A New Insight on Augmented Lagrangian Method with Applications in Machine Learning

Jianchao Bai [*] Linyuan Jia [†] Zheng Peng [‡]

Abstract. By exploiting double-penalty terms for the primal subproblem, we develop a novel relaxed augmented Lagrangian method for solving a family of convex optimization problems subject to equality or inequality constraints. This new method is then extended to solve a general multi-block separable convex optimization problem, and two related primal-dual hybrid gradient algorithms are also discussed. Convergence results about the sublinear and linear convergence rates are established by variational characterizations for both the saddle-point of the problem and the first-order optimality conditions of involved subproblems. A large number of experiments on testing the linear support vector machine problem and the robust principal component analysis problem arising from machine learning indicate that our proposed algorithms perform much better than several state-of-the-art algorithms.

Keywords: convex optimization, augmented Lagrangian method, relaxation step, convergence complexity, machine learning

Mathematics Subject Classification(2010): 65K10; 65Y20; 90C25

1 Introduction

An advanced interesting work, that is the Balanced Augmented Lagrangian Method (abbreviated by B-ALM) proposed by He-Yuan [22], aims to solve the following convex optimization problem subject to linear equality or inequality constraints:

$$\min\left\{\theta(\mathbf{x}) \mid A\mathbf{x} = b \text{ (or } \geq b),\ \mathbf{x} \in \mathcal{X}\right\}, \tag{1}$$

where $\theta : \mathcal{R}^n \to \mathcal{R}$ is a closed proper convex function; $\mathcal{X} \subseteq \mathcal{R}^n$ is a closed convex set; $A \in \mathcal{R}^{m\times n}$ and $b \in \mathcal{R}^m$ are given. Hereafter, the symbols $\mathcal{R}, \mathcal{R}^n(\mathcal{R}^n_+), \mathcal{R}^{m\times n}$ denote the sets of real numbers, n dimensional real (nonnegative) column vectors, and $m \times n$ real matrices, respectively. The bold $\mathbf{I}$ denotes the identity matrix and $\mathbf{0}$ stands for zero matrix/vector with proper dimensions. $Q \succ \mathbf{0}$ means Q is symmetric positive definite matrix, and $\nabla f(x)$ denotes the gradient of differentiable function f at x. We use $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ to denote the standard

[*] Research & Development Institute of Northwestern Polytechnical University in Shenzhen, Shenzhen 518057, China; School of Mathematics and Statistics, Northwestern Polytechnical University, Xi'an 710129, China. (`jianchaobai@nwpu.edu.cn`).

[†] School of Power and Energy, Northwestern Polytechnical University, Xi'an 710129, China. (`jialinyuan@nwpu.edu.cn`).

[‡] Corresponding author. School of Mathematics and Computational Science, Xiangtan University, Xiangtan 411105, China. (`pzheng@xtu.edu.cn`).

Euclidean norm and inner product, respectively. Given $H \succ \mathbf{0}$, we define $\|\mathbf{w}\|_H = \sqrt{\langle \mathbf{w}, H\mathbf{w}\rangle}$. Throughout this paper, the solution set of the problem (1) is assumed to be nonempty.

A fundamental tool to solve the problem (1) is the Augmented Lagrangian Method (ALM, [16, 33]) by exploring the following two steps:

$$\begin{cases} \mathbf{x}^{k+1} = \arg\min\limits_{\mathbf{x}\in\mathcal{X}} L(\mathbf{x}, \lambda^k) + \frac{r}{2}\|A\mathbf{x} - b\|^2, \\ \lambda^{k+1} = \lambda^k - r(A\mathbf{x}^{k+1} - b), \end{cases}$$

where $r > 0$ denotes penalty parameter for the violation of the linear constraints and $L(\mathbf{x}, \lambda) = \theta(\mathbf{x}) - \langle \lambda, A\mathbf{x} - b\rangle$ denotes the corresponding Lagrangian function. With simple algebra, the core subproblem of ALM amounts to

$$\mathbf{x}^{k+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \theta(\mathbf{x}) - \left\langle \mathbf{x}, A^{\mathsf{T}}\lambda^k + rA^{\mathsf{T}}b \right\rangle + \frac{1}{2}\|x\|^2_{rA^{\mathsf{T}}A}. \tag{2}$$

It is often complicated and has no efficient solution if without employing some inexact approximate techniques. As described in [22], the B-ALM reads

$$\begin{cases} \mathbf{x}^{k+1} = \arg\min\limits_{\mathbf{x}\in\mathcal{X}} L(\mathbf{x}, \lambda^k) + \frac{r}{2}\left\|\mathbf{x} - \mathbf{x}^k\right\|^2, \\ \lambda^{k+1} = \arg\max L([2\mathbf{x}^{k+1} - \mathbf{x}^k], \lambda) - \frac{1}{2}\left\|\lambda - \lambda^k\right\|^2_{\frac{1}{r}AA^{\mathsf{T}} + \delta\mathbf{I}}, \end{cases}$$

whose convergence depends on the positive definiteness of the block matrix $\begin{bmatrix} r\mathbf{I} & A^{\mathsf{T}} \\ A & AA^{\mathsf{T}}/r + \delta\mathbf{I} \end{bmatrix}$ for any $r, \delta > 0$. One may use a general form $Q + AA^{\mathsf{T}}/r$ for any $Q \succ \mathbf{0}$ to replace the above lower-right block so as to guarantee the convergence. The major merit of B-ALM is that it greatly weakens the convergence conditions of some ALM and related first-order splitting algorithms [5, 11, 14, 18, 20, 31, 39]. Namely, the parameter r does not depend on $\rho(A^{\mathsf{T}}A)$, where $\rho(\cdot)$ represents the spectrum radius of a matrix. Another merit of B-ALM is that it simplifies the solving difficult of the subproblems to a relatively easier proximal estimation, which may have a closed-form solution in many practical applications. However, it will needs an inner solver to tackle the dual subproblem or the Cholesky factorization to deal with an equivalent linear equation of the dual subproblem.

Motivated by these observations, we will develop and analyze a new double-penalty ALM with a relaxation step (abbreviated by **P-rALM**) for solving the problem (1). This method could reduce the difficulty of updating the dual variable while still maintaining the weak convergence condition and having a fast convergence behavior. For the sake of conciseness, we first present the framework of P-rALM as the following:

```
Initialize $(\mathbf{x}^0, \lambda^0)$, and choose $\gamma \in (0,2),\ r > 0,\ Q \succ \mathbf{0}$;
While stopping criteria is not satisfied do
  $\widetilde{\mathbf{x}}^k = \arg\min_{\mathbf{x}\in\mathcal{X}} \theta(\mathbf{x}) - \langle \lambda^k, A\mathbf{x} - b\rangle + \frac{r}{2}\left\|A(\mathbf{x} - \mathbf{x}^k)\right\|^2 + \frac{1}{2}\left\|\mathbf{x} - \mathbf{x}^k\right\|^2_Q$;
  $\widetilde{\lambda}^k = \mathcal{P}_\Lambda\Big(\lambda^k - r\left[A(2\widetilde{\mathbf{x}}^k - \mathbf{x}^k) - b\right]\Big)$;
  $\begin{pmatrix} \mathbf{x}^{k+1} \\ \lambda^{k+1} \end{pmatrix} = \begin{pmatrix} \mathbf{x}^k \\ \lambda^k \end{pmatrix} + \gamma \begin{pmatrix} \widetilde{\mathbf{x}}^k - \mathbf{x}^k \\ \widetilde{\lambda}^k - \lambda^k \end{pmatrix}$;
End while
```

In the above framework, $\mathcal{P}_{\Lambda}(\cdot)$ denotes the projection operator onto the set

$$\Lambda = \begin{cases} \mathcal{R}^m, & \text{if } A\mathbf{x} = b, \\ \mathcal{R}^m_+, & \text{if } A\mathbf{x} \geq b. \end{cases}$$

For the problem (1) subject to $A\mathbf{x} \geq b$, the dual update reduces to $\widetilde{\lambda}^k = \max\big(\lambda^k - r\big[A(2\widetilde{\mathbf{x}}^k - \mathbf{x}^k) - b\big], \mathbf{0}\big)$. Main features of this P-rALM are summarized as four aspects:

- Unlike the construction of the classical ALM, the $\mathbf{x}$-subproblem of P-rALM utilizes two different quadratic penalty terms $\frac{r}{2}\left\|A(\mathbf{x} - \mathbf{x}^k)\right\|^2$ and $\frac{1}{2}\left\|\mathbf{x} - \mathbf{x}^k\right\|_Q^2$. Equivalently,

$$\widetilde{\mathbf{x}}^k = \arg\min_{\mathbf{x}\in\mathcal{X}} \theta(\mathbf{x}) - \left\langle \mathbf{x}, A^{\mathsf{T}}\lambda^k + (Q + rA^{\mathsf{T}}A)\mathbf{x}^k \right\rangle + \frac{1}{2}\|\mathbf{x}\|^2_{Q+rA^{\mathsf{T}}A}, \tag{3}$$

  which is clearly different from (2) since the iterate $\widetilde{\mathbf{x}}^k$ doesn't depend on the data b but the previous iterate $\mathbf{x}^k$. In particular, taking $Q = \tau\mathbf{I} - rA^{\mathsf{T}}A$ with $\tau > r\|A^{\mathsf{T}}A\|$ could convert (3) to the following proximity operator

$$\mathbf{prox}_{\theta,\tau}(x) = \arg\min_{\mathbf{x}\in\mathcal{X}} \theta(\mathbf{x}) + \frac{\tau}{2}\left\|\mathbf{x} - \mathbf{x}^k - \frac{1}{\tau}A^{\mathsf{T}}\lambda^k\right\|^2,$$

  which has a unique global solution and further allows a closed-form solution when $\mathcal{X}$ is simple. For this case, by taking $\gamma = 1, \tau = \frac{r}{\alpha}$ for some $\alpha > 0$, our P-rALM reduces to [29, the scheme (38)], which indicates that the new algorithm is more general than some in the literature. If $\mathbf{prox}_{\theta,\tau}(x)$ is not available but θ is smooth, then user could exploit linearization technique or select an inner solver such as conjugate gradient method to solve the $\mathbf{x}$-subproblem inexactly, or use the formula provided by [32] for accurately approximating the proximal operator.

- The dual update $\widetilde{\lambda}^k$ is the same as that in [11] but is comparatively easier than that of B-ALM. It combines the information of both the current iterate $\widetilde{\mathbf{x}}^k$ and the extrapolation iterate $\widetilde{\mathbf{x}}^k - \mathbf{x}^k$. Moreover, after the primal-dual updates, a relaxation step is adopted to accelerate the convergence of the algorithm from theoretical and numerical interests.

- As said before, compared to some existing splitting algorithms, the global convergence of P-rALM will no longer depend on $\rho(A^{\mathsf{T}}A)$, although P-rALM with the choice $Q = \tau\mathbf{I} - rA^{\mathsf{T}}A$ reduces to [27, Algorothm 1] with $\eta = 0$ involved (in fact, the convergence of this algorithm depends on $\rho(A^{\mathsf{T}}A)$). However, as analyzed in Section 2.3, the double-penalty terms in P-rALM can be extended to the general Bregman distance while still ensure the convergence of P-rALM under a proper assumption.

- We show two elegant results as in Theorem 2.2 and Corollary 2.1, that is, the primal residual and the objective gap converge in a sublinear convergence rate. Motivated by the structure of H in (8), we also discuss a generalization of P-rALM and two new-types of Primal-Dual Hybrid Gradient algorithm (PDHG) for solving the multiple block separable convex optimization and the saddle-point problem, respectively. The connection between P-rALM and its variants is that all of them can be analyzed by variational characterization with similar structured matrices to the block matrix H in the sense of primal-dual and dual-primal frameworks, respectively. The linear convergence rate of P-rALM is indicated under similar assumptions as the analysis for the new PDHG in the appendix. Performance the proposed P-rALM and its two-block extensions are verified by testing two popular examples in machine learning and by comparing with several well-established algorithms in the literature.

The paper is organized as follows. In Section 2, we analyze the global convergence and sublinear convergence rate of P-rALM. A linearized P-rALM is also discussed when the objective function is smooth. Section 3 extends the proposed P-rALM to solve the multi-block separable convex programming and also shows a dual-primal version of the extended P-rALM. Section 4 investigates the performance of the proposed algorithm and its extensions. In the appendix, we further discuss the convergence complexity of two related PDHG algorithms based on the construction of P-rALM for solving a family of convex-concave saddle-point problems.

2 Convergence analysis of P-rALM

2.1 Variational characterization

Let us first recall the following fundamental lemma in e.g. [21] which will be used to characterize the saddle-point of (1) and the iterates of P-rALM.

Lemma 2.1 *Let* $\Omega \subseteq \mathcal{R}^n$ *be a closed convex set,* $f(x)$ *and* $h(x)$ *be convex functions. If* h *is differentiable on an open set which contains* Ω*, and the solution set of the minimization problem* $\min\{f(x)+h(x) \mid x\in\Omega\}$ *is nonempty. Then, we have*

$$x^* \in \arg\min\big\{f(x)+h(x) \mid x\in\Omega\big\}$$

if and only if

$$x^*\in\Omega,\quad f(x)-f(x^*)+\big\langle x-x^*, \nabla h(x^*)\big\rangle \ge 0,\ \forall x\in\Omega.$$

From the perspective of optimization, a point $\mathbf{w}^* = (\mathbf{x}^*;\lambda^*)\in\mathcal{M} := \mathcal{X}\times\Lambda$ is called the saddle-point of (1) if

$$L(\mathbf{x}^*,\lambda) \le L(\mathbf{x}^*,\lambda^*) \le L(\mathbf{x},\lambda^*),\quad \forall\lambda\in\Lambda, \mathbf{x}\in\mathcal{X},$$

which, by Lemma 2.1, is explicitly rewritten as

$$\begin{cases} \mathbf{x}^*\in\mathcal{X}, \quad \theta(\mathbf{x})-\theta(\mathbf{x}^*) & + \quad \langle \mathbf{x}-\mathbf{x}^*, -A^{\mathsf{T}}\lambda^*\rangle \ge 0, \quad \forall\mathbf{x}\in\mathcal{X}, \\ \lambda^*\in\Lambda, & \qquad\quad \langle \lambda-\lambda^*, A\mathbf{x}^*-b\rangle \ge 0, \quad \forall\lambda\in\Lambda. \end{cases}$$

These inequalities can be further expressed as the following more compact form

$$\mathrm{VI}(\theta,\mathcal{J},\mathcal{M}):\quad \theta(\mathbf{x})-\theta(\mathbf{x}^*)+\big\langle \mathbf{w}-\mathbf{w}^*, \mathcal{J}(\mathbf{w}^*)\big\rangle \ge 0,\ \forall\mathbf{w}\in\mathcal{M}, \tag{4}$$

where

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

Similar characterization can be found in e.g. [3, 19]. An equivalent expression of (4) is

$$\theta(\mathbf{x})-\theta(\mathbf{x}^*)+\big\langle \mathbf{w}-\mathbf{w}^*, \mathcal{J}(\mathbf{w})\big\rangle \ge 0,\ \forall\mathbf{w}\in\mathcal{M}, \tag{5}$$

since it holds by the monotonicity of $\mathcal{J}(\mathbf{w})$ that

$$\big\langle \mathbf{w} - \bar{\mathbf{w}}, \mathcal{J}(\mathbf{w}) - \mathcal{J}(\bar{\mathbf{w}})\big\rangle = 0,\ \forall \mathbf{w}, \bar{\mathbf{w}} \in \mathcal{M}. \tag{6}$$

Notice that the solution set of (4) is nonempty by the previous assumption that the solution set of (1) is nonempty, and it can be characterized as (see [17, Theorem 2.1])

$$\mathcal{M}^* = \bigcap_{\mathbf{w}\in\mathcal{M}} \Big\{ \hat{\mathbf{w}} \mid \theta(\mathbf{x}) - \theta(\hat{\mathbf{x}}) + \big\langle \mathbf{w} - \hat{\mathbf{w}}, \mathcal{J}(\mathbf{w})\big\rangle \geq 0 \Big\}.$$

Obviously, $\mathbf{w}^*$ satisfies (4) if and only if it is a primal-dual solution of (1). Next, we characterize the sequence generated by P-rALM as a mixed variational inequality with the aid of the auxiliary notation $\widetilde{\mathbf{w}}^k = (\widetilde{\mathbf{x}}^k; \widetilde{\lambda}^k)$.

Lemma 2.2 *The sequence $\{\mathbf{w}^k\}$ generated by P-rALM satisfies*

$$\widetilde{\mathbf{w}}^k \in \mathcal{M},\ \theta(\mathbf{x}) - \theta(\widetilde{\mathbf{x}}^k) + \big\langle \mathbf{w} - \widetilde{\mathbf{w}}^k, \mathcal{J}(\mathbf{w})\big\rangle \geq \big\langle \mathbf{w} - \widetilde{\mathbf{w}}^k, H(\mathbf{w}^k - \widetilde{\mathbf{w}}^k)\big\rangle \tag{7}$$

for any $\mathbf{w} \in \mathcal{M}$, where

$$\widetilde{\mathbf{w}}^k = \begin{pmatrix} \widetilde{\mathbf{x}}^k \\ \widetilde{\lambda}^k \end{pmatrix} \quad \textit{and} \quad H = \begin{bmatrix} rA^\mathsf{T}A + Q & A^\mathsf{T} \\ A & \frac{1}{r}\mathbf{I} \end{bmatrix} \tag{8}$$

is symmetric positive definite for any $r > 0$ and $Q \succ \mathbf{0}$.

Proof. By Lemma 2.1 the first-order optimality condition of the $\mathbf{x}$-subproblem in P-rALM is

$$\widetilde{\mathbf{x}}^k \in \mathcal{X},\ \theta(\mathbf{x}) - \theta(\widetilde{\mathbf{x}}^k) + \big\langle \mathbf{x} - \widetilde{\mathbf{x}}^k, -A^\mathsf{T}\lambda^k + (rA^\mathsf{T}A + Q)(\widetilde{\mathbf{x}}^k - \mathbf{x}^k)\big\rangle \geq 0,\ \forall \mathbf{x} \in \mathcal{X},$$

equivalently,

$$\begin{aligned} &\theta(\mathbf{x}) - \theta(\widetilde{\mathbf{x}}^k) + \big\langle \mathbf{x} - \widetilde{\mathbf{x}}^k, -A^\mathsf{T}\widetilde{\lambda}^k\big\rangle \\ \geq &\Big\langle \mathbf{x} - \widetilde{\mathbf{x}}^k, (rA^\mathsf{T}A + Q)(\mathbf{x}^k - \widetilde{\mathbf{x}}^k) + A^\mathsf{T}(\lambda^k - \widetilde{\lambda}^k)\Big\rangle,\ \forall \mathbf{x} \in \mathcal{X}. \end{aligned} \tag{9}$$

Besides, it follows from the dual update that $\widetilde{\lambda}^k \in \Lambda$ and

$$\Big\langle \lambda - \widetilde{\lambda}^k, A\widetilde{\mathbf{x}}^k - b\Big\rangle \geq \Big\langle \lambda - \widetilde{\lambda}^k, A(\mathbf{x}^k - \widetilde{\mathbf{x}}^k) + \frac{1}{r}(\lambda^k - \widetilde{\lambda}^k)\Big\rangle,\ \forall \lambda \in \Lambda. \tag{10}$$

Combine the above inequalities (9)-(10) together with the structure of H given in (8) and the property in (6) to ensure the result in (7).

Observing that the symmetric matrix H has the following decomposition:

$$H = \begin{bmatrix} rA^\mathsf{T}A & A^\mathsf{T} \\ A & \frac{1}{r}\mathbf{I} \end{bmatrix} + \begin{bmatrix} Q & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \begin{pmatrix} \sqrt{r}A^\mathsf{T} \\ \frac{1}{\sqrt{r}}\mathbf{I} \end{pmatrix} \Big(\sqrt{r}A, \frac{1}{\sqrt{r}}\mathbf{I}\Big) + \begin{bmatrix} Q & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

For any $\mathbf{w} = (\mathbf{x}; \lambda) \neq \mathbf{0}$ we have

$$\mathbf{w}^\mathsf{T}H\mathbf{w} = \Big\|\sqrt{r}A\mathbf{x} + \frac{1}{\sqrt{r}}\lambda\Big\|^2 + \|\mathbf{x}\|_Q^2 > 0$$

and therefore H is a positive definite matrix. ■

2.2 Convergence and convergence rate

The following theorem shows that the sequence $\{\mathbf{w}^k\}$ generated by P-rALM is contractive under the H-weighted norm and thus converges to the solution point of VI$(\theta, \mathcal{J}, \mathcal{M})$.

Theorem 2.1 *For any $\gamma \in (0,2)$, the sequence $\{\mathbf{w}^k\}$ generated by P-rALM satisfies*

$$\big\|\mathbf{w}^{k+1} - \mathbf{w}^*\big\|_H^2 \le \big\|\mathbf{w}^k - \mathbf{w}^*\big\|_H^2 - \frac{2-\gamma}{\gamma}\big\|\mathbf{w}^k - \mathbf{w}^{k+1}\big\|_H^2, \ \forall \mathbf{w}^* \in \mathcal{M}^*. \tag{11}$$

Moreover, there exists a point $\mathbf{w}^\infty \in \mathcal{M}^*$ such that $\lim\limits_{k\to\infty} \mathbf{w}^k = \mathbf{w}^\infty$.

Proof. Setting $\mathbf{w} = \mathbf{w}^*$ in (7) together with (4) is to achieve

$$\big\langle \widetilde{\mathbf{w}}^k - \mathbf{w}^*, H(\mathbf{w}^k - \widetilde{\mathbf{w}}^k)\big\rangle \ge \theta(\widetilde{\mathbf{x}}^k) - \theta(\mathbf{x}^*) + \big\langle \widetilde{\mathbf{w}}^k - \mathbf{w}^*, \mathcal{J}(\mathbf{w}^*)\big\rangle \ge 0.$$

Combining the above property with the update

$$\mathbf{w}^{k+1} = \mathbf{w}^k + \gamma(\widetilde{\mathbf{w}}^k - \mathbf{w}^k), \tag{12}$$

we have

$$\begin{aligned}
& \big\|\mathbf{w}^k - \mathbf{w}^*\big\|_H^2 - \big\|\mathbf{w}^{k+1} - \mathbf{w}^*\big\|_H^2 \\
= \ & \big\|\mathbf{w}^k - \mathbf{w}^*\big\|_H^2 - \big\|\mathbf{w}^k - \mathbf{w}^* + \mathbf{w}^{k+1} - \mathbf{w}^k\big\|_H^2 \\
= \ & 2\gamma\big\langle \mathbf{w}^k - \mathbf{w}^*, H(\mathbf{w}^k - \widetilde{\mathbf{w}}^k)\big\rangle - \gamma^2\big\|\widetilde{\mathbf{w}}^k - \mathbf{w}^k\big\|_H^2 \\
= \ & 2\gamma\big\langle \mathbf{w}^k - \widetilde{\mathbf{w}}^k + \widetilde{\mathbf{w}}^k - \mathbf{w}^*, H(\mathbf{w}^k - \widetilde{\mathbf{w}}^k)\big\rangle - \gamma^2\big\|\mathbf{w}^k - \widetilde{\mathbf{w}}^k\big\|_H^2 \\
= \ & \gamma(2-\gamma)\big\|\mathbf{w}^k - \widetilde{\mathbf{w}}^k\big\|_H^2 + 2\gamma\big\langle \widetilde{\mathbf{w}}^k - \mathbf{w}^*, H(\mathbf{w}^k - \widetilde{\mathbf{w}}^k)\big\rangle \\
\ge \ & \gamma(2-\gamma)\big\|\mathbf{w}^k - \widetilde{\mathbf{w}}^k\big\|_H^2 = \frac{2-\gamma}{\gamma}\big\|\mathbf{w}^k - \mathbf{w}^{k+1}\big\|_H^2.
\end{aligned}$$

Then, rearrange this inequality to obtain the inequality (11).

Note that the inequality (11) implies that the sequence $\{\mathbf{w}^k\}$ is bounded and $\lim\limits_{k\to\infty} \big\|\mathbf{w}^k - \mathbf{w}^{k+1}\big\|_H^2 = 0$. Namely, $\lim\limits_{k\to\infty} (\mathbf{w}^k - \mathbf{w}^{k+1}) = \mathbf{0}$, which by the relation (12) also shows

$$\lim_{k\to\infty} (\widetilde{\mathbf{w}}^k - \mathbf{w}^k) = \mathbf{0}. \tag{13}$$

Let $\mathbf{w}^\infty$ be any accumulation point of $\{\widetilde{\mathbf{w}}^k\}$. Then, (taking a subsequence of $\{\widetilde{\mathbf{w}}^k\}$ if necessary) it follows from (7) and (13) that

$$\theta(\mathbf{x}) - \theta(\mathbf{x}^\infty) + \langle \mathbf{w} - \mathbf{w}^\infty, \mathcal{J}(\mathbf{w}^\infty)\rangle \ge 0, \ \forall \mathbf{w} \in \mathcal{M}.$$

This indicates $\mathbf{w}^\infty \in \mathcal{M}^*$ compared to (4). So, by (11) again we have

$$\big\|\mathbf{w}^k - \mathbf{w}^\infty\big\|_H^2 \le \big\|\mathbf{w}^j - \mathbf{w}^\infty\big\|_H^2 \qquad \text{for all } k \ge j.$$

Finally, it follows from $\mathbf{w}^\infty$ being an accumulation point that $\lim\limits_{k\to\infty} \mathbf{w}^k = \mathbf{w}^\infty$. ■

Before establishing the sublinear convergence rate of P-rALM for the following average iterates (it was firstly used in [6] to accelerate the convergence of a stochastic method)

$$\mathbf{w}_T := \frac{1}{T+1}\sum_{k=\kappa}^{T+\kappa}\widetilde{\mathbf{w}}^k \quad \text{and} \quad \mathbf{x}_T := \frac{1}{T+1}\sum_{k=\kappa}^{T+\kappa}\widetilde{\mathbf{x}}^k, \quad \forall\kappa \geq 0, T > 0, \tag{14}$$

we would analyze the convergence complexity of the pointwise iterates (see also [7, Theorem 6]) and the primal residual, where the notation $\partial\theta(\mathbf{x})$ represents its sub-differential at $\mathbf{x}$, and $\mathcal{N}_{\mathcal{X}}(\mathbf{x})$ denotes the normal cone of $\mathcal{X}$ at $\mathbf{x}$.

Theorem 2.2 *For any $k > 0$, there exists an integer $t \leq k$ such that*

$$\left\|\mathbf{w}^{t+1} - \mathbf{w}^t\right\|_H^2 \leq \frac{\varrho}{k+1} \qquad \textit{and} \qquad \left\|\mathbf{s}^t\right\|_H^2 \leq \frac{\varrho}{k+1}\frac{\left\|\operatorname{diag}(rA^{\mathsf{T}}A + Q, A^{\mathsf{T}})\right\|_H^2}{\gamma^2}, \tag{15}$$

where $\mathbf{s}^t \in \mathcal{R}^n$ satisfies $A^{\mathsf{T}}\widetilde{\lambda}^t - \mathbf{s}^t \in \partial\theta(\widetilde{\mathbf{x}}^t) + \mathcal{N}_{\mathcal{X}}(\widetilde{\mathbf{x}}^t)$ and $\varrho = \frac{\gamma}{2-\gamma}\left\|\mathbf{w}^* - \mathbf{w}^0\right\|_H^2$.

Proof. Let $k > 0$ be a fixed constant and $t \leq k$ be a positive integer such that

$$\left\|\mathbf{w}^{t+1} - \mathbf{w}^t\right\|_H^2 = \min\left\{\left\|\mathbf{w}^{l+1} - \mathbf{w}^l\right\|_H^2 \mid l = 0, \ldots, k\right\}.$$

Summing up (11) over $k = 0, \cdots, \infty$ immediately gives

$$\sum_{k=0}^{\infty}\left\|\mathbf{w}^k - \mathbf{w}^{k+1}\right\|_H^2 \leq \frac{\gamma}{2-\gamma}\left\|\mathbf{w}^* - \mathbf{w}^0\right\|_H^2 < \infty,$$

which further shows

$$\left\|\mathbf{w}^{t+1} - \mathbf{w}^t\right\|_H^2 \leq \frac{\varrho}{k+1}. \tag{16}$$

Now, it follows from (9) that by defining

$$\mathbf{s}^t = (rA^{\mathsf{T}}A + Q)(\widetilde{\mathbf{x}}^t - \mathbf{x}^t) + A^{\mathsf{T}}(\widetilde{\lambda}^t - \lambda^t) \in \mathcal{R}^n, \tag{17}$$

we have

$$\theta(\mathbf{x}) - \theta(\widetilde{\mathbf{x}}^t) + (\widetilde{\mathbf{x}}^t - \mathbf{x})^{\mathsf{T}}(A^{\mathsf{T}}\widetilde{\lambda}^t - \mathbf{s}^t) \geq 0, \quad \forall\mathbf{x} \in \mathcal{X},$$

which implies $A^{\mathsf{T}}\widetilde{\lambda}^t - \mathbf{s}^t \in \partial\theta(\widetilde{\mathbf{x}}^t) + \mathcal{N}_{\mathcal{X}}(\widetilde{\mathbf{x}}^t)$. By (17) and (12) again, we have

$$\begin{aligned}\left\|\mathbf{s}^t\right\|_H &= \left\|\operatorname{diag}(rA^{\mathsf{T}}A + Q, A^{\mathsf{T}})(\widetilde{\mathbf{w}}^t - \mathbf{w}^t)\right\|_H = \left\|\frac{1}{\gamma}\operatorname{diag}(rA^{\mathsf{T}}A + Q, A^{\mathsf{T}})(\mathbf{w}^{t+1} - \mathbf{w}^t)\right\|_H \\ &\leq \frac{1}{\gamma}\left\|\operatorname{diag}(rA^{\mathsf{T}}A + Q, A^{\mathsf{T}})\right\|_H\left\|\mathbf{w}^{t+1} - \mathbf{w}^t\right\|_H,\end{aligned}$$

which together with (16) ensures the right inequality in (15). ■

The forthcoming remark suggests that our analysis of Theorem 2.2 is much easier than the analysis in e.g. [19], although the iteration-complexity results are consistent. And the second remark shows that the pointwise iteration complexity can be still ensured by regarding the relaxation factor γ as some special sequences.

Remark 2.1 *Analogous to the technique [19] to establish the sublinear convergence rate in a pointwise sense, by adding the inequality (7) with* $\mathbf{w} := \widetilde{\mathbf{w}}^{k+1}$*, i.e.,*

$$\theta(\widetilde{\mathbf{x}}^{k+1}) - \theta(\widetilde{\mathbf{x}}^{k}) + \big\langle \widetilde{\mathbf{w}}^{k+1} - \widetilde{\mathbf{w}}^{k}, \mathcal{J}(\widetilde{\mathbf{w}}^{k+1})\big\rangle \geq \big\langle \widetilde{\mathbf{w}}^{k+1} - \widetilde{\mathbf{w}}^{k}, H(\mathbf{w}^{k} - \widetilde{\mathbf{w}}^{k})\big\rangle$$

to the inequality (7) at $(k+1)$*-th iteration with* $\mathbf{w} := \widetilde{\mathbf{w}}^{k}$*, i.e.,*

$$\theta(\widetilde{\mathbf{x}}^{k}) - \theta(\widetilde{\mathbf{x}}^{k+1}) + \big\langle \widetilde{\mathbf{w}}^{k} - \widetilde{\mathbf{w}}^{k+1}, \mathcal{J}(\widetilde{\mathbf{w}}^{k})\big\rangle \geq \big\langle \widetilde{\mathbf{w}}^{k} - \widetilde{\mathbf{w}}^{k+1}, H(\mathbf{w}^{k+1} - \widetilde{\mathbf{w}}^{k+1})\big\rangle$$

together with the property (6), we have

$$\big\langle \widetilde{\mathbf{w}}^{k} - \widetilde{\mathbf{w}}^{k+1}, H(\mathbf{w}^{k} - \widetilde{\mathbf{w}}^{k} + \widetilde{\mathbf{w}}^{k+1} - \mathbf{w}^{k+1})\big\rangle \geq 0.$$

Then, adding the term $\big\|\mathbf{w}^{k} - \widetilde{\mathbf{w}}^{k} + \widetilde{\mathbf{w}}^{k+1} - \mathbf{w}^{k+1}\big\|_H^2$ *to both sides of the above inequality and using* $\mathbf{w}^{\mathsf{T}} H \mathbf{w} = \frac{1}{2}\big\|\mathbf{w}\big\|_{H+H^{\mathsf{T}}}^2$ *and (12) will lead to*

$$\begin{aligned}\frac{1}{2}\big\|\mathbf{w}^{k} - \widetilde{\mathbf{w}}^{k} + \widetilde{\mathbf{w}}^{k+1} - \mathbf{w}^{k+1}\big\|_{H+H^{\mathsf{T}}}^2 &\leq \big\langle \mathbf{w}^{k} - \mathbf{w}^{k+1}, H(\mathbf{w}^{k} - \widetilde{\mathbf{w}}^{k} + \widetilde{\mathbf{w}}^{k+1} - \mathbf{w}^{k+1})\big\rangle \\ &= \gamma\big\langle \mathbf{w}^{k} - \widetilde{\mathbf{w}}^{k}, H(\mathbf{w}^{k} - \widetilde{\mathbf{w}}^{k} + \widetilde{\mathbf{w}}^{k+1} - \mathbf{w}^{k+1})\big\rangle. \quad (18)\end{aligned}$$

So, we have from (18) and the identity $\|a\|_H^2 - \|b\|_H^2 = 2\langle a, H(a-b)\rangle - \|a-b\|_H^2$ *that*

$$\begin{aligned}&\big\|\mathbf{w}^{k} - \widetilde{\mathbf{w}}^{k}\big\|_H^2 - \big\|\mathbf{w}^{k+1} - \widetilde{\mathbf{w}}^{k+1}\big\|_H^2 \\ =\;& 2\big\langle \mathbf{w}^{k} - \widetilde{\mathbf{w}}^{k}, H(\mathbf{w}^{k} - \widetilde{\mathbf{w}}^{k} + \widetilde{\mathbf{w}}^{k+1} - \mathbf{w}^{k+1})\big\rangle - \big\|\mathbf{w}^{k} - \widetilde{\mathbf{w}}^{k} + \widetilde{\mathbf{w}}^{k+1} - \mathbf{w}^{k+1}\big\|_H^2 \\ \geq\;& \big\|\mathbf{w}^{k} - \widetilde{\mathbf{w}}^{k} + \widetilde{\mathbf{w}}^{k+1} - \mathbf{w}^{k+1}\big\|_{\widetilde{H}}^2,\end{aligned}$$

where $\widetilde{H} = \frac{1}{\gamma}(H + H^{\mathsf{T}}) - H = \frac{2-\gamma}{\gamma} H \succ \mathbf{0}$*. By (12) again, the above inequality further shows*

$$\big\|\mathbf{w}^{k} - \mathbf{w}^{k+1}\big\|_H^2 \geq \big\|\mathbf{w}^{k+1} - \mathbf{w}^{k+2}\big\|_H^2, \quad \forall k \geq 0,$$

which using (11) gives

$$\frac{\gamma}{2-\gamma}\big\|\mathbf{w}^{\kappa} - \mathbf{w}^{*}\big\|_H^2 \geq \sum_{k=\kappa}^{T+\kappa} \big\|\mathbf{w}^{k} - \mathbf{w}^{k+1}\big\|_H^2 \geq (T+1)\big\|\mathbf{w}^{T+\kappa} - \mathbf{w}^{T+\kappa+1}\big\|_H^2.$$

That is, $\big\|\mathbf{w}^{T+\kappa} - \mathbf{w}^{T+\kappa+1}\big\|_H^2 \leq \frac{1}{T+1}\frac{\gamma}{2-\gamma}\big\|\mathbf{w}^{\kappa} - \mathbf{w}^{*}\big\|_H^2$ *and it is consistent with the left result of (15). Especially, taking* $\kappa = 0$ *is to obtain* $\big\|\mathbf{w}^{T} - \mathbf{w}^{T+1}\big\|_H^2 \leq \frac{\varrho}{T+1}$*.*

Remark 2.2 *Because the relaxation parameter* γ *appears in the iteration complexity results of (15), one may select a proper fixed value belonging to* $(0,2)$ *to do experiments or update it adaptively by the following strategies for any* $k > 0$*:*

(S1) By updating $\gamma^k = \frac{2k}{2k+c}$ *where* $c > 0$ *is any real number, we will have* $\frac{\gamma^k}{2-\gamma^k} = \frac{k}{k+c} < 1$*; or*

(S2) By updating $\gamma^k = \frac{2k(k+2)}{2k^2+6k+3}$*, we will obtain* $\frac{\gamma^k}{2-\gamma^k} = \frac{k(k+2)}{(k+1)(k+3)} < 1$*;*

Then, it follows from (11) that $\big\|\mathbf{w}^{k} - \mathbf{w}^{k+1}\big\|_H^2 \leq \big\|\mathbf{w}^{k} - \mathbf{w}^{*}\big\|_H^2 - \big\|\mathbf{w}^{k+1} - \mathbf{w}^{*}\big\|_H^2$ *and then* $\big\|\mathbf{w}^{t-1} - \mathbf{w}^{t}\big\|^2 \leq \frac{1}{k}\big\|\mathbf{w}^{*} - \mathbf{w}^{0}\big\|_H^2$*. Meanwhile, the convergence rates of* $\big\|\mathbf{s}^t\big\|^2$ *and the residual* $\theta(\mathbf{x}_t) - \theta(\mathbf{x}^*) + \eta\big\|A\mathbf{x}_t - b\big\|$ *are still* $\mathcal{O}(1/k)$*, where* $\eta \geq \|\lambda\|$ *is a constant as shown in (23).*

(S3) *Finally, at each step we may choose a random number from 0 to 2. In this case, P-rALM still converges with a sublinear convergence rate.*

Theorem 2.3 *Let $T > 0, \kappa \geq 0$. Then, for any $r > 0, Q \succ \mathbf{0}$, the sequence $\{\mathbf{w}^k\}$ generated by P-rALM satisfies*

$$\theta(\mathbf{x}_T) - \theta(\mathbf{x}) + \big\langle \mathbf{w}_T - \mathbf{w}, \mathcal{J}(\mathbf{w}) \big\rangle \leq \frac{1}{2\gamma(T+1)} \big\|\mathbf{w}^\kappa - \mathbf{w}\big\|_H^2, \ \forall \mathbf{w} \in \mathcal{M}. \tag{19}$$

Proof. The inequality (7) and the relation (12) indicate

$$\begin{aligned} \gamma \Big[\theta(\mathbf{x}) - \theta(\widetilde{\mathbf{x}}^k) + \big\langle \mathbf{w} - \widetilde{\mathbf{w}}^k, \mathcal{J}(\mathbf{w}) \big\rangle\Big] &\geq \big\langle \widetilde{\mathbf{w}}^k - \mathbf{w}, H(\mathbf{w}^{k+1} - \mathbf{w}^k) \big\rangle \\ = \ \frac{1}{2}\Big\{ \big\|\widetilde{\mathbf{w}}^k - \mathbf{w}^k\big\|_H^2 - \big\|\mathbf{w}^{k+1} - \widetilde{\mathbf{w}}^k\big\|_H^2 &+ \big\|\mathbf{w}^{k+1} - \mathbf{w}\big\|_H^2 - \big\|\mathbf{w}^k - \mathbf{w}\big\|_H^2 \Big\}, \end{aligned} \tag{20}$$

in which the equality uses the identity

$$\big\langle \mathbf{p} - \mathbf{q}, H(\mathbf{u} - \mathbf{v}) \big\rangle = \frac{1}{2}\{\big\|\mathbf{p} - \mathbf{v}\big\|_H^2 - \big\|\mathbf{p} - \mathbf{u}\big\|_H^2 + \big\|\mathbf{q} - \mathbf{u}\big\|_H^2 - \big\|\mathbf{q} - \mathbf{v}\big\|_H^2\} \tag{21}$$

with specifications $\mathbf{p} := \widetilde{\mathbf{w}}^k, \mathbf{q} = \mathbf{w}, \mathbf{u} = \mathbf{w}^{k+1}$ and $\mathbf{v} := \mathbf{w}^k$. Note that

$$\begin{aligned} \big\|\widetilde{\mathbf{w}}^k - \mathbf{w}^k\big\|_H^2 - \big\|\mathbf{w}^{k+1} - \widetilde{\mathbf{w}}^k\big\|_H^2 &= \big\|\widetilde{\mathbf{w}}^k - \mathbf{w}^k\big\|_H^2 - \big\|\mathbf{w}^k - \widetilde{\mathbf{w}}^k + \mathbf{w}^{k+1} - \mathbf{w}^k\big\|_H^2 \\ &= \big\|\widetilde{\mathbf{w}}^k - \mathbf{w}^k\big\|_H^2 - \big\|\mathbf{w}^k - \widetilde{\mathbf{w}}^k + \gamma(\widetilde{\mathbf{w}}^k - \mathbf{w}^k)\big\|_H^2 \\ &= \gamma(2-\gamma)\big\|\widetilde{\mathbf{w}}^k - \mathbf{w}^k\big\|_H^2 \geq 0. \end{aligned}$$

Summing the inequality (20) over $k = \kappa, \kappa+1, \ldots, \kappa+T$, we have

$$(T+1)\theta(\mathbf{x}) - \sum_{k=\kappa}^{\kappa+T} \theta(\widetilde{\mathbf{x}}^k) + \Big\langle (T+1)\mathbf{w} - \sum_{k=\kappa}^{\kappa+T} \widetilde{\mathbf{w}}^k, \mathcal{J}(\mathbf{w}) \Big\rangle + \frac{1}{2\gamma}\big\|\mathbf{w} - \mathbf{w}^\kappa\big\|_H^2 \geq 0,$$

which, by the definition of $\mathbf{w}_T$ and $\mathbf{x}_T$, gives

$$\frac{1}{T+1}\sum_{k=\kappa}^{\kappa+T} \theta(\widetilde{\mathbf{x}}^k) - \theta(\mathbf{x}) + \big\langle \mathbf{w}_T - \mathbf{w}, \mathcal{J}(\mathbf{w}) \big\rangle \leq \frac{1}{2\gamma(T+1)}\big\|\mathbf{w} - \mathbf{w}^\kappa\big\|_H^2. \tag{22}$$

Because θ is convex function and has the property $\theta(\mathbf{x}_T) \leq \frac{1}{T+1}\sum_{k=\kappa}^{\kappa+T} \theta(\widetilde{\mathbf{x}}^k)$, the inequality (19) is obtained by plugging this property into (22). ■

Theorem 2.3 illustrates that the proposed P-rALM converges in a sublinear ergodic convergence rate. Furthermore, for any $\eta > 0$, by letting $\Gamma_\eta = \{\lambda \mid \|\lambda\| \leq \eta\}$ and

$$\gamma_\eta = \inf_{\mathbf{x}^* \in \mathcal{X}} \sup_{\lambda \in \Lambda} \big\|\mathbf{w} - \mathbf{w}^\kappa\big\|_H^2, \tag{23}$$

we can get the following tight result whose proof is similar to that of [4, 36] and thus is omitted here. We can see from Corollary 2.1 that the residual $\theta(\mathbf{x}_T) - \theta(\mathbf{x}^*) + \eta\big\|A\mathbf{x}_T - b\big\|$ converges still in the worst $\mathcal{O}(1/T)$ convergence rate.

Corollary 2.1 *For any $\eta > 0$, let γ_η be defined in (23) and $\mathbf{x}_T$ be defined in (14). Then, the sequence $\{\mathbf{w}^k\}$ generated by P-rALM satisfies*

$$\theta(\mathbf{x}_T) - \theta(\mathbf{x}^*) + \eta\big\|A\mathbf{x}_T - b\big\| \leq \frac{\gamma_\eta}{2\gamma(T+1)}, \ \forall \mathbf{x}^* \in \mathcal{X}.$$

2.3 Two special cases

In this section, we would like to discuss two interesting cases on exploiting the Bregman distance and developing a linearized version of P-rALM, respectively.

Case 1: The double-penalty terms in P-rALM can be extended to the Bregman distance

$$D_\varphi(\mathbf{x},\mathbf{x}^k)=\varphi(\mathbf{x})-\varphi(\mathbf{x}^k)-\big\langle\nabla\varphi(\mathbf{x}^k),\mathbf{x}-\mathbf{x}^k\big\rangle,\quad \forall\mathbf{x}\in\mathcal{X},$$

where φ is a strictly convex and continuously differentiable function. For this case, by combining the optimality condition (see e.g. [40, lemma 2]) of the corresponding $\mathbf{x}$-subproblem and the previous inequality (10), we can deduce

$$\begin{aligned}&\big\langle\widetilde{\mathbf{w}}^k-\mathbf{w}^*,H(\mathbf{w}^k-\widetilde{\mathbf{w}}^k)\big\rangle-\big\langle\widetilde{\mathbf{x}}^k-\mathbf{x}^*,(rA^\mathsf{T}A+Q)(\mathbf{x}^k-\widetilde{\mathbf{x}}^k)\big\rangle\\ \geq\ & D_\varphi(\widetilde{\mathbf{x}}^k,\mathbf{x}^k)+D_\varphi(\mathbf{x}^*,\widetilde{\mathbf{x}}^k)-D_\varphi(\mathbf{x}^*,\mathbf{x}^k),\end{aligned}$$

that is

$$\begin{aligned}&\big\|\mathbf{w}^*-\mathbf{w}^k\big\|_H^2+2D_\varphi(\mathbf{x}^*,\mathbf{x}^k)-\big\|\mathbf{x}^*-\mathbf{x}^k\big\|_{rA^\mathsf{T}A+Q}^2\\ \geq\ &\big\|\mathbf{w}^*-\widetilde{\mathbf{w}}^k\big\|_H^2+2D_\varphi(\mathbf{x}^*,\widetilde{\mathbf{x}}^k)-\big\|\mathbf{x}^*-\widetilde{\mathbf{x}}^k\big\|_{rA^\mathsf{T}A+Q}^2\\ &+\big\|\widetilde{\mathbf{w}}^k-\mathbf{w}^k\big\|_H^2+2D_\varphi(\widetilde{\mathbf{x}}^k,\mathbf{x}^k)-\big\|\widetilde{\mathbf{x}}^k-\mathbf{x}^k\big\|_{rA^\mathsf{T}A+Q}^2.\end{aligned}\tag{24}$$

Denote $\tilde{H}=H-\mathrm{diag}(Q,\mathbf{0})$. Under the assumption that

$$D_\varphi(\mathbf{x},\bar{\mathbf{x}})\geq\frac{1}{2}\big\{\big\|\mathbf{x}-\bar{\mathbf{x}}\big\|_{rA^\mathsf{T}A}^2-\big\|\mathbf{w}-\bar{\mathbf{w}}\big\|_{\tilde{H}}^2\big\}\tag{25}$$

for any $\mathbf{w},\bar{\mathbf{w}}\in\mathcal{M}$, the corresponding modified P-rALM is convergent. When $\mathbf{w}=\bar{\mathbf{w}}$, (25) holds obviously; when $\mathbf{w}$ and $\bar{\mathbf{w}}$ are different points, we have $\big\|\mathbf{w}-\bar{\mathbf{w}}\big\|_{\tilde{H}}^2\geq 0$ and hence there exists a constant $c>0$ such that $\big\|\mathbf{w}-\bar{\mathbf{w}}\big\|_{\tilde{H}}^2\leq\frac{c}{2}\big\|\mathbf{x}-\bar{\mathbf{x}}\big\|_{rA^\mathsf{T}A}^2$. As a result, (25) reduces to the assumption on the kernel function φ:

$$\varphi(\mathbf{x})-\varphi(\bar{\mathbf{x}})-\big\langle\nabla\varphi(\bar{\mathbf{x}}),\mathbf{x}-\bar{\mathbf{x}}\big\rangle\geq\frac{1-c}{2}\big\|\mathbf{x}-\bar{\mathbf{x}}\big\|_{rA^\mathsf{T}A}^2.$$

If we take $\varphi=\frac{1}{2}\big\|\mathbf{x}\big\|_{rA^\mathsf{T}A}^2$, then $D_\varphi(\mathbf{x},\bar{\mathbf{x}})=\frac{1}{2}\big\|\mathbf{x}-\bar{\mathbf{x}}\big\|_{rA^\mathsf{T}A}^2$, showing that the assumption holds clearly, and finally the inequality (24) with simple algebra will reduce to (11).

Case 2: If the objective function $\theta(\mathbf{x})$ is smooth and its gradient is Lipschitz continuous with constant L_θ, which implies

$$\theta(y)\leq\theta(z)+\langle\nabla\theta(z),y-z\rangle+\frac{L_\theta}{2}\|y-z\|^2\tag{26}$$

for every $y,z\in\mathcal{X}$, then one may update the $\widetilde{\mathbf{x}}^k$-subproblem as the following

$$\widetilde{\mathbf{x}}^k=\arg\min_{\mathbf{x}\in\mathcal{X}}\langle\nabla\theta(\mathbf{x}^k)-A^\mathsf{T}\lambda^k,\mathbf{x}\rangle+\frac{r}{2}\big\|A(\mathbf{x}-\mathbf{x}^k)\big\|^2+\frac{1}{2}\big\|\mathbf{x}-\mathbf{x}^k\big\|_Q^2,\tag{27}$$

which, by taking $Q=\tau I-rA^\mathsf{T}A$ with $\tau>r\|A^\mathsf{T}A\|$, will become $\widetilde{\mathbf{x}}^k=P_\mathcal{X}\left(\mathbf{x}^k-[\nabla\theta(\mathbf{x}^k)-A^\mathsf{T}\lambda^k]/\tau\right)$, where $P_\mathcal{X}(\mathbf{x})$ denotes the projection of $\mathbf{x}\in\mathcal{R}^n$ onto $\mathcal{X}$. With the new update (27), a similar inequality to (11) can be also obtained. In fact, the first-order optimality condition of (27) are $\widetilde{\mathbf{x}}^k \in \mathcal{X}$ and

$$\left\langle \mathbf{x} - \widetilde{\mathbf{x}}^k, \nabla\theta(\mathbf{x}^k) - A^\mathsf{T}\lambda^k + (rA^\mathsf{T}A + Q)(\widetilde{\mathbf{x}}^k - \mathbf{x}^k) \right\rangle \geq 0, \ \forall \mathbf{x} \in \mathcal{X},$$

which, by using the convexity of θ and the inequality (26) with $(y, z) := (\widetilde{\mathbf{x}}^k, \mathbf{x}^k)$, shows

$$\begin{aligned} & \left\langle \mathbf{x} - \widetilde{\mathbf{x}}^k, A^\mathsf{T}\lambda^k + (rA^\mathsf{T}A + Q)(\mathbf{x}^k - \widetilde{\mathbf{x}}^k) \right\rangle \\ \leq \ & \left\langle \mathbf{x} - \widetilde{\mathbf{x}}^k, \nabla\theta(\mathbf{x}^k) \right\rangle = \left\langle \mathbf{x} - \mathbf{x}^k + \mathbf{x}^k - \widetilde{\mathbf{x}}^k, \nabla\theta(\mathbf{x}^k) \right\rangle \\ \leq \ & \theta(\mathbf{x}) - \theta(\mathbf{x}^k) + \theta(\mathbf{x}^k) - \theta(\widetilde{\mathbf{x}}^k) + \frac{L_\theta}{2}\left\|\mathbf{x}^k - \widetilde{\mathbf{x}}^k\right\|^2. \end{aligned}$$

Rearrange it to get

$$\theta(\mathbf{x}) - \theta(\widetilde{\mathbf{x}}^k) + \langle \mathbf{x} - \widetilde{\mathbf{x}}^k, -A^\mathsf{T}\lambda^k \rangle \geq \langle \mathbf{x} - \widetilde{\mathbf{x}}^k, (rA^\mathsf{T}A + Q)(\mathbf{x}^k - \widetilde{\mathbf{x}}^k) \rangle - \frac{L_\theta}{2}\|\mathbf{x}^k - \widetilde{\mathbf{x}}^k\|^2.$$

Combine the last inequality with the inequality (10) to achieve

$$\widetilde{\mathbf{w}}^k \in \mathcal{M}, \ \theta(\mathbf{x}) - \theta(\widetilde{\mathbf{x}}^k) + \left\langle \mathbf{w} - \widetilde{\mathbf{w}}^k, \mathcal{J}(\mathbf{w}) \right\rangle \geq \left\langle \mathbf{w} - \widetilde{\mathbf{w}}^k, H(\mathbf{w}^k - \widetilde{\mathbf{w}}^k) \right\rangle - \frac{L_\theta}{2}\|\mathbf{x}^k - \widetilde{\mathbf{x}}^k\|^2$$

with H given by (8). Similar to the proof of Theorem 2.1, we will deduce

$$\left\|\mathbf{w}^{k+1} - \mathbf{w}^*\right\|_H^2 \leq \left\|\mathbf{w}^k - \mathbf{w}^*\right\|_H^2 - \frac{2-\gamma}{\gamma}\left\|\mathbf{w}^k - \mathbf{w}^{k+1}\right\|_{\tilde{H}}^2, \ \forall \mathbf{w}^* \in \mathcal{M}^*, \tag{28}$$

where $\tilde{H} = H - \mathrm{diag}\left(\frac{L_\theta}{2-\gamma}\mathbf{I}, \mathbf{0}\right)$. Clearly, if $r > 0$ and $Q \succ \frac{L_\theta}{2-\gamma}\mathbf{I}$, then (28) implies that this linearized P-rALM converges with the same convergence rate as P-rALM.

3 Extensions of P-rALM for multi-block problem

In this section, we discuss two interesting extensions of P-rALM for solving the following multi-block separable convex minimization problem

$$\min\left\{\theta(\mathbf{x}) := \sum_{i=1}^p \theta_i(\mathbf{x}_i) | \ \sum_{i=1}^p A_i\mathbf{x}_i = b \ (\text{or} \geq b), \ \mathbf{x}_i \in \mathcal{X}_i\right\}, \tag{29}$$

where $\theta_i : \mathcal{R}^{n_i} \to \mathcal{R}, i = 1, 2, \cdots, p$ are closed proper convex functions; $\mathcal{X}_i \subseteq \mathcal{R}^{n_i}$ are closed convex sets; $A_i \in \mathcal{R}^{m \times n_i}$ and $b \in \mathcal{R}^m$ are given data. For this problem, we denote

$$\mathcal{M} := \prod_{i=1}^p \mathcal{X}_i \times \Lambda, \quad \text{where } \Lambda := \begin{cases} \mathcal{R}^m, & \text{if } \sum_{i=1}^p A_i\mathbf{x}_i = b, \\ \mathcal{R}^m_+, & \text{if } \sum_{i=1}^p A_i\mathbf{x}_i \geq b. \end{cases}$$

An extended primal-dual version of P-rALM (denoted by **PD-rALM**) is described as follows.

```
Initialize $(\mathbf{x}_1^0,\dots,\mathbf{x}_p^0,\lambda^0)$, choose $\gamma\in(0,2), r_i>0,\ Q_i\succ\mathbf{0}$ for $i=1,2,\dots,p$;
While stopping criteria is not satisfied do
    For $i=1,2,\cdots,p$, parallelly update
        $\widetilde{\mathbf{x}}_i^k=\arg\min_{\mathbf{x}_i\in\mathcal{X}_i}\left\{\theta_i(\mathbf{x}_i)-\langle\lambda^k,A_i\mathbf{x}_i-b\rangle+\frac{r_i}{2}\left\|A_i(\mathbf{x}_i-\mathbf{x}_i^k)\right\|^2+\frac{1}{2}\left\|\mathbf{x}_i-\mathbf{x}_i^k\right\|_{Q_i}^2\right\}$;
    End for
    $\widetilde{\lambda}^k=\mathcal{P}_\Lambda\Big(\lambda^k-\frac{1}{\sum_{j=1}^p\frac{1}{r_j}}\left[\sum_{i=1}^pA_i(2\widetilde{\mathbf{x}}_i^k-\mathbf{x}_i^k)-b\right]\Big)$;
    $\begin{pmatrix}\mathbf{x}_1^{k+1}\\ \vdots\\ \mathbf{x}_p^{k+1}\\ \lambda^{k+1}\end{pmatrix}=\begin{pmatrix}\mathbf{x}_1^{k}\\ \vdots\\ \mathbf{x}_p^{k}\\ \lambda^{k}\end{pmatrix}+\gamma\begin{pmatrix}\widetilde{\mathbf{x}}_1^k-\mathbf{x}_1^k\\ \vdots\\ \widetilde{\mathbf{x}}_p^k-\mathbf{x}_p^k\\ \widetilde{\lambda}^k-\lambda^k\end{pmatrix}$;
End while
```

Analogous to the analysis in Section 2, the saddle-point $\mathbf{w}^*=(\mathbf{x}_1^*;\cdots;\mathbf{x}_p^*;\lambda^*)\in\mathcal{M}$ of the Lagrangian function of (29) will satisfy the previous VI$(\theta,\mathcal{J},\mathcal{M})$ but with new notations

$$\mathbf{w}=\begin{pmatrix}\mathbf{x}\\ \lambda\end{pmatrix}=\begin{pmatrix}\mathbf{x}_1\\ \vdots\\ \mathbf{x}_p\\ \lambda\end{pmatrix},\quad \mathcal{J}(\mathbf{w})=\begin{pmatrix}-A_1^\top\lambda\\ \vdots\\ -A_p^\top\lambda\\ \sum_{i=1}^pA_i\mathbf{x}_i-b\end{pmatrix}.$$

Compared to the traditional ALM, the above PD-rALM is capable of exploiting the properties of each component objective function and could make the subproblems much easier. Next, we briefly analyze the convergence of this PD-rALM.

Lemma 3.1 *The sequence $\{\mathbf{w}^k\}$ generated by PD-rALM satisfies*

$$\widetilde{\mathbf{w}}^k\in\mathcal{M},\ \theta(\mathbf{x})-\theta(\widetilde{\mathbf{x}}^k)+\left\langle\mathbf{w}-\widetilde{\mathbf{w}}^k,\mathcal{J}(\mathbf{w})\right\rangle\geq\left\langle\mathbf{w}-\widetilde{\mathbf{w}}^k,H(\mathbf{w}^k-\widetilde{\mathbf{w}}^k)\right\rangle \tag{30}$$

for any $\mathbf{w}\in\mathcal{M}$, where

$$\widetilde{\mathbf{w}}^k=\begin{pmatrix}\widetilde{\mathbf{x}}^k\\ \widetilde{\lambda}^k\end{pmatrix},\ H=\begin{bmatrix}r_1A_1^\top A_1+Q_1 & \mathbf{0} & \cdots & \mathbf{0} & A_1^\top\\ \mathbf{0} & r_2A_2^\top A_2+Q_2 & \cdots & \mathbf{0} & A_2^\top\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ \mathbf{0} & \mathbf{0} & \cdots & r_pA_p^\top A_p+Q_p & A_p^\top\\ A_1 & A_2 & \cdots & A_p & \sum_{i=1}^p\frac{1}{r_i}\mathbf{I}\end{bmatrix} \tag{31}$$

is symmetric positive definite for any $r_i>0$ and $Q_i\succ\mathbf{0}$. Moreover, we have

$$\left\|\mathbf{w}^{k+1}-\mathbf{w}^*\right\|_H^2\leq\left\|\mathbf{w}^k-\mathbf{w}^*\right\|_H^2-\frac{2-\gamma}{\gamma}\left\|\mathbf{w}^k-\mathbf{w}^{k+1}\right\|_H^2,\ \forall\mathbf{w}^*\in\mathcal{M}^*. \tag{32}$$

Proof. First of all, it follows from first-order optimality condition of the $\mathbf{x}_i$-subproblem ($i=1,2,\dots,p,$) in PD-rALM that $\widetilde{\mathbf{x}}_i^k\in\mathcal{X}_i$ and

$$\theta_i(\mathbf{x}_i)-\theta_i(\widetilde{\mathbf{x}}_i^k)+\left\langle\mathbf{x}_i-\widetilde{\mathbf{x}}_i^k,-A_i^\top\lambda^k+\left(r_iA_i^\top A_i+Q_i\right)(\widetilde{\mathbf{x}}_i^k-\mathbf{x}_i^k)\right\rangle\geq0,\ \forall\mathbf{x}_i\in\mathcal{X}_i,$$

in other words,

$$\theta_i(\mathbf{x}_i)-\theta_i(\widetilde{\mathbf{x}}_i^k)+\Big\langle \mathbf{x}_i-\widetilde{\mathbf{x}}_i^k,-A_i^{\mathsf{T}}\widetilde{\lambda}^k\Big\rangle \geq \Big\langle \mathbf{x}_i-\widetilde{\mathbf{x}}_i^k,\Big(r_iA_i^{\mathsf{T}}A_i+Q_i\Big)(\mathbf{x}_i^k-\widetilde{\mathbf{x}}_i^k)+A_i^{\mathsf{T}}(\lambda^k-\widetilde{\lambda}^k)\Big\rangle .\quad(33)$$

Besides, it follows from the update of $\widetilde{\lambda}^k$ that $\widetilde{\lambda}^k\in\Lambda$ and

$$\Big\langle \lambda-\widetilde{\lambda}^k,\sum_{i=1}^p A_i\widetilde{\mathbf{x}}_i^k-b\Big\rangle\geq\Big\langle \lambda-\widetilde{\lambda}^k,\sum_{i=1}^p A_i(\mathbf{x}_i^k-\widetilde{\mathbf{x}}_i^k)+\sum_{j=1}^p\frac{1}{r_j}(\lambda^k-\widetilde{\lambda}^k)\Big\rangle,\ \forall\lambda\in\Lambda. \quad (34)$$

Finally, combine the inequalities (33)-(34) together with the structure of H and the monotonicity of $\mathcal{J}(\mathbf{w})$ to confirm the result in (30).

Note that the matrix $H=\bar{H}+\mathrm{diag}(Q_1,\cdots,Q_p,\mathbf{0})$ where

$$\begin{aligned}\bar{H} &= \begin{bmatrix} r_1A_1^{\mathsf{T}}A_1 & \cdots & \mathbf{0} & A_1^{\mathsf{T}}\\ \vdots & \ddots & \vdots & \vdots\\ \mathbf{0} & \cdots & r_pA_p^{\mathsf{T}}A_p & A_p^{\mathsf{T}}\\ A_1 & \cdots & A_p & \sum\limits_{i=1}^p\frac{1}{r_i}\mathbf{I}\end{bmatrix}\\ &= \begin{bmatrix} r_1A_1^{\mathsf{T}}A_1 & \cdots & \mathbf{0} & A_1^{\mathsf{T}}\\ \vdots & \ddots & \vdots & \vdots\\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0}\\ A_1 & \cdots & \mathbf{0} & \frac{1}{r_1}\mathbf{I}\end{bmatrix}+\cdots+\begin{bmatrix} \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0}\\ \vdots & \ddots & \vdots & \vdots\\ \mathbf{0} & \cdots & r_pA_p^{\mathsf{T}}A_p & A_p^{\mathsf{T}}\\ \mathbf{0} & \cdots & A_p & \frac{1}{r_p}\mathbf{I}\end{bmatrix}\end{aligned}$$

$$=\begin{pmatrix}\sqrt{r_1}A_1^{\mathsf{T}}\\ \mathbf{0}\\ \vdots\\ \mathbf{0}\\ \frac{1}{\sqrt{r_1}}\mathbf{I}\end{pmatrix}\Big(\sqrt{r_1}A_1,\mathbf{0},\ldots,\mathbf{0},\frac{1}{\sqrt{r_1}}\mathbf{I}\Big)+\ldots+\begin{pmatrix}\mathbf{0}\\ \vdots\\ \mathbf{0}\\ \sqrt{r_p}A_p^{\mathsf{T}}\\ \frac{1}{\sqrt{r_p}}\mathbf{I}\end{pmatrix}\Big(\mathbf{0},\ldots,\mathbf{0},\sqrt{r_p}A_p,\frac{1}{\sqrt{r_p}}\mathbf{I}\Big)$$

For any $\mathbf{w}=(\mathbf{x};\lambda)\neq\mathbf{0}$, we have

$$\mathbf{w}^{\mathsf{T}}H\mathbf{w}=\sum_{i=1}^p\Big\|\sqrt{r_i}A_i\mathbf{x}_i+\frac{1}{\sqrt{r_i}}\lambda\Big\|^2+\sum_{i=1}^p\|\mathbf{x}_i\|_{Q_i}^2>0$$

and hence the matrix H is symmetric positive definite. Similar to the proof of Theorem 2.1, the inequality (32) can be also obtained. ■

According to the preliminary Lemma 3.1, the global convergence of PD-rALM and its sublinear convergence rate can be established as the rest parts of Section 2. Motivated by the structure of H in (31), we next present a dual-primal update of PD-rALM, which can be also regarded as a dual-primal extension of P-rALM.

Suppose $r_i>0,s_i>0$ and $Q_i\succeq r_iA_i^{\mathsf{T}}A_i$ for $i=1,2,\cdots,p$. Consider the following block matrix

$$H=\begin{bmatrix} Q_1+s_1\mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} & -A_1^{\mathsf{T}}\\ \mathbf{0} & Q_2+s_2\mathbf{I} & \cdots & \mathbf{0} & -A_2^{\mathsf{T}}\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ \mathbf{0} & \mathbf{0} & \cdots & Q_p+s_p\mathbf{I} & -A_p^{\mathsf{T}}\\ -A_1 & -A_2 & \cdots & -A_p & \sum\limits_{i=1}^p\frac{1}{r_i}\mathbf{I}\end{bmatrix}. \quad (35)$$

This new matrix H is symmetric positive definite since

$$H \succeq \underbrace{\begin{bmatrix} r_1A_1^\mathsf{T}A_1 & \cdots & \mathbf{0} & -A_1^\mathsf{T} \\ \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ -A_1 & \cdots & \mathbf{0} & \frac{1}{r_1}\mathbf{I} \end{bmatrix} + \cdots + \begin{bmatrix} \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \cdots & r_pA_p^\mathsf{T}A_p & -A_p^\mathsf{T} \\ \mathbf{0} & \cdots & -A_p & \frac{1}{r_p}\mathbf{I} \end{bmatrix}} + \begin{bmatrix} s_1\mathbf{I} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \cdots & \vdots & \vdots \\ \mathbf{0} & \cdots & s_p\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \end{bmatrix},$$

$$= \begin{pmatrix} \sqrt{r_1}A_1^\mathsf{T} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \\ -\frac{1}{\sqrt{r_1}}\mathbf{I} \end{pmatrix} \Big(\sqrt{r_1}A_1, \mathbf{0}, \ldots, \mathbf{0}, -\tfrac{1}{\sqrt{r_1}}\mathbf{I}\Big) + \ldots + \begin{pmatrix} \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \sqrt{r_p}A_p^\mathsf{T} \\ -\frac{1}{\sqrt{r_p}}\mathbf{I} \end{pmatrix} \Big(\mathbf{0}, \ldots, \mathbf{0}, \sqrt{r_p}A_p, -\tfrac{1}{\sqrt{r_p}}\mathbf{I}\Big)$$

and

$$\mathbf{w}^\mathsf{T}H\mathbf{w} = \sum_{i=1}^{p} \big\|\sqrt{r_i}A_i\mathbf{x}_i - \frac{1}{\sqrt{r_i}}\lambda\big\|^2 + \sum_{i=1}^{p} s_i\big\|\mathbf{x}_i\big\|^2 > 0$$

for any $\mathbf{w} = (\mathbf{x}; \lambda) \neq \mathbf{0}$. Substituting the above H into (30), it is not difficulty to obtain the following dual-primal updates (denoted by **DP-rALM**).

```
**Initialize** $(\mathbf{x}_1^0, \ldots, \mathbf{x}_p^0, \lambda^0)$, choose $\gamma \in (0,2), r_i > 0, s_i > 0,\ Q_i \succ r_iA_i^\mathsf{T}A_i$ for $i = 1, 2, \ldots, p$;
**While** stopping criteria is not satisfied **do**
    $\widetilde{\lambda}^k = \mathcal{P}_\Lambda\Big(\lambda^k - \frac{1}{\sum_{j=1}^{p}\frac{1}{r_j}}\Big[\sum_{i=1}^{p} A_i\mathbf{x}_i^k - b\Big]\Big)$;
    **For** $i = 1, 2, \cdots, p$, **parallelly update**
        $\widetilde{\mathbf{x}}_i^k = \arg\min_{\mathbf{x}_i \in \mathcal{X}_i}\Big\{\theta_i(\mathbf{x}_i) - \langle 2\widetilde{\lambda}^k - \lambda^k, A_i\mathbf{x}_i - b\rangle + \frac{1}{2}\big\|\mathbf{x}_i - \mathbf{x}_i^k\big\|_{Q_i + s_i\mathbf{I}}^2\Big\}$;
    **End for**

$\begin{pmatrix} \mathbf{x}_1^{k+1} \\ \vdots \\ \mathbf{x}_p^{k+1} \\ \lambda^{k+1} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^{k} \\ \vdots \\ \mathbf{x}_p^{k} \\ \lambda^{k} \end{pmatrix} + \gamma \begin{pmatrix} \widetilde{\mathbf{x}}_1^k - \mathbf{x}_1^k \\ \vdots \\ \widetilde{\mathbf{x}}_p^k - \mathbf{x}_p^k \\ \widetilde{\lambda}^k - \lambda^k \end{pmatrix}$;
**End while**
```

Although the condition of Q_i is strict than that in the previous PD-rALM, similar results to Lemma 3.1 can be obtained. In fact, For each $i = 1, 2, \cdots, p$, the first-order optimality conditions of $\widetilde{\mathbf{x}}_i^k$-subproblem is $\widetilde{\mathbf{x}}_i^k \in \mathcal{X}_i$ and

$$\theta_i(\mathbf{x}_i) - \theta_i(\widetilde{\mathbf{x}}_i^k) + \big\langle \mathbf{x}_i - \widetilde{\mathbf{x}}_i^k, -A^\mathsf{T}\widetilde{\lambda}^k\big\rangle \geq \big\langle \mathbf{x}_i - \widetilde{\mathbf{x}}_i^k, (Q_i + s_i\mathbf{I})(\mathbf{x}_i^k - \widetilde{\mathbf{x}}_i^k) - A_i^\mathsf{T}(\lambda^k - \widetilde{\lambda}^k)\big\rangle \tag{36}$$

for any $\mathbf{x}_i \in \mathcal{X}_i$. Meanwhile, the update of $\widetilde{\lambda}^k$ implies $\widetilde{\lambda}^k \in \Lambda$ and

$$\big\langle \lambda - \widetilde{\lambda}^k, \sum_{i=1}^{p} A_i\widetilde{\mathbf{x}}_i^k - b\big\rangle \geq \big\langle \lambda - \widetilde{\lambda}^k, -\sum_{i=1}^{p} A_i(\mathbf{x}_i^k - \widetilde{\mathbf{x}}_i^k) + \sum_{j=1}^{p}\frac{1}{r_j}(\lambda^k - \widetilde{\lambda}^k)\big\rangle,\ \forall\lambda \in \Lambda. \tag{37}$$

Combining (36) and (37), the inequalities (30) and (32) also hold but with the matrix H replaced by (35). So, this DP-rALM also converges with a sublinear convergence rate.

4 Numerical experiments

In this section, we investigate the performance of the proposed algorithm for solving two popular optimization problems in machine learning. All the forthcoming experiments are implemented in MATLAB R2019b (64-bit) and performed on a PC with Windows 10 operating system, with an Intel i7-8565U CPU and 16GB RAM.

4.1 Linear support vector machine

One fundamental function of machine learning is to make classification from a number of labeled training data. Suppose these training data are $\{(x_i, y_i)\}_{i=1}^m$, where $x_i \in \mathcal{R}^n$ are feature vectors and $y_i \in \{-1, 1\}$ are the labels of sample. If these two kinds of examples formulate two disjoint convex hulls in $\mathcal{R}^n$, then we can find a hyperplane $\{x \mid w^\mathsf{T} x + a = 0\}$ to separate them because of the well-known strong separation theorem. The linear support vector machine (abbreviated by SVM, see e.g. [28]) is to find the maximum margin hyperplane separating two classes of data as much as possible, which leads to the following optimization problem

$$\min_{w\in\mathcal{R}^n, a\in\mathcal{R}} \Big\{\frac{1}{2}\|w\|^2 \mid y_i\big(w^\mathsf{T} x_i + a\big) \geq 1, i = 1, \cdots, m\Big\}.$$

Introduce the following new notations

$$u = \begin{pmatrix} w \\ a \end{pmatrix}, \; F = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \; A = \begin{pmatrix} y_1(x_1^\mathsf{T}, 1) \\ \vdots \\ y_m(x_m^\mathsf{T}, 1) \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

to reformulate the above SVM model as

$$\min_{u\in\mathcal{R}^{m+1}} \Big\{\frac{1}{2}\|Fu\|^2 \mid Au \geq b\Big\} \tag{38}$$

which is clearly a special case of the problem (1). In fact, SVM had been successfully applied in aero-engine fault diagnosis [10, 15]. This subsection aims to test the numerical performance of our proposed method for solving the problem (38). Applying the proposed P-rALM with $Q = \varrho\mathbf{I} - rA^\mathsf{T} A$ to (38), the resulting key iterations are

$$\begin{cases} \widetilde{u}^k = (F^\mathsf{T} F + \varrho\mathbf{I})^{-1}(A^\mathsf{T}\lambda^k + \varrho u^k), \\ \widetilde{\lambda}^k = \mathcal{P}_{\mathcal{R}_+^m}\left[\lambda^k - r(A(2\widetilde{u}^k - u^k) - b)\right]. \end{cases}$$

An inverse operation is involved in the update of $\widetilde{u}^k$, but it is a fixed constant in the loop. The penalty parameter r plays an important role in the performance of ALM. To achieve a relatively better value of the penalty parameter r, we first test its effect on the performance of our basic algorithms P-rALM and DP-rALM for solving the problem (38) with different data number m. Throughout this subsection, the training data is generated by random numbers satisfying a normal distribution, and the following stopping criterion

$$\text{Opt\_err}(k) = \max\Big\{\|F^\mathsf{T} F u^k - A^\mathsf{T}\lambda^k\|, \|\min(Au^k - b, \mathbf{0})\|\Big\} < tol \tag{39}$$

is used to terminate P-rALM and DP-rALM under the maximal iteration number 2×10^6. With the same initial points $(u^0,\lambda^0)=(\texttt{ones(3,1)},\texttt{zeros(m,1)})$, Figure 1 shows the results about "Iter" (the number of iterations) and "CPU" (CPU time in seconds) along with the increase of r. We fix the tolerance $tol=10^{-2}$ and the parameters $(\varrho,\gamma)=(r(\rho(A^\mathsf{T}A)+0.1),1.8)$ for P-rALM, while $(Q,s,\gamma)=(\varrho\mathbf{I},10^{-3},1.8)$ for DP-rALM. We didn't test other values of r since the reporting iteration numbers and CPU time are worse than the results in Figure 1. It can be seen from Figure 1 that both P-rALM and DP-rALM are competitive and sensitive to r. After checking the reporting results of both iter and CPU in Figure 1, we find that $r=10^{-3}$ is relatively reasonable for $m=600$ to 900, while $r=3\times10^{-4}$ for $m=200$ to 500.

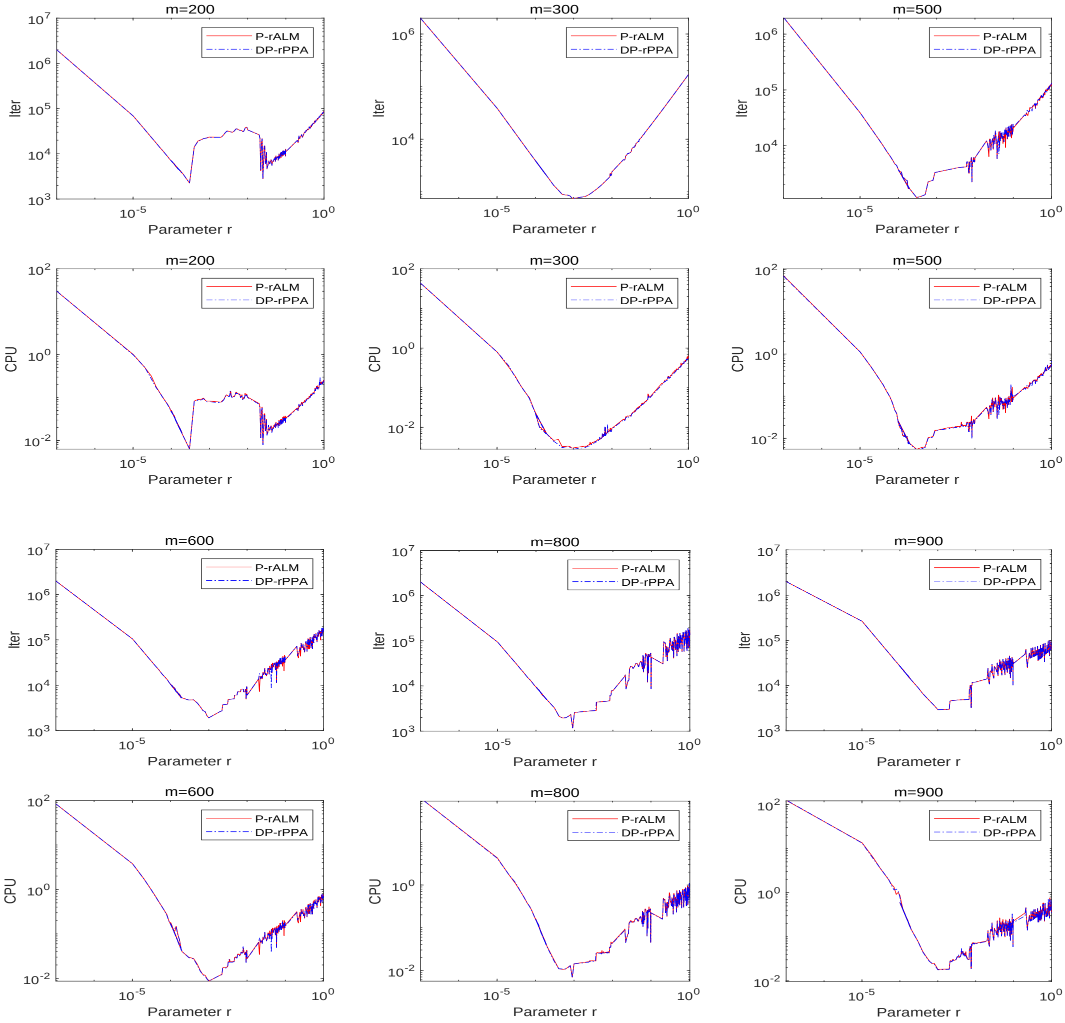

Figure 1: Effect of the parameter r on the performance of P-rALM and DP-rALM for solving (38).

To investigate the effect of the relaxation factor $\gamma\in(0,2)$ on the performance of P-rALM and DP-rALM, Figure 2 presents some comparative results of using different γ under $tol=10^{-8}$, from which we can see that $\gamma=1.9$ performs relatively better and it is set as the default value in the following experiments. Besides, it can be seen from Figure 2 that both P-rALM and DP-rALM enjoying a smaller $\gamma\in(0,1)$ leads to much worse results.

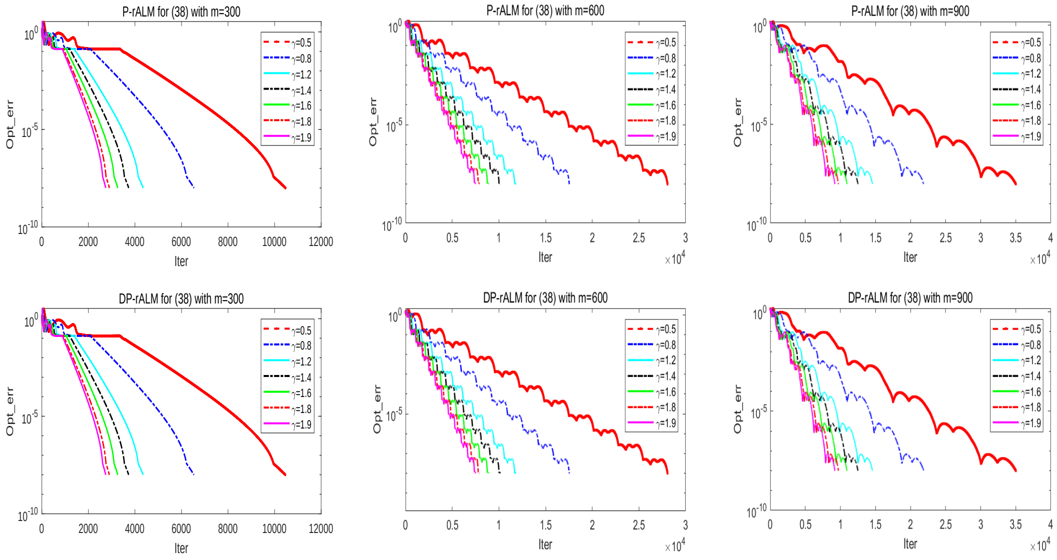

Figure 2: Effect of the relaxation parameter γ on the performance of P-rALM and DP-rALM.

A number of comparative experiments are further presented by comparing the following well-established first-order algorithms with our proposed algorithms P-rALM, DP-rALM, P-rALM with S1-S3, DP-rALM with S1-S3 (the strategies S1-S3 are emphasized in Remark 2.2), N-PDHG1 and N-PDHG2 together with the same relaxation step as P-rALM:

- I-IDL-ALM [23] with parameters $(\tau, r) = (0.75, \beta\rho(A^{\mathsf{T}}A) + 0.1)$ as the authors mentioned but with $\beta = 0.004$, which gives better performance than the original setting $\beta = 0.01$;
- C-PPA [18] with $(\gamma, s) = (1.9, 1.01\rho(A^{\mathsf{T}}A)/r)$ where we use adaptive value $r = m/8.5$;
- Generalized Primal-Dual Algorithm (G-PDA, [25]) with parameters

$$\tau = c_1/\sqrt{(1-\alpha+\alpha^2)\rho(A^{\mathsf{T}}A)}, \quad \sigma = c_2/\sqrt{(1-\alpha+\alpha^2)\rho(A^{\mathsf{T}}A)}$$

  to satisfy the convergence condition $\frac{1}{\tau\sigma} > (1-\alpha+\alpha^2)\rho(A^{\mathsf{T}}A)$, and we set $(c_1, c_2, \alpha) = (2, 1/c_1 - 0.001, 0.5)$ in the next experiments;
- G-PDHG [26] with the same setting as G-PDA for the parameters (τ, σ) to satisfy the convergence condition $\frac{1}{\tau\sigma} > 0.75\rho(A^{\mathsf{T}}A)$.

Here, we emphasize that although the parameters of G-PDHG and G-PDA are chosen in the same way, their frameworks are different. More specifically, the dual subproblem of G-PDA enjoys an expansion step with stepsize parameter $\alpha \in [0, 1]$, while G-PDA doesn't exploit this step; G-PDA has a correction step for the dual variable, while G-PDHG has a correction step for the primal variable. The primal-dual algorithms G-PDA, G-PDHG, N-PDHG1 and N-PDHG2 are used to solve the saddle-point problem of (38): $\min_u \max_{\lambda \geq 0} \frac{1}{2}\|Fu\|^2 - \langle \lambda, Au - b\rangle$. The parameters of our proposed algorithms use the tuned values as we mentioned before, and the parameter c in Remark 2.2 is taken as 0.1.

Size	**P-rALM**				**DP-rALM**			
m	Iter	CPU	It_err	Opt_err	Iter	CPU	It_err	Opt_err
300	*2747*	**0.0163**	1.0250e-9	9.9916e-9	**2746**	*0.0169*	1.0363e-9	9.9119e-9
400	*11352*	0.1108	6.0521e-10	9.7526e-9	*11352*	**0.0911**	5.9308e-10	9.5964e-9
500	**8115**	0.0942	2.3510e-10	9.9957e-9	*8116*	**0.0671**	2.3558e-10	9.8216e-9
600	*7518*	**0.0567**	1.4976e-10	9.9693e-9	*7518*	0.0650	1.4987e-10	9.9922e-9
700	*13464*	0.1546	1.3101e-10	9.9949e-9	*13464*	**0.1205**	1.3062e-10	9.9744e-9
800	15411	0.1324	6.0571e-11	9.8826e-9	15411	0.1268	6.0614e-11	9.8904e-9
900	*9213*	0.0932	8.2811e-11	9.9745e-9	*9213*	0.1169	8.2823e-11	9.9762e-9
Size	**P-rALM with S1**				**DP-rALM with S1**			
m	Iter	CPU	It_err	Opt_err	Iter	CPU	It_err	Opt_err
300	5229	0.0284	5.3706e-10	9.9946e-9	5226	0.0302	5.4623e-10	9.9787e-9
400	21576	0.1711	2.8475e-10	9.8869e-9	21575	0.1350	2.7884e-10	9.9433e-9
500	15427	0.1048	1.0254e-10	9.8898e-9	15428	0.1034	1.0238e-10	9.9169e-9
600	14028	0.0982	7.6785e-11	9.9603e-9	14028	0.1158	7.6858e-11	9.9523e-9
700	25591	0.1787	6.8756e-11	9.9926e-9	25591	0.1642	6.8552e-11	9.9717e-9
800	29286	0.1856	3.2832e-11	9.9382e-9	29286	0.2233	3.2879e-11	9.9459e-9
900	17508	0.1449	4.3011e-11	9.9947e-9	17508	0.1421	4.3017e-11	9.9966e-9
Size	**P-rALM with S2**				**DP-rALM with S2**			
m	Iter	CPU	It_err	Opt_err	Iter	CPU	It_err	Opt_err
300	5237	0.0329	5.3541e-10	9.9660e-9	5234	0.0284	5.4445e-10	9.9477e-9
400	21585	0.1284	2.8481e-10	9.8627e-9	21584	0.1182	2.7889e-10	9.9193e-9
500	15435	0.1122	1.0230e-10	9.9962e-9	15437	0.1154	1.0255e-10	9.8377e-9
600	14037	0.1453	7.6807e-11	9.8755e-9	14037	0.1030	7.6880e-11	9.8675e-9
700	25600	0.2132	6.8760e-11	9.9934e-9	25600	0.1822	6.8555e-11	9.9726e-9
800	29295	0.1888	3.2903e-11	9.9494e-9	29295	0.1926	3.2950e-11	9.9571e-9
900	17517	0.1580	4.2983e-11	9.9851e-9	17517	0.1441	4.2990e-11	9.9870e-9
Size	**P-rALM with S3**				**DP-rALM with S3**			
m	Iter	CPU	It_err	Opt_err	Iter	CPU	It_err	Opt_err
300	5245	0.0310	4.0011e-10	9.9876e-9	5257	0.0448	7.2148e-10	9.9437e-9
400	21370	0.1374	3.3819e-10	9.8724e-9	21593	0.1257	4.7754e-10	9.8583e-9
500	15386	0.1312	1.8824e-10	9.7337e-9	15483	0.1071	1.8188e-10	9.7578e-9
600	14288	0.1476	1.1117e-10	9.9965e-9	14160	0.1205	7.0168e-12	9.9993e-9
700	25430	0.1951	4.7062e-11	9.9807e-9	25717	0.1783	1.3591e-10	9.9731e-9
800	29274	0.2409	4.1498e-11	9.9694e-9	29288	0.2682	4.8997e-11	9.9513e-9
900	17522	0.2200	8.3521e-11	9.9658e-9	17522	0.2013	6.3857e-11	9.9893e-9
Size	**I-IDL-ALM**				**C-PPA**			
m	Iter	CPU	It_err	Opt_err	Iter	CPU	It_err	Opt_err
300	43099	0.2509	5.4987e-11	9.9979e-9	4791	0.0347	2.7731e-10	9.9823e-9
400	166303	0.5714	4.0977e-11	9.9895e-9	18504	0.1404	2.6511e-10	9.8792e-9
500	88062	0.3580	2.1486e-11	9.9728e-9	9840	0.0788	1.4402e-10	9.9608e-9
600	35955	0.2108	4.2482e-11	9.9628e-9	17021	0.1465	2.4382e-10	9.9742e-9
700	77014	0.4201	3.1677e-11	9.7316e-9	**13047**	0.1274	2.8562e-11	9.9909e-9
800	80489	0.4837	2.0271e-11	9.9714e-9	*8872*	*0.0826*	8.2374e-11	9.9742e-9
900	47650	0.3991	1.9483e-11	9.9859e-9	10383	0.1382	1.4948e-10	9.9613e-9
Size	**G-PDA**				**G-PDHG**			
m	Iter	CPU	It_err	Opt_err	Iter	CPU	It_err	Opt_err
300	27323	0.1607	9.3168e-11	9.9999e-9	27590	0.5680	9.3168e-11	9.9999e-9
400	91678	0.3795	7.0041e-11	9.9984e-9	87249	1.0196	7.9698e-11	9.9924e-9
500	47391	0.2808	1.5299e-11	9.9853e-9	42990	0.6013	2.3046e-11	9.9574e-9
600	19717	0.1511	6.0752e-11	9.9740e-9	13448	0.2437	5.2588e-11	9.9894e-9
700	29052	0.2019	6.0593e-11	9.9939e-9	31107	0.5585	6.0509e-11	9.9908e-9
800	32069	0.2502	2.8833e-11	9.9958e-9	31693	0.6043	5.6301e-11	9.9095e-9
900	14741	0.1709	4.4148e-11	9.9775e-9	17512	0.3928	5.3862e-11	9.9896e-9
Size	**N-PDHG1**				**N-PDHG2**			
m	Iter	CPU	It_err	Opt_err	Iter	CPU	It_err	Opt_err
300	*2747*	0.0231	1.0250e-9	9.9916e-9	3410	0.0212	1.4921e-10	9.9875e-9
400	*11352*	0.1095	6.0521e-10	9.7526e-9	**11342**	*0.1076*	5.9802e-10	9.7723e-9
500	**8115**	*0.0691*	2.3510e-10	9.9957e-9	9590	0.0916	2.5845e-10	9.5993e-9
600	*7518*	*0.0570*	1.4976e-10	9.9693e-9	**5720**	0.0595	9.2440e-11	9.9107e-9
700	*13464*	*0.1211*	1.3101e-10	9.9949e-9	13519	0.1520	1.2677e-10	9.9502e-9
800	15411	0.1339	6.0571e-11	9.8826e-9	**6721**	**0.0538**	3.1910e-11	9.9920e-9
900	*9213*	*0.0843*	8.2811e-11	9.9745e-9	**7081**	**0.0574**	1.2773e-10	9.9970e-9

Table 1: Comparative results of different algorithms for solving the problem (38).

Table 1 reports some comparative results of with different settings of total data number under given tolerance $tol = 10^{-8}$, where Opt_err denotes the final obtained residuals defined in (39) and It_err denotes the final obtained primal residuals defined as It_err$(k) = \|u^{k+1} - u^k\|$. The bold value in Table 1 denotes the smallest one for each problem and the italic value means a relatively smaller value. We also set $n = 2$ to visualize the comparative classification results of $m \in \{300, 900\}$ as shown in Figures 3-4. Figure 5 depicts the convergence curves of Opt_err(k) versus Iter when applying these state-of-the-art algorithms to solve (38) with $m = 900$. We didn't depict the curves about I-IDL-ALM in Figure 5 since it performs much worse than others. Figure 6 further shows the comparison of different algorithms for solving (38) with $m = 900$ under lower and higher tolerance errors. It can be seen from experimental results listed in Table 1 and convergence curves shown in Figures 5-6 that:

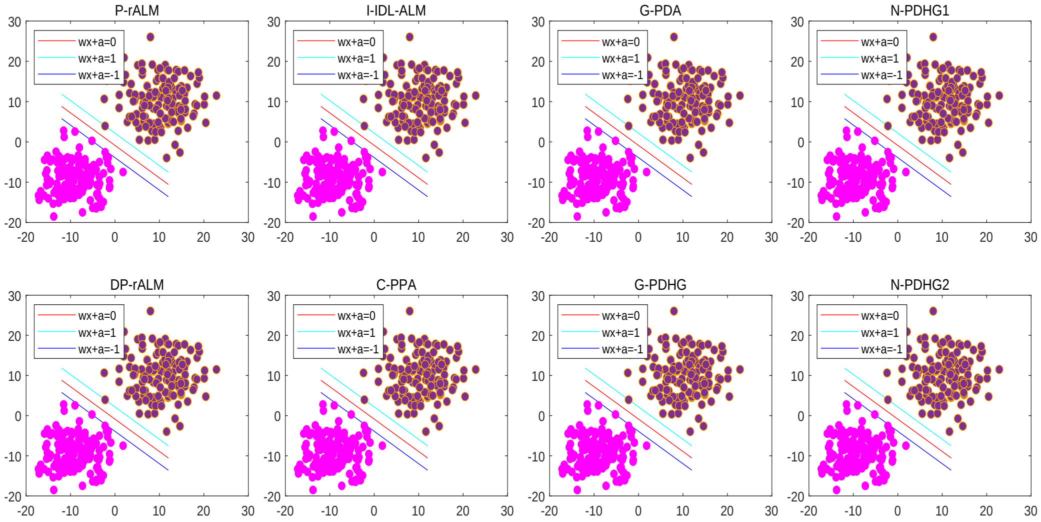

Figure 3: Classification results with 300 data points by state-of-the-art algorithms.

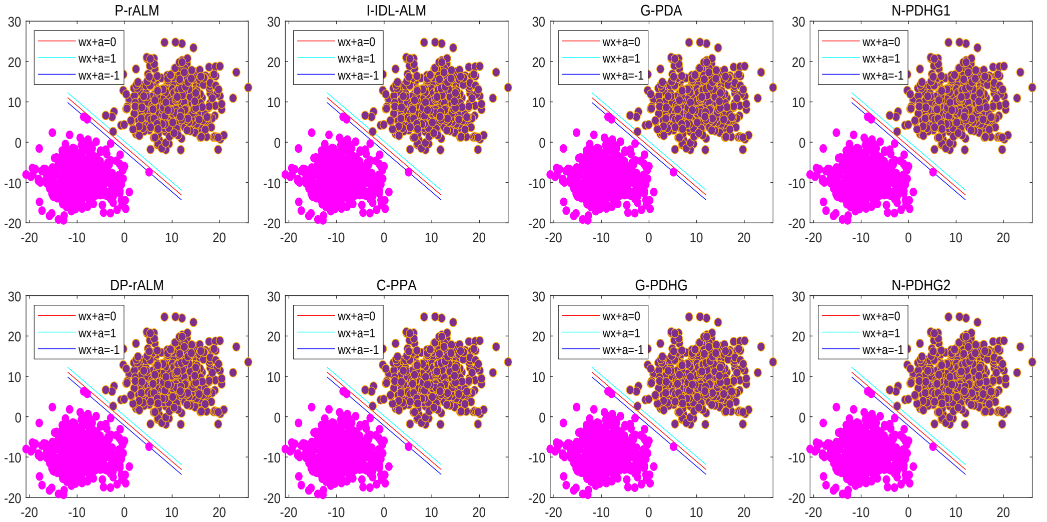

Figure 4: Classification results with 900 data points by state-of-the-art algorithms.

- The proposed methods P-rALM and DP-rALM are competitive and perform significantly better than each of them with S1-S3, which suggests that using a large fixed relaxation factor is better than using an adaptive sequence.
- Both P-rALM and DP-rALM perform significantly better than the recent developed methods I-IDL-ALM, G-PDA and G-PDHG for solving SVM (38) whether a lower or higher tolerance is required. In addition, both N-PDHG1 and N-PDHG2 (see appendix) are competitive to P-rALM and perform also better than I-IDL-ALM, G-PDA and G-PDHG, while N-PDHG2 seems more suitable for solving the large-scale SVM.
- Although C-PPA was proposed ten years ago, it performs sometimes competitive with our P-rALM, which is perhaps due to the adaptive update rules we suggested. However, in many cases this method performs worse than our P-rALM and DP-rALM, which can be observed from the reported Iter and CPU. Compared to all algorithms, I-IDL-ALM needs relatively more iteration numbers to satisfy the stopping condition (39).

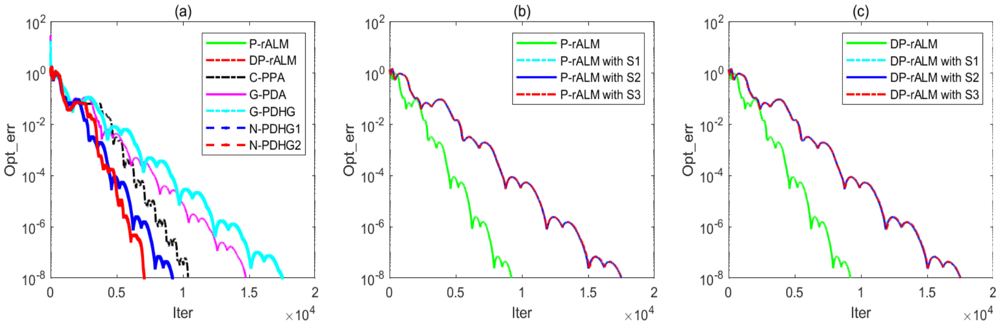

Figure 5: Comparison of Opt_err versus Iter by different algorithms for solving (38) with $m = 900$.

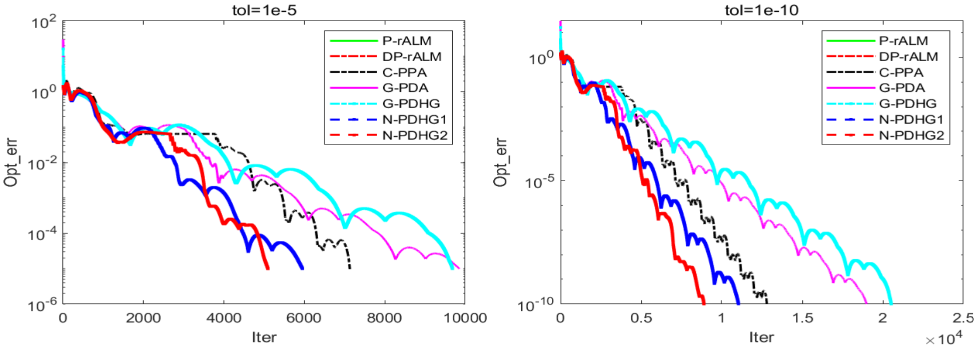

Figure 6: Comparison of different algorithms for solving (38) with $m = 900$ under different *tol*.

4.2 Robust principal component analysis

The Robust Principal Component Analysis (RPCA) was developed originally by Candes et al. [12], which aims to decompose a data matrix $D \in \mathcal{R}^{m \times n}$ into a low-rank matrix L and a sparse matrix S containing outliers and corrupt data. The principal components of L are robust to the outliers and corrupt data in S. This decomposition has a wide range of applications in e.g. video surveillance, face recognition, latent semantic indexing, machine learning and so forth, see e.g. [1, 9, 30, 34]. Mathematically, the goal is to find L and S satisfying the following separable nonconvex optimization problem

$$\min_{L,S\in\mathcal{R}^{m\times n}} \big\{\operatorname{rank}(L) + \|S\|_0 \mid L + S = D\big\}.$$

However, it is not a tractable optimization due to the non-convexity of rank function rank(L) and the sparse norm $\|S\|_0$. Similar to the technique to reformulate the compressed sensing problem, most researchers turn to the following convex relaxation form:

$$\min_{L,S\in\mathcal{R}^{m\times n}} \big\{\|L\|_* + \nu\|S\|_1 \mid L + S = D\big\}, \tag{40}$$

where $\|\cdot\|_*$ denotes the nuclear norm of a matrix (the sum of its singular values), $\|\cdot\|_1$ denotes the so-called l_1 norm of a matrix (the sum of its absolute values), and ν is a positive weighting parameter that provides a trade-off between the sparse and low rank components and usually it takes $1/\sqrt{\max(m,n)}$. Clearly, the problem (40) is a special case of the previous model (29) and hence the extended algorithms PD-rALM and DP-rALM in section 3 can be applied to solve it. For example, applying PD-rALM with $Q_i = \varrho_i\mathbf{I}$ for $i = 1, 2$, the resulting key iterations are

$$\begin{cases} \widetilde{L}^k = \arg\min\limits_{L\in\mathcal{R}^{m\times n}} \|L\|_* + \dfrac{r_1+\varrho_1}{2}\Big\|L - L^k - \dfrac{\Lambda^k}{r_1+\varrho_1}\Big\|_F^2, \\ \widetilde{S}^k = \arg\min\limits_{S\in\mathcal{R}^{m\times n}} \nu\|S\|_1 + \dfrac{r_2+\varrho_2}{2}\Big\|S - S^k - \dfrac{\Lambda^k}{r_2+\varrho_2}\Big\|_F^2 \\ \qquad = \operatorname{Shrink}\Big(S^k + \dfrac{\Lambda^k}{r_2+\varrho_2}, \dfrac{\nu}{r_2+\varrho_2}\Big), \\ \widetilde{\Lambda}^k = \Lambda^k - \dfrac{1}{1/r_1 + 1/r_2}\Big[(2\widetilde{L}^k - L^k) + (2\widetilde{S}^k - S^k) - D\Big], \end{cases}$$

where Shrink$(\cdot,\cdot)$ is the soft shrinkage operator (see e.g. [37]). And the L-subproblem admits the following explicit solution

$$\widetilde{L}^k = U^k \operatorname{diag}\Big(\max\Big\{\sigma_i^k - \frac{1}{r_1+\varrho_1}, 0\Big\}\Big)(V^k)^T,$$

where $U^k \in \mathcal{R}^{l\times r}, V^k \in \mathcal{R}^{n\times r}$ are obtained by the singular value decomposition: $L^k + \frac{\Lambda^k}{r_1+\varrho_1} = U^k\Sigma^k(V^k)^T$ with $\Sigma^k = \operatorname{diag}\big(\sigma_1^k, \sigma_1^k, \cdots, \sigma_r^k\big) \in \mathcal{R}^{r\times r}$.

By comparing to several well-established algorithms, in what follows we test the performance of the preliminary algorithms PD-rALM and DP-rALM for solving the problem (40) with Yale B database which consists of cropped and aligned images of 38 individuals under 9 poses and 64 lighting conditions[1]. The penalty parameter of the standard ADMM is fixed as $\frac{mn}{4\|D\|_1}$ according to [9, Page 109]. Similar to the parameter choice in ADMM, we set $r_1 = r_2, Q_1 = Q_2 = \varrho\mathbf{I}$ with $(r_1, \gamma, \varrho) = (\frac{mn}{5\|D\|_1}, 1.75, 10^{-6})$ for PD-rALM, while we set $s_1 = s_2$ with $(\varrho, s_1) = (r_1(1 + 10^{-3}), 10^{-4})$ for DP-rALM. The existing methods DP-BALM [38], PDHG [11], G-PDHG and G-PDA are used to solve the corresponding saddle-point problem $\min\limits_{L,S\in\mathbb{R}^{m\times n}} \max\limits_{Z\in\mathbb{R}^{m\times n}} \{\|L\|_* +$

[1] The database can be downloaded at http://vision.ucsd.edu/~iskwak/ExtYaleDatabase/ExtYaleB.html.

$\nu\|S\|_1 + \langle L+S-D, Z\rangle\}$. The parameters of DP-BALM is fixed as $(\alpha,\beta,\delta) = (1.9, 100, 10^{-3})$, both G-PDA and G-PDHG use the suggested values as mentioned in the second experiments [38] since the involved parameters (τ,σ) are restricted by the same condition $\tau\sigma\|KK^{\mathsf{T}}\| < 4/3$ and here $K = [\mathbf{I};\mathbf{I}]$. Due to the recent work [2], we set the inertial parameter $\theta = 0.8$ for PDHG. All mentioned algorithms are terminated when the following criterions

$$\text{RelChg(k)} := \frac{\left\|L^{k+1}-L^k\right\|_F + \left\|S^{k+1}-S^k\right\|_F}{\left\|L^k\right\|_F + \left\|S^k\right\|_F + 1} < \epsilon_1, \quad \text{Res(k)} := \frac{\|D - L^{k+1} - S^{k+1}\|_F}{\|D\|_F} < \epsilon_2$$

are satisfied with the same initial feasible points $(\Lambda^0, S^0) = (L^0, D - L^0)$, where (ϵ_1, ϵ_2) are given tolerance and L^0 is obtained by the truncated singular value decomposition:

$$L^0 = \texttt{F(:,1:l)Sigma(1:l,1:l)N(:,1:l)} \text{ where } \texttt{[F,Sigma,N]=svd(D,'econ');l=3.}$$

(ϵ_1, ϵ_2)	Methods	Iter	rank(L)	Time(s)	RelChg	Res
$(10^{-4}, 10^{-5})$	PD-rALM	*194*	37	**17.6648**	9.9673e-5	6.8302e-6
	DP-rALM	**192**	37	*18.4554*	9.9661e-5	6.8722e-6
	ADMM	254	31	21.0341	9.8391e-5	4.0032e-6
	DP-BALM	225	31	21.7985	5.9275e-5	9.9039e-6
	PDHG	280	31	23.0072	1.9046e-5	9.9674e-6
	G-PDHG	271	31	22.5811	2.0347e-5	9.9877e-6
	G-PDA	271	31	22.0591	1.9011e-5	9.9877e-6
$(10^{-5}, 10^{-6})$	PD-rALM	**343**	31	**30.8044**	9.8724e-6	7.4143e-7
	DP-rALM	*360*	31	*32.4381*	9.9899e-6	5.6868e-7
	ADMM	395	31	38.1146	9.8640e-6	4.7239e-7
	DP-BALM	469	31	46.5738	8.7892e-6	9.9287e-7
	PDHG	516	31	53.5624	1.8649e-6	9.9971e-7
	G-PDHG	508	31	49.6827	2.0340e-6	9.9375e-7
	G-PDA	508	31	48.8102	1.9017e-6	9.9375e-7
$(10^{-6}, 10^{-7})$	PD-rALM	**582**	31	**53.7565**	9.9349e-7	9.5168e-8
	DP-rALM	*605*	31	56.6440	9.9954e-7	8.6355e-8
	ADMM	619	31	57.4334	9.9702e-7	4.4109e-8
	DP-BALM	864	31	85.6948	9.9842e-7	7.0061e-8
	PDHG	988	31	93.3656	1.2053e-7	9.9848e-8
	G-PDHG	976	31	89.9472	1.4532e-7	9.9761e-8
	G-PDA	976	31	86.8632	1.2596e-7	9.9761e-8

Table 2: Comparative results of the state-of-the-art algorithms under different tolerances.

We report some comparative results of several state-of-the-art algorithms in Table 4.2 under different tolerances (ϵ_1, ϵ_2). The original columns of D, along with the low-rank and sparse components decomposed by different algorithms under $(\epsilon_1, \epsilon_2) = (10^{-5}, 10^{-6})$, are shown in Figure 7, and Figure 8 shows the comparative convergence curves of the relative residual Res(k) obtained along with the increasing of the CPU time and the iteration number. From Table 4.2 and Figure 8, it can be seen that both PD-rALM and DP-rALM perform better than other comparative methods in terms of the number of iterations and CPU time, and they can effectively fill in occluded regions of the image, corresponding to shadows. In the low-rank component L as shown in Figure 7, shadows under different lighting conditions are removed and filled in with the most consistent low-rank features from the eigenfaces.

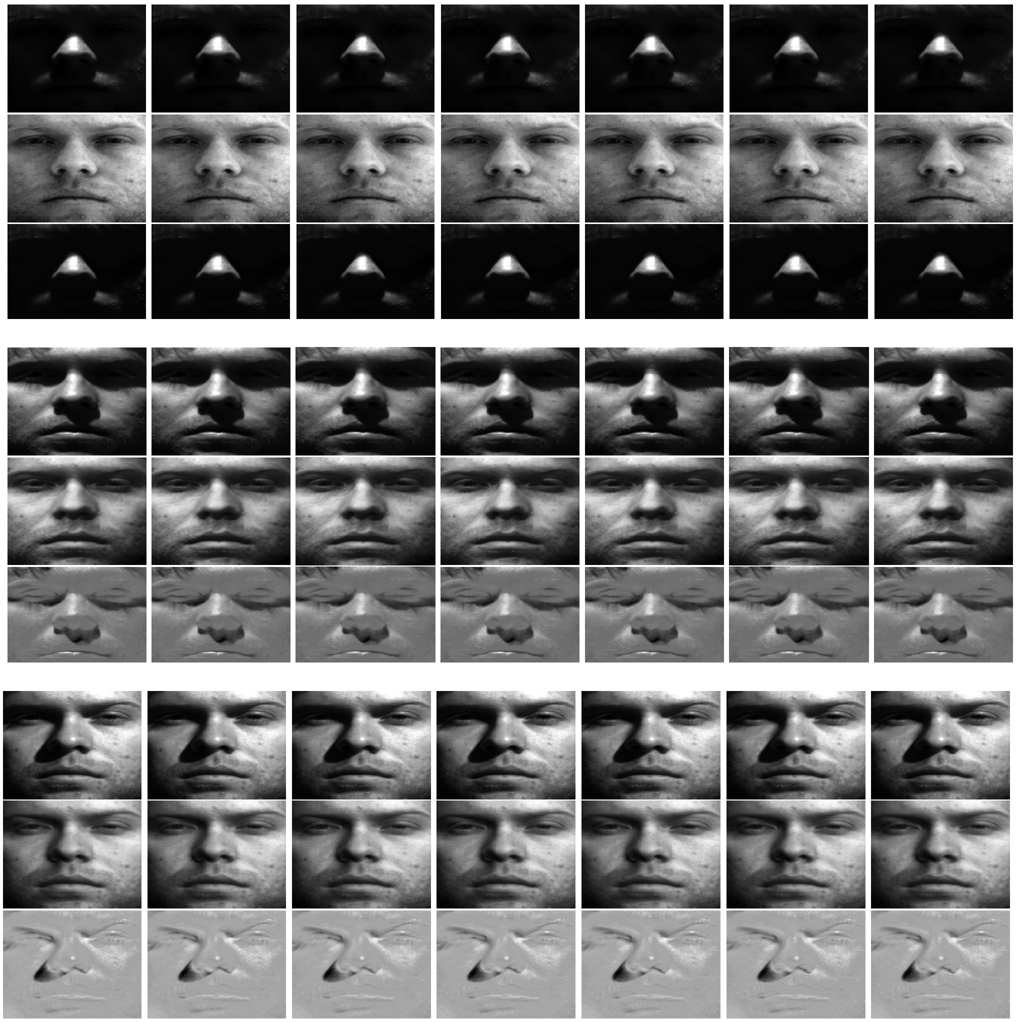

Figure 7: Output of different algorithms for the 4th(rows 1-3), 18th(rows 4-6) and 46th(rows 7-9) images in the Yale B database. From left to right: PD-rALM, DP-rALM, ADMM, DP-BALM, PDHG, G-PDHG, G-PDA, respectively.

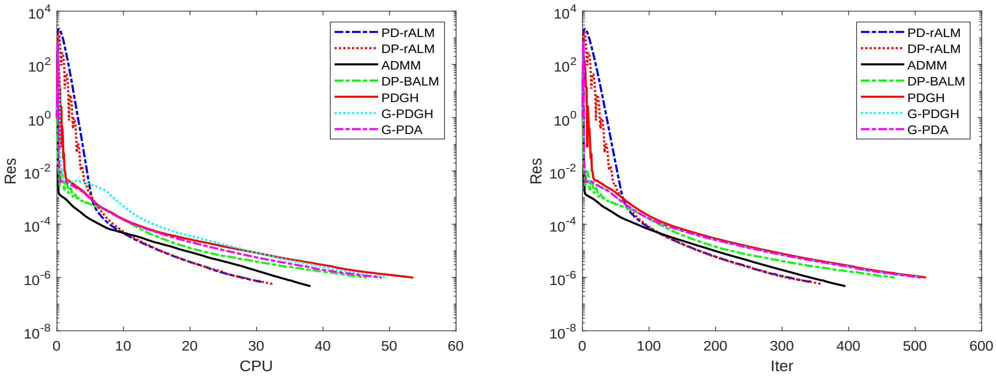

Figure 8: Comparative convergence curves of Res versus CPU time and iteration, respectively.

Funding This work was supported by Guangdong Basic and Applied Basic Research Foundation (2023A1515012405), Shaanxi Fundamental Science Research Project for Mathematics and Physics (22JSQ001), and National Natural Science Foundation of China (Grants 12071398 and 52372397).

Data availability The link https://github.com/pzheng4218/new-ALM-in-ML provides the matlab codes for experiments.

Conflict of interest The authors declared that, they do not have any commercial or associative interest that represents a conflict of interest in connection with this paper.

5 Appendix: discussions on two new PDHG

In this appendix, we discuss two new types of PDHG algorithm without relaxation step for solving the convex-concave saddle-point problem of the form

$$\min_{\mathbf{x}\in\mathcal{X}}\max_{\mathbf{y}\in\mathcal{Y}} \Phi(\mathbf{x},\mathbf{y}) := \theta_1(\mathbf{x}) - \mathbf{y}^\mathsf{T} A\mathbf{x} - \theta_2(\mathbf{y}), \tag{41}$$

or, equivalently, the composite problem $\min\limits_{\mathbf{x}\in\mathcal{X}} \big\{\theta_1(\mathbf{x}) + \theta_2^*(-A\mathbf{x})\big\}$, where $\mathcal{X} \subseteq \mathcal{R}^n, \mathcal{Y} \subseteq \mathcal{R}^m$ are closed convex sets, both $\theta_1 : \mathcal{X} \to \mathcal{R}$ and $\theta_2 : \mathcal{Y} \to \mathcal{R}$ are convex but possibly nonsmooth functions, θ_2^* is the conjugate function of θ_2, and $A \in \mathcal{R}^{m\times n}$ is a given data. A lot of practical examples can be reformulated as a special case of (41), see e.g. [27, Section 5]. Note that Problem (41) could reduce to the dual of (1) by letting $\theta_2 = -\lambda^\mathsf{T} b, \mathbf{y} = \lambda$ and $\mathcal{Y} = \Lambda$. Hence, the convergence results also hold for the previous P-rALM. Throughout the forthcoming discussions, the solution set of this problem is assumed to be nonempty.

The original PDHG proposed in [41] is to solve some TV image restorations models. Extending it to the problem (41), we get the following scheme:

$$\begin{cases} \mathbf{x}^{k+1} = \arg\min\limits_{\mathbf{x}\in\mathcal{X}} \Phi(\mathbf{x},\mathbf{y}^k) + \frac{r}{2}\|\mathbf{x}-\mathbf{x}^k\|^2, \\ \mathbf{y}^{k+1} = \arg\max\limits_{\mathbf{y}\in\mathcal{Y}} \Phi(\mathbf{x}^{k+1},\mathbf{y}) - \frac{s}{2}\|\mathbf{y}-\mathbf{y}^k\|^2, \end{cases}$$

where r, s are positive scalars. He, et al. [20] pointed out that convergence of the above PDHG can be ensured if θ_1 is strongly convex and $rs > \rho(A^\mathsf{T} A)$. To weaken these convergence conditions, e.g., the function θ_1 is only convex and the parameters r, s do not depend on $\rho(A^\mathsf{T} A)$, we develop the following novel PDHG (N-PDHG1) for solving the problem (41):

```
**Initialize** $(\mathbf{x}^0, \lambda^0)$ and choose $r > 0,\ Q \succ \mathbf{0}$;
**While** stopping criteria is not satisfied **do**
    $\mathbf{x}^{k+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \Phi(\mathbf{x},\mathbf{y}^k) + \frac{1}{2}\left\|\mathbf{x}-\mathbf{x}^k\right\|^2_{rA^\mathsf{T}A+Q}$;
    $\mathbf{y}^{k+1} = \arg\max_{\mathbf{y}\in\mathcal{Y}} \Phi(2\mathbf{x}^{k+1}-\mathbf{x}^k,\mathbf{y}) - \frac{1}{2r}\|\mathbf{y}-\mathbf{y}^k\|^2$;
**End while**
```

Another related algorithm (denoted by N-PDHG2) is just to modify the final subproblem of PDHG, whose framework is described as follows. A similar quadratic term was adopted in [24] to solve a special case of the problem (41). We can observe that N-PDHG1 has certain connections with P-rALM, since the first-order optimality conditions of their involved subproblems are reformulated as similar variational inequalities with the same block matrix H, see (7) and the next (42). Actually, their $\mathbf{x}$-subproblems enjoy the same proximal term. Another observation is that N-PDHG2 is developed from N-PDHG1 by just modifying the involved proximal parameters, and one of their subproblems could enjoy a proximity operator directly.

```
**Initialize** $(\mathbf{x}^0, \lambda^0)$ and choose $r > 0,\ Q \succ \mathbf{0}$;
**While** stopping criteria is not satisfied **do**
    $\mathbf{x}^{k+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \Phi(\mathbf{x}, \mathbf{y}^k) + \frac{r}{2}\left\|\mathbf{x} - \mathbf{x}^k\right\|^2;$
    $\mathbf{y}^{k+1} = \arg\max_{\mathbf{y}\in\mathcal{Y}} \Phi(2\mathbf{x}^{k+1} - \mathbf{x}^k, \mathbf{y}) - \frac{1}{2}\|\mathbf{y} - \mathbf{y}^k\|^2_{AA^\mathsf{T}/r+Q};$
**End while**
```

5.1 Sublinear convergence under general convex assumption

Because the above two algorithms are very similar, in the following we just analyze basic convergence properties of N-PDHG1 under general convex assumption and then briefly discuss the convergence of the second algorithm. For convenience, we denote $\mathcal{U} := \mathcal{X} \times \mathcal{Y}$ and

$$\theta(\mathbf{u}) = \theta_1(\mathbf{x}) + \theta_2(\mathbf{y}), \quad \mathbf{u} = \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}, \quad \mathbf{u}^k = \begin{pmatrix} \mathbf{x}^k \\ \mathbf{y}^k \end{pmatrix} \quad \text{and} \quad M = \begin{bmatrix} \mathbf{0} & -A^\mathsf{T} \\ A & \mathbf{0} \end{bmatrix}.$$

Lemma 5.1 *The sequence $\{\mathbf{u}^k\}$ generated by N-PDHG1 satisfies*

$$\mathbf{u}^{k+1} \in \mathcal{U},\ \theta(\mathbf{u}) - \theta(\mathbf{u}^{k+1}) + \langle \mathbf{u} - \mathbf{u}^{k+1}, M\mathbf{u}\rangle \geq \langle \mathbf{u} - \mathbf{u}^{k+1}, H(\mathbf{u}^k - \mathbf{u}^{k+1})\rangle \tag{42}$$

for any $\mathbf{u} \in \mathcal{U}$, where H is given by (8). Moreover, we have

$$\theta(\mathbf{u}) - \theta(\mathbf{u}^{k+1}) + \langle \mathbf{u} - \mathbf{u}^{k+1}, M\mathbf{u}\rangle \geq \frac{1}{2}\Big(\left\|\mathbf{u} - \mathbf{u}^{k+1}\right\|^2_H - \left\|\mathbf{u} - \mathbf{u}^k\right\|^2_H\Big) + \frac{1}{2}\left\|\mathbf{u}^k - \mathbf{u}^{k+1}\right\|^2_H. \tag{43}$$

Proof. According to the first-order optimality condition of the $\mathbf{x}$-subproblem in N-PDHG1, we have $\mathbf{x}^{k+1} \in \mathcal{X}$ and

$$\theta_1(\mathbf{x}) - \theta_1(\mathbf{x}^{k+1}) + \Big\langle \mathbf{x} - \mathbf{x}^{k+1}, -A^\mathsf{T}\mathbf{y}^k + \Big(rA^\mathsf{T}A + Q\Big)(\mathbf{x}^{k+1} - \mathbf{x}^k)\Big\rangle \geq 0,\ \forall \mathbf{x} \in \mathcal{X}, \tag{44}$$

that is,

$$\begin{aligned} &\theta_1(\mathbf{x}) - \theta_1(\mathbf{x}^{k+1}) + \Big\langle \mathbf{x} - \mathbf{x}^{k+1}, -A^\mathsf{T}\mathbf{y}^{k+1}\Big\rangle \\ \geq &\Big\langle \mathbf{x} - \mathbf{x}^{k+1}, \Big(rA^\mathsf{T}A + Q\Big)(\mathbf{x}^k - \mathbf{x}^{k+1}) + A^\mathsf{T}(\mathbf{y}^k - \mathbf{y}^{k+1})\Big\rangle. \end{aligned} \tag{45}$$

Similarly, we have $\mathbf{y}^{k+1} \in \mathcal{Y}$ and

$$\theta_2(\mathbf{y}) - \theta_2(\mathbf{y}^{k+1}) + \Big\langle \mathbf{y} - \mathbf{y}^{k+1}, A(2\mathbf{x}^{k+1} - \mathbf{x}^k) + \frac{1}{r}(\mathbf{y}^{k+1} - \mathbf{y}^k)\Big\rangle \geq 0,\ \forall \mathbf{y} \in \mathcal{Y}, \tag{46}$$

that is,

$$\theta_2(\mathbf{y})-\theta_2(\mathbf{y}^{k+1})+\left\langle \mathbf{y}-\mathbf{y}^{k+1}, A^{\mathsf{T}}\mathbf{x}^{k+1}\right\rangle \geq \left\langle \mathbf{y}-\mathbf{y}^{k+1}, A(\mathbf{x}^k-\mathbf{x}^{k+1})+\frac{1}{r}(\mathbf{y}^k-\mathbf{y}^{k+1})\right\rangle. \quad (47)$$

Combine the inequalities (45)-(47) and the structure of H given by (8) to have

$$\theta(\mathbf{u})-\theta(\mathbf{u}^{k+1})+\left\langle \mathbf{u}-\mathbf{u}^{k+1}, M\mathbf{u}^{k+1}\right\rangle \geq \left\langle \mathbf{u}-\mathbf{u}^{k+1}, H(\mathbf{u}^k-\mathbf{u}^{k+1})\right\rangle,$$

which together with the the property $\langle \mathbf{u}-\mathbf{u}^{k+1}, M(\mathbf{u}-\mathbf{u}^{k+1})\rangle = 0$ confirms (42). The inequality (43) can be obtained by applying (42) and the identity in (21). ■

Now, we discuss the global convergence and sublinear convergence rate of N-PDHG1. Let $\mathbf{u}^*=(\mathbf{x}^*;\mathbf{y}^*)\in\mathcal{U}$ be a solution point of the problem (41). Then, it holds

$$\Phi_{\mathbf{y}\in\mathcal{Y}}(\mathbf{x}^*,\mathbf{y}) \leq \Phi(\mathbf{x}^*,\mathbf{y}^*) \leq \Phi_{\mathbf{x}\in\mathcal{X}}(\mathbf{x},\mathbf{y}^*),$$

namely,

$$\begin{cases} \mathbf{x}^*\in\mathcal{X} & \theta_1(\mathbf{x})-\theta_1(\mathbf{x}^*) & + & \langle \mathbf{x}-\mathbf{x}^*, -A^{\mathsf{T}}\mathbf{y}^*\rangle \geq 0, & \forall \mathbf{x}\in\mathcal{X}, \\ \mathbf{y}^*\in\mathcal{Y} & \theta_2(\mathbf{y})-\theta_2(\mathbf{y}^*) & + & \langle \mathbf{y}-\mathbf{y}^*, A\mathbf{x}^*\rangle \geq 0, & \forall \mathbf{x}\in\mathcal{Y}. \end{cases}$$

So, finding a solution point of (41) amounts to finding $\mathbf{u}^*\in\mathcal{U}$ such that

$$\mathbf{u}^*\in\mathcal{U},\ \theta(\mathbf{u})-\theta(\mathbf{u}^*)+\langle \mathbf{u}-\mathbf{u}^*, M\mathbf{u}^*\rangle \geq 0, \quad \forall \mathbf{u}\in\mathcal{U}. \quad (48)$$

Setting $\mathbf{u}:=\mathbf{u}^*$ in (43) together with (48) gives

$$\left\|\mathbf{u}^*-\mathbf{u}^{k+1}\right\|_H^2 \leq \left\|\mathbf{u}^*-\mathbf{u}^k\right\|_H^2 - \left\|\mathbf{u}^k-\mathbf{u}^{k+1}\right\|_H^2, \quad (49)$$

that is, the sequence generated by N-PDHG1 is contractive that thus N-PDHG1 converges globally. The last inequality together with the analysis of P-rALM indicates that N-PDHG1 with a relaxation step also converges, and the sublinear convergence rate of N-PDHG1 is similar to the proof of P-rALM. Note that convergence of N-PDHG1 does not need the strongly convexity of θ_1 and allows more flexibility on choosing the proximal parameter r.

Finally, it is not difficulty from the first-order optimality conditions of the involved subproblems in N-PDHG2 that

$$\mathbf{u}^{k+1}\in\mathcal{U},\ \theta(\mathbf{u})-\theta(\mathbf{u}^{k+1})+\left\langle \mathbf{u}-\mathbf{u}^{k+1}, M\mathbf{u}\right\rangle \geq \left\langle \mathbf{u}-\mathbf{u}^{k+1}, \widetilde{H}(\mathbf{u}^k-\mathbf{u}^{k+1})\right\rangle$$

for any $\mathbf{u}\in\mathcal{U}$, where

$$\widetilde{H}=\begin{bmatrix} r\mathbf{I} & A^{\mathsf{T}} \\ A & \frac{1}{r}AA^{\mathsf{T}}+Q \end{bmatrix}$$

and $\widetilde{H}$ is positive definite for any $r>0$ and $Q\succ\mathbf{0}$. So, N-PDHG2 also converges globally with a sublinear convergence rate. This matrix $\widetilde{H}$ is what we discussed in Section 1 and could reduce to that in [22] with $Q=\delta\mathbf{I}$ for any $\delta>0$.

5.2 Linear convergence under strongly convexity assumption

The linear convergence rate of N-PDHG1 will be investigated in this subsection under the following assumptions:

(a1) The matrix A is full row rank and $\mathcal{X} = \mathcal{R}^n$;

(a2) The function θ_1 is strongly convex with modulus $\nu > 0$ and $\nabla\theta_1$ is Lipschitz continuous with constant $L > 0$.

From the second part of (a2) and the first-order optimality condition of $\mathbf{x}^{k+1}$-subproblem in N-PDHG1, we have

$$-\nabla\theta_1(\mathbf{x}^{k+1}) = -A^\mathsf{T}\mathbf{y}^k + \Big(rA^\mathsf{T}A + Q\Big)(\mathbf{x}^{k+1} - \mathbf{x}^k). \tag{50}$$

Together with this equation and the first part of (a2), it holds

$$\begin{gathered}
\theta_1(\mathbf{x}) - \theta_1(\mathbf{x}^{k+1}) \geq \big\langle \mathbf{x} - \mathbf{x}^{k+1}, \nabla\theta_1(\mathbf{x}^{k+1})\big\rangle + \frac{\nu}{2}\big\|\mathbf{x} - \mathbf{x}^{k+1}\big\|^2 \Rightarrow \\
\theta_1(\mathbf{x}) - \theta_1(\mathbf{x}^{k+1}) + \big\langle \mathbf{x} - \mathbf{x}^{k+1}, -A^\mathsf{T}\mathbf{y}^{k+1}\big\rangle \geq \frac{\nu}{2}\big\|\mathbf{x} - \mathbf{x}^{k+1}\big\|^2 + \\
\big\langle \mathbf{x} - \mathbf{x}^{k+1}, \big(rA^\mathsf{T}A + Q\big)(\mathbf{x}^k - \mathbf{x}^{k+1}) + A^\mathsf{T}(\mathbf{y}^k - \mathbf{y}^{k+1})\big\rangle,
\end{gathered}$$

which implies that the extra term $\frac{\nu}{2}\big\|\mathbf{x} - \mathbf{x}^{k+1}\big\|^2$ will be added to the right-hand-side of (42) and finally gives

$$\big\|\mathbf{u}^* - \mathbf{u}^{k+1}\big\|_H^2 \leq \big\|\mathbf{u}^* - \mathbf{u}^k\big\|_H^2 - \big\|\mathbf{u}^k - \mathbf{u}^{k+1}\big\|_H^2 - \nu\big\|\mathbf{x}^* - \mathbf{x}^{k+1}\big\|^2. \tag{51}$$

Note that the equation (50) can be equivalently rewritten as

$$A^\mathsf{T}\mathbf{y}^{k+1} = \nabla\theta_1(\mathbf{x}^{k+1}) + A^\mathsf{T}(\mathbf{y}^{k+1} - \mathbf{y}^k) + \big(rA^\mathsf{T}A + Q\big)(\mathbf{x}^{k+1} - \mathbf{x}^k). \tag{52}$$

Besides, the solution $(\mathbf{x}^*; \mathbf{y}^*)$ satisfies

$$\nabla\theta_1(\mathbf{x}^*) = A^\mathsf{T}\mathbf{y}^*. \tag{53}$$

Combining the equations (52) and (53) together with (a1)-(a2) is to obtain

$$\begin{aligned}
& \sigma_A\big\|\mathbf{y}^{k+1} - \mathbf{y}^*\big\|^2 \leq \big\|A^\mathsf{T}(\mathbf{y}^{k+1} - \mathbf{y}^*)\big\|^2 \\
= \;& \big\|\nabla\theta_1(\mathbf{x}^{k+1}) - \nabla\theta_1(\mathbf{x}^*) + A^\mathsf{T}(\mathbf{y}^{k+1} - \mathbf{y}^k) + \big(rA^\mathsf{T}A + Q\big)(\mathbf{x}^{k+1} - \mathbf{x}^k)\big\|^2 \\
\leq \;& 3\Big\{\big\|\nabla\theta_1(\mathbf{x}^{k+1}) - \nabla\theta_1(\mathbf{x}^*)\big\|^2 + \big\|A^\mathsf{T}(\mathbf{y}^{k+1} - \mathbf{y}^k)\big\|^2 + \big\|\big(rA^\mathsf{T}A + Q\big)(\mathbf{x}^{k+1} - \mathbf{x}^k)\big\|^2\Big\} \\
\leq \;& 3\Big\{L^2\big\|\mathbf{x}^{k+1} - \mathbf{x}^*\big\|^2 + \|A\|^2\big\|\mathbf{y}^{k+1} - \mathbf{y}^k\big\|^2 + \big\|\big(rA^\mathsf{T}A + Q\big)(\mathbf{x}^{k+1} - \mathbf{x}^k)\big\|^2\Big\},
\end{aligned}$$

where $\sigma_A > 0$ denotes the smallest eigenvalue of AA^T due to (a1). So, we have

$$\begin{aligned}
\big\|\mathbf{u}^* - \mathbf{u}^{k+1}\big\|_H^2 \;\leq\;& \|H\|\Big\{\big\|\mathbf{x}^* - \mathbf{x}^{k+1}\big\|^2 + \big\|\mathbf{y}^* - \mathbf{y}^{k+1}\big\|^2\Big\} \\
\leq\;& \|H\|\Big\{(1 + 3L^2\sigma_A^{-1})\big\|\mathbf{x}^* - \mathbf{x}^{k+1}\big\|^2 + 3\sigma_A^{-1}\|A\|^2\big\|\mathbf{y}^{k+1} - \mathbf{y}^k\big\|^2 \\
& + 3\sigma_A^{-1}\big\|\big(rA^\mathsf{T}A + Q\big)(\mathbf{x}^{k+1} - \mathbf{x}^k)\big\|^2\Big\}.
\end{aligned} \tag{54}$$

By the structure of H and the Young's inequality, it follows that

$$
\begin{aligned}
& \big\|\mathbf{u}^k-\mathbf{u}^{k+1}\big\|_H^2 \\
= \ & \big\|\big(rA^\mathsf{T}A+Q\big)(\mathbf{x}^{k+1}-\mathbf{x}^k)\big\|^2+\frac{1}{r}\big\|\mathbf{y}^{k+1}-\mathbf{y}^k\big\|^2+2\big\langle\mathbf{x}^{k+1}-\mathbf{x}^k,A^\mathsf{T}(\mathbf{y}^{k+1}-\mathbf{y}^k)\big\rangle \\
\geq \ & \big\|\big(rA^\mathsf{T}A+Q\big)(\mathbf{x}^{k+1}-\mathbf{x}^k)\big\|^2+\frac{1}{r}\big\|\mathbf{y}^{k+1}-\mathbf{y}^k\big\|^2 \\
& -\Big\{\delta_0\big\|\mathbf{x}^{k+1}-\mathbf{x}^k\big\|^2+\frac{1}{\delta_0}\|A^\mathsf{T}A\|\big\|\mathbf{y}^{k+1}-\mathbf{y}^k\big\|^2\Big\}, \\
\geq \ & \big\|\big(rA^\mathsf{T}A+Q\big)(\mathbf{x}^{k+1}-\mathbf{x}^k)\big\|^2-\delta_0\big\|\mathbf{x}^{k+1}-\mathbf{x}^k\big\|^2+\big(\frac{1}{r}-\frac{\|A^\mathsf{T}A\|}{\delta_0}\big)\big\|\mathbf{y}^{k+1}-\mathbf{y}^k\big\|^2, \quad (55)
\end{aligned}
$$

where $\delta_0\in(r\|A^\mathsf{T}A\|,\|rA^\mathsf{T}A+Q\|^2)$ exists for proper choices of r and Q. Now, Let

$$
\delta^k=\min\left\{\frac{\nu}{(1+3L^2\sigma_A^{-1})\|H\|},\ \frac{\delta_0-r\|A^\mathsf{T}A\|}{3r\delta_0\sigma_A^{-1}\|A\|^2\|H\|},\ \frac{\|rA^\mathsf{T}A+Q\|^2-\delta_0}{3\sigma_A^{-1}\|H\|\,\|(rA^\mathsf{T}A+Q)\,(\mathbf{x}^{k+1}-\mathbf{x}^k)\|^2}\right\}.
$$

Then, combining the above inequalities (51) and (54)-(55), we can deduce

$$
\begin{aligned}
& (1+\delta^k)\big\|\mathbf{u}^*-\mathbf{u}^{k+1}\big\|_H^2-\big\|\mathbf{u}^*-\mathbf{u}^k\big\|_H^2 \\
\leq \ & \delta^k\big\|\mathbf{u}^*-\mathbf{u}^{k+1}\big\|_H^2-\big\|\mathbf{u}^k-\mathbf{u}^{k+1}\big\|_H^2-\nu\big\|\mathbf{x}^*-\mathbf{x}^{k+1}\big\|^2 \\
\leq \ & \Big\{\delta^k(1+3L^2\sigma_A^{-1})\|H\|-\nu\Big\}\big\|\mathbf{x}^*-\mathbf{x}^{k+1}\big\|^2+\Big\{3\delta^k\sigma_A^{-1}\|A\|^2\|H\|-\frac{1}{r}+\frac{\|A^\mathsf{T}A\|}{\delta_0}\Big\}\big\|\mathbf{y}^{k+1}-\mathbf{y}^k\big\|^2 \\
& +\big(3\delta^k\sigma_A^{-1}\|H\|-1\big)\big\|\big(rA^\mathsf{T}A+Q\big)(\mathbf{x}^{k+1}-\mathbf{x}^k)\big\|^2+\delta_0\big\|\mathbf{x}^{k+1}-\mathbf{x}^k\big\|^2. \quad (56)
\end{aligned}
$$

Observing from the definition of δ^k, it holds

$$
\begin{cases}
\delta^k(1+3L^2\sigma_A^{-1})\|H\|-\nu\leq 0, \\
3\delta^k\sigma_A^{-1}\|A\|^2\|H\|-\frac{1}{r}+\frac{\|A^\mathsf{T}A\|}{\delta_0}\leq 0, \\
\big(3\delta^k\sigma_A^{-1}\|H\|-1\big)\ \big\|\big(rA^\mathsf{T}A+Q\big)\ (\mathbf{x}^{k+1}-\mathbf{x}^k)\big\|^2+\delta_0\ \big\|\mathbf{x}^{k+1}-\mathbf{x}^k\big\|^2\leq 0,
\end{cases}
$$

and finally ensures the following Q-linear convergence rate:

$$
\big\|\mathbf{u}^*-\mathbf{u}^{k+1}\big\|_H^2\leq\frac{1}{1+\delta^k}\big\|\mathbf{u}^*-\mathbf{u}^k\big\|_H^2.
$$

The above analysis also indicates that our proposed P-rALM for solving the problem (1) will converge Q-linearly under the similar assumptions that $\theta_1(\mathbf{x})$ is strongly convex, its gradient $\nabla\theta_1$ is Lipschitz continuous, the matrix A has full row rank and $\mathcal{X}=\mathcal{R}^n$.

5.3 Linear convergence under error bound condition

In this section, we use $\partial f(x)$ to denote the sub-differential of the convex function f at x. f is said to be a piecewise linear multifunction if its graph $Gr(f):=\{(x,y)\mid y\in f(x)\}$ is a union of finitely many polyhedra. The previous projection $P_\mathcal{C}(x)$ is nonexpansive, i.e.,

$$
\|P_\mathcal{C}(x)-P_\mathcal{C}(z)\|\leq\|x-z\|,\quad\forall x,z\in\mathcal{R}^n. \quad (57)
$$

For any $H \succ \mathbf{0}$, we define $\text{dist}_H(x, \mathcal{C}) := \min_{z\in\mathcal{C}} \|x - z\|_H$. When $H = \mathbf{I}$, we simply denote it $\text{dist}(x, \mathcal{C})$. For any $\mathbf{u} \in \mathcal{U}$ and $\alpha > 0$, the set-valued mapping $e_{\mathcal{U}}(\mathbf{u}, \alpha)$ is defined as

$$e_{\mathcal{U}}(\mathbf{u}, \alpha) := \begin{pmatrix} e_{\mathcal{X}}(\mathbf{u}, \alpha) := \mathbf{x} - P_{\mathcal{X}}\left[\mathbf{x} - \alpha(\xi_{\mathbf{x}} - A^{\mathsf{T}}\mathbf{y})\right] \\ e_{\mathcal{Y}}(\mathbf{u}, \alpha) := \mathbf{y} - P_{\mathcal{Y}}\left[\mathbf{y} - \alpha(\xi_{\mathbf{y}} + A\mathbf{x})\right] \end{pmatrix}, \tag{58}$$

where $\xi_{\mathbf{x}} \in \partial\theta_1(\mathbf{x}), \xi_{\mathbf{y}} \in \partial\theta_2(\mathbf{y})$. Note that a point

$$\mathbf{u}^* \in \mathcal{U}^* = \{\hat{\mathbf{u}} \in \mathcal{U} \mid \text{dist}\left(\mathbf{0}, e_{\mathcal{U}}(\hat{\mathbf{u}}, \alpha)\right) = 0\}$$

is the solution of (41) if and only if $e_{\mathcal{U}}(\mathbf{u}^*, \alpha) = \mathbf{0}$. Different from the assumptions (a1)-(a2), we next investigate the linear convergence rate of N-PDHG1 under the following error bound condition in terms of the mapping $e_{\mathcal{U}}(\mathbf{u}, 1)$:

(a3) Assume that there exists a constant $\zeta > 0$ such that

$$\text{dist}\left(\mathbf{u}, \mathcal{U}^*\right) \le \zeta\, \text{dist}\left(\mathbf{0}, e_{\mathcal{U}}(\mathbf{u}, 1)\right), \ \forall \mathbf{u} \in \mathcal{U}. \tag{59}$$

The condition (59) is generally weaker than the strong convexity assumption and hence can be satisfied by some problems that have non-strongly convex objective functions. Note that if the sub-differentials $\partial\theta_1(\mathbf{x})$ and $\partial\theta_2(\mathbf{y})$ are piecewise linear multifunctions and the constraint sets $\mathcal{X}, \mathcal{Y}$ are polyhedral, then both $P_{\mathcal{X}}$ and $P_{\mathcal{Y}}$ are piecewise linear multifunctions by [13, Prop. 4.1.4] and hence $e_{\mathcal{U}}(\mathbf{u}, \alpha)$ is also a piecewise linear multifunction. Followed by Robinson's continuity property [35] for polyhedral multifunctions, the assumption (a3) holds automatically. For convenience of the sequel analysis, we denote

$$\mathcal{Q} = \begin{bmatrix} (rA^{\mathsf{T}}A + Q)^{\mathsf{T}}(rA^{\mathsf{T}}A + Q) + A^{\mathsf{T}}A & \mathbf{0} \\ \mathbf{0} & \frac{1}{r}\mathbf{I} + AA^{\mathsf{T}} \end{bmatrix}. \tag{60}$$

It is easy to check that $\mathcal{Q}$ is symmetric positive definite because $\|\mathbf{u}\|_{\mathcal{Q}}^2 > 0$ for any $\mathbf{u} \neq \mathbf{0}$. By equivalent expressions for the first-order optimality conditions (44) and (46) together with the structure of $\mathcal{Q}$, we have the following estimation on the distance of $\mathbf{0}$ to $e_{\mathcal{U}}(\mathbf{u}^{k+1}, 1)$, which follows the similar proof as that in [8, Sec. 2.2].

Lemma 5.2 *Let $\mathcal{Q}$ be given in (60). Then, the iterates generated by N-PDHG1 satisfy*

$$\text{dist}^2\left(\mathbf{0}, e_{\mathcal{U}}(\mathbf{u}^{k+1}, 1)\right) \le 2\left\|\mathbf{u}^k - \mathbf{u}^{k+1}\right\|_{\mathcal{Q}}^2. \tag{61}$$

Proof. The first-order optimality condition in (44) implies

$$\mathbf{x}^{k+1} = P_{\mathcal{X}}\left\{\mathbf{x}^{k+1} - \left[\xi_{\mathbf{x}}^{k+1} - A^{\mathsf{T}}\mathbf{y}^k + \left(rA^{\mathsf{T}}A + Q\right)(\mathbf{x}^{k+1} - \mathbf{x}^k)\right]\right\}.$$

This, together with the definition of $\text{dist}_H(\cdot, \cdot)$ and the property in (57), shows

$$\begin{aligned} &\text{dist}^2\left(\mathbf{0}, e_{\mathcal{X}}(\mathbf{u}^{k+1}, 1)\right) = \text{dist}^2\left(\mathbf{x}^{k+1}, P_{\mathcal{X}}\left\{\mathbf{x}^{k+1} - \left[\xi_{\mathbf{x}}^{k+1} - A^{\mathsf{T}}\mathbf{y}^{k+1}\right]\right\}\right) \\ \le\ & \left\|A^{\mathsf{T}}(\mathbf{y}^k - \mathbf{y}^{k+1}) + \left(rA^{\mathsf{T}}A + Q\right)(\mathbf{x}^k - \mathbf{x}^{k+1})\right\|^2 \\ \le\ & 2\Big(\left\|A^{\mathsf{T}}(\mathbf{y}^k - \mathbf{y}^{k+1})\right\|^2 + \left\|\left(rA^{\mathsf{T}}A + Q\right)(\mathbf{x}^k - \mathbf{x}^{k+1})\right\|^2\Big) = 2\left\|\mathbf{u}^k - \mathbf{u}^{k+1}\right\|_{\mathcal{Q}_1}^2, \end{aligned} \tag{62}$$

where $\mathcal{Q}_1 = \mathrm{diag}\left((rA^{\mathsf{T}}A+Q)^{\mathsf{T}}(rA^{\mathsf{T}}A+Q), AA^{\mathsf{T}}\right)$. Similarly, we have from (46) that

$$\mathbf{y}^{k+1} = P_{\mathcal{Y}}\left\{\mathbf{y}^{k+1} - \left[\xi_{\mathbf{y}}^{k+1} + A(2\mathbf{x}^{k+1}-\mathbf{x}^k) + \frac{1}{r}(\mathbf{y}^{k+1}-\mathbf{y}^k)\right]\right\}$$

and

$$\begin{aligned} &\mathrm{dist}^2\left(\mathbf{0}, e_{\mathcal{Y}}(\mathbf{u}^{k+1},1)\right) = \mathrm{dist}^2\left(\mathbf{y}^{k+1}, P_{\mathcal{Y}}\{\mathbf{y}^{k+1} - [\xi_{\mathbf{y}}^{k+1} + A\mathbf{x}^{k+1}]\}\right) \\ \leq\ & \left\|A(\mathbf{x}^k-\mathbf{x}^{k+1}) + \frac{1}{r}(\mathbf{y}^k-\mathbf{y}^{k+1})\right\|^2 \\ \leq\ & 2\Big(\left\|A(\mathbf{x}^k-\mathbf{x}^{k+1})\right\|^2 + \left\|\frac{1}{r}(\mathbf{y}^k-\mathbf{y}^{k+1})\right\|^2\Big) = 2\left\|\mathbf{u}^k-\mathbf{u}^{k+1}\right\|_{\mathcal{Q}_2}^2, \end{aligned} \tag{63}$$

where $\mathcal{Q}_2 = \mathrm{diag}\left(A^{\mathsf{T}}A, \frac{1}{r}\mathbf{I}\right)$. The above inequalities (62)-(63) immediately gives (61) due to the relation $\mathcal{Q} = \mathcal{Q}_1 + \mathcal{Q}_2$. ■

Based on Lemma 5.2 and the conclusion (49), we next provide a global linear convergence rate of N-PDHG1 with the aid of the notations $\lambda_{\min}(H)$ and $\lambda_{\max}(H)$ which denote the smallest and largest eigenvalue of the positive definite matrix H, respectively.

Theorem 5.1 *Let $\mathcal{Q}$ be given in (60). Then, there exists a constant $\zeta > 0$ such that the iterates generated by N-PDHG1 satisfies*

$$\mathrm{dist}_H^2(\mathbf{u}^{k+1},\mathcal{U}^*) \leq \frac{1}{1+\hat{\zeta}}\,\mathrm{dist}_H^2(\mathbf{u}^k,\mathcal{U}^*), \tag{64}$$

where the constant $\hat{\zeta} = \frac{\lambda_{\min}(H)}{2\zeta^2\lambda_{\max}(\mathcal{Q})\lambda_{\max}(H)} > 0$.

Proof Because $\mathcal{U}^*$ is a closed convex set, there exists a $\mathbf{u}_k^* \in \mathcal{U}^*$ satisfying

$$\mathrm{dist}_H(\mathbf{u}^k,\mathcal{U}^*) = \left\|\mathbf{u}^k - \mathbf{u}_k^*\right\|_H. \tag{65}$$

By the condition (59) and Lemma 5.2 there exists a constant $\zeta > 0$ such that

$$\mathrm{dist}^2\left(\mathbf{u}^{k+1},\mathcal{U}^*\right) \leq 2\zeta^2\left\|\mathbf{u}^k-\mathbf{u}^{k+1}\right\|_{\mathcal{Q}}^2 \leq \frac{2\zeta^2\lambda_{\max}(\mathcal{Q})}{\lambda_{\min}(H)}\left\|\mathbf{u}^k-\mathbf{u}^{k+1}\right\|_H^2. \tag{66}$$

Note by the definition of $\mathrm{dist}_H(\cdot,\cdot)$, we have

$$\frac{1}{\lambda_{\max}(H)}\,\mathrm{dist}_H^2\left(\mathbf{u}^{k+1},\mathcal{U}^*\right) \leq \mathrm{dist}^2\left(\mathbf{u}^{k+1},\mathcal{U}^*\right). \tag{67}$$

Combine (66)-(67) and (49) to have

$$\begin{aligned} &\mathrm{dist}_H^2(\mathbf{u}^{k+1},\mathcal{U}^*) \leq \left\|\mathbf{u}^{k+1}-\mathbf{u}_k^*\right\|_H^2 \\ \leq\ & \left\|\mathbf{u}^k-\mathbf{u}_k^*\right\|_H^2 - \left\|\mathbf{u}^k-\mathbf{u}^{k+1}\right\|_H^2 \\ \leq\ & \mathrm{dist}_H^2(\mathbf{u}^k,\mathcal{U}^*) - \frac{\lambda_{\min}(H)}{2\zeta^2\lambda_{\max}(\mathcal{Q})\lambda_{\max}(H)}\,\mathrm{dist}_H^2\left(\mathbf{u}^{k+1},\mathcal{U}^*\right). \end{aligned}$$

Rearranging the above inequality is to confirm (64). ■

Corollary 5.1 *Let $\hat{\zeta} > 0$ be given in Theorem 5.1 and the sequence $\{\mathbf{u}^k\}$ be generated by N-PDHG1. Then, there exists a point $\mathbf{u}^\infty \in \mathcal{U}^*$ such that*

$$\|\mathbf{u}^k - \mathbf{u}^\infty\|_H \le C\epsilon^k, \tag{68}$$

where

$$C = \frac{2\,\mathrm{dist}_H(\mathbf{u}^0, \mathcal{U}^*)}{1-\epsilon} > 0 \quad \textit{and} \quad \epsilon = \frac{1}{\sqrt{1+\hat{\zeta}}} \in (0,1).$$

Proof Let $\mathbf{u}^* \in \mathcal{U}^*$ such that (65) holds and let

$$\mathbf{u}^{k+1} = \mathbf{u}^k + \mathbf{d}^k. \tag{69}$$

Then, it follows from (49) that $\|\mathbf{u}^{k+1} - \mathbf{u}^*\|_H \le \|\mathbf{u}^k - \mathbf{u}^*\|_H$ which further implies

$$\begin{aligned} \|\mathbf{d}^k\|_H &= \|\mathbf{u}^{k+1} - \mathbf{u}^k\|_H \le \|\mathbf{u}^{k+1} - \mathbf{u}^*\|_H + \|\mathbf{u}^k - \mathbf{u}^*\|_H \\ &\le 2\|\mathbf{u}^k - \mathbf{u}^*\|_H = 2\,\mathrm{dist}_H(\mathbf{u}^k, \mathcal{U}^*) \\ &\le 2\epsilon^k\,\mathrm{dist}_H\big(\mathbf{u}^0, \mathcal{U}^*\big), \end{aligned} \tag{70}$$

where the final inequality follows from (64). Because the sequence $\{\mathbf{u}^k\}$ generated by N-PDHG1 converges to a $\mathbf{u}^\infty \in \mathcal{U}^*$, we have by (69) that $\mathbf{u}^\infty = \mathbf{u}^k + \sum_{j=k}^\infty \mathbf{d}^j$, which by (70) indicates

$$\begin{aligned} \|\mathbf{u}^k - \mathbf{u}^\infty\|_H &\le \sum_{j=k}^\infty \|\mathbf{d}^j\|_H \le 2\,\mathrm{dist}_H(\mathbf{u}^0, \mathcal{U}^*)\sum_{j=k}^\infty \epsilon^j \\ &= 2\,\mathrm{dist}_H(\mathbf{u}^0, \mathcal{U}^*)\epsilon^k \sum_{j=0}^\infty \epsilon^j \le \epsilon^k\Big[2\,\mathrm{dist}_H(\mathbf{u}^0, \mathcal{U}^*)\frac{1}{1-\epsilon}\Big]. \end{aligned}$$

So, the inequality (68) holds, that is, $\mathbf{u}^k$ converges $\mathbf{u}^\infty$ R-linearly. ■

Remark 5.1 *Consider the following general saddle-point problem*

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \Phi(\mathbf{x}, \mathbf{y}) := f(\mathbf{x}) + \theta_1(\mathbf{x}) - \mathbf{y}^\mathsf{T} A\mathbf{x} - \theta_2(\mathbf{y}),$$

or, equivalently, the composite problem $\min_{\mathbf{x}\in\mathcal{X}} \big\{f(\mathbf{x}) + \theta_1(\mathbf{x}) + \theta_2^*(-A\mathbf{x})\big\}$, where $f(\mathbf{x}) : \mathcal{X} \to \mathcal{R}$ is a smooth convex function and its gradient is Lipschitz continuous with constant L_f, and the remaining notations have the same meanings as before. For this problem, similar to the previous case 2 in Section 2.3 we can develop the following iterative scheme

$$\begin{cases} \mathbf{x}^{k+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \theta_1(\mathbf{x}) + \langle \nabla f(\mathbf{x}^k) - A^\mathsf{T}\mathbf{y}^k, \mathbf{x}\rangle + \frac{1}{2}\left\|\mathbf{x} - \mathbf{x}^k\right\|^2_{rA^\mathsf{T}A+Q}, \\ \mathbf{y}^{k+1} = \arg\max_{\mathbf{y}\in\mathcal{Y}} \Phi(2\mathbf{x}^{k+1} - \mathbf{x}^k, \mathbf{y}) - \frac{1}{2r}\|\mathbf{y} - \mathbf{y}^k\|^2, \end{cases}$$

Its global convergence and linear convergence rate can be also established by the above analysis.

References

[1] A. Argyriou, T. Evgeniou, M. Pontil, Multi-task feature learning, Adv. Neural Inform. Process. Syst. 19 (2007), pp. 41-48.

[2] S. Banert, M. Upadhyaya, P. Giselsson, The Chambolle-Pock method converges weakly with $\theta > 1/2$ and $\tau\sigma\|L\|^2 < 4/(1+2\theta)$, arXiv:2309.03998v1, (2023).

[3] J. Bai, J. Li, F. Xu, H. Zhang, Generalized symmetric ADMM for separable convex optimization, Comput. Optim. Appl. 70 (2018), pp. 129-170.

[4] J. Bai, K. Guo, X. Chang, A family of multi-parameterized proximal point algorithms, IEEE Access, 7 (2019), pp. 164021-164028.

[5] J. Bai, J. Li, Z. Wu, Several variants of the primal-dual hybrid gradient algorithm with applications, Numer. Math. Theor. Meth. Appl. 13 (2020), pp. 176-199.

[6] J. Bai, W. Hager, H. Zhang, An inexact accelerated stochastic ADMM for separable convex optimization, Comput. Optim. Appl. 81 (2022), pp. 479-518.

[7] J. Bai, Y. Ma, H. Sun, M. Zhang, Iteration complexity analysis of a partial LQP-based alternating direction method of multipliers, Appl. Numer. Math. 165 (2021), pp. 500-518.

[8] J. Bai, X. Chang, J. Li, F. Xu, Convergence revisit on generalized symmetric ADMM, Optimization, 70 (2021), pp. 149-168.

[9] S. Brunton, J. Nathan Kutz, Machine Learning, Dynamical Systems, and Control, Cambridge University Press, Cambridge, 2019.

[10] J. Cui, X. Yan, X. Pu, et al., Aero-engine fault diagnosis based on dynamic PCA and improved SVM, J. Vib. Meas. Diano. 35 (2015), pp. 94-99.

[11] A. Chambolle, T. Pock, A first-order primal-dual algorithms for convex problem with applications to imaging, J. Math. Imaging Vison, 40 (2011), pp. 120-145.

[12] E. Candes, X. Li, Y. Ma, J. Wright, Robust principal component analysis? J. ACM, 58 (2011), pp. 1-37.

[13] F. Facchinei, J. Pang, Finite-Dimensional Variational Inequalities and Complementarity Problems, Springer-Verlag, Berlin, 2003.

[14] G. Gu, B. He, X. Yuan, Customized proximal point algorithms for linearly constrained convex minimization and saddle-point problems: A unified approach, Comput. Optim. Appl. 59 (2014), pp. 135-161.

[15] Y. Hao, J. Sun, G. Yang, J. Bai, The application of support vector machines to gas turbine performance diagnosis, Chinese J. Aeronaut. 18 (2005), pp. 15-19.

[16] M. Hestenes, Multiplier and gradient methods, J. Optim. Theory Appl. 4 (1969), pp. 303-320.

[17] B. He, X. Yuan, On the $O(1/n)$ convergence rate of the Douglas-Rachford alternating direction method, SIAM J. Numer. Anal. 50 (2012), pp. 700-709.

[18] B. He, X. Yuan, W. Zhang, A customized proximal point algorithm for convex minimization with linear constraints, Comput. Optim. Appl. 56 (2013), pp. 559-572.

[19] B. He, X. Yuan, A class of ADMM-based algorithms for three-block separable convex programming, Comput. Optim. Appl. 70 (2018), pp. 791-826.

[20] B. He, Y. You, X. Yuan, On the convergence of primal-dual hybrid gradient algorithms, SIAM J. Imaging Sci. 7 (2014), pp. 2526-2537.

[21] B. He, On the convergence properties of alternating direction method of multipliers, Numer. Math.: A Journal of Chinese Universities (Chinese Series), 39 (2017), pp. 81-96.

[22] B. He, X. Yuan, Balanced augmented Lagrangian method for convex programming, (2021) arXiv:2108.08554v1.

[23] B. He, S. Xu, J. Yuan, Indefinite linearized augmented Lagrangian method for convex programming with linear inequality constraints, (2021) arXiv:2105.02425v1.

[24] H. He, J. Desai, K. Wang, A primal-dual prediction-correction algorithm for saddle point optimization, J. Global Optim. 66 (2016), pp. 573-583.

[25] B. He, F. Ma, S. Xu, X. Yuan, A generalized primal-dual algorithm with improved convergence condition for saddle point problems, SIAM J. Imaging Sci. 15 (2022), pp. 1157-1183.

[26] F. Jiang, Z. Zhang, H. He, Solving saddle point problems: a landscape of primal-dual algorithm with larger stepsizes, J. Global Optim. 85 (2023), pp. 821-846.

[27] F. Jiang, Z. Wu, X. Cai, H. Zhang, A first-order inexact primal-dual algorithm for a class of convex-concave saddle point problems, Numer. Algor. 88 (2021), pp. 1109-1136.

[28] L. Li, Selected Applications of Convex Optimization, Tsinghua University Press, Beijing, (2015), pp. 17-18.

[29] Q. Li, Y. Xu, N. Zhang, Two-step fixed-point proximity algorithms for multi-block separable convex problems, J. Sci. Comput. 70 (2017), pp. 1204-1228.

[30] Z. Liu, J. Li, G. Li, et al., A new model for sparse and low rank matrix decomposition, J. Appl. Anal. Comput. 7 (2017), pp. 600-616.

[31] F. Ma, M. Ni, A class of customized proximal point algorithms for linearly constrained convex optimization, Comput. Appl. Math. 37 (2018), pp. 896-911.

[32] S. Osher, H. Heaton, S. Fung. A HamiltonCJacobi-based proximal operator, PANS, 120 (2023), pp. e2220469120.

[33] M. Powell, A method for nonlinear constraints in minimization problems, In Optimization edited by R. Fletcher, pp. 283-298, Academic Press, New York, 1969.

[34] C. Papadimitriou, P. Raghavan, H. Tamaki, S. Vempala, Latent semantic indexing, a probabilistic analysis, J. Comput. Syst. Sci. 61(2000), pp. 217-235.

[35] S. Robinson, Some continuity properties of polyhedral multifunctions, Math. Program. Stud. 14 (1981), pp. 206-241.

[36] Y. Shen, Y. Zuo, A. Yu, A partially proximal S-ADMM for separable convex optimization with linear constraints, Appl. Numer. Math. 160 (2021), pp. 65-83.

[37] M. Tao, X. Yuan, Recovering low-rank and sparse components of matrices from incomplete and noisy observations, SIAM J. Optim. 21 (2011), pp. 57-81.

[38] S. Xu, A dual-primal balanced augmented Lagrangian method for linearly constrained convex programming, J. Appl. Math. Comput. 69 (2023), pp. 1015–1035.

[39] Y. Zhu, J. Wu, G. Yu, A fast proximal point algorithm for l_1-minimization problem in compressed sensing, Appl. Math. Comput. 270 (2015), pp. 777-784.

[40] X. Zhang, Bregman Divergence and Mirror Descent, Lecture Notes, (2013) http://users.cecs.anu.edu.au/ xzhang/teaching/bregman.pdf.

[41] M. Zhu, T. F. Chan, An efficient primal-dual hybrid gradient algorithm for total variation image restoration, CAM Report 08-34, UCLA, Los Angeles, CA, 2008.