Computational Optimization and Applications
https://doi.org/10.1007/s10589-021-00338-8

An inexact accelerated stochastic ADMM for separable convex optimization

Jianchao Bai[1] · **William W. Hager**[2] · **Hongchao Zhang**[3]

Received: 14 February 2021 / Accepted: 23 November 2021

Abstract
An inexact accelerated stochastic Alternating Direction Method of Multipliers (AS-ADMM) scheme is developed for solving structured separable convex optimization problems with linear constraints. The objective function is the sum of a possibly nonsmooth convex function and a smooth function which is an average of many component convex functions. Problems having this structure often arise in machine learning and data mining applications. AS-ADMM combines the ideas of both ADMM and the stochastic gradient methods using variance reduction techniques. One of the ADMM subproblems employs a linearization technique while a similar linearization could be introduced for the other subproblem. For a specified choice of the algorithm parameters, it is shown that the objective error and the constraint violation are $\mathcal{O}(1/k)$ relative to the number of outer iterations k. Under a strong convexity assumption, the expected iterate error converges to zero linearly. A linearized variant of AS-ADMM and incremental sampling strategies are also discussed. Numerical experiments with both stochastic and deterministic ADMM algorithms show that AS-ADMM can be particularly effective for structured optimization arising in big data applications.

Keywords Convex optimization · Separable structure · Accelerated stochastic ADMM · Inexact stochastic ADMM · AS-ADMM · Accelerated gradient method · Complexity · Big data

1 Introduction

We consider the following structured separable convex optimization problems with linearly equality constraints:

✉ William W. Hager
hager@ufl.edu
http://people.clas.ufl.edu/hager/

Extended author information available on the last page of the article

Published online: 11 January 2022

$$\min\{f(\mathbf{x})+g(\mathbf{y}) \,:\, \mathbf{x}\in\mathcal{X},\ \mathbf{y}\in\mathcal{Y},\ A\mathbf{x}+B\mathbf{y}=\mathbf{b}\},\tag{1.1}$$

where $\mathcal{X}\subset\mathbb{R}^{n_1}$ and $\mathcal{Y}\subset\mathbb{R}^{n_2}$ are closed convex subsets, $g:\mathcal{Y}\to\mathbb{R}\cup\{+\infty\}$ is a convex, but not necessarily smooth function, $A\in\mathbb{R}^{n\times n_1}$, $B\in\mathbb{R}^{n\times n_2}$, and $\mathbf{b}\in\mathbb{R}^n$ are given, and f is an average of N real-valued convex functions:

$$f(\mathbf{x})=\frac{1}{N}\sum_{j=1}^{N}f_j(\mathbf{x}).$$

It is assumed that each f_j is defined on an open set containing $\mathcal{X}$ and that $f_j:\mathcal{X}\to\mathbb{R}$ is Lipschitz continuously differentiable. Problem (1.1) corresponds to the regularized empirical risk minimization in big data applications, including classification and regression models in machine learning, where N denotes the sample size and f_j is the empirical loss. A major difficulty in problems of the form (1.1) is that N can be very large, and hence, it would be expensive to evaluate either f or its gradient in each iteration.

The Lagrangian associated with (1.1) is

$$\mathcal{L}(\mathbf{x},\mathbf{y},\lambda)=f(\mathbf{x})+g(\mathbf{y})+\lambda^{\mathsf{T}}(\mathbf{b}-A\mathbf{x}-B\mathbf{y}),\tag{1.2}$$

while the augmented Lagrangian with penalty $\beta>0$ is

$$\mathcal{L}_\beta(\mathbf{x},\mathbf{y},\lambda)=\mathcal{L}(\mathbf{x},\mathbf{y},\lambda)+\frac{\beta}{2}\|\mathbf{b}-A\mathbf{x}-B\mathbf{y}\|^2.\tag{1.3}$$

The Alternating Direction Method of Multipliers (ADMM) [15, 16] is an effective approach to exploit the separable structure of the objective function. Assuming the existence of a solution to the first-order KKT optimality system for (1.1), Gabay [14, pp. 316–322] shows that the following ADMM scheme

$$\begin{cases}\mathbf{x}^{k+1} & \in\ \arg\min_{\mathbf{x}\in\mathcal{X}}\mathcal{L}_\beta\big(\mathbf{x},\mathbf{y}^k,\lambda^k\big),\\ \mathbf{y}^{k+1} & \in\ \arg\min_{\mathbf{y}\in\mathcal{Y}}\mathcal{L}_\beta\big(\mathbf{x}^{k+1},\mathbf{y},\lambda^k\big),\\ \lambda^{k+1} & =\ \lambda^k+\beta\big(\mathbf{b}-A\mathbf{x}^{k+1}-B\mathbf{y}^{k+1}\big),\end{cases}\tag{1.4}$$

is a special case of the Douglas-Rachford splitting method [10, 11] applied to the stationary system for the dual of (1.1). ADMM was proved convergent for the problem with two-block variables [15], while the direct extension to more than two blocks is not necessarily convergent [6], although its efficiency has been observed in some applications [22, 34].

ADMM and its variants have been extensively studied in the literature and applied to a wide range of applications in signal and image processing, and in statistical and machine learning. Here, we briefly review some of the ADMM literature. Classes of ADMM-type methods include proximal ADMM [2, 31], inexact ADMM [18–20, 27], and linearized/relaxed ADMM [37, 39]. Most of these are globally convergent with an $\mathcal{O}(1/k)$ ergodic convergence rate, where k denotes the iteration number. Some improvements in the convergence rate of ADMM have been obtained including [9] where the same $\mathcal{O}(1/k)$ convergence rate is obtained in a multi-block setting with a Jacobi-proximal implementation. For either a linear or a quadratic programming problem, the classic ADMM scheme and its variant have a linear convergence rate [3]. Under the assumption that the subdifferential of each component objective function is piecewise linear, the global linear convergence of ADMM for two-block separable convex optimization has been established in [38]. Assuming that an error bound condition holds and that the dual stepsize is sufficiently small, Hong and Luo [23] showed an R-linear convergence rate of their multi-block ADMM. Under the hypothesis that some of the underlying functions are strongly convex, global linear convergence of ADMM-type algorithms and their corresponding proximal/generalized versions have been established [4, 17, 21, 25, 30].

Notice that in standard deterministic ADMM for (1.1), gradient methods are often used to solve the subproblem involving f. Hence, the gradient of f needs to be evaluated at each iteration, which requires the gradient of each component function f_i. This could be expensive or impossible when N is large in big data applications. Hence, ADMM type algorithms have been designed in recent years to solve structured optimization problems of the form (1.1) using stochastic inexact gradients. Research in the stochastic gradient ADMM area includes [1, 26, 28, 32, 36, 40, 42].

The algorithm analyzed in this paper is the inexact accelerated ADMM, denoted AS-ADMM, given in Algorithm 1.1. Note that AS-ADMM contains a routine `xsub` to generate an approximation to the solution of the $\mathbf{x}$-subproblem in (1.4), and two steps corresponds to updates $\mathbf{y}^{k+1}$ and $\boldsymbol{\lambda}^{k+1}$ in (1.4). The algorithm is inexact since the solution of the $\mathbf{x}$-subproblem is approximated in `xsub`. The algorithm is stochastic since in each step of `xsub`, the gradient is computed at a randomly chosen component f_j of f. The outcome of AS-ADMM is stochastic since it depends on the randomly chosen component f_j where the gradient is evaluated. The structure of AS-ADMM is somewhat typical of the structure for stochastic gradient ADMM algorithms.

Parameters: $\beta > 0$, $s \in (0, (1+\sqrt{5})/2]$ and $\mathcal{H} \succ \mathbf{0}$.
Initialization: $(\mathbf{x}^0, \mathbf{y}^0, \boldsymbol{\lambda}^0) \in \mathcal{X} \times \mathcal{Y} \times \mathbb{R}^n$, $\breve{\mathbf{x}}^0 = \mathbf{x}^0$.

```
For k = 0, 1, ...
    Choose M_k, η_k > 0 and 𝓜_k such that 𝓜_k − βAᵀA ⪰ 0.
```

$$\mathbf{h}^k := -A^\mathsf{T}\left[\boldsymbol{\lambda}^k - \beta(A\mathbf{x}^k + B\mathbf{y}^k - \mathbf{b})\right].$$

$$(\mathbf{x}^{k+1}, \breve{\mathbf{x}}^{k+1}) = \textbf{xsub}\ (\mathbf{x}^k, \breve{\mathbf{x}}^k, \mathbf{h}^k).$$

$$\mathbf{y}^{k+1} \in \arg\min\left\{g(\mathbf{y}) + \frac{\beta}{2}\left\|A\mathbf{x}^{k+1} + B\mathbf{y} - \mathbf{b} - \frac{\boldsymbol{\lambda}^k}{\beta}\right\|^2 : \mathbf{y} \in \mathcal{Y}\right\}.$$

$$\boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k - s\beta\left(A\mathbf{x}^{k+1} + B\mathbf{y}^{k+1} - \mathbf{b}\right).$$

end

$(\mathbf{x}^+, \breve{\mathbf{x}}^+) = \textbf{xsub}\ (\mathbf{x}_1, \breve{\mathbf{x}}_1, \mathbf{h})$.

For $t = 1, 2, \ldots, M_k$

Randomly select $\xi_t \in \{1, 2, \ldots, N\}$ with uniform probability.

$\beta_t = 2/(t+1)$, $\gamma_t = 2/(t\eta_k)$, $\widehat{\mathbf{x}}_t = \beta_t \breve{\mathbf{x}}_t + (1-\beta_t)\mathbf{x}_t$.

$\mathbf{d}_t = \widehat{\mathbf{g}}_t + \mathbf{e}_t$, where $\widehat{\mathbf{g}}_t = \nabla f_{\xi_t}(\widehat{\mathbf{x}}_t)$ and $\mathbf{e}_t$ is a random vector satisfying $\mathbb{E}[\mathbf{e}_t] = \mathbf{0}$.

$$\breve{\mathbf{x}}_{t+1} = \arg\min\left\{\langle \mathbf{d}_t + \mathbf{h}, \mathbf{x}\rangle + \frac{\gamma_t}{2}\|\mathbf{x} - \breve{\mathbf{x}}_t\|^2_{\mathcal{H}} + \frac{1}{2}\left\|\mathbf{x} - \mathbf{x}^k\right\|^2_{\mathcal{M}_k} : \mathbf{x} \in \mathcal{X}\right\}.$$

$$\mathbf{x}_{t+1} = \beta_t \breve{\mathbf{x}}_{t+1} + (1-\beta_t)\mathbf{x}_t.$$

end

Return $(\mathbf{x}^+, \breve{\mathbf{x}}^+) = (\mathbf{x}_{M_k+1}, \breve{\mathbf{x}}_{M_k+1})$.

ALG. 1.1. *Accelerated Stochastic ADMM (AS-ADMM)*

It seems that the first development of a stochastic gradient ADMM scheme is given in [28]. In the context of (1.1), the algorithm computes the gradient of a single randomly chosen component f_j, and uses this gradient to linearize f_j at the current iterate. The solution of the linearized problem yields $\mathbf{x}^{k+1}$. If $\mathbb{E}$ denotes expectation, $(\mathbf{x}^*, \mathbf{y}^*)$ denotes a solution of (1.1), and $\mathcal{X}$ is compact, then it is shown that

$$\mathbb{E}\left[f(\overline{\mathbf{x}}_k) + g(\overline{\mathbf{x}}_k) - f(\mathbf{x}^*) - g(\mathbf{x}^*) + \|A\overline{\mathbf{x}}_k + B\overline{\mathbf{y}}_k - \mathbf{b}\|\right] \le c/\sqrt{k}, \tag{1.5}$$

where the bar over an iterate means the average of the first k iterates. Without some additional information, such as $f(\overline{\mathbf{x}}_k) + g(\overline{\mathbf{x}}_k) \ge f(\mathbf{x}^*) + g(\mathbf{x}^*)$, this bound is not strong enough to ensure that the expected objective value or constraint violation tends to zero. In [32] the same algorithm is considered, but in the special case that $B = -\mathbf{I}$ and $\mathcal{A}\mathbf{x} \in \mathcal{Y}$ for all $\mathbf{x} \in \mathcal{X}$. For this special case, $\overline{\mathbf{y}}^k$ can be replaced by $A\overline{\mathbf{x}}^k$ to obtain a feasible point, and (1.5) yields an $\mathcal{O}(1/\sqrt{k})$ bound for the objective error. In [1] the error bound (1.5) is sharpened to $\mathcal{O}(1/k)$ by further developing the algorithm in [28] by introducing a more complex averaging process and additional assumptions such as both $\mathcal{X}$ and $\mathcal{Y}$ compact, and the dual multipliers are bounded. Another variation of the method in [28] is given in [42] with an error bound of $\mathcal{O}(1/k)$.

The paper [40] seems to be the first to realize the potential benefit of solving the **x**-subproblem with greater accuracy. Using $M_k = \mathcal{O}(k^{2\varrho})$ inner iterations for the **x** -subproblem with $\varrho > 1$, an $\mathcal{O}(1/k)$ bound was established for the left side of (1.5). The paper [26] seems to represent the current state-of-the-art for problems of the form (1.1) with smooth f_j and potentially nonsmooth g. There were two fundamental innovations. First, for their algorithm ASVRG-ADMM, the **x**-subproblem takes advantage of both a momentum acceleration trick from [35] and variance reduction techniques from [24] when performing a fixed number m inner iterations with a fixed batch size for the stochastic gradients. Second, in the analysis of ASVRG-ADMM, the authors exploit an observation from [41] to obtain an $\mathcal{O}(1/k)$ bound for both the objective error and constraint violation.

In comparing AS-ADMM to the previous work, the STOC-ADMM scheme proposed in [28], and the various modifications of it, use one stochastic gradient step in each ADMM iteration to approximately solve the **x**-subproblem, while the scheme ASVRG-ADMM proposed in [26] uses a fixed number m inner iterations. In contrast, our AS-ADMM uses a dynamic M_k (see (4.10)) accelerated stochastic gradient iterations to solve the **x**-subproblem with increasing accuracy as the iterations progress. We found this strategy particularly effective in our earlier work [19] on an inexact, adaptive ADMM scheme. The number of iterations is chosen so as to achieve a convergence rate of either $\mathcal{O}(1/k)$ or $\mathcal{O}(k^{-1}\log k)$, based on the theory in our paper. In a specific adaptive scheme that we analyze, $M_k = \mathcal{O}(k^{\varrho})$ with $\varrho \geq 1$.

In the ASVRG-ADMM scheme, the **y**-subproblem is solved at each inner iteration; hence, in k iterations, ASVRG-ADMM will solve the **y**-subproblem mk times. In contrast, AS-ADMM treats the **y**-subproblem as a single step in the outer iteration, and it is only solved k times during k iterations.

Another fundamental difference between these schemes is that AS-ADMM does not require an estimate for the Lipschitz constant of ∇f, while ASVRG-ADMM uses the Lipschitz constant within the algorithm, as is typical in stochastic gradient techniques. In ASVRG-ADMM the Lipschitz constant is used to compute the momentum parameter which appears within the steps of the algorithm. Hence, a poor estimate of the Lipschitz constant could significantly affect the performance of ASVRG-ADMM and other stochastic ADMMs. If a good estimate of either the local or global Lipschitz constant were known, then it can be exploited in AS-ADMM, but it is not required in the algorithm.

A fundamental difference between the stochastic and deterministic ADMM literature is that in the deterministic setting, the literature typically establishes convergence of the iterates to a stationary point for (1.1), assuming the gradient of f is Lipschitz continuous. The corresponding convergence results in the stochastic setting have not yet been established; what is established is the convergence of the expected objective error and constraint violation. However, under strengthened assumptions, such as strong convexity, convergence of the expected ergodic error as well as convergence of the expected iterate error can be deduced (see Appendix).

A very recent paper [36] developed an inexact stochastic gradient algorithm SI-ADMM for a different version of (1.1), where not only f is viewed as stochastic, but also g. To incorporate the setting of [36] in (1.1), one should also view g as the sum of component functions g_j, just like f. The algorithm in [36] differs from our algorithm in that SI-ADMM is based on gradient steps for the augmented Lagrangian and proximal term, while AS-ADMM employs a linearization technique described in item 3 below. The assumptions in [36] imply that both f_j and g_j are strongly convex and Lipschitz continuous, that $\mathcal{X} = \mathbb{R}^{n_1}$ and $\mathcal{Y} = \mathbb{R}^{n_2}$, and that the linear constraint in (1.1) has full row rank. In this very smooth and strongly convex setting, a linear convergence rate for the expected error in the SI-ADMM iterates is established. In the Appendix of our paper, we also show that the expected error in the AS-ADMM iterates converges to zero at a linear rate when f and g are strongly convex.

In more detail, some features of AS-ADMM are the following:

1. The memory cost of AS-ADMM is low since the prior stochastic gradients and iterates are not saved, which is advantageous in big data applications. For a specific choice of η_k and M_k given in (4.10), we show in Theorem 4.2 that the expectation of the objective error and constraint violation for an ergodic mean of the AS-ADMM iterates is $\mathcal{O}(1/k)$. The Appendix introduces additional assumptions to obtain results concerning the convergence of the expected error in the iterates. For example, when $\mathcal{M}_k - \beta A^\mathsf{T} A$ is uniformly positive definite, then the iterates are bounded in expectation, and when f and g are strongly convex, the expected error in the iterates converges linearly to zero.
2. Although the AS-ADMM algorithm does not require knowledge of the Lipschitz constant for the gradient of f, faster convergence may be possible when a good estimate of the Lipschitz constant ν for ∇f_j, $1 \le j \le N$, is known and exploited. In particular, the convergence results apply when η_k reaches the interval $(0, 1/\nu)$; for the choice of η_k given in (4.10), η_k tends to zero, so it eventually lies in the interval where convergence is guaranteed. But if the Lipschitz constant is known, we could always take $\eta_k \in (0, 1/\nu)$ and the convergence rates would be valid from the start of the iterations.
3. The routine `xsub` is obtained from the deterministic inexact ADMM scheme in [20] by replacing the full gradient by a stochastic gradient. In the deterministic setting, it is shown in [20] (see Lemma 3.1 and the parameter choice (2.4) in [20]) that this inexact ADMM is an accelerated scheme for solving the problem

$$\arg\min_{\mathbf{x}\in\mathcal{X}} \mathcal{L}_\beta\left(\mathbf{x}, \mathbf{y}^k, \lambda^k\right) + \frac{1}{2}\|\mathbf{x} - \mathbf{x}^k\|^2_{\mathcal{D}_k}, \quad \mathcal{D}_k = \mathcal{M}_k - \beta A^\mathsf{T} A. \tag{1.6}$$

   Note that both the objective function and the penalty term of (1.6) are linearized to some degree in the optimization problem contained in `xsub`. The objective function of (1.6) is linearized by replacing the objective f by ∇f_j for some j, while the penalty term is partly linearized by including a proximal term of the form $(1/2)\|\mathbf{x} - \mathbf{x}^k\|^2_{\mathcal{M}_k - \beta A^\mathsf{T} A}$ in (1.6). This proximal term annihilates $(\beta/2)\|A\mathbf{x}\|^2$ in the penalty term. If $\mathcal{M}_k$ and $\mathcal{H}$ in `xsub` were a multiple of the identity, then the Hessian of the objective for the optimization problem in `xsub` would be a multiple of the identity. The constraint $\mathcal{M}_k - \beta A^\mathsf{T} A \succeq \mathbf{0}$ in AS-ADMM arises from the proximal term in (1.6).

4. AS-ADMM allows for variance reduction techniques. In each iteration of `xsub`, a stochastic gradient $\widehat{\mathbf{g}}_t$ of the function f at $\widehat{\mathbf{x}}_t$ is generated, and the user has the flexibility of choosing a zero mean random vector $\mathbf{e}_t$ to reduce the variance of $\widehat{\mathbf{g}}_t$. A trivial choice is $\mathbf{e}_t = \mathbf{0}$; however, faster convergence is observed in the numerical experiments when a variance reduction technique is employed.
5. In the standard deterministic Gauss-Seidel version of ADMM, a dual step $s \in (0, (1+\sqrt{5})/2)$ (the open interval) is used. In the stochastic AS-ADMM, the stepsize constraint is $s \in (0, (1+\sqrt{5})/2]$ (the half-open interval) since we only show convergence of the function values. If $M_k = 1$ and $N = 1$, then AS-ADMM becomes the standard linearized ADMM. If $M_k > 1$ and $N = 1$, then AS-ADMM is a deterministic inexact ADMM, where the $\mathbf{x}$-subproblems of ADMM are solved inexactly using M_k accelerated gradient iterations. Hence, our convergence results for AS-ADMM also imply convergence results for an inexact deterministic ADMM based on M_k accelerated gradient iterations. Similar to the Gauss-Seidel version of ADMM, $s \in (0, (1+\sqrt{5})/2)$ guarantees convergence of the iterates for this inexact deterministic ADMM, a result not previously known in the literature. In fact, the more general multi-block convergence results in [19, 20] require that $s \in (0, 1)$.
6. As shown in the analysis, the constraint in AS-ADMM that $\mathcal{D}_k = \mathcal{M}_k - \beta A^\mathsf{T} A$ is positive semidefinite can be weakened to $(\mathbf{x}^{k+1} - \mathbf{x}^k)^\mathsf{T} \mathcal{D}_k (\mathbf{x}^{k+1} - \mathbf{x}^k) \geq 0$ for all k sufficiently large. In Remark 4.2, we show that when $\mathcal{M}_k = \rho_k \mathbf{I}$, there is an easy and effective way to adjust ρ_k during the iterations, based on an underestimate of the largest eigenvalue of $\beta A^\mathsf{T} A$, so as to satisfy the weakened constraint on $\mathcal{D}_k$ when k is sufficiently large.
7. Our numerical experiments show that AS-ADMM performs much better than deterministic ADMM methods for solving problem (1.1) when it is expensive to compute the exact gradient of f, and it is competitive or faster than other state-of-the-art stochastic ADMM type algorithms [16, 26, 28, 29], especially when the linear constraints are not simple.

The paper is organized as follows. Section 2 introduces some notation and assumptions. Detailed convergence analysis of AS-ADMM is given in Sections 3 and 4. Incremental sampling strategies and a linearized variant of AS-ADMM are also briefly discussed in Sections 5 and 6. Numerical experiments comparing AS-ADMM with both deterministic and stochastic ADMM type algorithms are given in Section 7. The Appendix develops properties for the expected iterates under stronger assumptions. In particular, the AS-ADMM iterates are bounded in expectation when the proximal term is uniformly positive definite, while the expected error in the iterates converges to zero at a linear rate under a strong convexity assumption.

2 Notation and assumptions

Let $\mathbb{R}$, $\mathbb{R}^n$, and $\mathbb{R}^{n\times m}$ be the sets of real numbers, n dimensional real column vectors, and $n \times m$ real matrices, respectively. Let $\mathbf{I}$ denote the identity matrix and $\mathbf{0}$ denote zero matrix/vector. For symmetric matrices A and B of the same dimension, $A \succ B$ ($A \succeq B$) means $A - B$ is a positive definite (semidefinite) matrix. For any symmetric matrix G, $\|\mathbf{x}\|_G^2 := \mathbf{x}^\mathsf{T} G \mathbf{x}$, where the superscript $^\mathsf{T}$ denotes the transpose. Note that G could be indefinite with $\mathbf{x}^\mathsf{T} G \mathbf{x} < 0$ for some $\mathbf{x}$. If G is positive definite, then $\|\mathbf{x}\|_G$ is a norm. We use $\|\cdot\|$ and $\langle \cdot, \cdot \rangle$ to denote the standard Euclidean norm and inner product; $\nabla f(\mathbf{x})$ is the gradient of f at $\mathbf{x}$. For convenience in the analysis, we define

$$\mathbf{w} = \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \\ \lambda \end{pmatrix} \quad \text{and} \quad \mathcal{J}(\mathbf{w}) = \begin{pmatrix} -A^\mathsf{T}\lambda \\ -B^\mathsf{T}\lambda \\ A\mathbf{x} + B\mathbf{y} - \mathbf{b} \end{pmatrix}. \tag{2.1}$$

We also define $F(\mathbf{w}) = f(\mathbf{x}) + g(\mathbf{y})$ and $\mathbf{w}^k = (\mathbf{x}^k, \mathbf{y}^k, \lambda^k)$. The affine map $\mathcal{J}(\cdot)$ is skew-symmetric in the sense that

$$(\mathbf{w} - \mathbf{v})^\mathsf{T}[\mathcal{J}(\mathbf{w}) - \mathcal{J}(\mathbf{v})] = 0 \tag{2.2}$$

for all $\mathbf{v}$ and $\mathbf{w} \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \mathbb{R}^n$. In other words, the matrix associated with $\mathcal{J}$ is skew symmetric.

The point $\mathbf{w}^* := (\mathbf{x}^*, \mathbf{y}^*, \lambda^*) \in \Omega := \mathcal{X} \times \mathcal{Y} \times \mathbb{R}^n$ is a saddle-point of the Lagrangian $\mathcal{L}$, given in (1.2), if

$$\mathcal{L}(\mathbf{x}^*, \mathbf{y}^*, \lambda) \le \mathcal{L}\left(\mathbf{x}^*, \mathbf{y}^*, \lambda^*\right) \le \mathcal{L}\left(\mathbf{x}, \mathbf{y}, \lambda^*\right)$$

for every $\mathbf{w} = (\mathbf{x}, \mathbf{y}, \lambda) \in \Omega$. It follows that

$$\begin{array}{rcll} f(\mathbf{x}) - f(\mathbf{x}^*) & + & (\mathbf{x} - \mathbf{x}^*)^\mathsf{T}\left(-A^\mathsf{T}\lambda^*\right) & \ge 0, \\ g(\mathbf{y}) - g(\mathbf{y}^*) & + & (\mathbf{y} - \mathbf{y}^*)^\mathsf{T}\left(-B^\mathsf{T}\lambda^*\right) & \ge 0, \\ & & A\mathbf{x}^* + B\mathbf{y}^* - \mathbf{b} & = \mathbf{0}. \end{array}$$

These inequalities are equivalent to the variational inequality

$$F(\mathbf{w}) - F(\mathbf{w}^*) + (\mathbf{w} - \mathbf{w}^*)^\mathsf{T}\mathcal{J}(\mathbf{w}^*) \ge 0 \tag{2.3}$$

for all $\mathbf{w} \in \Omega$. Note that $\mathbf{w}^*$ satisfies (2.3) if and only if $\mathbf{w}^*$ is a primal-dual solution of problem (1.1). Let $\mathcal{W}^*$ denote the set of $\mathbf{w}^* \in \Omega$ satisfying (2.3).

Throughout the paper, we make the following assumptions:

(a1) The primal-dual solution set $\mathcal{W}^*$ of the problem (1.1) is nonempty.

(a2) The problem

$$\min\left\{g(\mathbf{y}) + (\beta/2)\mathbf{y}^\mathsf{T}\mathbf{B}^\mathsf{T}\mathbf{B}\mathbf{y} + \mathbf{z}^\mathsf{T}\mathbf{y} \,:\, \mathbf{y} \in \mathcal{Y}\right\}$$

has a minimizer for any $\mathbf{z} \in \mathbb{R}^{n_2}$.

(a3) For some $\nu > 0$ and $\mathcal{H} \succ \mathbf{0}$, the gradients ∇f_j satisfy the Lipschitz condition

$$\|\nabla f_j(\mathbf{x}_1) - \nabla f_j(\mathbf{x}_2)\|_{\mathcal{H}^{-1}} \le \nu \|\mathbf{x}_1 - \mathbf{x}_2\|_{\mathcal{H}} \tag{2.4}$$

for every $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ and $j = 1, 2, \ldots, N$.

By a Taylor expansion, (a3) implies that f is ν-bounded in the following sense:

$$f(\mathbf{x}_1) \le f(\mathbf{x}_2) + \langle \nabla f(\mathbf{x}_2), \mathbf{x}_1 - \mathbf{x}_2 \rangle + \frac{\nu}{2}\|\mathbf{x}_1 - \mathbf{x}_2\|_{\mathcal{H}}^2 \tag{2.5}$$

for every $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$.

3 Variational characterization

In this section we show that the iterates generated by AS-ADMM satisfy a variational inequality that is similar to (2.3), but with some additional error terms. The following lemma provides a key recursive property for the iterates $\{\mathbf{x}_t\}$ generated by `xsub`. Note that ϕ_k below is the objective function for (1.6), which `xsub` is minimizing.

Lemma 3.1 *Let us define* $\Gamma_t = 2/(t(t+1))$ *and*

$$\phi_k(\mathbf{x}) = f(\mathbf{x}) + \psi_k(\mathbf{x}), \quad \text{where} \quad \psi_k(\mathbf{x}) = \frac{1}{2}\left\|\mathbf{x} - \mathbf{x}^k\right\|_{\mathcal{M}_k}^2 + \langle \mathbf{h}^k, \mathbf{x} \rangle, \tag{3.1}$$

and $\mathbf{h}^k = -A^{\mathsf{T}}\left[\lambda^k - \beta(A\mathbf{x}^k + B\mathbf{y}^k - \mathbf{b})\right]$. *Then, for any* $\mathbf{x} \in \mathcal{X}$ *and* k *with* $\eta_k \in (0, 1/\nu)$, *we have*

$$\frac{1}{\Gamma_t}\left[\phi_k(\mathbf{x}_{t+1}) - \phi_k(\mathbf{x})\right] \le \begin{cases} \theta_1, & t = 1, \\ \frac{1}{\Gamma_{t-1}}\left[\phi_k(\mathbf{x}_t) - \phi_k(\mathbf{x})\right] + \theta_t, & t \ge 2, \end{cases} \tag{3.2}$$

where

$$\theta_t = \frac{1}{\eta_k}\left[\|\mathbf{x} - \check{\mathbf{x}}_t\|_{\mathcal{H}}^2 - \|\mathbf{x} - \check{\mathbf{x}}_{t+1}\|_{\mathcal{H}}^2\right] - \frac{t}{2}\|\mathbf{x} - \check{\mathbf{x}}_{t+1}\|_{\mathcal{M}_k}^2 \tag{3.3}$$

$$+ t\langle \boldsymbol{\delta}_t, \check{\mathbf{x}}_t - \mathbf{x} \rangle + \frac{\eta_k t^2}{4}\frac{\|\boldsymbol{\delta}_t\|_{\mathcal{H}^{-1}}^2}{(1 - \eta_k \nu)}, \quad t \ge 1, \quad \text{and}$$
$$\boldsymbol{\delta}_t = \nabla f(\widehat{\mathbf{x}}_t) - \mathbf{d}_t. \tag{3.4}$$

Proof By the updates of $\mathbf{x}_{t+1}$ and $\widehat{\mathbf{x}}_t$, we have

$$\beta_t(\check{\mathbf{x}}_{t+1} - \widehat{\mathbf{x}}_t) + (1 - \beta_t)(\mathbf{x}_t - \widehat{\mathbf{x}}_t) = \mathbf{x}_{t+1} - \widehat{\mathbf{x}}_t = \beta_t \mathbf{s}_t, \quad \mathbf{s}_t = \check{\mathbf{x}}_{t+1} - \check{\mathbf{x}}_t. \tag{3.5}$$

Since f is ν-bounded (2.5), the following relations hold due to (3.5) and the convexity of f:

$$
\begin{aligned}
f(\mathbf{x}_{t+1}) &\le f(\widehat{\mathbf{x}}_t) + \left\langle \nabla f(\widehat{\mathbf{x}}_t), \mathbf{x}_{t+1} - \widehat{\mathbf{x}}_t \right\rangle + \frac{\nu}{2}\|\mathbf{x}_{t+1} - \widehat{\mathbf{x}}_t\|_{\mathcal{H}}^2 \\
&= f(\widehat{\mathbf{x}}_t) + \left\langle \nabla f(\widehat{\mathbf{x}}_t), \beta_t(\check{\mathbf{x}}_{t+1} - \widehat{\mathbf{x}}_t) + (1-\beta_t)(\mathbf{x}_t - \widehat{\mathbf{x}}_t) \right\rangle + \frac{\nu\beta_t^2}{2}\|\mathbf{s}_t\|_{\mathcal{H}}^2 \\
&= (1-\beta_t)\left[f(\widehat{\mathbf{x}}_t) + \left\langle \nabla f(\widehat{\mathbf{x}}_t), \mathbf{x}_t - \widehat{\mathbf{x}}_t \right\rangle\right] + \beta_t R_f + \frac{\nu\beta_t^2}{2}\|\mathbf{s}_t\|_{\mathcal{H}}^2 \\
&\le (1-\beta_t) f(\mathbf{x}_t) + \beta_t R_f + \frac{\nu\beta_t^2}{2}\|\mathbf{s}_t\|_{\mathcal{H}}^2,
\end{aligned} \tag{3.6}
$$

where $R_f = f(\widehat{\mathbf{x}}_t) + \left\langle \nabla f(\widehat{\mathbf{x}}_t), \check{\mathbf{x}}_{t+1} - \widehat{\mathbf{x}}_t \right\rangle$. For any $\mathbf{x} \in \mathcal{X}$, it again follows from the convexity of f that

$$
\begin{aligned}
R_f &= f(\widehat{\mathbf{x}}_t) + \left\langle \nabla f(\widehat{\mathbf{x}}_t), \mathbf{x} - \widehat{\mathbf{x}}_t \right\rangle + \left\langle \nabla f(\widehat{\mathbf{x}}_t), \check{\mathbf{x}}_{t+1} - \mathbf{x} \right\rangle \\
&\le f(\mathbf{x}) + \left\langle \nabla f(\widehat{\mathbf{x}}_t), \check{\mathbf{x}}_{t+1} - \mathbf{x} \right\rangle.
\end{aligned} \tag{3.7}
$$

By the update formula $\mathbf{x}_{t+1} = \beta_t \check{\mathbf{x}}_{t+1} + (1-\beta_t)\mathbf{x}_t$ and the convexity of ψ_k, we have

$$
\psi_k(\mathbf{x}_{t+1}) \le \beta_t \psi_k(\check{\mathbf{x}}_{t+1}) + (1-\beta_t)\psi_k(\mathbf{x}_t). \tag{3.8}
$$

Combine (3.6), (3.7), and (3.8) with the definition of $\phi_k(\mathbf{x})$ in (3.1), to obtain

$$
\begin{aligned}
\phi_k(\mathbf{x}_{t+1}) \le& (1-\beta_t) f(\mathbf{x}_t) + \beta_t\left[f(\mathbf{x}) + \left\langle \nabla f(\widehat{\mathbf{x}}_t), \check{\mathbf{x}}_{t+1} - \mathbf{x} \right\rangle\right] + \frac{\nu\beta_t^2}{2}\|\mathbf{s}_t\|_{\mathcal{H}}^2 + \psi_k(\mathbf{x}_{t+1}) \\
\le& (1-\beta_t)\phi_k(\mathbf{x}_t) + \beta_t\left[f(\mathbf{x}) + \left\langle \nabla f(\widehat{\mathbf{x}}_t), \check{\mathbf{x}}_{t+1} - \mathbf{x} \right\rangle\right] + \frac{\nu\beta_t^2}{2}\|\mathbf{s}_t\|_{\mathcal{H}}^2 + \beta_t \psi_k(\check{\mathbf{x}}_{t+1}).
\end{aligned} \tag{3.9}
$$

In `xsub` of AS-ADMM,

$$
\check{\mathbf{x}}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \ H(\mathbf{x}) := \langle \mathbf{d}_t, \mathbf{x} \rangle + \frac{\gamma_t}{2}\|\mathbf{x} - \check{\mathbf{x}}_t\|_{\mathcal{H}}^2 + \psi_k(\mathbf{x}),
$$

where ψ_k is defined in (3.1). Since H is a quadratic with $\nabla^2 H = \gamma_t \mathcal{H} + \mathcal{M}_k$, we have

$$
H(\mathbf{x}) = H(\check{\mathbf{x}}_{t+1}) + \nabla H(\check{\mathbf{x}}_{t+1})(\mathbf{x} - \check{\mathbf{x}}_{t+1}) + \frac{1}{2}\|\mathbf{x} - \check{\mathbf{x}}_{t+1}\|_{\gamma_t\mathcal{H}+\mathcal{M}_k}^2.
$$

By the first-order optimality condition, we have $\nabla H(\check{\mathbf{x}}_{t+1})(\mathbf{x} - \check{\mathbf{x}}_{t+1}) \ge 0$ for all $\mathbf{x} \in \mathcal{X}$, which implies that $H(\mathbf{x}) \ge H(\check{\mathbf{x}}_{t+1}) + 0.5\|\mathbf{x} - \check{\mathbf{x}}_{t+1}\|_{\gamma_t\mathcal{H}+\mathcal{M}_k}^2$ for all $\mathbf{x} \in \mathcal{X}$. Rearrange this inequality to obtain

$$
\begin{aligned}
&\langle \mathbf{d}_t, \check{\mathbf{x}}_{t+1} - \mathbf{x} \rangle + \frac{\gamma_t}{2}\|\mathbf{s}_t\|_{\mathcal{H}}^2 + \psi_k(\check{\mathbf{x}}_{t+1}) \\
&\quad\le \frac{\gamma_t}{2}\|\mathbf{x} - \check{\mathbf{x}}_t\|_{\mathcal{H}}^2 + \psi_k(\mathbf{x}) - \frac{1}{2}\|\mathbf{x} - \check{\mathbf{x}}_{t+1}\|_{\gamma_t\mathcal{H}+\mathcal{M}_k}^2.
\end{aligned} \tag{3.10}
$$

Substituting $\nabla f(\widehat{\mathbf{x}}_t) = \boldsymbol{\delta}_t + \mathbf{d}_t$ in (3.9) and utilizing (3.10) yields

$$\begin{aligned}\phi_k(\mathbf{x}_{t+1}) \leq& \beta_t\left[f(\mathbf{x}) + \langle \mathbf{d}_t, \check{\mathbf{x}}_{t+1} - \mathbf{x}\rangle + \frac{\gamma_t}{2}\|\mathbf{s}_t\|_{\mathcal{H}}^2 + \psi_k(\check{\mathbf{x}}_{t+1})\right] \\ &+ (1-\beta_t)\phi_k(\mathbf{x}_t) + \beta_t\langle \boldsymbol{\delta}_t, \check{\mathbf{x}}_{t+1} - \mathbf{x}\rangle + \frac{\nu\beta_t^2}{2}\|s_t\|_{\mathcal{H}}^2 - \frac{\beta_t\gamma_t}{2}\|\mathbf{s}_t\|_{\mathcal{H}}^2 \\ \leq& \beta_t\left[f(\mathbf{x}) + \psi_k(\mathbf{x}) + \frac{\gamma_t}{2}\|\mathbf{x} - \check{\mathbf{x}}_t\|_{\mathcal{H}}^2 - \frac{1}{2}\|\mathbf{x} - \check{\mathbf{x}}_{t+1}\|_{\gamma_t\mathcal{H}+\mathcal{M}_k}^2\right] \\ &+ (1-\beta_t)\phi_k(\mathbf{x}_t) + R_d \\ =& \beta_t\left[\phi_k(\mathbf{x}) + \frac{\gamma_t}{2}\|\mathbf{x} - \check{\mathbf{x}}_t\|_{\mathcal{H}}^2 - \frac{1}{2}\|\mathbf{x} - \check{\mathbf{x}}_{t+1}\|_{\gamma_t\mathcal{H}+\mathcal{M}_k}^2\right] + (1-\beta_t)\phi_k(\mathbf{x}_t) + R_d,\end{aligned} \tag{3.11}$$

where

$$\begin{aligned}R_d =& \beta_t\langle \boldsymbol{\delta}_t, \check{\mathbf{x}}_{t+1} - \mathbf{x} - \check{\mathbf{x}}_t + \check{\mathbf{x}}_t\rangle + \frac{\nu\beta_t^2}{2}\|s_t\|_{\mathcal{H}}^2 - \frac{\beta_t\gamma_t}{2}\|\check{\mathbf{x}}_{t+1} - \check{\mathbf{x}}_t\|_{\mathcal{H}}^2 \\ =& \beta_t\langle \boldsymbol{\delta}_t, \check{\mathbf{x}}_t - \mathbf{x}\rangle + \beta_t\langle \boldsymbol{\delta}_t, \mathbf{s}_t\rangle - \frac{\beta_t\gamma_t - \nu\beta_t^2}{2}\|\mathbf{s}_t\|_{\mathcal{H}}^2 \\ \leq& \beta_t\langle \boldsymbol{\delta}_t, \check{\mathbf{x}}_t - \mathbf{x}\rangle + \beta_t\|\boldsymbol{\delta}_t\|_{\mathcal{H}^{-1}}\|s_t\|_{\mathcal{H}} - \frac{\beta_t\gamma_t - \nu\beta_t^2}{2}\|\mathbf{s}_t\|_{\mathcal{H}}^2 \\ =& \beta_t\langle \boldsymbol{\delta}_t, \check{\mathbf{x}}_t - \mathbf{x}\rangle + \beta_t\gamma_t\left[\frac{1}{\gamma_t}\|\boldsymbol{\delta}_t\|_{\mathcal{H}^{-1}}\|\mathbf{s}_t\|_{\mathcal{H}} - \frac{1 - \nu\beta_t/\gamma_t}{2}\|\mathbf{s}_t\|_{\mathcal{H}}^2\right].\end{aligned} \tag{3.12}$$

By the choice for β_t and γ_t, we have

$$1 - \frac{\nu\beta_t}{\gamma_t} = 1 - \nu\frac{2}{t+1}\frac{t\eta_k}{2} = 1 - \frac{t}{t+1}\eta_k\nu > 1 - \eta_k\nu > 0. \tag{3.13}$$

For $c > 0$, use the inequality

$$0 \leq \left(\frac{a}{2\sqrt{c}}\|\mathbf{x}\|_{\mathcal{H}^{-1}} - \sqrt{c}\|\mathbf{y}\|_{\mathcal{H}}\right)^2 = \frac{a^2}{4c}\|\mathbf{x}\|_{\mathcal{H}^{-1}}^2 + c\|\mathbf{y}\|_{\mathcal{H}}^2 - a\|\mathbf{x}\|_{\mathcal{H}^{-1}}\|\mathbf{y}\|_{\mathcal{H}}$$

to obtain

$$a\|\boldsymbol{\delta}_t\|_{\mathcal{H}^{-1}}\|\mathbf{s}_t\|_{\mathcal{H}} - c\|\mathbf{s}_t\|_{\mathcal{H}}^2 \leq \frac{a^2}{4c}\|\boldsymbol{\delta}_t\|_{\mathcal{H}^{-1}}^2. \tag{3.14}$$

Note that $c = [1 - \nu\beta_t/\gamma_t]/2 > 0$ by (3.13). Insert this choice for c and $a = 1/\gamma_t$ in (3.14), and use the resulting inequality in (3.12) to obtain

$$R_d \leq \beta_t\langle \boldsymbol{\delta}_t, \check{\mathbf{x}}_t - \mathbf{x}\rangle + \frac{\beta_t}{2(\gamma_t - \nu\beta_t)}\|\boldsymbol{\delta}_t\|_{\mathcal{H}^{-1}}^2 \leq \beta_t\langle \boldsymbol{\delta}_t, \check{\mathbf{x}}_t - \mathbf{x}\rangle + \frac{\beta_t}{2\gamma_t(1 - \nu\eta_k)}\|\boldsymbol{\delta}_t\|_{\mathcal{H}^{-1}}^2,$$

where the last inequality is due to (3.13). Combining this inequality with (3.11) gives

$$\phi_k(\mathbf{x}_{t+1}) \leq (1-\beta_t)\phi_k(\mathbf{x}_t) + \beta_t\phi_k(\mathbf{x}) + \frac{\beta_t\gamma_t}{2}\left[\|\mathbf{x}-\check{\mathbf{x}}_t\|_{\mathcal{H}}^2 - \|\mathbf{x}-\check{\mathbf{x}}_{t+1}\|_{\mathcal{H}}^2\right] \\ - \frac{\beta_t}{2}\|\mathbf{x}-\check{\mathbf{x}}_{t+1}\|_{\mathcal{M}_k}^2 + \beta_t\langle\boldsymbol{\delta}_t, \check{\mathbf{x}}_t - \mathbf{x}\rangle + \frac{\beta_t}{2\gamma_t}\frac{\|\boldsymbol{\delta}_t\|_{\mathcal{H}^{-1}}^2}{1-\nu\eta_k}.$$

Now, by subtracting $\phi_k(\mathbf{x})$ from each side of the above inequality, we obtain

$$\phi_k(\mathbf{x}_{t+1}) - \phi_k(\mathbf{x}) \leq (1-\beta_t)[\phi_k(\mathbf{x}_t) - \phi_k(\mathbf{x})] + \frac{\beta_t\gamma_t}{2}\left[\|\mathbf{x}-\check{\mathbf{x}}_t\|_{\mathcal{H}}^2 - \|\mathbf{x}-\check{\mathbf{x}}_{t+1}\|_{\mathcal{H}}^2\right] \\ - \frac{\beta_t}{2}\|\mathbf{x}-\check{\mathbf{x}}_{t+1}\|_{\mathcal{M}_k}^2 + \beta_t\langle\boldsymbol{\delta}_t, \check{\mathbf{x}}_t - \mathbf{x}\rangle + \frac{\beta_t}{2\gamma_t}\frac{\|\boldsymbol{\delta}_t\|_{\mathcal{H}^{-1}}^2}{1-\nu\eta_k}. \tag{3.15}$$

Finally, by the definitions $\Gamma_t = \frac{2}{t(t+1)}$, $\beta_t = 2/(t+1)$, and $\gamma_t = \frac{2}{t\eta_k}$, we have

$$\beta_t\gamma_t = \frac{4}{t(t+1)\eta_k}, \quad \frac{\beta_t\gamma_t}{\Gamma_t} = \frac{2}{\eta_k}, \quad \frac{\beta_t}{\Gamma_t} = t, \quad \text{and} \quad \frac{\beta_t}{\Gamma_t\gamma_t} = \frac{\eta_k t^2}{2}. \tag{3.16}$$

Dividing each side of (3.15) by Γ_t and exploiting these relations, we deduce that (3.2) holds for $t \geq 2$. Since $\Gamma_1 = \beta_1 = 1$, it also follows from (3.15) that (3.2) holds for $t = 1$. □

Based on Lemma 3.1, we are able to give a variational characterization of the AS-ADMM iterates.

Lemma 3.2 *Let $\mathcal{D}_k$ and $\boldsymbol{\delta}_t$ be as defined in* (1.6) *and* (3.4) *respectively, and suppose the $\eta_k \in (0, 1/\nu)$. Then the iterates generated by AS-ADMM satisfy*

$$f(\mathbf{x}) - f(\mathbf{x}^{k+1}) - \left\langle \mathbf{x}-\mathbf{x}^{k+1}, A^{\mathsf{T}}\widetilde{\lambda}^k \right\rangle \geq \left\langle \mathbf{x}^{k+1}-\mathbf{x}, \mathcal{D}_k(\mathbf{x}^{k+1}-\mathbf{x}^k)\right\rangle + \zeta^k, \tag{3.17}$$

for all $\mathbf{x} \in \mathcal{X}$, where

$$\widetilde{\lambda}^k = \lambda^k - \beta\left(A\mathbf{x}^{k+1} + B\mathbf{y}^k - \mathbf{b}\right), \text{ and} \tag{3.18}$$

$$\zeta^k = \frac{2}{M_k(M_k+1)}\left[\frac{1}{\eta_k}\left(\left\|\mathbf{x}-\check{\mathbf{x}}^{k+1}\right\|_{\mathcal{H}}^2 - \left\|\mathbf{x}-\check{\mathbf{x}}^k\right\|_{\mathcal{H}}^2\right) \\ - \sum_{t=1}^{M_k} t\langle\boldsymbol{\delta}_t, \check{\mathbf{x}}_t - \mathbf{x}\rangle - \frac{\eta_k}{4(1-\eta_k\nu)}\sum_{t=1}^{M_k} t^2\|\boldsymbol{\delta}_t\|_{\mathcal{H}^{-1}}^2\right]. \tag{3.19}$$

Proof Let us define $T = M_k$. Summing (3.2) over $1 \leq t \leq T$ and recalling that $\check{\mathbf{x}}^k = \check{\mathbf{x}}_1$, $\mathbf{x}^{k+1} = \mathbf{x}_{T+1}$, and $\check{\mathbf{x}}^{k+1} = \check{\mathbf{x}}_{T+1}$, we obtain

$$\begin{aligned}\frac{1}{\Gamma_T}\left[\phi_k(\mathbf{x}^{k+1})-\phi_k(\mathbf{x})\right] &\le \sum_{t=1}^{T}\theta_t \\ &= \frac{1}{\eta_k}\left[\left\|\mathbf{x}-\breve{\mathbf{x}}^k\right\|_{\mathcal{H}}^2-\left\|\mathbf{x}-\breve{\mathbf{x}}^{k+1}\right\|_{\mathcal{H}}^2\right]-\frac{1}{2}\sum_{t=1}^{T}t\|\mathbf{x}-\breve{\mathbf{x}}_{t+1}\|_{\mathcal{M}_k}^2 \\ &\quad+\sum_{t=1}^{T}t\langle\boldsymbol{\delta}_t,\breve{\mathbf{x}}_t-\mathbf{x}\rangle+\frac{\eta_k}{4(1-\eta_k\nu)}\sum_{t=1}^{T}t^2\|\boldsymbol{\delta}_t\|_{\mathcal{H}^{-1}}^2\end{aligned} \tag{3.20}$$

for any $\mathbf{x}\in\mathcal{X}$, where θ_t is defined in (3.3). Dividing the update formula $\mathbf{x}_{t+1}=\beta_t\breve{\mathbf{x}}_{t+1}+(1-\beta_t)\mathbf{x}_t$ by Γ_t and exploiting the identity $\beta_t/\Gamma_t=t$ from (3.16) yields

$$\frac{1}{\Gamma_t}\mathbf{x}_{t+1}=\frac{1}{\Gamma_{t-1}}\mathbf{x}_t+t\breve{\mathbf{x}}_{t+1}.$$

We sum over $2\le t\le T$ and recall that $\Gamma_1=\beta_1=1$ to obtain

$$\begin{aligned}\mathbf{x}^{k+1} &=\Gamma_T\left\{\frac{1}{\Gamma_1}\mathbf{x}_2+\sum_{t=2}^{T}t\breve{\mathbf{x}}_{t+1}\right\}=\Gamma_T\left\{\mathbf{x}_2-\breve{\mathbf{x}}_2+\sum_{t=1}^{T}t\breve{\mathbf{x}}_{t+1}\right\} \\ &=\Gamma_T\left\{[\beta_1\breve{\mathbf{x}}_2+(1-\beta_1)\mathbf{x}_1]-\breve{\mathbf{x}}_2+\sum_{t=1}^{T}t\breve{\mathbf{x}}_{t+1}\right\}=\sum_{t=1}^{T}(t\Gamma_T)\breve{\mathbf{x}}_{t+1}.\end{aligned} \tag{3.21}$$

Since $(t\Gamma_T)$ for $1\le t\le T$ sums to 1 and the quadratic term $\|\mathbf{z}-\mathbf{x}\|_{\mathcal{M}_k}^2$ is convex in $\mathbf{z}$, it follows from (3.21) that for any choice of $\mathbf{x}$, we have

$$\left\|\mathbf{x}^{k+1}-\mathbf{x}\right\|_{\mathcal{M}_k}^2\le\sum_{t=1}^{T}(t\Gamma_T)\|\breve{\mathbf{x}}_{t+1}-\mathbf{x}\|_{\mathcal{M}_k}^2.$$

Inserting this inequality into (3.20) gives

$$\begin{aligned}\frac{1}{\Gamma_T}\left[\phi_k(\mathbf{x}^{k+1})-\phi_k(\mathbf{x})+\frac{1}{2}\left\|\mathbf{x}^{k+1}-\mathbf{x}\right\|_{\mathcal{M}_k}^2\right] &\le\frac{1}{\eta_k}\left[\left\|\mathbf{x}-\breve{\mathbf{x}}^k\right\|_{\mathcal{H}}^2-\left\|\mathbf{x}-\breve{\mathbf{x}}^{k+1}\right\|_{\mathcal{H}}^2\right] \\ &\quad+\sum_{t=1}^{T}t\langle\boldsymbol{\delta}_t,\breve{\mathbf{x}}_t-\mathbf{x}\rangle+\frac{\eta_k}{4(1-\eta_k\nu)}\sum_{t=1}^{T}t^2\|\boldsymbol{\delta}_t\|_{\mathcal{H}^{-1}}^2.\end{aligned} \tag{3.22}$$

Now, by the definition of ϕ_k and ψ_k, we have

$$\begin{aligned}\phi_k(\mathbf{x}^{k+1})-\phi_k(\mathbf{x}) &=f(\mathbf{x}^{k+1})-f(\mathbf{x})+\psi_k(\mathbf{x}^{k+1})-\psi_k(\mathbf{x})\quad\text{and} \\ \psi_k(\mathbf{x}^{k+1})-\psi_k(\mathbf{x}) &=\left\langle\mathbf{h}^k,\mathbf{x}^{k+1}-\mathbf{x}\right\rangle+\frac{1}{2}\left[\|\mathbf{x}^{k+1}-\mathbf{x}^k\|_{\mathcal{M}_k}^2-\|\mathbf{x}-\mathbf{x}^k\|_{\mathcal{M}_k}^2\right].\end{aligned}$$

By the definition of $\mathbf{h}^k$, it follows that

$$\begin{aligned}\mathbf{h}^k &= -A^\top\left[\lambda^k - \beta\left(A\mathbf{x}^k + B\mathbf{y}^k - \mathbf{b}\right)\right]\\ &= -A^\top\left[\lambda^k - \beta\left(A\mathbf{x}^{k+1} + B\mathbf{y}^k - \mathbf{b}\right)\right] - \beta A^\top A\left(\mathbf{x}^{k+1} - \mathbf{x}^k\right)\\ &= -A^\top\widetilde{\lambda}^k - \beta A^\top A\left(\mathbf{x}^{k+1} - \mathbf{x}^k\right).\end{aligned}$$

The identity

$$(\mathbf{a}-\mathbf{b})^\top \mathcal{M}_k(\mathbf{a}-\mathbf{c}) = \frac{1}{2}\left\{\|\mathbf{a}-\mathbf{c}\|^2_{\mathcal{M}_k} - \|\mathbf{c}-\mathbf{b}\|^2_{\mathcal{M}_k} + \|\mathbf{a}-\mathbf{b}\|^2_{\mathcal{M}_k}\right\}$$

with $\mathbf{a} = \mathbf{x}^{k+1}$, $\mathbf{b} = \mathbf{x}^k$, and $\mathbf{c} = \mathbf{x}$ implies that

$$\frac{1}{2}\left[\left\|\mathbf{x}^{k+1}-\mathbf{x}^k\right\|^2_{\mathcal{M}_k} - \left\|\mathbf{x}-\mathbf{x}^k\right\|^2_{\mathcal{M}_k} + \left\|\mathbf{x}^{k+1}-\mathbf{x}\right\|^2_{\mathcal{M}_k}\right] = \left(\mathbf{x}^{k+1}-\mathbf{x}^k\right)^\top \mathcal{M}_k\left(\mathbf{x}^{k+1}-\mathbf{x}\right).$$

Insert all these relations in (3.22) and make the substitutions $T = M_k$ and $\Gamma_T = 2/(T(T+1))$ to obtain (3.17), which completes the proof. □

We now establish the following variational inequality similar to (2.3).

Lemma 3.3 *If* $\eta_k \in (0, 1/\nu)$, *then the iterates generated by AS-ADMM satisfy*

$$F(\mathbf{w}) - F(\widetilde{\mathbf{w}}^k) + \left\langle \mathbf{w}-\widetilde{\mathbf{w}}^k, \mathcal{J}(\widetilde{\mathbf{w}}^k)\right\rangle \geq \langle \mathbf{w}-\widetilde{\mathbf{w}}^k, Q_k(\mathbf{w}^k - \mathbf{w}^{k+1})\rangle + \zeta^k \quad (3.23)$$

for all $\mathbf{w} \in \Omega$, *where* ζ^k *is defined in* (3.19), $\widetilde{\lambda}^k$ *is defined in* (3.18), *and*

$$\widetilde{\mathbf{w}}^k := \begin{pmatrix}\widetilde{\mathbf{x}}^k\\ \widetilde{\mathbf{y}}^k\\ \widetilde{\lambda}^k\end{pmatrix} := \begin{pmatrix}\mathbf{x}^{k+1}\\ \mathbf{y}^{k+1}\\ \widetilde{\lambda}^k\end{pmatrix}, \quad Q_k = \begin{bmatrix}\mathcal{D}_k & & \\ & \beta B^\top B & \\ & & \frac{1}{s\beta}I\end{bmatrix}. \quad (3.24)$$

Proof Since the objective in the $\mathbf{y}$-subproblem is the sum of a nonsmooth and a smooth term, the first-order optimality condition can be expressed as

$$g(\mathbf{y}) - g(\mathbf{y}^{k+1}) + \left\langle \mathbf{y}-\mathbf{y}^{k+1}, \mathbf{p}_k\right\rangle \geq 0 \quad (3.25)$$

for all $\mathbf{y} \in \mathcal{Y}$, where $\mathbf{p}_k$ is the gradient with respect to $\mathbf{y}$, evaluated at $(\mathbf{x}^{k+1}, \mathbf{y}^{k+1})$, of the smooth term:

$$\begin{aligned}\mathbf{p}_k &= -B^\top\lambda^k + \beta B^\top\left(A\mathbf{x}^{k+1} + B\mathbf{y}^{k+1} - \mathbf{b}\right)\\ &= -B^\top\left[\lambda^k - \beta\left(A\mathbf{x}^{k+1} + B\mathbf{y}^k - \mathbf{b}\right) - \beta B\left(\mathbf{y}^{k+1} - \mathbf{y}^k\right)\right]\\ &= -B^\top\widetilde{\lambda}^k + \beta B^\top B\left(\mathbf{y}^{k+1} - \mathbf{y}^k\right).\end{aligned}$$

Here $\widetilde{\lambda}^k$ is defined in (3.18). Substituting $\mathbf{p}_k$ into (3.25) gives

$$g(\mathbf{y}) - g(\mathbf{y}^{k+1}) - \left\langle \mathbf{y}-\mathbf{y}^{k+1}, B^\top\widetilde{\lambda}^k\right\rangle \geq \beta\langle \mathbf{y}^{k+1}-\mathbf{y}, B^\top B(\mathbf{y}^{k+1}-\mathbf{y}^k)\rangle \quad (3.26)$$

for all $\mathbf{y} \in \mathcal{Y}$.

The update formula for λ^{k+1} yields the relation

$$A\mathbf{x}^{k+1} + B\mathbf{y}^{k+1} - \mathbf{b} = \frac{\lambda^k - \lambda^{k+1}}{s\beta}.$$

Take the inner product of the above equality with $\lambda - \widetilde{\lambda}^k$ to obtain

$$\langle \lambda - \widetilde{\lambda}^k, A\mathbf{x}^{k+1} + B\mathbf{y}^{k+1} - \mathbf{b}\rangle = \frac{1}{s\beta}\langle \lambda - \widetilde{\lambda}^k, \lambda^k - \lambda^{k+1}\rangle. \tag{3.27}$$

Adding (3.17), (3.26), and (3.27) yields (3.23). □

For the convergence analysis, we need to further analyze the right side of (3.23).

Corollary 3.4 *If $\eta_k \in (0, 1/\nu)$, then the iterates of AS-ADMM satisfy the following relation*:

$$\begin{aligned} &F(\mathbf{w}) - F(\widetilde{\mathbf{w}}^k) + (\mathbf{w} - \widetilde{\mathbf{w}}^k)^\top \mathcal{J}(\mathbf{w}) \\ &\geq \frac{1}{2}\left\{ \left\|\mathbf{w} - \mathbf{w}^{k+1}\right\|_{Q_k}^2 - \left\|\mathbf{w} - \mathbf{w}^k\right\|_{Q_k}^2 + \left\|\mathbf{w}^k - \widetilde{\mathbf{w}}^k\right\|_{G_k}^2 \right\} + \zeta^k, \end{aligned} \tag{3.28}$$

for any $\mathbf{w} \in \Omega$, where ζ^k is defined in (3.19), Q_k is given by (3.24) and

$$G_k = \begin{bmatrix} \mathcal{D}_k & & \\ & (1-s)\beta B^\top B & (s-1)B^\top \\ & (s-1)B & \frac{2-s}{\beta} I \end{bmatrix}. \tag{3.29}$$

Proof The identity

$$2(\mathbf{a} - \mathbf{b})^\top Q_k(\mathbf{c} - \mathbf{d}) = \|\mathbf{a} - \mathbf{d}\|_{Q_k}^2 - \|\mathbf{a} - \mathbf{c}\|_{Q_k}^2 + \|\mathbf{c} - \mathbf{b}\|_{Q_k}^2 - \|\mathbf{b} - \mathbf{d}\|_{Q_k}^2$$

with the choices $\mathbf{a} = \mathbf{w}$, $\mathbf{b} = \widetilde{\mathbf{w}}^k$, $\mathbf{c} = \mathbf{w}^k$, and $\mathbf{d} = \mathbf{w}^{k+1}$ gives

$$\begin{aligned} &\langle \mathbf{w} - \widetilde{\mathbf{w}}^k, Q_k(\mathbf{w}^k - \mathbf{w}^{k+1})\rangle \\ &= \frac{1}{2}\left\{ \left\|\mathbf{w} - \mathbf{w}^{k+1}\right\|_{Q_k}^2 - \left\|\mathbf{w} - \mathbf{w}^k\right\|_{Q_k}^2 + \left\|\mathbf{w}^k - \widetilde{\mathbf{w}}^k\right\|_{Q_k}^2 - \left\|\mathbf{w}^{k+1} - \widetilde{\mathbf{w}}^k\right\|_{Q_k}^2 \right\}. \end{aligned} \tag{3.30}$$

The update formula for λ^{k+1}, together with the definition of $\widetilde{\lambda}^k$ in (3.18), yield the relation

$$\lambda^{k+1} - \lambda^k = s\beta B(\mathbf{y}^k - \mathbf{y}^{k+1}) - s(\lambda^k - \widetilde{\lambda}^k). \tag{3.31}$$

Hence, we have

$$\begin{aligned} \lambda^{k+1} - \widetilde{\lambda}^k =& \lambda^{k+1} - \lambda^k + \lambda^k - \widetilde{\lambda}^k \\ =& s\beta B(\mathbf{y}^k - \mathbf{y}^{k+1}) + (1-s)(\lambda^k - \widetilde{\lambda}^k). \end{aligned}$$

Since the only nonzero component of $\mathbf{w}^{k+1} - \widetilde{\mathbf{w}}^k$ is the $\lambda^{k+1} - \widetilde{\lambda}^k$ component, we have

$$\left\|\mathbf{w}^{k+1} - \widetilde{\mathbf{w}}^k\right\|_{Q_k}^2 = \frac{1}{s\beta}\left\|s\beta B(\mathbf{y}^k - \mathbf{y}^{k+1}) + (1-s)(\lambda^k - \widetilde{\lambda}^k)\right\|^2.$$

With this substitution, it follows that

$$\left\|\mathbf{w}^k - \widetilde{\mathbf{w}}^k\right\|_{Q_k}^2 - \left\|\mathbf{w}^{k+1} - \widetilde{\mathbf{w}}^k\right\|_{Q_k}^2 = \left\|\mathbf{w}^k - \widetilde{\mathbf{w}}^k\right\|_{G_k}^2.$$

Combine this identity with (3.30), Lemma 3.3, and the skew symmetry of $\mathcal{J}$ to complete the proof. □

Comparing (2.3) with (3.28), the convergence of AS-ADMM can be analyzed relatively easily if the matrix G_k given by (3.29) is positive semidefinite, which ensures that $\|\mathbf{w}^k - \widetilde{\mathbf{w}}^k\|_{G_k}^2 \geq 0$. However, G_k is not always positive semidefinite when $s \in (0, (1+\sqrt{5})/2]$. Consequently, the convergence analysis requires the following lower bound for $\|\mathbf{w}^k - \widetilde{\mathbf{w}}^k\|_{G_k}^2$.

Lemma 3.5 *The iterates of AS-ADMM satisfy*

$$\begin{aligned}\left\|\mathbf{w}^k - \widetilde{\mathbf{w}}^k\right\|_{G_k}^2 \geq & (2-s)\beta\left\|A\mathbf{x}^{k+1} + B\mathbf{y}^{k+1} - \mathbf{b}\right\|^2 + \left\|\mathbf{x}^k - \mathbf{x}^{k+1}\right\|_{\mathcal{D}_k}^2 \\ & - (1-s)^2\beta\left\|A\mathbf{x}^k + B\mathbf{y}^k - \mathbf{b}\right\|^2,\end{aligned} \tag{3.32}$$

where G_k and $\mathcal{D}_k$ are defined in (3.29) *and* (1.6), *respectively.*

Proof By the definition of G_k in (3.29) and direct calculation, we have

$$\begin{aligned}\frac{1}{\beta}\left\|\mathbf{w}^k - \widetilde{\mathbf{w}}^k\right\|_{G_k}^2 = & \frac{1}{\beta}\left\|\mathbf{x}^k - \mathbf{x}^{k+1}\right\|_{\mathcal{D}_k}^2 + (1-s)\left\|B\left(\mathbf{y}^k - \mathbf{y}^{k+1}\right)\right\|^2 \\ & + \frac{2(s-1)}{\beta}\left(\lambda^k - \widetilde{\lambda}^k\right)^\top B\left(\mathbf{y}^k - \mathbf{y}^{k+1}\right) + \frac{2-s}{\beta^2}\left\|\lambda^k - \widetilde{\lambda}^k\right\|^2.\end{aligned}$$

Since $\widetilde{\lambda}^k - \lambda^k = -\beta(A\mathbf{x}^{k+1} + B\mathbf{y}^{k+1} - \mathbf{b}) + \beta B(\mathbf{y}^{k+1} - \mathbf{y}^k)$, it follows that

$$\begin{aligned}\frac{1}{\beta}\left\|\mathbf{w}^k - \widetilde{\mathbf{w}}^k\right\|_{G_k}^2 = & \frac{1}{\beta}\left\|\mathbf{x}^k - \mathbf{x}^{k+1}\right\|_{\mathcal{D}_k}^2 + (2-s)\left\|A\mathbf{x}^{k+1} + B\mathbf{y}^{k+1} - \mathbf{b}\right\|^2 + \\ & \left\|B\left(\mathbf{y}^k - \mathbf{y}^{k+1}\right)\right\|^2 + 2\left(A\mathbf{x}^{k+1} + B\mathbf{y}^{k+1} - \mathbf{b}\right)^\top B\left(\mathbf{y}^k - \mathbf{y}^{k+1}\right).\end{aligned} \tag{3.33}$$

Choosing $\mathbf{y} = \mathbf{y}^k$ in the first-order optimality condition (3.25), we have

$$g(\mathbf{y}^k) - g(\mathbf{y}^{k+1}) + \left\langle B\left(\mathbf{y}^k - \mathbf{y}^{k+1}\right), -\lambda^k + \beta\left(A\mathbf{x}^{k+1} + B\mathbf{y}^{k+1} - \mathbf{b}\right)\right\rangle \geq 0.$$

Similarly, choosing $\mathbf{y} = \mathbf{y}^{k+1}$ in the first-order optimality condition (3.25) at the $(k-1)$-th iteration, we have

$$g(\mathbf{y}^{k+1}) - g(\mathbf{y}^k) + \left\langle B\left(\mathbf{y}^{k+1} - \mathbf{y}^k\right), -\lambda^{k-1} + \beta\left(A\mathbf{x}^k + B\mathbf{y}^k - \mathbf{b}\right) \right\rangle \geq 0.$$

Adding these two inequalities and substituting $\lambda^k = \lambda^{k-1} - s\beta\left(A\mathbf{x}^k + B\mathbf{y}^k - \mathbf{b}\right)$, we have

$$\begin{aligned}
&\left(A\mathbf{x}^{k+1} + B\mathbf{y}^{k+1} - \mathbf{b}\right)^{\mathsf{T}} B\left(\mathbf{y}^k - \mathbf{y}^{k+1}\right) \\
&\quad \geq (1-s)\left(A\mathbf{x}^k + B\mathbf{y}^k - \mathbf{b}\right)^{\mathsf{T}} B\left(\mathbf{y}^k - \mathbf{y}^{k+1}\right) \\
&\quad \geq -\frac{1}{2}\left((1-s)^2\left\|A\mathbf{x}^k + B\mathbf{y}^k - \mathbf{b}\right\|^2 + \left\|B\left(\mathbf{y}^k - \mathbf{y}^{k+1}\right)\right\|^2\right),
\end{aligned}$$

where the last inequality comes from the relation $\mathbf{x}^{\mathsf{T}}\mathbf{y} \geq -\frac{1}{2}\left[c\|\mathbf{x}\|^2 + \frac{1}{c}\|\mathbf{y}\|^2\right]$ for any $c > 0$. Inserting this lower bound for the last term in (3.33) yields (3.32). □

4 Convergence analysis

In this section, we analyze the convergence properties of AS-ADMM. The following theorem explores how closely an ergodic average of the iterates satisfies the first-order optimality condition (2.3).

Theorem 4.1 *Suppose that for some integers $\kappa \geq 0$ and $T > 0$ and for all $k \in [\kappa, \kappa + T]$, the following conditions are satisfied*:

(A1) $\mathcal{D}_k \succeq \mathcal{D}_{k+1} \succeq \mathbf{0}$ *and* $\mathbb{E}\left[\|\boldsymbol{\delta}_t\|^2_{\mathcal{H}^{-1}}\right] \leq \sigma^2$ *for some* $\sigma > 0$, *independent of t and the iteration number k, where* $\boldsymbol{\delta}_t$ *is defined in* (3.4).

(A2) $\eta_k \in (0, 1/(2\nu)]$, *where* $\nu > 0$ *is the Lipschitz constant given in* (*a*3), *and the sequence* $\{\eta_k M_k(M_k + 1)\}$ *is nondecreasing*.

Then for every $\mathbf{w} \in \Omega$, *we have*

$$\begin{aligned}
\mathbb{E}\left[F(\mathbf{w}_T) - F(\mathbf{w}) + (\mathbf{w}_T - \mathbf{w})^{\mathsf{T}}\mathcal{J}(\mathbf{w})\right] \leq \frac{1}{2(1+T)}\Bigg\{\sigma^2 \sum_{k=\kappa}^{\kappa+T} \eta_k M_k \\
+ \|\mathbf{w} - \mathbf{w}^\kappa\|^2_{Q_\kappa} + \beta(1-s)^2\|A\mathbf{x}^\kappa + B\mathbf{y}^\kappa - \mathbf{b}\|^2 + \frac{4}{M_\kappa(M_\kappa+1)\eta_\kappa}\|\mathbf{x} - \mathbf{x}^\kappa\|^2_{\mathcal{H}}\Bigg\}
\end{aligned} \tag{4.1}$$

where

$$\mathbf{w}_T := \frac{1}{1+T}\sum_{k=\kappa}^{\kappa+T}\widetilde{\mathbf{w}}^k. \tag{4.2}$$

Proof Since $s \in (0, (1+\sqrt{5})/2]$ and $\beta > 0$ in AS-ADMM, we have

$$\xi_1 := \beta((2-s)-(1-s)^2) \geq 0 \quad \text{and} \quad \xi_2 := \beta(1-s)^2 \geq 0.$$

The inequality (3.32) can be rearranged into the form

$$\begin{aligned}\left\|\mathbf{w}^k - \widetilde{\mathbf{w}}^k\right\|_{G_k}^2 \geq & \left\|\mathbf{x}^k - \mathbf{x}^{k+1}\right\|_{\mathcal{D}_k}^2 + \xi_1\left\|A\mathbf{x}^{k+1} + B\mathbf{y}^{k+1} - \mathbf{b}\right\|^2 \\ & + \xi_2\left(\left\|A\mathbf{x}^{k+1} + B\mathbf{y}^{k+1} - \mathbf{b}\right\|^2 - \left\|A\mathbf{x}^k + B\mathbf{y}^k - \mathbf{b}\right\|^2\right).\end{aligned} \tag{4.3}$$

By (A1) and the fact that $s > 0$, it follows that Q_k in (3.24) satisfies $Q_k \succeq Q_{k+1} \succeq \mathbf{0}$. Substituting (4.34.3) into (3.28) and utilizing the relation $Q_k \succeq Q_{k+1}$, we have

$$\begin{aligned}& F(\widetilde{\mathbf{w}}^k) - F(\mathbf{w}) + (\widetilde{\mathbf{w}}^k - \mathbf{w})^\top \mathcal{J}(\mathbf{w}) \\ & \leq \frac{1}{2}\left\{\left\|\mathbf{w} - \mathbf{w}^k\right\|_{Q_k}^2 - \left\|\mathbf{w} - \mathbf{w}^{k+1}\right\|_{Q_{k+1}}^2\right\} \\ & \quad + \frac{\xi_2}{2}\left\{\left\|A\mathbf{x}^k + B\mathbf{y}^k - \mathbf{b}\right\|^2 - \left\|A\mathbf{x}^{k+1} + B\mathbf{y}^{k+1} - \mathbf{b}\right\|^2\right\} \\ & \quad - \frac{1}{2}\left\{\left\|\mathbf{x}^k - \mathbf{x}^{k+1}\right\|_{\mathcal{D}_k}^2 + \xi_1\left\|A\mathbf{x}^{k+1} + B\mathbf{y}^{k+1} - \mathbf{b}\right\|^2\right\} - \zeta^k,\end{aligned} \tag{4.4}$$

where ζ^k is defined in (3.19).

Sum the inequality (4.4) over k between κ and $\kappa + T$. Notice that the sum associated with the first two bracketed terms are telescoping series while the sum associated with the third bracketed expression is negative and can be neglected. Thus by the definition of $\mathbf{w}_T$ in (4.2), we obtain

$$\begin{aligned}& \sum_{k=\kappa}^{\kappa+T} F(\widetilde{\mathbf{w}}^k) - (1+T)\left\{F(\mathbf{w}) + \left(\mathbf{w}_T - \mathbf{w}\right)^\top \mathcal{J}(\mathbf{w})\right\} \\ & \quad \leq \frac{1}{2}\left\{\|\mathbf{w} - \mathbf{w}^\kappa\|_{Q_\kappa}^2 + \xi_2\|A\mathbf{x}^\kappa + B\mathbf{y}^\kappa - \mathbf{b}\|^2\right\} - \sum_{k=\kappa}^{\kappa+T}\zeta^k.\end{aligned} \tag{4.5}$$

It further follows from the convexity of F that

$$F(\mathbf{w}_T) \leq \frac{1}{1+T}\sum_{k=\kappa}^{\kappa+T} F(\widetilde{\mathbf{w}}^k). \tag{4.6}$$

Dividing (4.5) by $T+1$ and utilizing (4.6), we obtain

$$
\begin{aligned}
&F(\mathbf{w}_T) - F(\mathbf{w}) + (\mathbf{w}_T - \mathbf{w})^\top \mathcal{J}(\mathbf{w}) \\
&\quad \le \frac{1}{1+T}\left[\frac{1}{2}\left\{\|\mathbf{w}-\mathbf{w}^\kappa\|_{Q_\kappa}^2 + \xi_2\|A\mathbf{x}^\kappa + B\mathbf{y}^\kappa - \mathbf{b}\|^2 - 2\sum_{k=\kappa}^{\kappa+T}\zeta^k\right\}\right].
\end{aligned} \tag{4.7}
$$

Let us now focus on the ζ^k summation in (4.7). By assumption (A2), the sequence $\{M_k(M_k+1)\eta_k\}$ is nondecreasing for $k \in [\kappa, \kappa+T]$; hence, by the telescoping nature of the sum, we have

$$
\begin{aligned}
&\sum_{k=\kappa}^{\kappa+T} \frac{2}{M_k(M_k+1)\eta_k}\left(\|\mathbf{x}-\breve{\mathbf{x}}^k\|_{\mathcal{H}}^2 - \|\mathbf{x}-\breve{\mathbf{x}}^{k+1}\|_{\mathcal{H}}^2\right) \\
&\quad \le \sum_{k=\kappa}^{\kappa+T}\left(\frac{2\|\mathbf{x}-\breve{\mathbf{x}}^k\|_{\mathcal{H}}^2}{M_k(M_k+1)\eta_k} - \frac{2\|\mathbf{x}-\breve{\mathbf{x}}^{k+1}\|_{\mathcal{H}}^2}{M_{k+1}(M_{k+1}+1)\eta_{k+1}}\right) \le \frac{2\|\mathbf{x}-\mathbf{x}^\kappa\|_{\mathcal{H}}^2}{M_\kappa(M_\kappa+1)\eta_\kappa}.
\end{aligned} \tag{4.8}
$$

For $\boldsymbol{\delta}_t$ defined in (3.4), we have

$$
\boldsymbol{\delta}_t = \nabla f(\widehat{\mathbf{x}}_t) - \mathbf{d}_t = \nabla f(\widehat{\mathbf{x}}_t) - \nabla f_{\xi_t}(\widehat{\mathbf{x}}_t) - \mathbf{e}_t.
$$

Since the random variable $\xi_t \in \{1, 2, \dots, N\}$ is chosen with uniform probability and $\mathbb{E}[\mathbf{e}_t] = \mathbf{0}$, it follows that $\mathbb{E}[\boldsymbol{\delta}_t] = \mathbf{0}$. Also, since $\boldsymbol{\delta}_t$ only depends on the index ξ_t while $\breve{\mathbf{x}}_t$ depends on $\xi_{t-1}, \xi_{t-2}, \cdots$, we have

$$
\mathbb{E}\left[\langle \boldsymbol{\delta}_t, \breve{\mathbf{x}}_t - \mathbf{x}\rangle\right] = \mathbf{0}.
$$

By (A1), we have $\mathbb{E}(\|\boldsymbol{\delta}_t\|_{\mathcal{H}^{-1}}^2) \le \sigma^2$. Since $M_k \ge 1$, it follows that

$$
\mathbb{E}\left[\sum_{t=1}^{M_k} t^2\|\boldsymbol{\delta}_t\|_{\mathcal{H}^{-1}}^2\right] \le \frac{\sigma^2 M_k(M_k+1)(2M_k+1)}{6} \le M_k^2(M_k+1)\left(\frac{\sigma^2}{2}\right).
$$

Combining these bounds for the terms in ζ^k defined in (3.19) with the condition $\eta_k \le 1/(2\nu)$ in (A2) yields

$$
-\mathbb{E}\left[\sum_{k=\kappa}^{\kappa+T}\zeta^k\right] \le \frac{2\|\mathbf{x}-\mathbf{x}^\kappa\|_{\mathcal{H}}^2}{M_\kappa(M_\kappa+1)\eta_\kappa} + \frac{\sigma^2}{2}\sum_{k=\kappa}^{\kappa+T}\eta_k M_k.
$$

To complete the proof, apply the expectation operator to (4.7) and substitute this bound for the ζ^k term. □

Analogous to the definition (4.2), we define

$$
\lambda_T = \frac{1}{1+T}\sum_{k=\kappa}^{\kappa+T}\widetilde{\lambda}^k, \quad \mathbf{x}_T = \frac{1}{1+T}\sum_{k=\kappa}^{\kappa+T}\widetilde{\mathbf{x}}^k \quad \text{and} \quad \mathbf{y}_T = \frac{1}{1+T}\sum_{k=\kappa}^{\kappa+T}\widetilde{\mathbf{y}}^k. \tag{4.9}
$$

Theorem 4.1 yields a convergence result for AS-ADMM when we make the following choice for η_k and M_k in (4.1):

$$\eta_k = \min\left\{\frac{c_1}{M_k(M_k+1)}, c_2\right\} \quad \text{and} \quad M_k = \max\left\{\lceil c_3 k^{\varrho}\rceil, M\right\}, \tag{4.10}$$

where $c_1, c_2, c_3 > 0$ and $\varrho \geq 1$ are constants, and $M > 0$ is a given integer. Choose k large enough so that $M_k = \lceil c_3 k^{\varrho}\rceil$. As k tends to infinity, M_k tends to infinity and η_k tends to zero. Choose k larger if necessary to ensure that $\eta_k = c_1/(M_k(M_k+1)) \leq 1/(2\nu)$, where ν is the Lipschitz constant in (A2). Since $\eta_k M_k(M_k+1) = c_1$, a constant, and $\eta_k \in (0, 1/2\nu]$ for $k \geq \kappa$, condition (A2) of Theorem 4.1 is satisfied for this choice of k.

Theorem 4.2 *If* (A1) *of Theorem* 4.1 *holds for all k and the parameters η_k and M_k are chosen according to* (4.10), *then for $\mathbf{w}^* \in \mathcal{W}^*$, we have*

$$\left|\mathbb{E}\left[F(\mathbf{w}_T)\right] - F(\mathbf{w}^*)\right| = E_{\varrho}(T) = \mathbb{E}\left[\|A\mathbf{x}_T + B\mathbf{y}_T - \mathbf{b}\|\right], \tag{4.11}$$

where $E_{\varrho}(T) = \mathcal{O}(1/T)$ for $\varrho > 1$ and $E_{\varrho}(T) = \mathcal{O}(T^{-1}\log T)$ for $\varrho = 1$.

Proof Suppose that κ is chosen by the procedure explained beneath (4.10), which ensures that condition (A2) of Theorem 4.1 is satisfied for all $k \geq \kappa$. By assumption, (A1) holds. Hence, the conclusion (4.1) of Theorem 4.1 holds.

First, let us analyze the left side of (4.1). By the definition of $\mathcal{J}$ (see (2.1)), it follows that

$$(\mathbf{w}_T - \mathbf{w})^{\mathsf{T}}\mathcal{J}(\mathbf{w}) = \lambda_T^{\mathsf{T}}(A\mathbf{x} + B\mathbf{y} - \mathbf{b}) - \lambda^{\mathsf{T}}\left(A\mathbf{x}_T + B\mathbf{y}_T - \mathbf{b}\right). \tag{4.12}$$

For any $\mathbf{w}^* = (\mathbf{x}^*, \mathbf{y}^*, \lambda^*) \in \mathcal{W}^*$, let us choose $\mathbf{w} = (\mathbf{x}^*, \mathbf{y}^*, \boldsymbol{\mu}_T)$, where $\boldsymbol{\mu}_T = \lambda^* + \overline{\boldsymbol{\mu}}_T$ and $\overline{\boldsymbol{\mu}}_T$ is a unit vector chosen so that

$$(\overline{\boldsymbol{\mu}}_T)^{\mathsf{T}}\left[A\mathbf{x}_T + B\mathbf{y}_T - \mathbf{b}\right] = -\|A\mathbf{x}_T + B\mathbf{y}_T - \mathbf{b}\|. \tag{4.13}$$

Note that $\mathbf{x}_T$ and $\mathbf{y}_T$ are stochastic variables. In equations such as (4.12) and (4.13), the vectors $\mathbf{x}_T$ and $\mathbf{y}_T$ represent a specific realization of the stochastic vectors. For each possible realization of $\mathbf{x}_T$ and $\mathbf{y}_T$, $\overline{\boldsymbol{\mu}}_T$ should be chosen so that (4.13) holds. Thus the choice for $\overline{\boldsymbol{\mu}}_T$ depends on the realization of $\mathbf{x}_T$ and $\mathbf{y}_T$. Since $A\mathbf{x}^* + B\mathbf{y}^* = \mathbf{b}$, the choice $\mathbf{w} = (\mathbf{x}^*, \mathbf{y}^*, \boldsymbol{\mu}_T)$ in (4.12) yields

$$(\mathbf{w}_T - \mathbf{w})^{\mathsf{T}}\mathcal{J}(\mathbf{w}) = -\|A\mathbf{x}_T + B\mathbf{y}_T - \mathbf{b}\| - (\lambda^*)^{\mathsf{T}}(A\mathbf{x}_T + B\mathbf{y}_T - \mathbf{b}). \tag{4.14}$$

Since $F(\mathbf{w}) = F(\mathbf{w}^*)$ when $\mathbf{w} = (\mathbf{x}^*, \mathbf{y}^*, \boldsymbol{\mu}_T)$, (4.14) yields

$$\begin{aligned} &F(\mathbf{w}_T) - F(\mathbf{w}) + (\mathbf{w}_T - \mathbf{w})^{\mathsf{T}}\mathcal{J}(\mathbf{w}) \\ &\quad = F(\mathbf{w}_T) - F(\mathbf{w}^*) - (\lambda^*)^{\mathsf{T}}(A\mathbf{x}_T + B\mathbf{y}_T - \mathbf{b}) + \|A\mathbf{x}_T + B\mathbf{y}_T - \mathbf{b}\|. \end{aligned} \tag{4.15}$$

By the variational inequality (2.3) with $\mathbf{w} = \mathbf{w}_T$, we have

$$\begin{aligned} &F(\mathbf{w}_T) - F(\mathbf{w}^*) + (\mathbf{w}_T - \mathbf{w}^*)^{\mathsf{T}} \mathcal{J}(\mathbf{w}^*) \\ &\quad = F(\mathbf{w}_T) - F(\mathbf{w}^*) - (\boldsymbol{\lambda}^*)^{\mathsf{T}} \left[A\mathbf{x}_T + B\mathbf{y}_T - \mathbf{b}\right] \geq 0. \end{aligned} \tag{4.16}$$

Use this inequality in (4.15) to obtain the lower bound

$$F(\mathbf{w}_T) - F(\mathbf{w}) + (\mathbf{w}_T - \mathbf{w})^{\mathsf{T}} \mathcal{J}(\mathbf{w}) \geq \|A\mathbf{x}_T + B\mathbf{y}_T - \mathbf{b}\| \tag{4.17}$$

for $\mathbf{w} = (\mathbf{x}^*, \mathbf{y}^*, \boldsymbol{\mu}_T)$.

Next, let us analyze the right side of (4.1) when $\mathbf{w} = (\mathbf{x}^*, \mathbf{y}^*, \boldsymbol{\mu}_T)$. By the choice (4.10) for η_k and M_k, we see that

$$\sum_{k=\kappa}^{\kappa+T} \eta_k M_k \leq \sum_{k=\kappa}^{\kappa+T} c_1/(1 + c_3 k^{\varrho}).$$

This sum is $\mathcal{O}(1)$ if $\varrho > 1$, while it is $\mathcal{O}(\log T)$ if $\varrho = 1$. Since κ was chosen so that $\eta_\kappa M_\kappa (M_\kappa + 1) = c_1$, it follows that the other terms in brackets on the right side of (4.1) are all $\mathcal{O}(1)$. Consequently, we have

$$\mathbb{E}\left[F(\mathbf{w}_T) - F(\mathbf{w}) + (\mathbf{w}_T - \mathbf{w})^{\mathsf{T}} \mathcal{J}(\mathbf{w})\right] = E_{\varrho}(T).$$

We combine this upper bound $E_{\varrho}(T)$ with the lower bound (4.17) and take expectation to obtain

$$\mathbb{E}\left[\|A\mathbf{x}_T + B\mathbf{y}_T - \mathbf{b}\|\right] = E_{\varrho}(T), \tag{4.18}$$

which establishes the right side of (4.11).

The optimality condition (4.16) implies that

$$F(\mathbf{w}_T) - F(\mathbf{w}^*) \geq -\|\boldsymbol{\lambda}^*\| \, \|A\mathbf{x}_T + B\mathbf{y}_T - \mathbf{b}\|.$$

Again, take expectation and utilize (4.18) to obtain

$$\mathbb{E}\left[F(\mathbf{w}_T) - F(\mathbf{w}^*)\right] \geq -E_{\varrho}(T). \tag{4.19}$$

Similar to our observation above, the right side of (4.1) is bounded by $E_{\varrho}(T)$ when $\mathbf{w} = \mathbf{w}^*$:

$$\mathbb{E}\left[F(\mathbf{w}_T) - F(\mathbf{w}^*) + (\mathbf{w}_T - \mathbf{w}^*)^{\mathsf{T}} \mathcal{J}(\mathbf{w}^*)\right] \leq E_{\varrho}(T). \tag{4.20}$$

Since $(\mathbf{w}_T - \mathbf{w}^*)^{\mathsf{T}} \mathcal{J}(\mathbf{w}^*) = -(\boldsymbol{\lambda}^*)^{\mathsf{T}} \left[A\mathbf{x}_T + B\mathbf{y}_T - \mathbf{b}\right]$, it follows that

$$\mathbb{E}\left[F(\mathbf{w}_T) - F(\mathbf{w}^*)\right] \leq E_{\varrho}(T) + \mathbb{E}\left[(\boldsymbol{\lambda}^*)^{\mathsf{T}} \left[A\mathbf{x}_T + B\mathbf{y}_T - \mathbf{b}\right]\right] = E_{\varrho}(T),$$

where the last equality is by (4.18). Combine this with (4.19) to obtain the left side of (4.11), which completes the proof. □

We now have the following remarks.

Remark 4.1 The objective error and the constraint violation converge to zero in expectation due to (4.11); however, this does not imply the convergence or boundedness of the ergodic iterates. If there exists $c > 0$ such that $\mathcal{D}_k \succeq c\mathbf{I}$, then the iterates $(\mathbf{x}_k, \mathbf{y}_k, \lambda_k)$ are bounded in expectation, and under a strong convexity assumption, the ergodic iterates converge in expectation (see Appendix).

Remark 4.2 In AS-ADMM, it was required that $\mathcal{D}_k = \mathcal{M}_k - \beta A^\top A \succeq \mathbf{0}$, however, in Theorem 4.1, the proof only requires that $\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|^2_{\mathcal{D}_k} \geq 0$ so that the third bracketed expression in (4.4) can be dropped while preserving the inequality. Hence, for numerical efficiency, at any iteration k, we could set $\mathcal{M}_k = \rho_k \mathbf{I}$ and then adjust ρ_k based on an underestimate $\beta\delta_2^k/\delta_1^k$ for the largest eigenvalue of $\beta A^\top A$, where

$$\delta_1^k = \left\|\mathbf{x}^k - \mathbf{x}^{k-1}\right\|^2 \quad \text{and} \quad \delta_2^k = \left\|A(\mathbf{x}^k - \mathbf{x}^{k-1})\right\|^2.$$

In particular, given parameters ρ_0 and $\rho_{\min} > 0$, and $\eta > 1$, we multiply $\rho_{\min}$ by η in any iteration where $\rho_{k-1} < \beta\delta_2^k/\delta_1^k$, and in each iteration, we set

$$\rho_k = \max\{\rho_{\min},\ \beta\delta_2^k/\delta_1^k\}.$$

The increase in $\rho_{\min}$ can only happen a finite number of times since $\rho_{k-1} \geq \beta\delta_2^k/\delta_1^k$ whenever $\rho_{k-1} \geq \beta\|A^\top A\|$; in fact, the increase in $\rho_{\min}$ can happen at most $\lceil \log_\eta \frac{\beta\|A^\top A\|}{\rho_0} \rceil$ times. Hence, for k large enough, ρ_k, $\mathcal{M}_k$, and $\mathcal{D}_k$ are all unchanged, and $\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|^2_{\mathcal{D}_k} \geq 0$. Related techniques were first used in [8] in the context of a line search.

Remark 4.3 Theorem 4.1 holds under the assumption that $\left\{\eta_k M_k(M_k + 1)\right\}$ is nondecreasing. We now point out that Theorem 4.1 can be reformulated so as to hold when $\left\{\eta_k M_k(M_k + 1)\right\}$ is *nonincreasing* if $\mathcal{X}$ is a bounded set. Let $\mathcal{N}_\mathcal{X}$ denote the diameter of $\mathcal{X}$:

$$\mathcal{N}_\mathcal{X} = \sup\{\|\mathbf{x}_1 - \mathbf{x}_2\|_\mathcal{H} : \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}\}. \tag{4.21}$$

If $\mathcal{N}_\mathcal{X}$ is finite and $\left\{\eta_k M_k(M_k + 1)\right\}$ is nonincreasing for $k \in [\kappa, \kappa + T]$, then the term (4.8) in the proof of Theorem 4.1 has the following bound: For any $\mathbf{x} \in \mathcal{X}$, we have

$$\begin{aligned}
&\sum_{k=\kappa}^{\kappa+T} \frac{2}{M_k(M_k+1)\eta_k}\left(\|\mathbf{x}-\breve{\mathbf{x}}^k\|_{\mathcal{H}}^2 - \|\mathbf{x}-\breve{\mathbf{x}}^{k+1}\|_{\mathcal{H}}^2\right) \\
&= \frac{2}{M_\kappa(M_\kappa+1)\eta_\kappa}\|\mathbf{x}-\mathbf{x}^\kappa\|_{\mathcal{H}}^2 - \frac{2}{M_T(M_T+1)\eta_T}\|\mathbf{x}-\breve{\mathbf{x}}^{T+1}\|_{\mathcal{H}}^2 \\
&\quad - \sum_{k=\kappa}^{\kappa+T-1}\left(\frac{2}{M_k(M_k+1)\eta_k} - \frac{2}{M_{k+1}(M_{k+1}+1)\eta_{k+1}}\right)\|\mathbf{x}-\breve{\mathbf{x}}^{k+1}\|_{\mathcal{H}}^2 \\
&\le \frac{2}{M_\kappa(M_\kappa+1)\eta_\kappa}\mathcal{N}_{\mathcal{X}}^2 - \sum_{k=\kappa}^{\kappa+T-1}\left(\frac{2}{M_k(M_k+1)\eta_k} - \frac{2}{M_{k+1}(M_{k+1}+1)\eta_{k+1}}\right)\mathcal{N}_{\mathcal{X}}^2 \\
&= \frac{2}{M_T(M_T+1)\eta_T}\mathcal{N}_{\mathcal{X}}^2.
\end{aligned} \tag{4.22}$$

By using (4.22), a bound similar to (4.1) can be established.

Remark 4.4 If $N = 1$ and $\mathbf{e}_t = \mathbf{0}$, then AS-ADMM is a deterministic ADMM with multiple accelerated gradient steps to solve the $\mathbf{x}$-subproblem inexactly, and the expectation operator can be removed from (4.11). If additional assumptions hold, such as $s \in (0, (1+\sqrt{5})/2)$ (the open interval), $\mathcal{D}_k \succeq c\mathbf{I}$ for some $c > 0$, and B has full column rank, then the iterates $\mathbf{w}^k$ are uniformly bounded and convergent to some $\mathbf{w}^* \in \mathcal{W}^*$.

5 Incremental sampling of stochastic gradient with variance reduction

In this section, we discuss AS-ADMM algorithm with incremental sampling of the stochastic gradient. These techniques can potentially reduce the number of stochastic gradient steps and can be beneficial when the subproblems for computing the stochastic gradient step is expensive. Suppose that at the t-th inner iteration of subroutine **xsub** in the k-th outer iteration of AS-ADMM, when calculating the stochastic gradient of function f, we randomly select an index sample set

$$U_t \subset \{1, 2, \cdots, N\} \quad \text{of size} \quad |U_t| = m_k \le N$$

with uniform probability. We define

$$\hat{\mathbf{g}}_t = \frac{1}{m_k}\sum_{i\in U_t}\nabla f_i(\hat{\mathbf{x}}_t) \quad \text{and} \quad \mathbf{e}_t = \nabla f(\overline{\mathbf{x}}^k) - \frac{1}{m_k}\sum_{i\in U_t}\nabla f_i(\overline{\mathbf{x}}^k)$$

for some choice of $\overline{\mathbf{x}}^k \in \Omega$. Since the elements of U_t are chosen with uniform probability, $\mathbb{E}\left[\mathbf{e}_t\right] = \mathbf{0}$. Also, we define

$$\mathbf{d}_t = \hat{\mathbf{g}}_t + \mathbf{e}_t = \frac{1}{m_k}\sum_{i\in U_t}\left[\nabla f_i(\hat{\mathbf{x}}_t) - \nabla f_i(\overline{\mathbf{x}}^k)\right] + \nabla f(\overline{\mathbf{x}}^k), \tag{5.1}$$

and $\boldsymbol{\delta}_t = \nabla f(\widehat{\mathbf{x}}_t) - \mathbf{d}_t$. Again, since U_t is chosen with uniform probability, we have $\mathbb{E}[\boldsymbol{\delta}_t] = \mathbf{0}$. Moreover, if the diameter $\mathcal{N}_\chi$ of $\mathcal{X}$, defined in (4.21), is finite, then the variance of $\boldsymbol{\delta}_t$ has the following bound:

$$\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{\delta}_t\|^2_{\mathcal{H}^{-1}}\right] =&\mathbb{E}\left[\left\|\frac{1}{m_k}\sum_{i\in U_t}\left[\nabla f_i(\widehat{\mathbf{x}}_t) - \nabla f_i(\overline{\mathbf{x}}^k)\right] + \nabla f(\overline{\mathbf{x}}^k) - \nabla f(\widehat{\mathbf{x}}_t)\right\|^2_{\mathcal{H}^{-1}}\right] \\
=&\frac{N-m_k}{m_k(N-1)}\mathbb{E}_\xi\left[\left\|\nabla f_\xi(\overline{\mathbf{x}}^k) - \nabla f_\xi(\widehat{\mathbf{x}}_t) - [\nabla f(\overline{\mathbf{x}}^k) - \nabla f(\widehat{\mathbf{x}}_t)]\right\|^2_{\mathcal{H}^{-1}}\right] \\
=&\frac{N-m_k}{m_k(N-1)}\left\{\mathbb{E}_\xi\left[\left\|\nabla f_\xi(\overline{\mathbf{x}}^k) - \nabla f_\xi(\widehat{\mathbf{x}}_t)\right\|^2_{\mathcal{H}^{-1}}\right] - \left\|\nabla f(\overline{\mathbf{x}}^k) - \nabla f(\widehat{\mathbf{x}}_t)\right\|^2_{\mathcal{H}^{-1}}\right\} \\
\le&\frac{1}{m_k}\mathbb{E}_\xi\left[\left\|\nabla f_\xi(\overline{\mathbf{x}}^k) - \nabla f_\xi(\widehat{\mathbf{x}}_t)\right\|^2_{\mathcal{H}^{-1}}\right] = \frac{1}{m_k N}\sum_{j=1}^N\left\|\nabla f_j(\overline{\mathbf{x}}^k) - \nabla f_j(\widehat{\mathbf{x}}_t)\right\|^2_{\mathcal{H}^{-1}} \\
\le&\frac{1}{m_k N}\sum_{j=1}^N \nu^2\left\|\overline{\mathbf{x}}^k - \widehat{\mathbf{x}}_t\right\|^2_{\mathcal{H}} = \frac{\nu^2}{m_k}\left\|\overline{\mathbf{x}}^k - \widehat{\mathbf{x}}_t\right\|^2_{\mathcal{H}} \le \frac{\nu^2\mathcal{N}^2_{\mathcal{X}}}{m_k},
\end{aligned} \tag{5.2}$$

where the second equality follows from [12, Page 183], $\mathbb{E}_\xi[\cdot]$ is taken with respect to a random drawing of $\xi \in \{1, 2, \dots, N\}$ with uniform probability, and ν is the Lipschitz constant for the f_j given in (a3). Consequently, $\mathbb{E}[\|\boldsymbol{\delta}_t\|^2_{\mathcal{H}^{-1}}] \le \sigma^2$ with $\sigma = \nu\mathcal{N}_{\mathcal{X}}/\sqrt{m_k}$.

On the other hand, if we obtain information during the computation by choosing $\overline{\mathbf{x}}^k$ such that $\|\overline{\mathbf{x}}^k - \widehat{\mathbf{x}}_t\|$ is small, then we see from (5.2) that the variance of $\boldsymbol{\delta}_t$ could be reduced significantly. Note that the full gradient $\nabla f(\overline{\mathbf{x}}^k)$ is only calculated in the outer iteration. In our numerical experiments, we choose $\overline{\mathbf{x}}^k$ to be the ergodic mean of the iterates at certain iterations. Furthermore, under the conditions of Theorem 4.1, we can show that

$$\begin{aligned}
\mathbb{E}\left[F(\mathbf{w}_T) - F(\mathbf{w}) + (\mathbf{w}_T - \mathbf{w})^\top \mathcal{J}(\mathbf{w})\right] \le \frac{1}{2(1+T)}\left\{(\nu\mathcal{N}_{\mathcal{X}})^2\sum_{k=\kappa}^{\kappa+T}\frac{\eta_k M_k}{m_k}\right. \\
\left. + \|\mathbf{w} - \mathbf{w}^\kappa\|^2_Q + \xi_2\|A\mathbf{x}^\kappa + B\mathbf{y}^\kappa - \mathbf{b}\|^2 + \frac{4}{M_\kappa(M_\kappa+1)\eta_\kappa}\|\mathbf{x} - \mathbf{x}^\kappa\|^2_{\mathcal{H}}\right\},
\end{aligned} \tag{5.3}$$

where $\xi_2 = \beta(1-s)^2$.

Suppose we choose the parameters

$$\eta_k = \eta \in \left(0, \frac{1}{2\nu}\right], \quad M_k = M \quad \text{and} \quad m_k = \min\left\{\lceil c(1+k)^\varrho\rceil, N\right\}, \tag{5.4}$$

where $M \ge 1$ is an integer and $c > 0$ and $\varrho \ge 1$ are real scalars. In the case that the total data size N is large, with $m_k < N$, we can deduce from (5.3) that

$$\mathbb{E}\left[F(\mathbf{w}_T) - F(\mathbf{w}) + (\mathbf{w}_T - \mathbf{w})^\top \mathcal{J}(\mathbf{w})\right] = \mathcal{O}\left(\frac{1}{T}\left(1 + \sum_{k=0}^{T} \frac{1}{(1+k)^\rho}\right)\right). \tag{5.5}$$

Hence, the convergence rate with the incremental sampling of the stochastic gradient will be the same as the rate (4.11) of AS-ADMM with parameter setting (4.10).

In addition, it can be observed that with the parameter settings (5.4), the total number of sample gradients used in the inner iteration when N is large and $m_k < N$ is given by

$$\sum_{k=0}^{T} M_k m_k = \mathcal{O}\left(\sum_{k=0}^{T} (1+k)^\rho\right),$$

which is on the same order as that of AS-ADMM with parameter settings (4.10). However, the stepsize parameter η_k in (5.4) can be larger than that in (4.10), and the total number of stochastic gradient steps performed in AS-ADMM is

$$\sum_{k=0}^{T} M_k = MT,$$

which can be significantly smaller than the total number of stochastic gradient steps $\mathcal{O}\left(\sum_{k=0}^{T} k^\rho\right)$ performed by AS-ADMM with parameter settings (4.10); this would greatly reduce the computational cost in the case that the subproblem for calculating the stochastic gradient step is expensive.

6 Linearized AS-ADMM

When B is a relatively complicated matrix, a closed-form solution of the $\mathbf{y}$-subproblem may not exist, even when g is simple. A common approach, in this case, is to modify the $\mathbf{y}$-subproblem by linearizing its quadratic penalty term so that a closed-form solution may exist, similar to what is done in `xsub` for the $\mathbf{x}$-subproblem. The corresponding proximal term is

$$\frac{1}{2}\left\|\mathbf{y} - \mathbf{y}^k\right\|^2_{\tau\mathbf{I} - \beta B^\top B},$$

where $\tau > 0$ is large enough that $\tau\mathbf{I} \succeq \beta B^\top B$. This proximal term, when added to the penalty term in the $\mathbf{y}$-subproblem, will annihilate the penalty term $(\beta/2)\|B\mathbf{y}\|^2$. For $\tau > 0$, the $\mathbf{y}$-subproblem reduces to the following proximal mapping:

$$\mathbf{y}^{k+1} = \mathbf{prox}_{g,\tau}(\mathbf{q}^k) := \arg\min_{\mathbf{y}\in\mathcal{Y}} \left\{g(\mathbf{y}) + (\tau/2)\|\mathbf{y} - \mathbf{q}^k\|^2\right\},$$

where $\mathbf{q}^k = \mathbf{y}^k - B^\top[\beta(A\mathbf{x}^{k+1} + B\mathbf{y}^k - \mathbf{b}) - \lambda^k]/\tau$. Note that the assumption (a2) is not required in this case since strong convexity of the y-subproblem implies a unique global solution. The complexity analysis when the $\mathbf{y}$-subproblem is linearized is the same as that of the original AS-ADMM given in Theorem 4.2 for appropriate choices of the parameters. It may be possible to relax the constraint $\tau\mathbf{I} \succeq \beta B^\top B$ using ideas from [7, 33].

7 Numerical experiments

This section provides numerical experiments to investigate the performance of AS-ADMM.

7.1 Test problem and parameter settings

Given a number of training samples $\{(\mathbf{a}_j, b_j)\}_{j=1}^N$ where $\mathbf{a}_j \in \mathbb{R}^l$ and $b_j \in \{-1, 1\}$, we solve the following generalized lasso problem (called the graph-guided fused lasso model):

$$\min_{\mathbf{x}} \frac{1}{N} \sum_{j=1}^N f_j(\mathbf{x}) + \mu \|A\mathbf{x}\|_1,$$

where $f_j(\mathbf{x}) = \log\left(1 + \exp(-b_i \mathbf{a}_i^\top \mathbf{x})\right)$ denotes the logistic loss function on the feature-label pair $(\mathbf{a}_j, b_j)$, N is the data size (usually large), $\mu > 0$ is a given regularization parameter, and $A = \mathbf{I}$ or $A = [\mathbf{G};\mathbf{I}]$, where $\mathbf{G}$ is obtained from a sparse inverse covariance estimation given in [13]. Although the generalized lasso problem is used to compare the ADMM algorithms, this specific problem is potentially solved more efficiently using a stochastic primal-dual algorithm such as the one developed in [5].

By introducing an auxiliary variable $\mathbf{y}$, the above problem can be reduced to a special case of problem (1.1):

$$\min_{\mathbf{x},\mathbf{y}} \{F(\mathbf{x}, \mathbf{y}) = \frac{1}{N} \sum_{j=1}^N f_j(\mathbf{x}) + \mu \|\mathbf{y}\|_1 \;:\; A\mathbf{x} - \mathbf{y} = \mathbf{0}\}. \tag{7.1}$$

We use AS-ADMM to solve (7.1); the closed-form solutions of the subproblems are

$$\left\{ \begin{array}{ll} \breve{\mathbf{x}}_{t+1} & = \left[\gamma_t \mathcal{H} + \mathcal{M}_k\right]^{-1} \left[\gamma_t \mathcal{H} \breve{\mathbf{x}}_t + \mathcal{M}_k \mathbf{x}^k - \mathbf{d}_t - \mathbf{h}^k\right], \\ \mathbf{y}^{k+1} & = \mathrm{Shrink}\left(\frac{\mu}{\beta}, A\mathbf{x}^{k+1} - \frac{\lambda^k}{\beta}\right). \end{array} \right. \tag{7.2}$$

Here, Shrink$(\cdot, \cdot)$ denotes the so-called soft shrinkage operator, which can be evaluated using the built-in MATLAB function "`wthresh`".

The datasets of Table 1 and the Lipschitz constants of f are taken from the LIBSVM website. The parameter settings used in AS-ADMM are as follows. The stepsize s is taken as $s = 1.618$ (approximately its largest value), the penalty parameter is $\beta = 0.04$, and the values of η_k and M_k are given by (4.10) with $c_1 = 1/\nu$, $c_2 = 1/(2\nu)$, $c_3 = 0.01$, $\varrho = 1.1$, and $M = 200$. The choice for c_2 ensures that the condition $\eta_k \in (0, 1/2\nu]$ of Theorem 4.1 is satisfied from the start of the iterations, while c_1 was chosen so that it scaled in the same way as c_2. A small value was used

Table 1 Real-world datasets and regularization parameters used in the experiments

Dataset	Number of samples	Dimensionality	μ
a9a	32,561	123	1e-5
ijcnn1	49,990	23	1e-5
w8a	49,749	300	1e-5
mnist	11,791	784	1e-5

for c_3 so that the growth of M_k would be delayed, and $M = 200$ since N is on the order of tens of thousands, and we wanted at least several hundred inner iterations. The matrices $\mathcal{M}_k$ are updated adaptively by the strategy in Remark 4.2 with initial values $\rho_0 = 1$, $\eta = 1.1$, $\rho_{\min} = 10^{-5}$; these were the same parameter values that seemed to work well in [19, 20] when we solved image reconstruction problems. In particular, $\eta = 1.1$ so that the lower bound $\rho_{\min}$ for the largest eigenvalue would grow slowly. We set $\mathcal{H} := \sigma \mathbf{I}$ with $\sigma = 2 \times 10^{-5}$. Thus both $\mathcal{M}_k$ and $\mathcal{H}$ are diagonal matrices. We set the regularization parameter $\mu = 10^{-5}$ since ASVRG-ADMM set all the regularization parameters to 10^{-5}. We found that it is expensive and unnecessary to calculate one full gradient at each outer iteration for reducing the variance of the stochastic gradient. Hence, in numerical experiments, we only do the variance reduction when the number of inner iterations M_k is larger than the dimension of the $\mathbf{x}$-variable. More precisely, at the k-th outer iteration of AS-ADMM, in the t-th inner iteration of subroutine **xsub**, we set

$$\mathbf{e}_t = \begin{cases} \nabla f(\mathbf{x}_{k-1}) - \nabla f_{\xi_t}(\mathbf{x}_{k-1}), & \text{if } M_k > n_1, \\ \mathbf{0}, & \text{otherwise}, \end{cases}$$

where $\mathbf{x}_k$ is the ergodic mean of the $\mathbf{x}$-iterates. All comparison algorithms are implemented in MATLAB R2018a (64-bit) with the same starting point $(\mathbf{x}^0, \mathbf{y}^0, \lambda^0) = (\mathbf{0}, \mathbf{0}, \mathbf{0})$, and all experiments are performed on a PC with Windows 10 operating system, with an Intel i7-8700K CPU, and with 16GB RAM.

7.2 Comparative experiments

In this section, we compare the following algorithms for solving problem (7.1) using the four data sets of Table 1:

- Accelerated stochastic ADMM, Algorithm 1.1 (AS-ADMM).
- Stochastic ADMM ([28], STOC-ADMM).
- Accelerated variance reduced stochastic ADMM ([26, Alg. 2], ASVRG-ADMM).
- Accelerated Linearized ADMM with $\chi = 1$([29, Alg. 2], ALP-ADMM).
- The classic ADMM [16] with f linearized (L-ADMM):

$$\mathbf{x}^{k+1} = \arg\min_{\mathbf{x} \in \mathcal{R}^l} \left\langle \nabla f(\mathbf{x}^k), \mathbf{x} - \mathbf{x}^k \right\rangle + \frac{\nu}{2}\left\|\mathbf{x} - \mathbf{x}^k\right\|^2 + \frac{\beta}{2}\left\|A\mathbf{x} - \mathbf{y}^k - \lambda^k/\beta\right\|^2.$$

We did not compare AS-ADMM with many other stochastic algorithms mentioned in this paper since their performance has been shown in the literature to be worse than that of ASVRG-ADMM. We compare AS-ADMM with STOC-ADMM [28] since STOC-ADMM only applies one stochastic gradient step to solve the **x**-subproblem in each outer iteration, while AS-ADMM applies a multiple number of accelerated gradient steps as determined by the theory, and ASVRG-ADMM preforms a fixed number $m = N/200$ inner iterations. Note that both ALP-ADMM and L-ADMM are deterministic ADMM-type algorithms using the full gradient.

In comparing algorithms, we plot Opt_err, the maximum of the relative objective error (Obj_err) and constraint violation (Equ_err), versus CPU time in seconds, where

$$\text{Obj\_err} = \frac{|F(\mathbf{x},\mathbf{y}) - F^*|}{\max\{F^*, 1\}} \quad \text{and} \quad \text{Equ\_err} = \|A\mathbf{x} - \mathbf{y}\|.$$

Here F^* is the approximate optimal objective function value obtained by running AS-ADMM for more than 10 minutes. Experimental results are averaged over 10 successive runs for the three stochastic algorithms. For AS-ADMM, we plot the error associated with the iterates over the first 1/3 of the total CPU time budget, followed by the error associated with the ergodic iterates over the last 2/3 of the budget. Note that the convergence theory describes the error for $k \geq \kappa$, where κ is the iteration number where the assumptions in the analysis are satisfied. An advantage of AS-ADMM is that the algorithm is completely adaptive, and the user does not need to provide Lipschitz constants or eigenvalue bounds; and in theory, convergence is guaranteed. Nonetheless, the initial iterates may be less reliable than later iterates.

Figures 1, 2 and 3 show results for the data sets `a9a`, `ijcnn1` and `w8a` and $A = \mathbf{I}$, while Figure 4 is the corresponding plot for the `mnist` data set with the more complicated choice $A = [\mathbf{G};\mathbf{I}]$ explained in subsect. 7.1. We can see that both AS-ADMM and ASVRG-ADMM perform better than STOC-ADMM [28], where only one stochastic gradient step is used in each iteration to solve the **x**

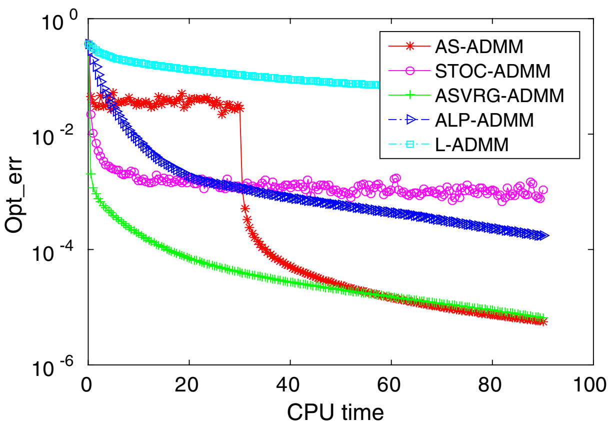

Fig. 1 Comparison of Opt _ err vs CPU time for Problem (7.1) and the `a9a` dataset

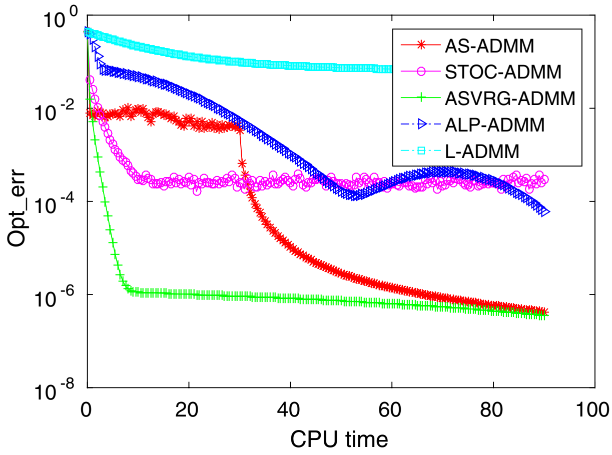

Fig. 2 Comparison of Opt _ err vs CPU time for Problem (7.1) and the ijcnn1 dataset

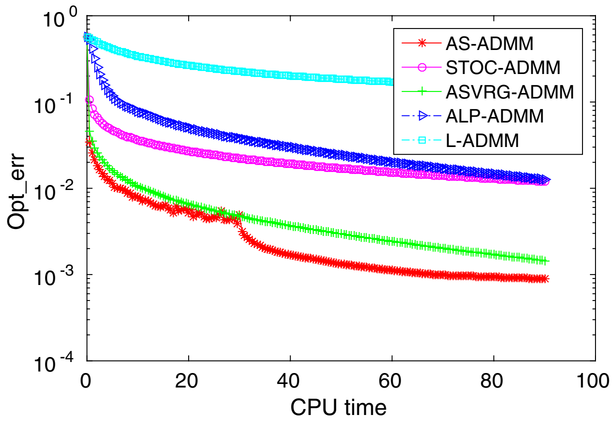

Fig. 3 Comparison of Opt _ err vs CPU time for Problem (7.1) and the w8a dataset

-subproblem. We also see that AS-ADMM and ASVRG-ADMM achieve comparable performance on the lasso problems for the first three data sets, while AS-ADMM performs significantly better than ASVRG-ADMM on the last data set, where the constraint is more complex. Note that Opt _ err for AS-ADMM has a big drop at around 1/3 of the CPU time budget, the point where we start to utilize the ergodic iterates when reporting the objective value. Observe that both stochastic algorithms, AS-ADMM and ASVRG-ADMM, perform significantly better than the deterministic methods ALP-ADMM and L-ADMM, while the accelerated nature of ALP-ADMM leads to much better performance than that of the classic L-ADMM.

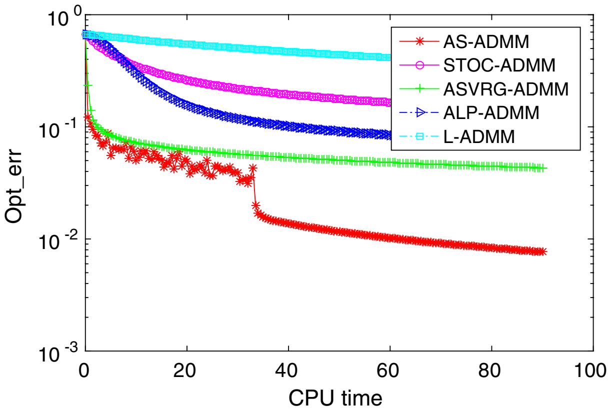

Fig. 4 Comparison of Opt _ err vs CPU time for Problem (7.1) and the `mnist` dataset

8 Conclusion

We have developed an accelerated stochastic ADMM for solving a type of regularized empirical risk minimization problem arising in machine and statistical learning. We also discussed incremental sampling techniques, which are potentially beneficial when the subproblems for computing the stochastic gradient step are expensive, and a variant of AS-ADMM that was achieved by linearizing the **y**-subproblem. The proposed algorithm AS-ADMM combines both the variance reduction technique and an accelerated gradient method for fast convergence. Using a unified variational analysis, the expected objective error and constraint violation for ergodic iterates are $\mathcal{O}(1/k)$ or $\mathcal{O}(k^{-1}\log k)$, depending on the choice of parameters. Numerical experiments on group lasso problems using well-established stochastic and deterministic ADMM algorithms show that AS-ADMM can be very effective for solving data mining and machine learning problems with large data sets. With stronger assumptions, bounds for the iterates in expectation, as well as linear convergence results, are established.

Appendix: Additional properties of the iterates

In the appendix, we derive additional properties of AS-ADMM which involve new assumptions that do not appear in the previous analysis.

Iteration bounds

When $\mathcal{D}_k := \mathcal{M}_k - \beta A^{\mathsf{T}}A$ is uniformly positive definite, the expectation of the (non-ergodic) iterates $\mathbf{w}^k = (\mathbf{x}^k, \mathbf{y}^k, \lambda^k)$ is uniformly bounded.

Proposition 9.1 *If* (A1) *and* (A2) *in Theorem* 4.1 *are satisfied, the parameters* η_k *and* M_k *are chosen according to* (4.10) *with* $\varrho > 1$ *and there exists*

$c > 0$ *such that* $\mathcal{D}_k \geq c\mathbf{I}$ *for every* k, *then* $\mathbb{E}\left[\|\mathbf{w}^k\|^2\right]$ *is bounded uniformly in* k; *moreover, if* $s \in (0, (1+\sqrt{5})/2)$, *then* $\mathbb{E}\left[\|A\mathbf{x}^k + B\mathbf{y}^k - \mathbf{b}\|^2\right]$ *tends to* 0, *while* $\mathbb{E}\left[\|A\mathbf{x}^k + B\mathbf{y}^k - \mathbf{b}\|^2\right]$ *is uniformly bounded if* $s = (1+\sqrt{5})/2$.

Proof Insert $\mathbf{w} = \mathbf{w}^* \in \mathcal{W}^*$ in (4.4) and utilize (2.3) to obtain

$$\left\|\mathbf{w}^{k+1} - \mathbf{w}^*\right\|^2_{Q_{k+1}} - \left\|\mathbf{w}^k - \mathbf{w}^*\right\|^2_{Q_k} \leq \xi_2(\gamma_k - \gamma_{k+1}) - \xi_1\gamma_{k+1} - 2\zeta^k, \tag{9.1}$$

where $\gamma_k = \|A\mathbf{x}^k + B\mathbf{y}^k - \mathbf{b}\|^2$, ζ^k is defined in (3.19) with $\mathbf{x} = \mathbf{x}^*$, $\xi_2 := \beta(1-s)^2 \geq 0$, and $\xi_1 := \beta((2-s) - (1-s)^2) > 0$ if $s \in (0, (1+\sqrt{5})/2)$. Let E_k be defined by

$$E_k = \left\|\mathbf{w}^k - \mathbf{w}^*\right\|^2_{Q_k} + \xi_2\gamma_k + \frac{4}{M_k(M_k+1)\eta_k}\left\|\check{\mathbf{x}}^k - \mathbf{x}^*\right\|^2_{\mathcal{H}}.$$

Since $\{M_k(M_k+1)\eta_k\}$ is nondecreasing, it follows from (9.1) and the definition of ζ^k that

$$E_{k+1} - E_k + \xi_1\gamma_{k+1} \leq \frac{4}{M_k(M_k+1)}\left[\sum_{t=1}^{M_k} t\langle\boldsymbol{\delta}_t, \check{\mathbf{x}}_t - \mathbf{x}^*\rangle + \frac{\eta_k}{4(1-\eta_k\nu)}\sum_{t=1}^{M_k} t^2\left\|\boldsymbol{\delta}_t\right\|^2_{\mathcal{H}^{-1}}\right].$$

As shown in the proof of Theorem 4.1, the expectation of the first term on the right side vanishes, while the expectation of the second term is bounded by $\sigma^2\eta_k M_k$. Hence, we have

$$\mathbb{E}\left[E_{k+1} - E_k + \xi_1\gamma_{k+1}\right] \leq \sigma^2\eta_k M_k.$$

We sum this inequality over $k \in [\kappa, j)$ to obtain

$$\mathbb{E}\left[E_j\right] + \xi_1\sum_{k=\kappa+1}^{j}\mathbb{E}\left[\gamma_k\right] \leq \mathbb{E}\left[E_\kappa\right] + \sigma^2\sum_{k=\kappa}^{j-1}\eta_k M_k.$$

Since $\varrho > 1$, it follows that $\eta_k M_k$ is summable when (4.10) holds, and $\mathbb{E}\left[E_j\right]$ is bounded, uniformly in j. Moreover, when $s \in (0, (1+\sqrt{5})/2)$, the bound for the sum of $\mathbb{E}\left[\gamma_k\right]$ over $k \geq \kappa$ implies that $\mathbb{E}\left[\gamma_k\right]$ tends to zero. If $s = (1+\sqrt{5})/2$, then $\xi_2 > 0$ and the uniform bound for $\mathbb{E}\left[E_j\right]$ implies that $\mathbb{E}\left[\gamma_j\right]$ is uniformly bounded. □

Convergence of ergodic iterates under strong convexity

We now show that an error bound such as (4.11) implies convergence of the ergodic iterates in expectation when strong convexity holds.

Proposition 9.2 *Suppose* (4.11) *holds. If either* f *and* g *are strongly convex or* f *is strongly convex and the columns of* B *are linearly independent, then*

$$\mathbb{E}\left[\|\mathbf{x}_T - \mathbf{x}^*\|^2 + \|\mathbf{y}_T - \mathbf{y}^*\|^2\right] = E_\varrho(T),$$

where $(\mathbf{x}^*, \mathbf{y}^*)$ *is the unique solution of* (1.1) *and* $E_\varrho(T)$ *is defined in Theorem* 4.2.

Proof If f is strongly convex with modulus α, then it follows from strong convexity, the first-order optimality conditions for a stationary point $(\mathbf{x}^*, \mathbf{y}^*, \lambda^*)$, and the inclusion $\mathbf{x}_T \in \mathcal{X}$ that

$$\begin{aligned} f(\mathbf{x}_T) - f(\mathbf{x}^*) \geq& \nabla f(\mathbf{x}^*)(\mathbf{x}_T - \mathbf{x}^*) + \frac{\alpha}{2}\|\mathbf{x}_T - \mathbf{x}^*\|^2 \\ \geq& (\lambda^*)^\mathsf{T} A(\mathbf{x}_T - \mathbf{x}^*) + \frac{\alpha}{2}\|\mathbf{x}_T - \mathbf{x}^*\|^2. \end{aligned}$$

If g is also strongly convex, with the same modulus α, then the same inequality holds, but with $\mathbf{x}$ replaced by $\mathbf{y}$, A replaced by B, and gradient replaced by subgradient. Together, these inequalities yield

$$F(\mathbf{w}_T) - F(\mathbf{w}^*) \geq (\lambda^*)^\mathsf{T}\left[A\mathbf{x}_T + B\mathbf{y}_T - \mathbf{b}\right] + \frac{\alpha}{2}\left[\|\mathbf{x}_T - \mathbf{x}^*\|^2 + \|\mathbf{y}_T - \mathbf{y}^*\|^2\right].$$

Taking expectations and utilizing (4.11) gives

$$\begin{aligned} \mathbb{E}\left[\|\mathbf{x}_T - \mathbf{x}^*\|^2 + \|\mathbf{y}_T - \mathbf{y}^*\|^2\right] \leq& \frac{2}{\alpha}\mathbb{E}\left[\|A\mathbf{x}_T + B\mathbf{y}_T - \mathbf{b}\|\right]\|\lambda^*\| + E_\varrho(T) \\ =& E_\varrho(T). \end{aligned} \tag{9.2}$$

On the other hand, if g is only convex, not strongly convex, then the term $\|\mathbf{y}_T - \mathbf{y}^*\|^2$ in this last inequality is lost. But if the columns of B were linearly independent, then the equation error can be manipulated as follows:

$$\begin{aligned} \mathbb{E}\left[\|A\mathbf{x}_T + B\mathbf{y}_T - \mathbf{b}\|\right] =& \mathbb{E}\left[\|A(\mathbf{x}_T - \mathbf{x}^*) + B(\mathbf{y}_T - \mathbf{y}^*)\|\right] \\ \geq& \mathbb{E}\left[\|B(\mathbf{y}_T - \mathbf{y}^*)\|\right] - \mathbb{E}\left[\|A(\mathbf{x}_T - \mathbf{x}^*)\|\right]. \end{aligned}$$

Thus we have

$$\mathbb{E}\left[\|B(\mathbf{y}_T - \mathbf{y}^*)\|\right] \leq E_\varrho(T) + \mathbb{E}\left[\|A(\mathbf{x}_T - \mathbf{x}^*)\|\right].$$

Hence, the bound for $\mathbb{E}\left[\|\mathbf{x}_T - \mathbf{x}^*\|^2\right]$ from (9.2) and the independence of the columns in B imply again that $\mathbb{E}\left[\|\mathbf{y}_T - \mathbf{y}^*\|^2\right] = E_\varrho(T)$. □

Linear convergence of iterates under strong convexity

In this section, it is proved that AS-ADMM is linearly convergent when both f and g are strongly convex. The analysis requires a geometric growth rate for the inner iterations, similar to the geometric growth rate employed in [36] for the analysis of a much different ADMM. In detail, the linear convergence result needs the following assumptions:

(L1) f and g are strongly convex with modulus $\alpha > 0$, and both f and g have Lipschitz continuous gradients, with Lipschitz constant ν.
(L2) The sets $\mathcal{X} = \mathbb{R}^{n_1}$ and $\mathcal{Y} = \mathbb{R}^{n_2}$.
(L3) The stepsize $s \in (0, (1+\sqrt{5})/2)$ and for some c_1 and $c_2 > 0$ and for all k, we have

$$c_1 \mathbf{I} \preceq \mathcal{D}_{k+1} \preceq \mathcal{D}_k := \mathcal{M}_k - \beta A^{\mathsf{T}} A \preceq c_2 \mathbf{I}.$$

Moreover, $\mathbb{E}\left[\|\boldsymbol{\delta}_t\|^2_{\mathcal{H}^{-1}}\right] \le \sigma^2$ for some $\sigma > 0$, independent of t and the iteration number k, where $\boldsymbol{\delta}_t$ is defined in (3.4).
(L4) For some $\theta > 0$, we have

$$M_k = \lceil (1+\theta)^2 M_{k-1} (\|\breve{\mathbf{x}}^k\|^2 + 1) \rceil \quad \text{and} \quad \eta_k = (1+\theta)^{-k}/M_k.$$

Here $M_0 > 0$ is an integer chosen large enough that $\eta_0 < 1/(2\nu)$, where ν is the Lipschitz constant given in (a3).

By (L4), we have:

(L4a) $M_k \ge (1+\theta)^{2k} M_0$,
(L4b) $M_k \eta_k = (1+\theta)^{-k}$,
(L4c) $(\|\breve{\mathbf{x}}^k\|^2 + 1)/[M_k(M_k+1)\eta_k] \le (1+\theta)^{-k}/M_0$,
(L4d) $M_k(M_k+1)\eta_k$ is nondecreasing.

The condition (L4a) follows from the first equation in (L4) after erasing $\|\breve{\mathbf{x}}^k\|^2$ and the ceiling operators, and replacing the equality by an inequality. The equation (L4b) is simply the definition of η_k. After substituting for η_k, we see that (L4c) is equivalent to

$$\frac{\|\breve{\mathbf{x}}^k\|^2 + 1}{M_k + 1} \le \frac{(1+\theta)^{-2k}}{M_0}. \tag{9.3}$$

Erasing the ceiling operation in (L4) and replacing the equality by an inequality gives

$$\frac{\|\breve{\mathbf{x}}^k\|^2 + 1}{M_k} \le \frac{(1+\theta)^{-2}}{M_{k-1}}.$$

Replace M_k by $M_k + 1$ and use the lower bound $M_{k-1} \ge (1+\theta)^{2k-2} M_0$ of (L4a) to obtain (9.3) which is equivalent to (L4c). To establish (L4d), we need to show that the sequence $a_k = M_k(M_k+1)\eta_k$ is nondecreasing, or equivalently, that $a_{k+1} \ge a_k$. By (L4b), this is equivalent to

$$M_{k+1} + 1 \ge (1+\theta)(M_k + 1). \tag{9.4}$$

Erasing the ceiling operator and the $\|\breve{\mathbf{x}}^k\|^2$ in (L4), it follows that $M_{k+1} \ge (1+\theta)^2 M_k$, which implies that

$$M_{k+1} + 1 \ge (1+\theta)^2 M_k + 1 = (1+\theta)M_k + (\theta + \theta^2)M_k + 1 \ge (1+\theta)M_k + 1 + \theta$$

since $M_k \geq 1$. Hence, (9.4) holds and the sequence a_k is nondecreasing.

Proposition 9.3 *If* (*L*1)–(*L*4) *hold, then there exists* $c > 0$ *and* $0 < \tau < 1$ *such that*

$$\mathbb{E}\left[\|\mathbf{x}^k - \mathbf{x}^*\|^2 + \|\mathbf{y}^k - \mathbf{y}^*\|^2 + \|\lambda^k - \overline{\lambda}^k\|^2\right] \leq c\tau^k$$

for all $k \geq 0$, *where* $\overline{\lambda}$ *denotes the projection of* λ *onto* $\mathbf{\Lambda}^*$, *the set of all multipliers associated with the solution* $(\mathbf{x}^*, \mathbf{y}^*)$ *of* (1.1).

Proof Throughout the proof, c denotes a generic positive constant, independent of k, which typically has different values in different equations. If $\overline{\mathbf{w}}^k := (\mathbf{x}^*, \mathbf{y}^*, \overline{\lambda}^k)$, then by the strong convexity assumption and the first-order optimality conditions for $(\mathbf{x}^*, \mathbf{y}^*)$, we have

$$\begin{aligned} F(\widetilde{\mathbf{w}}^k) - F(\overline{\mathbf{w}}^k) + (\widetilde{\mathbf{w}}^k - \overline{\mathbf{w}}^k)^{\mathsf{T}}\mathcal{J}(\overline{\mathbf{w}}^k) =& F(\widetilde{\mathbf{w}}^k) - F(\overline{\mathbf{w}}^k) - (A\mathbf{x}^{k+1} + B\mathbf{y}^{k+1} - \mathbf{b})^{\mathsf{T}}\overline{\lambda}^k \\ \geq& \frac{\alpha}{2}\left(\|\mathbf{x}^{k+1} - \mathbf{x}^*\|^2 + \|\mathbf{y}^{k+1} - \mathbf{y}^*\|^2\right). \end{aligned}$$

Hence, by (4.4) with $\mathbf{w} = \overline{\mathbf{w}}^k$ and $\gamma_k = \|A\mathbf{x}^k + B\mathbf{y}^k - \mathbf{b}\|^2$, we have

$$\begin{aligned} \alpha\left(\|\mathbf{x}^{k+1} - \mathbf{x}^*\|^2 + \|\mathbf{y}^{k+1} - \mathbf{y}^*\|^2\right) \leq& \|\mathbf{w}^k - \overline{\mathbf{w}}^k\|_{Q_k} - \|\mathbf{w}^{k+1} - \overline{\mathbf{w}}^k\|_{Q_{k+1}} \\ &+ \xi_2(\gamma_k - \gamma_{k+1}) - c_1\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 - \xi_1\gamma_{k+1} - 2\zeta^k, \end{aligned} \tag{9.5}$$

where ζ^k is defined in (3.19) with $\mathbf{x} = \mathbf{x}^*$, and $\xi_1, \xi_2 \geq 0$. Since $\|\lambda^{k+1} - \overline{\lambda}^{k+1}\| \leq \|\lambda^{k+1} - \overline{\lambda}^k\|$, it follows that

$$\|\mathbf{w}^{k+1} - \overline{\mathbf{w}}^{k+1}\|_{Q_{k+1}} \leq \|\mathbf{w}^{k+1} - \overline{\mathbf{w}}^k\|_{Q_{k+1}}. \tag{9.6}$$

Add $0.5\xi_1(\gamma_k - \gamma_{k+1})$ to each side of (9.5) and combine with (9.6) to obtain

$$\begin{aligned} &\alpha\left(\|\mathbf{x}^{k+1} - \mathbf{x}^*\|^2 + \|\mathbf{y}^{k+1} - \mathbf{y}^*\|^2\right) + c_1\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 + 0.5\xi_1(\gamma_k + \gamma_{k+1}) \\ &\leq E_k - E_{k+1} - 2\zeta^k, \end{aligned}$$

where $E_k := \|\mathbf{w}^k - \overline{\mathbf{w}}^k\|^2_{Q_k} + (0.5\xi_1 + \xi_2)\gamma_k$ and $\xi_1 > 0$ since $s \in (0, (1 + \sqrt{5})/2)$. The left side of this inequality is bounded from below by a positive multiple $\overline{c}$ of

$$d_k = \|\mathbf{x}^{k+1} - \mathbf{x}^*\|^2 + \|\mathbf{y}^{k+1} - \mathbf{y}^*\|^2 + \|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 + \gamma_k + \gamma_{k+1}.$$

Hence, the inequality can be rearranged to yield $E_{k+1} \leq E_k - 2\zeta^k - \overline{c}d_k$. We will show that $|\mathbb{E}[\zeta^k]| \leq c(1 + \theta)^{-k}$, in which case, we have

$$\mathbb{E}\left[E_{k+1}\right] \leq \mathbb{E}\left[E_k\right] + c(1 + \theta)^{-k} - \overline{c}\mathbb{E}\left[d_k\right]. \tag{9.7}$$

Note that $\mathbb{E}[d_k]$ must approach zero. Otherwise, there exists $\epsilon > 0$ and an infinite number of indices k where $\mathbb{E}[d_k] \geq \epsilon$. If this were to hold, then $\mathbb{E}[E_k]$ is eventually negative, which is impossible.

For $\mathbf{x} = \mathbf{x}^*$, (3.19) gives $\zeta^k =$

$$\frac{\Gamma_k}{\eta_k}\left[\left\|\mathbf{x}^* - \check{\mathbf{x}}^k\right\|_{\mathcal{H}}^2 - \left\|\mathbf{x}^* - \check{\mathbf{x}}^{k+1}\right\|_{\mathcal{H}}^2\right] + \Gamma_k \sum_{t=1}^{M_k} t\langle \boldsymbol{\delta}_t, \check{\mathbf{x}}_t - \mathbf{x}^*\rangle + \frac{\Gamma_k \eta_k}{4(1-\eta_k \nu)} \sum_{t=1}^{T} t^2 \|\boldsymbol{\delta}_t\|_{\mathcal{H}^{-1}}^2,$$

where $\Gamma_k = 2/[M_k(M_k+1)]$. The first bracketed term in ζ^k has the upper bound

$$\frac{2\|\mathbf{x}^* - \check{\mathbf{x}}^k\|_{\mathcal{H}}^2}{\eta_k M_k(M_k+1)} \le \frac{4\left(\|\mathbf{x}^*\|_{\mathcal{H}}^2 + \|\check{\mathbf{x}}^k\|_{\mathcal{H}}^2\right)}{\eta_k M_k(M_k+1)},$$

which is bounded by $c(1+\theta)^{-k}$ using (L4c). Also, by (L4b), we have $M_k \eta_k = (1+\theta)^{-k}$. Taking the expectation of ζ^k and utilizing the estimates obtained in Theorem 4.1 for the last two terms in ζ^k, we obtain a bound of the form $\left|\mathbb{E}\left[\zeta^k\right]\right| \le c(1+\theta)^{-k}$.

To complete the proof, we will show that

$$\mathbb{E}\left[E_{k+1}\right] \le c(\mathbb{E}\left[d_k\right] + (1+\theta)^{-k}). \tag{9.8}$$

This is combined with the bound (9.7) to obtain for some $r > 0$,

$$\mathbb{E}\left[E_{k+1}\right] \le \frac{1}{1+r}\mathbb{E}\left[E_k\right] + c(1+\theta)^{-k}.$$

Consequently, $\mathbb{E}\left[E_k\right]$ converges to zero at linear convergence rate τ where

$$1 > \tau > \max\{1/(1+r), 1/(1+\theta)\}.$$

To establish (9.8), first define e_1 and e_2 by

$$e_1(\mathbf{x}, \lambda) = \|\nabla f(\mathbf{x}) - A^\top \lambda\| \quad \text{and} \quad e_2(\mathbf{y}, \lambda) = \|\nabla g(\mathbf{y}) - B^\top \lambda\|.$$

Also, define $\overline{\mathbf{x}}^k = \arg\min\{\phi_k(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}$, where ϕ_k is defined in (3.1). Since $\overline{\mathbf{x}}^k$ minimizes ϕ_k over $\mathcal{X} = \mathbb{R}^{n_1}$, $\nabla \phi_k(\overline{\mathbf{x}}^k) = \mathbf{0}$, which implies that

$$\nabla f(\overline{\mathbf{x}}^k) - A^\top \lambda^k + \beta A^\top (A\mathbf{x}^k + B\mathbf{y}^k - \mathbf{b}) + \mathcal{M}_k(\overline{\mathbf{x}}^k - \mathbf{x}^k) = \mathbf{0}.$$

Utilize this equality in the definition of e_1 and exploit the Lipschitz continuity of ∇f to obtain

$$e_1(\mathbf{x}^{k+1}, \lambda^k) \le c\left(\sqrt{\gamma_k} + \|\overline{\mathbf{x}}^k - \mathbf{x}^k\| + \|\overline{\mathbf{x}}^k - \mathbf{x}^{k+1}\|\right), \tag{9.9}$$

Inserting $\mathbf{x} = \overline{\mathbf{x}}^k$ and $T = M_k$ in (3.22) gives

$$\begin{aligned} \left\|\mathbf{x}^{k+1} - \overline{\mathbf{x}}^k\right\|_{\mathcal{M}_k}^2 \le& \frac{2\Gamma_k}{\eta_k}\left[\left\|\overline{\mathbf{x}}^k - \check{\mathbf{x}}^k\right\|_{\mathcal{H}}^2 - \left\|\overline{\mathbf{x}}^k - \check{\mathbf{x}}^{k+1}\right\|_{\mathcal{H}}^2\right] \\ &+ 2\Gamma_k \sum_{t=1}^{M_k} t\langle \boldsymbol{\delta}_t, \check{\mathbf{x}}_t - \overline{\mathbf{x}}^k\rangle + \frac{2\Gamma_k \eta_k}{4(1-\eta_k \nu)} \sum_{t=1}^{T} t^2 \|\boldsymbol{\delta}_t\|_{\mathcal{H}^{-1}}^2, \end{aligned} \tag{9.10}$$

where $\Gamma_k = 2/[M_k(M_k+1)]$. The right side of (9.10) has the same expression ζ^k that was analyzed previously, except that $\mathbf{x}^*$ is now replaced by $\overline{\mathbf{x}}^k$. By the previous analysis, the last two terms on the right side of (9.10) are bounded by $c(1+\theta)^{-k}$ in expectation. In the first term, observe that

$$\left\|\overline{\mathbf{x}}^k - \breve{\mathbf{x}}^k\right\|_{\mathcal{H}}^2 \le 2\left(\left\|\overline{\mathbf{x}}^k\right\|_{\mathcal{H}}^2 + \left\|\breve{\mathbf{x}}^k\right\|_{\mathcal{H}}^2\right).$$

Again, by the previous analysis, $(\Gamma_k/\eta_k)\|\breve{\mathbf{x}}^k\|_{\mathcal{H}}^2 \le c(1+\theta)^{-k}$ due to (L4c). We will show that $\mathbb{E}\left[\|\overline{\mathbf{x}}^k\|_{\mathcal{H}}^2\right]$ is uniformly bounded, which implies that $(\Gamma_k/\eta_k)\mathbb{E}\left[\|\overline{\mathbf{x}}^k\|_{\mathcal{H}}^2\right] \le c(1+\theta)^{-k}$. In this case, (L3) and (9.10) yield

$$c_1\mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \overline{\mathbf{x}}^k\right\|^2\right] \le \mathbb{E}\left[\left\|\mathbf{x}^{k+1} - \overline{\mathbf{x}}^k\right\|_{\mathcal{M}_k}^2\right] \le c(1+\theta)^{-k}.$$

Combine this with (9.9) to obtain

$$\begin{aligned}\mathbb{E}\left[e_1(\mathbf{x}^{k+1},\lambda^k)^2\right] &\le c\left(\mathbb{E}\left[\gamma_k + \|\overline{\mathbf{x}}^k - \mathbf{x}^k\|^2\right] + (1+\theta)^{-k}\right)\\ &\le c\left(\mathbb{E}\left[\gamma_k + \|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2\right] + (1+\theta)^{-k}\right).\end{aligned} \tag{9.11}$$

To obtain a bound for $\mathbb{E}\left[\|\overline{\mathbf{x}}^k\|_{\mathcal{H}}^2\right]$, we utilize strong convexity (L1) along with any $\mathbf{x} \in \mathbb{R}^{n_1}$ to obtain

$$0 \ge \phi_k(\overline{\mathbf{x}}^k) - \phi_k(\mathbf{x}) \ge \nabla\phi_k(\mathbf{x})(\overline{\mathbf{x}}^k - \mathbf{x}) + \left(\frac{\alpha + c_1}{2}\right)\|\overline{\mathbf{x}}^k - \mathbf{x}\|^2.$$

Hence, by the Schwarz and triangle inequalities, we have

$$\|\overline{\mathbf{x}}^k\| - \|\mathbf{x}\| \le \|\overline{\mathbf{x}}^k - \mathbf{x}\| \le \left(\frac{2}{\alpha + c_1}\right)\|\nabla\phi_k(\mathbf{x})\| = \left(\frac{2}{\alpha + c_1}\right)\|\nabla f(\mathbf{x}) + \mathbf{h}^k + \mathcal{M}_k(\mathbf{x} - \mathbf{x}^k)\|.$$

For $\mathbf{x} = \mathbf{x}_f^*$, the minimizer of f, we have $\nabla f(\mathbf{x}_f^*) = \mathbf{0}$ and

$$\|\overline{\mathbf{x}}^k\|^2 \le c\left(\|\mathbf{h}^k\|^2 + \|\mathbf{x}^k\|^2 + \|\mathbf{x}_f^*\|^2\right).$$

Since $\|\mathbf{h}^k\|^2$ and $\|\mathbf{x}^k\|^2$ are uniformly bounded in expectation by Proposition 9.1, it follows that $\|\overline{\mathbf{x}}^k\|^2$ is uniformly bounded in expectation.

Similar to the treatment of $\overline{\mathbf{x}}^k$, the optimality condition for $\mathbf{y}^{k+1}$, the solution of the $\mathbf{y}$-subproblem in AS-ADMM, is

$$\nabla g(\mathbf{y}^{k+1}) - B^\top\lambda^k + \beta B^\top(A\mathbf{x}^{k+1} + B\mathbf{y}^{k+1} - \mathbf{b}) = \mathbf{0}.$$

This identity is rearranged to give

$$e_2(\mathbf{y}^{k+1},\lambda^k) \le c\sqrt{\gamma_{k+1}}, \quad \text{which yields} \quad \mathbb{E}\left[e_2(\mathbf{y}^{k+1},\lambda^k)^2\right] \le c\mathbb{E}\left[\gamma_{k+1}\right]. \tag{9.12}$$

The set of multipliers $\boldsymbol{\Lambda}^*$ are those $\lambda \in \mathbb{R}^n$ that satisfy both of the equations $A^\top\lambda = \nabla f(\mathbf{x}^*)$ and $B^\top\lambda = \nabla g(\mathbf{y}^*)$. Hence, $\boldsymbol{\Lambda}^*$ is a particular solution plus any vector in the null space $\mathcal{N}$ of the matrix $[A \;\; B]^{\mathsf{T}}$. The projection $\overline{\lambda}$ onto $\mathbf{\Lambda}^*$ has the property that $\lambda - \overline{\lambda}$ is orthogonal to $\mathcal{N}$, which implies the existence of a constant c such that

$$\|\lambda - \overline{\lambda}\|^2 \le c\|[A \;\; B]^{\mathsf{T}}(\lambda - \overline{\lambda})\|^2 = c\Big(\|A^{\mathsf{T}}(\lambda - \overline{\lambda})\|^2 + \|B^{\mathsf{T}}(\lambda - \overline{\lambda})\|^2 \Big).$$

Since $A^{\mathsf{T}}\overline{\lambda} = \nabla f(\mathbf{x}^*)$ and $B^{\mathsf{T}}\overline{\lambda} = \nabla g(\mathbf{y}^*)$, this bound can be rewritten

$$\|\lambda - \overline{\lambda}\|^2 \le c\big(e_1(\mathbf{x}^*, \lambda)^2 + e_2(\mathbf{y}^*, \lambda)^2\big). \tag{9.13}$$

The following inequality is deduced from the triangle inequality, and the Lipschitz assumption for the gradient of f:

$$\begin{aligned} e_1(\mathbf{x}^*, \lambda)^2 \le& 2e_1(\mathbf{x}^{k+1}, \lambda^k)^2 + 2\big(e_1(\mathbf{x}^*, \lambda) - e_1(\mathbf{x}^{k+1}, \lambda^k)\big)^2 \\ \le& c\big(e_1(\mathbf{x}^{k+1}, \lambda^k)^2 + \|\mathbf{x}^{k+1} - \mathbf{x}^*\|^2 + \|\lambda^k - \lambda\|^2\big). \end{aligned}$$

An analogous inequality holds for e_2. Hence, by (9.13),

$$\|\lambda - \overline{\lambda}\|^2 \le c\delta_{k+1}, \tag{9.14}$$

where

$$\delta_{k+1} = e_1(\mathbf{x}^{k+1}, \lambda^k)^2 + e_2(\mathbf{y}^{k+1}, \lambda^k)^2 + \|\mathbf{x}^{k+1} - \mathbf{x}^*\|^2 + \|\mathbf{y}^{k+1} - \mathbf{x}^*\|^2 + \|\lambda^k - \lambda\|^2.$$

Now insert $\lambda = \lambda^{k+1}$ in (9.14), take expectation, and utilize the bounds (9.11) and (9.12) to obtain

$$\mathbb{E}\big[\|\lambda^{k+1} - \overline{\lambda}^{k+1}\|^2\big] \le c\big(\mathbb{E}\big[d_k\big] + (1+\theta)^{-k}\big).$$

Since

$$\|\mathbf{w}^{k+1} - \overline{\mathbf{w}}^{k+1}\|^2_{\mathcal{Q}_{k+1}} = \|\mathbf{x}^{k+1} - \mathbf{x}^*\|^2_{\mathcal{D}_{k+1}} + \beta\|B(\mathbf{y}^{k+1} - \mathbf{y}^*)\|^2 + \|\lambda^{k+1} - \overline{\lambda}^{k+1}\|^2/(s\beta)$$

and

$$E_{k+1} = \|\mathbf{w}^{k+1} - \overline{\mathbf{w}}^{k+1}\|^2_{\mathcal{Q}_{k+1}} + 0.5(\xi_1 + \xi_2)\gamma_{k+1},$$

it follows that $\mathbb{E}\big[E_{k+1}\big] \le c(\mathbb{E}\big[d_k\big] + (1+\theta)^{-k})$, which completes the proof of (9.8). □

A careful analysis of the proof of Proposition 9.3 reveals that the strong convexity assumption (L1) can be relaxed to the following: Both f and g have Lipschitz continuous gradients and any one of the following conditions hold:

Case 1: Both f and g are strongly convex (current version of (L1)).
Case 2: f is strongly convex and B has full column rank.
Case 3: g is strongly convex and A has full column rank.

Acknowledgement This research was partially supported by the National Natural Science Foundation of China under grant 12001430, by the China Postdoctoral Science Foundation under grant 2020M683545, by the USA National Science Foundation under grants 1522629, 1522654, 1819002, 1819161, 2031213, and 2110722, and by the USA Office of Naval Research under grant N00014-18-1-2100.

References

1. Azadi, S., Sra, S.: Towards an optimal stochastic alternating direction method of multipliers. In: Xing, EP., Jebara, T. (eds.) Proceedings of the 31st International Conference on Machine Learning, vol. 32 of Proceedings of Machine Learning Research, Bejing, China, 22–24 , PMLR, pp. 620–628 (2014)
2. Bai, J.C., Li, J.C., Xu, F.M., Zhang, H.: Generalized symmetric ADMM for separable convex optimization. Comput. Optim. Appl. **70**, 129–170 (2018)
3. Boley, D.: Local linear convergence of the alternating direction method of multipliers on quadratic or linear programs. SIAM J. Optim. **23**, 2183–2207 (2013)
4. Cai, X., Han, D., Yuan, X.: On the convergence of the direct extension of ADMM for three-block separable convex minimization models with one strongly convex function. Comput. Optim. Appl. **66**, 39–73 (2017)
5. Chambolle, A., Ehrhardt, M.J., Richtárik, P., Schönlieb, C.-B.: Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging applications. SIAM J. Optim. **28**, 2783–2808 (2018)
6. Chen, C., He, B., Ye, Y., Yuan, X.: The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent. Math. Program. **155**, 57–79 (2016)
7. Chen, J.W., Wang, Y.Y., He, H.J., Lv, Y.B.: Convergence analysis of positive-indefinite proximal ADMM with a Glowinski's relaxation factor. Numer. Algor. **83**, 1415–1440 (2020)
8. Chen, Y., Hager, W.W., Yashtini, M., Ye, X., Zhang, H.: Bregman operator splitting with variable stepsize for total variation image reconstruction. Comput. Optim. Appl. **54**, 317–342 (2013)
9. Deng, W., Lai, M.J., Peng, Z.M., Yin, W.T.: Parallel multi-block ADMM with $o(1/k)$ convergence. J. Sci. Comput. **71**, 712–736 (2017)
10. Douglas, J., Rachford, H.: On the numerical solution of heat conduction problems in two and three space variables. Trans. Amer. Math. Soc. **82**, 421–439 (1956)
11. Eckstein, J., Bertsekas, D.P.: On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. Math. Program. **55**, 293–318 (1992)
12. Freund, J.E.: Mathematical Statistics. Prentice-Hall, Englewood Cliffs, NJ (1962)
13. Friedman, J., Hastie, T., Tibshiran, R.: Sparse inverse covariance estimation with the graphical lasso. Biostatistics **9**, 432–441 (2008)
14. Gabay, D.: Applications of the method of multipliers to variational inequalities. In: Fortin, M., Glowinski, R. (eds) Augmented Lagrange methods: applications to the solution of boundary-valued problems, Amsterdam, North Holland, pp. 299–331 (1983)
15. Gabay, D., Mercier, B.: A dual algorithm for the solution of nonlinear variational problems via finite-element approximations. Comput. Math. Appl. **2**, 17–40 (1976)
16. Glowinski, R., Marrocco, A.: Sur l'approximation par éléments finis d'ordre un, et la résolution par pénalisation-dualité d'une classe de problèmes de dirichlet nonlinéaires. RAIRO Anal. Numé. **9**(R2), 41–76 (1975)
17. Goldstein, T., Donoghue, B., Setzer, S., Baraniuk, R.: Fast alternating direction optimization methods. SIAM J. Imaging Sci. **7**, 1588–1623 (2014)
18. Gu, G.Y., He, B.S., Yang, J.F.: Inexact alternating-direction-based contraction methods for separable linearly constrained convex optimization. J. Optim. Theory Appl. **163**, 105–129 (2014)
19. Hager, W.W., Zhang, H.: Inexact alternating direction methods of multipliers for separable convex optimization. Comput. Optim. Appl. **73**, 201–235 (2019)
20. Hager, W.W., Zhang, H.: Convergence rates for an inexact ADMM applied to separable convex optimization. Comput. Optim. Appl. **77**, 729–754 (2020). https://doi.org/10.1007/s10589-020-00221
21. Han, D., Sun, D., Zhang, L.: Linear rate convergence of the alternating direction method of multipliers for convex composite programming. Math. Oper. Res. **43**, 622–637 (2018)

22. He, B., Yuan, X.: A class of ADMM-based algorithms for three-block separable convex programming. Comput. Optim. Appl. **70**, 791–826 (2018)
23. Hong, M., Luo, Z.: On the linear convergence of the alternating direction method of multipliers. Math. Program. **162**, 165–199 (2017)
24. Johnson, R., Zhang, T.: Accelerating stochastic gradient descent using predictive variance reduction. In: Conference on Neural Information Processing Systems (NIPS), pp. 315–323 (2013)
25. Lin, T., Ma, S., Zhang, S.: On the global linear convergence of the ADMM with multiblock variables. SIAM J. Optim. **25**, 1478–1497 (2015)
26. Liu, Y., Shang, F., Cheng, J.: Accelerated variance reduced stochastic ADMM. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI'17, AAAI Press, pp. 2287–2293 (2017)
27. Ng, M.K., Wang, F., Yuan, X.: Inexact alternating direction methods for image recovery. SIAM J. Sci. Comput. **33**, 1643–1668 (2011)
28. Ouyang, H., He, N., Tran, L., Gray, A.: Stochastic alternating direction method of multipliers. In: Dasgupta, S., McAllester, D. (eds.) Proceedings of the 30th International Conference on Machine Learning, vol. 28 of Proceedings of Machine Learning Research, Atlanta, Georgia, USA, 17–19, PMLR, pp. 80–88 (2013)
29. Ouyang, Y.Y., Chen, Y.M., Lan, G.H., Pasiliao, E.: An accelerated linearized alternating direction method of multipliers. SIAM J. Imaging Sci. **8**, 644–681 (2015)
30. Peng, J., Zhang, X.: Linear convergence rate of the generalized alternating direction method of multipliers for a class of convex optimization problems (2017). http://www.optimization-online.org/DB_FILE/2017/11/6354.pdf
31. Sun, H., Sun, M., Wang, Y.: Proximal ADMM with larger step size for two-block separable convex programming and its application to the correlation matrices calibrating problems. J. Nonlinear Sci. Appl. **10**, 5038–5051 (2017)
32. Suzuki, T.: Dual averaging and proximal gradient descent for online alternating direction multiplier method. In: Dasgupta, S., McAllester, D. (eds.) Proceedings of the 30th International Conference on Machine Learning, vol. 28 of Proceedings of Machine Learning Research, Atlanta, Georgia, USA, 17–19 Jun 2013, PMLR, pp. 392–400
33. Tao, M.: Convergence study of indefinite proximal ADMM with a relaxation factor. Comput. Optim. Appl. **77**, 91–123 (2020). https://doi.org/10.1007/s10589-020-00206-x
34. Tao, M., Yuan, X.: Recovering low-rank and sparse components of matrices from incomplete and noisy observations. SIAM J. Optim. **21**, 57–81 (2011)
35. Tseng, P.: Approximation accuracy, gradient methods, and error bound for structured convex optimization. Math. Program. **125**, 263–295 (2010)
36. Xie, Y., Shanbhag, U.V.: SI-ADMM: a stochastic inexact ADMM framework for stochastic convex programs. IEEE Trans. Automat. Control **65**, 2355–2370 (2020)
37. Xu, M.H., Wu, T.: A class of linearized proximal alternating direction methods. J. Optim. Theory Appl. **151**, 321–337 (2011)
38. Yang, W.H., Han, D.: Linear convergence of the alternating direction method of multipliers for a class of convex optimization problems. SIAM J. Numer. Anal. **54**, 625–640 (2016)
39. Yang, Z., Yan, Z.: Fast linearized alternating direction method of multipliers for the augmented l_1 -regularized problem. Signal, Image Video Process. **9**, 1601–1612 (2015)
40. Zhao, S.-Y., Li, W.-J., Zhou, Z.-H.: Scalable stochastic alternating direction method of multipliers (2015), arXiv: 1502.03529
41. Zheng, S., Kwok, J. T.: Fast-and-light stochastic admm. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI'16, AAAI Press, pp. 2407–2413 (2016)
42. Zhong, W., Kwok, J.: Fast stochastic alternating direction method of multipliers. In Xing, E. P., Jebara, T. (eds) Proceedings of the 31st International Conference on Machine Learning,, vol. 32 of Proceedings of Machine Learning Research, Bejing, China, 22–24 Jun 2014, PMLR, pp. 46–54

Authors and Affiliations

Jianchao Bai[1] · William W. Hager[2] · Hongchao Zhang[3]

Jianchao Bai
jianchaobai@nwpu.edu.cn

Hongchao Zhang
hozhang@math.lsu.edu
http://www.math.lsu.edu/~hozhang

1 MIIT Key Laboratory of Dynamics and Control of Complex Systems, School of Mathematics and Statistics, Northwestern Polytechnical University, Xi'an 710072, China

2 Department of Mathematics, University of Florida, PO Box 118105, Gainesville, FL 32611-8105, USA

3 Department of Mathematics, Louisiana State University, Baton Rouge, LA 70803-4918, USA