Peaceman-Rachford splitting for a class of nonconvex optimization problems

Guoyin Li * Tianxiang Liu† Ting Kei Pong ‡

January 9th, 2017

Abstract

We study the applicability of the Peaceman-Rachford (PR) splitting method for solving nonconvex optimization problems. When applied to minimizing the sum of a strongly convex Lipschitz differentiable function and a proper closed function, we show that if the strongly convex function has a large enough strong convexity modulus and the step-size parameter is chosen below a threshold that is computable, then any cluster point of the sequence generated, if exists, will give a stationary point of the optimization problem. We also give sufficient conditions guaranteeing boundedness of the sequence generated. We then discuss one way to split the objective so that the proposed method can be suitably applied to solving optimization problems with a coercive objective that is the sum of a (not necessarily strongly) convex Lipschitz differentiable function and a proper closed function; this setting covers a large class of nonconvex feasibility problems and constrained least squares problems. Finally, we illustrate the proposed algorithm numerically.

1 Introduction

Consider the following optimization problem with competing structure:

$$\min_u \ f(u) + g(u), \tag{1}$$

where f and g are proper closed possibly nonconvex functions. Optimization problems of this form arise in many important modern applications such as signal processing, machine learning and statistics [6, 10, 17, 32]. A typical application of (1) is to solve some ill-posed inverse problems where the function f represents the data fitting term and the function g is the regularization term. To solve problems with competing structures, an important and powerful class of algorithms is the class of splitting methods. In these methods, the objective function is decomposed into simpler individuals which are then processed separately in the subproblems. Two classical splitting methods in the literature are the Douglas-Rachford (DR) splitting method [15, 16, 26] and the Peaceman-Rachford (PR) splitting method [26, 30].

The PR splitting method was originally introduced in [30] for solving linear heat flow equations, and was later generalized to deal with nonlinear equations in [26]. In the case when f and g are

*Department of Applied Mathematics, University of New South Wales, Sydney 2052, Australia. E-mail: g.li@unsw.edu.au. This author was partially supported by a research grant from Australian Research Council.

†Department of Applied Mathematics, the Hong Kong Polytechnic University, Hong Kong. E-mail: tiskyliu@polyu.edu.hk. This author was supported partly by the AMSS-PolyU Joint Research Institute Postdoctoral Scheme.

‡Department of Applied Mathematics, the Hong Kong Polytechnic University, Hong Kong. E-mail: tk.pong@polyu.edu.hk. This author's work was supported partly by Hong Kong Research Grant Council PolyU253008/15p.

both convex, the PR splitting method can be described conveniently by the following update:

$$x^{t+1} = (2\text{prox}_{\gamma g} - I) \circ (2\text{prox}_{\gamma f} - I)(x^t), \tag{2}$$

where I is the identity mapping, $\gamma > 0$ and

$$\text{prox}_{\gamma h}(z) := \underset{u}{\text{Arg min}} \left\{ \gamma h(u) + \frac{1}{2}\|u - z\|^2 \right\},$$

i.e., the set of minimizers of the problem $\min_u \gamma h(u) + \frac{1}{2}\|u - z\|^2$; we note that this set is a singleton when h is convex. Although the PR splitting method can be faster than the DR splitting method (see, for example, [18] and Example 1 in Appendix), the PR splitting method was not as popular as the DR splitting method. This is also witnessed by the fact that the PR splitting method is not discussed nor mentioned in the recent monograph [5] on operator splitting methods. One of the main reasons for the unpopularity is that, even in the convex settings, the PR splitting method is not convergent in general. To guarantee convergence, typically one would require either the operator $(2\text{prox}_{\gamma f} - I)$ or $(2\text{prox}_{\gamma g} - I)$ to be a *contraction mapping*. In applications where f, g are both convex, this requirement typically needs f or g to be strongly convex, which largely limits the applicability of the PR splitting method; see, for example, [12, 26]. In contrast, under a commonly used constraint qualification which can be easily satisfied, the DR splitting method converges in the convex case [13, Theorem 20]. Moreover, recently, it has been shown in [25] that the DR splitting method can be adapted to a nonconvex setting with global convergence guaranteed under some assumptions. This broadens the applicability of the DR splitting method to cover many nonconvex feasibility problems and many important nonconvex optimization problems arising in statistical machine learning such as the $\ell_{1/2}$ regularized least squares problem.

In this paper, to broaden the applicability of the PR splitting method, we extend it to a nonconvex setting. By constructing a merit function which captures the progress of the PR splitting method, we extend the global convergence of the PR splitting method from the known convex setting to the case where the objective function can be decomposed as the sum of a strongly convex Lipschitz differentiable function and a nonconvex function, under suitable assumptions. As a by-product, this extension also allows us to establish the global convergence and iteration complexity of a new PR splitting method for convex optimization problems in the *absence* of strong convexity. The underlying intuitive idea is that one can decompose a non-strongly convex function $F + G$ into the sum of a strongly convex function $f = F + \gamma\|\cdot\|^2$ and a nonconvex function $g = G - \gamma\|\cdot\|^2$, if a $\gamma > 0$ can be chosen so that f is strongly convex.

The contributions of this paper are two-fold. First, we establish that, for the sequence generated by the PR splitting method applied to minimizing the sum of a strongly convex Lipschitz differentiable function and a proper closed function, if the strongly convex function has a sufficiently large strong convexity modulus and the step-size parameter is chosen below a threshold that is computable, then any cluster point, if exists, gives a stationary point of the optimization problem. We also provide sufficient conditions to guarantee boundedness of the sequence generated. To our knowledge, this is the first work that studies the convergence of the PR splitting method for nonconvex optimization problems. Second, we demonstrate how the method can be suitably applied to minimizing a coercive function $F + G$, where G is a proper closed function, and F is convex Lipschitz differentiable but *not* necessarily strongly convex. Even in the case when G is also convex, it was previously unknown in the literature how the PR splitting method can be suitably applied to solving it. Our study largely broadens the applicability of the PR splitting method. We also discuss global iteration complexity of this new PR splitting method under the additional assumption that G is convex, and establish *global* linear convergence of the sequence generated if $F + G$ is further assumed to be strongly convex.

The rest of the paper is organized as follows. In Section 1.1, we fix the notation and recall some basic definitions which will be used throughout this paper. In Section 2, we establish the convergence of the PR splitting method for nonconvex optimization problems where the objective function can be decomposed as the sum of a strongly convex function and a proper closed function, under suitable assumptions. In Section 3, we demonstrate how the PR splitting method can be applied in the absence of strong convexity. In Section 4, as applications, we illustrate how the PR splitting method can be applied to solving two important classes of nonconvex optimization problems that arise in the area of statistics and machine learning: constrained least squares problem and feasibility problems. We also demonstrate our approach numerically. Our concluding remarks are in Section 5. Finally, in the Appendix, we provide simple and concrete examples illustrating the different behaviors of the classical PR splitting method, the classical DR splitting method and our proposed PR splitting method.

1.1 Notation

In this paper, the n-dimensional Euclidean space is denoted by $\mathbb{R}^n$, with the associated inner product denoted by $\langle\cdot,\cdot\rangle$ and the induced norm denoted by $\|\cdot\|$. For an extended-real-valued function $f:\mathbb{R}^n\to(-\infty,\infty]$, we say that f is proper if it is never $-\infty$ and its domain, $\mathrm{dom} f:=\{x\in\mathbb{R}^n: f(x)<+\infty\}$, is nonempty. Such a function is said to be closed if it is lower semicontinuous. For a proper function f, we let $z\overset{f}{\to}x$ denote $f(z)\to f(x)$ and $z\to x$. The limiting *subdifferential* of f at $x\in\mathrm{dom} f$ is defined by [31]

$$\partial f(x):=\left\{v\in\mathbb{R}^n:\ \exists x^t\overset{f}{\to}x,\ v^t\to v\ \text{ with }\liminf_{z\to x^t}\frac{f(z)-f(x^t)-\langle v^t,z-x^t\rangle}{\|z-x^t\|}\geq 0\text{ for each }t\right\}. \tag{3}$$

From the above definition, one immediately obtains the following robustness property:

$$\left\{v\in\mathbb{R}^n:\ \exists x^t\overset{f}{\to}x,\ v^t\to v\ , v^t\in\partial f(x^t)\right\}\subseteq\partial f(x). \tag{4}$$

The subdifferential (3) reduces to the derivative of f (denoted by ∇f) if f is continuously differentiable, and the classical subdifferential in convex analysis if f is convex (see, for example, [31, Proposition 8.12]). For a function f having more than one group of variables, we let $\partial_x f$ (resp., $\nabla_x f$) denote the subdifferential (resp., derivative) of f with respect to the variable x.

We say that a function f is a *strongly convex function* with modulus $\sigma>0$ if $f-\frac{\sigma}{2}\|\cdot\|^2$ is a convex function. A function f is said to be coercive if $\liminf_{\|x\|\to\infty}f(x)=\infty$. For a nonempty closed set $S\subseteq\mathbb{R}^n$, its indicator function δ_S is defined by

$$\delta_S(x)=\begin{cases}0 & \text{if } x\in S,\\ +\infty & \text{if } x\notin S.\end{cases}$$

We use the notation $d_S(x)$ or $\mathrm{dist}(x,S)$ to denote the distance from an $x\in\mathbb{R}^n$ to S, i.e., $d_S(x):=\inf_{y\in S}\|x-y\|$. Moreover, we use $P_S(x)$ to denote the points in S that are closest to x: note that $P_S(x)$ is a singleton set if S is, in addition, convex.

Finally, for an optimization problem $\min_{x\in\mathbb{R}^n}f(x)$, we use $\operatorname{Arg\,min}_x f(x)$ to denote the set consisting of all its minimizers. If $\operatorname{Arg\,min}_x f(x)$ turns out to be a singleton, we simply denote it as $\arg\min_x f(x)$.

2 Peaceman-Rachford splitting for structured nonconvex problems

Recall that the class of problems we consider is

$$\min_u \ f(u) + g(u), \tag{5}$$

where f and g are proper closed possibly nonconvex functions. As discussed in the introduction, even in the case when both f and g are convex, typically one would need f (or g) to be strongly convex to guarantee convergence of the PR splitting method. Moreover, we recall that the Lipschitz differentiability of f played an important role in the recent convergence analysis of the closely related DR splitting method in [25] for (5) in the nonconvex settings. Motivated by these, we make the following blanket assumption on f throughout this paper.

Assumption 1 (**Blanket assumption on** f). *The function f is strongly convex with a strong convexity modulus at least $\sigma > 0$, and is Lipschitz differentiable so that ∇f has a Lipschitz continuity modulus at most $L > 0$.*

Notice that the proximal mapping $\mathrm{prox}_{\gamma f}(z)$ of a strongly convex function f is well defined for any $\gamma > 0$ at any point z. Thus, in order for the iterates in (2) to be well defined, we only need to make additionally the following blanket assumption on g in this paper.

Assumption 2 (**Blanket assumption on** g). *The function g is proper closed with a nonempty proximal mapping $\mathrm{prox}_{\gamma g}(z)$ for any z and for the $\gamma > 0$ we use in the algorithm.*

Under the blanket assumptions, we consider the following adaptation of the PR splitting method to solve the possibly nonconvex problem (5), which can be easily shown to be equivalent to (2) in the case when f and g are convex (so that the proximal mappings are single-valued).

PR splitting method

Step 0. Input x^0 and $\gamma > 0$.

Step 1. Set

$$\begin{cases} y^{t+1} = \arg\min_y \left\{ f(y) + \frac{1}{2\gamma}\|y - x^t\|^2 \right\}, \\ z^{t+1} \in \mathrm{Arg}\min_z \left\{ g(z) + \frac{1}{2\gamma}\|2y^{t+1} - x^t - z\|^2 \right\}, \\ x^{t+1} = x^t + 2(z^{t+1} - y^{t+1}). \end{cases} \tag{6}$$

Step 2. If a termination criterion is not met, go to Step 1.

Our convergence analysis follows a similar line of arguments (with some intricate modifications) for showing convergence for the Douglas-Rachford splitting method as in our recent work [25], and has to make extensive use of the following merit function:

$$\begin{aligned} \mathfrak{P}_\gamma(y, z, x) &:= f(y) + g(z) - \frac{3}{2\gamma}\|y - z\|^2 + \frac{1}{\gamma}\langle x - y, z - y\rangle \\ &= \mathfrak{D}_\gamma(y, z, x) - \frac{1}{\gamma}\|y - z\|^2, \end{aligned} \tag{7}$$

where $\mathfrak{D}_\gamma$ is the so-called Douglas-Rachford merit function given by $\mathfrak{D}_\gamma(y,z,x) = f(y) + g(z) - \frac{1}{2\gamma}\|y-z\|^2 + \frac{1}{\gamma}\langle x-y, z-y\rangle$ (see [25, Definition 2.1]), motivated by [29, Eq. 35].

Before proceeding, we make two important observations. First, it is not hard to see that the merit function $\mathfrak{P}_\gamma$ can alternatively be written as

$$\begin{aligned}\mathfrak{P}_\gamma(y,z,x) &= f(y) + g(z) + \frac{1}{2\gamma}\|2y-z-x\|^2 - \frac{1}{2\gamma}\|x-y\|^2 - \frac{2}{\gamma}\|y-z\|^2 \\ &= f(y) + g(z) + \frac{1}{2\gamma}(\|x-y\|^2 - \|x-z\|^2 - 2\|y-z\|^2),\end{aligned} \tag{8}$$

where the first relation follows from the elementary relation $\langle u, v\rangle = \frac{1}{2}(\|u+v\|^2 - \|u\|^2 - \|v\|^2)$ applied with $u = x-y$ and $v = z-y$ in (7), while the second relation is obtained by using the elementary relation $\langle u, v\rangle = \frac{1}{2}(\|u\|^2 + \|v\|^2 - \|u-v\|^2)$ in (7) with $u = x-y$ and $v = z-y$. We will make use of these equivalent formulations in the convergence analysis. Second, we note by using the optimality conditions for the y and z-updates in (6) that:

$$\begin{aligned}0 &= \nabla f(y^{t+1}) + \frac{1}{\gamma}(y^{t+1} - x^t), \\ 0 &\in \partial g(z^{t+1}) + \frac{1}{\gamma}(z^{t+1} - y^{t+1}) - \frac{1}{\gamma}(y^{t+1} - x^t),\end{aligned} \tag{9}$$

where we made use of the subdifferential calculus rule [31, Exercise 8.8]. Consequently, for all $t \geq 1$,

$$0 \in \nabla f(y^t) + \partial g(z^t) + \frac{1}{\gamma}(z^t - y^t). \tag{10}$$

To establish convergence and characterize the cluster point of the sequence generated, we will subsequently show that $\lim_{t\to\infty}\|z^t - y^t\| = 0$ and that g is "continuous" at the cluster point along the sequence generated.

We are now ready to state and prove a convergence result for the PR splitting method (6). We would like to point out that our proof is following exactly the same line of arguments as [25, Theorem 1]. However, there are two crucial differences. First, we now make use of the merit function (7) in place of the Douglas-Rachford merit function. Second, as we will see in the upper estimate in (20), the factor of γ in the denominator is canceled, and thus the strong convexity modulus σ comes into play in establishing the non-increasing property of the sequence $\{\mathfrak{P}_\gamma(y^t, z^t, x^t)\}_{t\geq 1}$.

Theorem 1 (**Global subsequential convergence**). *Suppose that $3\sigma > 2L$ and the parameter γ is chosen so that*

$$0 < \gamma < \frac{3\sigma - 2L}{L^2}. \tag{11}$$

Then the sequence $\{\mathfrak{P}_\gamma(y^t, z^t, x^t)\}_{t\geq 1}$ is nonincreasing. Moreover, if a cluster point (y^*, z^*, x^*) of the sequence exists, then we have

$$\lim_{t\to\infty}\|x^{t+1} - x^t\| = 2\lim_{t\to\infty}\|z^{t+1} - y^{t+1}\| = 0, \tag{12}$$

the cluster point satisfies $z^* = y^*$, and

$$0 \in \nabla f(z^*) + \partial g(z^*).$$

Remark 1. *We note that the condition $3\sigma > 2L$ indicates that this convergence result can only be applied when f has a relatively large strong convexity modulus, i.e., when $\sigma > \frac{2}{3}L$. It seems restrictive at first glance, but we will demonstrate in the next section how this theorem can be*

applied in a wide range of problems that do not explicitly contain a strongly convex part in the objective. Specifically, we will show that the method can be suitably applied to minimizing a coercive function $F + G$, where G is a proper closed function and F is convex Lipschitz differentiable but not *necessarily strongly convex; see Corollary 1.*

Proof. We study the behavior of $\mathfrak{P}_\gamma$ along the sequence generated from the PR splitting method. First, using (7) and the definition of the x-update, we see that

$$\mathfrak{P}_\gamma(y^{t+1}, z^{t+1}, x^{t+1}) - \mathfrak{P}_\gamma(y^{t+1}, z^{t+1}, x^t) = \frac{1}{\gamma}\langle x^{t+1} - x^t, z^{t+1} - y^{t+1}\rangle = \frac{1}{2\gamma}\|x^{t+1} - x^t\|^2. \quad (13)$$

Second, making use of the first relation in (8) and the definition of z^{t+1} as a minimizer, we have

$$\begin{aligned}
&\mathfrak{P}_\gamma(y^{t+1}, z^{t+1}, x^t) - \mathfrak{P}_\gamma(y^{t+1}, z^t, x^t) \\
&= g(z^{t+1}) + \frac{1}{2\gamma}\|2y^{t+1} - z^{t+1} - x^t\|^2 - \frac{2}{\gamma}\|y^{t+1} - z^{t+1}\|^2 \\
&\quad - g(z^t) - \frac{1}{2\gamma}\|2y^{t+1} - z^t - x^t\|^2 + \frac{2}{\gamma}\|y^{t+1} - z^t\|^2 \\
&\leq \frac{2}{\gamma}\left(\|y^{t+1} - z^t\|^2 - \|y^{t+1} - z^{t+1}\|^2\right) = \frac{2}{\gamma}\left(\|y^{t+1} - z^t\|^2 - \frac{1}{4}\|x^{t+1} - x^t\|^2\right),
\end{aligned} \quad (14)$$

where the last relation is due to the definition of x^{t+1}. Consequently, summing (13) and (14), we have

$$\mathfrak{P}_\gamma(y^{t+1}, z^{t+1}, x^{t+1}) - \mathfrak{P}_\gamma(y^{t+1}, z^t, x^t) \leq \frac{2}{\gamma}\|y^{t+1} - z^t\|^2. \quad (15)$$

Next, making use of the second relation in (8), we see that

$$\begin{aligned}
&\mathfrak{P}_\gamma(y^{t+1}, z^t, x^t) - \mathfrak{P}_\gamma(y^t, z^t, x^t) \\
&= f(y^{t+1}) + \frac{1}{2\gamma}\|x^t - y^{t+1}\|^2 - f(y^t) - \frac{1}{2\gamma}\|x^t - y^t\|^2 - \frac{1}{\gamma}\|y^{t+1} - z^t\|^2 + \frac{1}{\gamma}\|y^t - z^t\|^2 \\
&\leq -\frac{1}{2}\left(\frac{1}{\gamma} + \sigma\right)\|y^{t+1} - y^t\|^2 - \frac{1}{\gamma}\|y^{t+1} - z^t\|^2 + \frac{1}{\gamma}\|y^t - z^t\|^2,
\end{aligned} \quad (16)$$

where, in the last inequality, we used the definition of y^{t+1} as a minimizer and the strong convexity of the objective in the minimization problem that defines the y-update. Combining (16) with (15) gives further that

$$\mathfrak{P}_\gamma(y^{t+1}, z^{t+1}, x^{t+1}) - \mathfrak{P}_\gamma(y^t, z^t, x^t) \leq -\frac{1}{2}\left(\frac{1}{\gamma} + \sigma\right)\|y^{t+1} - y^t\|^2 + \frac{1}{\gamma}\|y^{t+1} - z^t\|^2 + \frac{1}{\gamma}\|y^t - z^t\|^2. \quad (17)$$

To further upper estimate (17), observe from the first relation in (9) that

$$\nabla f(y^{t+1}) = \frac{1}{\gamma}(x^t - y^{t+1}).$$

Since f is strongly convex with modulus $\sigma > 0$ by assumption, we see that for all $t \geq 1$,

$$\begin{aligned}
&\left\langle \frac{1}{\gamma}(x^t - y^{t+1}) - \frac{1}{\gamma}(x^{t-1} - y^t), y^{t+1} - y^t \right\rangle \geq \sigma\|y^{t+1} - y^t\|^2 \\
&\Longrightarrow \langle x^t - x^{t-1}, y^{t+1} - y^t\rangle \geq (1 + \gamma\sigma)\|y^{t+1} - y^t\|^2.
\end{aligned}$$

Thus, making use of the definition of x^t and the above relation, we obtain further that

$$\begin{aligned}
\|y^{t+1} - z^t\|^2 &= \|y^{t+1} - y^t + y^t - z^t\|^2 = \left\|y^{t+1} - y^t - \frac{1}{2}(x^t - x^{t-1})\right\|^2 \\
&= \|y^{t+1} - y^t\|^2 - \langle y^{t+1} - y^t, x^t - x^{t-1}\rangle + \frac{1}{4}\|x^t - x^{t-1}\|^2 \\
&\leq -\gamma\sigma\|y^{t+1} - y^t\|^2 + \frac{1}{4}\|x^t - x^{t-1}\|^2.
\end{aligned} \quad (18)$$

In addition, observe also from the definition of the x-update, the first relation in (9) and the Lipschitz continuity of ∇f that for $t \geq 1$

$$2\|y^t - z^t\| = \|x^t - x^{t-1}\| \leq (1+\gamma L)\|y^{t+1} - y^t\|. \tag{19}$$

Combining (18), (19) with (17), we conclude that for any $t \geq 1$

$$\begin{aligned} \mathfrak{P}_\gamma(y^{t+1}, z^{t+1}, x^{t+1}) - \mathfrak{P}_\gamma(y^t, z^t, x^t) &\leq \frac{1}{2\gamma}\left((1+\gamma L)^2 - 3\gamma\sigma - 1\right)\|y^{t+1} - y^t\|^2 \\ &= \frac{1}{2}\left(-3\sigma + 2L + \gamma L^2\right)\|y^{t+1} - y^t\|^2. \end{aligned} \tag{20}$$

By our choice of γ, $-3\sigma + 2L + \gamma L^2 < 0$. From this we see immediately that $\{\mathfrak{P}_\gamma(y^t, z^t, x^t)\}$ is nonincreasing. Summing (20) from $t = 1$ to $N - 1 \geq 1$, we obtain that

$$\mathfrak{P}_\gamma(y^N, z^N, x^N) - \mathfrak{P}_\gamma(y^1, z^1, x^1) \leq \frac{1}{2}\left(-3\sigma + 2L + \gamma L^2\right)\sum_{t=1}^{N-1}\|y^{t+1} - y^t\|^2. \tag{21}$$

Using this, the closedness of $\mathfrak{P}_\gamma$ and the existence of cluster points, we conclude immediately from (21) that $\lim_{t\to\infty}\|y^{t+1} - y^t\| = 0$. Combining this with (19), we conclude that (12) holds. Furthermore, combining these with the third relation in (6), we obtain further that $\lim_{t\to\infty}\|z^{t+1} - z^t\| = 0$.

Consequently, if (y^*, z^*, x^*) is a cluster point of $\{(y^t, z^t, x^t)\}$ with a convergent subsequence $\{(y^{t_j}, z^{t_j}, x^{t_j})\}$ such that $\lim_{j\to\infty}(y^{t_j}, z^{t_j}, x^{t_j}) = (y^*, z^*, x^*)$, then we must have

$$\lim_{j\to\infty}(y^{t_j}, z^{t_j}, x^{t_j}) = \lim_{j\to\infty}(y^{t_j-1}, z^{t_j-1}, x^{t_j-1}) = (y^*, z^*, x^*). \tag{22}$$

Since z^t is a minimizer of the subproblem,

$$g(z^t) + \frac{1}{2\gamma}\|2y^t - z^t - x^{t-1}\|^2 \leq g(z^*) + \frac{1}{2\gamma}\|2y^t - z^* - x^{t-1}\|^2.$$

Taking limit along the convergent subsequence and using (22) yields

$$\limsup_{j\to\infty} g(z^{t_j}) \leq g(z^*).$$

Conversely, we have $\liminf_{j\to\infty} g(z^{t_j}) \geq g(z^*)$ by the lower semicontinuity of g. Thus,

$$\lim_{j\to\infty} g(z^{t_j}) = g(z^*). \tag{23}$$

Using (4), (12), (23) and passing to the limit in (10) along the convergent subsequence above, we conclude that the cluster point gives a stationary point of (5), i.e., $y^* = z^*$ and

$$0 \in \nabla f(z^*) + \partial g(z^*).$$

This completes the proof. □

In the next theorem, we study sufficient conditions to guarantee boundedness of the sequence generated from the PR splitting method. Thus, a cluster point will necessarily exist under these conditions.

Theorem 2 (**Boundedness of sequence**). *Suppose that $3\sigma > 2L$ and the γ is chosen to satisfy* (11). *Suppose in addition that $f + g$ is coercive, i.e., $\liminf_{\|u\|\to\infty}(f+g)(u) = \infty$. Then the sequence $\{(y^t, z^t, x^t)\}$ generated from* (6) *is bounded.*

Proof. Recall from Theorem 1 that the merit function is nonincreasing along the sequence generated from (6). In particular,

$$\mathfrak{P}_\gamma(y^t, z^t, x^t) \leq \mathfrak{P}_\gamma(y^1, z^1, x^1) \tag{24}$$

whenever $t \geq 1$, where

$$\mathfrak{P}_\gamma(y^t, z^t, x^t) = f(y^t) + g(z^t) - \frac{1}{2\gamma}\|x^t - z^t\|^2 + \frac{1}{2\gamma}\|x^t - y^t\|^2 - \frac{1}{\gamma}\|y^t - z^t\|^2 \tag{25}$$

from the second relation in (8). Next, recall from the definition of x-update that $x^t = x^{t-1} + 2(z^t - y^t)$, which together with the first relation in (9) gives

$$\nabla f(y^t) = \frac{1}{\gamma}(x^{t-1} - y^t) = \frac{1}{\gamma}([x^t - z^t] - [z^t - y^t]). \tag{26}$$

Moreover, for the function f whose gradient is Lipschitz continuous with modulus L, we have

$$f(z^t) \leq f(y^t) + \langle \nabla f(y^t), z^t - y^t \rangle + \frac{L}{2}\|z^t - y^t\|^2. \tag{27}$$

Combining these with (25) and (24), we see further that

$$\begin{aligned}
&\mathfrak{P}_\gamma(y^1, z^1, x^1) \geq f(y^t) + g(z^t) - \frac{1}{2\gamma}\|x^t - z^t\|^2 + \frac{1}{2\gamma}\|x^t - y^t\|^2 - \frac{1}{\gamma}\|y^t - z^t\|^2 \\
&\geq f(z^t) + g(z^t) - \langle \nabla f(y^t), z^t - y^t \rangle - \frac{1}{2\gamma}\|x^t - z^t\|^2 + \frac{1}{2\gamma}\|x^t - y^t\|^2 - \left(\frac{L}{2} + \frac{1}{\gamma}\right)\|y^t - z^t\|^2 \\
&= f(z^t) + g(z^t) - \frac{1}{\gamma}\langle x^t - z^t, z^t - y^t \rangle - \frac{1}{2\gamma}\|x^t - z^t\|^2 + \frac{1}{2\gamma}\|x^t - y^t\|^2 - \frac{L}{2}\|y^t - z^t\|^2 \\
&= f(z^t) + g(z^t) + \frac{1}{2}\left(\frac{1}{\gamma} - L\right)\|y^t - z^t\|^2,
\end{aligned} \tag{28}$$

where the second inequality follows from (27), the first equality follows from (26), while the last equality follows from the elementary relation $\langle u, v \rangle = \frac{1}{2}(\|u+v\|^2 - \|u\|^2 - \|v\|^2)$ applied to $u = x^t - z^t$ and $v = z^t - y^t$. From (28), the coerciveness of $f+g$ and the fact that $\gamma < \frac{3\sigma - 2L}{L^2} \leq \frac{1}{L}$, we conclude that $\{z^t\}$ and $\{y^t\}$ are bounded. The boundedness of $\{x^t\}$ now follows from these and the first relation in (9). This completes the proof. □

Remark 2 (**Comments on the proof of Theorem 2**). (i) *The technique of using* (27) *for establishing* (28) *was also used previously in [20, Lemma 3.3] for showing that the augmented Lagrangian function is bounded below along the sequence generated from the alternating direction method of multipliers for a special class of problems. Here, we applied the technique to the new merit function* $\mathfrak{P}_\gamma$.

(ii) *The same technique used here can be applied to establishing the boundedness of the sequence generated by the DR splitting method studied in [25] under a condition which is slightly weaker than the one used in [25]. In fact, one can show that, the DR splitting method in [25] generates a bounded sequence under the blanket assumptions of* f *and* g *in [25, Section 3], the condition that* $f + g$ *is coercive and the choice of parameter specified in [25, Theorem 4]* [1].

To see this, recall that for the DR splitting method, we also have $\nabla f(y^t) = \frac{1}{\gamma}(x^{t-1} - y^t)$ *but have* $x^t = x^{t-1} + (z^t - y^t)$ *instead of the third relation in* (6). *Thus,* $\nabla f(y^t) = \frac{1}{\gamma}(x^t - z^t)$

[1] This slightly improves [25, Theorem 4] because [25, Theorem 4] assumed a slightly stronger condition that f and g are bounded below and one of them is coercive.

and we have the following estimate for the DR merit function, making use of (27)*:*

$$\begin{aligned}
&\mathfrak{D}_\gamma(y^t, z^t, x^t) = f(y^t) + g(z^t) - \frac{1}{2\gamma}\|x^t - z^t\|^2 + \frac{1}{2\gamma}\|x^t - y^t\|^2 \\
&\geq f(z^t) + g(z^t) - \langle \nabla f(y^t), z^t - y^t\rangle - \frac{L}{2}\|z^t - y^t\|^2 - \frac{1}{2\gamma}\|x^t - z^t\|^2 + \frac{1}{2\gamma}\|x^t - y^t\|^2 \\
&= f(z^t) + g(z^t) - \frac{1}{\gamma}\langle x^t - z^t, z^t - y^t\rangle - \frac{L}{2}\|z^t - y^t\|^2 - \frac{1}{2\gamma}\|x^t - z^t\|^2 + \frac{1}{2\gamma}\|x^t - y^t\|^2 \\
&= f(z^t) + g(z^t) + \frac{1}{2}\left(\frac{1}{\gamma} - L\right)\|y^t - z^t\|^2,
\end{aligned}$$

where the last equality follows from the elementary relation $\langle u, v\rangle = \frac{1}{2}(\|u+v\|^2 - \|u\|^2 - \|v\|^2)$ *applied to* $u = x^t - z^t$ *and* $v = z^t - y^t$*. The boundedness of the sequence can then be deduced under the choice of* γ *in [25, Theorem 4], which guarantees* $\gamma < \frac{1}{L}$*, and the assumption that* $f + g$ *is coercive.*

As in [24, Theorem 4] and [25, Theorem 2], one can also show that the whole sequence generated is convergent under the additional assumption that $\mathfrak{P}_\gamma(y, z, x)$ is a KL function.[2] To this end, note that for any $t \geq 1$, we have from (7) and the third relation in (6) that

$$\nabla_x \mathfrak{P}_\gamma(y^t, z^t, x^t) = \frac{1}{\gamma}(z^t - y^t) = \frac{1}{2\gamma}(x^t - x^{t-1}). \tag{29}$$

Moreover, using the second relation in (8), one can obtain

$$\nabla_y \mathfrak{P}_\gamma(y^t, z^t, x^t) = \nabla f(y^t) + \frac{1}{\gamma}(y^t - x^t) - \frac{2}{\gamma}(y^t - z^t) = \frac{1}{\gamma}(x^{t-1} - x^t) - \frac{2}{\gamma}(y^t - z^t) = 0 \tag{30}$$

where the second equality follows from the first relation in (9), and the last equality follows again from the third relation in (6). Finally, using the second relation in (8), one can compute that

$$\begin{aligned}
&\partial_z \mathfrak{P}_\gamma(y^t, z^t, x^t) = \partial g(z^t) - \frac{1}{\gamma}(z^t - x^t) - \frac{2}{\gamma}(z^t - y^t) \\
&= \partial g(z^t) + \frac{1}{\gamma}(z^t - y^t) - \frac{1}{\gamma}(y^t - x^{t-1}) - \frac{1}{\gamma}(z^t - y^t) + \frac{1}{\gamma}(y^t - x^{t-1}) - \frac{1}{\gamma}(z^t - x^t) - \frac{2}{\gamma}(z^t - y^t) \\
&\ni -\frac{4}{\gamma}(z^t - y^t) + \frac{1}{\gamma}(x^t - x^{t-1}) = -\frac{1}{\gamma}(x^t - x^{t-1}),
\end{aligned} \tag{31}$$

where the inclusion follows from the second relation in (9) and the last equality follows from the third relation in (6). Consequently, by combining (29), (30), (31) and (19), we see the existence of $\kappa > 0$ so that

$$\operatorname{dist}(0, \partial \mathfrak{P}_\gamma(y^t, z^t, x^t)) \leq \kappa \|y^{t+1} - y^t\|.$$

Using this, (20) and following the arguments as in the proof of [25, Theorem 2], it is not hard to prove the following result. We omit the detailed proof here.

Theorem 3 (**Global convergence of the whole sequence**)**.** *Suppose that* $3\sigma > 2L$*, the parameter* $\gamma > 0$ *is chosen as in* (11) *and that the sequence* $\{(y^t, z^t, x^t)\}$ *generated from* (6) *has a cluster point* (y^*, z^*, x^*)*. Suppose also that* $\mathfrak{P}_\gamma$ *is a KL function. Then the whole sequence* $\{(y^t, z^t, x^t)\}$ *is convergent.*

[2] We refer the readers to, for example, [1, 2, 7, 8], for the definition and examples of KL functions. In particular, if f and g are proper closed semi-algebraic functions, then $\mathfrak{P}_\gamma$ is a KL function for any $\gamma > 0$.

As we have seen from Theorems 1 and 2, our convergence analysis of the PR splitting method requires that the nonconvex objective function can be decomposed as $f + g$ where f is strongly convex. It should be noted that if the strong convexity assumption on f is dropped, then the sequence generated is not necessarily converging to/clustering at a stationary point even when g is also convex. On the other hand, in the next section, we will demonstrate how the method can be suitably applied to minimizing a coercive function $F + G$, where G is a proper closed function and F is convex Lipschitz differentiable but *not* necessarily strongly convex.

3 Peaceman-Rachford splitting methods for nonconvex problems with non-strongly convex decomposition

In many applications, the underlying optimization problem can be formulated as

$$\min_u \quad F(u) + G(u) \tag{32}$$

where $F+G$ is *coercive*, F is a convex smooth function with a Lipschitz continuous gradient whose modulus is at most $L_F > 0$, and G is a proper and closed function with a nonempty proximal mapping $\text{prox}_{\tau G}(z)$ for any z and any $\tau > 0$. For example, when F is the least squares loss function for linear regression and G is the indicator function of the ℓ_1 norm ball, the problem (32) reduces to the LASSO [32]. This and various related (possibly nonconvex) models have been studied extensively in the statistical literature; see, for example, [2, 6, 11, 17, 22]. We will also provide more concrete examples and simulation results later in Section 4.

In view of the structure of (32), a natural way of applying a splitting method would be to set $f(y) = F(y)$ and $g(z) = G(z)$. However, since this choice of f is not strongly convex, our convergence theory in Section 2 cannot be applied to deducing convergence of the resulting PR splitting method.

Thus, we consider an alternative way of splitting the objective in order to obtain a strongly convex f. To this end, we start by fixing any $\alpha > 0$ and defining $f(y) = F(y) + \frac{\alpha}{2}\|y\|^2$, $g(z) = G(z) - \frac{\alpha}{2}\|z\|^2$. Then ∇f is Lipschitz continuous with a modulus at most $L = L_F + \alpha$, and f is strongly convex with modulus at least $\sigma = \alpha$. Thus, one only needs to pick $\alpha > 2L_F$ so that $3\sigma > 2L$. Let $\alpha = \beta L_F$ for some $\beta > 2$. Then the upper bound of γ in (11) is given by

$$\frac{\alpha - 2L_F}{(L_F + \alpha)^2} = \frac{\beta - 2}{(\beta + 1)^2 L_F}.$$

Consequently, if we set

$$f(y) = F(y) + \frac{\beta L_F}{2}\|y\|^2 \text{ and } g(z) = G(z) - \frac{\beta L_F}{2}\|z\|^2,$$

then we can pick $0 < \gamma < \frac{\beta-2}{(\beta+1)^2 L_F}$.[3] Moreover, for this choice of γ, the Assumption 2 is satisfied for the above choice of g. Hence, it follows from Theorem 2 that the sequence generated by applying the PR splitting method to this pair of f and g is bounded, and then any cluster point gives a stationary point of (32), according to Theorem 1. For concreteness and easy reference for our subsequent discussion, we present this algorithm explicitly below:

[3]One natural choice of β is to set $\beta = 5$ so that $\max_{\beta>2} \frac{\beta-2}{(\beta+1)^2 L_F} = \frac{1}{12 L_F}$ is attained. However, we discover in our numerical experiments that a smaller $\beta > 2$ coupled with a suitable heuristic for updating γ leads to faster convergence.

PR splitting method for (32)

Step 0. Input x^0, $\beta > 2$ and $\gamma \in \left(0, \frac{\beta-2}{(\beta+1)^2 L_F}\right)$.

Step 1. Set

$$\begin{cases} y^{t+1} = \arg\min\limits_{y} \left\{ F(y) + \dfrac{\beta L_F}{2}\|y\|^2 + \dfrac{1}{2\gamma}\|y - x^t\|^2 \right\}, \\ z^{t+1} \in \operatorname{Arg}\min\limits_{z} \left\{ G(z) - \dfrac{\beta L_F}{2}\|z\|^2 + \dfrac{1}{2\gamma}\|2y^{t+1} - x^t - z\|^2 \right\}, \\ x^{t+1} = x^t + 2(z^{t+1} - y^{t+1}). \end{cases} \tag{33}$$

Step 2. If a termination criterion is not met, go to Step 1.

To the best of our knowledge, the global convergence of the sequence generated from (33) is new, which we summarize below for concreteness.

Corollary 1. *Consider optimization problem* (32) *and let* $\{(y^t, z^t, x^t)\}$ *be the sequence generated from* (33)*. Then the sequence is bounded, and any cluster point* $(\bar{y}, \bar{z}, \bar{x})$ *would satisfy* $\bar{y} = \bar{z}$*, and* $\bar{z}$ *is a stationary point of* (32)*, that is,*

$$0 \in \nabla F(\bar{z}) + \partial G(\bar{z}).$$

Proof. We first note that since (33) is just (6) applied to $f(y) = F(y) + \frac{\beta L_F}{2}\|y\|^2$ and $g(z) = G(z) - \frac{\beta L_F}{2}\|z\|^2$, we obtain immediately from the above discussion and Theorem 1 that $\bar{y} = \bar{z}$ and $\bar{z}$ is a stationary point of (32) for any cluster point $(\bar{y}, \bar{z}, \bar{x})$. In addition, the objective function $f + g = F + G$ is coercive by assumption. The boundedness of the sequence $\{(y^t, z^t, x^t)\}$ now follows from Theorem 2. This completes the proof. □

3.1 Peaceman-Rachford splitting method for convex problems

In this subsection, we suppose in addition that the G in (32) is also convex. Hence, (32) is a convex problem. We first establish the following global (ergodic) complexity result for the sequence generated from (33). Similar kinds of complexity results have also been established for other primal-dual methods for convex optimization problems; see, for example, [33, Theorem 2]. We would like to emphasize that the PR splitting method we discuss here is different from the classical PR splitting method in the literature: we split the convex objective $F + G$ into the sum of a strongly convex function f and a possibly *nonconvex* function g, while the classical PR splitting method only admits splitting into a sum of convex functions.

Theorem 4 (**Global iteration complexity under convexity**)**.** *Consider optimization problem* (32) *with* G *being convex. Let* $\{(y^t, z^t, x^t)\}$ *be the sequence generated from* (33) *and* $(\bar{y}, \bar{z}, \bar{x})$ *be any cluster point of this sequence. Then,* $\bar{y} = \bar{z}$ *and* $\bar{z}$ *is a solution of* (32)*. Moreover, for any* $N \geq 1$*, we have*

$$F(\bar{z}^N) + G(\bar{z}^N) - F(\bar{z}) - G(\bar{z}) \leq \frac{1}{8\beta\gamma N L_F}\left(\frac{1}{\gamma} - \beta L_F\right)\|x^0 - \bar{x}\|^2, \tag{34}$$

where $\bar{z}^N := \frac{1}{N}\sum_{t=1}^{N} z^t$ *and*

$$\min_{0 \leq t \leq N}\{\|x^{t+1} - x^t\|\} = o\left(\frac{1}{\sqrt{N}}\right).$$

Proof. Since (32) is convex, we conclude that $\bar{z}$ is actually optimal. We now establish the inequality (34). First, from the first-order optimality conditions for the y and z-updates in (33), we have

$$\begin{aligned} -\left(\beta L_F + \frac{1}{\gamma}\right) y^{t+1} + \frac{1}{\gamma} x^t &= \nabla F(y^{t+1}), \\ \left(\beta L_F - \frac{1}{\gamma}\right) z^{t+1} - \frac{1}{\gamma} x^t + \frac{2}{\gamma} y^{t+1} &\in \partial G(z^{t+1}). \end{aligned} \tag{35}$$

Moreover, it is not hard to see from the definition of cluster point and (12) that (35) is also satisfied with $\bar{x}$ in place of x^t and $(\bar{y}, \bar{z})$ in place of (y^{t+1}, z^{t+1}). Write $w_e^t = w^t - \bar{w}$ for $w = x$, y or z for notational simplicity. We have from (35) (and its counterpart at $(\bar{y}, \bar{z}, \bar{x})$) and the monotonicity of convex subdifferentials that

$$\left\langle -\left(\beta L_F + \frac{1}{\gamma}\right) y_e^{t+1} + \frac{1}{\gamma} x_e^t, y_e^{t+1} \right\rangle \geq 0, \quad \left\langle \left(\beta L_F - \frac{1}{\gamma}\right) z_e^{t+1} - \frac{1}{\gamma} x_e^t + \frac{2}{\gamma} y_e^{t+1}, z_e^{t+1} \right\rangle \geq 0.$$

Summing these two relations and rearranging terms, we obtain that

$$\langle x_e^t, y^{t+1} - z^{t+1} \rangle + 2\langle y_e^{t+1}, z_e^{t+1} \rangle \geq (1 + \beta\gamma L_F)\|y_e^{t+1}\|^2 + (1 - \beta\gamma L_F)\|z_e^{t+1}\|^2. \tag{36}$$

Next, observe that

$$\begin{aligned} \langle x_e^t, y^{t+1} - z^{t+1} \rangle &= \frac{1}{2}\langle x_e^t, x^t - x^{t+1} \rangle = \frac{1}{4}(\|x_e^t\|^2 + \|x^t - x^{t+1}\|^2 - \|x_e^{t+1}\|^2) \\ &= \frac{1}{4}(\|x_e^t\|^2 - \|x_e^{t+1}\|^2) + \|z^{t+1} - y^{t+1}\|^2 \\ &= \frac{1}{4}(\|x_e^t\|^2 - \|x_e^{t+1}\|^2) + \|z_e^{t+1}\| + \|y_e^{t+1}\|^2 - 2\langle y_e^{t+1}, z_e^{t+1} \rangle, \end{aligned} \tag{37}$$

where the first and third equalities follow from the third relation in (33), the second equality follows from the elementary relation $\langle u, v \rangle = \frac{1}{2}(\|u\|^2 + \|v\|^2 - \|u - v\|^2)$ as applied to $u = x_e^t$ and $v = x^t - x^{t+1}$. Combining (37) with (36), we see further that

$$\frac{1}{4}\|x_e^t\|^2 - \frac{1}{4}\|x_e^{t+1}\|^2 \geq \beta\gamma L_F(\|y_e^{t+1}\|^2 - \|z_e^{t+1}\|^2) \tag{38}$$

Next, using the fact that ∇F is Lipschitz continuous with modulus at most L_F, we have

$$F(z^{t+1}) \leq F(y^{t+1}) + \langle \nabla F(y^{t+1}), z^{t+1} - y^{t+1} \rangle + \frac{L_F}{2}\|z^{t+1} - y^{t+1}\|^2. \tag{39}$$

From this we see further that

$$
\begin{aligned}
&F(z^{t+1}) + G(z^{t+1}) - F(\bar{z}) - G(\bar{z}) \\
&\leq F(y^{t+1}) - F(\bar{y}) + G(z^{t+1}) - G(\bar{z}) + \langle \nabla F(y^{t+1}), z^{t+1} - y^{t+1} \rangle + \frac{L_F}{2}\|z^{t+1} - y^{t+1}\|^2 \\
&\leq \langle \nabla F(y^{t+1}), y_e^{t+1} \rangle + \left\langle \left(\beta L_F - \frac{1}{\gamma}\right) z^{t+1} - \frac{1}{\gamma} x^t + \frac{2}{\gamma} y^{t+1}, z_e^{t+1} \right\rangle \\
&\quad + \langle \nabla F(y^{t+1}), z^{t+1} - y^{t+1} \rangle + \frac{L_F}{2}\|z^{t+1} - y^{t+1}\|^2 \\
&= \left\langle \nabla F(y^{t+1}) + \left(\beta L_F - \frac{1}{\gamma}\right) z^{t+1} - \frac{1}{\gamma} x^t + \frac{2}{\gamma} y^{t+1}, z_e^{t+1} \right\rangle + \frac{L_F}{2}\|z^{t+1} - y^{t+1}\|^2 \\
&= \left\langle - \left(\beta L_F - \frac{1}{\gamma}\right) y^{t+1} + \left(\beta L_F - \frac{1}{\gamma}\right) z^{t+1}, z_e^{t+1} \right\rangle + \frac{L_F}{2}\|z^{t+1} - y^{t+1}\|^2 \\
&= \left(\frac{1}{\gamma} - \beta L_F\right) \langle y^{t+1} - z^{t+1}, z_e^{t+1} \rangle + \frac{L_F}{2}\|z^{t+1} - y^{t+1}\|^2 \\
&= \frac{1}{2}\left(\frac{1}{\gamma} - \beta L_F\right) (\|y_e^{t+1}\|^2 - \|z_e^{t+1}\|^2) + \frac{1}{2}\left((1+\beta)L_F - \frac{1}{\gamma}\right)\|z^{t+1} - y^{t+1}\|^2 \\
&\leq \frac{1}{2}\left(\frac{1}{\gamma} - \beta L_F\right) (\|y_e^{t+1}\|^2 - \|z_e^{t+1}\|^2) \leq \frac{1}{8\beta\gamma L_F}\left(\frac{1}{\gamma} - \beta L_F\right) (\|x_e^t\|^2 - \|x_e^{t+1}\|^2),
\end{aligned}
\tag{40}
$$

where: the first inequality follows from (39) and the fact that $\bar{z} = \bar{y}$; the second inequality follows from the subdifferential inequalities applied to F and G at the points y^{t+1} and z^{t+1} respectively, and also the second relation in (35); the second equality follows from the first relation in (35); the fourth equality follows from the elementary relation $\langle u, v \rangle = \frac{1}{2}(\|u+v\|^2 - \|u\|^2 - \|v\|^2)$ as applied to $u = z_e^{t+1}$ and $v = y^{t+1} - z^{t+1}$; the second last inequality follows from the fact that $0 < \gamma < \frac{\beta - 2}{(\beta+1)^2 L_F}$ so that $(1+\beta)L_F - \frac{1}{\gamma} < 0$, while the last inequality follows from (38).

Summing both sides of (40) from $t = 0$ to $N - 1 \geq 0$ and using the convexity of $F + G$, we have

$$
\begin{aligned}
F(\bar{z}^N) + G(\bar{z}^N) - F(\bar{z}) - G(\bar{z}) &\leq \frac{1}{N}\sum_{t=0}^{N-1} (F(z^{t+1}) + G(z^{t+1}) - F(\bar{z}) - G(\bar{z})) \\
&\leq \frac{1}{8\beta\gamma N L_F}\left(\frac{1}{\gamma} - \beta L_F\right)\|x^0 - \bar{x}\|^2,
\end{aligned}
$$

where $\bar{z}^N$ is defined in the statement of the theorem. This proves (34).

Finally, observe from the last equality in (40) that for all $t \geq 1$

$$
\begin{aligned}
0 &\leq F(z^{t+1}) + G(z^{t+1}) - F(\bar{z}) - G(\bar{z}) \\
&\leq \frac{1}{2}\left(\frac{1}{\gamma} - \beta L_F\right) (\|y_e^{t+1}\|^2 - \|z_e^{t+1}\|^2) + \frac{1}{2}\left((1+\beta)L_F - \frac{1}{\gamma}\right)\|z^{t+1} - y^{t+1}\|^2,
\end{aligned}
$$

where the first inequality follows from the optimality of $\bar{z}$. Rearranging terms in the above relation, we see further that

$$
\left(\frac{1}{\gamma} - (1+\beta)L_F\right)\|z^{t+1} - y^{t+1}\|^2 \leq \left(\frac{1}{\gamma} - \beta L_F\right) (\|y_e^{t+1}\|^2 - \|z_e^{t+1}\|^2).
$$

Using this relation and the definition of the x-update, we obtain

$$
\begin{aligned}
\frac{1}{4}\sum_{t=0}^{N-1} \|x^{t+1} - x^t\|^2 = \sum_{t=0}^{N-1} \|z^{t+1} - y^{t+1}\|^2 &\leq \frac{\gamma}{1 - (1+\beta)\gamma L_F}\left(\frac{1}{\gamma} - \beta L_F\right) \sum_{t=0}^{N-1} (\|y_e^{t+1}\|^2 - \|z_e^{t+1}\|^2) \\
&\leq \frac{1}{4\beta L_F (1 - (1+\beta)\gamma L_F)}\left(\frac{1}{\gamma} - \beta L_F\right)\|x^0 - \bar{x}\|^2,
\end{aligned}
$$

where the last inequality is due to (38). Thus, $\sum_{t=0}^{+\infty} \|x^{t+1} - x^t\|^2 < +\infty$ and so, $\sum_{t=N}^{2N-1} \|x^{t+1} - x^t\|^2 \to 0$ as $N \to \infty$. Now consider $\alpha_N := \min_{0\le t\le N}\{\|x^{t+1} - x^t\|^2\}$ for all $N \ge 0$. Then, we have $\alpha_{N+1} \le \alpha_N$ for all $N \ge 0$ and,

$$N\,\alpha_{2N} \le \alpha_N + \ldots \alpha_{2N-1} \le \sum_{t=N}^{2N-1} \|x^{t+1} - x^t\|^2 \to 0.$$

This implies that $\alpha_N = o(1/N)$. Therefore, the conclusion follows. This completes the proof. □

Next, we show that the PR splitting method exhibits linear convergence in solving (32) if G is convex and $F + G$ is strongly convex. We note that, for the classical PR splitting method, linear convergence under strongly convexity is known; see [26, Remark 10 and Proposition 4]. As explained before, here we are considering a different PR splitting method.

Proposition 1. (Linear convergence under strong convexity) *Consider optimization problem* (32) *with G being convex. Suppose that $F + G$ is indeed strongly convex. Let $\{(y^t, z^t, x^t)\}$ be the sequence generated from* (33). *Then $\{(y^t, z^t, x^t)\}$ converges linearly to $(\bar{y}, \bar{z}, \bar{x})$ with $\bar{y} = \bar{z}$ and $\bar{z}$ being the unique optimal solution for* (32), *i.e., there exist $M > 0$ and $r \in (0, 1)$ such that for all $t \ge 1$,*

$$\max\{\|y^t - \bar{y}\|^2, \|z^t - \bar{z}\|^2, \|x^t - \bar{x}\|^2\} \le M\, r^t.$$

Proof. Let $(\bar{y}, \bar{z}, \bar{x})$ be any cluster point of the sequence $\{(y^t, z^t, x^t)\}$. As before, we write $w_e^t = w^t - \bar{w}$ for $w = x$, y or z for notational simplicity. From the preceding theorem $\bar{y} = \bar{z}$ and $\bar{z}$ is optimal for (32). Note that $F+G$ is strongly convex. Hence, the optimal solution of (32) exists and is unique. Consequently, the whole sequence $\{(y^t, z^t)\}$ converges to the *unique* limit $(\bar{z}, \bar{z})$, where $\bar{z}$ is the unique solution of (32). From this and (35) one can deduce that $\{x^t\}$ is also convergent, and hence, converges to $\bar{x}$. We next establish linear convergence.

Denote the strong convexity modulus of $F+G$ by σ_1. From (40), the strong convexity of $F+G$ and the fact that $\bar{z}$ is the solution of (32), we see that for all $t \ge 1$,

$$\frac{\sigma_1}{2}\|z_e^{t+1}\|^2 \le F(z^{t+1}) + G(z^{t+1}) - F(\bar{z}) - G(\bar{z}) \le C(\|x_e^t\|^2 - \|x_e^{t+1}\|^2), \tag{41}$$

where $C := \frac{1}{8\beta\gamma L_F}\left(\frac{1}{\gamma} - \beta L_F\right)$. Moreover, from the last inequality in (40), we have for all $t \ge 1$,

$$C_1(\|y_e^{t+1}\|^2 - \|z_e^{t+1}\|^2) \le C(\|x_e^t\|^2 - \|x_e^{t+1}\|^2),$$

where $C_1 = \frac{1}{2}\left(\frac{1}{\gamma} - \beta L_F\right)$. It then follows that

$$\|y_e^{t+1}\|^2 - \frac{C}{C_1}(\|x_e^t\|^2 - \|x_e^{t+1}\|^2) \le \|z_e^{t+1}\|^2.$$

This together with (41) gives us that for all $t \ge 1$,

$$\|y_e^{t+1}\|^2 \le \left(\frac{2C}{\sigma_1} + \frac{C}{C_1}\right)(\|x_e^t\|^2 - \|x_e^{t+1}\|^2). \tag{42}$$

On the other hand, note from the first relation in (35) that

$$-\left(\beta L_F + \frac{1}{\gamma}\right) y_e^{t+1} + \frac{1}{\gamma} x_e^t = \nabla F(y^{t+1}) - \nabla F(\bar{y}).$$

This together with the Lipschitz continuity of ∇F implies that

$$-\left(\beta L_F + \frac{1}{\gamma}\right)\|y_e^{t+1}\| + \frac{1}{\gamma}\|x_e^t\| \le \|\nabla F(y^{t+1}) - \nabla F(\bar{y})\| \le L_F\|y_e^{t+1}\|$$

and consequently, $\|x_e^t\| \le ((1+\beta)\gamma L_F + 1)\|y_e^{t+1}\|$. Thus, we obtain that, for all $t \ge 1$

$$\frac{1}{((1+\beta)\gamma L_F + 1)^2}\|x_e^t\|^2 \le \|y_e^{t+1}\|^2 \le \left(\frac{2C}{\sigma_1} + \frac{C}{C_1}\right)(\|x_e^t\|^2 - \|x_e^{t+1}\|^2).$$

This shows that there exists $r \in (0,1)$ such that

$$\|x_e^{t+1}\|^2 \le r\|x_e^t\|^2 \text{ for all } t \ge 1.$$

It follows that

$$\|x_e^t\|^2 \le \|x^0 - \bar{x}\|^2 r^t \text{ for all } t \ge 1.$$

Moreover, from (41) and (42), this further yields that, for all $t \ge 1$,

$$\|z_e^{t+1}\|^2 \le \frac{2C}{\sigma_1}\|x_e^t\|^2 \le \frac{2C\|x^0 - \bar{x}\|^2}{\sigma_1} r^t.$$

and

$$\|y_e^{t+1}\|^2 \le \left(\frac{2C}{\sigma_1} + \frac{C}{C_1}\right)\|x_e^t\|^2 \le \left(\frac{2C}{\sigma_1} + \frac{C}{C_1}\right)\|x^0 - \bar{x}\|^2 r^t.$$

Therefore, the conclusion follows. □

4 Applications

In this section, we apply the PR splitting method (33) to solving two important class of nonconvex optimization problems: constrained least squares problem and feasibility problems, based on our discussion in Section 3.

Constrained least squares problems. A common type of problems that arises in the area of statistics and machine learning is the following constrained least squares problem:

$$\min_{u \in D} \quad \tfrac{1}{2}\|Au - b\|^2, \tag{43}$$

where A is a linear map, b is a vector of suitable dimension, and D is a nonempty *compact* set that is not necessarily convex. See [23, 32] for concrete examples of (43).

The classical PR splitting method applied to (43) does not have a convergence guarantee. As an alternative, as discussed in Section 3, we can set $f(y) = \frac{1}{2}\|Ay - b\|^2 + \frac{\beta\lambda_{\max}(A^TA)}{2}\|y\|^2$ and $g(z) = \delta_D(z) - \frac{\beta\lambda_{\max}(A^TA)}{2}\|z\|^2$ and apply the PR splitting method accordingly.

We next discuss computation of the proximal mappings. We start with the proximal mapping of γg. From the definition, for each w, the proximal mapping gives the set of minimizers of

$$\min_{z \in D}\left\{-\frac{\beta\lambda_{\max}(A^TA)}{2}\|z\|^2 + \frac{1}{2\gamma}\|z - w\|^2\right\}.$$

It is clear that this set is given by $P_D\left(\frac{w}{1-\beta\lambda_{\max}(A^TA)\gamma}\right)$ since $\gamma < \frac{1}{\beta\lambda_{\max}(A^TA)}$. On the other hand, to compute the proximal mapping for γf, we consider the following optimization problem for each w

$$\min_y \left\{\frac{1}{2}\|Ay - b\|^2 + \frac{\beta\lambda_{\max}(A^TA)}{2}\|y\|^2 + \frac{1}{2\gamma}\|y - w\|^2\right\},$$

whose unique minimizer is given by

$$y = [(\beta\gamma\lambda_{\max}(A^TA) + 1)I + \gamma A^TA]^{-1}(w + \gamma A^Tb).$$

Thus, the PR splitting method for (43) can be stated as follows:

PR splitting method for (43)

Step 0. Input x^0, $\beta > 2$ and $\gamma \in \left(0, \frac{\beta-2}{(\beta+1)^2\lambda_{\max}(A^TA)}\right)$.

Step 1. Set

$$\begin{cases} y^{t+1} = [(\beta\gamma\lambda_{\max}(A^TA)+1)I + \gamma A^TA]^{-1}(x^t + \gamma A^Tb), \\ z^{t+1} \in P_D\left(\dfrac{2y^{t+1} - x^t}{1-\beta\lambda_{\max}(A^TA)\gamma}\right), \\ x^{t+1} = x^t + 2(z^{t+1} - y^{t+1}). \end{cases} \tag{44}$$

Step 2. If a termination criterion is not met, go to Step 1.

As a consequence of Corollary 1, we see that Algorithm (44) generates a bounded sequence such that any of its cluster point gives a stationary point of (43). We note that this global convergence result of (44) is new even when D is convex.

To illustrate our proposed approach, we now test the PR splitting method (44) on solving (43). We compare our algorithm against the DR splitting method in [25]. Our initialization and termination criteria for both algorithms are the same as in [25, Section 5]; both algorithms are initialized at the origin and terminated when

$$\frac{\max\{\|x^t - x^{t-1}\|, \|y^t - y^{t-1}\|, \|z^t - z^{t-1}\|\}}{\max\{\|x^{t-1}\|, \|y^{t-1}\|, \|z^{t-1}\|, 1\}} < tol \tag{45}$$

for some $tol > 0$. Note that, in general, the upper bound of γ in algorithm (44) might be too small in practical computation. Thus, following a technique used in [25, Section 5] for the DR splitting method, we adopt a heuristic for PR splitting method in our numerical simulation, which combines algorithm (44) with a specific update rule of the parameter γ. In particular, we set $\beta = 2.2$ and start with $\gamma = 0.93/(\beta\lambda_{\max}(A^TA))$. We then update γ as $\max\{\frac{\gamma}{2}, 0.9999\cdot\gamma_1\}$ whenever $\gamma > \gamma_1 := \frac{\beta-2}{(\beta+1)^2\lambda_{\max}(A^TA)}$ and the sequence satisfies either $\|y^t - y^{t-1}\| > \frac{1000}{t}$ or $\|y^t\| > 10^{10}$. Following a similar discussion as in [25, Remark 4], one can show that this heuristic leads to a bounded sequence which clusters at a stationary point of (43). On the other hand, for the DR splitting method, we use the same heuristics described in [25, Section 5] for updating γ but we consider three different initial γ's: $k\cdot\gamma_0$ for $k = 10$, 30 and 50, with $\gamma_0 = (\sqrt{\frac{3}{2}} - 1)/\lambda_{\max}(A^TA)$. These variants are denoted by DR_{10}, DR_{30} and DR_{50}, respectively.

In our first numerical experiment, we first randomly generate an $m \times n$ matrix A, a noise vector $\epsilon \in \mathbb{R}^m$, and also an $\widehat{x} \in \mathbb{R}^r$ with $r = \lceil \frac{m}{10} \rceil$, all with i.i.d. standard Gaussian entries. We further scale each column of A to have norm 1. Next, we generate a random sparse vector $\widetilde{x} \in \mathbb{R}^n$ by first setting $\widetilde{x} = 0$ and then assigning randomly r entries in $\widetilde{x}$ to be $\widehat{x}$. Finally, we set $b = A\widetilde{x} + 0.01\cdot\epsilon$ and $D = \{x \in \mathbb{R}^n : \|x\|_0 \le r,\ \|x\|_\infty \le 10^6\}$; here $\|x\|_0$ denotes the cardinality of x and $\|x\|_\infty$ is the ℓ_∞ norm of x.

We generate 50 random instances as described above for each pair of (m, n), where $m \in \{100, 200, 300, 400, 500\}$ and $n \in \{4000, 5000, 6000\}$. Our results are reported in Table 1, where we present the number of iterations and the function value at termination[4] averaged over the 50 instances. One can observe that the PR splitting method is faster than the DR splitting methods for larger m. Besides, the function values obtained by the PR splitting method are usually comparable with DR_{30}, worse than DR_{50} and better than DR_{10}.

[4]We choose $tol = 10^{-8}$, and we report $\frac{1}{2}\|Az^t - b\|^2$ for both methods.

Table 1: Comparing DR_{10}, DR_{30}, DR_{50} and PR splitting for constrained least squares problem on random instances.

Data		DR_{10}		DR_{30}		DR_{50}		PR	
m	n	iter	fval	iter	fval	iter	fval	iter	fval
100	4000	805	5.00e-01	225	2.67e-01	274	7.73e-02	324	3.17e-01
100	5000	962	6.43e-01	252	4.96e-01	291	2.06e-01	370	4.95e-01
100	6000	1137	6.18e-01	326	5.02e-01	301	2.53e-01	436	4.76e-01
200	4000	508	5.32e-01	172	4.74e-02	217	9.20e-03	185	7.59e-02
200	5000	624	5.78e-01	195	6.93e-02	234	9.10e-03	224	2.06e-01
200	6000	723	6.93e-01	220	1.60e-01	250	8.94e-03	281	1.77e-01
300	4000	415	1.41e-01	141	1.33e-02	184	1.31e-02	123	1.39e-02
300	5000	489	2.70e-01	154	1.39e-02	201	1.35e-02	150	1.42e-02
300	6000	567	5.20e-01	170	1.36e-02	215	1.32e-02	187	1.44e-02
400	4000	322	4.35e-02	124	1.78e-02	166	1.75e-02	91	1.79e-02
400	5000	406	9.08e-02	137	1.77e-02	179	1.75e-02	115	1.83e-02
400	6000	481	1.48e-01	148	1.82e-02	194	1.77e-02	140	1.85e-02
500	4000	258	2.53e-02	114	2.26e-02	160	2.23e-02	75	2.27e-02
500	5000	314	2.97e-02	124	2.20e-02	166	2.17e-02	92	2.22e-02
500	6000	406	4.05e-02	135	2.25e-02	178	2.22e-02	112	2.27e-02

We also perform experiments using real data. We consider four sets of real data for the A and b used in (43): leukemia data, lymph node status data, breast cancer prognosis data and colon tumor gene expression data. We use the leukemia data pre-processed in [34], that has 3501 genes and 72 samples. The lymph node status data we use are pre-processed in [14], with 4514 genes and 148 samples. The breast cancer prognosis data we use are pre-processed in [34], containing 4919 genes and 76 samples. Finally, we use the data pre-processed in [19] with 2000 genes and 62 samples for the colon tumor gene expression data.

Similar to [21, Section 3.3], for all the data, we first standardize A and b to make each column have mean 0 and variance 1, and then scale the columns of A to have unit norm. For the A and b thus constructed, we solve (43) with $D = \{x \in \mathbb{R}^n : \|x\|_0 \leq r,\ \|x\|_\infty \leq 10^6\}$ for $r = 10$, 20, 30 by the PR splitting method (44) and compare it with DR_{10}, DR_{30} and DR_{50}. Our numerical results are presented in Table 2,[5] where one can see that PR is slower than DR_{50} and faster than DR_{10}. Moreover, it usually outperforms DR_{30} in terms of function values, and its speed is comparable with DR_{30} for the Breast and the Colon data.

Table 2: Comparing DR_{10}, DR_{30}, DR_{50} and PR splitting on real data.

Data	r	DR_{10}		DR_{30}		DR_{50}		PR	
		iter	fval	iter	fval	iter	fval	iter	fval
Leukemia	10	8242	2.40e+00	1805	3.92e+00	1229	3.92e+00	3461	2.47e+00
	20	7890	2.32e+00	3727	6.09e-01	3065	5.81e-01	6608	3.05e-01
	30	12530	2.24e-01	5011	3.01e-01	2988	1.47e-01	8265	1.20e-01
Lymph	10	1345	2.93e+01	758	2.90e+01	496	2.90e+01	1297	2.76e+01
	20	5912	2.26e+01	1910	1.91e+01	895	1.73e+01	2529	1.84e+01
	30	9354	7.91e+00	1883	1.34e+01	939	1.44e+01	2089	8.27e+00
Breast	10	2338	1.28e+01	2705	9.33e+00	1095	8.40e+00	1656	1.33e+01
	20	14359	2.90e+00	2345	3.53e+00	2824	4.11e+00	2906	2.81e+00
	30	9905	6.96e-01	5162	1.33e+00	3802	7.50e-01	8241	9.58e-01
Colon	10	7072	8.08e+00	4313	8.08e+00	3352	8.08e+00	4463	8.08e+00
	20	14393	3.20e+00	7011	1.95e+00	9798	2.29e+00	6187	1.89e+00
	30	18361	7.17e-01	8952	6.45e-01	4922	7.26e-01	10937	1.33e+00

[5] We choose $tol = 10^{-5}$, and we report $\frac{1}{2}\|Az^t - b\|^2$ for both methods.

Feasibility problems. Another important problem in optimization is the feasibility problem [2–4, 9, 28]. We consider the following simple version: finding a point in the intersection of a nonempty closed convex set C and a nonempty *compact* set D. It is well known that this problem can be modeled via (32) by setting $F(u) = \frac{1}{2}d_C^2(u)$ and $G(u) = \delta_D(u)$; see, for example, [27]. For this choice of F, we have $L_F = 1$.

As before, it can be shown that the proximal mapping of γg is given by $P_D\left(\frac{w}{1-\beta\gamma}\right)$ since $\gamma < \frac{1}{\beta}$. We next compute the proximal mapping for γf in this case. From the definition, for each w, we consider the following optimization problem

$$v := \min_y \left\{ \frac{1}{2}d_C^2(y) + \frac{\beta}{2}\|y\|^2 + \frac{1}{2\gamma}\|y-w\|^2 \right\} = \min_{u\in C} \min_y \left\{ \frac{1}{2}\|y-u\|^2 + \frac{\beta}{2}\|y\|^2 + \frac{1}{2\gamma}\|y-w\|^2 \right\}. \tag{46}$$

Notice that the inner minimization on the right hand side is attained at

$$y = \frac{\gamma u + w}{(1+\beta)\gamma + 1}. \tag{47}$$

Plugging (47) back into the (46), we see further that

$$v = \frac{1}{((1+\beta)\gamma+1)^2} \min_{u\in C} \left\{ \frac{1}{2}\|(1+\beta\gamma)u - w\|^2 + \frac{\beta}{2}\|\gamma u + w\|^2 + \frac{\gamma}{2}\|u-(1+\beta)w\|^2 \right\}. \tag{48}$$

It is routine to show that the minimum in (48) is attained at

$$u = P_C\left(\frac{w}{1+\beta\gamma}\right).$$

Combining this with (47), the proximal mapping of γf at w is given by

$$\frac{\gamma P_C\left(\frac{w}{1+\beta\gamma}\right) + w}{(1+\beta)\gamma+1}.$$

Thus, the PR splitting method for (32) with $F(u) = \frac{1}{2}d_C^2(u)$ and $G(u) = \delta_D(u)$ can be described as follows:

PR splitting method for (32) **with** $F(u) = \frac{1}{2}d_C^2(u)$ **and** $G(u) = \delta_D(u)$

Step 0. Input x^0, $\beta > 2$ and $\gamma \in \left(0, \frac{\beta-2}{(\beta+1)^2}\right)$.

Step 1. Set

$$\begin{cases} y^{t+1} = \dfrac{\gamma P_C\left(\frac{x^t}{1+\beta\gamma}\right) + x^t}{(1+\beta)\gamma+1}, \\ z^{t+1} \in P_D\left(\dfrac{2y^{t+1}-x^t}{1-\beta\gamma}\right), \\ x^{t+1} = x^t + 2(z^{t+1}-y^{t+1}). \end{cases} \tag{49}$$

Step 2. If a termination criterion is not met, go to Step 1.

Similarly, as an immediate consequence of Corollary 1, we see that Algorithm (49) generates a bounded sequence such that any of its cluster point gives a stationary point of (32). We would like to point out that this global convergence result of (49) is new even when D is also convex.

As an illustration of our proposed approach, we now test the PR splitting method (49) on solving (32) with $F(u) = \frac{1}{2}d_C^2(u)$ and $G(u) = \delta_D(u)$ via MATLAB experiments. We again benchmark our algorithm against the DR splitting method in [25]. Both algorithms are initialized at the origin and terminated when (45) is satisfied with $tol = 10^{-8}$. Also, as in the previous subsection, we adopt a heuristic for updating γ following the technique used in [25, Section 5]. Specifically, for the PR splitting method (49), we set $\beta = 2.2$ and start with $\gamma = 0.93/\beta$ and update γ as $\max\{\frac{\gamma}{2}, 0.9999\cdot\gamma_1\}$ whenever $\gamma > \gamma_1 := \frac{\beta-2}{(\beta+1)^2}$, and the sequence satisfies either $\|y^t - y^{t-1}\| > \frac{1000}{t}$ or $\|y^t\| > 10^{10}$. Following a similar discussion as in [25, Remark 4], this heuristic can be shown to give a bounded sequence that clusters at a stationary point of (32). On the other hand, for the DR splitting method, we adopt the same heuristics described in [25, Section 5] for updating γ but we consider three different initial γ's: $k\cdot\gamma_0$ for $k = 50$, 100 and 150, with $\gamma_0 := \sqrt{\frac{3}{2}} - 1$. These variants are denoted by DR_{50}, DR_{100} and DR_{150}, respectively.

As in [25, Section 5], we consider the problem of finding an r-sparse solution of a randomly generated linear system $Ax = b$. To be concrete, we set $C = \{x \in \mathbb{R}^n : Ax = b\}$ and $D = \{x \in \mathbb{R}^n : \|x\|_0 \le r,\ \|x\|_\infty \le 10^6\}$; here $\|x\|_0$ denotes the cardinality of x and $\|x\|_\infty$ is the ℓ_∞ norm of x. For the set C, we first generate an $m \times n$ matrix A and an $\widehat{x} \in \mathbb{R}^r$ with $r = \lceil \frac{m}{5} \rceil$, both with i.i.d. standard Gaussian entries. We then set $\widetilde{x}$ to be the n-dimensional zero vector and randomly assign r entries in $\widetilde{x}$ to be $\widehat{x}$. We further project this $\widetilde{x}$ onto $[-10^6, 10^6]^n$ so that $\widetilde{x} \in D$. Finally, we set $b = A\widetilde{x}$. Consequently, the intersection $C \cap D$ is nonempty for the instance generated because it contains $\widetilde{x}$. In particular, this means that the globally optimal value of $\min_u\{\frac{1}{2}d_C^2(u) : u \in D\}$ is zero.

In our experiments, we generate 50 random instances as described above for each pair of (m, n), where $m \in \{100, 200, 300, 400, 500\}$ and $n \in \{4000, 5000, 6000\}$. We report our results in Tables 3 and 4, where we present the number of iterations averaged over the 50 instances, the largest and smallest function values at termination,[6] and also the number of successes and failures in identifying a sparse solution of the linear system.[7] We also present the average number of iterations for successful instances ($\mathrm{iter_s}$) and failed instances ($\mathrm{iter_f}$).

In Table 3, we compare our PR splitting method with DR_{150}. One can observe that this version of DR splitting method outperforms the PR splitting method in terms of the solution quality in this setting. However, the PR splitting method is consistently faster and its performance becomes comparable with the DR splitting method for easier instances (larger m and smaller n/m).

We also present in Table 4 the numerical results for DR_{50} and DR_{100}. One can see that the DR splitting method becomes faster (while still slower than the PR splitting method) for these two smaller initial γ, at the price of fewer successful instances.

5 Concluding remarks

In this paper, we studied the applicability of the PR splitting method for solving nonconvex optimization problems. We established global convergence of the method when applied to minimizing the sum of a strongly convex Lipschitz differentiable function f and a proper closed function g, under suitable assumptions. Exploiting the *possible nonconvexity* of g, we showed how to suitably apply the PR splitting method to a large class of convex optimization problems whose objective function is not necessarily strongly convex. This significantly broadens the applicability of the PR splitting method to cover feasibility problems and many constrained least squares problems.

[6] For both methods, we report $\frac{1}{2}d_C^2(z^t)$.

[7] We declare a failure if the function value at termination is above 10^{-6}, and a success if the value is below 10^{-12}.

Table 3: Comparing DR_{150} and PR splitting on random instances.

Data		DR_{150}							PR						
m	n	iter	$\mathrm{fval}_{\max}$	$\mathrm{fval}_{\min}$	succ	fail	$\mathrm{iter}_{\mathrm{s}}$	$\mathrm{iter}_{\mathrm{f}}$	iter	$\mathrm{fval}_{\max}$	$\mathrm{fval}_{\min}$	succ	fail	$\mathrm{iter}_{\mathrm{s}}$	$\mathrm{iter}_{\mathrm{f}}$
100	4000	2073	3e-02	1e-16	36	14	1861	2617	297	6e-02	4e-05	0	50	-	297
100	5000	2931	3e-02	1e-16	12	38	1842	3275	367	5e-02	3e-05	0	50	-	367
100	6000	2014	2e-02	2e-16	5	45	1891	2028	431	5e-02	8e-08	0	49	-	423
200	4000	833	7e-02	3e-16	49	1	825	1219	189	2e-01	1e-15	15	35	227	173
200	5000	970	5e-02	2e-16	48	2	947	1528	230	1e-01	2e-15	11	39	297	211
200	6000	1254	4e-02	3e-16	44	6	1193	1704	277	1e-01	3e-15	4	46	344	271
300	4000	607	3e-15	2e-16	50	0	607	-	132	3e-01	9e-16	38	12	138	111
300	5000	705	3e-15	3e-16	50	0	705	-	163	2e-01	1e-15	24	26	181	146
300	6000	819	3e-15	4e-16	50	0	819	-	204	2e-01	2e-15	16	34	241	187
400	4000	523	3e-15	5e-17	50	0	523	-	95	2e-01	8e-16	44	6	96	91
400	5000	574	4e-15	2e-16	50	0	574	-	125	3e-01	1e-15	43	7	127	114
400	6000	655	4e-15	5e-16	50	0	655	-	156	3e-01	2e-15	27	23	165	145
500	4000	500	2e-16	7e-19	50	0	500	-	106	2e-01	6e-16	49	1	64	2173
500	5000	521	1e-15	4e-17	50	0	521	-	91	3e-01	1e-15	47	3	91	87
500	6000	560	4e-15	4e-16	50	0	560	-	123	3e-01	1e-15	47	3	124	108

Table 4: Computational results for DR_{50} and DR_{100}.

Data		DR_{50}							DR_{100}						
m	n	iter	$\mathrm{fval}_{\max}$	$\mathrm{fval}_{\min}$	succ	fail	$\mathrm{iter}_{\mathrm{s}}$	$\mathrm{iter}_{\mathrm{f}}$	iter	$\mathrm{fval}_{\max}$	$\mathrm{fval}_{\min}$	succ	fail	$\mathrm{iter}_{\mathrm{s}}$	$\mathrm{iter}_{\mathrm{f}}$
100	4000	336	4e-02	6e-16	1	49	423	334	854	2e-02	2e-16	5	45	716	870
100	5000	345	4e-02	3e-16	1	49	423	343	681	2e-02	4e-16	2	48	683	681
100	6000	349	3e-02	5e-03	0	50	-	349	647	2e-02	3e-16	1	49	715	646
200	4000	331	1e-01	4e-16	17	33	351	321	711	7e-02	8e-17	48	2	669	1728
200	5000	332	8e-02	9e-16	3	47	357	330	983	5e-02	1e-16	44	6	864	1857
200	6000	341	7e-02	5e-16	6	44	396	333	1186	4e-02	1e-16	24	26	802	1540
300	4000	319	2e-01	1e-16	45	5	315	353	489	3e-15	4e-16	50	0	489	-
300	5000	332	1e-01	5e-16	29	21	335	328	545	3e-15	4e-16	50	0	545	-
300	6000	341	1e-01	6e-16	16	34	378	323	674	5e-02	3e-16	49	1	651	1799
400	4000	271	3e-15	9e-16	50	0	271	-	405	4e-15	2e-16	50	0	405	-
400	5000	301	1e-01	8e-16	48	2	296	413	453	4e-15	5e-16	50	0	453	-
400	6000	329	1e-01	5e-16	40	10	330	329	516	4e-15	5e-16	50	0	516	-
500	4000	244	5e-15	2e-16	50	0	244	-	363	3e-15	2e-16	50	0	363	-
500	5000	269	4e-15	7e-16	50	0	269	-	404	5e-15	3e-16	50	0	404	-
500	6000	295	5e-15	4e-16	50	0	295	-	442	5e-15	9e-16	50	0	442	-

Appendix: Concrete numerical examples

In this appendix, we provide some simple and concrete examples illustrating the different behaviors of the classical PR splitting method, the classical DR splitting method and our proposed PR splitting method (33).

The first example shows that, even in the convex setting, the classical PR splitting method can be faster than the classical DR splitting method, and our proposed PR method can outperform the classical DR method for some particular choice of the parameter γ. The second example on nonconvex feasibility problem shows that the classical PR method can *diverge* while our proposed PR method *converges linearly* to a solution for the feasibility problem.

Example 1. (Classical DR splitting method vs classical/proposed PR method) *Consider $f(x) = \|x\|^2$ and $g(x) = 0$ for all $x \in \mathbb{R}^n$. Then, a direct verification shows that, for any $\gamma > 0$,*

$$\mathrm{prox}_{\gamma f}(z) = \arg\min_u \left\{ \gamma\|u\|^2 + \frac{1}{2}\|u - z\|^2 \right\} = \frac{z}{2\gamma + 1}$$

and

$$\mathrm{prox}_{\gamma g}(z) = \arg\min_u \left\{ \frac{1}{2}\|u - z\|^2 \right\} = z.$$

Thus, the classical DR method reads

$$x^{t+1} = \frac{I + (2\mathrm{prox}_{\gamma g} - I) \circ (2\mathrm{prox}_{\gamma f} - I)}{2}(x^t) = \frac{1}{2\gamma + 1}x^t = \cdots = \left(\frac{1}{2\gamma + 1}\right)^{t+1} x^0,$$

while the classical PR method reads

$$x^{t+1} = (2\mathrm{prox}_{\gamma g} - I) \circ (2\mathrm{prox}_{\gamma f} - I)(x^t) = \frac{1 - 2\gamma}{2\gamma + 1}x^t = \cdots = \left(\frac{1 - 2\gamma}{2\gamma + 1}\right)^{t+1} x^0.$$

Thus, for this example, the classical PR method converges faster than the classical DR method when $\gamma \in (0, 1)$.

Moreover, let $\beta = 2.5$ *and* $\gamma < \frac{\beta - 2}{(\beta+1)^2 L_F} = \frac{1}{49}$. *Then, the proposed PR method* (33) *reads*

$$\begin{cases} y^{t+1} = \arg\min_y \left\{ \frac{7}{2}\|y\|^2 + \frac{1}{2\gamma}\|y - x^t\|^2 \right\} = \frac{1}{1 + 7\gamma}x^t, \\ z^{t+1} = \arg\min_z \left\{ -\frac{5}{2}\|z\|^2 + \frac{1}{2\gamma}\|2y^{t+1} - x^t - z\|^2 \right\} = \frac{1}{1 - 5\gamma}(2y^{t+1} - x^t), \\ x^{t+1} = x^t + 2(z^{t+1} - y^{t+1}) = \left(1 - \frac{4\gamma}{(1 - 5\gamma)(1 + 7\gamma)}\right) x^t. \end{cases} \tag{50}$$

Note that, for $\gamma = 0.01 < \frac{\beta - 2}{(\beta+1)^2 L_F} = \frac{1}{49}$, *we have*

$$0 < 1 - \frac{4\gamma}{(1 - 5\gamma)(1 + 7\gamma)} \leq 0.97 < \frac{1}{2\gamma + 1}.$$

Thus, for $\gamma = 0.01$, *our proposed PR method* (33) *is faster than the classical DR method for this example.*

Example 2. (classical PR method vs the proposed PR method) *Let* $C = \{(0, 0)\}$ *and* $D = \big(\{0\} \times \mathbb{R}\big) \cup \big(\mathbb{R} \times \{0\}\big)$. *We consider the feasibility problem of finding a point in the intersection of* C *and* D. *We start with the initial point* $x^0 = (a, 0)$ *with* $a \neq 0$. *Then, the classical PR splitting method applies* (6) *to* $f(x) = \delta_C(x)$ *and* $g(x) = \delta_D(x)$ *for all* $x \in \mathbb{R}^2$, *and reduces to*

$$x^{t+1} = (2\mathrm{prox}_{\gamma g} - I) \circ (2\mathrm{prox}_{\gamma f} - I)(x^t) = (2P_D - I) \circ (2P_C - I)(x^t) = -x^t.$$

Thus, the classical PR splitting method diverges and cycles between two points $(a, 0)$ *and* $(-a, 0)$. *On the other hand, let* $\beta = 5$ *and* $\gamma \in \left(0, \frac{1}{12}\right)$ *and consider the proposed PR method* (49) *for feasibility problems. This algorithm reads*

$$\begin{cases} y^{t+1} = \dfrac{\gamma P_C\left(\frac{x^t}{1+\beta\gamma}\right) + x^t}{(1 + \beta)\gamma + 1} = \dfrac{x^t}{6\gamma + 1}, \\ z^{t+1} \in P_D\left(\dfrac{2y^{t+1} - x^t}{1 - \beta\gamma}\right) = \left\{\dfrac{2y^{t+1} - x^t}{1 - 5\gamma}\right\}, \\ x^{t+1} = x^t + 2(z^{t+1} - y^{t+1}) = \left(1 - \dfrac{2\gamma}{(1 - 5\gamma)(6\gamma + 1)}\right) x^t, \end{cases} \tag{51}$$

where the formula for the z*-update follows from the fact that* x^t, $y^t \in \mathbb{R} \times \{0\} \subset D$, *and so is* $2y^{t+1} - x^t$ *by the construction. Hence, the proposed PR method* (51) *converges to* $(0, 0) \in C \cap D$ *linearly in this case.*

References

[1] H. Attouch, J. Bolte, P. Redont and A. Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems. An approach based on the Kurdyka-Łojasiewicz inequality. *Math. Oper. Res.* 35, pp. 438–457 (2010).

[2] H. Attouch, J. Bolte and B. F. Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. *Math. Program.* 137, pp. 91–129 (2013).

[3] H. H. Bauschke and J. M. Borwein. On the convergence of von Neumann's alternating projection algorithm for two sets. *Set-Valued Anal.* 1, pp. 185–212 (1993).

[4] H. H. Bauschke and J. M. Borwein. On projection algorithms for solving convex feasibility problems. *SIAM Rev.* 38, pp. 367–426 (1996).

[5] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer (2011).

[6] M. Bogdan, E. van den Berg, W. Su and E. Candès. Statistical estimation and testing via the sorted L1 norm. Preprint (2013). Available at `http://arxiv.org/abs/1310.1969`.

[7] J. Bolte, A. Daniilidis and A. Lewis. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM J. Optim.* 17, pp. 1205–1223 (2007).

[8] J. Bolte, A. Daniilidis, A. Lewis and M. Shiota. Clarke subgradients of stratifiable functions. *SIAM J. Optim.* 18, pp. 556–572 (2007).

[9] J. M. Borwein, G. Li and L. J. Yao. Analysis of the convergence rate for the cyclic projection algorithm applied to basic semialgebraic convex sets. *SIAM J. Optim.* 24, pp. 498–527 (2014).

[10] A. M. Bruckstein, D. L. Donoho and M. Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Rev.* 51, pp. 34–81 (2009).

[11] E. Candès and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.* 35, pp. 2313–2351 (2007).

[12] P. L. Combettes. Iterative construction of the resolvent of a sum of maximal monotone operators. *J. Convex Anal.* 16, pp. 727–748 (2009).

[13] P. L. Combettes and J.-C. Pesquet. A Douglas-Rachford splitting approach to nonsmooth convex variational signal recovery. *IEEE J. Sel. Topics Signal Process.* 1, no. 4, pp. 564–574, (2007).

[14] A. Dobra. Variable selection and dependency networks for genomewide data. *Biostatistics.* 10, pp. 621-639 (2009).

[15] J. Douglas and H. H. Rachford. On the numerical solution of heat conduction problems in two or three space variables. *T. Am. Math. Soc.* 82, pp. 421–439 (1956).

[16] J. Eckstein and D. P. Bertsekas. On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Math. Program.* 55, pp. 293–318 (1992).

[17] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* 96, pp. 1348–1360 (2001).

[18] P. Giselsson and S. Boyd. Diagonal scaling in Douglas-Rachford splitting and ADMM. In *Proc. of the 53rd IEEE Conf. on Decision and Contr.*, pp. 5033–5039 (2014).

[19] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science.* 286, pp. 531–537 (1999).

[20] M. Hong, Z.-Q. Luo and M. Razaviyayn. Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems. *SIAM J. Optim.* 26, pp. 337–364 (2016).

[21] Z. Lu, T.K. Pong and Y. Zhang. An alternating direction method for finding Dantzig selectors. *Comput. Stat. Data An.* 56, pp. 4037–4046 (2012).

[22] K. Knight and W. Fu. Asymptotics for the lasso-type estimators. *Ann. Statist.* 28, pp. 1356–1378 (2000).

[23] A. Kyrillidis, S. Becker, V. Cevher and C. Koch. Sparse projections onto the simplex. *JMLR W&CP* 28, pp. 235–243 (2013).

[24] G. Li and T. K. Pong. Global convergence of splitting methods for nonconvex composite optimization. *SIAM J. Optim.* 25, pp. 2434–2460 (2015).

[25] G. Li and T. K. Pong. Douglas-Rachford splitting for nonconvex optimization with application to nonconvex feasibility problems. *Math. Program.* 159, pp. 371–401 (2016).

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

[27] D. R. Luke, Finding best approximation pairs relative to a convex and a prox-regular set in a Hilbert space. *SIAM J. Optim.* 19, pp. 714–739 (2008).

[28] R. Hesse and D. R. Luke. Nonconvex notions of regularity and convergence of fundamental algorithms for feasibility problems. *SIAM J. Optim.* 23, pp. 2397–2419 (2013).

[29] P. Patrinos, L. Stella and A. Bemporad. Douglas-Rachford splitting: complexity estimates and accelerated variants. In *Proc. of the 53rd IEEE Conf. on Decision and Contr.*, pp. 4234–4239 (2014).

[30] D. W. Peaceman and H. H. Rachford. The numerical solution of parabolic and elliptic differential equations. *J. Soc. Ind. Appl. Math.* 3, pp. 28–41 (1955).

[31] R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis.* Springer (1998).

[32] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B* 58, pp. 267–288 (1996).

[33] H. Wang and A. Banerjee. Bregman alternating direction method of multipliers. *NIPS* 27, pp. 2816–2824 (2014).

[34] K. Y. Yeung, R. E. Bumgarner and A. E. Raftery. Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data. *Bioinformatics.* 21, pp. 2394–2402 (2005).