Towards an optimal stochastic alternating direction method of multipliers

Samaneh Azadi SAZADI156@GMAIL.COM
UC Berkeley, Berkeley, CA
School of ECE, Shiraz University, Shiraz, Iran

Suvrit Sra SUVRIT@TUEBINGEN.MPG.DE
Carnegie Mellon University, Pittsburgh
Max Planck Institute for Intelligent Systems, Tübingen, Germany

Abstract

We study regularized stochastic convex optimization subject to linear equality constraints. This class of problems was recently also studied by Ouyang et al. (2013) and Suzuki (2013); both introduced similar stochastic alternating direction method of multipliers (SADMM) algorithms. However, the analysis of both papers led to suboptimal convergence rates. This paper presents two new SADMM methods: (i) the first attains the minimax optimal rate of $O(1/k)$ for nonsmooth strongly-convex stochastic problems; while (ii) the second progresses towards an *optimal* rate by exhibiting an $O(1/k^2)$ rate for the *smooth* part. We present several experiments with our new methods; the results indicate improved performance over competing ADMM methods.

1. Introduction

We study stochastic optimization problems of the form

$$\begin{aligned} \min \quad & (f(x) := \mathbb{E}[F(x,\xi)]) + h(y), \\ \text{s.t.} \quad & Ax + By = b, \text{ and } x \in \mathcal{X}, y \in \mathcal{Y}, \end{aligned} \tag{1}$$

where ξ follows some distribution over a space Ξ, so that $f(x) = \int_\Xi F(x,\xi)dP(\xi)$, and for each ξ, function $F(\cdot,\xi)$ is closed and convex. The function h is assumed to be closed and convex, while $\mathcal{X}$ and $\mathcal{Y}$ are compact convex sets.

Problem (1) enjoys great importance in machine learning: the function $f(x)$ typically represents a loss over all data, while $h(y)$ enforces structure or regularizes the learning model and aids generalization (Srebro & Tewari, 2010).

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

The linear constraints in (1) allow us to decouple f and h, and thereby consider sophisticated regularizers without having to rely on carefully tuned proximity operators—this can greatly simplify the overall optimization algorithm and is a substantial gain (Ouyang et al., 2013; Boyd et al., 2011).

Using linear constraints to "split variables" is an idea that has been most impressively exploited by the alternating direction method of multipliers (ADMM). As a consequence, ADMM leads to methods that are easy to implement, scale well, and are widely applicable—these benefits are eminently advocated in recent survey (Boyd et al., 2011), and is also the primary motivation of (Ouyang et al., 2013; Suzuki, 2013). Indeed, these benefits have also borne through in applications such large-scale lasso (Boyd et al., 2011), constrained image deblurring (Chan et al., 2013), and matrix completion (Goldfarb et al., 2012), to name a few.

But despite their broad applicability, traditional ADMM methods cannot handle stochastic optimization, a drawback recently circumvented by Ouyang et al. (2013) and (Suzuki, 2013), who approached (1) using an ADMM strategy combined with ideas from ordinary stochastic convex optimization. Both (Ouyang et al., 2013; Suzuki, 2013) showed some experiments that suggested benefits of combining stochastic ideas with an ADMM strategy. A key benefit of such a combination is that it allows one to tackle stochastic problems with sophisticated regularization penalties such as graph-structured norms and overlapping group norms (Parikh & Boyd, 2013), more easily than either batch or online proximal splitting methods (Ghadimi & Lan, 2012; Duchi & Singer, 2009; Beck & Teboulle, 2009).

We remark that (deterministic) ADMM family of methods are now widely used and substantial engineering effort has been invested into deploying them, both in research as well as industry (see e.g., (Boyd et al., 2011); and

also (Kraska et al., 2013)). Thus, enriching ADMM to handle stochastic optimization problems may be of great practical value.

Contributions. The main contributions of this paper are:

- A new SADMM algorithm (Alg. 2) for strongly convex stochastic optimization that achieves the minmax optimal $O(1/k)$ convergence rate; this improves on the previously shown suboptimal $O(\log k/k)$ rates (Ouyang et al., 2013; Suzuki, 2013).
- A new SADMM algorithm (Alg. 3) for stochastic convex problems with Lipschitz continuous gradients; this method achieves an $O(1/k^2)$ rate in the smooth component, improving on the previous $O(1/k)$ rates of Ouyang et al. (2013); Suzuki (2013).

Empirical results (§4) indicate that when f is strongly convex or smooth, our new methods outperform basic SADMM. We note in passing that our analysis extends to yield high-probability bounds assuming light-tailed on the errors; this extension is fairly routine, so we omit it for lack of space (see §5 also for other possible extensions).

2. Background

ADMM is a special case of the Douglas-Rachford (Douglas & Rachford, 1956) splitting method, which itself may be viewed as an instance of the proximal-point algorithm (Rockafellar, 1976; Eckstein & Bertsekas, 1992). But before discussing SADMM, let us first recall some material about ADMM.

Consider the classic convex optimization problem

$$\min_x \quad f(x) + h(Ax - b), \qquad x \in \mathcal{X}, \tag{2}$$

where f, h are closed convex functions and $\mathcal{X}$ is a closed convex set. Introducing $y = Ax - b$, problem (2) becomes

$$\min \quad f(x)+h(y), \quad x \in \mathcal{X},\ Ax-y-b = 0,\ y \in \operatorname{dom} h.$$

ADMM considers the slightly more general problem

$$\min_{x\in\mathcal{X}, y\in\mathcal{Y}} \quad f(x) + h(y), \quad Ax + By - b = 0, \tag{3}$$

to solve which it introduces an *augmented Lagrangian*

$$L_\beta(x, y, \lambda) := f(x) + h(y) - \langle \lambda,\ Ax + By - b\rangle + \tfrac{\beta}{2}\|Ax + By - b\|_2^2. \tag{4}$$

Here λ is the dual variable, and β is a penalty parameter. ADMM minimizes (4) over x and y in a Gauss-Seidel manner, followed by a step that updates the dual variable. The resulting method is summarized as Alg. 1.

For stochastic problems with $f(x) = \int_\Xi F(x,\xi)dP(\xi)$ over a potentially unknown distribution P, the standard ADMM scheme is no longer applicable. Ouyang et al. (2013) suggest *linearizing* $f(x)$ by considering a modified augmented Lagrangian that now depends on subgradients of f. Specifically, the augmented Lagrangian they use is

$$L_\beta^k(x, y, \lambda) := f(x_k)+\langle g_k,\ x\rangle+h(y)-\langle \lambda,\ Ax + By - b\rangle + \tfrac{\beta}{2}\|Ax + By - b\|_2^2 + \tfrac{1}{2\eta_k}\|x - x_k\|_2^2, \tag{5}$$

where g_k is a stochastic (sub)gradient of f, i.e., $\mathbb{E}[g_k] \in \partial f(x_k)$, where $g_k \in \partial F(x, \xi_{k+1})$. Replacing L_β by this iteration dependent L_β^k in Alg. 1 one obtains the SADMM method of (Ouyang et al., 2013). The $\|x - x_k\|_2^2$ prox-term ensures that (5) has a unique solution, even if the augmented Lagrangian (AL) fails to be strictly convex; it also aids the convergence analysis.

1 Initialize: x_0, y_0, and λ_0.
2 **for** $k \geq 0$ **do**
3 $\quad x_{k+1} \leftarrow \operatorname{argmin}_{x\in\mathcal{X}}\{L_\beta(x, y_k, \lambda_k)\}$
4 $\quad y_{k+1} \leftarrow \operatorname{argmin}_{y\in\mathcal{Y}}\{L_\beta(x_{k+1}, y, \lambda_k)\}$
5 $\quad \lambda_{k+1} \leftarrow \lambda_k - \beta(Az_{k+1} + By_{k+1} - b)$
6 **end**

Algorithm 1: ADMM

Since SADMM borrows techniques from stochastic subgradient methods, it is natural to expect similarities in convergence guarantees. Previous authors (Ouyang et al., 2013; Suzuki, 2013) showed that for uniformly averaged iterates ($\bar{x}_k := \frac{1}{k}\sum_j x_j$, etc.) one obtains the approximation bound

$$\mathbb{E}[f(\bar{x}_k)+h(\bar{y}_k)-f(x^*)-h(y^*)+\rho\|A\bar{x}_k + B\bar{y}_k - b\|_2] = O(\tfrac{1}{\sqrt{k}}).$$

Exploiting properties of the stochastic part f, we can obtain more refined rates. Specifically, if f is strongly convex, the rate $E[\cdot] = O(\frac{\log k}{k})$, and if f is Lipschitz smooth, the rate $E[\cdot] = O(\frac{1}{k}) + O(\frac{1}{\sqrt{k}})$ was shown by both Ouyang et al. (2013) and Suzuki (2013) (though for the RDA-ADMM variant in (Suzuki, 2013), no refined rate for strongly convex losses was proved).

However, these rates are suboptimal. It is of great interest to obtain optimal rates—see the long line of work in stochastic optimization (Nemirovski et al., 2009; Chen et al., 2012; Shamir & Zhang, 2013; Ghadimi & Lan, 2012), where impressive effort has been expended to obtain optimal rates; this effort can even translate into improved empirical performance. For deterministic settings, notable examples of optimal methods are given by (Beck & Teboulle, 2009; Nesterov, 2007), which often substantially outperform their non-optimal counterparts.

In light of this background, we are now ready to present new SADMM methods, which achieve the minimax opti-

mal $O(\frac{1}{k})$ rate for strongly convex losses. Without strong convexity, the rate for the nonsmooth and stochastic parts is $O(\frac{1}{k})+O(\frac{1}{\sqrt{k}})$, with a more refined (and optimal) $O(\frac{1}{k^2})$ contribution from the smooth part of the objective.

3. SADMM

We begin by stating our key structural assumptions:

1. Bounded subgradients: $\mathbb{E}[\|g_k\|_2^2] \leq G^2$ (for the strongly convex case)
2. Bounded noise variance: $\mathbb{E}[\|g_k - \nabla f(x)\|_2^2] \leq \sigma^2$ (for the smooth case).
3. Compactness of $\mathcal{X}$, $\mathcal{Y}$; bounded dual variables.

Unfortunately, we must impose somewhat stricter assumptions than ordinary SADMM—this seems to be the price that we have to pay for faster convergence—the discussion in (Chambolle & Pock, 2011) sheds more light on these aspects (especially Assumption 3).

Following He & Yuan (2012) we introduce the notation

$$w := [x^T; y^T; \lambda^T]^T, \quad w_k := [x_k^T; y_k^T; \lambda_k^T]^T,$$
$$F(w) := \begin{bmatrix} -A^T\lambda \\ -B^T\lambda \\ Ax + By - b \end{bmatrix}, \quad \Omega := \begin{bmatrix} \mathcal{X} \\ \mathcal{Y} \\ \mathbb{R}^m \end{bmatrix}. \tag{6}$$

The operator $F(\cdot)$ satisfies a simple but useful property

$$\langle u - v,\, F(u) - F(v)\rangle = 0, \qquad u, v \in \operatorname{dom} F. \tag{7}$$

Moreover, for an optimal $w^* \in \Omega$, and any $w \in \Omega$, we have

$$f(x) - f(x^*) + h(y) - h(y^*) + \langle w - w^*,\, F(w^*)\rangle \geq 0. \tag{8}$$

Therefore, a vector $\bar{w} \in \Omega$ is ϵ-optimal for the deterministic ADMM problem (for $\epsilon > 0$) if it satisfies

$$f(\bar{x}) - f(x) + h(\bar{y}) - h(y) + \langle \bar{w} - w,\, F(w)\rangle \leq \epsilon, \quad \forall w \in \Omega.$$

As in (He & Yuan, 2012), Ouyang et al. (2013) also use this variational characterization of optimality and seek to bound it in expectation. We too use this characterization; we first estimate it after one step of our SADMM algorithm to eventually bound it in expectation. We are now ready to describe the our SADMM algorithms (§3.1 and §3.2).

3.1. SADMM for strongly convex f

When f is μ-strongly convex, we use essentially the same SADMM method as in (Ouyang et al., 2013) (shown as Alg. 2). The key difference lies in how the iterates generated by the Alg. 2 are averaged to obtain an optimal rate.

Initialize: x_0, y_0, and λ_0
for $k \geq 0$ **do**
  Obtain stochastic gradient g_k; build L_β^k via (5)
  $x_{k+1} \leftarrow \operatorname{argmin}_{x\in\mathcal{X}}\{L_\beta^k(x, y_k, \lambda_k)\}$
  $y_{k+1} \leftarrow \operatorname{argmin}_{y\in\mathcal{Y}}\{L_\beta^k(x_{k+1}, y, \lambda_k)\}$
  $\lambda_{k+1} \leftarrow \lambda_k - \beta(Ax_{k+1} + By_{k+1} - b)$
end

Algorithm 2: Stochastic ADMM (strongly convex)

We begin our analysis by introducing some more notation

$$\Delta_k := \|x_k - x\|_2^2, \quad k = 0, 1, \ldots,$$
$$\mathcal{A}_k := \|Ax + By_k - b\|_2^2, \quad \mathcal{L}_k := \|\lambda - \lambda_k\|_2^2$$
$$D_\mathcal{Y} := \sup_{y\in\mathcal{Y}} \|B(y - y^*)\|_2, \quad \|\lambda_k\|_2 \leq \rho.$$

Here, Δ_k is related to the diameter of the primal variable $x \in \mathcal{X}$, $\mathcal{A}_k$ measures how well the linear constraints are satisfied, $\mathcal{L}_k$ measures distance between dual variables, $D_\mathcal{Y}$ measures a diameter-like term for the primal variable $y \in \mathcal{Y}$, while ρ is a parameter that bounds the dual variables.

Lemma 1 is a key result that describes progress made at one step. Upon taking suitable expectations, it leads to Thm. 2.

Lemma 1. *Let f be μ-strongly convex, and let x_{k+1}, y_{k+1} and λ_{k+1} be computed as per Alg. 2. For all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, and $w \in \Omega$, it holds for $k \geq 0$ that*

$$f(x_k) - f(x) + h(y_{k+1}) - h(y) + \langle w_{k+1} - w,\, F(w_{k+1})\rangle$$
$$\leq \tfrac{\eta_k}{2}\|g_k\|_2^2 - \tfrac{\mu}{2}\Delta_k + \tfrac{1}{2\eta_k}[\Delta_k - \Delta_{k+1}] + \tfrac{\beta}{2}[\mathcal{A}_k - \mathcal{A}_{k+1}]$$
$$+ \tfrac{1}{2\beta}[\mathcal{L}_k - \mathcal{L}_{k+1}] + \langle \delta_k,\, x_k - x\rangle.$$

To use Lemma 1 to obtain an optimal SADMM method, we use an idea that has also found success for the stochastic subgradient method—see e.g., (Lacoste-Julien et al., 2012; Shamir & Zhang, 2013)—the idea is to use nonuniform averaging of the iterates where more recent iterates are given higher weight. For SADMM, some care is required to ensure that the nonuniform weighting does not conflict with the augmented Lagrangian (AL) terms.

We propose to use the following weighted iterates:[1]

$$\bar{x}_k := \tfrac{2}{k(k+1)}\sum_{j=0}^{k-1}(j+1)x_j, \quad \bar{y}_k := \tfrac{2}{k(k+1)}\sum_{j=1}^{k} j y_j,$$
$$\bar{\lambda}_k := \tfrac{2}{k(k+1)}\sum_{j=1}^{k} j\lambda_j. \tag{9}$$

It is important to note that these weighted averages can be maintained in an online manner. Indeed, given $\bar{x}_{k-1}$, we

[1] Since x is treated asymmetrically from y by all SADMM variants (optimizing x involves subgradients, y does not), it is no surprise that the weighted average of the previous iterates that we use for x is slightly different from what we use for y.

can update the weighted average as

$$\bar{x}_k = (1-\theta_k)\bar{x}_{k-1} + \theta_k x_k, \qquad k \geq 1, \qquad (10)$$

where $\theta_k = 2/(k+2)$; similar updates apply for $\bar{y}_k$ and $\bar{\lambda}_k$. These weighted averages in combination with Lemma 1 help prove our main theorem on SADMM.

Theorem 2. *Let f be μ-strongly convex. Let $\eta_k = \frac{2}{\mu(k+2)}$, let x, y_j, λ_j be generated by Alg. 2, and $\bar{x}_k, \bar{y}_k, \bar{\lambda}_k$ computed by (9). Let x^*, y^* be the optimal; then for $k \geq 1$,*

$$\mathbb{E}[f(\bar{x}_k) - f(x^*) + h(\bar{y}_k) - h(y^*) + \rho\|A\bar{x}_k + B\bar{y}_k - b\|_2] \leq \tfrac{2G^2}{\mu(k+1)} + \tfrac{\beta}{2(k+1)}D_{\mathcal{Y}}^2 + \tfrac{2\rho^2}{\beta(k+1)}.$$

3.2. SADMM for smooth f

Obtaining an optimal version of SADMM for Lipschitz-smooth $f \in C_L^1$ proves considerably harder.

Input: Sequence (γ_k) of interpolation parameters; $(\eta_k = (L+\alpha_k)^{-1})$, stepsizes
1 Initialize: $x_0 = z_0$, y_0.
2 **for** $k \geq 0$ **do**
3 $\quad p_k \leftarrow (1-\gamma_k)x_k + \gamma_k z_k$
4 $\quad$ Get stochastic gradient g_k s.t. $E[g_k] = \nabla f(p_k)$
5 $\quad z_{k+1} \leftarrow \operatorname{argmin}_{x\in\mathcal{X}}\{\hat{L}_\beta^k(x, y_k, \lambda_k)\}$
6 $\quad x_{k+1} \leftarrow (1-\gamma_k)x_k + \gamma_k z_{k+1}$
7 $\quad y_{k+1} \leftarrow \operatorname{argmin}_{y\in\mathcal{Y}}\{\hat{L}_\beta^k(z_{k+1}, y, \lambda_k)\}$
8 $\quad \lambda_{k+1} \leftarrow \lambda_k - \beta(Az_{k+1} + By_{k+1} - b)$
9 **end**

Algorithm 3: SADMM for smooth $f(x)$

Alg. 3 depends on several careful modifications to the basic SADMM scheme. First, it uses interpolatory sequences (p_k) and (z_k), as well as "stepsizes" γ_k (this is inspired by techniques from fast-gradient methods (Tseng, 2008; Nesterov, 2004)). Second, x is updated (*cf.* Line 4 in Alg. 2) by first computing z_{k+1}, which in turn uses a weighted prox-term that enforces proximity to z_k instead of to x_k. Third, the update to y uses an AL term that depends on z_{k+1} instead of x_{k+1}—this change is for simplifying the analysis; one could continue to use an AL term based on x_{k+1}, but at the expense of much more tedious analysis. Finally, an important modification is to the augmented Lagrangian, which is now defined as

$$\hat{L}_\beta^k(x,y,\lambda) := f(x_k) + \langle g_k, x\rangle + h(y) - \theta_k\langle\lambda, Ax+By-b\rangle + \tfrac{\beta\theta_k}{2}\|Ax+By-b\|_2^2 + \tfrac{\gamma_k}{2\eta_k}\|x-z_k\|_2^2, \quad (11)$$

for suitable parameters (θ_k, γ_k).

We begin our analysis by again stating a key lemma that measures per-step progress; here we use slightly different notation by redefining in w and w_k (6) as

$$w := [z^T, y^T, \lambda^T]^T, \quad w_k := [z_k^T, y_k^T, \lambda_k^T]^T.$$

Lemma 3. *Let $x_{k+1}, y_{k+1}, z_{k+1}$ be generated by Alg. 3. For $x \in \mathcal{X}$, $y \in \mathcal{Y}$ and $w \in \Omega$, and with $\eta_k = (L+\alpha_k)^{-1}$ the following bound holds for all $k \geq 0$:*

$$\begin{aligned} &f(x_{k+1}) + \theta_k[h(y_{k+1}) - h(y)] \\ &\qquad + \theta_k\langle w_{k+1} - w,\, F(w_{k+1})\rangle \\ &\leq (1-\gamma_k)f(x_k) + \gamma_k f(x) + \tfrac{\gamma_k^2}{2\eta_k}[\Delta_k - \Delta_{k+1}] \\ &+ \tfrac{1}{2\alpha_k}\|\delta_k\|_2^2 + \gamma_k\langle\delta_k,\, z_k - x\rangle + \tfrac{\beta\gamma_k}{2}[\mathcal{A}_k - \mathcal{A}_{k+1}] \\ &+ \tfrac{\gamma_k}{2\beta}\left[\mathcal{L}_k - \mathcal{L}_{k+1}\right]. \end{aligned}$$

The proof of this inequality is lengthy and tedious, so we leave it in the supplement. Lemma 3 proves crucial for doing the induction to obtain the next main step towards our convergence proof. Let $R := \sup_{x\in\mathcal{X}}\|x - x^*\|_2$. Then,

Lemma 4. *Using the notation of Lemma 3, for $\frac{1-\gamma_{k+1}}{\gamma_{k+1}^2} \leq \frac{1}{\gamma_k^2}$ and x such that $f(x_k) \geq f(x)$ $(\forall k)$, we have*

$$\begin{aligned} &\tfrac{1}{\gamma_k^2}(f(x_{k+1}) - f(x)) + \sum_{j=1}^k \tfrac{1}{\gamma_j}\left[h(y_{j+1}) - h(y)\right] \\ &\qquad + \tfrac{\theta_j}{\gamma_j}\left[\langle w_{j+1} - w,\, F(w_{j+1})\rangle\right] \\ &\leq \tfrac{L+\alpha_k}{2}R^2 + \tfrac{\beta}{2}\sum_{j=1}^k \mathcal{A}_j + \tfrac{1}{2\beta}\sum_{j=1}^k \mathcal{L}_j \\ &+ \sum_{j=1}^k \tfrac{1}{\gamma_j^2\alpha_j}\|\delta_j\|_2^2 + \tfrac{1}{\gamma_j}\langle\delta_j,\, z_j - x\rangle. \end{aligned}$$

As before, to state our convergence theorem, we average the iterates generated by Alg. 3 non-uniformly. This technique is borrowed from the analysis of accelerated methods, see e.g., (Ghadimi & Lan, 2012). To use Lemma 4 to obtain our main convergence result, we introduce weighted candidate solution vectors. For $k \geq 0$, we define the weighted iterates (it is important to note that these weighted averages can be easily maintained in an online manner (*cf.* formula (10)):

$$\begin{aligned} \bar{x}_k &:= x_{k+1}, \quad \bar{y}_k := \sum_{j=1}^k \nu_j y_{j+1}, \\ \bar{z}_k &:= \sum_{j=1}^k \nu_j z_{j+1}, \quad \bar{\lambda}_k := \sum_{j=1}^k \nu_j \lambda_j, \end{aligned} \qquad (12)$$

where $\nu_j = 2(j+1)/(k+1)(k+2)$. Since $f(x)$ is smooth, it turns out that in (12) we do not need to average over x_{k+1}, thereby obtaining "non-ergodic" convergence in expectation for the smooth part. This is an interesting technical difference from nonsmooth $f(x)$, where one needs to average over the x iterates too unless one is willing to pay an additional $\log k$ penalty (Shamir & Zhang, 2013). Finally, we have the following theorem.

Theorem 5. *Let $\bar{x}_k$, $\bar{z}_k$, $\bar{y}_k$, and $\bar{\lambda}_k$ be as defined in (12). Then for $\theta_j = 1$ and $k \geq 0$,*

$$\frac{1}{\gamma_k^2}\mathbb{E}[f(\bar{x}_k) - f(x^*) + h(\bar{y}_k) - h(y^*) + \rho\|A\bar{z}_k + B\bar{y}_k - b\|_2] \leq \frac{(L+\alpha_k)R^2}{2} + \frac{\beta(k+1)}{2}D_{\mathcal{Y}}^2 + \frac{(k+1)}{\beta}\rho^2 + \sum_{j=1}^{k}\frac{\sigma^2}{\gamma_j^2\alpha_j}.$$

An immediate corollary is our refined result on the convergence rate of SADMM with a smooth stochastic objective (notice that h is assumed to be nonsmooth):

Corollary 6. *Let $\alpha_j = c^{-1}\sigma(j+1)^{3/2}$ (for a constant c), and $\gamma_j = 2/(j+1)$; then*

$$\mathbb{E}[f(\bar{x}_k) - f(x^*) + h(\bar{y}_k) - h(y^*) + \rho\|A\bar{z}_k + B\bar{y}_k - b\|_2] \leq \frac{2LR^2}{(k+1)^2} + \frac{2\beta D_{\mathcal{Y}}^2}{k+1} + \frac{2\rho^2}{\beta(k+1)} + \frac{2\sigma(c^{-1}+c)}{\sqrt{k+1}}.$$

Observe that when there is no noise ($\sigma = 0$), our analysis can be slightly modified to yield the bound

$$\mathbb{E}[f(\bar{x}_k) - f(x^*) + h(\bar{y}_k) - h(y^*) + \rho\|A\bar{z}_k + B\bar{y}_k - b\|_2] \leq \frac{2LR^2}{(k+1)^2} + \frac{2\beta D_{\mathcal{Y}}^2}{k+1} + \frac{2\rho^2}{\beta(k+1)}.$$

4. Experiments

In this section we present experiments that illustrate performance of our SADMM variants. The results indicate that our methods converge faster (on the generalization error) than previous SADMM approaches. We note that for all all experiments, we set the AL parameter $\beta = 1$, as also done in Ouyang et al. (2013).

4.1. GFLasso with smooth loss

Our first experiment follows Ouyang et al. (2013), wherein we consider the *Graph-guided fused lasso* (GFlasso). This problem uses a graph-based regularizer where variables are considered as vertices of the graph and the difference between two adjacent variables is penalized according to the edge weight. This leads to the optimization problem:

$$\min \mathbb{E}[L(x,\xi)] + \lambda\|x\|_1 + \nu\sum\nolimits_{\{i,j\}\in\mathcal{E}} w_{ij}|x_i - x_j|, \tag{13}$$

where $\mathcal{E}$ is the set of edges in the graph, and w_{ij} is the weight for the edge between x_i and x_j. To verify performance of Alg. 3 we consider the following "large-margin" modification to (13):

$$\min \mathbb{E}[L(x,\xi)] + \tfrac{\lambda}{2}\|x\|_2^2 + \nu\|y\|_1, \quad \text{s.t.} \quad Fx - y = 0,$$

where $F_{ij} = w_{ij}$, $F_{ji} = -w_{ij}$ for all edges $\{i,j\} \in \mathcal{E}$, and $L(x,\xi) = \frac{1}{2}(l - x^T s)^2$ for (s,l) feature label pair in the training sample ξ. This formulation is a special case of (1) with $A = F$, $B = -I$ and $b = 0$. The corresponding steps of Alg. 3 assume the form

$$\begin{aligned}
p_k &\leftarrow (1-\gamma_k)x_k + \gamma_k z_k, \quad g_k \leftarrow L'(p_k, \xi_{k+1}) + \gamma_k p_k \\
z_{k+1} &\leftarrow (\tfrac{\gamma_k}{\eta_k}I + \beta\theta_k F^T F)^{-1}[\theta_k F^T(\beta y_k + \lambda_k) \\
&\qquad + \tfrac{\gamma_k}{\eta_k}z_k - g_k] \\
x_{k+1} &\leftarrow (1-\gamma_k)x_k + \gamma_k z_k \\
y_{k+1} &\leftarrow S_{\frac{\nu}{\beta\theta_k}}(Fz_{k+1} - \tfrac{\lambda_k}{\beta}) \\
\lambda_{k+1} &\leftarrow \lambda_k - \beta F z_{k+1} + \beta y_{k+1},
\end{aligned}$$

where $S_\alpha(x)$ denotes the standard soft-thresholding operator As in Ouyang et al. (2013), we obtain F by sparse inverse covariance selection Banerjee et al. (2008) to determine the adjacency matrix of the graph by thresholding the sparsity pattern of the inverse covariance matrix.

We compare the following methods: SADMM (Ouyang et al., 2013), Alg. 3 (called Optimal-SADMM[2]), ordinary stochastic gradient descent (SGD), proximal-SGD (aka FOBOS (Duchi & Singer, 2009)), and online RDA (Xiao, 2010). We compare these methods on a version of the well-known 20newsgroups dataset[3]. This dataset consists of binary occurrence data of 100 words for 16,242 instances, and the samples are labeled into four categories for which one can do classification by one-vs rest scheme multiclass classification. In Fig. 1, we show prediction accuracy on test data (20% of samples) and the training performance as measured by the objective function value.

To implement proximal-SGD and online-RDA, the two methods that require computing the proximity operator $\frac{1}{2}\|x-y\|_2^2 + \lambda\|Fx\|_1$, we implemented an inexact QP-solver that solves the corresponding dual problem[4]:

$$\min \|F^T u - y\|_2^2 \quad \text{s.t.} \quad \|u\|_\infty \leq \beta.$$

If u^* is the optimal dual solution, we can recover the primal solution by setting $x^* = y - f^T u^*$.

Fig. 1 shows that on the training data SADMM and Optimal-SADMM converge faster than the other methods. The classification performance of all methods is similar, except Optimal-SADMM which achieves higher test accuracy (notice #-iterations refers to number of training data points seen). Also, once we made a *single pass* through the training data we terminate all the methods.

[2]We refer to both our SADMM variants as 'Optimal-SADMM'.

[3]Obtained from http://www.cs.nyu.edu/ roweis/data.html

[4]This dual problem is just a box-constrained quadratic program, which we solved using the well-known freely available implementation of the LBFGS-B method.

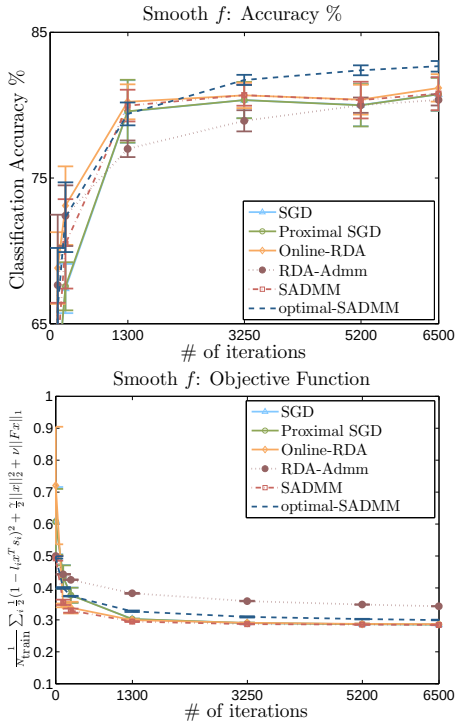

Figure 1. Graph Regularization for 20newsgroups dataset with smooth loss. #iterations refers to #-training data points seen. Upper figure shows the test data accuracy (after seeing about 1300 training points, Optimal-SADMM outperforms the other methods); the lower one shows training data objectives.

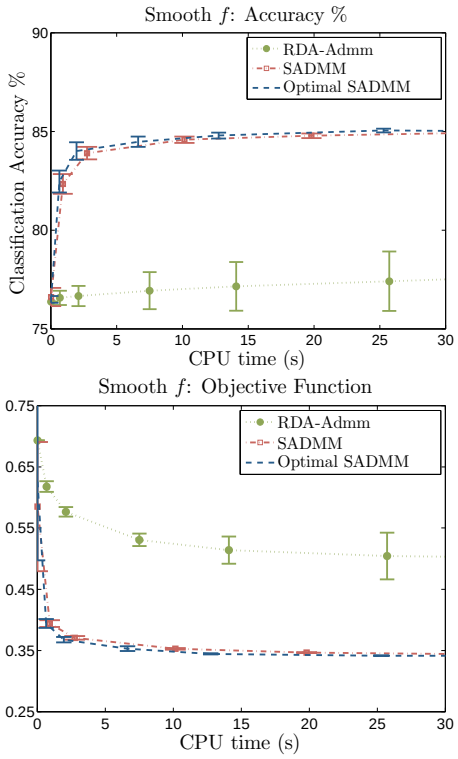

Figure 2. Graph Regularization for adult dataset with smooth loss. For this problem SADMM and Optimal-SADMM perform similarly; both substantially outperform RDA-ADMM.

4.2. Overlapped group lasso

In our second experiment, we present overlapped group lasso results, as explained in (Suzuki, 2013). Here,

$$h(x) = C \sum_{g \in \mathcal{G}} \|x_g\| =: C\|x\|_{\mathcal{G}}, \tag{14}$$

where $\mathcal{G}$ is a set of groups of indices. Feature selection using non-overlapping groups of features by the Lasso can be extended to the group Lasso. But using only non-overlapping groups limits the discoverable structures in practice. One of the solutions to handle this problem is to allow overlapping groups accompanied with the following settings. We divide $\mathcal{G}$ into m non-overlapping subsets $\mathcal{G}_1, \cdots, \mathcal{G}_m$, and let Ax be a concatenation of m-repetitions of x. Thus, $h([x; \cdots; x]) = h(Ax) = C \sum_{i=1}^{m} \|x\|_{\mathcal{G}_i}$. With this formulation, we can easily solve the optimization problem using a proximal operation for each subset; see (Qin & Goldfarb, 2012) for more details.

We applied our optimal SADMM on a dataset for a binary classification task in which $f(x, \xi)$ and $h(y)$ are defined as:

$$f(x, \xi) = 0.1 \sum_{j=1}^{10} L(x, \xi_j), \tag{15}$$
$$h(y) = C \left(\|x^{(1)}\|_1 + \tfrac{1}{\sqrt{123}} \|x^{(2)}\|_{\text{block}} \right),$$

where $L(x, \xi_j) = \log(1 + e^{-l_j s_j^T x})$, $\|x\|_{\text{block}} = \sum_i \|X_{i,\cdot}\|_2 + \sum_j \|X_{\cdot,j}\|_2$ where X denotes a reshaped version of x as a square matrix; observe that $L(x, \xi)$ is a logistic loss and $h(y)$ is the overlapping group lasso regularizer.

We used the dataset 'adult'[5] which contains 123 dimensional feature vectors. Following Suzuki (2013) we also augmented the feature space by taking products of features resulting in $(123+123^2)$ dimensions. Vector $x^{(1)}$ in (15) is related to the 123-first elements of x, while $x^{(2)}$ represents the rest of x. Hyperparameter C is set to 0.01. Moreover, we used mini-batches of size 10 for each iteration.

We present plots on the test data classification accuracy as well as the training data objective functions. We compare ordinary SADMM, Optimal-SADMM, and RDA-ADMM. On this task, the difference between SADMM and Optimal-SADMM is not remarkable, but both substantially outperform RDA-SADMM, as seen from Fig. 2.

[5] Obtained from the LIBSVM datasets webpage.

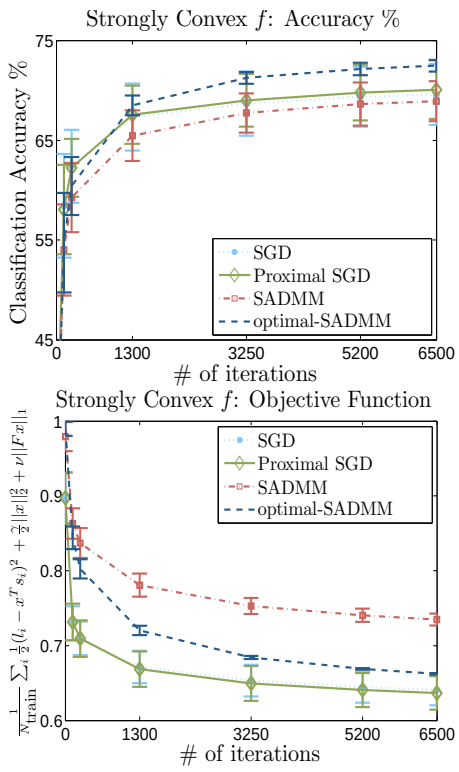

Figure 3. Graph Regularization for 20newsgroups dataset with strongly convex loss.

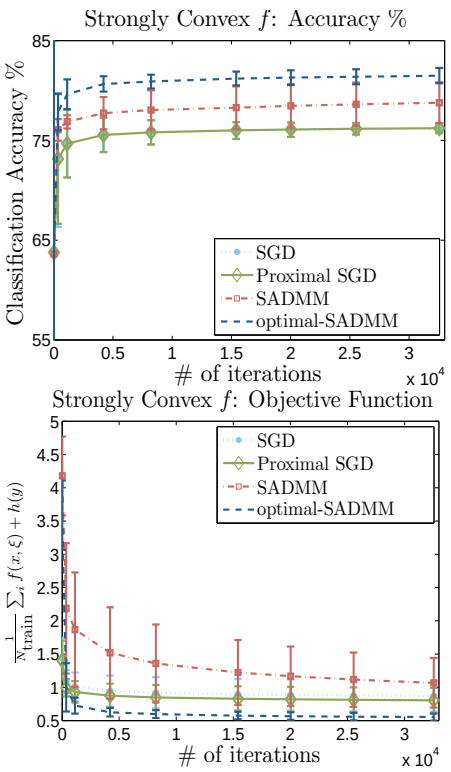

Figure 4. Group Regularization for adult dataset with strongly convex loss.

4.3. Strongly Convex Loss Functions

Now we show a more detailed comparisons between SGD, Proximal SGD, strongly convex SADMM from (Ouyang et al., 2013), and our Optimal-SADMM version of Alg. 2. In this case, we compare the mentioned algorithms on a nonsmooth but strongly convex GFLasso and Group Lasso problems, which use the *hinge loss* for $L(x,\xi)$ in (4.1) and (15), respectively ($L(x,\xi) = \max\{0, 1 - ls^T x\}$). Other terms remain the same. The closed form updates are similar to those in the previous section, except that x_k is used instead of z_k; also $\bar{x}_k$ is computed as per (10). Step size equal to $\frac{1}{k}$ is used for the SGD and proximal SGD methods. The classification accuracy on hold out test data and the objective function value on the training data are plotted in Figures 3 and 4. The plots indicate that the proposed algorithm significantly outperforms other methods, both in terms of training objective value and classification accuracy except for the objective function value on the training data in Fig. 3 in which our optimal-SADMM training performance dips a bit in comparison with SGD and proximal-SGD methods.

4.4. Experiment on Synthetic Data

Here, we intend to explore the behavior of smooth SADMM variants as the number of features in the data varies. For this experiment, we generated a synthetic dataset, and performed classification using the smooth formulation (4.1). The data is generated as follows: we sample a matrix of m samples with n features following a multivariate Gaussian with a random covariance matrix. The true weight vector x^* is chosen from an i.i.d. standard normal; the labels are defined according to $l_m = \text{sgn}(s_m^T x^* + \epsilon_m)$ in which ϵ_m is a mean zero Gaussian noise with standard deviation of 2.

Fig. 5 reports percentage improvement of Optimal-SADMM over SADMM in terms of classification accuracy as a function of number of features. This experiment suggests that our optimal SADMM may use features more efficiently, especially with increasing feature dimension. Exploring this phenomenon more closely is ongoing work.

5. Conclusions and future work

We presented two new accelerated versions the stochastic ADMM (Ouyang et al., 2013). In particular, we presented a variant that attains the theoretically optimal $O(1/k)$ convergence rate for strongly convex stochastic problems. When the stochastic part is smooth, we showed another SADMM algorithm that has an optimal $O(1/k^2)$ dependence on the smooth part.

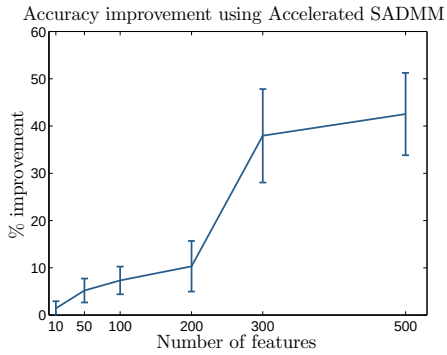

Figure 5. Feature efficiency of Optimal-SADMM vs SADMM (smooth loss).

Our initial experiments reveal that our accelerated variants do exhibit notable performance gains over their non-accelerated counterparts (see §4, also Fig. 3)—though, obviously as also seen from the experiments in (Lacoste-Julien et al., 2012; Shamir & Zhang, 2013), gains in stochastic settings are less dramatic than in the deterministic case (Beck & Teboulle, 2009). This is not surprising, since accelerated methods are more sensitive to stochastic noise than their deterministic counterparts (Devolder et al., 2011).

There will be more results and details available in the longer arXiv version of the paper. We mention below a list of extensions to the present paper:

- Transfer the $O(\log k/k)$ convergence rate of the last iterate as done for SGD by Shamir & Zhang (2013) to the SADMM setting.
- Obtaining high-probability bounds under light-tailed assumptions on the stochastic error.
- Incorporate the impact of sampling multiple stochastic gradients to decrease the variance in the gradient estimates.
- Derive a mirror-descent version.
- *Improve rate dependence of the augmented Lagrangian part to $O(1/k^2)$ for smooth problems.*

Most, except the last, of these extensions are easy (though tedious) and follow by invoking standard techniques from the analysis of stochastic convex optimization. We hope to address these in a longer version of the present paper.

We conclude by highlighting that our empirical results are encouraging and suggest that for strongly convex or smooth losses, our accelerated SADMM variants outperform the other known SADMM methods.

References

Banerjee, O., Ghaoui, L. E., and d'Aspremont, A. Model selection through sparse maximum likelihood estimation for multivarite Gaussian or binary data. *JMLR*, 9:485–516, 2008.

Beck, A. and Teboulle, M. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM J. Imaging Sciences*, 2(1):183–202, 2009.

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.

Chambolle, A. and Pock, T. A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging. *J. Math. Imaging Vis.*, 40(1):120–145, 2011.

Chan, R. H., Tao, M., and Yuan, X. Constrained total variation models and fast algorithms based on alternating direction method of multipliers. *SIAM J. Imaging Sci.*, 6 (1), 2013.

Chen, Xi, Lin, Qihang, and Pena, Javier. Optimal regularized dual averaging methods for stochastic optimization. In *Advances in Neural Information Processing Systems (NIPS-12)*, pp. 404–412, 2012.

Devolder, O., Glineur, F., and Nesterov, Yu. First-order methods of smooth convex optimization with inexact oracle. Technical Report 2011/17, UCL, 2011.

Douglas, J. and Rachford, H. H. On the numerical solution of the heat conduction problem in 2 and 3 space variables. *Tran. Amer. Math. Soc.*, 82:421–439, 1956.

Duchi, J. and Singer, Y. Online and Batch Learning using Forward-Backward Splitting. *J. Mach. Learning Res. (JMLR)*, 2009.

Eckstein, J. and Bertsekas, D. P. On the Douglas-Rachoford splitting method and the proximal point algorithm for maximal monotone operators. *Math. Prog.*, 55(3):292–318, 1992.

Ghadimi, S. and Lan, G. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, i: a generic algorithmic framework. *SIAM J. Optimization*, 22:1469–1492, 2012.

Goldfarb, D., Ma, S., and Scheinberg, K. Fast alternating linearization methods for minimizing the sum of two convex functions. *Math. Prog. Ser. A*, 2012.

He, B. and Yuan, X. On non-ergodic convergence rate of Douglas-Rachford alternating direction method of multipliers. *Unpublished*, 2012.

Kraska, T., Talwalkar, A., J.Duchi, Griffith, R., Franklin, M., and Jordan, M.I. MLbase:A Distributed Machine Learning System. In *Conf. Innovative Data Systems Research*, 2013.

Lacoste-Julien, S., Schmidt, M., and Bach, F. A simpler approach to obtaining an $O(1/t)$ convergence rate for projected subgradient descent. *arXiv:1212.2002*, 2012.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Nesterov, Y. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer, 2004.

Nesterov, Yu. Gradient methods for minimizing composite objective function. Technical Report 2007/76, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), 2007.

Ouyang, Hua, He, Niao, Tran, Long, and Gray, Alexander G. Stochastic alternating direction method of multipliers. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pp. 80–88, 2013.

Parikh, N. and Boyd, S. *Proximal Algorithms*, volume 1. NOW, 2013.

Qin, Z. and Goldfarb, D. Structured sparsity via alternating direction methods. *J. Machine Learning Research*, 13: 1435–1468, 2012.

Rockafellar, R. T. Monotone Operators and the Proximal Point Algorithm. *SIAM J. Control Optim.*, 14(5):877–989, 1976.

Shamir, Ohad and Zhang, Tong. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pp. 71–79, 2013.

Srebro, N. and Tewari, A. Stochastic Optimization for Machine Learning. ICML 2010 Tutorial, 2010.

Suzuki, Taiji. Dual averaging and proximal gradient descent for online alternating direction multiplier method. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pp. 392–400, 2013.

Tseng, P. On accelerated proximal gradient methods for convex-concave optimization. *Unpublished*, 2008.

Xiao, L. Dual averaging methods for regularized stochastic learning and online optimization. *JMLR*, 11:2543–2596, 2010.