadimi, A. Teixeira, I. Shames, and M. Johansson. Optimal parameter selection for the alternating direction method of multipliers (ADMM): Quadratic problems. *IEEE Transactions on Automatic Control*, 60(3):644–658, 2015.

[Giselsson and Boyd, 2014] P. Giselsson and S. Boyd. Diagonal scaling in Douglas-Rachford splitting and ADMM. In *Proceedings of the 53rd IEEE Conference on Decision and Control*, 2014.

[Golub and Van Loan, 2012] G.H. Golub and C.F. Van Loan. *Matrix Computations*. JHU Press, 2012.

[He and Yuan, 2012] B. He and X. Yuan. 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 *et al.*, 2009] L. Jacob, G. Obozinski, and J.-P. Vert. Group lasso with overlap and graph lasso. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 433–440, 2009.

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

[Kim *et al.*, 2009] S. Kim, K. A. Sohn, and E. P. Xing. A multivariate regression approach to association analysis of a quantitative trait network. *Bioinformatics*, 25(12):i204–i212, 2009.

[Nishihara *et al.*, 2015] R. Nishihara, L. Lessard, B. Recht, A. Packard, and M. I. Jordan. A general analysis of the convergence of ADMM. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 343–352, 2015.

[Ouyang *et al.*, 2013] H. Ouyang, N. He, L. Tran, and A. Gray. Stochastic alternating direction method of multipliers. In *Proceedings of the 30th International Conference on Machine Learning*, pages 80–88, 2013.

[Parikh and Boyd, 2014] Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):127–239, 2014.

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

[Russakovsky *et al.*, 2015] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and Li F.-F. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[Shalev-Shwartz and Zhang, 2013] S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss. *Journal of Machine Learning Research*, 14(1):567–599, 2013.

[Simonyan and Zisserman, 2014] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. Technical Report arXiv:1409.1556, 2014.

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

[Suzuki, 2014] T. Suzuki. Stochastic dual coordinate ascent with alternating direction method of multipliers. In *Proceedings of the 31st International Conference on Machine Learning*, pages 736–744, 2014.

[Wang and Banerjee, 2012] H. Wang and A. Banerjee. Online alternating direction method. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1119–1126, 2012.

[Xiao and Zhang, 2014] L. Xiao and T. Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4), 2014.

[Zhang *et al.*, 2011] X. Zhang, M. Burger, and S. Osher. A unified primal-dual algorithm framework based on Bregman iteration. *Journal of Scientific Computing*, 46(1):20–46, 2011.

[Zhao *et al.*, 2015] S. Y. Zhao, W. J. Li, and Z. H. Zhou. Scalable stochastic alternating direction method of multipliers. Technical Report arXiv:1502.03529, 2015.

[Zhong and Kwok, 2014] W. Zhong and J. Kwok. Fast stochastic alternating direction method of multipliers. In *Proceedings of the 31st International Conference on Machine Learning*, pages 46–54, 2014.