Proceedings of the 2016 Winter Simulation Conference
T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, eds.

SI-ADMM: A STOCHASTIC INEXACT ADMM FRAMEWORK FOR RESOLVING STRUCTURED STOCHASTIC CONVEX PROGRAMS

Yue Xie
Uday V. Shanbhag

Department of Industrial and Manufacturing Engineering
Pennsylvania State University, PA 16803, USA.

ABSTRACT

We consider the resolution of the structured stochastic convex program: min $\mathbb{E}[\tilde{f}(x,\xi)]+\mathbb{E}[\tilde{g}(y,\xi)]$ such that $Ax+By=b$. To exploit problem structure and allow for developing distributed schemes, we propose an inexact stochastic generalization in which the subproblems are solved inexactly via stochastic approximation schemes. Based on this framework, we prove the following: (i) when the inexactness sequence satisfies suitable summability properties, the proposed stochastic inexact ADMM (**SI-ADMM**) scheme produces a sequence that converges to the unique solution almost surely; (ii) if the inexactness is driven to zero at a polynomial (geometric) rate, the sequence converges to the unique solution in a mean-squared sense at a prescribed polynomial (geometric) rate.

1 INTRODUCTION

In the context of large datasets, it has become increasingly important to process the data in a parallel and decentralized fashion. Therefore, distributed optimization is often considered an option, and a simple yet powerful algorithm of this kind is the alternating direction method of multipliers (ADMM). ADMM schemes may be traced to mid-70s to the work by Glowinski and Marroco (1975) and subsequently Gabay and Mercier (1976). It has grown immensely in popularity and has been utilized for resolving a host of structured machine learning and image processing problems such as image recovery (Afonso, Bioucas-Dias, and Figueiredo 2010), robust PCA (Lin, Chen, and Ma 2010), low-rank representation (Lin, Liu, and Su 2011); see Boyd, Parikh, Chu, Peleato, and Eckstein (2011) for a comprehensive review. Typically, ADMM is applied towards structured deterministic convex optimization problems of the form:

$$\begin{aligned} \min_{x,y} \quad & f(x)+g(y) \\ \text{subject to} \quad & Ax+By=b, \end{aligned} \tag{1}$$

We consider a stochastic generalization leading to a structured stochastic convex program:

$$\begin{aligned} \min_{x,y} \quad & \mathbb{E}[\tilde{f}(x,\xi)]+\mathbb{E}[\tilde{g}(y,\xi)] \\ \text{subject to} \quad & Ax+By=b, \end{aligned} \tag{SOpt}$$

where $\xi:\Omega\to\mathbb{R}^d$, $\tilde{f}:\mathbb{R}^n\times\mathbb{R}^d\to\mathbb{R}$, $\tilde{g}:\mathbb{R}^m\times\mathbb{R}^d\to\mathbb{R}$, $A\in\mathbb{R}^{p\times n}$, $B\in\mathbb{R}^{p\times m}$, $b\in\mathbb{R}^p$, and $(\Omega,\mathscr{F},\mathbb{P})$ denotes the probability space. Furthermore, we assume that $\tilde{f}(.,\xi)$ and $\tilde{g}(.,\xi)$ are convex in $(.)$ for every $\xi\in\Xi$.

A rather popular approach for the solution of (SOpt) is by utilizing Monte-Carlo sampling schemes, such as sample-average approximation (see Shapiro, Dentcheva, and Ruszczyński (2009)) or stochastic approximation schemes (Robbins and Monro 1951). In fact, over the last decade, there has been significant study of stochastic approximation schemes, particularly from the standpoint of the tuning of steplengths (Nemirovski, Juditsky, Lan, and Shapiro (2009); Yousefian, Nedić, and Shanbhag (2012)), the resolution

978-1-5090-4486-3/16/$31.00 ©2016 IEEE

of stochastic variational inequality problems (Koshal, Nedić, and Shanbhag (2013)), amongst others. We also note that recent work by Pasupathy et al. (2014) considers sequential sampling concerns, an issue that assumes relevance in this paper. The present work considers addressing (SOpt) by adapting the existing ADMM architecture to the stochastic setting, and is characterized by the following benefits: (1) ADMM schemes display strong theoretical properties and computational performance in the resolution of structured constrained optimization problems (see Boyd et al. (2011)); (2) this avenue allows for problem structure to be exploited via distributed computation.

It is noteworthy that alternative stochastic generalizations of ADMM were studied by Wang and Banerjee (2013); Ouyang, He, Tran, and Gray (2013). They considered the following problem:

$$\min_{x\in\mathscr{X},y\in\mathscr{Y}} \left\{\mathbb{E}_{\xi}[\theta_1(x,\xi)]+\theta_2(y) : Ax+By=b\right\}.$$

This problem can be regarded as minimizing regularized expected risk function where $\mathbb{E}[\theta_1(x,\xi)]$ denotes the expected loss function while $\theta_2(y)$ represents the regularizer. In order to solve this problem, Wang and Banerjee (2013) considered a scheme, referred to as online ADMM (or OADMM), in which $\mathbb{E}_{\xi}[\theta_1(x,\xi)]$ is substituted by the sampled function $\theta_1(x,\xi_k)$ to compute the x update at the kth iterate. Ouyang, He, Tran, and Gray (2013) extended this scheme to the inexact regime and showed that the convergence rates in terms of sub-optimality and infeasibility are $\mathscr{O}(1/\sqrt{T})$ and $\mathscr{O}(1/T)$ for convex and strongly convex functions, respectively, where T denotes the iteration index.

The problem (SOpt) considered here can be viewed as a generalization of the problem considered by Wang and Banerjee (2013) in the sense that θ_1 and θ_2 are both expected-value functions. To solve this (SOpt), we extend a generalized ADMM scheme proposed by Deng and Yin (2012). Their scheme was shown to generate a sequence of iterates that converge at a linear rate on suitable strong convexity assumptions. Notably, their scheme can be reduced to a prox-linear ADMM (cf. Lin, Liu, and Su (2011)) as well as a gradient-descent ADMM (cf. Ouyang, He, Tran, and Gray (2013)). We modify their scheme to an implementable one for (SOpt) that requires computing inexact solutions to the x and y update in each iteration. In fact, since each subproblem is a stochastic optimization problem, each update requires a finite (but) increasing number of gradient steps (assuming the use of stochastic approximation). Based on such a framework, we make the following contributions:

(i) **a.s. convergence:** It is shown that under suitable summability assumptions on the inexactness sequence η_k, the sequence of iterates converges almost surely to the unique solution of the problem.

(ii) **Convergence of mean error at polynomial rate:** It is demonstrated that when the inexactness sequence is driven to zero at a polynomial rate given by $\sqrt{\eta_k}=k^{-\alpha}$, the mean-squared error diminishes to zero at a rate $K(\alpha)/k^{\alpha}$ where $K(\alpha)$ is a specified constant.

(iii) **Convergence of mean error at geometric rate:** When the inexactness sequence is driven to zero at a geometric rate, the mean-squared error diminishes to zero at a prescribed geometric rate.

The remainder of the paper is organized into three sections. In Section 2, we introduce the stochastic ADMM framework. In section 3, the asymptotic convergence is analyzed and the associated rate statements are obtained. We conclude with a short summary in Section 4.

2 A STOCHASTIC ADMM SCHEME

In this section, we begin by providing an introduction in Section 2.1 to some recent linear convergence statements for a deterministic ADMM scheme. Unfortunately, direct application of this scheme to the stochastic regime requires exact resolution of stochastic optimization subproblems. To obviate this challenge, in Section 2.2, we present an implementable generalization allows for accommodating this stochastic generalization, referred to as stochastic inexact ADMM (SI-ADMM).

2.1 A Generalized ADMM Scheme

A generalized ADMM scheme (Algorithm 1) that can resolve (SOpt) was suggested by Deng and Yin (2012) where the augmented Lagrangian function $\mathscr{L}_A(x,y,\lambda)$ is defined as

$$\mathscr{L}_A(x,y,\lambda) \triangleq f(x)+g(y)-\lambda^T(Ax+By-b)+\tfrac{1}{2}\rho\|Ax+By-b\|^2, \tag{2}$$

where $f(x) \triangleq \mathbb{E}[\tilde{f}(x,\xi)]$ and $g(y) \triangleq \mathbb{E}[\tilde{g}(y,\xi)]$, both of which are assumed to be convex throughout this paper. Moreover, we make the following two assumptions, both of which are necessary for **global linear** convergence property of the sequence of iterates produced by Algorithm 1. Of these, the first pertains to the existence of a KKT point to the original optimization problem:

Algorithm 1 g-ADMM: Generalized ADMM scheme

```
(0)   Choose matrices $P,Q$ and let $k=0$;
(1)   Given $x_0,y_0,\lambda_0,\rho>0$, $\gamma>0$;
(2)   Let $x_{k+1},y_{k+1},\lambda_{k+1}$ be given by the following:
          $y_{k+1} := \operatorname*{argmin}_y \left(\mathscr{L}_A(x_k,y,\lambda_k)+\tfrac{1}{2}(y-y_k)^TQ(y-y_k)\right)$        $(y^{\text{exact}})$
          $x_{k+1} := \operatorname*{argmin}_x \left(\mathscr{L}_A(x,y_{k+1},\lambda_k)+\tfrac{1}{2}(x-x_k)^TP(x-x_k)\right)$    $(x^{\text{exact}})$
          $\lambda_{k+1} := \lambda_k-\gamma\rho(Ax_{k+1}+By_{k+1}-b).$                        $(\lambda^{\text{exact}})$
(3)   $k:=k+1$; If $k<K$, return to (1); else STOP.
```

Assumption 1 There exists a KKT point $u^*=(x^*,y^*,\lambda^*)$ to problem (SOpt); i.e., $\{x^*,y^*,\lambda^*\}$ satisfies the KKT conditions:

$$\begin{aligned} A^T\lambda^* &\in \partial f(x^*),\\ B^T\lambda^* &\in \partial g(y^*),\\ Ax^*+By^*-b &= 0. \end{aligned}$$

The second assumption imposes suitable convexity and Lipschitzian assumptions about f and g as well as suitable rank assumptions about P and Q:

Assumption 2 ρ is a positive scalar. Q and $\hat{P}=P+\rho A^TA$ are symmetric and positive semidefinite matrices. Additionally, either (a) or (b) holds:
(a): The function $f(x)$ is strongly convex in x with a Lipschitz continuous gradient. Additionally, A has full row rank, B has full column rank, and Q is a positive definite matrix.
(b): The functions f and g are strongly convex in x and y respectively, $\nabla_x f$ is Lipschitz continuous in x, and A has full row rank.

Next we provide the main convergence statement presented by Deng and Yin (2012).

Theorem 1 (Theorem 3.4, 3.5 (Deng and Yin 2012)) *Suppose Assumptions 1 and 2 hold. In addition, suppose γ satisfies one of following: (i) $P \neq 0$ and $(2-\gamma)P \succ (\gamma-1)\rho A^TA$; and (ii) $P=0$ and $\gamma=1$. If sequence $\{u_k\}$ generated by Algorithm 1 is bounded, then $\|u_k-u^*\|_G \to 0$ as $k\to\infty$. Furthermore, there exists a constant $\delta>0$ such that*

$$\|u_{k+1}-u^*\|_G^2 \le \frac{1}{1+\delta}\|u_k-u^*\|_G^2, \tag{3}$$

where $u_k \triangleq (x_k,y_k,\lambda_k)$, and $u^* \triangleq (x^*,y^*,\lambda^*)$ denotes a KKT point of (SOpt), *$\|z\|_G \triangleq \sqrt{z^TGz}$,*

$$G_0 \triangleq \begin{pmatrix} I_n & & \\ & I_m & \\ & & \gamma I_p \end{pmatrix}, G_1 \triangleq \begin{pmatrix} P+\rho A^TA & & \\ & Q & \\ & & 1/\rho I_p \end{pmatrix}, \text{ and } G \triangleq G_0^{-1}G_1 = \begin{pmatrix} P+\rho A^TA & & \\ & Q & \\ & & \frac{1}{\rho\gamma}I_p \end{pmatrix}.$$

Note that if Q and $\hat{P}$ are chosen to be positive definite, then $\|\bullet\|_G$ reduces to a norm, rather than a semi-norm. A map Γ^* is defined as follows:

$$\Gamma^*(u) = \begin{pmatrix} y^{\text{exact}}(x,\lambda) \\ x^{\text{exact}}(\tilde{y}^{\text{exact}},\lambda) \\ \lambda^{\text{exact}}(\tilde{x}^{\text{exact}},\tilde{y}^{\text{exact}}) \end{pmatrix}, \tag{4}$$

where $u \triangleq (x,y,\lambda)$, $\tilde{y}^{\text{exact}} \triangleq y^{\text{exact}}(x,\lambda)$ and $\tilde{x}^{\text{exact}} \triangleq x^{\text{exact}}(\tilde{y}^{\text{exact}},\lambda)$. Furthermore, on Assumption 2, it can be seen that y update and x update in Algorithm 1 have unique solutions. Thus, for each given input u_k, each iteration of algorithm 1 provides a unique output, denoted as u^*_{k+1}, i.e.

$$u^*_{k+1} := \Gamma^*(u_k).$$

Thus, the map $\Gamma^*(\bullet)$ is a well-defined single-valued map and u^* is a fixed-point of this map or $u^* = \Gamma^*(u^*)$. Unfortunately, when the expectation $\mathbb{E}[\bullet]$ is over a general measure space, the ADMM scheme is not practically implementable since the x and y updates require exact solutions of stochastic optimization problems. This motivates an implementable stochastic generalization of this scheme.

2.2 Stochastic ADMM Scheme

Since x^{exact} and y^{exact} in Algorithm 1 necessitate exact solutions impossible to obtain in stochastic regimes, we propose a *stochastic inexact* extension of the generalized ADMM scheme. In this framework, the sequence of iterates generated by the scheme are random variables and require taking a finite (but) increasing number of (stochastic) gradient steps. To formally define this stochastic ADMM scheme, we denote the history of the process as follows. First, $\mathscr{F}_0 \triangleq \{x_0,y_0,\lambda_0\}$. Then at the $(k+1)$th epoch, $\mathscr{F}^y_{k+1} \triangleq \mathscr{F}_k \cup \{\xi^y_{k,1},\ldots,\xi^y_{k,N^y_k}\}$ where the union is taken with the previous history $\mathscr{F}_k$ and the samples generated for the $y-$update. Similarly, $\mathscr{F}_{k+1} = \mathscr{F}^y_{k+1} \cup \{\xi^x_{k,1},\ldots,\xi^x_{k,N^x_k}\}$. Note that N^x_k and N^y_k denote the number of samples generated within the x and y updates to ensure meeting a suitable error criterion. The stochastic ADMM scheme is then defined in algorithm 2 and referred to as stochastic inexact ADMM (**SI-ADMM**). Note that the sequence of iterates $u_k = (x_k,y_k,\lambda_k)$ are inexact in that their mean-squared error with respect to their exact counterpart is bounded by η_k. Such an inexact solution is obtainable by using standard stochastic approximation schemes. In fact, we may define a map $\Gamma_k(u)$, akin to the map $\Gamma^*(u)$ specified in (4).

$$\Gamma_k(u) \triangleq \begin{pmatrix} y_k^{\text{s}-\text{inex}}(x,\lambda) \\ x_k^{\text{s}-\text{inex}}(y_k^{\text{s}-\text{inex}},\lambda) \\ \lambda_k^{\text{s}-\text{inex}}(x_k^{\text{s}-\text{inex}},y_k^{\text{s}-\text{inex}}) \end{pmatrix}, \tag{5}$$

where $y_k^{\text{s}-\text{inex}}(x,\lambda)$ is any solution to the $y_k^{\text{s}-\text{inex}}$ update while $x_k^{\text{s}-\text{inex}}(y_k^{\text{s}-\text{inex}},\lambda)$ is any solution to the $x_k^{\text{s}-\text{inex}}$ update. Consequently, $\Gamma_k(u)$ is a well-defined single-valued map. To ensure convergence of this sequence, we make the following assumption about f and g.

Assumption 3 The functions f and g are both differentiable and strongly convex in their respective arguments with constants μ and σ, respectively. Furthermore, the matrix $[A\ B]$ has full row rank.

Algorithm 2 SI-ADMM: A stochastic inexact ADMM scheme

```
(0)  Choose $Q$ and $P$, $k=0$, choose the inexactness sequence $\eta_i$, for $i \in \mathbb{N}$;
(1)  Given $x_0, y_0, \lambda_0, \rho > 0$, $\gamma > 0$;
(2)  Let $x_{k+1}, y_{k+1}, \lambda_{k+1}$ be given by the following:
         $\mathbb{E}[\|y_{k+1} - y^*_{k+1}\|^2 \mid \mathscr{F}_k] \leq \eta_{k+1}$                                    $(y^{\mathrm{s-inex}}_{k+1})$
             where $y^*_{k+1} := \operatorname*{argmin}_y \big(\mathscr{L}_A(x_k, y, \lambda_k) + \frac{1}{2}(y - y_k)^T Q (y - y_k)\big)$
         $\mathbb{E}[\|x_{k+1} - \tilde{x}^*_{k+1}\|^2 \mid \mathscr{F}^y_{k+1}] \leq \eta_{k+1}$                                $(x^{\mathrm{s-inex}}_{k+1})$
             where $\tilde{x}^*_{k+1} := \operatorname*{argmin}_x \big(\mathscr{L}_A(x, y_{k+1}, \lambda_k) + \frac{1}{2}(x - x_k)^T P (x - x_k)\big)$
             $\lambda_{k+1} := \lambda_k - \gamma\rho(Ax_{k+1} + By_{k+1} - b).$                               $(\lambda^{\mathrm{s-inex}}_{k+1})$
(3)  $k := k+1$; If $k < K$, return to (1); else STOP.
```

Remark: Assumption 3 is sufficient to claim that there exist only one triple (x^*, y^*, λ^*) that satisfies the KKT conditions:

$$\begin{aligned} A^T \lambda^* &= \nabla f(x^*) \\ B^T \lambda^* &= \nabla g(y^*) \\ Ax^* + By^* - b &= 0. \end{aligned}$$

In fact, since f and g are strongly convex, problem (SOpt) has a unique primal optimal solution. Furthermore, the KKT conditions are necessary and sufficient. This indicates that KKT conditions admit a unique pair (x^*, y^*). Since $[A\ B]$ has linearly independent rows, λ^* is uniquely determined by (x^*, y^*). As a matter of fact, Assumptions 2 and 3 share a significant overlap. If g is differentiable, then Assumption 2 implies Assumption 3.

2.3 Inexact Solutions of Subproblems

The x and y updates in Algorithm 2 requires computing inexact solutions to stochastic problems. Consider the problem in the $y-$update of $k+1$th iteration (discussion for $x-$update is similar thus omitted):

$$\min_y \quad h(y; x_k, y_k, \lambda_k) \triangleq \left[\mathbb{E}[\tilde{g}(y;\xi)] - \lambda_k^T(Ax_k + By - b) + \tfrac{1}{2}\rho\|Ax_k + By - b\|^2 + \tfrac{1}{2}(y - y_k)Q(y - y_k)\right]. \tag{6}$$

Recall that $g(\cdot) = \mathbb{E}[\tilde{g}(\cdot, \xi)]$ is a strongly convex objective, implying that $h(y; x_k, y_k, \lambda_k)$ is strongly convex in y for any given x_k, y_k and λ_k, with a parameter σ that is not related to x_k, y_k or λ_k. Suppose that $\nabla_y g(.)$ is also Lipschitz continuous in y with constant L_y. Then it indicates that $\nabla_y h(x; x_k, y_k)$ is Lipschitz continuous in y. This can be seen to hold by noting the following:

$$\nabla_y h(y; x_k, y_k, \lambda_k) = \nabla_y g(y) - B^T \lambda_k + \rho B^T(Ax_k + By - b) + Q(y - y_k).$$

This implies that

$$\begin{aligned} \|\nabla_y h(y_1; x_k, y_k, \lambda_k) - \nabla_y h(y_2; x_k, y_k, \lambda_k)\| &\leq \|\nabla_y g(y_1) - \nabla_y g(y_2)\| + (\rho\|B\|^2 + \|Q\|)\|y_1 - y_2\| \\ &\leq (L_y + \rho\|B\|^2 + \|Q\|)\|y_1 - y_2\|, \end{aligned}$$

for any y_1, y_2. It follows that the $\nabla_y h(y; x_k, y_k, \lambda_k)$ is Lipschitz continuous in y for any x_k, y_k and λ_k with a constant $L = L_y + \rho\|B\|^2 + \|Q\|$. Based on the Lipschitz continuity of the expected gradient and the strong convexity of the objective, a recursive relationship (7) holds when a stochastic approximation scheme like the following is applied to the resolution of (6): given a z_1,

$$z_{t+1} = z_t - \gamma_t \nabla_z \tilde{h}(z_t, \xi_t; x_k, y_k, \lambda_k), \quad t \geq 1, \tag{SA}$$

where $\nabla_z \tilde{h}(z_t, \xi_t; x_k, y_k, \lambda_k)$ refers to the sampled gradient associated with the tth sample. Let $w_t = \nabla_z \tilde{h}(z_t, \xi_t; x_k, y_k, \lambda_k) - \mathbb{E}[\nabla_z \tilde{h}(z_t, \xi; x_k, y_k, \lambda_k)]$, and assume that $\mathbb{E}[w_t \mid \mathscr{F}_{t-1}] = 0$ and $\mathbb{E}[\|w_t\|^2 \mid \mathscr{F}_{t-1}] \leq \nu^2, \quad \forall z_t \in \mathbb{R}^m$. Next, we provide a rate statement from stochastic approximation schemes for strongly convex stochastic optimization.

Lemma 1 Consider the application of a stochastic approximation scheme given by (SA) on the stochastic optimization problem (6): $\min_{z \in \mathbb{R}^n} h(z; x_k, y_k, \lambda_k)$, where $h(z; x_k, y_k, \lambda_k) \triangleq \mathbb{E}[\tilde{h}(z, \xi; x_k, y_k, \lambda_k)]$. Suppose $h(z; x_k, y_k, \lambda_k)$ is a strongly convex function with convexity constant $\sigma > 0$ and has Lipschitz continuous gradients with constant L. Suppose $\gamma_t = \theta/t$ and $\theta > 1/2\sigma$, z^* is the optimal solution of (6). Then the following holds for any $t \geq 1$:

$$\mathbb{E}[\|z_{t+1} - z^*\|^2] \leq (1 - 2\sigma\theta/t)\mathbb{E}[\|z_t - z^*\|^2] + (L + \nu^2)\theta^2/t^2, \tag{7}$$

Furthermore

$$\mathbb{E}[\|z_t - z^*\|^2] \leq \frac{\max\left\{\frac{(L+\nu^2)\theta^2}{2\sigma\theta - 1}, \mathbb{E}[\|z_1 - z^*\|^2]\right\}}{t}. \tag{8}$$

Proof sketch: Expression 7 can be derived from (Yousefian, Nedić, and Shanbhag 2012) (See expression (7)). Then by invoking the inductive argument in (Shapiro, Dentcheva, and Ruszczyński 2009) (See expression (5.291)), the bound (8) follows. ∎

Remark:

1. Based on Lemma 1, we may derive the number of gradient steps T_k^y and T_k^x at the kth step required to achieve a mean-squared error of η_k for first two subproblems at the kth iteration. Note that parameters L, ν, θ, σ do not depend on iterates.
2. Any solution produced by this stochastic approximation scheme uses samples $\xi_{k,1}, \ldots, \xi_{k,N_k}$ and is a random variable. In fact, even the function $h(z; x_k, y_k, \lambda_k)$ is a random function since x_k, y_k and λ_k are random variables. This provides a basis for why the update rule uses a conditional expectation rather than an unconditional expectation.
3. In effect, the scheme requires a finite but increasing number of gradient steps at each epoch.
4. We assume that we have an M such that $\mathbb{E}[\|x_0 - x^*\|^2] \leq M, \mathbb{E}[\|y_0 - y^*\|^2] \leq M$ and $\mathbb{E}[\|\lambda_0 - \lambda^*\|^2] \leq M$. This bound is then employed to construct similar bounds $\mathbb{E}[\|z_1 - z^*\|^2]$ at the kth iterate which will then be used to derive T_k^y and T_k^x.

3 CONVERGENCE ANALYSIS

In this section, we examine the convergence of the (random) sequence $\{u_k\}$ generated by (**SI-ADMM**), both from an asymptotic and a rate standpoint, based on the choice of the inexactness sequence $\{\eta_k\}$. In Section 3.1, it is proven that the sequence $\{u_k\}$ converges almost surely to the unique solution for suitable choices of $\{\eta_k\}$. Then in Sections 3.2 and 3.3, asymptotics and rate statements are derived when η_k decays polynomially and geometrically.

3.1 Almost Sure Convergence of $\{u_k\}$

The sequence $\{u_k\}$ created by Algorithm 2 has the following relationship:

$$u_{k+1} := \Gamma_{k+1}(u_k), \tag{9}$$

where $\Gamma_k(\bullet)$ is defined in (5). Prior to proceeding, we state the supermartingale convergence lemma from (Polyak 1987) for the proof of almost sure convergence of SI-ADMM.

Lemma 2 (Lemma 10, pg. 49 (Polyak 1987)) *Let $\{v_k\}$ be a sequence of nonnegative random variables, where $\mathbb{E}[v_0] < \infty$, and let $\{u_k\}$ and $\{\mu_k\}$ be deterministic scalar sequences such that:*

$$\mathbb{E}[v_{k+1} \mid v_0, \ldots, v_k] \leq (1-u_k)v_k + \mu_k, \quad a.s. \quad \forall k \geq 0,$$
$$0 \leq u_k \leq 1, \quad \mu_k \geq 0, \quad \forall k \geq 0, \quad \sum_{k=0}^{\infty} u_k = \infty, \quad \sum_{k=0}^{\infty} \mu_k < \infty, \quad \lim_{k\to\infty} \frac{\mu_k}{u_k} = 0.$$

Then $v_k \to 0$ almost surely as $k \to \infty$.

Based on the above lemma, we would obtain almost sure convergence if the error bound η_k in each iteration is properly chosen.

Theorem 2 (a.s. convergence) Consider Algorithm 2. Suppose assumptions in theorem 1 and Assumption 3 hold and $P \succ 0, Q \succ 0$, $\sum_{k=1}^{\infty} \sqrt{\eta_k} < \infty$. Then $\|u_k - u^*\|_G \to 0$ almost surely as $k \to \infty$.

Proof. We begin by developing a bound on the (conditional) expectation of the error between u_{k+1} and its exact counterpart or $\mathbb{E}[\|\Gamma_{k+1}(u_k) - \Gamma^*(u_k)\|_G \mid \mathscr{F}_k]$:

$$\begin{aligned}\mathbb{E}[\|\Gamma_{k+1}(u_k) - \Gamma^*(u_k)\|_G \mid \mathscr{F}_k] &= \mathbb{E}\left[\sqrt{(u_{k+1} - u^*_{k+1})^T G (u_{k+1} - u^*_{k+1})}\middle| \mathscr{F}_k\right] \quad \text{(by the definition of } G) \\ &= \mathbb{E}\left[\sqrt{\|x_{k+1} - x^*_{k+1}\|^2_{\hat{P}} + \|y_{k+1} - y^*_{k+1}\|^2_Q + \frac{1}{\rho\gamma}\|\lambda_{k+1} - \lambda^*_{k+1}\|^2}\middle| \mathscr{F}_k\right]. \quad (10)\end{aligned}$$

By invoking Jensen's inequality, the concavity of $\sqrt{\bullet}$, and the definition of $\hat{P}$ and Q, (10) may be bounded as follows:

$$\begin{aligned}&\mathbb{E}\left[\sqrt{\|x_{k+1} - x^*_{k+1}\|^2_{\hat{P}} + \|y_{k+1} - y^*_{k+1}\|^2_Q + \frac{1}{\rho\gamma}\|\lambda_{k+1} - \lambda^*_{k+1}\|^2}\middle| \mathscr{F}_k\right] \\ \leq &\sqrt{\mathbb{E}\left[\|x_{k+1} - x^*_{k+1}\|^2_{\hat{P}} + \|y_{k+1} - y^*_{k+1}\|^2_Q + \frac{1}{\rho\gamma}\|\lambda_{k+1} - \lambda^*_{k+1}\|^2 | \mathscr{F}_k\right]} \\ \leq &\sqrt{\Lambda_{\hat{P}}\mathbb{E}\left[\|x_{k+1} - x^*_{k+1}\|^2 | \mathscr{F}_k\right] + \Lambda_Q \mathbb{E}\left[\|y_{k+1} - y^*_{k+1}\|^2 | \mathscr{F}_k\right] + \frac{1}{\rho\gamma}\mathbb{E}\left[\|\lambda_{k+1} - \lambda^*_{k+1}\|^2 | \mathscr{F}_k\right]},\end{aligned}$$

where $\Lambda_{\hat{P}}$ and Λ_Q are the maximimum eigenvalues of $\hat{P}$ and Q, respectively. From the definition of the updates in x, in Algorithm 1 and 2, the gradient of the augmented Lagrangian is given by the following:

$$\nabla_x \mathscr{L}_A(x^*_{k+1}, y^*_{k+1}, \lambda_k) = \nabla f(x^*_{k+1}) - A^T\lambda_k + \rho A^T(Ax^*_{k+1} + By^*_{k+1} - b) + P(x^*_{k+1} - x_k) = 0, \quad (11)$$
$$\nabla_x \mathscr{L}_A(\tilde{x}^*_{k+1}, y_{k+1}, \lambda_k) = \nabla f(\tilde{x}^*_{k+1}) - A^T\lambda_k + \rho A^T(A\tilde{x}^*_{k+1} + By_{k+1} - b) + P(\tilde{x}^*_{k+1} - x_k) = 0. \quad (12)$$

Recall from Assumption 3 that f is strongly convex with parameter μ; therefore, so is $\mathscr{L}(\bullet, y, \lambda)$ implying:

$$\begin{aligned}&\mu\|\tilde{x}^*_{k+1} - x^*_{k+1}\|^2 \leq \left\langle \tilde{x}^*_{k+1} - x^*_{k+1}, \nabla_x \mathscr{L}(\tilde{x}^*_{k+1}, y_{k+1}, \lambda_k) - \nabla_x \mathscr{L}(x^*_{k+1}, y_{k+1}, \lambda_k)\right\rangle \\ \Rightarrow \quad &\mu\|\tilde{x}^*_{k+1} - x^*_{k+1}\|^2 \leq \|\tilde{x}^*_{k+1} - x^*_{k+1}\|\|-\rho A^T B(y_{k+1} - y^*_{k+1})\| \\ \Rightarrow \quad &\mu\|\tilde{x}^*_{k+1} - x^*_{k+1}\| \leq \rho\|A^T B\|\|y_{k+1} - y^*_{k+1}\| \\ \Rightarrow \quad &\mathbb{E}[\|\tilde{x}^*_{k+1} - x^*_{k+1}\|^2 \mid \mathscr{F}_k] \leq \left(\frac{\rho}{\mu}\|A^T B\|\right)^2 \mathbb{E}[\|y_{k+1} - y^*_{k+1}\|^2 \mid \mathscr{F}_k] \leq \left(\frac{\rho}{\mu}\|A^T B\|\right)^2 \eta_{k+1}.\end{aligned}$$

Then, we have that the following holds:

$$\begin{aligned}
\mathbb{E}[\|x_{k+1}-x^*_{k+1}\|^2 \mid \mathscr{F}_k] &\leq \mathbb{E}[2\|x_{k+1}-\tilde{x}^*_{k+1}\|^2 + 2\|x^*_{k+1}-\tilde{x}^*_{k+1}\|^2 \mid \mathscr{F}_k] \\
&\leq 2\mathbb{E}[\mathbb{E}[\|x_{k+1}-\tilde{x}^*_{k+1}\|^2 \mid \mathscr{F}^y_{k+1}] \mid \mathscr{F}_k] + 2\mathbb{E}[\|x^*_{k+1}-\tilde{x}^*_{k+1}\|^2 \mid \mathscr{F}_k] \\
&\leq \left(2+2\left(\frac{\rho}{\mu}\|A^TB\|\right)^2\right)\eta_{k+1} = \tau\eta_{k+1}, \quad \text{where } \tau \triangleq \left(2+2\left(\tfrac{\rho}{\mu}\|A^TB\|\right)^2\right).
\end{aligned}$$

We recall that λ^*_{k+1} is denoted by $\lambda^*_{k+1} = \lambda_k - \gamma\rho(Ax^*_{k+1} + By^*_{k+1} - b)$, implying that

$$\begin{aligned}
\mathbb{E}[\|\lambda^*_{k+1}-\lambda_{k+1}\|^2|\mathscr{F}_k] &\leq 2\gamma^2\rho^2\left(\mathbb{E}[\|A\|^2\|x^*_{k+1}-x_{k+1}\|^2|\mathscr{F}_k] + \mathbb{E}[\|B\|^2\|y_{k+1}-y^*_{k+1}\|^2|\mathscr{F}_k]\right) \\
&\leq 2\gamma^2\rho^2\left(\|A\|^2\tau + \|B\|^2\right)\eta_{k+1}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}[\|\Gamma_{k+1}(u_k)-\Gamma^*(u_k)\|_G \mid \mathscr{F}_k] \\
\leq &\sqrt{\Lambda_{\hat{P}}\mathbb{E}[\|x_{k+1}-x^*_{k+1}\|^2|\mathscr{F}_k] + \Lambda_Q\mathbb{E}[\|y_{k+1}-y^*_{k+1}\|^2|\mathscr{F}_k] + \frac{1}{\rho\gamma}\mathbb{E}[\|\lambda_{k+1}-\lambda^*_{k+1}\|^2|\mathscr{F}_k]} \\
= &\sqrt{\Lambda_{\hat{P}}\tau\eta_{k+1} + \Lambda_Q\eta_{k+1} + 2\rho\gamma(\|A\|^2\tau + \|B\|^2)\eta_{k+1}} = C\sqrt{\eta_{k+1}},
\end{aligned} \tag{13}$$

where $C \triangleq \sqrt{\Lambda_{\hat{P}}\tau + \Lambda_Q + 2\rho\gamma(\|A\|^2\tau + \|B\|^2)}$. Suppose $v_k \triangleq \|u_k - u^*\|_G$. Then

$$\begin{aligned}
\mathbb{E}[v_{k+1} \mid \mathscr{F}_k] &= \mathbb{E}[\|u_{k+1}-u^*\|_G \mid \mathscr{F}_k] \\
&= \mathbb{E}[\|\Gamma_{k+1}(u_k)-u^*\|_G \mid \mathscr{F}_k] \\
&= \mathbb{E}[\|\Gamma_{k+1}(u_k)-\Gamma^*(u_k)+\Gamma^*(u_k)-\Gamma^*(u^*)\|_G \mid \mathscr{F}_k] \\
&\leq \mathbb{E}[\|\Gamma_{k+1}(u_k)-\Gamma^*(u_k)\|_G \mid \mathscr{F}_k] + \mathbb{E}[\|\Gamma^*(u_k)-\Gamma^*(u^*)\|_G \mid \mathscr{F}_k] \\
&\leq \mathbb{E}[\|\Gamma_{k+1}(u_k)-\Gamma^*(u_k)\|_G \mid \mathscr{F}_k] + (1/\sqrt{1+\delta})\mathbb{E}[\|u_k-u^*\|_G \mid \mathscr{F}_k].
\end{aligned}$$

where the second inequality follows from the property $u^* = \Gamma^*(u^*)$ and inequality (3). It follows that

$$\mathbb{E}[v_{k+1} \mid \mathscr{F}_k] \leq C\sqrt{\eta_{k+1}} + (1/\sqrt{1+\delta})v_k, \tag{14}$$

by inequality (13). Since $\sum_{k=0}^{\infty}(1-\frac{1}{\sqrt{1+\delta}}) = \infty$ and $\sum_{k=1}^{\infty} C\sqrt{\eta_k} < \infty$ by assumption, we have that $\eta_k \to 0$. Therefore $\sqrt{\eta_k}/(1-\frac{1}{\sqrt{1+\delta}}) \to 0$ as $k \to \infty$. Finally, by invoking the supermartingale convergence lemma (Lemma 2), $v_k = \|u_k - u^*\|_G \to 0$ almost surely as $k \to \infty$. ∎

Next, we provide a simple result that is a consequence of Theorem 2 and finds usage in the subsequent results.

Lemma 3 Consider Algorithm 2. Suppose assumptions in theorem 1 and Assumption 3 hold. Then the following holds for any iterate u_k:

$$\mathbb{E}[\|u_{k+1}-u^*\|_G] \leq a\mathbb{E}[\|u_k-u^*\|_G] + C\sqrt{\eta_{k+1}}, \tag{15}$$

where $C \triangleq \sqrt{\Lambda_{\hat{P}}\tau + \Lambda_Q + 2\rho\gamma(\|A\|^2\tau + \|B\|^2)}, \tau \triangleq \left(2+2\left(\frac{\rho}{\mu}\|A^TB\|\right)^2\right)$, μ is the strongly convex constant of $f(x)$, $a \triangleq (1/\sqrt{1+\delta})$, and $\Lambda_{\hat{P}}$ and Λ_Q are the maximum eigenvalues of $\hat{P}$ and Q, respectively.

Proof. By invoking the definition of u_{k+1} in (9) and by taking unconditional expectations on both sides of (14), we obtain the result. ∎

Now we are ready to investigate how the choice of η_k affects the rate of convergence of the iterates to the unique solution.

3.2 Rate Analysis for Polynomial Decay of $\sqrt{\eta_k}$

First consider the case when $\sqrt{\eta_k} = \frac{1}{k^\alpha}, \alpha > 0$. Based on the inequality (15), we have the following results:

Theorem 3 (**Rate of convergence under polynomial decay of** η_k) Consider Algorithm 2. Suppose (15) holds and $\sqrt{\eta_k} = \frac{1}{k^\alpha}$ where $\alpha > 0$. Let $K(\alpha)$ and r_0 be defined as follows:

$$K(\alpha) \triangleq \max\left\{ \left(\frac{C}{1-a} + ar_0\right)\left(\frac{2a^{-1/\alpha}-1}{a^{-1/\alpha}-1}\right)^{\alpha}, \frac{C}{1-a^2(2a^{-1/\alpha}-1)^{\alpha}} \right\} \text{ and } r_0 = \mathbb{E}[\|u_0 - u^*\|_G].$$

Then for every $k > 0$, we have that $\mathbb{E}[\|u_k - u^*\|_G] \leq \frac{K(\alpha)}{k^\alpha}$. Furthermore, $\mathbb{E}[\|u_k - u^*\|_G] \to 0$ as $k \to \infty$.

Proof. Let $r_k = \mathbb{E}[\|u_k - u^*\|_G]$ and let $k^* \triangleq \left\lceil \frac{1}{a^{-1/\alpha}-1} \right\rceil + 1$. We prove the result by considering two cases:

(i) $k \leq k^*$: Then, from (15), we have for $k \in \{1, \ldots, k^*\}$:

$$\begin{aligned} r_k &\leq C\sum_{i=0}^{k-1} a^i (k-i)^{-\alpha} + a^k r_0 \leq C \sum_{i=0}^{k-1} a^i + a^k r_0 \\ &\leq C\left(\frac{1-a^k}{1-a}\right) + ar_0 \leq \frac{C}{1-a} + ar_0 \leq \left(\frac{C}{1-a} + ar_0\right)\left(\frac{k^*}{k}\right)^{\alpha}, \end{aligned}$$

since $k \leq k^*$ and $\alpha > 0$. Consequently, the right-hand side of the above inequality can be further bounded as follows by invoking the definition of k^*:

$$\left(\frac{C}{1-a} + ar_0\right)\left(\frac{k^*}{k}\right)^{\alpha} \leq \left(\frac{C}{1-a} + ar_0\right)\left(\frac{1}{a^{-\frac{1}{\alpha}}-1} + 2\right)^{\alpha}\left(\frac{1}{k}\right)^{\alpha} \leq \frac{K(\alpha)}{k^\alpha}.$$

(ii) $k > k^*$: We begin by noting that $K(\alpha)$ is bounded from below as follows:

$$K(\alpha) \geq \frac{C}{1-a^2(2a^{-1/\alpha}-1)^{\alpha}} = \frac{C}{1-a(2-\frac{1}{a^{-1/\alpha}})^{\alpha}} = \frac{C}{1-a(1+\frac{a^{-1/\alpha}-1}{a^{-1/\alpha}})^{\alpha}}. \tag{16}$$

Next by recalling that $C, \alpha > 0$, the following set of statements are equivalent:

$$\begin{aligned} k^* \geq \frac{1}{a^{-\frac{1}{\alpha}}-1} + 1 = \frac{a^{-\frac{1}{\alpha}}}{a^{-\frac{1}{\alpha}}-1} &\Leftrightarrow 1 + \frac{1}{k^*} \leq 1 + \frac{a^{-\frac{1}{\alpha}}-1}{a^{-\frac{1}{\alpha}}} \\ &\Leftrightarrow 1 - a\left(1+\frac{1}{k^*}\right)^{\alpha} \geq 1 - a\left(1 + \frac{a^{-\frac{1}{\alpha}}-1}{a^{-\frac{1}{\alpha}}}\right)^{\alpha} \\ &\Leftrightarrow \frac{C}{1-a\left(1+\frac{1}{k^*}\right)^{\alpha}} \leq \frac{C}{1-a\left(1+\frac{a^{-\frac{1}{\alpha}}-1}{a^{-\frac{1}{\alpha}}}\right)^{\alpha}} \end{aligned} \tag{17}$$

Then by utilizing the inequality (17) to bound the right hand side of (16), we obtain the following:

$$\frac{C}{1-a(1+\frac{a^{-1/\alpha}-1}{a^{-1/\alpha}})^{\alpha}} \geq \frac{C}{1-a(1+\frac{1}{k^*})^{\alpha}} \geq \frac{C}{1-a(1+\frac{1}{k})^{\alpha}}, \tag{18}$$

where the second inequality follows from $k > k^*$. As a result, combining (16) and (18) leads to the following:

$$C \leq K(\alpha) - a\left(1+\frac{1}{k}\right)^{\alpha} K(\alpha) \Rightarrow C + a\left(1+\frac{1}{k}\right)^{\alpha} K(\alpha) \leq K(\alpha). \tag{19}$$

We inductively prove the remainder of this result. For any $k \leq k^*$, we have $r_k \leq \frac{K(\alpha)}{k^\alpha}$ as shown in (i). This forms the inductive hypothesis. Then, from (15), we have that $\forall k \geq k^*$,

$$r_{k+1} \leq \frac{C}{(k+1)^\alpha} + a r_k \leq \frac{C}{(k+1)^\alpha} + a\frac{K(\alpha)}{k^\alpha} = \frac{C + a(1+\frac{1}{k})^\alpha K(\alpha)}{(k+1)^\alpha} \leq \frac{K(\alpha)}{(k+1)^\alpha}.$$

where the second inequality is a consequence of the inductive hypothesis and the third inequality arises from (19). The result follows.

The convergence of $\mathbb{E}[\|u_k - u^*\|_G^2]$ follows immediately when $\alpha > 0$ is fixed. ∎

Remark: Several points require emphasis. Note that $K(\alpha)$ is a function of α. In fact, $K(\alpha)$ is an increasing function of α and $1 \leq k < k^* = \left\lceil \frac{1}{a^{-\frac{1}{\alpha}}-1} \right\rceil + 1$, $K(\alpha)/k^\alpha$ can be bounded from below as follows:

$$\begin{aligned} \frac{K(\alpha)}{k^\alpha} &\geq \left(\frac{C}{1-a} + a r_0\right)\left(\frac{2a^{-1/\alpha}-1}{a^{-1/\alpha}-1}\right)^{\alpha}\left(\frac{1}{k}\right)^{\alpha} \geq \left(\frac{C}{1-a} + a r_0\right)\left(\frac{2a^{-1/\alpha}-1}{a^{-1/\alpha}-1}\right)^{\alpha}\left(\frac{a^{-1/\alpha}-1}{a^{-1/\alpha}}\right)^{\alpha} \\ &= \underbrace{\left(\frac{C}{1-a} + a r_0\right)\left(2 - a^{1/\alpha}\right)^{\alpha}}_{\triangleq h(\alpha)} \geq \left(\frac{C}{1-a} + a r_0\right), \end{aligned}$$

where the last inequality follows from the fact that $h(\alpha)$ is a concave increasing function over $\alpha \geq 0$ and $\lim_{\alpha \to 0} h(\alpha) = \frac{C}{1-a} + a r_0$. Consequently, we see that when $k < k^*$, one cannot accrue arbitrarily large benefit in terms of the error bound by increasing α.

3.3 Rate of Convergence for Geometric Decay of $\sqrt{\eta_k}$

Next we consider the case when η_k decreases geometrically and prove that the mean error converges to zero as $k \to \infty$ and does so at a geometric rate. We start by restating the following Lemma recently proved in prior work in (Ahmadi 2016).

Lemma 4 Given a function zc^z where $c < 1$. Then for all $z \geq 0$, we have that

$$zc^z < Dq^z, \text{ where } c < q < 1 \text{ and } D > \frac{1}{\ln(q/c)^e}.$$

Proof. Consider a function Dq^z where $D > 0$ and $q < 1$. Then zc^z cannot be bounded by Dq^z if $Dq^z - zc^z$ has a real positive root. Specifically, we have that any root has to satisfy

$$h(z) = \ln(D) + z\ln(q/c) - \ln(z) = 0.$$

We note that $h'(z) = \ln(q/c) - 1/z$, $h''(z) = 1/z^2 > 0$ for $z \in [0, \infty)$ implying that h is a convez function. Furthermore, $h(0) = +\infty$ and the minimizer $z^{\min} \in (0, \infty)$ where $h'(z^{\min}) = 0$ or

$$+\ln(q/c) = 1/z^{\min} \implies z^{\min} = 1/\ln(q/c).$$

If $q > c$, then $z^{\min} > 0$ and $h(z^{\min})$ is given by

$$h(z^{\min}) = \ln(D) + \frac{1}{\ln(q/c)} \ln(q/c) - \ln(1/\ln(q/c)) = 1 + \ln(D) + \ln(\ln(q/c)).$$

It suffices that $h(z^{\min}) > 0$ since the function $h(z)$ increases in z for $z \geq z^{\min}$ and there can be no real root. Consequently, if $1 > q > c$ and D is chosen such that $\ln(D) > -1 - \ln(\ln(q/c)$, then $Dq^z = zc^z$ does not have a root. Equivalently, we have that

$$D > e^{-1-\ln(\ln(q/c))} = \frac{1}{e}(\ln(q/c))^{-1} = \frac{1}{\ln(q/c)^e}.$$

Since $h(0) > 0$, it follows that $h(z) > 0$ for all $z \geq 0$ ∎

Based on this lemma, the following statement holds:

Theorem 4 (Rate of convergence under geometric decay of η_k) Consider Algorithm 2. Suppose (15) holds and suppose $\sqrt{\eta_k} = \eta^k$ for some $0 < \eta < 1$. Then for every $k > 0$, we have that $\mathbb{E}[\|u_k - u^*\|_G] \leq (CD + r_0)q^k$, where $q > b \triangleq \max\{a, \eta\}$ and D is chosen such that $D > \frac{1}{e\ln(q/b)}$. Furthemore, $\mathbb{E}[\|u_k - u^*\|_G] \to 0$ as $k \to \infty$.

Proof. Let $r_k = \mathbb{E}[\|u_k - u^*\|_G]$. Since $\sqrt{\eta_k} = \eta^k$ where $\eta < 1$, we have the following sequence of inequalities based on (15).

$$\begin{aligned} r_{k+1} &\leq ar_k + C\eta^{k+1} \leq a^2 r_{k-1} + aC\eta^k + C\eta^{k+1} \\ &\vdots \\ &\leq a^{k+1} r_0 + C\sum_{j=0}^{k} a^{k-j}\eta^{j+1} \leq b^{k+1} r_0 + C\left(\sum_{j=0}^{k} b^{k+1}\right) = (r_0 + (k+1)C)b^{k+1} \\ \Rightarrow r_k &\leq (r_0 + kC)b^k. \end{aligned}$$

From Lemma 4, it can be shown that there exist scalars q and D satisfying $q \in (b,1)$ and $D > 1/\ln((q/b)^e)$ such that $r_k \leq r_0 b^k + Ckb^k < r_0 b^k + CDq^k < (r_0 + CD)q_k$. Finally, since $q < 1$, it follows that as $k \to \infty$, $\mathbb{E}[\|u_k - u^*\|_G^2] \to 0$. ∎

4 CONCLUDING REMARKS

We consider a structured stochastic convex optimization problem where standard stochastic approximation schemes appear inadvisable. Instead, we develop an implementable stochastic inexact ADMM scheme and provide almost sure convergence statements and rate guarantee on suitable assumptions about the decay rates of the inexactness sequence. In future work, we intend to analyze the overall iteration complexity and perform extensive numerical studies.

REFERENCES

Afonso, M. V., J. M. Bioucas-Dias, and M. A. Figueiredo. 2010. “Fast Image Recovery Using Variable Splitting and Constrained Optimization”. *Image Processing, Institute of Electrical and Electronics Engineers Transactions on* 19 (9): 2345–2356.

Ahmadi, H. 2016. *On the Analysis of Data-driven and Distributed Algorithms for Convex Optimization Problems*. Ph. D. thesis, Department of Industrial and Manufacturing Engineering, Pennsylvania State University, University Park, Pennsylvania. Available via: https://etda.libraries.psu.edu/catalog/29502.

Boyd, S., N. Parikh, E. Chu, B. Peleato, and J. Eckstein. 2011. “Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers”. *Foundations and Trends® in Machine Learning* 3 (1): 1–122.

Deng, W., and W. Yin. 2012. "On the Global and Linear Convergence of the Generalized Alternating Direction Method of Multipliers". *Journal of Scientific Computing*:1–28.

Eckstein, J. 1989. *Splitting Methods for Monotone Operators with Applications to Parallel Optimization*. Ph. D. thesis, Massachusetts Institute of Technology. Available via: https://dspace.mit.edu/handle/1721.1/14356.

Gabay, D., and B. Mercier. 1976. "A Dual Algorithm for the Solution of Nonlinear Variational Problems via Finite Element Approximation". *Computers & Mathematics with Applications* 2 (1): 17–40.

Giselsson, P., and S. Boyd. 2014. "Metric Selection in Douglas-Rachford Splitting and ADMM". Available via: https://arxiv.org/abs/1410.8479.

Glowinski, R., and A. Marroco. 1975. "Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de Dirichlet non linéaires". *Revue française d'automatique, informatique, recherche opérationnelle. Analyse numérique* 9 (2): 41–76.

Koshal, J., A. Nedić, and U. V. Shanbhag. 2013. "Regularized Iterative Stochastic Approximation Methods for Variational Inequality Problems". *Institute of Electrical and Electronics Engineers Transactions on Automatic Control* 58(3):594–609.

Lin, Z., M. Chen, and Y. Ma. 2010. "The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-rank Matrices". Available via: https://arxiv.org/abs/1009.5055.

Lin, Z., R. Liu, and Z. Su. 2011. "Linearized Alternating Direction Method with Adaptive Penalty for Low-rank Representation". In *Advances in neural information processing systems*, 612–620.

Nemirovski, A., A. Juditsky, G. Lan, and A. Shapiro. 2009. "Robust Stochastic Approximation Approach to Stochastic Programming". *Society for Industrial and Applied Mathematics Journal on Optimization* 19 (4): 1574–1609.

Ouyang, H., N. He, L. Tran, and A. Gray. 2013. "Stochastic Alternating Direction Method of Multipliers". In *Proceedings of the 30th International Conference on Machine Learning*, 80–88.

Pasupathy, R., P. Glynn, S. Ghosh, and F. Hashemi. 2014. "How Much to Sample in Simulation-based Stochastic Recursions?". *Under review at Society of Industrial and Applied Mathematics Journal on Optimization*.

Polyak, B. T. 1987. *Introduction to Optimization*. Optimization Software New York.

Robbins, H., and S. Monro. 1951. "A Stochastic Approximation Method". *Annals of Mathematical Statistics* 22:400–407.

Shapiro, A., D. Dentcheva, and A. Ruszczyński. 2009. *Lectures on Stochastic Programming: Modeling and Theory*, Volume 9 of *MPS/SIAM Series on Optimization*. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM) and Mathematical Programming Society (MPS).

Wang, H., and A. Banerjee. 2013. "Online Alternating Direction Method (longer version)". Available via: http://arxiv.org/abs/1306.3721.

Yousefian, F., A. Nedić, and U. V. Shanbhag. 2012. "On Stochastic Gradient and Subgradient Methods with Adaptive Steplength Sequences". *Automatica* 48 (1): 56–67. An extended version of the paper is available via http://arxiv.org/abs/1105.4549.

AUTHOR BIOGRAPHIES

YUE XIE is a Ph.D. student in the Harold and Inge Marcus Department of Industrial and Manufacturing Engineering at Pennsylvania State University and can be reached at yux111@psu.edu

UDAY V. SHANBHAG is Professor in the Harold and Inge Marcus Department of Industrial and Manufacturing Engineering at Pennsylvania State University and can be reached at udaybag@psu.edu.