On the Convergence of the Structural Estimation of Proximal Operator with Gaussian Processes (STEP-GP) Method with Adaptive Quantization for Communication-Efficient Distributed Optimization

Aldo Duarte Vera Tudela [1], Truong Nghiem [1], and Shuangqing Wei [1]

[1]Affiliation not available

December 7, 2023

Abstract

This technical note presents proof of the convergence of the Alternating Direction Method of Multipliers (ADMM) for addressing the sharing problem when applied in conjunction with two algorithms: 1) the stochastic STEP-GP algorithm and 2) its variant named LGP, which includes adaptive uniform quantization. For the case using LGP, the coordinator can assign different quantization resolutions at each iteration and we assume that the number of bits that can be assigned is unrestricted and can go to infinity. This document describes and analyzes the two methods for integrating learning and uniform quantization into the ADMM to reduce its communication overhead and a general formulation of their communication decision method. The problems are formulated for a multi-agent setting.

On the Convergence of the Structural Estimation of Proximal Operator with Gaussian Processes (STEP-GP) Method with Adaptive Quantization for Communication-Efficient Distributed Optimization[⋆,⋆⋆]

Aldo Duarte[a,*], Truong X. Nghiem[b], Shuangqing Wei[a]

[a]*Department of Electrical Engineering, Louisiana State University, Baton Rouge, LA 70803, United States*
[b]*School of Informatics, Computing, and Cyber Systems, Northern Arizona University, Flagstaff, AZ 86011, United States*

Abstract

This technical note presents proof of the convergence of the Alternating Direction Method of Multipliers (ADMM) for addressing the sharing problem when applied in conjunction with two algorithms: 1) the stochastic STEP-GP algorithm and 2) its variant named LGP, which includes adaptive uniform quantization. For the case using LGP, the coordinator can assign different quantization resolutions at each iteration and we assume that the number of bits that can be assigned is unrestricted and can go to infinity. This document describes and analyzes the two methods for integrating learning and uniform quantization into the ADMM to reduce its communication overhead and a general formulation of their communication decision method. The problems are formulated for a multi-agent setting.

1. Introduction

This technical note serves as a complementary discussion for our previous works [1] and [2]. In those works, the Alternating Direction Method of Multipliers (ADMM) is used to solve the sharing problem in a multi-agent setting. The main goal in both cases is to reduce the ADMM communication overhead. The work in [1] extends the work in [3] which proposed an approach called STEP (STructural Estimation of Proximal operator) that relies on the concept of the Moreau Envelope. The STEP approach estimates the unknown gradient of the Moreau Envelope by constructing a set of possible gradients based on past inforamtion and then selecting a gradient that is "most likely" the true gradient. The work presented in [1] improved STEP by learning the Moreau envelopes corresponding to the local proximal operators with GP, which are updated online from past query data and used to predict the gradient. The resulting algorithm of this work was named STEP-GP.

On the other hand, in [2] the work in [1] was extended to consider uniform quantization on the agent's reply to the coordinator. Following an analysis of the statistical properties of the uniform quantization noise, we were able to derive a mechanism to adapt the uniform quantizer relying on the regression's predicted mean and covariance. This adaptation is performed so the quantization error can be approximated to follow a uniform distribution and the correlation of such error with the quantizer's input can be considered negligible. However, the inclusion of uniform quantization violates the condition for GP where all the components are considered Gaussian. For that reason, we derived a new regression scheme constructed upon the concept of a Linear Minimum Mean Square Estimator (LMMSE). To further ensure the conditions for the uncorrelation

⋆This material is based upon work supported by the U.S. National Science Foundation (NSF) under Grant No. 2238296, and State of Louisiana Economic Develop Assistantships (EDA) under Grant No. PG001065.

⋆⋆E-mail addresses: aduart3@lsu.edu (Aldo Duarte), Truong.Nghiem@nau.edu (Truong X. Nghiem), swei@lsu.edu (Shuangqing Wei)

*Corresponding author

between quantizer's input and error, orthogonal transformation and additive dithering were included. The resulting algorithm was named Linear GP (LGP).

Those works present extensive simulations proving that a significant communication overhead can be obtained by both algorithms. However, we did no prove the convergence of both algorithms analytically. In this work, we complement our previous research by presenting convergence analysis for STEP-GP and LGP.

We initiate this note with a summary of key results pertaining to the standard ADMM and the stochastic inexact ADMM that are important for our derivations. Subsequently, we delve into a brief discussion on the learning-integrated ADMM, employing adaptive uniform quantization for the sharing problem. Finally, we conduct convergence analyses for the STEP-GP and LGP algorithms where we prove that the expected value of the ADMM residual goes to zero as the algorithmic iterations go to infinity and do so at a geometric rate. As a remark, for the LGP convergence analysis, we assume that the quantization resolution can be varied by the coordinator which differs from the results in [2] where the quantization resolution was fixed.

2. Preliminary Convergence Results

2.1. Generalized ADMM Convergence Analysis

This subsection summarizes useful results from [4]; however, because the notations used in [4] are different from those used in our previous papers [1] and [2], they will be adjusted to match with our notations.

The results from [4] are for the generalized ADMM algorithm solving the general problem:

$$\begin{aligned} \text{minimize}_{x,y} \quad & f(x)+h(y) \\ \text{subject to} \quad & Ax+By=c \end{aligned}$$

The algorithm is different from the standard ADMM by introducing smoothing terms based on the norms. Let the augmented Lagrangian be

$$\mathcal{L}(x,y,m)=f(x)+h(y)-m^\top(Ax+By-c)+\frac{1}{2\rho}\|Ax+By-c\|_2^2$$

Choose $Q\succeq 0$ and a symmetric matrix P (possibly indefinite). Then each algorithm's iteration consists of:

$$\begin{aligned} y^{k+1} &= \text{argmin}_y \mathcal{L}(x^k,y,m^k)+\frac{1}{2}\|y-y^k\|_Q^2 \\ x^{k+1} &= \text{argmin}_x \mathcal{L}(x,y^{k+1},m^k)+\frac{1}{2}\|x-x^k\|_P^2 \\ m^{k+1} &= m^k-\frac{\zeta}{\rho}(Ax^{k+1}+By^{k+1}-c). \end{aligned}$$

Note that the standard ADMM is a special case of the generalized ADMM where $P=Q=0$ and $\zeta=1$.

Let $s=[x,y,m]$ with corresponding versions s^* for the optimal solutions and s^k for the algorithm iterations. Also, define the following matrices:

$$\hat{P}=P+(1/\rho)A^\top A, \qquad G=\begin{bmatrix} \hat{P} & & \\ & Q & \\ & & \frac{\rho}{\zeta}I_p \end{bmatrix}$$

where p is the dimension of m. For standard ADMM, with $P=Q=0$ and $\zeta=1$, we have

$$\hat{P}=\beta A^\top A, \qquad G=\begin{bmatrix} \beta A^\top A & & \\ & 0 & \\ & & \rho I_p \end{bmatrix}=(1/\rho)G_0^\top G_0, \qquad G_0=\begin{bmatrix} A & & \\ & 0 & \\ & & \rho I_p \end{bmatrix}$$

Also define the norm $\|s\|_G=\sqrt{s^\top G s}$. Assumptions 1 and 2 in [4] are standard for ADMM convergence.

- **Assumption 1**: There exists a saddle point $s^* = (x^*, y^*, m^*)$ to the problem, namely, x^*, y^*, and m^* satisfying the KKT conditions:

$$A^\top m^* \in \partial f(x^*)$$
$$B^\top m^* \in \partial h(y^*)$$
$$Ax^* + By^* - c = 0.$$

- **Assumption 2**: Functions f and h are convex.

Under these assumptions and another technical assumption, Theorem 3.1 in [4] provides a bound on the convergence rate of generalized ADMM. The theorem is reproduced below.

Theorem 1 (Theorem 3.1 in [4]). *Assume Assumptions 1 and 2, $\zeta = 1$, and that s^k of the Generalized ADMM is bounded (see remark below). For all scenarios in Table 1, there exists $\delta > 0$ such that*

$$\|s^k - s^*\|_G^2 \geqslant (1+\delta)\|s^{k+1} - s^*\|_G^2$$

Remark 1. *On the boundedness of $\{s^k\}$, Remark 2.2 in [4] provides several conditions. For example, if the objective functions are coercive then the boundedness is guaranteed. Also, for the standard ADMM, the boundedness is guaranteed if A and B have full column rank.*

We now apply the above results to standard ADMM and more specifically the sharing problem, as defined in [5, 6] having the form

$$\text{minimize} \quad \sum_{i=1}^{n} f_i(x_i) + h\left(\sum_{i=1}^{n} x_i\right). \tag{1}$$

Theorem 2.2 in [4] states the convergence of $\{s^k\}$ to the KKT point. In particular, for the standard ADMM as a special case of the generalized ADMM, under the same assumptions as above, we have $m^k \to m^*$, $Ax^k \to Ax^*$, and $By^k \to By^*$. The sharing problem is a special case of the standard ADMM problem with $A = I$ and $B = -I$; therefore we have $u^k \to u^*$, $x_i^k \to x_i^*$, and $\bar{y}^k \to \bar{y}^*$ (or $y_i^k \to y_i^*$). Note that here u is just the scaled version of the dual variables m.

Applying Theorem 1 to the standard ADMM and the sharing problem, we have:

- For standard ADMM: there exists $\delta > 0$ such that

$$\left\| \begin{bmatrix} A(x^k - x^*) \\ u^k - u^* \end{bmatrix} \right\|_2^2 \geqslant (1+\delta) \left\| \begin{bmatrix} A(x^{k+1} - x^*) \\ u^{k+1} - u^* \end{bmatrix} \right\|_2^2$$

- For the sharing problem with standard ADMM: there exists $\delta > 0$ such that

$$\left\| \begin{bmatrix} x_i^k - x_i^* \\ u^k - u^* \end{bmatrix} \right\|_2^2 \geqslant (1+\delta) \left\| \begin{bmatrix} x_i^{k+1} - x_i^* \\ u^{k+1} - u^* \end{bmatrix} \right\|_2^2,$$

  for all i stacked vertically.

2.2. Stochastic inexact ADMM (SI-ADMM) Convergence Analysis

This subsection summarizes the stochastic inexact ADMM for the general ADMM problem and its convergence result in [7]. The paper considers the general stochastic ADMM problem (of which the sharing problem is a special case):

$$\begin{aligned} \text{minimize}_{x,y} \quad & \mathbb{E}[\tilde{f}(x,\xi)] + \mathbb{E}[\tilde{h}(y,\xi)] \\ \text{subject to} \quad & Ax + By = c \end{aligned}$$

for some random variable ξ with known distribution. This problem can be solved by the standard ADMM if $f(x) = \mathbb{E}[\tilde{f}(x,\xi)]$ and $h(y) = \mathbb{E}[\tilde{h}(y,\xi)]$ can be calculated analytically and easily. However, this is not true in

many cases, and $f(x)$ and $g(y)$ can only be approximated. This means that the steps of the ADMM where the proximal operators of f and g are evaluated cannot be done exactly. For example, $\operatorname{argmin}_x f(x) + \frac{1}{2\rho}\|x - z_k\|_2^2$ cannot be solved exactly easily. To overcome this issue, the paper proposes to apply a sampled gradient descent approach to solve the proximal minimization problems. For instance, given a sequence of N_k^x samples $\{\xi_{k,1}^x, \ldots, \xi_{k,N_k^x}^x\}$ of ξ, then the above stochastic proximal minimization can be solved approximately by iterative gradient descent steps: $x_{k+1}^{i+1} = x_{k+1}^i - \gamma \nabla_x \left(\mathbb{E}[\tilde{f}(x, \xi_{k,i}^x)] + \frac{1}{2\rho}\|x - v_k\|_2^2 \right)$. It can be shown that as N_k^x increases, the error between $x_{k+1}^{N_k^x}$ and the true solution $x_{k+1}^\star$ decreases and can be bounded.

Given the above approach, the stochastic ADMM algorithm is modified as follows (simplified version of Algorithm 2 SI-ADMM in [7]):

$$\begin{aligned} y_{k+1} \quad & \text{is such that } \mathbb{E}[\|y_{k+1} - y_{k+1}^*\|^2] \leqslant \eta_{k+1} \\ x_{k+1} \quad & \text{is such that } \mathbb{E}[\|x_{k+1} - \tilde{x}_{k+1}^*\|^2] \leqslant \eta_{k+1} \\ m_{k+1} \quad & = m_k - \gamma\rho(Ax_{k+1} + By_{k+1} - c) \end{aligned}$$

where y_{k+1}^* is the unknown true solution of the generalized proximal minimization (with an extra smoothing term), $\tilde{x}_{k+1}^*$ is the unknown true solution of the generalized proximal minimization that uses y_{k+1} instead of y_{k+1}^* (hence the tilde).

The key theorem of the papers can be stated below (Theorem 2 in [7]).

Theorem 2. *Consider the above SI-ADMM algorithm. Under certain technical assumptions (see [7]) and* $\sum_{k=1}^\infty \sqrt{\eta_k} < \infty$*, we have that* $\|s_k - s^*\|_G \to 0$ *almost surely as* $k \to \infty$*.*

Here, $s = [x, y, u]$ and G is a matrix derived in the work on generalized ADMM [4].

Lemma 1. *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:*

$$\begin{gathered} \mathbb{E}[v_{k+1}|v_{0,\ldots,v_k}] \leqslant (1 - u_k)v_k + \mu_k \ \ a.s. \ \forall k \geqslant 0, \\ 0 \leqslant u_k \leqslant 1, \quad \mu_k \geqslant 0, \quad \forall k \geqslant 0, \quad \sum_{k=0}^\infty u_k = \infty, \quad \sum_{k=0}^\infty \mu_k < \infty, \quad \lim_{k\to\infty} \tfrac{\mu_k}{u_k} = 0. \end{gathered}$$

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

3. Problem Formulation

In the setting of a multi-agent optimization problem where the structure resembles the sharing problem as defined in (1), each of the n agents has local decision variables $x_i \in \mathbb{R}^p$ and a convex local cost function f_i: $\mathbb{R}^p \mapsto \mathbb{R}$. Their objective is to minimize the overall system cost, which comprises their local costs and a convex shared global cost function h: $\mathbb{R}^p \mapsto \mathbb{R}$. However, due to privacy concerns, each agent only knows its own cost function and cannot share it with other agents or the coordinator. The problem is solved through information exchange solely between the coordinator and the agents.

The problem presented in (1) can be solved with the ADMM. By introducing auxiliary variables y_i for each x_i, the problem can be reformulated equivalently as

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^n f_i(x_i) + h\left(\sum_{i=1}^n y_i\right) \\ \text{subject to} \quad & x_i - y_i = 0, \quad \forall i = 1, \ldots, N. \end{aligned}$$

Because the agents must keep their local cost function f_i private, each agent i only provides the solution to the following local *proximal minimization problem* to the coordinator

$$\mathbf{prox}_{\frac{1}{\rho} f_i}(z_i^k) = \underset{x_i \in \mathbb{R}^p}{\arg\min} \left\{ f_i(x_i) + \frac{\rho}{2}\|x_i - z_i^k\|^2 \right\}, \tag{2}$$

in response to a value (a query) z_i^k sent to it by the coordinator at iteration k, where $\rho > 0$ is a penalty parameter. The ADMM works in a query-response manner as follows. At iteration k, a query point z_i^k is generated by the coordinator and sent to an agent i. Each agent solves its proximal minimization problem at its query point z_i^k and replies with the response vector $\mathbf{prox}_{\frac{1}{\rho}f_i}(z_i^k)$ to the coordinator. The coordinator then updates the dual variables and generates the query points at the next iteration. Mathematically, each ADMM iteration k involves the following updates:

1. The coordinator updates the average of y_i

$$\bar{y}^{k+1} = \underset{\bar{y}\in\mathbb{R}^p}{\arg\min} \left\{ h(n\bar{y}) + (n\rho/2)\|\bar{y} - \bar{x}^k - u^k\|^2 \right\}$$

   then sends a query $z_i^k = x_i^k - \bar{x}^k + \bar{y}^{k+1} - u^k$ to each agent i.
2. Each agent i updates and sends its response $x_i^{k+1} = \mathbf{prox}_{\frac{1}{\rho}f_i}\left(z_i^k\right)$ to the coordinator.
3. The coordinator calculates the average $\bar{x}^{k+1} = (1/n)\sum_{i=1}^n x_i^{k+1}$ and updates the scaled dual vector $u^{k+1} = u^k + \bar{x}^{k+1} - \bar{y}^{k+1}$.

This process is repeated until convergence is achieved or until a maximum number of iterations is reached.

The concept of the Moreau envelope of a function f underlies the STEP approach [3], which is the foundation of the STEP-GP algorithm. For brevity, we drop the subscript i and the superscript k in the subsequent equations. For $1/\rho > 0$, the Moreau envelope $f^{\frac{1}{\rho}}$ of f is defined as

$$f^{\frac{1}{\rho}}(z) = \min_{x\in\mathbb{R}^n} \left\{ f(x) + \frac{\rho}{2}\|x - z\|^2 \right\}.$$

When f is a convex function, the Moreau envelope $f^{\frac{1}{\rho}}$ is *convex and differentiable with Lipschitz continuous gradient with constant* ρ. Moreover, the unique solution to the proximal minimization $\mathbf{prox}_{\frac{1}{\rho}f}(z)$ is [8, Proposition 5.1.7]

$$\mathbf{prox}_{\frac{1}{\rho}f}(z) = z - \frac{1}{\rho}\nabla f^{\frac{1}{\rho}}(z). \tag{3}$$

Consequently, the gradient $\nabla f^{\frac{1}{\rho}}(z)$ is all that is required to reconstruct the optimizer of (2) following from (3).

The STEP approach estimates the unknown gradient $\nabla f^{\frac{1}{\rho}}(z)$ at any query point z by constructing a set of possible gradients at z based on past queries and then selecting a gradient that is "most likely" the true gradient. The work presented in [1] improved STEP by learning the Moreau envelopes corresponding to the local proximal operators with GP, which are updated online from past query data and used to predict the gradient $\nabla f^{\frac{1}{\rho}}(z)$ for estimating the proximal operators (2) of the agents by (3). This approach is named STEP-GP.

The STEP and STEP-GP methods only consider reducing the number of agents communicating simultaneously but do not consider the *payload size* of each transmission. Our approach named LGP considers adding adaptive uniform quantization as explained in our previous work [2]. We assume in this work, that the adaptation of the quantizer is not only done over the middle point and the window length but also over the quantization resolution. This means that the coordinator can adjust the bits used for quantization as needed. The LGP approach considers first defining the communication decision variable for agent i at iteration k as

$$\gamma_i^k = \begin{cases} 1, & \text{if agent } i \text{ is queried} \\ 0, & \text{otherwise.} \end{cases}$$

When $\gamma_i^k = 1$, the query z_i^k is sent to agent i to get the quantized value of $\nabla f_i^{\frac{1}{\rho}}(z_i^k)$. On the contrary, when $\gamma_i^k = 0$, we use the predicted value $\mu_i^k(z_i^k)$ given by the LGP regression. We then define an expression β_i^k as

$$\beta_i^k = \gamma_i^k \mathbb{Q}(\nabla f_i^{\frac{1}{\rho}}(z_i^k)) + (1 - \gamma_i^k)\mu_i^k(z_i^k).$$

Contrary to the regular STEP-GP algorithm, we always have a source of inexactness either from the LGP prediction when there is no query or from the adaptive uniform quantization when a query is made. This is because the decision mechanism used in [2], where we aim for the overall inexactness to be bounded by ψ^k. The value of the threshold ψ^k determines which agents to be queried This decision mechanism is expressed in the following optimization problem: However, because of the decision mechanism used expressed in the optimization problem:

$$\begin{aligned} \underset{\gamma^k, b^k}{\text{minimize}} \quad & \sum_{i=1}^{n} \left[(\gamma_i^k) b_i^k\right] \\ \text{subject to} \quad & b_i^k \in \mathcal{N}, \\ & \gamma_i^k \in \{0,1\}, \\ & \sum_{i=1}^{n} \left[\gamma_i^k \frac{\theta}{2^{2b_i^k}} \text{trace}(\Sigma_i^k(z_i^k)) + (1-\gamma_i^k)\text{trace}(\Sigma_i^k(z_i^k))\right] < \psi^k. \end{aligned} \tag{4}$$

This threshold will decrease at each iteration to keep up with the decrease of $\text{trace}(\Sigma_i^k(z_i^k))$. Also, it's important to point out that $\gamma_i^k \frac{\theta}{2^{2b_i^k}} < (1-\gamma_i^k)$ and, depending on the value of b_i^k, the uncertainty coming from the prediction might be much bigger than the one coming from quantization for a particular agent. Furthermore, b_i^k is assumed to be unbounded so its value can be as big as necessary to satisfy the constraint. Thus, the value of this variable can go to infinity making the quantization uncertainty vanish if necessary.

Finally, the sharing ADMM expression considering the communication reduction can be expressed as:

$$\begin{aligned} \bar{y}^{k+1} &= \underset{\bar{y}\in\mathbb{R}^p}{\arg\min} \left\{h(n\bar{y}) + (n\rho/2)\|\bar{y} - \bar{x}^k - u^k\|^2\right\} \\ x_i^{k+1} &= z_i^k - (1/\rho)\beta_i^k \\ u^{k+1} &= u^k + \bar{x}^{k+1} - \bar{y}^{k+1}. \end{aligned} \tag{5}$$

4. Convergence Analysis of the STEP-GP algorithm

In this section, we present a convergence analysis for the STEP-GP algorithm presented in [1]. This algorithm will have its ADMM updates similar to (5) when solving the sharing problem. However, it considers the following problem for its querying decision-making

$$\begin{aligned} \underset{\gamma^k}{\text{minimize}} \quad & \|\gamma^k\|_1 \\ \text{subject to} \quad & \gamma_i^k \in \{0,1\}. \\ & \sum_{i=1}^{n} \left[(1-\gamma_i^k)\text{trace}(\Sigma_i^k(z_i^k))\right] < \psi^k, \end{aligned} \tag{6}$$

Here, we want to minimize the number of communicating agents while keeping the global prediction uncertainty bounded. The threshold ψ^k decreases at each iteration, ensuring that the uncertainty given to the system also reduces over time until it eventually vanishes. However, at the moment the threshold ψ^k decreases too much then we will query all agents at each iteration impacting the communication reduction in the last rounds before reaching convergence.

4.1. Preliminaries

Define $s^k = [\bar{x}^k; \bar{y}^k; u^k]$ and that $\mathcal{I}_i^k$ collects the query information from each agent i up to iteration k. The STEP-GP algorithm defines the mapping $\Gamma^{k+1} : s^k \to s^{k+1}$ which gives a mixture of inexact and exact values depending on the value of the decision variable γ_i^k. On the other hand, the exact ADMM algorithm defines the exact mapping $\Gamma_*^{k+1} : s^k \to s_*^{k+1}$ where $s_*^{k+1} = [\bar{x}_*^k; \bar{y}_*^k; u_*^k]$ are the exact values. Note that $\bar{y_*}^k = \bar{y}^k$, therefore it is always known exactly.

The convergence proof is constructed upon the querying policy in (6) and the mean square error between the inexact and exact values of x_i^{k+1} and u^{k+1}. Those expressions are given by:

- We know that $(x_i^{k+1} - x_{*,i}^{k+1}) = (1/\rho)(\beta_i^{k+1} - \beta_{*,i}^{k+1})$, and it can be shown that $\mathbb{E}[||x_i^{k+1} - x_{*,i}^{k+1}||^2|\gamma_i^k] = \text{trace}(\text{Cov}(x_i^{k+1} - x_{*,i}^{k+1}|\gamma_i^k)) = (1/\rho)^2((1-\gamma_i^k)\text{trace}(\Sigma_i^k(z_i^k)))$. Thus,

$$\mathbb{E}[||x^{k+1} - x_*^{k+1}||^2|\gamma^k] = (1/\rho)^2 \sum_{i=1}^{n} \left((1-\gamma_i^k)\text{trace}(\Sigma_i^k(z_i^k))\right)$$

- We can express $u^{k+1} = -(1/\rho)\bar{\beta}^k$, so

$$\mathbb{E}[||u^{k+1} - u_*^{k+1}||^2|\gamma^k] = \left(\frac{(1/\rho)}{n}\right)^2 \sum_{i=1}^{n} \left((1-\gamma_i^k)\text{trace}(\Sigma_i^k(z_i^k))\right) \tag{7}$$

- Due to the querying policy defined in problem (6), we have that

$$\mathbb{E}[||x^{k+1} - x_*^{k+1}||^2|\gamma^k] \leq (1/\rho)^2\psi^k$$

4.2. Upper-bound on the expected value of the ADMM residual for STEP-GP

In each iteration, we consider the following variables: s^k is the algorithm's state at the beginning; s^{k+1} is the output of the STEP-GP algorithm, which is a random variable; s_*^{k+1} is implicitly produced by the exact ADMM algorithm. Therefore, the STEP-GP algorithm produces a sequence of samples of random variables $\{s^k\}$.

Let s^* be the KKT solution (which the exact ADMM converges to). Note that s^* is a fixed point of the mapping Γ^*, that is $s^* = \Gamma^*(s^*)$. Define $\hat{s}^k = [x^k; u^k]$, which is part of s^k (excluding $\bar{y}^k$). We will consider the residual $\epsilon^k = ||\hat{s}^k - \hat{s}^*||_2 = ||s^k - s^*||_G$. Henceforth, we will omit the conditioning on γ^k for brevity. Let $\mathcal{I}^k$ denote the total collected information, i.e., the total history of the queries, of all agents i up to iteration k; in other words, $\mathcal{I}^k = \bigcup_i \mathcal{I}_i^k$.

Theorem 3. *Consider the STEP-GP algorithm for the sharing problem. Suppose the 3 assumptions in [7] hold, and* $\sum_{i=1}^{\infty} \sqrt{\psi^k} < \infty$. *Then* $\mathbb{E}[\epsilon^{k+1}|\mathcal{I}^k]$ *is bounded by*

$$\mathbb{E}[\epsilon^{k+1}|\mathcal{I}^k] \leq c\sqrt{\psi^k} + \frac{1}{\sqrt{1+\delta}}\mathbb{E}[\epsilon^k|\mathcal{I}^{k-1}],$$

where $c = (1/\rho)\sqrt{(1 + \left(\frac{1}{n^2}\right))}$.

Proof: This proof follows the proof of Theorem 2 in [7]. We first develop a bound on $\mathbb{E}[||s^{k+1} - s^*||_G]$.

$$\mathbb{E}[||s^{k+1} - s_*^{k+1}||_G|\mathcal{I}^k] = \mathbb{E}\left[\sqrt{\sum_{i=1}^{n} ||x_i^{k+1} - x_{*,i}^{k+1}||_2^2 + ||u^{k+1} - u_*^{k+1}||_2^2}\middle|\mathcal{I}^k\right]$$

$$\leq \sqrt{\sum_{i=1}^{n} \mathbb{E}[||x_i^{k+1} - x_{*,i}^{k+1}||_2^2|\mathcal{I}^k] + \mathbb{E}[||u^{k+1} - u_*^{k+1}||_2^2|\mathcal{I}^k]}$$

$$= \sqrt{(1/\rho)^2\psi^k + \mathbb{E}[||u^{k+1} - u_*^{k+1}||_2^2|\mathcal{I}^k]}$$

where the inequality comes from applying Jensen's inequality, the concavity of the square root, and the querying policy condition. For the second term we apply (7) and the constraint in (4), giving that

$$\mathbb{E}[||u^{k+1} - u_*^{k+1}||_2^2|\mathcal{I}^k] \leq \left(\frac{(1/\rho)}{n}\right)^2 \psi^k.$$

Therefore,

$$\mathbb{E}[||s^{k+1} - s_*^{k+1}||_G|\mathcal{I}^k] \leq \sqrt{(1/\rho)^2\psi^k + \left(\frac{(1/\rho)}{n}\right)^2 \psi^k} = c\sqrt{\psi^k},$$

where $c = (1/\rho)\sqrt{(1 + \left(\frac{1}{n^2}\right))}$.

Now, consider the residual ϵ^k. We have that

$$\begin{aligned}\mathbb{E}[\epsilon^{k+1}|\mathcal{I}^k] &= \mathbb{E}[||s^{k+1} - s^*||_G|\mathcal{I}^k] = \mathbb{E}[||s^{k+1} - s_*^{k+1} + s_*^{k+1} - s^*||_G|\mathcal{I}^k] \\ &\leq \mathbb{E}[||s^{k+1} - s_*^{k+1}||_G|\mathcal{I}^k] + \mathbb{E}[||\Gamma^*(s^k) - \Gamma^*(s^*)||_G|\mathcal{I}^k] \\ &\leq c\sqrt{\psi^k} + \frac{1}{\sqrt{1+\delta}}\mathbb{E}[||s^k - s^*||_G|\mathcal{I}^k]\end{aligned}$$

for some $\delta > 0$, where the last inequality comes from Theorem 3 in [4]. It follows that

$$\mathbb{E}[\epsilon^{k+1}|\mathcal{I}^k] \leq c\sqrt{\psi^k} + \frac{1}{\sqrt{1+\delta}}\mathbb{E}[\epsilon^k|\mathcal{I}^{k-1}].$$

□

4.3. Rate of Convergence and Convergence of the Expected Value of the Residual of STEP-GP

In the following lines, we prove that if ψ^k decreases geometrically then the expected value of the mean square error of the residual converges to zero as $k \to \infty$ and does so at a geometric rate. First, we restate Lemma 4 proven in [9] as:

Lemma 2. *Given a function $f(z) = zw^z$ where $w < 1$. Then, for all $z \geq 0$, we have that*

$$zw^z < Dq^z,$$

where $w < q < 1$ and $D > \frac{1}{\ln(q/w)^e}$.

This lemma makes the following Theorem to hold:

Theorem 4. *Consider the STEP-GP algorithm. Suppose that Theorem 3 holds and $\sqrt{\psi^k} = (\alpha)^k$ for some $0 < \alpha < 1$ (note that $(\alpha)^k$ refers to a constant raised to the power k). Then for every $k > 0$, we have that*

$$\mathbb{E}[\varepsilon^k|\mathcal{I}^{k-1}] \leq (cD + \varepsilon^0)(q)^k,$$

where $q > r \triangleq \max(\frac{1}{\sqrt{1+\delta}}, \alpha)$ and D is chosen such that $D > \frac{1}{e\ln(q/r)}$. Furthermore, $\mathbb{E}[\varepsilon^k|\mathcal{I}^{k-1}] \to 0$ as $k \to \infty$.

Proof: Let $a = \frac{1}{\sqrt{1+\delta}}$. Since $\sqrt{\psi^k} = (\alpha)^k$ where $\alpha < 1$, we have the following sequence of inequalities based on the bound $\mathbb{E}[\varepsilon^{k+1}|\mathcal{I}^k] \leqslant c(\alpha)^k + a\mathbb{E}[\varepsilon^k|\mathcal{I}^{k-1}]$.

$$\begin{aligned}\mathbb{E}[\varepsilon^{k+1}|\mathcal{I}^k] &\leqslant a\mathbb{E}[\varepsilon^k|\mathcal{I}^{k-1}] + c(\alpha)^k \leqslant (a)^2\mathbb{E}[\varepsilon^{k-1}|\mathcal{I}^{k-2}] + ac(\alpha)^{k-1} + c(\alpha)^k \\ &\leqslant (a)^3\mathbb{E}[\varepsilon^{k-2}|\mathcal{I}^{k-3}] + (a)^2c(\alpha)^{k-2} + ac(\alpha)^{k-1} + c(\alpha)^k \\ &\vdots \\ &\leqslant (a)^{k+1}\varepsilon^0 + c\sum_{j=0}^{k}(a)^{k-j}(\alpha)^j \leqslant (r)^k\varepsilon^0 + c\sum_{j=0}^{k}(r)^k \\ &= (\varepsilon^0 + c(k+1))(r)^k \leqslant (\varepsilon^0 + ck)(r)^k \\ \Rightarrow \mathbb{E}[\varepsilon^k|\mathcal{I}^{k-1}] &\leqslant (\varepsilon^0 + c(k-1))(r)^{k-1}.\end{aligned}$$

From Lemma 2, it can be shown that there exist scalars q and D satisfying $q \in (r, 1)$ and $D > 1/\ln((q/r)^e)$ such that

$$\mathbb{E}[\varepsilon^k|\mathcal{I}^{k-1}] \leqslant \varepsilon^0(r)^{k-1} + c(k-1)(r)^{k-1} < \varepsilon^0(r)^{k-1} + cD(r)^{k-1} < (\varepsilon^0 + cD)(q)^{k-1}.$$

Finally, since $q < 1$, it follows that as $k \to \infty$ then $\mathbb{E}[\varepsilon^k|\mathcal{I}^{k-1}] \to 0$. □

5. Convergence Analysis of the LGP algorithm with Unbounded Quantization Resolution

In this section, we present a convergence analysis for the LGP algorithm presented in [2]. However, we consider the case where the coordinator can vary the quantization resolution at each iteration and it is not fixed as in our previous study. Moreover, we assume that we can assign an infinitely large quantization resolution if needed. At the end of this section, we provide a brief discussion on what happens when there is a maximum number of bits that can be assigned for quantization.

5.1. Preliminaries

The mapping and definitions of s^*, s_*^k, and s^k are the same as in Section 4.1.

The convergence proof is constructed upon the querying policy in (4) and the mean square error between the inexact and exact values of x_i^{k+1} and u^{k+1}. Those expressions are given by:

- We know that $(x_i^{k+1} - x_{*,i}^{k+1}) = (1/\rho)(\beta_i^{k+1} - \beta_{*,i}^{k+1})$, and it can be shown that $\mathbb{E}[||x_i^{k+1} - x_{*,i}^{k+1}||^2|\gamma_i^k] = \text{trace}(\text{Cov}(x_i^{k+1} - x_{*,i}^{k+1}|\gamma_i^k)) = (1/\rho)^2(\gamma_i^k \frac{\theta}{2^{2b_i^k}}\text{trace}(\Sigma_i^k(z_i^k)) + (1-\gamma_i^k)\text{trace}(\Sigma_i^k(z_i^k)))$. Thus,

$$\mathbb{E}[||x^{k+1} - x_*^{k+1}||^2|\gamma^k] = (1/\rho)^2 \sum_{i=1}^{n} \left(\gamma_i^k \frac{\theta}{2^{2b_i^k}}\text{trace}(\Sigma_i^k(z_i^k)) + (1-\gamma_i^k)\text{trace}(\Sigma_i^k(z_i^k))\right)$$

- We can express $u^{k+1} = -(1/\rho)\bar{\beta}^k$, so

$$\mathbb{E}[||u^{k+1} - u_*^{k+1}||^2|\gamma^k] = \left(\frac{(1/\rho)}{n}\right)^2 \sum_{i=1}^{n} \left(\gamma_i^k \frac{\theta}{2^{2b_i^k}}\text{trace}(\Sigma_i^k(z_i^k)) + (1-\gamma_i^k)\text{trace}(\Sigma_i^k(z_i^k))\right) \tag{8}$$

- Due to the querying policy defined in the problem (4), we have that

$$\mathbb{E}[||x^{k+1} - x_*^{k+1}||^2|\gamma^k] \leq (1/\rho)^2\psi^k$$

5.2. Upper-bound on the expected value of the ADMM residual

In each iteration, we consider the following variables: s^k is the algorithm's state at the beginning; s^{k+1} is the output of the LGP algorithm, which is a random variable; s_*^{k+1} is implicitly produced by the exact ADMM algorithm. Therefore, the LGP algorithm produces a sequence of samples of random variables $\{s^k\}$.

Let s^* be the KKT solution (which the exact ADMM converges to). Note that s^* is a fixed point of the mapping Γ^*, that is $s^* = \Gamma^*(s^*)$. Define $\hat{s}^k = [x^k; u^k]$, which is part of s^k (excluding $\bar{y}^k$). We will consider the residual $\epsilon^k = ||\hat{s}^k - \hat{s}^*||_2 = ||s^k - s^*||_G$. Henceforth, we will omit the conditioning on γ^k for brevity. Let $\mathcal{I}^k$ denote the total collected information, i.e., the total history of the queries, of all agents i up to iteration k; in other words, $\mathcal{I}^k = \bigcup_i \mathcal{I}_i^k$.

Theorem 5. *Consider the LGP algorithm for the sharing problem. Suppose the 3 assumptions in [7] hold, and* $\sum_{k=1}^{\infty}\sqrt{\psi^k} < \infty$. *Then* $\mathbb{E}[\epsilon^{k+1}|\mathcal{I}^k]$ *is bounded by*

$$\mathbb{E}[\epsilon^{k+1}|\mathcal{I}^k] \leq c\sqrt{\psi^k} + \frac{1}{\sqrt{1+\delta}}\mathbb{E}[\epsilon^k|\mathcal{I}^{k-1}],$$

where $c = (1/\rho)\sqrt{(1 + \left(\frac{1}{n^2}\right))}$.

Proof: This proof follows the proof of Theorem 2 in [7]. We first develop a bound on $\mathbb{E}[||s^{k+1} - s^*||_G]$.

$$\mathbb{E}[||s^{k+1} - s_*^{k+1}||_G|\mathcal{I}^k] = \mathbb{E}\left[\sqrt{\sum_{i=1}^{n}||x_i^{k+1} - x_{*,i}^{k+1}||_2^2 + ||u^{k+1} - u_*^{k+1}||_2^2}\middle|\mathcal{I}^k\right]$$

$$\leq \sqrt{\sum_{i=1}^{n} \mathbb{E}[||x_i^{k+1} - x_{*,i}^{k+1}||_2^2 | \mathcal{I}^k] + \mathbb{E}[||u^{k+1} - u_*^{k+1}||_2^2 | \mathcal{I}^k]}$$

$$= \sqrt{(1/\rho)^2 \psi^k + \mathbb{E}[||u^{k+1} - u_*^{k+1}||_2^2 | \mathcal{I}^k]}$$

where the inequality comes from applying Jensen's inequality, the concavity of the square root, and the querying policy condition. For the second term we apply (8) and the constraint in (4), giving that

$$\mathbb{E}[||u^{k+1} - u_*^{k+1}||_2^2 | \mathcal{I}^k] \leq \left(\frac{(1/\rho)}{n}\right)^2 \psi^k$$

. Therefore,

$$\mathbb{E}[||s^{k+1} - s_*^{k+1}||_G | \mathcal{I}^k] \leq \sqrt{(1/\rho)^2 \psi^k + \left(\frac{(1/\rho)}{n}\right)^2 \psi^k} = c\sqrt{\psi^k},$$

where $c = (1/\rho)\sqrt{(1 + \left(\frac{1}{n^2}\right))}$.

Now, consider the residual ϵ^k. We have that

$$\mathbb{E}[\epsilon^{k+1} | \mathcal{I}^k] = \mathbb{E}[||s^{k+1} - s^*||_G | \mathcal{I}^k] = \mathbb{E}[||s^{k+1} - s_*^{k+1} + s_*^{k+1} - s^*||_G | \mathcal{I}^k]$$

$$\leq \mathbb{E}[||s^{k+1} - s_*^{k+1}||_G | \mathcal{I}^k] + \mathbb{E}[||\Gamma^*(s^k) - \Gamma^*(s^*)||_G | \mathcal{I}^k]$$

$$\leq c\sqrt{\psi^k} + \frac{1}{\sqrt{1+\delta}} \mathbb{E}[||s^k - s^*||_G | \mathcal{I}^k]$$

for some $\delta > 0$, where the last inequality comes from Theorem 3 in [4]. It follows that

$$\mathbb{E}[\epsilon^{k+1} | \mathcal{I}^k] \leq c\sqrt{\psi^k} + \frac{1}{\sqrt{1+\delta}} \mathbb{E}[\epsilon^k | \mathcal{I}^{k-1}].$$

□

5.3. Rate of Convergence and Convergence of the Expected Value of the Residual

In the following lines, we prove that if ψ^k decreases geometrically then the expected value of the mean square error of the residual converges to zero as $k \to \infty$ and does so at a geometric rate. Considering Lemma 2, we formulate the following theorem:

Theorem 6. *Consider the LGP algorithm. Suppose that* $\mathbb{E}[\varepsilon^{k+1} | \mathcal{I}^k] \leqslant c\sqrt{\psi^k} + \frac{1}{\sqrt{1+\delta}} \mathbb{E}[\varepsilon^k | \mathcal{I}^{k-1}]$ *holds and* $\sqrt{\psi^k} = (\alpha)^k$ *for some* $0 < \alpha < 1$ *(note that* $(\alpha)^k$ *refers to a constant raised to the power* k*). Then for every* $k > 0$*, we have that*

$$\mathbb{E}[\varepsilon^k | \mathcal{I}^{k-1}] \leq (cD + \varepsilon^0)(q)^k,$$

where $q > r \triangleq \max(\frac{1}{\sqrt{1+\delta}}, \alpha)$ *and* D *is chosen such that* $D > \frac{1}{e \ln(q/r)}$*. Furthermore,* $\mathbb{E}[\varepsilon^k | \mathcal{I}^{k-1}] \to 0$ *as* $k \to \infty$*.*

Proof: Let $a = \frac{1}{\sqrt{1+\delta}}$. Since $\sqrt{\psi^k} = (\alpha)^k$ where $\alpha < 1$, we have the following sequence of inequalities based on the bound $\mathbb{E}[\varepsilon^{k+1} | \mathcal{I}^k] \leqslant c(\alpha)^k + a\mathbb{E}[\varepsilon^k | \mathcal{I}^{k-1}]$.

$$\begin{aligned}
\mathbb{E}[\varepsilon^{k+1} | \mathcal{I}^k] &\leqslant a\mathbb{E}[\varepsilon^k | \mathcal{I}^{k-1}] + c(\alpha)^k \leqslant (a)^2 \mathbb{E}[\varepsilon^{k-1} | \mathcal{I}^{k-2}] + ac(\alpha)^{k-1} + c(\alpha)^k \\
&\leqslant (a)^3 \mathbb{E}[\varepsilon^{k-2} | \mathcal{I}^{k-3}] + (a)^2 c(\alpha)^{k-2} + ac(\alpha)^{k-1} + c(\alpha)^k \\
&\vdots \\
&\leqslant (a)^{k+1} \varepsilon^0 + c \sum_{j=0}^{k} (a)^{k-j} (\alpha)^j \leqslant (r)^k \varepsilon^0 + c \sum_{j=0}^{k} (r)^k \\
&= (\varepsilon^0 + c(k+1))(r)^k \leqslant (\varepsilon^0 + ck)(r)^k \\
\Rightarrow \mathbb{E}[\varepsilon^k | \mathcal{I}^{k-1}] &\leqslant (\varepsilon^0 + c(k-1))(r)^{k-1}.
\end{aligned}$$

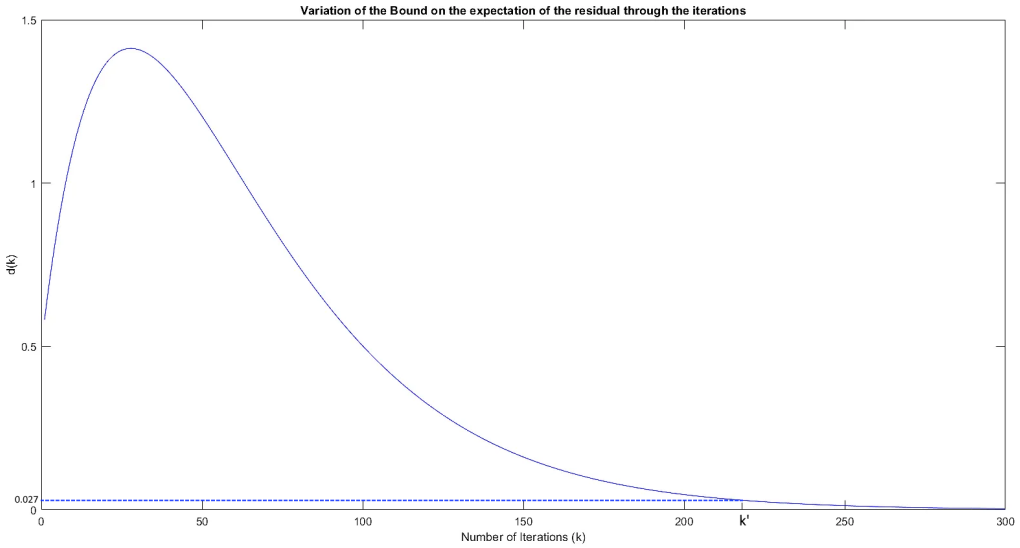

Figure 1: Upper bound of the expected value of the residual through the iterations for values of $\alpha = 0.97$, $n = 10$, $\rho = 10$.

From Lemma 2, it can be shown that there exist scalars q and D satisfying $q \in (r, 1)$ and $D > 1/\ln((q/r)^e)$ such that

$$\mathbb{E}[\varepsilon^k|\mathcal{I}^{k-1}] \leqslant \varepsilon^0(r)^{k-1} + c(k-1)(r)^{k-1} < \varepsilon^0(r)^{k-1} + cD(q)^{k-1} < (\varepsilon^0 + cD)(q)^{k-1}.$$

Finally, since $q < 1$, it follows that as $k \to \infty$ then $\mathbb{E}[\varepsilon^k|\mathcal{I}^{k-1}] \to 0$. □

5.4. Remark on the Bounded Scenario

The convergence analysis in the previous subsections assumes that the quantization resolution can be infinitely large, eventually making the uncertainty zero if all agents communicate. In real life, having infinite quantization resolution defies the purpose of using quantization. However, not allowing the quantization resolution to be infinitely large goes against the condition to conclude convergence that the uncertainty eventually reaches zero as k goes to infinity.

Let's define the maximum possible value of b_i^k as $b_{\max}$. At the limit case, let's say that iteration k' is the last iteration where the constraint is met having the form:

$$\sum_{i=1}^{n} \left(\frac{\theta}{2^{2b_{\max}}} \text{trace}(\Sigma_i^{k'}(z_i^{k'})) \right) < \psi_i^{k'},$$

where we assume the lowest possible uncertainty that is attained when all agents are queried. This is because, for any given agent, the quantization uncertainty is smaller than the GP prediction uncertainty following the expression in (4). Iteration k' is the last one where the constraint is met. Therefore, at iteration $k'+1$ we are forced to stop the algorithm because the threshold $\psi_i^{k'+1}$ will be smaller than our uncertainty measure that can't decrease any further. However, at iteration k' we still satisfy the condition so the results of Theorem 5 hold leading to

$$\mathbb{E}[\epsilon^{k'+1}|\mathcal{I}^{k'}] \leq c\sqrt{\psi^{k'}} + \frac{1}{\sqrt{1+\delta}}\mathbb{E}[\epsilon^{k'}|\mathcal{I}^{k'-1}].$$

Then, we can do the same analysis as in the proof of Theorem 6 leading to the inequality

$$\mathbb{E}[\varepsilon^{k'+1}|\mathcal{I}^{k'}] \leq (\varepsilon^0 + ck')(r)^{k'}.$$

Figure 1 shows the plot of the bound $(\varepsilon^0 + ck)(r)^k$ where we can see that it increments up to a certain iteration at first and then starts decreasing indefinitely. The region where the function $d(k)$ decreases is given by

$$k \geq \frac{r}{(1-r)} - \frac{\varepsilon^0}{c}.$$

The constant $\frac{r}{(1-r)} - \frac{\varepsilon^0}{c}$ defines the iteration where $d(k)$ starts to be decreasing.

Unfortunately, having a bound for k' involves having a bound on trace$(\Sigma_i^{k'}(z_i^{k'}))$ which depends on k'. However, the error bound is not infinitely large even in the worst-case scenario meaning that the uncertainty is always bounded. Considering that iteration k' is an extreme case and assuming a well-designed threshold mechanism, we anticipate that by the time we reach this moment the bound on the residual is not significantly large as shown in Figure 1, so the solution we have at that moment is not too far away from the true solution. In the hypothetical case presented in Figure 1, considering that k' happens at iteration 220 then the bound on the residual is really small at 0.027. This is mostly because our algorithms make sure the overall uncertainty keeps decreasing at each iteration.

6. Conclusion

In this document we presented a convergence analysis for the STEP-GP and LGP algorithms. These analyses had their foundation in the convergence analysis of the generalized ADMM and the SI-ADMM algorithms. For the case of the LGP algorithm analysis, we assumed that the coordinator can vary the quantization resolution at each iteration and that it can assign infinitely large bits for quantization. A brief discussion was presented about the case when the quantization resolution is upper-bounded.

References

[1] T. X. Nghiem, G. Stathopoulos, C. Jones, Learning Proximal Operators with Gaussian Processes, in: Annual Allerton Conference on Communication, Control, and Computing, Illinois, USA, 2018.

[2] A. Duarte, T. X. Nghiem, S. Wei, Communication-efficient ADMM using Quantization-aware Gaussian Process Regression (8 2022). doi:10.36227/techrxiv.20448222.v1.

[3] G. Stathopoulos, C. Jones, Communication reduction in distributed optimization via estimation of the proximal operator, arXiv preprint arXiv:1803.07143 (03 2018).

[4] W. Deng, W. Yin, On the global and linear convergence of the generalized alternating direction method of multipliers, J. Sci. Comput. 66 (3) (2016) 889–916. doi:10.1007/s10915-015-0048-x.
URL https://doi.org/10.1007/s10915-015-0048-x

[5] X. Cao, K. J. R. Liu, Dynamic sharing through the ADMM, IEEE Transactions on Automatic Control 65 (5) (2020) 2215–2222. doi:10.1109/TAC.2019.2940317.

[6] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, Found. Trends Mach. Learn. (2011).

[7] Y. Xie, U. V. Shanbhag, Si-admm: A stochastic inexact admm framework for resolving structured stochastic convex programs, in: 2016 Winter Simulation Conference (WSC), 2016, pp. 714–725. doi:10.1109/WSC.2016.7822135.

[8] D. P. Bertsekas, Convex Optimization Algorithms, Athena Scientific, 2015.

[9] Y. Xie, U. V. Shanbhag, Si-admm: A stochastic inexact admm framework for resolving structured stochastic convex programs, in: 2016 Winter Simulation Conference (WSC), 2016, pp. 714–725. doi:10.1109/WSC.2016.7822135.