Convergence analysis of the Peaceman-Rachford splitting method for nonsmooth convex optimization

Deren Han * **and Xiaoming Yuan**†

July 11, 2013

Abstract

In this paper, we focus on the convergence analysis for the application of the Peaceman-Rachford splitting method to a convex minimization model whose objective function is the sum of a smooth and nonsmooth convex functions. The sublinear convergence rate in term of the worst-case $O(1/t)$ iteration complexity is established if the gradient of the smooth objective function is assumed to be Lipschitz continuous; and the linear convergence rate is derived if this smooth function is further assumed to be strongly convex. A key technique to obtain these convergence rate results is that we use the primal-dual gap, rather than the objective function value to measure the accuracy of iterates. We also propose a modified Peaceman-Rachford splitting method for this convex minimization model which does not require to know the involved Lipschitz constant. Convergence analysis is conducted for this modified Peaceman-Rachford splitting method.

Key words: Peaceman-Rachford splitting method, convex optimization, nonsmooth, iteration complexity, linear convergence rate.

1 Introduction

In this paper, we consider the convex minimization model whose objective function is the sum of two functions:

$$\min_{x\in\mathcal{R}^n} \quad F(x) \equiv f(x) + g(x), \tag{1.1}$$

*School of Mathematical Sciences, Key Laboratory for NSLSCS of Jiangsu Province, Nanjing Normal University, Nanjing 210023, P.R. China. Email: handeren@njnu.edu.cn. This author was supported by the NSFC grants 11071122 and 11171159.

†Department of Mathematics, Institute of Computational and Theoretical Studies, Hong Kong Baptist University, Hong Kong, P.R. China. Email: xmyuan@hkbu.edu.hk. This author was supported by the General Research Fund from Hong Kong Research Grants Council: 203613.

where f, g: $\mathcal{R}^n \to \mathcal{R}$ are convex functions. Our discussion focuses on the case where one of the functions in (1.1) is nonsmooth (e.g., f) while the other (i.e., g) is smooth and its gradient is Lipschitz continuous. A typical application of (1.1) is solving some ill-posed inverse problems with regularization terms, where one of the functions in (1.1) represents the data fidelity term while the other is a regularization term.

A particularly interesting case of (1.1) with wide applications is that the nonsmooth function f is "simple" in the sense that its proximal problem has a closed-form solution (equivalently, the resolvent operator of f's subdifferential has a closed-form representation). That is, the following optimization problem

$$\min_{x\in\mathcal{R}^n}\left\{f(x)+\frac{1}{2\mu}\|x-z\|_2^2\right\} \tag{1.2}$$

has a closed-form solution for any $z \in \mathcal{R}^n$ and $\mu > 0$. Such a case is the popular $l_1 - l_2$ model:

$$\min_{x\in\mathcal{R}^n} \quad \|x\|_1+\frac{1}{2\mu}\|Ax-b\|_2^2, \tag{1.3}$$

where $x \in \mathcal{R}^n$, $A \in \mathcal{R}^{m\times n}$, $b \in \mathcal{R}^m$ and $\|x\|_1 := \sum_{i=1}^n |x_i|$. When $m \ll n$, the model (1.3) can be explained as pursuing a sparse solution of the under-determined system of linear equations $Ax = b$ and it captures many applications arising in compressive sensing, image processing, statistical learning, computer vision, etc. Note when $f(x) = \|x\|_1$, the closed-form solution of the proximal problem (1.2) is given by the soft-shrinkage operator (see e.g. [4]):

$$x = \text{shrinkage}(z,\mu) = \text{sign}(z)\cdot\max\{0,|z|-\mu\}.$$

To solve such a particular case of (1.1) with a "simple" $f(x)$, an iterative scheme effective for using the simplicity of $f(x)$ is to solve the following subproblem iteratively:

$$\min_{x\in\mathcal{R}^n} \quad f(x)+g_L^k(x), \tag{1.4}$$

where

$$g_L^k(x) := g(x^k)+\langle\nabla g(x^k), x-x^k\rangle+\frac{1}{2\mu}\|x-x^k\|_2^2$$

is a linear approximation of $g(x)$ at $x = x^k$ with a quadratic proximal term, ∇g denotes the gradient of g and $\mu > 0$ is the linearization parameter. Obviously, subject to a difference of constant in the objective, solving (1.4) amounts to

$$\min_{x\in\mathcal{R}^n}\left\{f(x)+\frac{1}{2\mu}\|x-(x^k-\mu\nabla g(x^k))\|_2^2\right\}, \tag{1.5}$$

which is exactly a proximal problem in form of (1.2) with the proximal center $x^k - \mu\nabla g(x^k)$. More specifically, the iterative scheme of ISTA is

$$x^{k+1} = (I+\mu\partial f)^{-1}(x^k-\mu\nabla g(x^k)), \tag{1.6}$$

where $\partial(\cdot)$ denotes the subdifferential of a nonsmooth convex function. This is the essential idea of the iterative shrinkage-thresholding algorithm (ISTA) in [1]. The simple algorithmic framework of ISTA not only makes its numerical implementation extremely easy, but also simplifies the theoretical analysis of its convergence rate. In [1], ISTA was shown to have a worst-case $O(1/t)$ convergence rate in term of the iteration complexity[1]. Moreover, a faster version of ISTA (FISTA) accelerated by Nesterov's schemes in [18] with a worst-case $O(1/t^2)$ iteration complexity was also proposed in [1]. Note this worst-case $O(1/t)$ or $O(1/t^2)$ iteration complexity amounts to a sublinear convergence rate, e.g., see [1]. In fact, if the smooth function g is assumed to be strongly convex, we can show a linear convergence rate of ISTA. See the appendix.[2]

The efficiency of ISTA has also inspired other impressive work in the literature. One example is the fast alternating linearization method proposed in [10]. To understand this work, we follow the procedure in [10] and consider a reformulation of (1.1):

$$\min_{x\in\mathcal{R}^n, y\in\mathcal{R}^n} \{f(x)+g(y):\ x=y\}, \tag{1.7}$$

where the auxiliary variable y is to split the variable in the objective function. The Lagrangian function for (1.7) is

$$\mathcal{L}(x,y,\lambda) := f(x)+g(y)-\langle\lambda, x-y\rangle,$$

where λ is the Lagrange multiplier, and solving (1.7) amounts to finding a saddle point of $\mathcal{L}$. That is, finding $(\hat{x},\hat{y},\hat{\lambda})\in\mathcal{R}^n\times\mathcal{R}^n\times\mathcal{R}^n$ such that the following inequalities hold:

$$\mathcal{L}(\hat{x},\hat{y},\lambda)\le\mathcal{L}(\hat{x},\hat{y},\hat{\lambda})\le\mathcal{L}(x,y,\hat{\lambda}),\ \ \forall(x,y,\lambda)\in\mathcal{R}^n\times\mathcal{R}^n\times\mathcal{R}^n. \tag{1.8}$$

Let the set of saddle points be denoted by S. Throughout this paper, S is assumed to be nonempty. Moreover, let

$$\mathcal{L}_\mu(x,y,\lambda) := f(x)+g(y)-\langle\lambda,x-y\rangle+\frac{1}{2\mu}\|x-y\|^2, \tag{1.9}$$

be the augmented Lagrange function of (1.7) with $\mu>0$ being a penalty parameter. The constrained reformulation (1.7) falls into the applicable scope of the classical augmented Lagrangian method (ALM) in [15, 22], and makes it possible to develop splitting versions of ALM for (1.1) and borrow analytic framework in ALM literature to conduct theoretical analysis. In our analysis, we will stick to this reformulation of (1.1). Let us first recall the alternating linearization method in [16] for solving the case of (1.1) where both f and g are smooth:

$$\begin{cases} x^{k+1} = \arg\min\limits_{x\in\mathcal{R}^n} Q_\mu^g(x,y^k), & \text{(1.10a)}\\ y^{k+1} = \arg\min\limits_{y\in\mathcal{R}^n} Q_\mu^f(y,x^{k+1}), & \text{(1.10b)}\end{cases}$$

[1] As the work [18, 19] and many others, a worst-case $O(1/t)$ convergence rate means the accuracy to a solution under certain criteria is of the order $O(1/t)$ after t iterations of an iterative scheme; or equivalently, it requires at most $O(1/\epsilon)$ iterations to achieve an approximate solution with an accuracy of ϵ.

[2] Since the proof framework of the linear convergence rate of ISTA is technically different from our main result in Section 4, we only include it in the appendix.

where

$$Q_\mu^g(u,v) := f(u) + g(v) + \langle \nabla g(v), u - v\rangle + \frac{1}{2\mu}\|u - v\|^2,$$

and

$$Q_\mu^f(u,v) := g(u) + f(v) + \langle \nabla f(v), u - v\rangle + \frac{1}{2\mu}\|u - v\|^2.$$

To extend the scheme (1.10a)-(1.10b) to the case where f is nonsmooth, the authors of [10] proposed the following scheme

$$\begin{cases} y^{k+1} = \arg\min\limits_{y\in\mathcal{R}^n} \mathcal{L}_\mu(x^k, y, \lambda^k), & \text{(1.11a)}\\ \textbf{If } F(y^{k+1}) > \mathcal{L}_\mu(x^k, y^{k+1}, \lambda^k), \textbf{ then } y^{k+1} := x^k & \text{(1.11b)}\\ x^{k+1} = \arg\min\limits_{x\in\mathcal{R}^n} Q_\mu^g(x, y^{k+1}), & \text{(1.11c)}\\ \lambda^{k+1} = \nabla g(y^{k+1}) - \dfrac{1}{\mu}(y^{k+1} - x^{k+1}), & \text{(1.11d)} \end{cases}$$

where (1.11b) is a so-called "skipping step" —- if the objective function value is not reduced sufficiently, then the new iterate y^{k+1} computed by (1.11a) is skipped. Hence the name of alternating linearization method with skipping steps in [10] for the scheme (1.11a)-(1.11d). To see the relationship between the scheme (1.11a)-(1.11d) and the ISTA scheme (1.5), let us derive the first-order optimality condition for (1.11c):

$$-(\nabla g(y^{k+1}) + \frac{1}{\mu}(x^{k+1} - y^{k+1})) \in \partial f(x^{k+1}).$$

This is exactly the scheme of ISTA (1.6) since $y^{k+1} = x^k$. Therefore, when the skipping step (1.11b) takes place, we have $y^{k+1} = x^k$ and the scheme (1.11a)-(1.11d) reduces to ISTA. Alternatively, the scheme (1.11a)-(1.11d) can be understood as a hybrid method: If a sufficient reduction of the objective function value is guaranteed subject to the criterion (1.11b), then the new iterate y^{k+1} is obtained by solving the split augmented Lagrangian subproblem (1.11a); otherwise it just performs ISTA. Some sublinear convergence rates including a worst-case $O(1/t)$ iteration complexity of the scheme (1.11a)-(1.11d) and a worst-case $O(1/t^2)$ iteration complexity of its accelerated version were established in [10].

The iterative scheme of ISTA (1.5) can be explained also as an application of the forward-backward splitting method —- e.g. see [6, 20, 23]. In this paper, we focus on the application of another operator splitting method —- the Peaceman-Rachford splitting method (PRSM) which was originally proposed in [17, 21] for finding a root of the sum of two maximal monotone set-valued mappings and received intensive attention in PDE literature, to the model (1.1). Some rationale of applying PRSM can be found in, e.g., [8] where it was stated that "PRSM is always faster than the Douglas Rachford splitting method (DRSM) in [5, 17]"; and also [2, 9] where the efficiency of PRSM is verified numerically. Following the analysis in [9] and using the notation

in [10], the iterative scheme of PRSM for solving (1.7) is

$$\begin{cases} x^{k+1} = \arg\min\limits_{x\in\mathcal{R}^n} \mathcal{L}_\mu(x, y^k, \lambda^k), & \text{(1.12a)} \\ \lambda^{k+\frac{1}{2}} = \lambda^k - \dfrac{1}{\mu}(x^{k+1} - y^k), & \text{(1.12b)} \\ y^{k+1} = \arg\min\limits_{y\in\mathcal{R}^n} \mathcal{L}_\mu(x^{k+1}, y, \lambda^{k+\frac{1}{2}}), & \text{(1.12c)} \\ \lambda^{k+1} = \lambda^{k+\frac{1}{2}} - \dfrac{1}{\mu}(x^{k+1} - y^{k+1}), & \text{(1.12d)} \end{cases}$$

As analyzed in [10], the iterative scheme (1.12a)-(1.12d) reduces to the alternating linearization method [16] when both f and g are smooth.

In this paper, we analyze the convergence of the PRSM scheme (1.12a)-(1.12d) without any skipping step. In Section 3, we establish its sublinear convergence rate in term of the worst-case $O(1/t)$ iteration complexity under the conditions that both f and g are convex, f is nonsmooth, g is smooth, and ∇g is Lipschitz continuous. Then, in Section 4 we show its linear convergence rate when g is further assumed to be strongly convex. It can be easily observed that if the step (1.12b) is removed, then the iterative scheme (1.12a)-(1.12d) reduces to the alternating direction method of multipliers (ADMM) proposed in [7] for solving (1.7), which is a special application of DRSM to the dual of (1.7), as analyzed in [6]. We refer to [12] and [13] for the worst-case $O(1/t)$ iteration complexity of ADMM respectively in ergodic and nonergodic senses in optimization context; and [14] for its extension to DRSM in the more general setting of finding a root of the sum of two maximal monotone mappings. But, as analyzed in [11], the convergence rate analysis of PRSM is usually more complicated than that of DRSM. The reason can be explained that for the convex programming model under consideration, the iterative sequence generated by PRSM might not be strictly contractive with respect to the solution set under consideration while that of DRSM always is. Because of this significant difference, it seems that the techniques used in [11, 12, 13, 14] are not applicable to study the convergence rate of PRSM. Our key technique to be used in Sections 3 and 4 is that we use the primal-dual gap (see definition in (2.4)), rather than the objective function value in existing work [1, 10] to measure the accuracy of iterates to a solution point[3]. The reason is that the PRSM scheme (1.12a)-(1.12d) is not able to ensure that the objective function value is decreasing, while as we will show, it can guarantee that both the primal-dual gap and the distance between iterates and the solution set are decreasing iteratively.

The analysis in Sections 3 and 4 follows the traditional requirement in the literature (see e.g. [1, 10, 19]) that the linearization parameter μ must satisfy $\mu < 1/L$ where L denotes the Lipschitz constant of ∇g. For the case where L is hard to estimate, some backtracking searching strategies in [1] were proposed and computation for these searching strategies is often time-consuming. The numerical results reported in [10] even show that the restriction of μ leads to significantly slower convergence empirically, despite that it ensures the convergence

[3]Recall our analysis is based on the constrained reformulation (1.7), as [10]. Thus, it is not appropriate to measure the optimality only in term of the decrease of the objective function value.

theoretically. Moreover, these numerical results demonstrate that if μ is sufficiently small such that the sequence of objective function value is reduced monotonically (so the analysis of iteration complexity becomes possible in [10]), then the skipping step (1.11b) almost does not occur. The authors of [10] thus suggested to choose the value of μ in a more aggressive way empirically, disobeying this theoretical restriction. These facts inspire us to discuss how to relax the restriction of μ in absence of L for (1.12a)-(1.12d) and allow to choose any positive value for the linearization parameter μ. Consequently, a modified PRSM scheme with a moving proximal center is proposed in Section 5. We will show that all the convergence rate results in Sections 3 and 4 still hold for this modified PRSM scheme. Finally, we make some conclusions in Section 6.

2 Preliminaries

In this section, we summarize some definitions and notations useful for further analysis.

All vectors are column vectors in our disucssion. For any two vectors $x \in \mathcal{R}^n$ and $y \in \mathcal{R}^m$, we simply use $u = (x, y)$ to denote their adjunction, i.e., (x, y) denotes $(x^T, y^T)^T$. We denote by $\langle \cdot, \cdot \rangle$ the inner product, and by $\| \cdot \|$ the associated norm. For any symmetric and positive definite matrix M, we denote by $\|x\|_M := \sqrt{\langle x, Mx \rangle}$ the M-norm. Moreover, we denote by $\lambda_{\max}$ the maximum eigenvalue of a matrix. For a given matrix A, its norm is

$$\|A\| := \sup_{x \neq 0} \left\{ \frac{\|Ax\|}{\|x\|} \right\}.$$

Specially, for a symmetric matrix A, $\|A\|_2$ denotes its spectral norm.

A mapping $F : \mathcal{R}^n \to \mathcal{R}^n$ is Liptschitz continuous if there is a constant $L > 0$, such that

$$\|F(x) - F(x^{'})\| \leq L\|x - x^{'}\|, \ \forall x, \ x^{'} \in \mathcal{R}^n.$$

For any saddle point $(\hat{x}, \hat{y}, \hat{\lambda}) \in S$, the following conditions hold:

$$\hat{\lambda} \in \partial f(\hat{x}), \tag{2.1}$$

$$-\hat{\lambda} = \nabla g(\hat{y}), \tag{2.2}$$

$$\hat{x} - \hat{y} = 0. \tag{2.3}$$

As in [3], we first introduce the partial primal-dual gap defined as

$$\begin{aligned} \mathcal{G}_{\mathcal{B}_1 \times \mathcal{B}_2 \times \mathcal{B}_3}(x, y, \lambda) \ := \ & \max_{\lambda^{'} \in \mathcal{B}_3} \{\langle \lambda^{'}, y - x \rangle + f(x) + g(y)\} \\ & - \min_{(x^{'}, y^{'}) \in \mathcal{B}_1 \times \mathcal{B}_2} \{\langle \lambda, y^{'} - x^{'} \rangle + f(x^{'}) + g(y^{'})\}, \end{aligned} \tag{2.4}$$

where $\mathcal{B}_1 \times \mathcal{B}_2 \times \mathcal{B}_3$ is a subset of $\mathcal{R}^n \times \mathcal{R}^n \times \mathcal{R}^n$. If $\mathcal{B}_1 \times \mathcal{B}_2 \times \mathcal{B}_3$ contains a saddle point $(\hat{x}, \hat{y}, \hat{\lambda})$, then for any $(x, y, \lambda) \in \mathcal{B}_1 \times \mathcal{B}_2 \times \mathcal{B}_3$, we have

$$\begin{aligned} \mathcal{G}_{\mathcal{B}_1 \times \mathcal{B}_2 \times \mathcal{B}_3}(x, y, \lambda) &\geq (\langle \hat{\lambda}, y - x \rangle + f(x) + g(y)) - (\langle \lambda, \hat{y} - \hat{x} \rangle + f(\hat{x}) + g(\hat{y})) \\ &= (f(x) - f(\hat{x}) - \langle \hat{\lambda}, x - \hat{x} \rangle) + (g(y) - g(\hat{y}) + \langle \hat{\lambda}, y - \hat{y} \rangle) \\ &\geq 0, \end{aligned} \tag{2.5}$$

where the equality follows from the fact that $\hat{x} - \hat{y} = 0$, and the last equality follows from $\hat{\lambda} \in \partial f(\hat{x})$, $\hat{\lambda} = -\partial g(\hat{y})$, and the convexity of f and g.

The following lemma gives a sufficient condition for a point to be a saddle point. The assertion is an immediate conclusion of (2.5); we thus omit the proof.

Lemma 2.1. *Let $(\bar{x}, \bar{y}, \bar{\lambda})$ be an arbitrary point in $\mathcal{R}^n \times \mathcal{R}^n \times \mathcal{R}^n$ and let $(\hat{x}, \hat{y}, \hat{\lambda})$ be a saddle point in $\mathcal{B}_1 \times \mathcal{B}_2 \times \mathcal{B}_3$. If*

$$\mathcal{G}_{\mathcal{B}_1 \times \mathcal{B}_2 \times \mathcal{B}_3}(\bar{x}, \bar{y}, \bar{\lambda}) \leq 0, \tag{2.6}$$

then $(\bar{x}, \bar{y}, \bar{\lambda})$ is also a saddle point.

Remark 2.1. *As remarked in [3], in general $\mathcal{G}_{\mathcal{B}_1 \times \mathcal{B}_2 \times \mathcal{B}_3}(x, y, \lambda)$ cannot be used to measure for optimality, since we can not assert that (x, y, λ) is a saddle point even when $\mathcal{G}_{\mathcal{B}_1 \times \mathcal{B}_2 \times \mathcal{B}_3}(x, y, \lambda) = 0$. However, if $\mathcal{G}_{\mathcal{B}_1 \times \mathcal{B}_2 \times \mathcal{B}_3}(x, y, \lambda) = 0$ and (x, y, λ) lies in the interior of $\mathcal{B}_1 \times \mathcal{B}_2 \times \mathcal{B}_3$, then (x, y, λ) is a saddle point. Since the sequence generated by the PRSM scheme (1.12a)-(1.12d) or the modified PRSM scheme (5.1a)-(5.1d) is bounded (see Theorem 3.1 and Lemma 5.1), the primal-dual gap $\mathcal{G}_{\mathcal{B}_1 \times \mathcal{B}_2 \times \mathcal{B}_3}(x, y, \lambda)$ can be used to measure the optimality of iterates.*

Finally, let us define a matrix which will be used often in our analysis:

$$H_\mu = \begin{pmatrix} \frac{1}{\mu} I_{n \times n} & I_{n \times n} \\ I_{n \times n} & \mu I_{n \times n} \end{pmatrix}, \tag{2.7}$$

where μ is the linearization parameter used in (1.12a)-(1.12d). Obviously, H_μ is positive semidefinite under the condition that $\mu > 0$. We will use this matrix to define a matrix norm in our analysis later.

3 Sublinear convergence rate

In this section, we establish the sublinear converegence rate in term of a worst-case $O(1/t)$ iteration complexity for the PRSM scheme (1.12a)-(1.12d).

As we have mentioned, we shall use the primal-dual gap defined in (2.4) to measure the accuracy of an iterate generated by the PRSM scheme (1.12a)-(1.12d) to the set of saddle points

satisfying (2.1)-(2.3); and accordingly we analyze its convergence rate based on the constrained reformulation (1.7).

We first need to find an upper bound for the primal-dual gap for the iterate $(x^{k+1}, y^{k+1}, \lambda^{k+1})$ generated by (1.12a)-(1.12d). To manipulate recursively and thus derive an iteration complexity, this bound is expected to be expressed in quadratic terms and inner product terms involving only the iterates (i.e., the objective function values $f(x^{k+1})$ and $g(y^{k+1})$ should be absent in the bound). This is completed in Lemma 3.1.

Lemma 3.1. *For (1.1), let f and g be convex, f be nonsmooth, g be smooth and ∇g be Lipscitz continuous with a constant L. For any $(x, y, \lambda) \in \mathcal{R}^n \times \mathcal{R}^n \times \mathcal{R}^n$, the sequence $\{(x^k, y^k, \lambda^k)\}$ generated by the PRSM scheme (1.12a)-(1.12d) satisfies*

$$\begin{aligned}&[f(x^{k+1}) + g(y^{k+1}) - \langle \lambda, x^{k+1} - y^{k+1}\rangle] - [f(x) + g(y) - \langle \lambda^{k+1}, x - y\rangle] \\ &\leq \frac{1}{4}\left(\left\| \begin{array}{c} y^k - y \\ \lambda^k - \lambda \end{array} \right\|_{H_\mu}^2 - \left\| \begin{array}{c} y^{k+1} - y \\ \lambda^{k+1} - \lambda \end{array} \right\|_{H_\mu}^2 \right) + \frac{1}{\mu}\langle x^{k+1} - y^{k+1}, y - x\rangle. \end{aligned} \tag{3.1}$$

where H_μ is defined in (2.7).

Proof. It follows from the optimality condition for (1.12a) and the definition of $\lambda^{k+\frac{1}{2}}$ that

$$\lambda^{k+\frac{1}{2}} \in \partial f(x^{k+1}). \tag{3.2}$$

Similarly, from the optimality condition for (1.12c) and the definition of λ^{k+1} that

$$-\lambda^{k+1} = \nabla g(y^{k+1}). \tag{3.3}$$

Then, for any $(x, y, \lambda) \in \mathcal{R}^n \times \mathcal{R}^n \times \mathcal{R}^n$, we have

$$\begin{array}{rcl} f(x) & \geq & f(x^{k+1}) + \langle \lambda^{k+\frac{1}{2}}, x - x^{k+1}\rangle, \\ g(y) & \geq & g(y^{k+1}) - \langle \lambda^{k+1}, y - y^{k+1}\rangle. \end{array} \tag{3.4}$$

Adding both inequalities and rearranging terms, we get

$$\begin{aligned}&[f(x^{k+1}) + g(y^{k+1}) - \langle \lambda, x^{k+1} - y^{k+1}] - [f(x) + g(y) - \langle \lambda^{k+1}, x - y\rangle] \\ &\leq \langle \lambda^{k+1}, y - y^{k+1}\rangle - \langle \lambda^{k+\frac{1}{2}}, x - x^{k+1}\rangle - \langle \lambda, x^{k+1} - y^{k+1}\rangle + \langle \lambda^{k+1}, x - y\rangle \\ &= \langle \lambda^{k+\frac{1}{2}} - \lambda, x^{k+1} - y^{k+1}\rangle + \frac{1}{\mu}\langle y^{k+1} - y, x^{k+1} - y^{k+1}\rangle + \frac{1}{\mu}\langle x^{k+1} - y^{k+1}, y - x\rangle. \end{aligned} \tag{3.5}$$

Recall that

$$\lambda^k - \lambda^{k+1} = \frac{1}{\mu}(2x^{k+1} - y^k - y^{k+1}).$$

It follows from the definition of H_μ and (3.5) that

$$\begin{aligned}&[f(x^{k+1}) + g(y^{k+1}) - \langle \lambda, x^{k+1} - y^{k+1}\rangle] - [f(x) + g(y) - \langle \lambda^{k+1}, x - y\rangle] \\ &\leq \frac{1}{2}\left\langle \begin{array}{c} y^{k+1} - y \\ \lambda^{k+\frac{1}{2}} - \lambda \end{array}, H_\mu \left(\begin{array}{c} y^k - y^{k+1} \\ \lambda^k - \lambda^{k+1} \end{array} \right) \right\rangle + \frac{1}{\mu}\langle x^{k+1} - y^{k+1}, y - x\rangle. \end{aligned} \tag{3.6}$$

Now, let us focus on the first term on the right-hand side of (3.6). We shall show that this crossed term can be expressed as the difference of two quadratic terms. More specifically, it is easy to see that for any $\mu > 0$, H_μ is symmetric and positive semidefinite. Moreover, using the identity

$$\|x - y\|_{H_\mu}^2 = \|x\|_{H_\mu}^2 - \|y\|_{H_\mu}^2 - 2\langle x - y, H_\mu y\rangle, \ \forall x, y,$$

we have

$$\left\| \begin{array}{c} y^{k+1} - y \\ \lambda^{k+1} - \lambda \end{array} \right\|_{H_\mu}^2 = \left\| \begin{array}{c} y^k - y \\ \lambda^k - \lambda \end{array} \right\|_{H_\mu}^2 - \left\| \begin{array}{c} y^{k+1} - y^k \\ \lambda^{k+1} - \lambda^k \end{array} \right\|_{H_\mu}^2 - 2\left\langle \begin{array}{c} y^{k+1} - y \\ \lambda^{k+1} - \lambda \end{array}, H_\mu \left(\begin{array}{c} y^k - y^{k+1} \\ \lambda^k - \lambda^{k+1} \end{array} \right) \right\rangle$$

and

$$\begin{aligned} \left\| \begin{array}{c} y^{k+1} - y^k \\ \lambda^{k+1} - \lambda^k \end{array} \right\|_{H_\mu}^2 = & \left\| \begin{array}{c} y^{k+1} - y^k \\ \lambda^k - \lambda^{k+\frac{1}{2}} \end{array} \right\|_{H_\mu}^2 - \left\| \begin{array}{c} 0 \\ \lambda^{k+1} - \lambda^{k+\frac{1}{2}} \end{array} \right\|_{H_\mu}^2 \\ & -2\left\langle \begin{array}{c} y^{k+1} - y^k \\ \lambda^{k+1} - \lambda^k \end{array}, H_\mu \left(\begin{array}{c} 0 \\ \lambda^{k+1} - \lambda^{k+\frac{1}{2}} \end{array} \right) \right\rangle. \end{aligned}$$

Adding the above two inequalities, we get

$$\begin{aligned} \left\| \begin{array}{c} y^{k+1} - y \\ \lambda^{k+1} - \lambda \end{array} \right\|_{H_\mu}^2 = & \left\| \begin{array}{c} y^k - y \\ \lambda^k - \lambda \end{array} \right\|_{H_\mu}^2 - \left\| \begin{array}{c} y^{k+1} - y^k \\ \lambda^k - \lambda^{k+\frac{1}{2}} \end{array} \right\|_{H_\mu}^2 + \left\| \begin{array}{c} 0 \\ \lambda^{k+1} - \lambda^{k+\frac{1}{2}} \end{array} \right\|_{H_\mu}^2 \\ & -2\left\langle \begin{array}{c} y^{k+1} - y \\ \lambda^{k+\frac{1}{2}} - \lambda \end{array}, H_\mu \left(\begin{array}{c} y^k - y^{k+1} \\ \lambda^k - \lambda^{k+1} \end{array} \right) \right\rangle. \end{aligned}$$

That is,

$$\begin{aligned} \frac{1}{2}\left\langle \begin{array}{c} y^{k+1} - y \\ \lambda^{k+\frac{1}{2}} - \lambda \end{array}, H_\mu \left(\begin{array}{c} y^k - y^{k+1} \\ \lambda^k - \lambda^{k+1} \end{array} \right) \right\rangle = & \frac{1}{4}\left(\left\| \begin{array}{c} y^k - y \\ \lambda^k - \lambda \end{array} \right\|_{H_\mu}^2 - \left\| \begin{array}{c} y^{k+1} - y \\ \lambda^{k+1} - \lambda \end{array} \right\|_{H_\mu}^2 \right) \\ & - \frac{1}{4}\left(\left\| \begin{array}{c} y^{k+1} - y^k \\ \lambda^k - \lambda^{k+\frac{1}{2}} \end{array} \right\|_{H_\mu}^2 - \left\| \begin{array}{c} 0 \\ \lambda^{k+1} - \lambda^{k+\frac{1}{2}} \end{array} \right\|_{H_\mu}^2 \right) \end{aligned} \quad (3.7)$$

It follows from (1.12b) and (1.12d) that

$$\lambda^{k+1} - \lambda^{k+\frac{1}{2}} = \lambda^{k+\frac{1}{2}} - \lambda^k + \frac{1}{\mu}(y^{k+1} - y^k).$$

Hence,

$$\left\| \begin{array}{c} 0 \\ \lambda^{k+1} - \lambda^{k+\frac{1}{2}} \end{array} \right\|_{H_\mu}^2 = \mu\|\lambda^{k+1} - \lambda^{k+\frac{1}{2}}\|^2 = \mu\left\|\lambda^{k+\frac{1}{2}} - \lambda^k + \frac{1}{\mu}(y^{k+1} - y^k)\right\|^2.$$

A simple calculation yields

$$\begin{aligned}\left\| \begin{array}{c} y^{k+1}-y^k \\ \lambda^k-\lambda^{k+\frac{1}{2}} \end{array} \right\|_{H_\mu}^2 &= \frac{1}{\mu}\|y^{k+1}-y^k\|^2+\frac{2}{\mu}\langle y^{k+1}-y^k,\lambda^{k+\frac{1}{2}}-\lambda^k\rangle+\mu\|\lambda^{k+\frac{1}{2}}-\lambda^k\|^2 \\ &= \mu\left\|\lambda^{k+\frac{1}{2}}-\lambda^k+\frac{1}{\mu}(y^{k+1}-y^k)\right\|^2.\end{aligned}$$

Consequently,

$$\left\| \begin{array}{c} y^{k+1}-y^k \\ \lambda^k-\lambda^{k+\frac{1}{2}} \end{array} \right\|_{H_\mu}^2 - \left\| \begin{array}{c} 0 \\ \lambda^{k+1}-\lambda^{k+\frac{1}{2}} \end{array} \right\|_{H_\mu}^2 = 0.$$

Thus,

$$\frac{1}{2}\left\langle \begin{array}{c} y^{k+1}-y \\ \lambda^{k+\frac{1}{2}}-\lambda \end{array}, H_\mu \left(\begin{array}{c} y^k-y^{k+1} \\ \lambda^k-\lambda^{k+1} \end{array} \right) \right\rangle = \frac{1}{4}\left(\left\| \begin{array}{c} y^k-y \\ \lambda^k-\lambda \end{array} \right\|_{H_\mu}^2 - \left\| \begin{array}{c} y^{k+1}-y \\ \lambda^{k+1}-\lambda \end{array} \right\|_{H_\mu}^2 \right).$$

The assertion (3.1) then immediately follows from the above inequality and (3.6). □

Based on Lemma 3.1, it is easy to show the convergence of the PRSM scheme (1.12a)-(1.12d). Although the convergence of PRSM has been proved in the literature for generic settings (e.g.[9, 17]), here we include the detail of proof in the specific setting of (1.7) for completeness.

Theorem 3.1. *For (1.1), let f and g be convex, f be nonsmooth, g be smooth and ∇g be Lipscitz continuous with a constant L in (1.1). Suppose the set of saddle points S is nonempty. Then the sequence $\{(x^k, y^k, \lambda^k)\}$ generated by PRSM scheme (1.12a)-(1.12d) with $0 < \mu < 1/L$ converges to a saddle-point in S.*

Proof. We first show that the sequence $\{(x^k, y^k, \lambda^k)\}$ is bounded. Let $(\hat{x}, \hat{y}, \hat{\lambda}) \in S$ be an arbitrary saddle point. Recall the relations in (3.3) and (2.2). The Lipschitz continuity of ∇g then implies that

$$\|(y^{k+1}-\hat{y})+\mu(\lambda^{k+1}-\hat{\lambda})\| \geq \|y^{k+1}-\hat{y}\|-\mu\|\lambda^{k+1}-\hat{\lambda}\| \geq (1-\mu/L)\|y^{k+1}-\hat{y}\|. \tag{3.8}$$

Thus, we have

$$\left\| \begin{array}{c} y^{k+1}-\hat{y} \\ \lambda^{k+1}-\hat{\lambda} \end{array} \right\|_{H_\mu}^2 = \frac{1}{\mu}\|(y^{k+1}-\hat{y})+\mu(\lambda^{k+1}-\hat{\lambda})\|^2 \geq \frac{(1-\mu/L)^2}{\mu}\|y^{k+1}-\hat{y}\|^2,$$

which, together with (3.14), means that $\{y^k\}$ is bounded. The continuity of ∇g and (3.3) gurantee that $\{\lambda^k\}$ is bounded. Finally, it follows from (1.12d) that

$$x^{k+1} = \frac{1}{\mu}(\lambda^k - \lambda^{k+1}) + y^{k+1}. \tag{3.9}$$

Since $\{(y^k, \lambda^k)\}$ is bounded, $\{x^k\}$ is also bounded.

Setting $(x, y, \lambda) := (\hat{x}, \hat{y}, \hat{\lambda})$ in (3.1), it follows that

$$\begin{aligned}&[f(x^{k+1}) + g(y^{k+1}) - \langle \hat{\lambda}, x^{k+1} - y^{k+1}\rangle] - [f(\hat{x}) + g(\hat{y}) - \langle \lambda^{k+1}, \hat{x} - \hat{y}\rangle] \\ &\leq \frac{1}{4}\left(\left\| \begin{array}{c} y^k - \hat{y} \\ \lambda^k - \hat{\lambda} \end{array} \right\|_{H_\mu}^2 - \left\| \begin{array}{c} y^{k+1} - \hat{y} \\ \lambda^{k+1} - \hat{\lambda} \end{array} \right\|_{H_\mu}^2 \right) + \frac{1}{\mu}\langle x^{k+1} - y^{k+1}, \hat{y} - \hat{x}\rangle. \end{aligned} \tag{3.10}$$

Since $\{(x^k, y^k, \lambda^k)\}$ is bounded, it has at least one cluster point. Let $(\tilde{x}, \tilde{y}, \tilde{\lambda})$ be an arbitrary cluster point of $\{(x^k, y^k, \lambda^k)\}$ and let $\{(x^{k_j}, y^{k_j}, \lambda^{k_j})\}$ be the corresponding subsequence that converges to $(\tilde{x}, \tilde{y}, \tilde{\lambda})$. Then, taking limit along the subsequence and using $\hat{x} = \hat{y}$, we have from (3.10) that

$$[f(\tilde{x}) + g(\tilde{y}) - \langle \hat{\lambda}, \tilde{x} - \tilde{y}\rangle] - [f(\hat{x}) + g(\hat{y}) - \langle \tilde{\lambda}, \hat{x} - \hat{y}\rangle] = 0.$$

Hence, it follows from Lemma 2.1 that $(\tilde{x}, \tilde{y}, \tilde{\lambda})$ is a saddle point.

Now, we prove that $(\tilde{x}, \tilde{y}, \tilde{\lambda})$ is the unique cluster point of $\{(x^k, y^k, \lambda^k)\}$. From (3.9), it is sufficient to prove that $(\tilde{y}, \tilde{\lambda})$ is the unique cluster point of $\{(y^k, \lambda^k)\}$. Suppose there is another cluster point $(\check{y}, \check{\lambda}) \neq (\tilde{y}, \tilde{\lambda})$. Since $(\tilde{y}, \tilde{\lambda})$ is a saddle point and $(\hat{y}, \hat{\lambda})$ is arbitrary, we can set $(\hat{y}, \hat{\lambda}) = (\tilde{y}, \tilde{\lambda})$ in (3.14), yielding

$$\left\| \begin{array}{c} y^{k+1} - \tilde{y} \\ \lambda^{k+1} - \tilde{\lambda} \end{array} \right\|_{H_\mu}^2 \leq \left\| \begin{array}{c} y^k - \tilde{y} \\ \lambda^k - \tilde{\lambda} \end{array} \right\|_{H_\mu}^2. \tag{3.11}$$

Moreover, the fact that $(\tilde{y}, \tilde{\lambda})$ is a cluster point of $\{(y^k, \lambda^k)\}$ and (3.11) imply that there is $K_1 > 0$, such that for all $k \geq K_1$,

$$\left\| \begin{array}{c} y^{k+1} - \tilde{y} \\ \lambda^{k+1} - \tilde{\lambda} \end{array} \right\|_{H_\mu} \leq \frac{1}{3} \left\| \begin{array}{c} \check{y} - \tilde{y} \\ \check{\lambda} - \tilde{\lambda} \end{array} \right\|_{H_\mu}. \tag{3.12}$$

Thus, for all $k \geq K_1$,

$$\left\| \begin{array}{c} y^{k+1} - \check{y} \\ \lambda^{k+1} - \check{\lambda} \end{array} \right\|_{H_\mu} \geq \left\| \begin{array}{c} \check{y} - \tilde{y} \\ \check{\lambda} - \tilde{\lambda} \end{array} \right\|_{H_\mu} - \left\| \begin{array}{c} y^{k+1} - \tilde{y} \\ \lambda^{k+1} - \tilde{\lambda} \end{array} \right\|_{H_\mu} > 0, \tag{3.13}$$

which indicates that $(\check{y}, \check{\lambda})$ can not be a cluster point of $\{(y^k, \lambda^k)\}$. Thus, the generated sequence has just one cluster point and the whole sequence converges. □

Lemma 3.1 suffices to establish the sublinear convergence rate for the PRSM scheme (1.12a)-(1.12d). We complete it in the following theorem.

Theorem 3.2. *For (1.1), let f and g be convex, f be nonsmooth, g be smooth and ∇g be Lipscitz continuous with a constant L. Suppose the set of saddle points S is nonempty. After t iterations, the PRSM scheme (1.12a)-(1.12d) with $0 < \mu < 1/L$ yields an approximate solution of (1.1) with an accuracy of $O(1/t)$.*

Proof. Since $(\hat{x}, \hat{y}, \hat{\lambda})$ is a saddle point, it follows from (2.5) that the left-hand side of (3.1) is nonegative. We thus have

$$\left\| \begin{array}{c} y^{k+1} - \hat{y} \\ \lambda^{k+1} - \hat{\lambda} \end{array} \right\|_{H_\mu}^2 \leq \left\| \begin{array}{c} y^{k} - \hat{y} \\ \lambda^{k} - \hat{\lambda} \end{array} \right\|_{H_\mu}^2 \leq \cdots \leq \left\| \begin{array}{c} y^{0} - \hat{y} \\ \lambda^{0} - \hat{\lambda} \end{array} \right\|_{H_\mu}^2. \tag{3.14}$$

Summing (3.10) from $k = 0$ to $t - 1$ yields

$$\begin{aligned} &\sum_{k=0}^{t-1} [f(x^{k+1}) + g(y^{k+1}) - \langle \hat{\lambda}, x^{k+1} - y^{k+1} \rangle] - [f(\hat{x}) + g(\hat{y}) - \langle \lambda^{k+1}, \hat{x} - \hat{y} \rangle] \\ &\leq \frac{1}{4} \left\| \begin{array}{c} y^{0} - \hat{y} \\ \lambda^{0} - \hat{\lambda} \end{array} \right\|_{H_\mu}^2. \end{aligned} \tag{3.15}$$

Define

$$x_t := \frac{1}{t} \sum_{k=0}^{t} x^k \quad \text{and} \quad y_t := \frac{1}{t} \sum_{k=0}^{t} y^k. \tag{3.16}$$

It then follows from the convexity of f and g that

$$f(x_t) \leq \frac{1}{t} \sum_{k=0}^{t-1} f(x^k) \quad \text{and} \quad g(y_t) \leq \frac{1}{t} \sum_{k=0}^{t-1} g(y^k). \tag{3.17}$$

Combining (3.15) and (3.17),

$$t[f(x_t) + g(y_t) - \langle \hat{\lambda}, x_t - y_t \rangle] - t[f(\hat{x}) + g(\hat{y})] \leq \frac{1}{4} \left\| \begin{array}{c} y^{0} - \hat{y} \\ \lambda^{0} - \hat{\lambda} \end{array} \right\|_{H_\mu}^2. \tag{3.18}$$

Since $\{(x^k, y^k, \lambda^k)\}$ converges to a saddle point, $\{(x_t, y_t, \lambda_t)\}$ also converges to the same saddle point, and (3.18) implies that

$$\mathcal{G}_{\mathcal{B}_1 \times \mathcal{B}_2 \times \mathcal{B}_3}(x_t, y_t, \lambda_t) \leq \frac{1}{4t} \left\| \begin{array}{c} y^{0} - \hat{y} \\ \lambda^{0} - \hat{\lambda} \end{array} \right\|_{H_\mu}^2.$$

That is, x_t is an approximate solution of (1.1) with an accuracy of $O(1/t)$. A worst-case $O(1/t)$ iteration complexity in the ergodic sense is thus established for the PRSM scheme (1.12a)-(1.12d). This completes the proof. □

Remark 3.1. *Along this line of research on estimating worst-case convergence rate in term of iteration complexity, the important thing is to show the existence (x_t in Theorem 3.2) of an approximate solution with the accuracy $O(1/t)$, and this approximate solution could be a combination of the iterates generated by PRSM (In Theorem 3.2, it is the average of all generated iterates).*

Remark 3.2. *Under certain additional conditions, the technique in [12] which requires to use a variational inequality characterization as the measurement of optimality could be extended to establish similar worst-case $O(1/t)$ iteration complexity in ergodic sense for PRSM. But our new result using the primal-dual-gap technique measures directly the rate of convergence of the iterative sequence $\{(x^k, y^k, \lambda^k)\}$ generated by PRSM.*

4 Linear convergence rate

In this section, we show that the linear convergence rate of the PRSM scheme (1.12a)-(1.12d) can be derived if g is further assumed to be strongly convex with a modulus γ in (1.1). Note this assumption is satisfied for some concrete applications of (1.1). For example, for the $l_1 - l_2$ model (1.3), if the matrix A is full column-rank, then this assumption is satisfied.

First, with the convexity assumption of f and strong convexity assumption of g, it follows from (3.2) and (3.3) that

$$\begin{array}{rcl} f(x) & \geq & f(x^{k+1}) + \langle \lambda^{k+\frac{1}{2}}, x - x^{k+1} \rangle \\ g(y) & \geq & g(y^{k+1}) - \langle \lambda^{k+1}, y - y^{k+1} \rangle + \frac{\gamma}{2}\|y - y^{k+1}\|^2. \end{array} \tag{4.1}$$

We thus can establish a result similar as (3.1) except that one more quadratic term appears on the right-hand side of the inequality in (4.2). Since its proof is analogous to that of (3.1), we omit it.

Lemma 4.1. *For (1.1), let f be convex and nonsmooth, g be strongly convex with a modulus γ and smooth, and ∇g be Lipscitz continuous with a constant L. For any $(x, y, \lambda) \in \mathcal{R}^n \times \mathcal{R}^n \times \mathcal{R}^n$, the sequence $\{(x^k, y^k, \lambda^k)\}$ generated by the PRSM scheme (1.12a)-(1.12d) satisfies*

$$\begin{aligned} & [f(x^{k+1}) + g(y^{k+1}) - \langle \lambda, x^{k+1} - y^{k+1} \rangle] - [f(x) + g(y) - \langle \lambda^{k+1}, x - y \rangle] \\ & \leq \frac{1}{4}\left(\left\| \begin{array}{c} y^k - y \\ \lambda^k - \lambda \end{array} \right\|_{H_\mu}^2 - \left\| \begin{array}{c} y^{k+1} - y \\ \lambda^{k+1} - \lambda \end{array} \right\|_{H_\mu}^2 \right) + \frac{1}{\mu}\langle x^{k+1} - y^{k+1}, y - x \rangle \\ & \quad - \frac{\gamma}{2}\|y - y^{k+1}\|^2. \end{aligned} \tag{4.2}$$

Now, with Lemma 4.1, we are able to establish the linear convergence rate for the PRSM scheme (1.12a)-(1.12d).

Theorem 4.1. *For (1.1), let f be convex and nonsmooth, g be strongly convex with a modulus γ and smooth, and ∇g be Lipscitz continuous with a constant L. Suppose the set of saddle points S is nonempty. Then the sequence $\{(y^k, \lambda^k)\}$ generated by the PRSM scheme (1.12a)-(1.12d) with $0 < \mu < 1/L$ converges linearly to $(\hat{y}, \hat{\lambda})$ in the following sense:*

$$\left\| \begin{array}{c} y^{k+1} - \hat{y} \\ \lambda^{k+1} - \hat{\lambda} \end{array} \right\|_{H_\mu} \leq \xi \left\| \begin{array}{c} y^k - \hat{y} \\ \lambda^k - \hat{\lambda} \end{array} \right\|_{H_\mu} \tag{4.3}$$

where

$$0 < \xi := \sqrt{\frac{(1+\mu)(1+L^2)}{\mu + (1+\mu)(1+L^2)}} < 1.$$

Proof. Setting $(x, y, \lambda) = (\hat{x}, \hat{y}, \hat{\lambda})$ in (4.2) and using $\hat{x} = \hat{y}$, we obtain

$$\begin{aligned}&[f(x^{k+1}) + g(y^{k+1}) - \langle \hat{\lambda}, x^{k+1} - y^{k+1}\rangle] - [f(\hat{x}) + g(\hat{y})]\\ &\leq \frac{1}{4}\left(\left\| \begin{array}{c} y^k - \hat{y} \\ \lambda^k - \hat{\lambda} \end{array} \right\|_{H_\mu}^2 - \left\| \begin{array}{c} y^{k+1} - \hat{y} \\ \lambda^{k+1} - \hat{\lambda} \end{array} \right\|_{H_\mu}^2\right) - \frac{\gamma}{2}\|\hat{y} - y^{k+1}\|^2. \end{aligned} \tag{4.4}$$

Recall that the left-hand side of the above inequality is nonnegative, it then follows that

$$\left\| \begin{array}{c} y^{k+1} - \hat{y} \\ \lambda^{k+1} - \hat{\lambda} \end{array} \right\|_{H_\mu}^2 \leq \left\| \begin{array}{c} y^k - \hat{y} \\ \lambda^k - \hat{\lambda} \end{array} \right\|_{H_\mu}^2 - \frac{\gamma}{2}\|\hat{y} - y^{k+1}\|^2. \tag{4.5}$$

Since $\lambda^{k+1} = -\nabla g(y^{k+1})$, $\hat{\lambda} = -\nabla g(\hat{y})$

$$\|\lambda^{k+1} - \hat{\lambda}\|^2 \leq L^2\|y^{k+1} - \hat{y}\|^2.$$

Then,

$$\|y^{k+1} - \hat{y}\|^2 \geq \frac{1}{1+L^2}(\|y^{k+1} - \hat{y}\|^2 + \|\lambda^{k+1} - \hat{\lambda}\|^2). \tag{4.6}$$

On the other hand,

$$\|y^{k+1} - \hat{y}\|^2 + \|\lambda^{k+1} - \hat{\lambda}\|^2 = \left\| \begin{array}{c} y^{k+1} - \hat{y} \\ \lambda^{k+1} - \hat{\lambda} \end{array} \right\|^2 \geq \frac{1}{\lambda_{\max}(H_\mu)} \left\| \begin{array}{c} y^{k+1} - \hat{y} \\ \lambda^{k+1} - \hat{\lambda} \end{array} \right\|_{H_\mu}^2, \tag{4.7}$$

and

$$\lambda_{\max}(H_\mu) = \mu + \frac{1}{\mu}. \tag{4.8}$$

Combining the inequalities (4.5)-(4.8), we obtain

$$\left\| \begin{array}{c} y^{k+1} - \hat{y} \\ \lambda^{k+1} - \hat{\lambda} \end{array} \right\|_{H_\mu}^2 \leq \frac{(1+\mu)(1+L^2)}{\mu + (1+\mu)(1+L^2)} \left\| \begin{array}{c} y^k - \hat{y} \\ \lambda^k - \hat{\lambda} \end{array} \right\|_{H_\mu}^2,$$

and the assersion (4.3) follows immediately. This completes the proof. □

Remark 4.1. *In Theorem 4.1 we show the linear convergence rate in term of the sequence* $\{(y^k, \lambda^k)\}$. *Recall the PRSM scheme* (1.12a)-(1.12d). *It is easy to see that the variable* x^k *is not required to perform the* $(k+1)$*-th iteration; it is thus an intermediate variable. Thereofore, it is meaningful to measure the convergence rate of the PRSM scheme* (1.12a)-(1.12d) *only in term of the sequence* $\{(y^k, \lambda^k)\}$.

5 A modified PRSM with a moving proximal center and its convergence rates

In Sections 3 and 4, we analyze the convergence of the PRSM scheme (1.12a)-(1.12d) under the conventional assumption that $\mu < 1/L$. As we have mentioned, sometimes it is too expensive

or impractical to obey this restrictive assumption in implementation. In this section, we show that this requirement can be eliminated for the PRSM scheme (1.12a)-(1.12d). Moreover, the conclusions established in Sections 3 and 4 can be extended easily to this modified PRSM scheme.

We first present the iterative scheme of this modified PRSM scheme:

$$\begin{cases} x^{k+1} = \arg\min\limits_{x\in\mathcal{R}^n} \mathcal{L}_\mu(x, y^k, \lambda^k), & (5.1a)\\ \lambda^{k+\frac{1}{2}} = \lambda^k - \dfrac{1}{\mu}(x^{k+1} - y^k), & (5.1b)\\ y^{k+1} = \arg\min\limits_{y\in\mathcal{R}^n} \mathcal{L}_\mu(x^{k+1}, y, \lambda^{k+\frac{1}{2}}) + \dfrac{\beta}{2}\|y - y^k\|^2, & (5.1c)\\ \lambda^{k+1} = \lambda^{k+\frac{1}{2}} - \dfrac{1}{\mu}(x^{k+1} - y^{k+1}), & (5.1d) \end{cases}$$

where $\beta > 0$ is a proximal parameter. In (5.1a)-(5.1d), both the parameters μ and β can be taken as any positive constants to ensure convergence, and the requirement $\mu < 1/L$ is not required. With the relaxed condition on μ, it becomes possible to tune this parameter to find an appropriate value for a given specific application of (1.1).

Let us make some remarks for the modified PRSM scheme (5.1a)-(5.1d). First, its only difference from the PRSM scheme (1.12a)-(1.12d) is the y^{k+1}-subproblem where an additional proximal term is added. In fact, the subproblem (1.12c) can be written as (subject to a difference of constant in the objective function):

$$y^{k+1} = \arg\min\{g(y) + \frac{1}{2\mu}\|y - (x^{k+1} - \mu\lambda^{k+\frac{1}{2}})\|_2^2\};$$

while the subproblem (5.1c) can be written as (subject to a difference of constant in the objective function):

$$y^{k+1} = \arg\min\{g(y) + \frac{1}{2\mu}\|y - (\frac{1}{1+\mu\beta}(x^{k+1} - \mu\lambda^{k+\frac{1}{2}} + \mu\beta y^k))\|_2^2\}.$$

That is, the proximal center $(x^{k+1} - \mu\lambda^{k+\frac{1}{2}})$ in (1.12c) is now changed to $(\frac{1}{1+\mu\beta}(x^{k+1} - \mu\lambda^{k+\frac{1}{2}} + \mu\beta y^k))$. Thus, we name the scheme (5.1a)-(5.1d) as a modified PRSM with a moving proximal center. Moreover, the subproblem (5.1c) is of the same difficulty as (1.12c).

To extend the conclusions in Sections 3 and 4 to the modified PRSM scheme (5.1a)-(5.1d), we note that the first two steps are the same as those of (1.12a)-(1.12d). Consequently, (3.2) holds. The optimality condition for the y-subproblem and the update scheme of λ imply that

$$-\lambda^{k+1} + \beta(y^{k+1} - y^k) = \nabla g(y^{k+1}). \tag{5.2}$$

As a result, (3.4) turns to be

$$\begin{array}{rcl} f(x) & \geq & f(x^{k+1}) + \langle\lambda^{k+\frac{1}{2}}, x - x^{k+1}\rangle \\ g(y) & \geq & g(y^{k+1}) - \langle\lambda^{k+1}, y - y^{k+1}\rangle + \beta\langle y^{k+1} - y^k, y - y^{k+1}\rangle. \end{array} \tag{5.3}$$

In the following lemma, we show that the sequence $\{(x^k, y^k, \lambda^k)\}$ generated by the modified PRSM scheme (5.1a)-(5.1d) is bounded.

Lemma 5.1. *Suppose the set of saddle points S is nonempty. Then the sequence $\{(x^k, y^k, \lambda^k)\}$ generated by the modified PRSM scheme (5.1a)-(5.1d) is bounded.*

Proof. Using (5.3), (3.6) now becomes that for any $(x, y, \lambda) \in \mathcal{R}^n \times \mathcal{R}^n \times \mathcal{R}^n$,

$$\begin{aligned}&[f(x^{k+1}) + g(y^{k+1}) - \langle \lambda, x^{k+1} - y^{k+1}\rangle] - [f(x) + g(y) - \langle \lambda^{k+1}, x - y\rangle] \\ &\leq \frac{1}{2}\left\langle \begin{pmatrix} y^{k+1} - y \\ \lambda^{k+\frac{1}{2}} - \lambda \end{pmatrix}, H_{\mu\beta}\begin{pmatrix} y^k - y^{k+1} \\ \lambda^k - \lambda^{k+1} \end{pmatrix}\right\rangle + \frac{1}{\mu}\langle x^{k+1} - y^{k+1}, y - x\rangle, \end{aligned} \tag{5.4}$$

where

$$H_{\mu\beta} = \begin{pmatrix} (\frac{1}{\mu} + 2\beta)I & I \\ I & \mu I \end{pmatrix}. \tag{5.5}$$

Similar to the proof of Lemma 3.1, we can prove that for any $(x, y, \lambda) \in \mathcal{R}^n \times \mathcal{R}^n \times \mathcal{R}^n$,

$$\begin{aligned}&[f(x^{k+1}) + g(y^{k+1}) - \langle \lambda, x^{k+1} - y^{k+1}\rangle] - [f(x) + g(y) - \langle \lambda^{k+1}, x - y\rangle] \\ &\leq \frac{1}{4}\left(\left\| \begin{pmatrix} y^k - y \\ \lambda^k - \lambda \end{pmatrix}\right\|^2_{H_{\mu\beta}} - \left\| \begin{pmatrix} y^{k+1} - y \\ \lambda^{k+1} - \lambda \end{pmatrix}\right\|^2_{H_{\mu\beta}}\right) + \frac{1}{\mu}\langle x^{k+1} - y^{k+1}, y - x\rangle. \end{aligned} \tag{5.6}$$

Setting $(x, y, \lambda) = (\hat{x}, \hat{y}, \hat{\lambda})$ in (5.6) and using the nonnegativeness of the left hand, we obtain

$$\left\| \begin{pmatrix} y^{k+1} - \hat{y} \\ \lambda^{k+1} - \hat{\lambda} \end{pmatrix}\right\|^2_{H_{\mu\beta}} \leq \left\| \begin{pmatrix} y^k - \hat{y} \\ \lambda^k - \hat{\lambda} \end{pmatrix}\right\|^2_{H_{\mu\beta}} \leq \cdots \leq \left\| \begin{pmatrix} y^0 - \hat{y} \\ \lambda^0 - \hat{\lambda} \end{pmatrix}\right\|^2_{H_{\mu\beta}}. \tag{5.7}$$

From the definition of $H_{\mu\beta}$, we have

$$\left\| \begin{pmatrix} y^{k+1} - \hat{y} \\ \lambda^{k+1} - \hat{\lambda} \end{pmatrix}\right\|^2_{H_{\mu\beta}} = \frac{1}{\mu}\|(y^{k+1} - \hat{y}) + \mu(\lambda^{k+1} - \hat{\lambda})\|^2 + \beta\|y^{k+1} - \hat{y}\|^2.$$

Hence, $\{y^k\}$ is bounded and the continuity of ∂g and (5.2) implies that $\{\lambda^k\}$ is bounded. Finally, the boundedness of $\{x^k\}$ follows from its relationship to $\{(y^k, \lambda^k)\}$. □

Based on Lemma 5.1, we can easily prove the global convergence of the modified PRSM scheme (5.1a)-(5.1d), and establish its convergence rate. Since all the proofs are analogous to those in Sections 3 and 4, we only list the conclusions in the following theorem and omit the proof.

Theorem 5.1. *Suppose that (1.7) has an optimal $(\hat{x}, \hat{y})$ with $\hat{\lambda}$ being the associated optimal multiplier to the linear constraint. Then*

1). Let f and g be convex, f be nonsmooth, g be smooth and ∇g be Lipscitz continuous. The modified PRSM scheme (5.1a)-(5.1d) with any $\beta > 0$ and $\mu > 0$ converges to a saddle point in S.

2). Let f and g be convex, f be nonsmooth, g be smooth and ∇g be Lipscitz continuous. The modified PRSM scheme (5.1a)-(5.1d) with any $\beta > 0$ and $\mu > 0$ yields an approximate solution of (1.1) with an accuracy of $O(1/t)$.

3). Let f be convex and nonsmooth, g be strongly convex with a modulus γ and smooth, and ∇g be Lipscitz continuous with a constant L in (1.1). The modified PRSM scheme (5.1a)-(5.1d) with any $\beta > 0$ and $\mu > 0$ converges linearly to a saddle point of in S.

6 Conclusion

In this paper, we analyze the convergence of the Peaceman-Rachford splitting method when it is applied to solve a convex minimization model whose objection is the sum of a smooth and nonsmmoth convex functions. The sublinear and linear convergence rates are derived under different conditions. The implementation of the original Peaceman-Rachford splitting method requires to know the Lipchitz constant of the gradient of the smooth objective function. We consider to remove this restrictive requirement and thus propose a modified Peaceman-Rachford splitting method. Its convergence rates are also analyzed under different conditions.

It would be interesting to investigate the convergence rate of the Peaceman-Rachford splitting method in the generic setting of finding a root of two maximal monotone set-valued mappings. To the best of our knowledge, results along this line are very limited, and we will try to extend the technique used in this paper to the generic setting in our future research.

Appendix: Linear convergence of ISTA

In this appendix, we show the linear convergence of ISTA for solving (1.1) under appropriate assumptions.

Theorem *In (1.1), let f be convex and nonsmooth, g be strongly convex with a modulus γ and smooth, and ∇g be Lipscitz continuous with a constant L. If the linearization parameter μ is chosen as $0 < \mu < \frac{2\gamma}{L^2}$, then the sequence $\{x^k\}$ generated by ISTA converges to the solution x^* of (1.1) linearly. In particular, it holds*

$$\|x^{k+1} - x^*\| \leq \xi \|x^k - x^*\|,$$

where

$$0 < \xi := \sqrt{1 - 2\mu\gamma + \mu^2 L^2} < 1.$$

Proof. First, the strong convexity of g implies the strong monotonicity of ∇g, i.e.,

$$\langle x_1 - x_2, \nabla g(x_1) - \nabla g(x_2)\rangle \geq \gamma \|x_1 - x_2\|^2, \ \forall x_1, x_2 \in \mathcal{R}^n.$$

Due to the strong convexity of g and the Lipschitz continuity of ∇g, we have $L \geq \gamma$. Since f is convex, $(I + \mu\partial f)^{-1}$ is nonexpansive. On the other hand, since x^* is the (unique) solution of (1.1), we have

$$x^* = (I + \mu\partial f)^{-1}(x^* - \mu\nabla g(x^*)), \ \ \forall \mu > 0.$$

Consequently,

$$\begin{aligned}
\|x^{k+1} - x^*\|^2 &= \left\|(I + \mu\partial f)^{-1}(x^k - \mu\nabla g(x^k)) - (I + \mu\partial f)^{-1}(x^* - \mu\nabla g(x^*))\right\|^2 \\
&\leq \|x^k - \mu\nabla g(x^k) - (x^* - \mu\nabla g(x^*))\|^2 \\
&= \|x^k - x^*\|^2 - 2\mu\langle x^k - x^*, \nabla g(x^k) - \nabla g(x^*)\rangle + \mu^2\|\nabla g(x^k) - \nabla g(x^*)\|^2 \\
&\leq (1 - 2\mu\gamma + \mu^2 L^2)\|x^k - x^*\|^2,
\end{aligned}$$

where the last inequality follows from the strong monotonicity and Lipschitz continuity of ∇g. In addition, because of the choice of μ, we have $0 < \sqrt{1 - 2\mu\gamma + \mu^2 L^2} < 1$. This completes the proof. □

References

[1] A. Beck and M. Teboulle. *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci., 2(1) (2009), pp. 183-202.

[2] D. P. Bertsekas, Constrained Optimization and Lagrange Multiplier Methods, Academic Press, 1982.

[3] A. Chamboulle and T. Pock, *A first-order primal-dual algorithm for convex problems with applications to imaging*, J. Math. Imaging Vis., 40 (2011), pp. 120-145.

[4] S. S. Chen, D. Donoho and M. A. Saunders, *Atomic decomposition by basis pursuit*, SIAM J. Sci. Comput. 20(1) (1998), pp. 33-61.

[5] J. Douglas and H. H. Rachford, *On the numerical solution of the heat conduction problem in 2 and 3 space variables*, Trans. Amer. Math. Soc., 82 (1956), pp. 421-439.

[6] D. Gabay, *Applications of the method of multipliers to variational inequalities*, In: Fortin, M., Glowinski, R. (eds.) Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems, pp. 299-331, North-Holland, Amsterdam, 1983.

[7] R. Glowinski and A. Marrocco, *Approximation par éléments finis d'ordre un et résolution par pénalisation-dualité d'une classe de problèmes non linéaires*, R.A.I.R.O., R2 (1975), pp. 41-76.

[8] R. Glowinski, T. Kärkkäinen and K. Majava, *On the convergence of operator-splitting methods*, In: Kuznetsov, Y., Neittanmaki, P., and Pironneau O. (eds.) Numerical Methods for Scienfic computing, Variational Problems and Applications, Barcelona, 2003.

[9] R. Glowinski and P. Le Tallec, Augmented Lagrangian and Operator-Splitting Methods in Nonlinear Mechanics, SIAM Studies in Applied Mathematics, Philadelphia, 1989.

[10] D. Goldfarb, S. Q. Ma, and K. Scheinberg, *Fast alternating linearization methods for minimizing the sum of two convex functions*, Math. Prog., to appear.

[11] B. S. He, H. Liu, Z. R. Wang and X. M. Yuan, *A strictly contractive Peaceman-Rachford splitting method for convex programming*, SIAM J. Optim, under revision, 2013.

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

[13] B. S. He and X. M. Yuan, *On nonergodic convergence rate of Douglas-Rachford alternating direction method of multipliers*, submission, manuscript, 2012.

[14] B. S. He and X. M. Yuan, *On convergence rate of the Douglas-Rachford operator splitting method*, Math. Program., under revision, 2012.

[15] M. R. Hestenes, *Multiplier and gradient methods*, J. Optim. Theory Appli., 4 (1969), pp. 303-320.

[16] K. C. Kiwiel, C. H. Rosa, and A. Ruszczynski, *Proximal decomposition via alternating linearization*, SIAM J. Optim., 9 (1999), pp. 668-689.

[17] P. L. Lions and B. Mercier, *Splitting algorithms for the sum of two nonlinear operators*, SIAM J. Num. Anal., 16(1979), pp. 964-979.

[18] Y. E. Nesterov, *A method for solving the convex programming problem with convergence rate $O(1/k^2)$*, Dokl. Akad. Nauk SSSR, 269 (1983), pp. 543-547.

[19] Y. E. Nesterov, *Gradient methods for minimizing composite objective function*, CORE report 2007; available at http://www.ecore.be/DPs/dp-1191313936.pdf.

[20] G. B. Passty, *Ergodic convergence to a zero of the sum of monotone operators in Hilbert space*, J. Math. Analy. Applic. 72 (1979), pp. 383-390.

[21] D. H. Peaceman and H. H. Rachford, *The numerical solution of parabolic elliptic differential equations*, SIAM J. Appl. Math. 3 (1955), 28-41.

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

[23] P. Tseng, *A modified forward-backword splitting method for maximal monotone mappings*, SIAM. J. Con. Optim., 38 (2000), pp. 431-446.