Adv Comput Math (2021) 47:9
https://doi.org/10.1007/s10444-020-09840-9

New convergence analysis of a primal-dual algorithm with large stepsizes

Zhi Li[1] · Ming Yan[2]

Received: 2 February 2020 / Accepted: 17 December 2020 / Published online: 25 January 2021

Abstract

We consider a primal-dual algorithm for minimizing $f(\mathbf{x}) + h\Box l(\mathbf{A}\mathbf{x})$ with Fréchet differentiable f and l^*. This primal-dual algorithm has two names in literature: Primal-Dual Fixed-Point algorithm based on the Proximity Operator (PDFP^{2}O) and Proximal Alternating Predictor-Corrector (PAPC). In this paper, we prove its convergence under a weaker condition on the stepsizes than existing ones. With additional assumptions, we show its linear convergence. In addition, we show that this condition (the upper bound of the stepsize) is tight and can not be weakened. This result also recovers a recently proposed positive-indefinite linearized augmented Lagrangian method. In addition, we apply this result to a decentralized consensus algorithm PG-EXTRA and derive the weakest convergence condition.

Keywords Linearized augmented lagrangian · Primal-dual · Decentralized consensus

Mathematics Subject Classification (2010) 68Q25 · 68R10 · 68U05

1 Introduction

Minimizing the sum of two functions has applications in various areas including image processing, machine learning, and decentralized consensus optimization [4, 5,

Communicated by: Russell Luke

✉ Ming Yan
myan@msu.edu

Zhi Li
lizhiupc@gmail.com

[1] Shanghai Key Laboratory of Multidimensional Information Processing, Department of Computer Science and Technology, East China Normal University, Shanghai, 200062, China

[2] Department of Computational Mathematics, Science and Engineering and Department of Mathematics, Michigan State University, East Lansing, MI, 48824, USA

17, 26]. In this paper, we aim to minimize the sum of two functions in the following form:

$$\operatorname*{minimize}_{\mathbf{x}\in\mathcal{X}} \; f(\mathbf{x}) + h\Box l(\mathbf{A}\mathbf{x}), \tag{1}$$

where $\mathcal{X}$ and $\mathcal{S}$ are two real Hilbert spaces; $f(\mathbf{x}) : \mathcal{X} \mapsto (-\infty, +\infty]$, $h(\mathbf{s}) : \mathcal{S} \mapsto (-\infty, +\infty]$, and $l(\mathbf{s}) : \mathcal{S} \mapsto (-\infty, +\infty]$ are proper lower semi-continuous (lsc) convex functions; $h\Box l$ is the infimal convolution of h and l that is defined as $h\Box l(\mathbf{s}) = \inf_{\mathbf{t}\in\mathcal{S}} h(\mathbf{t}) + l(\mathbf{s}-\mathbf{t})$; the linear operator $\mathbf{A} : \mathcal{X} \mapsto \mathcal{S}$ is bounded. In addition, we assume that $f(\mathbf{x})$ is Fréchet differentiable with a Lipschitz continuous gradient, l is strongly convex in $\mathrm{dom}(l)$[1], and the proximal operator of h, which is defined as:

$$\mathbf{prox}_{\lambda h}(\mathbf{t}) = (\mathbf{I} + \lambda\partial h)^{-1}(\mathbf{t}) := \operatorname*{arg\,min}_{\mathbf{s}\in\mathcal{S}} \; h(\mathbf{s}) + \frac{1}{2\lambda}\|\mathbf{s}-\mathbf{t}\|^2,$$

has a closed-form solution or can be easily computed. Here, ∂h is the subdifferential of the convex function h.

Many existing papers considered a special case of (1) with $l(\mathbf{s})$ being the indicator function $\iota_{\{\mathbf{0}\}}(\mathbf{s})$ that returns 0 if $\mathbf{s} = \mathbf{0}$ and $+\infty$ otherwise. In this special case, the infimal convolution $h\Box l$ degenerates to h, and the problem (1) becomes:

$$\operatorname*{minimize}_{\mathbf{x}\in\mathcal{X}} \; f(\mathbf{x}) + h(\mathbf{A}\mathbf{x}). \tag{2}$$

The corresponding saddle-point problem is:

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{s}\in\mathcal{S}} \; f(\mathbf{x}) + \langle \mathbf{A}\mathbf{x}, \mathbf{s}\rangle - h^*(\mathbf{s}). \tag{3}$$

If a saddle point $(\mathbf{x}^\star, \mathbf{s}^\star)$ exists for (3), then $\mathbf{x}^\star$ is an optimal solution for (2).

In order to solve (2) (or (3)), a primal-dual algorithm was proposed in different fields under different names [7, 12, 23]. Loris and Verhoeven [23] focused on a particular smooth function $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{K}\mathbf{x} - \mathbf{y}\|^2$, where $\mathbf{K}$ is a linear operator and $\mathbf{y}$ is given. Chen, Huang, and Zhang [7] considered the general problem (2) and proposed a Primal-Dual Fixed-Point algorithm based on the Proximity Operator (PDFP^{2}O). Then, the same algorithm was rediscovered under the name Proximal Alternating Predictor-Corrector (PAPC) in [12] to solve (2) and its extension to a finite sum of composite functions when h is separable. One iteration of the algorithm is:

$$\mathbf{s}^{k+1} = \left(\mathbf{I} + \sigma\partial h^*\right)^{-1}\left((\mathbf{I} - \tau\sigma\mathbf{A}\mathbf{A}^\top)\mathbf{s}^k + \sigma\mathbf{A}\left(\mathbf{x}^k - \tau\nabla f(\mathbf{x}^k)\right)\right), \tag{4a}$$

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \tau\nabla f(\mathbf{x}^k) - \tau\mathbf{A}^\top\mathbf{s}^{k+1}. \tag{4b}$$

Here, τ and σ are the primal and dual stepsizes, respectively, and the convergence of this algorithm is shown when $\tau\sigma\|\mathbf{A}\mathbf{A}^\top\| \le 1$ and $2\tau/L < 1$ [7]. Here, L is the Lipschitz constant of $\nabla f(\mathbf{x})$ and $\|\mathbf{A}\mathbf{A}^\top\|$ is the operator norm of $\mathbf{A}\mathbf{A}^\top$. When $\mathbf{A}$ is a matrix, $\|\mathbf{A}\mathbf{A}^\top\|$ is the largest eigenvalue of $\mathbf{A}\mathbf{A}^\top$.

[1] It means that $l^*(\mathbf{s})$ (the Legendre-Fenchel conjugate of $l(\mathbf{s})$) is Fréchet differentiable with a Lipschitz continuous gradient.

There are many other algorithms for solving (2) and its extensions. For example, Condat-Vu [6, 10, 27] solves a more general problem than (2) with an additional non-differential function. However, the corresponding parameters τ and σ have to satisfy $\tau\sigma\|\mathbf{A}\mathbf{A}^\top\| + 2\tau/L \leq 1$ [18]. When $f(\mathbf{x}) = 0$, Condat-Vu reduces to Chambolle-Pock [4]. There are several other primal-dual algorithms for minimizing the sum of three functions, one of which is differentiable [2, 3, 8, 9, 11, 20, 31]. Interested readers are referred to [19, 31] for the comparison of different primal-dual algorithms for minimizing the sum of three functions. All the algorithms mentioned above solve bilinear saddle-point problems in the form of (3) or its variants. Recently, many algorithms have been developed to solve more general saddle-point problems with non-bilinear terms [13, 14, 16, 29, 30]. A review for primal-dual algorithms is beyond the scope of this paper, and we focus on the specific primal-dual algorithm PAPC here.

When there is only one function $f(\mathbf{x})$, i.e., $h(\mathbf{s}) = 0$, we let $\mathbf{A} = \mathbf{0}$, and the primal-dual algorithm reduces to the gradient descent with stepsize τ. Therefore, the condition $\tau < 2/L$ can not be relaxed. The remaining question is *can the condition* $\tau\sigma \leq 1/\|\mathbf{A}\mathbf{A}^\top\|$ *be relaxed?* In [7, Section 5.1], the authors numerically showed that a larger stepsize (e.g., $\tau\sigma = 4/(3\|\mathbf{A}\mathbf{A}^\top\|)$) gives a better performance than stepsizes satisfying the condition $\tau\sigma \leq 1/\|\mathbf{A}\mathbf{A}^\top\|$. The convergence for $\tau\sigma < 4/(3\|\mathbf{A}\mathbf{A}^\top\|)$ was an open problem, and this work resolves it.

For linearized Augmented Lagrangian Method (ALM) [32]—a special case of the primal-dual algorithm (4)—the condition $\tau\sigma \leq 1/\|\mathbf{A}\mathbf{A}^\top\|$ is relaxed in [15]. Consider the constrained optimization problem:

$$\begin{aligned} \underset{\mathbf{s}}{\text{minimize}} \quad & h^*(\mathbf{s}), \\ \text{subject to} \quad & -\mathbf{A}^\top\mathbf{s} = \mathbf{b}. \end{aligned}$$

Its dual problem is

$$\underset{\mathbf{x}}{\text{minimize}} \;\; \mathbf{b}^\top\mathbf{x} + h(\mathbf{A}\mathbf{x}),$$

which is the problem (2) with $f(\mathbf{x}) = \mathbf{b}^\top\mathbf{x}$. The linearized ALM is

$$\mathbf{s}^{k+1} = \arg\min_{\mathbf{s}} \; h^*(\mathbf{s}) + \frac{\beta}{2}\left\|\mathbf{s} - \mathbf{s}^k - \frac{1}{\beta}\mathbf{A}(\mathbf{x}^k - \tau(\mathbf{A}^\top\mathbf{s}^k + \mathbf{b}))\right\|^2, \tag{6a}$$

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \tau(\mathbf{A}^\top\mathbf{s}^{k+1} + \mathbf{b}). \tag{6b}$$

It is exactly the primal-dual algorithm (4) with $\beta = 1/\sigma$. Note that the step in (6a) can be rewritten as:

$$\arg\min_{\mathbf{s}} \; h^*(\mathbf{s}) - \langle \mathbf{x}^k, \mathbf{A}^\top\mathbf{s} + \mathbf{b}\rangle + \frac{\tau}{2}\|\mathbf{A}^\top\mathbf{s} + \mathbf{b}\|_2^2 + \frac{1}{2\sigma}\|\mathbf{s} - \mathbf{s}^k\|^2_{\mathbf{I}-\tau\sigma\mathbf{A}\mathbf{A}^\top}.$$

In [32], positive-definiteness of $\mathbf{I} - \tau\sigma\mathbf{A}\mathbf{A}^\top$ is required for showing the convergence. Then, the authors in [15] relaxed the condition and showed that $(4/3)\mathbf{I} - \tau\sigma\mathbf{A}\mathbf{A}^\top$ being positive definite is the necessary and sufficient condition for the convergence of linearized ALM. That is, this relaxed condition is sufficient for the convergence of linearized ALM, and if the condition is not satisfied, there exist a function $h^*(\mathbf{s})$, a linear operator $\mathbf{A}$, and an initialization such that the algorithm does not converge. This result motivates us to show the convergence of (4) under a weaker condition. In this paper, we provide the necessary and sufficient condition on $\tau\sigma$ for the convergence of algorithm (4). This extension from [15] is nontrivial because the function $f(\mathbf{x})$ from linearized ALM is linear, i.e., $f(\mathbf{x}) = \mathbf{b}^\top\mathbf{x}$, and the Lipschitz constant of ∇f is 0.

Furthermore, we consider the more general problem (1) with infimal convolution, which was not considered in [7, 12], because it provides a tight upper bound for the stepsize of Proximal Gradient EXact firsT-ordeR Algorithm (PG-EXTRA) in decentralized consensus optimization. More details are in Section 3.

In this paper, we relax the parameters for the primal-dual algorithm (4) and provide a tight bound for the primal and dual stepsizes. This result recovers one special case of the positive-indefinite ALM in [15]. Instead of using positive semidefinite operators for primal-dual variables in standard analysis, we allow the operator to be indefinite; see the operator in (8). Note that the analysis in this paper with indefinite operators is nontrivial because the standard techniques can not be applied. In addition, the linear convergence result is better than existing ones. Finally, we apply this result to a decentralized consensus algorithm and obtain its weakest convergence condition.

The rest of this paper is organized as follows. In Section 2, we present the algorithm to solve (1). We show its convergence for the general case in Section 2.3 and linear convergence rates under additional assumptions in Section 2.4. In Section 2.5, we provide one example to show that the upper bound for its stepsize is tight. The application to a decentralized consensus algorithm is provided in Section 3. Then, we end this paper with a short conclusion.

2 New convergence results with weaker conditions

2.1 A primal-dual algorithm

In this paper, we extend an existing primal-dual algorithm (4) to solve (1) with an infimal convolution and show its convergence results with weaker conditions. Firstly, we explain this algorithm via operator splitting, which is different from those in the literature. Instead of considering problem (1), we consider the corresponding saddle-point problem:

$$\min_{\mathbf{x}} \max_{\mathbf{s}} f(\mathbf{x}) + \langle \mathbf{A}\mathbf{x}, \mathbf{s}\rangle - h^*(\mathbf{s}) - l^*(\mathbf{s}), \tag{7}$$

whose optimality condition for a saddle point $(\mathbf{x}^\star, \mathbf{s}^\star)$ is

$$\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix} \in \begin{bmatrix} 0 & \mathbf{A}^\top \\ -\mathbf{A} & \partial h^* \end{bmatrix} \begin{bmatrix}\mathbf{x}^\star \\ \mathbf{s}^\star\end{bmatrix} + \begin{bmatrix}\nabla f(\mathbf{x}^\star) \\ \nabla l^*(\mathbf{s}^\star)\end{bmatrix}.$$

We apply the following forward-backward operator splitting with self-adjoint positive definite operators $\mathbf{P}$ and $\mathbf{D}-\tau\sigma\mathbf{A}\mathbf{P}^{-1}\mathbf{A}^\top$ defined on $\mathcal{X}$ and $\mathcal{S}$, respectively:

$$\begin{aligned}&\begin{bmatrix}\mathbf{P} & \mathbf{0}\\ \mathbf{0} & \mathbf{D}-\tau\sigma\mathbf{A}\mathbf{P}^{-1}\mathbf{A}^\top\end{bmatrix}\begin{bmatrix}\mathbf{x}^k\\ \mathbf{s}^k\end{bmatrix}-\begin{bmatrix}\tau\nabla f(\mathbf{x}^k)\\ \sigma\nabla l^*(\mathbf{s}^k)\end{bmatrix}\\ \in\ &\begin{bmatrix}\mathbf{P} & \mathbf{0}\\ \mathbf{0} & \mathbf{D}-\tau\sigma\mathbf{A}\mathbf{P}^{-1}\mathbf{A}^\top\end{bmatrix}\begin{bmatrix}\mathbf{x}^{k+1}\\ \mathbf{s}^{k+1}\end{bmatrix}+\begin{bmatrix}0 & \tau\mathbf{A}^\top\\ -\sigma\mathbf{A} & \sigma\partial h^*\end{bmatrix}\begin{bmatrix}\mathbf{x}^{k+1}\\ \mathbf{s}^{k+1}\end{bmatrix}.\end{aligned}\tag{8}$$

Here, τ and σ are two positive parameters. When $\mathbf{P}$ and $\mathbf{D}$ are the identity operators in $\mathcal{X}$ and $\mathcal{Y}$, respectively, τ and σ are the primal and dual stepsizes, respectively. Different operators $\mathbf{P}$ and $\mathbf{D}$ may be chosen in different scenarios. For example, we can choose $\mathbf{P}$ (or $\mathbf{D}$) to be a diagonal matrix such that the stepsize is different for different coordinates of $\mathbf{x}$ (or $\mathbf{s}$) when $\mathcal{X}$ (or $\mathcal{S}$) is finite dimensional. Define $\mathbf{M}=\frac{\tau}{\sigma}(\mathbf{D}-\tau\sigma\mathbf{A}\mathbf{P}^{-1}\mathbf{A}^\top)$. Then, we apply the Gaussian elimination and obtain:

$$\begin{bmatrix}\mathbf{P} & \mathbf{0}\\ \sigma\mathbf{A} & \frac{\sigma}{\tau}\mathbf{M}\end{bmatrix}\begin{bmatrix}\mathbf{x}^k\\ \mathbf{s}^k\end{bmatrix}-\begin{bmatrix}\tau\nabla f(\mathbf{x}^k)\\ \sigma\tau\mathbf{A}\mathbf{P}^{-1}\nabla f(\mathbf{x}^k)+\sigma\nabla l^*(\mathbf{s}^k)\end{bmatrix}\in\begin{bmatrix}\mathbf{P} & \tau\mathbf{A}^\top\\ \mathbf{0} & \mathbf{D}+\sigma\partial h^*\end{bmatrix}\begin{bmatrix}\mathbf{x}^{k+1}\\ \mathbf{s}^{k+1}\end{bmatrix}.$$

Given $(\mathbf{x}^k,\mathbf{s}^k)$, one iteration of the primal-dual algorithm is

$$\mathbf{s}^{k+1}=\left(\mathbf{D}+\sigma\partial h^*\right)^{-1}\left(\frac{\sigma}{\tau}\mathbf{M}\mathbf{s}^k+\sigma\mathbf{A}\left(\mathbf{x}^k-\tau\mathbf{P}^{-1}\nabla f(\mathbf{x}^k)\right)-\sigma\nabla l^*(\mathbf{s}^k)\right),\tag{9a}$$

$$\mathbf{x}^{k+1}=\mathbf{x}^k-\tau\mathbf{P}^{-1}\nabla f(\mathbf{x}^k)-\tau\mathbf{P}^{-1}\mathbf{A}^\top\mathbf{s}^{k+1}.\tag{9b}$$

From this analysis, we can easily see that a point $(\mathbf{x}^\star,\mathbf{s}^\star)$ is a saddle point of (7) if and only if it is a fixed point of (9). Therefore, we only need to show the convergence to a fixed point of (9). Note that we could store $\mathbf{A}^\top\mathbf{s}$ in the implementation, and the iteration is equivalent to

$$\begin{aligned}\mathbf{s}^{k+1}&=\left(\mathbf{D}+\sigma\partial h^*\right)^{-1}\left(\mathbf{D}\mathbf{s}^k+\sigma\mathbf{A}\left(\mathbf{x}^k-\tau\mathbf{P}^{-1}(\nabla f(\mathbf{x}^k)+\mathbf{A}^\top\mathbf{s}^k)\right)-\sigma\nabla l^*(\mathbf{s}^k)\right),\\ \mathbf{x}^{k+1}&=\mathbf{x}^k-\tau\mathbf{P}^{-1}\nabla f(\mathbf{x}^k)-\tau\mathbf{P}^{-1}\mathbf{A}^\top\mathbf{s}^{k+1}.\end{aligned}$$

Therefore, only one application of $\mathbf{A}$ and one application of $\mathbf{A}^\top$ are needed in each iteration.

Let $\mathbf{I}$ be the identity operator defined on a Hilbert space. For simplicity, we do not specify the space on which it is defined when it is clear from the context. When l is the indicator of a singleton[2], $\mathbf{P}=\mathbf{I}$, and $\mathbf{D}=\mathbf{I}$, the iteration of (9) reduces to (4), the existing primal-dual algorithm proposed in [7, 12, 23]. Its convergence is shown if $\mathbf{I}-\tau\sigma\mathbf{A}\mathbf{A}^\top$ is positive semidefinite and $\tau<2/L$ with L being the Lipschitz constant of ∇f.

If the operators $\mathbf{P}$ and $\mathbf{D}-\tau\sigma\mathbf{A}\mathbf{P}^{-1}\mathbf{A}^\top$ are positive definite, the convergence of (9) with an additional condition for τ can be shown easily from nonexpansive operators with metric [1, 25, 31]. To the best of our knowledge, this paper is the first one to show the convergence of a primal-dual algorithm when $\mathbf{D}-\tau\sigma\mathbf{A}\mathbf{P}^{-1}\mathbf{A}^\top$ is not positive definite, and the analysis is different from positive definite cases.

[2] It means that $\nabla l^*(\mathbf{s})\equiv 0$.

2.2 Assumptions for new analysis

An extension of this existing primal-dual algorithm (4) to (9) is derived to solve the problem (1) with an infimal convolution. In addition, we show the convergence of (4) with a larger $\tau\sigma$. Specifically, we can choose $\tau\sigma$ such that $(4/3)\mathbf{D} - \tau\sigma\mathbf{A}\mathbf{P}^{-1}\mathbf{A}^\top$ is positive semidefinite, i.e., the upper bound for $\tau\sigma$ is increased by $1/3$. It means that we can choose a larger stepsize σ when the primal stepsize τ is fixed.

For convenience, we introduce two operators as:

$$\mathbf{M}_1 := \frac{\tau}{\sigma}(\mathbf{D} - \theta\tau\sigma\mathbf{A}\mathbf{P}^{-1}\mathbf{A}^\top), \qquad \mathbf{M}_2 := \tau^2(1-\theta)\mathbf{A}\mathbf{P}^{-1}\mathbf{A}^\top.$$

Here, $\theta \in (3/4, 1]$ is chosen such that $\mathbf{M}_1$ is positive definite and $\mathbf{M}_2$ is positive semidefinite. We can find such $\theta \in (3/4, 1]$ whenever $(4/3)\mathbf{D} - \tau\sigma\mathbf{A}\mathbf{P}^{-1}\mathbf{A}^\top$ is positive semidefinite. We would like to emphasize here that $\theta > 3/4$ is crucial in the proof of the convergence because we need $4\theta - 3$ to be positive. On the other side, $\theta \le 1$ is required for $\mathbf{M}_2$ being positive semidefinite. With these two operators, we have $\mathbf{M} = \mathbf{M}_1 - \mathbf{M}_2$. In addition, we define a positive definite operator as follows:

$$\widetilde{\mathbf{M}} := \mathbf{M}_1 + \mathbf{M}_2.$$

Given a self-adjoint operator $\overline{\mathbf{M}}$, we let $\langle \mathbf{s}, \mathbf{t}\rangle_{\overline{\mathbf{M}}} := \langle \mathbf{s}, \overline{\mathbf{M}}\mathbf{t}\rangle$ and $\|\mathbf{s}\|^2_{\overline{\mathbf{M}}} = \langle \mathbf{s}, \overline{\mathbf{M}}\mathbf{s}\rangle$. Note that $\|\mathbf{s}\|^2_{\overline{\mathbf{M}}}$ can be negative if $\overline{\mathbf{M}}$ is not positive semidefinite. When $\overline{\mathbf{M}}$ is positive definite, we further define the induced norm as $\|\mathbf{s}\|_{\overline{\mathbf{M}}} = \sqrt{\langle \mathbf{s}, \mathbf{s}\rangle_{\overline{\mathbf{M}}}}$. Let $\lambda_{\min}(\overline{\mathbf{M}})$ be the smallest eigenvalue of $\overline{\mathbf{M}}$. For $(\mathbf{x}, \mathbf{s}) \in \mathcal{X} \times \mathcal{S}$, we define $\|(\mathbf{x}, \mathbf{s})\|^2_{\mathbf{P},\overline{\mathbf{M}}} = \|\mathbf{x}\|^2_{\mathbf{P}} + \|\mathbf{s}\|^2_{\overline{\mathbf{M}}}$.

Assumption 1 *Functions f, h, and l are proper lsc convex. In addition, f is Frechet differentiable and l is strictly convex (i.e., l^* is Frechet differentiable). Operators $\mathbf{P}$ and $\mathbf{M}_1$ are positive definite. The iteration (9) has at least one fixed point. Let $(\mathbf{x}^\star, \mathbf{s}^\star)$ be any fixed point of (9). For any $\mathbf{x} \in \mathcal{X}$ and $\mathbf{s} \in \mathcal{S}$, we have:*

$$\langle \mathbf{x} - \mathbf{x}^\star, \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}^\star)\rangle \ge \beta\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}^\star)\|^2_{\mathbf{P}^{-1}}, \tag{11}$$

$$\langle \mathbf{s} - \mathbf{s}^\star, \nabla l^*(\mathbf{s}) - \nabla l^*(\mathbf{s}^\star)\rangle \ge \beta\|\nabla l^*(\mathbf{s}) - \nabla l^*(\mathbf{s}^\star)\|^2_{\mathbf{M}_1^{-1}}, \tag{12}$$

for some $\beta > 0$.

Lemma 1 *When f and l^* have Lipschitz continuous gradients with parameters L_f and L_{l^*}, respectively, we can choose*

$$\beta = \min\left(\lambda_{\min}(\mathbf{P})L_f^{-1}, \frac{\tau}{\sigma}\lambda_{\min}(\mathbf{D} - \theta\tau\sigma\mathbf{A}\mathbf{P}^{-1}\mathbf{A}^\top)L_{l^*}^{-1}\right)$$

such that Assumption 1 is satisfied. When $\mathbf{D}$ and $\mathbf{P}$ are identity matrices, we can simplify it as

$$\beta = \min\left(L_f^{-1}, \frac{\tau}{\sigma}(1 - \theta\tau\sigma\lambda_{\max}(\mathbf{A}\mathbf{A}^\top))L_{l^*}^{-1}\right).$$

The proof for this lemma is simple and omitted.

Remark 1 We choose norms that are different from standard norms for simplicity. They come from the operators $\mathbf{P}$ and $\mathbf{M}$ in (8).

- The condition (11) usually comes from the cocoerciveness of ∇f. It is satisfied with $\beta = \frac{\min_{\mathbf{x}:\|\mathbf{x}\|=1}\|\mathbf{Px}\|}{L_f}$ if $f(\mathbf{x})$ has a Lipschitz continuous gradient with constant L_f [1, Theorem 18.15]. One example of $\mathbf{P}$ is the diagonal matrix when f is separable and the Lipschtiz continuous constants are different for different blocks. By choosing a diagonal matrix $\mathbf{P}$, we can have a fast algorithm. For example, in [22], we let different agents choose different stepsizes to improve the convergence speed.
- Note that the condition (12) depends on θ, which does not exist in the algorithm. We choose to have the same β in (11) and (12) for simplicity. From the definition of $\mathbf{M}_1$, we can see that the condition (12) depends on function l^*, $\mathbf{P}$, $\mathbf{D}$, $\mathbf{A}$, β, θ, τ, and σ. But it is not as complicated as it looks like. Let us assume that $\mathbf{D} = \mathbf{I}$ and $\mathbf{P} = \mathbf{I}$, f and l^* have Lipschitz continuous gradients with L_f and L_{l^*}, respectively. The condition (12) requires:

  $$\beta \le \lambda_{\min}(\mathbf{M}_1)/L_{l^*} = \tau(1-\theta\tau\sigma\|\mathbf{A}\mathbf{A}^\top\|)/(\sigma L_{l^*}).$$

  Therefore, we can also choose a small $\theta \in (3/4, 1]$ to make it valid if a larger β works. By making θ small, we can have a large dual stepsize σ for a given primal stepsize τ. In fact, we do not need to know β explicitly to determine both stepsizes. When we consider both conditions ((11) and (12)) and the condition $\tau < 2\beta$ in Theorem 1, we have:

  $$\tau L_f < 2, \; \sigma L_{l^*} < 2(1-\theta\tau\sigma\lambda_{\max}(\mathbf{A}\mathbf{A}^\top)). \tag{13}$$

  For comparison, the condition in [27] is $\max(\tau,\sigma)\max(L_f, L_{l^*}) < 2(1 - \sqrt{\tau\sigma\|\mathbf{A}\mathbf{A}^\top\|})$. Our condition has two benefits. One is that we consider τ and σ differently and can obtain a large stepsize even when the Lipschitz constants L_f and L_{l^*} have different scales. The other is the introduction of $\theta \in (3/4, 1]$, which may increase the upper bounds for the stepsizes. The best result in this paper comes from choosing a θ that is close to $3/4$ even when $\theta = 1$ is enough for $\mathbf{M}_1$ being positive definite. See the example in Section 2.5.
- (Special cases) The positiveness of $\mathbf{M}_1$ gives an upper bound for $\tau\sigma$ that depends on $\mathbf{P}$, $\mathbf{D}$, and $\mathbf{A}$. The convergence of (9) requires an upper bound for τ that is $\tau < 2\beta$; see Theorem 1. If ∇l^* is fixed for all $\mathbf{s}$, e.g., problem (2), then (12) is satisfied with any $\beta > 0$, and the upper bound of τ depends on $\mathbf{P}$ and L_f only, i.e., $\tau < 2\lambda_{\min}(\mathbf{P})L_f^{-1}$. The condition is strictly weaker than that in [27] and [2] because of the introduction of θ. If ∇f is fixed for all $\mathbf{x}$, e.g., the linear f in linearized ALM, then (11) is satisfied with any $\beta > 0$, and the upper bound for τ depends on σ, $\mathbf{A}$, $\mathbf{D}$, $\mathbf{P}$, and the Lipschitz constant of ∇l^* because of $\mathbf{M}_1$ in (12), i.e., $\sigma < 2\lambda_{\min}(\mathbf{D} - (3/4)\tau\sigma\mathbf{A}\mathbf{P}^{-1}\mathbf{A}^\top)L_{l^*}^{-1}$.

Assumption 2 *Let* $(\mathbf{x}^\star,\mathbf{s}^\star)$ *be any fixed point of (9). There exist* $\mu_f\geq 0$, $\mu_h\geq 0$, *and* $\mu_l\geq 0$, *such that, for any* $\mathbf{x}\in\mathcal{X}$ *and* $\mathbf{s}\in\mathcal{S}$,

$$\langle\mathbf{x}-\mathbf{x}^\star,\nabla f(\mathbf{x})-\nabla f(\mathbf{x}^\star)\rangle\geq\ \mu_f\|\mathbf{x}-\mathbf{x}^\star\|_{\mathbf{P}}^2, \tag{14}$$

$$\langle\mathbf{s}-\mathbf{s}^\star,\mathbf{p}_h(\mathbf{s})-\mathbf{p}_h(\mathbf{s}^\star)\rangle\geq\ \mu_h\|\mathbf{s}-\mathbf{s}^\star\|_{\mathbf{M}_1}^2, \tag{15}$$

$$\langle\mathbf{s}-\mathbf{s}^\star,\nabla l^*(\mathbf{s})-\nabla l^*(\mathbf{s}^\star)\rangle\geq\ \mu_l\|\mathbf{s}-\mathbf{s}^\star\|_{\mathbf{M}_1}^2, \tag{16}$$

where $\mathbf{p}_h(\mathbf{s})\in\partial h^*(\mathbf{s})$ *and* $\mathbf{p}_h(\mathbf{s}^\star)\in\partial h^*(\mathbf{s}^\star)$.

The assumption is satisfied if functions $f(\mathbf{x})$, $h(\mathbf{s})$, and $l(\mathbf{s})$ are convex, and in this case, $\mu_f=\mu_h=\mu_l=0$. We choose the norms $\|\cdot\|_{\mathbf{P}}$ and $\|\cdot\|_{\mathbf{M}_1}$ for the two spaces for simplicity. All the results in this paper also hold for standard norms, but the formulas are complicated. We will need this assumption with positive values to show the linear convergence for strongly convex functions. In this case, because $\mathbf{P}$ and $\mathbf{M}_1$ are positive definite, $\mu_f>0$ (or $\mu_h>0$, $\mu_l>0$) is implied from the strong convexity of the function $f(\mathbf{x})$ (or $g^*(\mathbf{s})$, $l^*(\mathbf{s})$).

2.3 Convergence for general convex functions

First of all, we find a subgradient of h^* at $\mathbf{s}^{k+1}$:

$$\mathbf{q}_h(\mathbf{s}^{k+1}):=\frac{1}{\tau}\mathbf{M}\mathbf{s}^k-\frac{1}{\tau}\mathbf{M}\mathbf{s}^{k+1}+\mathbf{A}\mathbf{x}^{k+1}-\nabla l^*(\mathbf{s}^k)\in\partial h^*(\mathbf{s}^{k+1}). \tag{17}$$

It can be easily obtained from (9), and its proof is omitted here. Let $(\mathbf{x}^\star,\mathbf{s}^\star)$ be any fixed point of (9), and we have a subgradient of h^* at $\mathbf{s}^\star$:

$$\mathbf{q}_h(\mathbf{s}^\star):=\mathbf{A}\mathbf{x}^\star-\nabla l^*(\mathbf{s}^\star)\in\partial h^*(\mathbf{s}^\star). \tag{18}$$

Lemma 2 (Fundamental inequality) *Let* $(\mathbf{x}^\star,\mathbf{s}^\star)$ *be any fixed point of (9), and* $\{(\mathbf{x}^k,\mathbf{s}^k)\}$ *a sequence generated by (9). Then, we have:*

$$\begin{aligned}
&\|(\mathbf{x}^{k+1},\mathbf{s}^{k+1})-(\mathbf{x}^\star,\mathbf{s}^\star)\|_{\mathbf{P},\widetilde{\mathbf{M}}}^2\\
\leq\ &\|(\mathbf{x}^{k},\mathbf{s}^{k})-(\mathbf{x}^\star,\mathbf{s}^\star)\|_{\mathbf{P},\widetilde{\mathbf{M}}}^2-\|\mathbf{s}^k-\mathbf{s}^{k+1}\|_{\mathbf{M}_1}^2\\
&-2\tau\langle\mathbf{s}^{k+1}-\mathbf{s}^\star,\mathbf{q}_h(\mathbf{s}^{k+1})-\mathbf{q}_h(\mathbf{s}^\star)+\nabla l^*(\mathbf{s}^k)-\nabla l^*(\mathbf{s}^\star)\rangle\\
&+2\tau\langle\nabla f(\mathbf{x}^k)-\nabla f(\mathbf{x}^\star),\mathbf{x}^\star-\mathbf{x}^k+(4\theta-3)(\mathbf{x}^k-\mathbf{x}^{k+1})\rangle\\
&-(4\theta-3)\|\mathbf{x}^k-\mathbf{x}^{k+1}\|_{\mathbf{P}}^2+4(1-\theta)\tau^2\|\nabla f(\mathbf{x}^k)-\nabla f(\mathbf{x}^\star)\|_{\mathbf{P}^{-1}}^2.
\end{aligned} \tag{19}$$

Proof The definitions of $\mathbf{q}_h(\mathbf{s}^{k+1})$ and $\mathbf{q}_h(\mathbf{s}^\star)$ in (17) and (18), respectively, and the update of $\mathbf{x}^{k+1}$ in (9b) show:

$$\begin{aligned}
&2\tau\langle\mathbf{s}^{k+1}-\mathbf{s}^\star,\mathbf{q}_h(\mathbf{s}^{k+1})-\mathbf{q}_h(\mathbf{s}^\star)+\nabla l^*(\mathbf{s}^k)-\nabla l^*(\mathbf{s}^\star)\rangle\\
\overset{(17)(18)}{=}\ &2\tau\langle\mathbf{s}^{k+1}-\mathbf{s}^\star,\frac{1}{\tau}\mathbf{M}\mathbf{s}^k-\frac{1}{\tau}\mathbf{M}\mathbf{s}^{k+1}+\mathbf{A}\mathbf{x}^{k+1}-\mathbf{A}\mathbf{x}^\star\rangle\\
=\ &2\langle\mathbf{s}^{k+1}-\mathbf{s}^\star,\mathbf{s}^k-\mathbf{s}^{k+1}\rangle_{\mathbf{M}}+2\tau\langle\mathbf{s}^{k+1}-\mathbf{s}^\star,\mathbf{A}\mathbf{x}^{k+1}-\mathbf{A}\mathbf{x}^\star\rangle\\
=\ &2\langle\mathbf{s}^{k+1}-\mathbf{s}^\star,\mathbf{s}^k-\mathbf{s}^{k+1}\rangle_{\mathbf{M}}+2\tau\langle\mathbf{A}^\top\mathbf{s}^{k+1}-\mathbf{A}^\top\mathbf{s}^\star,\mathbf{x}^{k+1}-\mathbf{x}^\star\rangle
\end{aligned}$$

$$\begin{aligned}
&\overset{(9b)}{=} 2\langle \mathbf{s}^{k+1}-\mathbf{s}^\star, \mathbf{s}^k-\mathbf{s}^{k+1}\rangle_{\mathbf{M}} + 2\langle \mathbf{x}^k-\mathbf{x}^{k+1}, \mathbf{x}^{k+1}-\mathbf{x}^\star\rangle_{\mathbf{P}} \\
&\quad -2\tau\langle \nabla f(\mathbf{x}^k)-\nabla f(\mathbf{x}^\star), \mathbf{x}^{k+1}-\mathbf{x}^\star\rangle \qquad (20)\\
&\quad = \|\mathbf{s}^k-\mathbf{s}^\star\|_{\mathbf{M}}^2 - \|\mathbf{s}^{k+1}-\mathbf{s}^\star\|_{\mathbf{M}}^2 - \|\mathbf{s}^k-\mathbf{s}^{k+1}\|_{\mathbf{M}}^2 \\
&\quad +\|\mathbf{x}^k-\mathbf{x}^\star\|_{\mathbf{P}}^2 - \|\mathbf{x}^{k+1}-\mathbf{x}^\star\|_{\mathbf{P}}^2 - \|\mathbf{x}^k-\mathbf{x}^{k+1}\|_{\mathbf{P}}^2 \\
&\quad +2\tau\langle \nabla f(\mathbf{x}^k)-\nabla f(\mathbf{x}^\star), \mathbf{x}^\star-\mathbf{x}^{k+1}\rangle,
\end{aligned}$$

where we expanded the first two terms in (20) using $2\langle a, b\rangle = \|a+b\|^2 - \|a\|^2 - \|b\|^2$ to obtain the last equality. Therefore, we have:

$$\begin{aligned}
&\|(\mathbf{x}^{k+1}, \mathbf{s}^{k+1}) - (\mathbf{x}^\star, \mathbf{s}^\star)\|_{\mathbf{P},\mathbf{M}}^2 \\
=\ & 2\tau\langle \nabla f(\mathbf{x}^k)-\nabla f(\mathbf{x}^\star), \mathbf{x}^\star-\mathbf{x}^{k+1}\rangle \\
& -2\tau\langle \mathbf{s}^{k+1}-\mathbf{s}^\star, \mathbf{q}_h(\mathbf{s}^{k+1})-\mathbf{q}_h(\mathbf{s}^\star)+\nabla l^*(\mathbf{s}^k)-\nabla l^*(\mathbf{s}^\star)\rangle \\
& +\|(\mathbf{x}^k, \mathbf{s}^k)-(\mathbf{x}^\star, \mathbf{s}^\star)\|_{\mathbf{P},\mathbf{M}}^2 - \|\mathbf{x}^k-\mathbf{x}^{k+1}\|_{\mathbf{P}}^2 - \|\mathbf{s}^k-\mathbf{s}^{k+1}\|_{\mathbf{M}}^2. \qquad (21)
\end{aligned}$$

The fact that $\mathbf{M} = \mathbf{M}_1 - \mathbf{M}_2$ gives us an upper bound for the last term of (21).

$$\begin{aligned}
-\|\mathbf{s}^k-\mathbf{s}^{k+1}\|_{\mathbf{M}}^2 &= -\|\mathbf{s}^k-\mathbf{s}^{k+1}\|_{\mathbf{M}_1}^2 + \|\mathbf{s}^k-\mathbf{s}^{k+1}\|_{\mathbf{M}_2}^2 \\
&= -\|\mathbf{s}^k-\mathbf{s}^{k+1}\|_{\mathbf{M}_1}^2 + \|\mathbf{s}^k-\mathbf{s}^\star+\mathbf{s}^\star-\mathbf{s}^{k+1}\|_{\mathbf{M}_2}^2 \\
&\le -\|\mathbf{s}^k-\mathbf{s}^{k+1}\|_{\mathbf{M}_1}^2 + 2\|\mathbf{s}^k-\mathbf{s}^\star\|_{\mathbf{M}_2}^2 + 2\|\mathbf{s}^{k+1}-\mathbf{s}^\star\|_{\mathbf{M}_2}^2. \qquad (22)
\end{aligned}$$

Adding $2\|\mathbf{s}^{k+1}-\mathbf{s}^\star\|_{\mathbf{M}_2}^2$ onto both sides of (21), recalling that $\widetilde{\mathbf{M}} = \mathbf{M}_1 + \mathbf{M}_2 = \mathbf{M} + 2\mathbf{M}_2$, and combining (22) and (21), we have:

$$\begin{aligned}
&\|(\mathbf{x}^{k+1}, \mathbf{s}^{k+1}) - (\mathbf{x}^\star, \mathbf{s}^\star)\|_{\mathbf{P},\widetilde{\mathbf{M}}}^2 \\
\le\ & 2\tau\langle \nabla f(\mathbf{x}^k)-\nabla f(\mathbf{x}^\star), \mathbf{x}^\star-\mathbf{x}^{k+1}\rangle \\
& -2\tau\langle \mathbf{s}^{k+1}-\mathbf{s}^\star, \mathbf{q}_h(\mathbf{s}^{k+1})-\mathbf{q}_h(\mathbf{s}^\star)+\nabla l^*(\mathbf{s}^k)-\nabla l^*(\mathbf{s}^\star)\rangle \\
& +\|(\mathbf{x}^k, \mathbf{s}^k)-(\mathbf{x}^\star, \mathbf{s}^\star)\|_{\mathbf{P},\widetilde{\mathbf{M}}}^2 - \|\mathbf{x}^k-\mathbf{x}^{k+1}\|_{\mathbf{P}}^2 - \|\mathbf{s}^k-\mathbf{s}^{k+1}\|_{\mathbf{M}_1}^2 \\
& +4\|\mathbf{s}^{k+1}-\mathbf{s}^\star\|_{\mathbf{M}_2}^2. \qquad (23)
\end{aligned}$$

With the definition of $\mathbf{M}_2$, the last term in (23) can be written as:

$$\begin{aligned}
4\|\mathbf{s}^{k+1}-\mathbf{s}^\star\|_{\mathbf{M}_2}^2 &= 4(1-\theta)\|\tau\mathbf{P}^{-1}\mathbf{A}^\top\mathbf{s}^{k+1} - \tau\mathbf{P}^{-1}\mathbf{A}^\top\mathbf{s}^\star\|_{\mathbf{P}}^2 \\
&= 4(1-\theta)\|\mathbf{x}^k - \tau\mathbf{P}^{-1}\nabla f(\mathbf{x}^k) - \mathbf{x}^{k+1} + \tau\mathbf{P}^{-1}\nabla f(\mathbf{x}^\star)\|_{\mathbf{P}}^2 \\
&= 4(1-\theta)\|\mathbf{x}^k-\mathbf{x}^{k+1}\|_{\mathbf{P}}^2 + 4(1-\theta)\tau^2\|\nabla f(\mathbf{x}^k)-\nabla f(\mathbf{x}^\star)\|_{\mathbf{P}^{-1}}^2 \\
&\quad -8(1-\theta)\tau\langle \mathbf{x}^k-\mathbf{x}^{k+1}, \nabla f(\mathbf{x}^k)-\nabla f(\mathbf{x}^\star)\rangle, \qquad (24)
\end{aligned}$$

where the second equality comes from (9b). Then, we plug (24) into (23) and obtain:

$$\begin{aligned}
&\|(\mathbf{x}^{k+1}, \mathbf{s}^{k+1}) - (\mathbf{x}^\star, \mathbf{s}^\star)\|_{\mathbf{P},\widetilde{\mathbf{M}}}^2 \\
\le\ & 2\tau\langle \nabla f(\mathbf{x}^k)-\nabla f(\mathbf{x}^\star), \mathbf{x}^\star-\mathbf{x}^k+(4\theta-3)(\mathbf{x}^k-\mathbf{x}^{k+1})\rangle
\end{aligned}$$

$$
\begin{aligned}
&-2\tau\langle \mathbf{s}^{k+1}-\mathbf{s}^\star, \mathbf{q}_h(\mathbf{s}^{k+1})-\mathbf{q}_h(\mathbf{s}^\star)+\nabla l^*(\mathbf{s}^k)-\nabla l^*(\mathbf{s}^\star)\rangle\\
&+\|(\mathbf{x}^k,\mathbf{s}^k)-(\mathbf{x}^\star,\mathbf{s}^\star)\|^2_{\mathbf{P},\widetilde{\mathbf{M}}}-\|\mathbf{s}^k-\mathbf{s}^{k+1}\|^2_{\mathbf{M}_1}\\
&-(4\theta-3)\|\mathbf{x}^k-\mathbf{x}^{k+1}\|^2_{\mathbf{P}}+4(1-\theta)\tau^2\|\nabla f(\mathbf{x}^k)-\nabla f(\mathbf{x}^\star)\|^2_{\mathbf{P}^{-1}}.
\end{aligned}
$$

The result is proved. □

Lemma 3 *Let* (12) *be satisfied, then*

$$
\begin{aligned}
&-\|\mathbf{s}^k-\mathbf{s}^{k+1}\|^2_{\mathbf{M}_1}-2\tau\langle \mathbf{s}^{k+1}-\mathbf{s}^\star,\nabla l^*(\mathbf{s}^k)-\nabla l^*(\mathbf{s}^\star)\rangle\\
\le&-(1-\tau/(2\beta))\,\|\mathbf{s}^k-\mathbf{s}^{k+1}\|^2_{\mathbf{M}_1}.
\end{aligned}
$$

Proof Because $\mathbf{M}_1$ is positive definite, we have

$$
\begin{aligned}
&-\|\mathbf{s}^k-\mathbf{s}^{k+1}\|^2_{\mathbf{M}_1}-2\tau\langle \mathbf{s}^{k+1}-\mathbf{s}^\star,\nabla l^*(\mathbf{s}^k)-\nabla l^*(\mathbf{s}^\star)\rangle\\
=&-\|\mathbf{s}^k-\mathbf{s}^{k+1}\|^2_{\mathbf{M}_1}-2\tau\langle \mathbf{s}^{k+1}-\mathbf{s}^k,\nabla l^*(\mathbf{s}^k)-\nabla l^*(\mathbf{s}^\star)\rangle\\
&-2\tau\langle \mathbf{s}^k-\mathbf{s}^\star,\nabla l^*(\mathbf{s}^k)-\nabla l^*(\mathbf{s}^\star)\rangle\\
\le&-\|\mathbf{s}^k-\mathbf{s}^{k+1}\|^2_{\mathbf{M}_1}+\frac{\tau}{2\beta}\|\mathbf{s}^k-\mathbf{s}^{k+1}\|^2_{\mathbf{M}_1}+2\tau\beta\|\nabla l^*(\mathbf{s}^k)-\nabla l^*(\mathbf{s}^\star)\|^2_{\mathbf{M}_1^{-1}}\\
&-2\tau\beta\|\nabla l^*(\mathbf{s}^k)-\nabla l^*(\mathbf{s}^\star)\|^2_{\mathbf{M}_1^{-1}}\\
=&-\|\mathbf{s}^k-\mathbf{s}^{k+1}\|^2_{\mathbf{M}_1}+\frac{\tau}{2\beta}\|\mathbf{s}^k-\mathbf{s}^{k+1}\|^2_{\mathbf{M}_1},
\end{aligned}
$$

where the inequality comes from the Cauchy-Schwarz inequality and (12). □

Theorem 1 *Let Assumption 1 hold,* $\theta\in(3/4,1]$*, and* $\tau\in(0,2\beta)$*. The sequence* $\{(\mathbf{x}^k,\mathbf{s}^k)\}$ *is generated by (9). For any fixed point* $(\mathbf{x}^\star,\mathbf{s}^\star)$ *of (9), we have:*

$$
\begin{aligned}
&\|(\mathbf{x}^{k+1},\mathbf{s}^{k+1})-(\mathbf{x}^\star,\mathbf{s}^\star)\|^2_{\mathbf{P},\widetilde{\mathbf{M}}}-\|(\mathbf{x}^k,\mathbf{s}^k)-(\mathbf{x}^\star,\mathbf{s}^\star)\|^2_{\mathbf{P},\widetilde{\mathbf{M}}}\\
\le&-\left(1-\tfrac{\tau}{2\beta}\right)\|\mathbf{s}^k-\mathbf{s}^{k+1}\|^2_{\mathbf{M}_1}-\tfrac{(4\theta-3)(2\beta-\tau)}{2\beta-4(1-\theta)\tau}\|\mathbf{x}^k-\mathbf{x}^{k+1}\|^2_{\mathbf{P}}.
\end{aligned}
\tag{25}
$$

Proof Applying Lemma 3 and h being convex to the inequality (19) in Lemma 2 gives:

$$
\begin{aligned}
&\|(\mathbf{x}^{k+1},\mathbf{s}^{k+1})-(\mathbf{x}^\star,\mathbf{s}^\star)\|^2_{\mathbf{P},\widetilde{\mathbf{M}}}\\
\le&\|(\mathbf{x}^k,\mathbf{s}^k)-(\mathbf{x}^\star,\mathbf{s}^\star)\|^2_{\mathbf{P},\widetilde{\mathbf{M}}}-(1-\tau/(2\beta))\,\|\mathbf{s}^k-\mathbf{s}^{k+1}\|^2_{\mathbf{M}_1}\\
&+\underbrace{2\tau\langle\nabla f(\mathbf{x}^k)-\nabla f(\mathbf{x}^\star),\mathbf{x}^\star-\mathbf{x}^k\rangle}_{A}+4(1-\theta)\tau^2\|\nabla f(\mathbf{x}^k)-\nabla f(\mathbf{x}^\star)\|^2_{\mathbf{P}^{-1}}\\
&-(4\theta-3)\|\mathbf{x}^k-\mathbf{x}^{k+1}\|^2_{\mathbf{P}}+\underbrace{2\tau(4\theta-3)\langle\nabla f(\mathbf{x}^k)-\nabla f(\mathbf{x}^\star),\mathbf{x}^k-\mathbf{x}^{k+1}\rangle}_{B}.
\end{aligned}
\tag{26}
$$

Next, we bound terms A and B. For term A, the assumption (11) implies

$$
2\tau\langle\nabla f(\mathbf{x}^k)-\nabla f(\mathbf{x}^\star),\mathbf{x}^\star-\mathbf{x}^k\rangle\le-2\tau\beta\|\nabla f(\mathbf{x}^k)-\nabla f(\mathbf{x}^\star)\|^2_{\mathbf{P}^{-1}},\tag{27}
$$

and the Cauchy-Schwarz inequality applied to term B implies

$$\begin{aligned}&2\tau(4\theta-3)\langle\nabla f(\mathbf{x}^k)-\nabla f(\mathbf{x}^\star),\mathbf{x}^k-\mathbf{x}^{k+1}\rangle\\ \leq\ &(2\tau\beta-4(1-\theta)\tau^2)\|\nabla f(\mathbf{x}^k)-\nabla f(\mathbf{x}^\star)\|^2_{\mathbf{P}^{-1}}\\ &+\frac{\tau(4\theta-3)^2}{2\beta-4(1-\theta)\tau}\|\mathbf{x}^k-\mathbf{x}^{k+1}\|^2_{\mathbf{P}},\end{aligned}\tag{28}$$

when $\theta\in(3/4,1]$ and $\tau\in(0,2\beta)$. The inequality holds because $2\beta-4(1-\theta)\tau>0$, owing to the bounds on τ and θ. Plugging (27) and (28) into (26), we have:

$$\begin{aligned}&\|(\mathbf{x}^{k+1},\mathbf{s}^{k+1})-(\mathbf{x}^\star,\mathbf{s}^\star)\|^2_{\mathbf{P},\widetilde{\mathbf{M}}}\\ \leq\ &\|(\mathbf{x}^{k},\mathbf{s}^{k})-(\mathbf{x}^\star,\mathbf{s}^\star)\|^2_{\mathbf{P},\widetilde{\mathbf{M}}}-(1-\tau/(2\beta))\,\|\mathbf{s}^k-\mathbf{s}^{k+1}\|^2_{\mathbf{M}_1}\\ &-(4\theta-3)\|\mathbf{x}^k-\mathbf{x}^{k+1}\|^2_{\mathbf{P}}+\frac{\tau(4\theta-3)^2}{2\beta-4(1-\theta)\tau}\|\mathbf{x}^k-\mathbf{x}^{k+1}\|^2_{\mathbf{P}}\\ =\ &\|(\mathbf{x}^{k},\mathbf{s}^{k})-(\mathbf{x}^\star,\mathbf{s}^\star)\|^2_{\mathbf{P},\widetilde{\mathbf{M}}}-(1-\tau/(2\beta))\,\|\mathbf{s}^k-\mathbf{s}^{k+1}\|^2_{\mathbf{M}_1}\\ &-\frac{(4\theta-3)(2\beta-\tau)}{2\beta-4(1-\theta)\tau}\|\mathbf{x}^k-\mathbf{x}^{k+1}\|^2_{\mathbf{P}}.\end{aligned}$$

The inequality (25) is proved. □

Remark 2 When $\beta=+\infty$, i.e., the Lipschitz constant of $\nabla f(\mathbf{x})$ and $\nabla l^*(\mathbf{s})$ is 0, then (25) becomes:

$$\begin{aligned}&\|(\mathbf{x}^{k+1},\mathbf{s}^{k+1})-(\mathbf{x}^\star,\mathbf{s}^\star)\|^2_{\mathbf{P},\widetilde{\mathbf{M}}}-\|(\mathbf{x}^{k},\mathbf{s}^{k})-(\mathbf{x}^\star,\mathbf{s}^\star)\|^2_{\mathbf{P},\widetilde{\mathbf{M}}}\\ \leq\ &-\|\mathbf{s}^k-\mathbf{s}^{k+1}\|^2_{\mathbf{M}_1}-(4\theta-3)\|\mathbf{x}^k-\mathbf{x}^{k+1}\|^2_{\mathbf{P}}.\end{aligned}$$

This is the key result in [15, Theorem 3.1] for linearized ALM. In [15], the authors also considered the case with a general dual stepsize.

Remark 3 (Large stepsizes) We let $\mathbf{P}=\mathbf{I}$ and $\mathbf{D}=\mathbf{I}$ for simplicity. Consider the problem (2) without function l. We have $\beta=1/L$, where L is the Lipschitz constant of ∇f. Then, we can choose $\tau<2/L$, and $\tau\sigma\leq 4/(3\|\mathbf{A}\mathbf{A}^\top\|)$.

However, for the problem (1) with function l, the choice of the primal stepsize τ also depends on σ because of the operator $\mathbf{M}_1$ in the assumption (12). For this case, how to choose τ and σ is complicated. From Remark 11, if f and l^* have Lipschitz continuous gradients with constants L_f and L_{l^*}, respectively, a sufficient condition for convergence is $\tau L_f<2$ and $\sigma L_{l^*}<2(1-(3/4)\tau\sigma\|\mathbf{A}\mathbf{A}^\top\|)$. Except the same conditions $\tau<2/L$ and $\tau\sigma\leq 4/(3\|\mathbf{A}\mathbf{A}^\top\|)$, there is an additional condition $\sigma<2(1-(3/4)\tau\sigma\|\mathbf{A}\mathbf{A}^\top\|)/L_{l^*}$.

Theorem 2 *Under the assumptions in Theorem 1, the sequence* $\{(\mathbf{x}^k,\mathbf{s}^k)\}$ *converges weakly to a fixed point of* (9). *If the iteration* (9) *is demicompact at* $\mathbf{0}$ [24],[3] *the sequence converges strongly.*

[3] An operator $\mathbf{T}$ is demicompact at $\mathbf{x}\in\mathcal{H}$ if for every bounded sequence $\{\mathbf{x}^k\}_{k\geq 0}$ in $\mathcal{H}$ such that $T\mathbf{x}^k-\mathbf{x}^k\to\mathbf{x}$, there exists a strongly convergent subsequence.

Proof Theorem 1 shows that the sequence $\{(\mathbf{x}^k, \mathbf{s}^k)\}$ is bounded, so weakly convergent subsequences of $\{(\mathbf{x}^k, \mathbf{s}^k)\}$ exist. For any weakly convergent subsequence such that $(\mathbf{x}^{k_i}, \mathbf{s}^{k_i}) \rightharpoonup (\mathbf{x}, \mathbf{s})$, the inequality (25) gives $(\mathbf{x}^{k_i-1} - \mathbf{x}^{k_i}, \mathbf{s}^{k_i-1} - \mathbf{s}^{k_i}) \to 0$. Then, based on the iteration (9), we obtain [1, Fact 1.37]:

$$\begin{aligned}\nabla f(\mathbf{x}^{k_i}) + \mathbf{A}^\top \mathbf{s}^{k_i} &= \frac{1}{\tau}\mathbf{P}(\mathbf{x}^{k_i-1} - \mathbf{x}^{k_i}) + \nabla f(\mathbf{x}^{k_i}) - \nabla f(\mathbf{x}^{k_i-1}) \to \mathbf{0},\\ -\mathbf{A}\mathbf{x}^{k_i} + \mathbf{q}_h(\mathbf{s}^{k_i}) + \nabla l^*(\mathbf{s}^{k_i}) &= \frac{1}{\tau}\mathbf{M}(\mathbf{s}^{k_i-1} - \mathbf{s}^{k_i}) - \nabla l^*(\mathbf{s}^{k_i-1}) + \nabla l^*(\mathbf{s}^{k_i}) \to \mathbf{0}.\end{aligned}$$

Because f, h^*, and l^* are convex, the operator:

$$\begin{bmatrix}\nabla f & \mathbf{A}^\top\\ -\mathbf{A} & \partial h^* + \nabla l^*\end{bmatrix}$$

is maximal monotone. Thus, $(\mathbf{x}, \mathbf{s})$ is a fixed point of (9) because of [1, Proposition 20.33(ii)].

The inequality (25) also shows that the sequence $\{(\mathbf{x}^k, \mathbf{s}^k)\}$ is Fejér monotone with respect to the set of fixed points of (9). Then [1, Theorem 5.5] shows that $\{(\mathbf{x}^k, \mathbf{s}^k)\}$ converges weakly to a fixed point of (9).

The inequality (25) shows that $\{(\mathbf{x}^k, \mathbf{s}^k)\}$ is a bounded sequence and $(\mathbf{x}^{k+1} - \mathbf{x}^k, \mathbf{s}^{k+1} - \mathbf{s}^k) \to 0$. Then, the demicompactness of the iteration in (9) at $\mathbf{0}$ shows that there is a strongly convergent subsequence $(\mathbf{x}^{k_n}, \mathbf{s}^{k_n}) \to (\bar{\mathbf{x}}^\star, \bar{\mathbf{s}}^\star)$, and $(\bar{\mathbf{x}}^\star, \bar{\mathbf{s}}^\star)$ is a fixed point of (9) because this subsequence is also weakly convergent. Then, the inequality (25) shows that the whole sequence $\{(\mathbf{x}^k, \mathbf{s}^k)\}$ converges to the fixed point $(\bar{\mathbf{x}}^\star, \bar{\mathbf{s}}^\star)$. □

Remark 4 When $\mathcal{X}$ and $\mathcal{S}$ are finite dimensional, the sequence $\{(\mathbf{x}^k, \mathbf{s}^k)\}$ converges strongly to a fixed point of (9).

In Theorem 2, we showed the convergence of this primal-dual algorithm without providing the convergence rate. The ergodic sublinear convergence rate is showed for primal-dual algorithms for more general problems [6, 31].

2.4 Linear convergence

In this subsection, we prove the linear convergence of the sequence $\{(\mathbf{x}^k, \mathbf{s}^k)\}$ in Theorem 3 under the additional Assumption 2.

Before showing the linear convergence, we prove the following lemma, which provides a different upper bound for the same object in Lemma 3.

Lemma 4 *Let* (12) *and* (16) *be satisfied, then*

$$\begin{aligned}&-\|\mathbf{s}^k - \mathbf{s}^{k+1}\|^2_{\mathbf{M}_1} - 2\tau\langle \mathbf{s}^{k+1} - \mathbf{s}^\star, \nabla l^*(\mathbf{s}^k) - \nabla l^*(\mathbf{s}^\star)\rangle\\ \le\ &-\|\mathbf{M}_2(\mathbf{s}^{k+1} - \mathbf{s}^k) + \tau\mathbf{A}\mathbf{x}^{k+1} - \tau\mathbf{A}\mathbf{x}^\star - \tau\mathbf{q}_h(\mathbf{s}^{k+1}) + \tau\mathbf{q}_h(\mathbf{s}^\star)\|^2_{\mathbf{M}_1^{-1}} \qquad (29)\\ &-\left(2\tau - \tau^2/\beta\right)\mu_l\|\mathbf{s}^k - \mathbf{s}^*\|^2_{\mathbf{M}_1}.\end{aligned}$$

Proof Because $\mathbf{M}_1$ is positive definite, we have:

$$\begin{aligned}
&-\|\mathbf{s}^k-\mathbf{s}^{k+1}\|^2_{\mathbf{M}_1}-2\tau\langle\mathbf{s}^{k+1}-\mathbf{s}^\star,\nabla l^*(\mathbf{s}^k)-\nabla l^*(\mathbf{s}^\star)\rangle\\
=&-\|\mathbf{s}^k-\mathbf{s}^{k+1}\|^2_{\mathbf{M}_1}-2\tau\langle\mathbf{M}_1^{1/2}(\mathbf{s}^{k+1}-\mathbf{s}^k),\mathbf{M}_1^{-1/2}(\nabla l^*(\mathbf{s}^k)-\nabla l^*(\mathbf{s}^\star))\rangle\\
&-2\tau\langle\mathbf{s}^k-\mathbf{s}^\star,\nabla l^*(\mathbf{s}^k)-\nabla l^*(\mathbf{s}^\star)\rangle\\
=&-\|\mathbf{M}_1^{1/2}(\mathbf{s}^{k+1}-\mathbf{s}^k)+\mathbf{M}_1^{-1/2}\tau(\nabla l^*(\mathbf{s}^k)-\nabla l^*(\mathbf{s}^\star))\|^2\\
&+\tau^2\|\nabla l^*(\mathbf{s}^k)-\nabla l^*(\mathbf{s}^\star)\|^2_{\mathbf{M}_1^{-1}}-2\tau\langle\mathbf{s}^k-\mathbf{s}^\star,\nabla l^*(\mathbf{s}^k)-\nabla l^*(\mathbf{s}^\star)\rangle. \qquad (30)
\end{aligned}$$

The first term on the right-hand side of (30) becomes:

$$\begin{aligned}
&-\|\mathbf{M}_1^{1/2}(\mathbf{s}^{k+1}-\mathbf{s}^k)+\mathbf{M}_1^{-1/2}\tau(\nabla l^*(\mathbf{s}^k)-\nabla l^*(\mathbf{s}^\star))\|^2\\
=\;&-\|\mathbf{M}_1(\mathbf{s}^{k+1}-\mathbf{s}^k)+\tau(\nabla l^*(\mathbf{s}^k)-\nabla l^*(\mathbf{s}^\star))\|^2_{\mathbf{M}_1^{-1}}\\
=\;&-\|\mathbf{M}_2(\mathbf{s}^{k+1}-\mathbf{s}^k)+\mathbf{M}(\mathbf{s}^{k+1}-\mathbf{s}^k)+\tau(\nabla l^*(\mathbf{s}^k)-\nabla l^*(\mathbf{s}^\star))\|^2_{\mathbf{M}_1^{-1}}\\
\overset{(17),(18)}{=}\;&-\|\mathbf{M}_2(\mathbf{s}^{k+1}-\mathbf{s}^k)+\tau\mathbf{A}\mathbf{x}^{k+1}-\tau\mathbf{A}\mathbf{x}^\star-\tau\mathbf{q}_h(\mathbf{s}^{k+1})+\tau\mathbf{q}_h(\mathbf{s}^\star)\|^2_{\mathbf{M}_1^{-1}},
\end{aligned}$$

where the second equality comes from $\mathbf{M}=\mathbf{M}_1-\mathbf{M}_2$.

For the other two terms on the right-hand side of (30), we have:

$$\begin{aligned}
&\tau^2\|\nabla l^*(\mathbf{s}^k)-\nabla l^*(\mathbf{s}^\star)\|^2_{\mathbf{M}_1^{-1}}-2\tau\langle\mathbf{s}^k-\mathbf{s}^\star,\nabla l^*(\mathbf{s}^k)-\nabla l^*(\mathbf{s}^\star)\rangle\\
\overset{(12),(16)}{\le}\;&-(2\tau-\tau^2/\beta)\mu_l\|\mathbf{s}^k-\mathbf{s}^\star\|^2_{\mathbf{M}_1}.
\end{aligned}$$

Combining both inequalities together with (30) gives (29). □

Theorem 3 *Let* $(\mathbf{x}^\star,\mathbf{s}^\star)$ *be a fixed point of (9) and Assumptions 1 and 2 hold. Define* $\widehat{\mathbf{M}}:=(1+2\tau\mu_h)\mathbf{M}_1+\mathbf{M}_2$, *and we have*

$$\|(\mathbf{x}^{k+1},\mathbf{s}^{k+1})-(\mathbf{x}^\star,\mathbf{s}^\star)\|^2_{\mathbf{P},\widehat{\mathbf{M}}}\le\rho_1\|(\mathbf{x}^k,\mathbf{s}^k)-(\mathbf{x}^\star,\mathbf{s}^\star)\|^2_{\mathbf{P},\widehat{\mathbf{M}}}, \qquad (31)$$

where

$$\rho_1=\max\left(\frac{1-(2\tau-\tau^2/\beta)\mu_l+C_1}{1+2\tau\mu_h+C_1},1-(2\tau-\tau^2/\beta)\mu_f\right).$$

Here, $C_1\equiv\|\mathbf{M}_1^{-1/2}\mathbf{M}_2\mathbf{M}_1^{-1/2}\|\ge 0$. *The sequence* $\{(\mathbf{x}^k,\mathbf{s}^k)\}$ *converges linearly to the fixed point* $(\mathbf{x}^\star,\mathbf{s}^\star)$ *with rate* $\rho_1<1$ *if* $\tau\in(0,2\beta)$, $\mu_h+\mu_l>0$, *and* $\mu_f>0$.

Proof Applying Lemma 4 to (19) in Lemma 2 gives:

$$
\begin{aligned}
&\|(\mathbf{x}^{k+1},\mathbf{s}^{k+1})-(\mathbf{x}^\star,\mathbf{s}^\star)\|^2_{\mathbf{P},\widetilde{\mathbf{M}}}\\
\le\ &\|(\mathbf{x}^{k},\mathbf{s}^{k})-(\mathbf{x}^\star,\mathbf{s}^\star)\|^2_{\mathbf{P},\widetilde{\mathbf{M}}}-\left(2\tau-\tau^2/\beta\right)\mu_l\|\mathbf{s}^k-\mathbf{s}^*\|^2_{\mathbf{M}_1}\\
&-2\tau\langle\mathbf{s}^{k+1}-\mathbf{s}^\star,\mathbf{q}_h(\mathbf{s}^{k+1})-\mathbf{q}_h(\mathbf{s}^\star)\rangle\\
&+2\tau\langle\nabla f(\mathbf{x}^k)-\nabla f(\mathbf{x}^\star),\mathbf{x}^\star-\mathbf{x}^k+(4\theta-3)(\mathbf{x}^k-\mathbf{x}^{k+1})\rangle\\
&-(4\theta-3)\|\mathbf{x}^k-\mathbf{x}^{k+1}\|^2_{\mathbf{P}}+4(1-\theta)\tau^2\|\nabla f(\mathbf{x}^k)-\nabla f(\mathbf{x}^\star)\|^2_{\mathbf{P}^{-1}}\\
=\ &\|(\mathbf{x}^{k},\mathbf{s}^{k})-(\mathbf{x}^\star,\mathbf{s}^\star)\|^2_{\mathbf{P},\widetilde{\mathbf{M}}}-\left(2\tau-\tau^2/\beta\right)\mu_l\|\mathbf{s}^k-\mathbf{s}^*\|^2_{\mathbf{M}_1}\\
&-2\tau\langle\mathbf{s}^{k+1}-\mathbf{s}^\star,\mathbf{q}_h(\mathbf{s}^{k+1})-\mathbf{q}_h(\mathbf{s}^\star)\rangle\\
&-2\tau\langle\nabla f(\mathbf{x}^k)-\nabla f(\mathbf{x}^\star),\mathbf{x}^k-\mathbf{x}^*\rangle+\tau^2\|\nabla f(\mathbf{x}^k)-\nabla f(\mathbf{x}^\star)\|^2_{\mathbf{P}^{-1}}\\
&-(4\theta-3)\|\mathbf{x}^k-\mathbf{x}^{k+1}-\tau\mathbf{P}^{-1}(\nabla f(\mathbf{x}^k)-\nabla f(\mathbf{x}^\star))\|^2_{\mathbf{P}}.
\end{aligned}
$$

Note that

$$
\begin{aligned}
&-2\tau\langle\nabla f(\mathbf{x}^k)-\nabla f(\mathbf{x}^\star),\mathbf{x}^k-\mathbf{x}^*\rangle+\tau^2\|\nabla f(\mathbf{x}^k)-\nabla f(\mathbf{x}^\star)\|^2_{\mathbf{P}^{-1}}\\
\overset{(11)}{\le}\ &-(2\tau-\tau^2/\beta)\langle\nabla f(\mathbf{x}^k)-\nabla f(\mathbf{x}^\star),\mathbf{x}^k-\mathbf{x}^*\rangle\\
\overset{(14)}{\le}\ &-(2\tau-\tau^2/\beta)\mu_f\|\mathbf{x}^k-\mathbf{x}^*\|^2_{\mathbf{P}}.
\end{aligned}
$$

Then we have, together with (15):

$$
\begin{aligned}
&\|(\mathbf{x}^{k+1},\mathbf{s}^{k+1})-(\mathbf{x}^\star,\mathbf{s}^\star)\|^2_{\mathbf{P},\widetilde{\mathbf{M}}}\\
\le\ &\|(\mathbf{x}^{k},\mathbf{s}^{k})-(\mathbf{x}^\star,\mathbf{s}^\star)\|^2_{\mathbf{P},\widetilde{\mathbf{M}}}-\left(2\tau-\tau^2/\beta\right)\mu_l\|\mathbf{s}^k-\mathbf{s}^*\|^2_{\mathbf{M}_1}\\
&-2\tau\mu_h\|\mathbf{s}^{k+1}-\mathbf{s}^\star\|^2_{\mathbf{M}_1}-(2\tau-\tau^2/\beta)\mu_f\|\mathbf{x}^k-\mathbf{x}^*\|^2_{\mathbf{P}}.
\end{aligned}
$$

That is

$$
\begin{aligned}
&\|\mathbf{x}^{k+1}-\mathbf{x}^\star\|^2_{\mathbf{P}}+\|\mathbf{s}^{k+1}-\mathbf{s}^\star\|^2_{(1+2\tau\mu_h)\mathbf{M}_1+\mathbf{M}_2}\\
\le\ &(1-(2\tau-\tau^2/\beta)\mu_f)\|\mathbf{x}^k-\mathbf{x}^\star\|^2_{\mathbf{P}}+\|\mathbf{s}^k-\mathbf{s}^\star\|^2_{(1-(2\tau-\tau^2/\beta)\mu_l)\mathbf{M}_1+\mathbf{M}_2}.
\end{aligned}
\tag{32}
$$

For the last term on the right hand of (32), we have:

$$
\begin{aligned}
&\|\mathbf{s}^k-\mathbf{s}^\star\|^2_{(1-(2\tau-\tau^2/\beta)\mu_l)\mathbf{M}_1+\mathbf{M}_2}\\
=\ &\|\mathbf{M}_1^{1/2}(\mathbf{s}^k-\mathbf{s}^\star)\|^2_{(1-(2\tau-\tau^2/\beta)\mu_l)\mathbf{I}+\mathbf{M}_1^{-1/2}\mathbf{M}_2\mathbf{M}_1^{-1/2}}\\
\le\ &\frac{1-(2\tau-\tau^2/\beta)\mu_l+C_1}{1+2\tau\mu_h+C_1}\|\mathbf{M}_1^{1/2}(\mathbf{s}^k-\mathbf{s}^\star)\|^2_{(1+2\tau\mu_h)\mathbf{I}+\mathbf{M}_1^{-1/2}\mathbf{M}_2\mathbf{M}_1^{-1/2}}\\
=\ &\frac{1-(2\tau-\tau^2/\beta)\mu_l+C_1}{1+2\tau\mu_h+C_1}\|\mathbf{s}^k-\mathbf{s}^\star\|^2_{(1+2\tau\mu_h)\mathbf{M}_1+\mathbf{M}_2}.
\end{aligned}
$$

Therefore, the inequality (31) is proved. □

Note that paper [7] proves the linear convergence rate for the case with $l^*(\mathbf{s}) \equiv 0$ and $\mathbf{M}_2 = 0$ as

$$\max\left(1 - \frac{\min_{\mathbf{x}:\|\mathbf{x}\|=1} \|\mathbf{A}^\top\mathbf{A}\mathbf{x}\|}{\|\mathbf{A}^\top\mathbf{A}\|}, 1 - (2\tau - \tau^2/\beta)\mu_f\right)$$

under the additional assumption that $\mathbf{A}^\top\mathbf{A}$ is surjective. However, $\mu_h > 0$ is not required.

Next, we compare this result with the linear convergence rate of Condat-Vu in [2] by letting $h^*(\mathbf{s}) \equiv 0$ and $\mathbf{M}_2 = \mathbf{0}$, $\mathbf{P} = \mathbf{I}$, $\mathbf{D} = \mathbf{I}$. For simplicity, we assume that f and l^* are both μ-strongly convex and have L-Lipschitz continuous gradients. The linear convergence rate of Condat-Vu is

$$\frac{4}{4 + \min\left(\frac{\mu^2}{L^2}, \sqrt{\frac{\mu^2}{\|\mathbf{A}\mathbf{A}^\top\|}}\right)},$$

with the primal and dual stepsizes in the order of μ/L^2. However, if we let $\tau = \sigma$ in (9), then we have $\mu_l \geq \mu = \mu_f$ and $\beta = (1 - \tau^2\|\mathbf{A}\mathbf{A}^\top\|)/L$ in Assumptions 1 and 2. In addition, we let $\tau = \beta$, then the linear convergence rate in Theorem 3 becomes:

$$1 - \tau\mu = 1 - \frac{2\mu}{\sqrt{L^2 + 4\|\mathbf{A}\mathbf{A}^\top\|} + L}.$$

We can see that the linear convergence rate of (9) is much better than that of Condata-Vu in [2].

2.5 Tight upper bound for the stepsizes

A very simple example was provided in [15] to show the upper bound's tightness for a case without infimal convolution. In this subsection, we provide another example to show the tightness for a case with infimal convolution. This result will be applied to decentralized consensus optimization in the next section. Given a self-adjoint positive definite operator $\mathbf{D}$, we consider the following optimization problem:

$$\underset{\mathbf{x}}{\text{minimize}}\ \mathbf{a}^\top\mathbf{x} + \frac{\mathbf{x}^\top\mathbf{A}^\top\mathbf{D}^{-1}\mathbf{A}\mathbf{x}}{2}.$$

It is a special case of (1) with $f(\mathbf{x}) = \mathbf{a}^\top\mathbf{x}$, $h^*(\mathbf{y}) = 0$, and $l^*(\mathbf{y}) = \mathbf{y}^\top\mathbf{D}\mathbf{y}/2$. The primal-dual iteration (9) after a change of order is:

$$\begin{aligned}
\mathbf{x}^{k+1} &= \mathbf{x}^k - \tau\mathbf{P}^{-1}\mathbf{a} - \tau\mathbf{P}^{-1}\mathbf{A}^\top\mathbf{s}^k,\\
\mathbf{s}^{k+1} &= (\mathbf{I} - \tau\sigma\mathbf{D}^{-1}\mathbf{A}\mathbf{P}^{-1}\mathbf{A}^\top - \sigma\mathbf{I})\mathbf{s}^k + \sigma\mathbf{D}^{-1}\mathbf{A}\mathbf{x}^{k+1} - \tau\sigma\mathbf{D}^{-1}\mathbf{A}\mathbf{P}^{-1}\mathbf{a}.
\end{aligned}$$

Denote $\widetilde{\mathbf{D}} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{P}^{-1}\mathbf{A}^\top\mathbf{D}^{-1/2}$. Then, the iteration is equivalent to

$$\begin{bmatrix}\mathbf{D}^{-1/2}\mathbf{A}\mathbf{x}^{k+1}\\ \mathbf{D}^{1/2}\mathbf{s}^{k+1}\end{bmatrix} = \begin{bmatrix}\mathbf{I} & -\tau\widetilde{\mathbf{D}}\\ \sigma\mathbf{I} & (1-\sigma)\mathbf{I} - 2\tau\sigma\widetilde{\mathbf{D}}\end{bmatrix}\begin{bmatrix}\mathbf{D}^{-1/2}\mathbf{A}\mathbf{x}^{k}\\ \mathbf{D}^{1/2}\mathbf{s}^{k}\end{bmatrix} - \begin{bmatrix}\tau\mathbf{D}^{-1/2}\mathbf{A}\mathbf{P}^{-1}\mathbf{a}\\ 2\tau\sigma\mathbf{D}^{-1/2}\mathbf{A}\mathbf{P}^{-1}\mathbf{a}\end{bmatrix}.$$

The convergence of this iteration for any given initial $(\mathbf{s}^0, \mathbf{x}^0)$ requires the magnitudes of the eigenvalues of the operator:

$$\begin{bmatrix} \mathbf{I} & -\tau\widetilde{\mathbf{D}} \\ \sigma\mathbf{I} & (1-\sigma)\mathbf{I} - 2\tau\sigma\widetilde{\mathbf{D}} \end{bmatrix}$$

being less than 1. Since $\widetilde{\mathbf{D}}$ is self-adjoint, we need the magnitudes of the eigenvalues: of

$$\widetilde{\mathbf{M}} := \begin{bmatrix} 1 & -\tau\lambda \\ \sigma & 1-\sigma-2\tau\sigma\lambda \end{bmatrix}$$

being less than 1 for all λ being the eigenvalues of $\widetilde{\mathbf{D}}$. We calculate the determinant of $\widetilde{\mathbf{M}} - d\mathbf{I}$ for any d below:

$$\det(\widetilde{\mathbf{M}} - d\mathbf{I}) = d^2 - (2-\sigma-2\tau\sigma\lambda)d + (1-\sigma-\tau\sigma\lambda).$$

Particularly, the convergence requires $\det(\widetilde{\mathbf{M}} + \mathbf{I}) > 0$, that is

$$1 + (2-\sigma-2\tau\sigma\lambda) + (1-\sigma-\tau\sigma\lambda) = 4 - 3\tau\sigma\lambda - 2\sigma > 0.$$

It is equivalent to

$$\sigma < 2\left(1 - \frac{3}{4}\tau\sigma\|\widetilde{\mathbf{D}}\|\right).$$

On the other hand, we proved the convergence of the primal-dual algorithm under the condition:

$$\tau < 2\beta = 2\lambda_{\min}(\mathbf{D}^{-1/2}\mathbf{M}_1\mathbf{D}^{-1/2}) = \frac{2\tau}{\sigma}(1 - \theta\tau\sigma\|\widetilde{\mathbf{D}}\|)$$

for some $\theta \in (3/4, 1]$. It shows that the upper bounds for the stepsizes in this paper are optimal.

3 Application in decentralized consensus optimization

In this section, we first show that algorithm (9) recovers PG-EXTRA [26] for decentralized consensus optimization. Then, we provide its convergence result under a weaker condition than that in [26] and a tight upper bound for the stepsize. Note that PG-EXTRA was shown to be equivalent to Condat-Vu for a problem without infimal convolution [28], but this equivalence can not give the weaker condition for convergence and the tight upper bound for the stepsize.

We use the same notation as [26]. The decentralized consensus problem is

$$\underset{x\in\mathbf{R}^p}{\text{minimize}} \sum_{i=1}^{n} s_i(x) + r_i(x),$$

where $s_i : \mathbf{R}^p \to \mathbf{R}$ and $r_i : \mathbf{R}^p \to (-\infty, +\infty]$ are proper lsc convex functions held privately by the node i to encode the node's objective function. The objective of decentralized consensus is minimizing the sum of all private objective functions while using information exchange between neighboring nodes in a network. Here, s_i

has a Lipschitz continuous gradient with parameter $L > 0$ and the proximal mapping of r_i is simple. We let x_i be one copy of x kept at node i. These $\{x_i\}_{i=1}^n$ are not the same in general, and we say that it is consensual if they are the same. Stacking all the copies together, we define:

$$\mathbf{x} := \begin{pmatrix} - & x_1^\top & - \\ - & x_2^\top & - \\ & \vdots & \\ - & x_n^\top & - \end{pmatrix} \in \mathbf{R}^{n\times p},$$

and

$$s(\mathbf{x}) = \sum_{i=1}^n s_i(x_i), \quad r(\mathbf{x}) = \sum_{i=1}^n r_i(x_i).$$

Then, the decentralized consensus problem becomes:

$$\underset{\mathbf{x}}{\text{minimize}} \quad s(\mathbf{x}) + r(\mathbf{x}), \text{ subject to } x_1 = x_2 = \cdots = x_n.$$

The gradient of s at $\mathbf{x}$ is written in the following matrix form:

$$\nabla s(\mathbf{x}) := \begin{pmatrix} - & (\nabla s_1(x_1))^\top & - \\ - & (\nabla s_2(x_2))^\top & - \\ & \vdots & \\ - & (\nabla s_n(x_n))^\top & - \end{pmatrix} \in \mathbf{R}^{n\times p},$$

and $\|\cdot\|_F$ is the Frobenius norm for a matrix in $\mathbf{R}^{n\times p}$. One iteration of PG-EXTRA reads as:

$$\mathbf{z}^{k+1} = \mathbf{z}^k - \mathbf{x}^k + \frac{\mathbf{I}+\mathbf{W}}{2}(2\mathbf{x}^k - \mathbf{x}^{k-1}) - \alpha\nabla s(\mathbf{x}^k) + \alpha\nabla s(\mathbf{x}^{k-1}), \tag{33a}$$

$$\mathbf{x}^{k+1} = \arg\min_{\mathbf{x}} \ r(\mathbf{x}) + \frac{1}{2\alpha}\|\mathbf{x} - \mathbf{z}^{k+1}\|_F^2, \tag{33b}$$

where α is the stepsize and $\mathbf{W}$ is a symmetric matrix that represents information exchange between neighboring nodes. We have $\mathbf{I} - \mathbf{W}$ being positive semidefinite, so we can find $\mathbf{A}$ such that $\mathbf{I} - \mathbf{W} = \mathbf{A}\mathbf{A}^\top$. In addition, we assume that $\mathbf{Null}(\mathbf{A}^\top) = \mathbf{Null}(\mathbf{I} - \mathbf{W}) = \mathbf{span}(\mathbf{1}_{n\times 1})$, which means that $\mathbf{A}^\top\mathbf{x} = \mathbf{0}$ is equivalent to $x_1 = x_2 = \cdots = x_n$. Therefore, the decentralized consensus problem becomes

$$\underset{\mathbf{x}}{\text{minimize}} \quad s(\mathbf{x}) + r(\mathbf{x}) \text{ subject to } \mathbf{A}^\top\mathbf{x} = \mathbf{0}.$$

The equivalence between PG-EXTRA and Condat-Vu can be obtained via considering the primal problem with an indicator function for the constraint [28]. Here, we consider its dual problem in the following form:

$$\underset{\mathbf{y}}{\text{minimize}} \ r^*\square s^*(\mathbf{A}\mathbf{y}), \tag{34}$$

where r^* and s^* are convex conjugate functions of r and s, respectively. We apply (9) to (34) ($h \Rightarrow r^*,\ l \Rightarrow s^*,\ \mathbf{x} \Rightarrow \mathbf{y},\ \mathbf{s} \Rightarrow \mathbf{t}$) and arrive at:

$$\mathbf{z}^{k+1} = (\mathbf{I} - \tau\sigma\mathbf{A}\mathbf{A}^\top)\mathbf{t}^k + \sigma\mathbf{A}\mathbf{y}^k - \sigma\nabla s(\mathbf{t}^k), \tag{35a}$$

$$\mathbf{t}^{k+1} = \arg\min_{\mathbf{t}} \{r(\mathbf{t}) + \frac{1}{2\sigma}\|\mathbf{t} - \mathbf{z}^{k+1}\|_F^2\}, \tag{35b}$$

$$\mathbf{y}^{k+1} = \mathbf{y}^k - \tau\mathbf{A}^\top\mathbf{t}^{k+1}. \tag{35c}$$

Combining (35a) and (35c), we get:

$$\mathbf{z}^{k+1} = \mathbf{z}^k - \mathbf{t}^k + (\mathbf{I} - \tau\sigma\mathbf{A}\mathbf{A}^\top)(2\mathbf{t}^k - \mathbf{t}^{k-1}) - \sigma\nabla s(\mathbf{t}^k) + \sigma\nabla s(\mathbf{t}^{k-1}). \tag{36}$$

We let $\tau\sigma = \frac{1}{2}$ and $\sigma = \alpha$, then (36) is exactly (33a) with $\mathbf{t} \Rightarrow \mathbf{x}$. Because $\mathbf{M} = 2\tau^2(\mathbf{I} - (1/2)\mathbf{A}\mathbf{A}^\top) = \tau^2(\mathbf{I} + \mathbf{W})$ is positive definite, we can let $\mathbf{M}_1 = \mathbf{M}$. If $\{\nabla s_i(x)\}_{i=1}^n$ are Lipschitz continuous with constant $L > 0$, the other condition for convergence is:

$$\tau < 2\beta \leq \frac{2}{L}\lambda_{\min}(\mathbf{M}_1) = \frac{2\tau^2}{L}\lambda_{\min}(\mathbf{I} + \mathbf{W}),$$

where the second inequality comes from:

$$\langle\nabla s(\tilde{\mathbf{x}}) - \nabla s(\bar{\mathbf{x}}), \tilde{\mathbf{x}} - \bar{\mathbf{x}}\rangle \geq \frac{1}{L}\|\tilde{\mathbf{x}} - \bar{\mathbf{x}}\|^2 \geq \frac{1}{L}\lambda_{\min}(\mathbf{M}_1)\|\tilde{\mathbf{x}} - \bar{\mathbf{x}}\|_{\mathbf{M}_1^{-1}}^2.$$

Therefore, we obtain the condition on the stepsize:

$$\alpha = \frac{1}{2\tau} < \lambda_{\min}(\mathbf{I} + \mathbf{W})/L.$$

This is exactly the upper bound in [26].

The previous upper bound is obtained with $\theta = 1$. As we mentioned before, we can choose θ to be close to $3/4$ to obtain large stepsizes. By letting $\theta = 3/4 + \epsilon$ with an arbitrary small $\epsilon > 0$, we have $\mathbf{M}_1 = 2\tau^2(\mathbf{I} - (3/4 + \epsilon)(1/2)\mathbf{A}\mathbf{A}^\top)$ and $\mathbf{M}_2 = (1/4 - \epsilon)\tau^2\mathbf{A}\mathbf{A}^\top$. Then, a larger upper bound for the stepsize:

$$\begin{aligned}\alpha = \frac{1}{2\tau} &\leq \lambda_{\min}(2\mathbf{I} - (3/4 + \epsilon)\mathbf{A}\mathbf{A}^\top)/L\\ &< \lambda_{\min}(2\mathbf{I} - (3/4)\mathbf{A}\mathbf{A}^\top)/L = ((3/4)\lambda_{\min}(\mathbf{I} + \mathbf{W}) + 1/2)/L,\end{aligned}$$

is derived.

The new relaxed condition for $\mathbf{W}$ is $\mathbf{M}_1 = \tau^2(2\mathbf{I} - (3/4 + \epsilon)\mathbf{A}\mathbf{A}^\top) = \tau^2((5/4 - \epsilon)\mathbf{I} + (3/4 + \epsilon)\mathbf{W})$ being positive definite. That is $5\mathbf{I} + 3\mathbf{W}$ is positive definite. Also, the special example in Section 2.5 shows that the condition for the stepsize of PG-EXTRA can not be weakened. Its linear convergence without $\{r_i\}$ is discussed in [21] under the relaxed condition for $\mathbf{W}$ and stepsize.

4 Conclusion

In this paper, we consider the primal-dual algorithm in [7, 12, 23] to solve the problem $f(\mathbf{x}) + h\Box l(\mathbf{x})$ and show its convergence under a weaker condition. We provide an example to show that this condition can not be weakened for a general problem. This result recovers and is more general than the positive-indefinite linear ALM proposed in [15]. Then, we apply this result to decentralized consensus optimization and obtain the tight upper bound for the stepsize in PG-EXTRA.

Acknowledgments The authors would like to thank the anonymous reviewers for the helpful comments and suggestions that improve this paper.

Funding This work was supported in part by the National Science Foundation (NSF) grants DMS-1621798 and DMS-2012439, the Natural Science Foundation of China grant 62001167.

References

1. Bauschke, H.H., Combettes, P.L.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces. Springer, New York (2011)
2. Boţ, R.I., Csetnek, E.R., Heinrich, A., Hendrich, C.: On the convergence rate improvement of a primal-dual splitting algorithm for solving monotone inclusion problems. Math. Program. **150**(2), 251–279 (2015)
3. Bot, R.I., Hendrich, C.: A douglas–Rachford type primal-dual method for solving inclusions with mixtures of composite and parallel-sum type monotone operators. SIAM J. Optim. **23**(4), 2541–2565 (2013)
4. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. J. Math. Imaging Vis. **40**(1), 120–145 (2011)
5. Chambolle, A., Pock, T.: An introduction to continuous optimization for imaging. Acta Numer. **25**, 161–319 (2016)
6. Chambolle, A., Pock, T.: On the ergodic convergence rates of a first-order primal–dual algorithm. Math. Program. **159**(1-2), 253–287 (2016)
7. Chen, P., Huang, J., Zhang, X.: A primal–dual fixed point algorithm for convex separable minimization with applications to image restoration. Inverse Probl. **29**(2), 025011 (2013)
8. Chen, P., Huang, J., Zhang, X.: A primal-dual fixed point algorithm for minimization of the sum of three convex separable functions. Fixed Point Theory and Applications **2016**(1), 54 (2016)
9. Combettes, P.L., Pesquet, J.C.: Primal-dual splitting algorithm for solving inclusions with mixtures of composite, Lipschitzian, and parallel-sum type monotone operators. Set-Valued Var Anal **20**(2), 307–330 (2012)
10. Condat, L.: A primal–dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms. J. Optim. Theory Appl. **158**(2), 460–479 (2013)
11. Davis, D., Yin, W.: A three-operator splitting scheme and its optimization applications. Set-Valued Var Anal **25**(4), 829–858 (2017)
12. Drori, Y., Sabach, S., Teboulle, M.: A simple algorithm for a class of nonsmooth convex–concave saddle-point problems. Oper. Res. Lett. **43**(2), 209–214 (2015)
13. Hamedani, E.Y., Aybat, N.S.: A primal-dual algorithm for general convex-concave saddle point problems. (2018)
14. Hamedani, E.Y., Jalilzadeh, A., Aybat, N., Shanbhag, U.: Iteration complexity of randomized primal-dual methods for convex-concave saddle point problems. arXiv:1806.04118 (2018)
15. He, B., Ma, F., Yuan, X.: Optimal proximal augmented Lagrangian method and its application to full Jacobian splitting for multi-block separable convex minimization problems. IMA J. Numer. Anal. **40**(2), 1188–1216 (2020)
16. Hien, L.T.K., Zhao, R., Haskell, W.B.: An inexact primal-dual smoothing framework for large-scale non-bilinear saddle point problems. arXiv:1711.03669 (2017)
17. Jaggi, M., Smith, V., Takác, M., Terhorst, J., Krishnan, S., Hofmann, T., Jordan, M.I.: Communication-Efficient Distributed Dual Coordinate Ascent. In: Advances in Neural Information Processing Systems, pp. 3068–3076 (2014)

18. Ko, S., Yu, D., Won, J.H.: Easily parallelizable and distributable class of algorithms for structured sparsity, with optimal acceleration. J. Comput. Graph. Stat. **28**(4), 821–833 (2019)
19. Komodakis, N., Pesquet, J.C.: Playing with duality: an overview of recent primal-dual approaches for solving large-scale optimization problems. IEEE Signal Process. Mag. **32**(6), 31–54 (2015)
20. Latafat, P., Patrinos, P.: Asymmetric forward–backward–adjoint splitting for solving monotone inclusions involving three operators. Comput. Optim. Appl. **68**(1), 57–93 (2017)
21. Li, Y., Yan, M.: On linear convergence of two decentralized algorithms. J. Optim. Theory Appl. (2019)
22. Li, Z., Shi, W., Yan, M.: A decentralized proximal-gradient method with network independent step-sizes and separated convergence rates. IEEE Trans. Signal Process. **67**(17), 4494–4506 (2019)
23. Loris, I., Verhoeven, C.: On a generalization of the iterative soft-thresholding algorithm for the case of non-separable penalty. Inverse Probl. **27**(12), 125007 (2011)
24. Petryshyn, W.: Construction of fixed points of demicompact mappings in Hilbert space. J. Math. Anal. Appl. **14**(2), 276–284 (1966)
25. Ryu, E.K., Boyd, S.: Primer on monotone operator methods. Appl Comput Math **15**(1), 3–43 (2016)
26. Shi, W., Ling, Q., Wu, G., Yin, W.: A proximal gradient algorithm for decentralized composite optimization. IEEE Trans. Signal Process. **63**(22), 6013–6023 (2015)
27. Vũ, B.C.: A splitting algorithm for dual monotone inclusions involving cocoercive operators. Adv. Comput. Math. **38**(3), 667–681 (2013)
28. Wu, T., Yuan, K., Ling, Q., Yin, W., Sayed, A.H.: Decentralized consensus optimization with asynchrony and delays. IEEE Trans. Signal Inf. Process Netw. **4**(2), 293–307 (2018)
29. Xu, Y.: First-order methods for constrained convex programming based on linearized augmented Lagrangian function. INFORMS Journal on Optimization to appear (2020)
30. Xu, Y.: Primal-dual stochastic gradient method for convex programs with many functional constraints. SIAM J. Optim. **30**(2), 1664–1692 (2020)
31. Yan, M.: A new primal-dual algorithm for minimizing the sum of three functions with a linear operator. J. Sci. Comput. **76**(3), 1698–1717 (2018)
32. Yang, J., Yuan, X.: Linearized augmented Lagrangian and alternating direction methods for nuclear norm minimization. Math Comput **82**(281), 301–329 (2013)