# A four-operator splitting algorithm for nonconvex and nonsmooth optimization

Jan Harold Alcantara* Ching-pei Lee† Akiko Takeda‡

March 21, 2025

**Abstract**

In this work, we address a class of nonconvex nonsmooth optimization problems where the objective function is the sum of two smooth functions (one of which is proximable) and two nonsmooth functions (one proper, closed and proximable, and the other continuous and weakly concave). We introduce a new splitting algorithm that extends the Davis-Yin splitting (DYS) algorithm to handle such four-term nonconvex nonsmooth problems. We prove that with appropriately chosen stepsizes, our algorithm exhibits global subsequential convergence to stationary points with a stationarity measure converging at a global rate of $1/T$, where $T$ is the number of iterations. When specialized to the setting of the DYS algorithm, our results allow for larger stepsizes compared to existing bounds in the literature. Experimental results demonstrate the practical applicability and effectiveness of our proposed algorithm.

**Keywords.** operator splitting; Davis-Yin splitting; nonconvex optimization; nonsmooth optimization

## 1 Introduction

We consider the nonsmooth and nonconvex problem

$$\min_{x\in\mathbb{R}^n} \ \Psi(x) \coloneqq f(x) + g(x) + h(x) + p(x) \tag{1.1}$$

under the following assumptions.

**Assumption 1.1.** *The functions, $f$, $g$, $h$ and $p$ satisfy the following:*[1]

*(a)* $f : \mathbb{R}^n \to (-\infty, \infty)$ *is an $L_{\mathrm{f}}$-smooth function and is "proximable", in the sense that its proximal mapping either has a closed form or is easily computable;*

*(b)* $g : \mathbb{R}^n \to (-\infty, \infty]$ *is a proper, closed, and proximable function;*

*(c)* $h : \mathbb{R}^n \to (-\infty, \infty)$ *is an $L_{\mathrm{h}}$-smooth function; and*

*janharold.alcantara@riken.jp. Center for Advanced Intelligence Project, RIKEN, Tokyo, Japan.

†chingpei@ism.ac.jp. Department of Advanced Data Science, Institute of Statistical Mathematics, Tokyo, Japan.

‡takeda@mist.i.u-tokyo.ac.jp. Department of Mathematical Informatics, Graduate School of Information Science and Technology, University of Tokyo, Tokyo, Japan, and Center for Advanced Intelligence Project, RIKEN, Tokyo, Japan

[1] See Section 2 for the definitions of $L$-smooth, $\rho_g$-weakly convex, and proximal mapping.

*(d)* $p : \mathbb{R}^n \to \mathbb{R}$ *is continuous (possibly nonsmooth) on an open set containing the domain of* $g$ *such that* $-p$ *is* $\rho_{\mathrm{p}}$*-weakly convex with* $\rho_{\mathrm{p}} \geq 0$.

A notable special case of (1.1) is the three-term optimization problem, where $p \equiv 0$. The Davis-Yin splitting (DYS) algorithm, introduced in Davis and Yin (2017), tackles this problem for convex functions $f$, $g$, and $h$, assuming $h$ additionally meets Assumption 1.1(c). To extend the algorithm's applicability to nonconvex problems, Bian and Zhang (2021) leveraged techniques from Li and Pong (2016) to show that with a sufficiently small stepsize, the DYS algorithm achieves global subsequential convergence. This requires $f$, $g$ and $h$ to satisfy conditions (a), (b) and (c) of Assumption 1.1, respectively. However, despite not requiring the functions to be convex, this extension of the DYS algorithm is still limited as it can only admit at most one nonsmooth function.

In scenarios involving objective functions comprised of multiple nonsmooth, nonconvex terms − such as those encountered in nonconvex fused Lasso Parekh and Selesnick (2015), $\ell_{1-2}$ regularized optimization problems Yin et al. (2015), simultaneous sparse and low-rank optimization problems Richard et al. (2012), among others − the DYS algorithm becomes inapplicable. Indeed, the problem setting of the DYS framework is generally not suited for several classes of nonsmooth nonconvex optimization problems such as the following ones.

**Nonsmooth nonconvex optimization with concave component.** The DYS algorithm is not applicable to problems that take the form

$$\min_{x \in \mathbb{R}^n} \; F(x) + G(x) - H(x), \tag{1.2}$$

where $F$ is $L$-smooth, $G$ is proper and closed, and $H$ is convex. The limitation of the DYS algorithm arises from the potential nonsmoothness of the concave component $-H$. Meanwhile, by setting $f = F$, $g = G$, $h \equiv 0$ and $p := -H$ in (1.1), we satisfy Assumption 1.1 with $L_{\mathrm{h}} = \rho_{\mathrm{p}} = 0$.

**General nonsmooth nonconvex regularized optimization.** Another example is the general class of regularized problems considered in Liu et al. (2019a), given by

$$\min_{x \in \mathbb{R}^n} \; h(x) + g(x) + \sum_{i=1}^{r} P_i(A_i x),$$

where $g$ and $P_i$ are nonsmooth, nonconvex and nonnegative regularizers, and $A_i \in \mathbb{R}^{n_i \times n}$. This difficult class of optimization problems was addressed by Liu et al. (2019a) through the use of the Moreau envelope $M_{\lambda_i P_i}$ of each $P_i$ to obtain an approximate problem:

$$\min_{x \in \mathbb{R}^n} \; h(x) + g(x) + \sum_{i=1}^{r} M_{\lambda_i P_i}(A_i x). \tag{1.3}$$

However, note that even after this reformulation, the DYS algorithm is still not applicable to (1.3), as the third term is in general a nonsmooth function due to the nonconvexity of the $P_i$'s. On the other hand, we can utilize the fact (see (Lucet, 2006, Proposition 3)) that for any function $P$,

$$\frac{1}{2\lambda}\|x\|^2 - M_{\lambda P}(x) = \sup_{y \in \mathrm{dom}(P)} \left\{ \frac{1}{\lambda} x^\top y - \frac{1}{2\lambda}\|y\|^2 - P(y) \right\}. \tag{1.4}$$

By noting the convexity of the function on the right-hand side of (1.4) (being the supremum of affine (convex) functions), we may then fit (1.3) in the form (1.1) by setting $f \equiv 0$ and $p := \sum_{i=1}^{r} M_{\lambda_i P_i} \circ A_i$ so that Assumption 1.1(d) holds with $\rho_{\mathrm{p}} = \sum_{i=1}^{r} \lambda_i^{-1}$.

**Nonconvex feasibility problems.** We can also draw examples from feasibility problems reformulated as optimization problems: Let $A, B, C$ and $D_i$ for $i \in \{1, 2, \ldots, r\}$ be closed nonempty sets where $A$ and $C$ are convex, and the sets $B$ and $D_i$ ($i \in \{1, \ldots, r\}$) are not necessarily convex. The *feasibility problem* is given by:

$$\text{Find } x \in A \cap B \cap C \cap \bigcap_{i=1}^{r} D_i.$$

By defining $d_S(x) \coloneqq \inf_{w \in S} \|w - x\|$ for any set $S$, this problem can be reformulated as the following problem that conforms the form of (1.1):

$$\min_{x \in \mathbb{R}^n} \underbrace{\frac{1}{2} d_A^2(x)}_{f(x)} + \underbrace{\delta_B(x)}_{g(x)} + \underbrace{\frac{1}{2} d_C^2(x)}_{h(x)} + \underbrace{\sum_{i=1}^{r} \frac{1}{2} d_{D_i}^2(x)}_{p(x)}. \tag{1.5}$$

The original setting of the DYS algorithm in Davis and Yin (2017) can handle this problem when $r = 0$. On the other hand, the extended setting considered in Bian and Zhang (2021) can handle the case $B = \mathbb{R}^n$ and $r = 1$ (*i.e.*, the feasibility problem with two convex sets and one nonconvex set), but it requires a very small stepsize. Meanwhile, by the convexity of $A$, $B$ and $C$, we see that Assumption 1.1(a), (b) and (c) hold for (1.5). The function $p$, on the other hand, is a nonsmooth function due to the nonconvexity of each $D_i$. Nevertheless, noting that the half-squared distance function is the Moreau envelope of the indicator function, we have from (1.4) that $p$ satisfies Assumption 1.1(d) with $\rho_{\mathrm{p}} = r$.

We note that the optimization problem (1.1) under Assumption 1.1 can be reformulated into the form (1.2) by setting $H \coloneqq p - \frac{\rho_{\mathrm{p}}}{2}\|\cdot\|^2$ and choosing either (i) $F \coloneqq f + h + \frac{\rho_{\mathrm{p}}}{2}\|\cdot\|^2$, $G \coloneqq g$, or (ii) $F \coloneqq f + h$, $G \coloneqq g + \frac{\rho_{\mathrm{p}}}{2}\|\cdot\|^2$. Under such reformulations, one may apply existing algorithms such as the generalized proximal point (GPP) method An and Nam (2016), which coincides with the proximal difference-of-convex (PDC) algorithm Wen et al. (2018) when $F$ and $G$ are convex.

However, one of the advantages of the framework we propose is that it enables a direct treatment of the problem (1.1) without requiring such reformulations, and it takes advantage of the decomposition of the smooth term of the objective into $f + h$, where $f$ and $h$ possess the structures specified in Assumption 1.1(a) and (c), respectively.[2]

## 1.1 Contributions

In this paper, we propose an extension of the Davis–Yin splitting (DYS) algorithm to handle optimization problems involving two nonsmooth terms, thereby expanding its application to a broader class of nonconvex and nonsmooth problems, including those described earlier. Specifically, we develop a four-operator splitting algorithm for the general problem (1.1) under the setting described in Assumption 1.1. Our main contributions are summarized below:

(I) We prove that our algorithm, which includes the DYS splitting as a special case, is globally subsequentially convergent to stationary points under appropriately chosen stepsizes. Furthermore, a first-order optimality measure is shown to converge at a global rate of $1/T$, where $T$ is the number of iterations.

[2] In practice, when both terms satisfy Assumption 1.1(a), it is often more practical to assign more structure to $f$, as the proximal step tends to exploit more information than the gradient step.

(II) A key theoretical contribution of this work is the derivation of upper bounds on the stepsizes that guarantee convergence. When specialized to the case of $p \equiv 0$, our results yield stepsize estimates for the DYS algorithm that are significantly larger than existing bounds in Bian and Zhang (2021). As a practical consequence, our experiments show that these larger stepsizes lead to faster convergence of the algorithm.

(III) Our experiments demonstrate that leveraging the structures of $f$ and $h$ separately leads to faster convergence compared to GPP/PDC, which treats the smooth term as a single component. Hence, while reformulation into (1.2) is possible, our direct approach offers practical advantages due to utilization of individual structures.

During the final stage of preparing the first draft of this paper, we came across the preprint Dao et al. (2024), which introduces the "doubly relaxed forward-Douglas–Rachford (DRFDR)" splitting algorithm. This method applies to a special case of our problem setting where $\rho_{\mathrm{p}} = 0$ in Assumption 1.1(d). When $\rho_{\mathrm{p}} > 0$, we note that DRFDR can also be applied to (1.1) by subtracting $\frac{\rho_{\mathrm{p}}}{2}\|\cdot\|^2$ from $p$, and then adding $\frac{\rho_{\mathrm{p}}}{2}\|\cdot\|^2$ to either $f$, $g$ or $h$. However, this transformation introduces an additional step, and determining where it is most practical to add this quadratic term is not straightforward. In contrast, our proposed algorithm accommodates the case $\rho_{\mathrm{p}} > 0$ without requiring such transformations. More importantly, our convergence analysis relies on entirely different proof techniques, enabling us to derive larger stepsize bounds compared to those in Dao et al. (2024). Our numerical experiments demonstrate that these larger stepsizes lead to improved practical performance, highlighting another key distinction between our work and Dao et al. (2024).

### 1.2 Organization

This paper is organized as follows. In Section 2, we summarize important definitions, notations and known results that we will use in this paper. We propose our algorithm and establish its global subsequential convergence and convergence rates in Section 3. Experiments to demonstrate the applicability and efficiency of our method are presented in Section 4, and concluding remarks are given in Section 5.

## 2 Preliminaries

Throughout the paper, $\mathbb{R}^n$ denotes the $n$-dimensional Euclidean space endowed with the inner product $\langle\cdot,\cdot\rangle$, and we denote its induced norm by $\|\cdot\|$. The set $(-\infty,\infty]$ is the extended real-line, and we adopt the conventions that $\frac{a}{\infty} = 0$, $\frac{a}{0} = \infty$ for any $a \neq 0$, and $\frac{\infty}{\infty} = 1$.

Let $\phi : \mathbb{R}^n \to (-\infty,\infty]$ be an extended-valued function. The *domain* of $\phi$ is given by the set $\mathrm{dom}(\phi) = \{x \in \mathbb{R}^n : \phi(x) < \infty\}$. We say that $\phi$ is a *proper* function if $\mathrm{dom}(\phi) \neq \emptyset$, and that $\phi$ is *closed* if it is lower semicontinuous. $\phi$ is said to be a *$\sigma_\phi$-convex* function if $\phi - \frac{\sigma_\phi}{2}\|\cdot\|^2$ is convex for some $\sigma_\phi \in \mathbb{R}$. If $\sigma_\phi > 0$, then $\phi$ is *$\sigma_\phi$-strongly convex*. If $\sigma_\phi \leq 0$, we denote $\rho_\phi := -\sigma_\phi$ and call $\phi$ a *$\rho_\phi$-weakly convex function*. In other words, $\phi$ is $\rho_\phi$-weakly convex for $\rho_\phi \geq 0$ if $\phi + \frac{\rho_\phi}{2}\|\cdot\|^2$ is convex.

The *subdifferential* of $\phi$ at a point $x \in \mathrm{dom}(\phi)$ is defined as

$$\partial\phi(x) := \left\{\xi \in \mathbb{R}^n : \exists\{(x^k,\xi^k)\} \text{ such that } x^k \xrightarrow{\phi} x,\ \xi^k \in \hat{\partial}\phi(x^k),\ \text{and } \xi^k \to \xi\right\}, \tag{2.1}$$

where $x^k \xrightarrow{\phi} x$ means $x^k \to x$ and $\phi(x^k) \to \phi(x)$, and

$$\hat{\partial}\phi(x) := \left\{\xi \in \mathbb{R}^n : \liminf_{\bar{x}\to x, \bar{x}\neq x} \frac{\phi(\bar{x}) - \phi(x) - \langle\xi, \bar{x} - x\rangle}{\|\bar{x} - x\|} \geq 0\right\}.$$

When $\phi$ is convex, (2.1) coincides with the classical subdifferential in convex analysis:

$$\partial\phi(x) = \{\xi \in \mathbb{R}^n : \phi(y) \geq \phi(x) + \langle \xi, y - x\rangle\,,\ \forall y \in \mathbb{R}^n\}.$$

If $\phi$ is continuously differentiable, the subdifferential reduces to a singleton containing the gradient of $\phi$. We also note that the definition of the subdifferential gives the following property:

$$\left\{\xi \in \mathbb{R}^n : \exists\{(x^k, \xi^k) \text{ such that } x^k \xrightarrow{\phi} x,\ \xi^k \in \partial\phi(x^k),\ \text{and } \xi^k \to \xi\right\} \subseteq \partial\phi(x). \tag{2.2}$$

For a proper and closed function $\phi$, we define the *proximal mapping* of $\phi$ as

$$\operatorname{prox}_{\gamma\phi}(x) \coloneqq \operatorname*{arg\,min}_{w\in\mathbb{R}^n} \phi(w) + \frac{1}{2\gamma}\|w - x\|^2, \quad \gamma > 0. \tag{2.3}$$

For a set $S \subseteq \mathbb{R}^n$, we define $\operatorname{prox}_{\gamma\phi}(S) \coloneqq \bigcup_{x\in S} \operatorname{prox}_{\gamma\phi}(x)$. From the optimality condition of (2.3), we have

$$y \in \operatorname{prox}_{\gamma\phi}(x) \quad \Longrightarrow \quad x - y \in \gamma\partial\phi(y), \tag{2.4}$$

and the converse holds if $\phi + \frac{1}{2\gamma}\|\cdot\|^2$ is convex. We also use the notation $T_{\gamma\phi} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ to denote

$$T_{\gamma\phi} \coloneqq Id - \gamma\partial\phi,$$

where $Id$ is the identity map on $\mathbb{R}^n$. Given a point $x$ and a subgradient $\xi \in \partial\phi(x)$, for any $\gamma > 0$, we define

$$Q_{\gamma\phi}(w; x, \xi) \coloneqq \phi(x) + \langle \xi, w - x\rangle + \frac{1}{2\gamma}\|w - x\|^2, \quad \forall w \in \mathbb{R}^n.$$

When $\phi$ is smooth, we write $Q_{\gamma\phi}(w; x)$, and $\xi$ is understood to be equal to $\nabla\phi(x)$.

Let $\phi : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function $\phi : \mathbb{R}^n \to \mathbb{R}$. We say that $\phi$ is *$L_\phi$-smooth* if its gradient satisfies

$$\|\nabla\phi(x) - \nabla\phi(y)\| \leq L_\phi\|x - y\|, \quad \forall x, y \in \mathbb{R}^n.$$

We collect some properties of $L_\phi$-smooth functions.

**Lemma 2.1.** *Let $\phi : \mathbb{R}^n \to \mathbb{R}$ be an $L_\phi$-smooth function. Then for any $x, y \in \mathbb{R}^n$,*

*(a)* $|\phi(y) - \phi(x) - \langle\nabla\phi(x), y - x\rangle| \leq \frac{L_\phi}{2}\|y - x\|^2$. *(Descent Lemma)*

*(b) If in addition, $\phi$ is $\rho_\phi$-weakly convex, then for any $L \geq L_\phi$ such that $L > \rho_\phi$, we have*

$$\phi(y) - \phi(x) - \langle\nabla\phi(x), y - x\rangle \geq \frac{1}{2(L - \rho_\phi)}\|\nabla\phi(y) - \nabla\phi(x)\|^2 - \frac{\rho_\phi L}{2(L - \rho_\phi)}\|y - x\|^2.$$

*Proof.* Parts (a) and (b) follow from (Bertsekas, 2016, Proposition A.24) and (Themelis and Patrinos, 2020, Theorem 2.2), respectively. □

**Remark 2.2.** By Lemma 2.1(a), $\phi$ is $\rho_\phi$-weakly convex for some $\rho_\phi \in [0, L_\phi]$ if $\phi$ is $L_\phi$-smooth. If $\phi$ is convex and $L_\phi$-smooth, we obtain from Lemma 2.1(b) that

$$\phi(y) - \phi(x) - \langle\nabla\phi(x), y - x\rangle \geq \frac{1}{2L_\phi}\|\nabla\phi(y) - \nabla\phi(x)\|^2, \quad \forall x, y \in \mathbb{R}^n. \tag{2.5}$$

# 3 Proposed algorithm and its convergence

In this section, we propose our algorithm and provide its convergence analysis under various parameter settings.

## 3.1 Proposed algorithm

We present our proposed algorithm for solving (1.1) in Algorithm 1.

---

**Algorithm 1:** A four-operator splitting algorithm for nonsmooth nonconvex optimization

---

**Step 0.** Choose an initial point $(y^0, z^0) \in \mathbb{R}^n \times \mathbb{R}^n$ and stepsize parameters $\tau > 0$, $\gamma \in (0, \infty)$, and $\alpha, \beta \in (0, \infty]$ such that $\frac{1}{\gamma} = \frac{1}{\alpha} + \frac{1}{\beta}$.

**Step 1.** Set

$$x^k \in \mathrm{prox}_{\alpha f}(z^k), \tag{3.1}$$

$$y^{k+1} \in \mathrm{prox}_{\gamma g}\left(\frac{\gamma}{\alpha}(2x^k - z^k - \alpha\nabla h(x^k)) + \frac{\gamma}{\beta}T_{\beta p}(y^k)\right), \tag{3.2}$$

$$z^{k+1} = z^k + \tau(y^{k+1} - x^k). \tag{3.3}$$

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

---

For convenience in later discussions, we define $P_\Lambda : \mathbb{R}^n \times \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ as

$$P_\Lambda(y, z) \coloneqq \bigcup_{x \in \mathrm{prox}_{\alpha f}(z)} \mathrm{prox}_{\gamma g}\left(\frac{\gamma}{\alpha}(2x - z - \alpha\nabla h(x)) + \frac{\gamma}{\beta}T_{\beta p}(y)\right), \tag{3.4}$$

where $\Lambda \coloneqq (\alpha, \beta, \gamma)$, $\alpha, \beta \in (0, \infty]$ and $\gamma \in (0, \infty)$. Hence, the $y$-step in (3.2) can be written as $y^{k+1} \in P_\Lambda(y^k, z^k)$. Observe that since $f$ is differentiable, $x \in \mathrm{prox}_{\alpha f}(z)$ implies that $z - x = \alpha\nabla f(x)$ by (2.4), so we may also write $P_\Lambda$ as

$$P_\Lambda(y, z) = \bigcup_{x \in \mathrm{prox}_{\alpha f}(z)} \mathrm{prox}_{\gamma g}\left(\frac{\gamma}{\alpha}T_{\alpha(f+h)}(x) + \frac{\gamma}{\beta}T_{\beta p}(y)\right). \tag{3.5}$$

When $\rho_{\mathrm{p}} = 0$, we in particular set $\beta = \infty$, so that $\gamma = \alpha$ by Step 0 of Algorithm 1. On the other hand, when $L_{\mathrm{f}} = L_{\mathrm{h}} = 0$, we set $\alpha = \infty$ and therefore $\gamma = \beta$.

**Remark 3.1.** The above algorithm covers several algorithms in the literature.

1. (Davis-Yin splitting algorithm). When $p \equiv 0$, $\tau = 1$, and $\gamma = \alpha$, the algorithm reduces to the Davis-Yin splitting (DYS) algorithm Davis and Yin (2017), which itself covers the gradient descent algorithm (when $f \equiv g \equiv 0$), the proximal gradient algorithm for the sum of a smooth and a nonsmooth function (when $f \equiv 0$), and the Douglas-Rachford algorithm Douglas and Rachford (1956) (when $h \equiv 0$).

2. (Proximal subgradient method for sum of two nonsmooth functions). When $f \equiv h \equiv 0$ so that the stepsizes satisfy $\gamma = \beta$ and $\alpha = \infty$, the algorithm reduces to the proximal subgradient algorithm:
$$y^{k+1} = \mathrm{prox}_{\gamma g}(y^k - \gamma\partial p(y^k)),$$

but covers a wider range of problems compared to DYS with $f \equiv 0$ since the function $p$ could be nonsmooth.

3. (Generalized proximal point algorithm/Proximal DC algorithm). When $f \equiv 0$, the algorithm simplifies to

$$y^{k+1} \in \operatorname{prox}_{\gamma g}\left(\frac{\gamma}{\alpha}(z^k - \alpha\nabla h(z^k)) + \frac{\gamma}{\beta}T_{\beta p}(y^k)\right),\ z^{k+1} = (1-\tau)z^k + \tau y^{k+1}.$$

Thus, when $\tau = 1$ and $\rho_{\mathrm{p}} = 0$ so that $\beta = \infty$ and $\gamma = \alpha$, we further obtain

$$y^{k+1} \in \operatorname{prox}_{\gamma g}(y^k - \gamma\nabla h(y^k) - \gamma\partial p(y^k)). \tag{3.6}$$

When $g$ is convex, this method is known as the proximal DC (PDC) algorithm Wen et al. (2018), while for a general (not necessarily convex) $g$, it is referred to as the generalized proximal point (GPP) algorithm in An and Nam (2016).

### 3.2 Stepsize bounds

We now derive appropriate stepsizes for Algorithm 1 that will guarantee sufficient descent of a certain merit function. First, we introduce some notations. Given $\xi \in \partial p(y)$, we define $V_{\Lambda,\xi} : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ as

$$V_{\Lambda,\xi}(y,z,x) \coloneqq \min_{w\in\mathbb{R}^n}\ \Phi_{\Lambda,\xi}(w;y,z,x), \tag{3.7}$$

where

$$\Phi_{\Lambda,\xi}(w;y,z,x) \coloneqq Q_{\alpha(f+h)}(w;x) + Q_{\beta p}(w;y,\xi) + g(w). \tag{3.8}$$

We will use (3.7) as a merit function for our proposed algorithm. It is not difficult to verify that when $p \equiv 0$ and $\alpha = \gamma$ as in the setting described in the first item of Remark 3.1, the function given by (3.7) with $\alpha < L_{\mathrm{f}}^{-1}$ and $x = \operatorname{prox}_{\alpha f}(z)$ simplifies to the Davis-Yin envelope function introduced in Liu and Yin (2019), which covers the forward-backward envelope Themelis et al. (2018) and the Douglas-Rachford envelope Patrinos et al. (2014).

**Lemma 3.2.** *Suppose that $f$ and $h$ are continuously differentiable functions, $g$ is proper and closed, and $\gamma > 0$ satisfies $\frac{1}{\gamma} = \frac{1}{\alpha} + \frac{1}{\beta}$ with $\alpha, \beta > 0$. Then $P_\Lambda(y,z) \neq \emptyset$ and*

$$P_\Lambda(y,z) = \bigcup_{\substack{x\in\operatorname{prox}_{\alpha f}(z)\\ \xi\in\partial p(y)}} \operatorname*{arg\,min}_{w\in\mathbb{R}^n} \Phi_{\Lambda,\xi}(w;y,z,x) \tag{3.9}$$

*for all $(y,z) \in \operatorname{dom}(g) \times \mathbb{R}^n$, where $P_\Lambda$ is given by* (3.4).

*Proof.* Given $x \in \operatorname{prox}_{\alpha f}(z)$ and $\xi \in \partial p(y)$, we have that

$$\begin{aligned}
&\operatorname*{arg\,min}_{w\in\mathbb{R}^n} \ \Phi_{\Lambda,\xi}(w;y,z,x)\\
=\ &\operatorname*{arg\,min}_{w\in\mathbb{R}^n} \ f(x)+h(x)+\langle \nabla f(x)+\nabla h(x), w-x\rangle + \frac{1}{2\alpha}\|w-x\|^2\\
&\qquad\qquad + p(y)+\langle \xi, w-y\rangle + \frac{1}{2\beta}\|w-y\|^2 + g(w)\\
=\ &\operatorname*{arg\,min}_{w\in\mathbb{R}^n} \ \langle \nabla f(x)+\nabla h(x), w\rangle + \frac{1}{2\alpha}\|w\|^2 - \frac{1}{\alpha}\langle w, x\rangle\\
&\qquad\qquad + \langle \xi, w\rangle + \frac{1}{2\beta}\|w\|^2 - \frac{1}{\beta}\langle w, y\rangle + g(w)\\
=\ &\operatorname*{arg\,min}_{w\in\mathbb{R}^n} \ \frac{1}{2\gamma}\|w\|^2 - \frac{1}{\alpha}\langle x-\alpha(\nabla f(x)+\nabla h(x)), w\rangle - \frac{1}{\beta}\langle y-\beta\xi, w\rangle + g(w)\\
=\ &\operatorname*{arg\,min}_{w\in\mathbb{R}^n} \ \frac{1}{2\gamma}\left\|w-\left(\frac{\gamma}{\alpha}(x-\alpha\nabla f(x)-\alpha\nabla h(x))+\frac{\gamma}{\beta}(y-\beta\xi)\right)\right\|^2 + g(w)\\
=\ &\operatorname{prox}_{\gamma g}\left(\frac{\gamma}{\alpha}(x-\alpha\nabla f(x)-\alpha\nabla h(x))+\frac{\gamma}{\beta}(y-\beta\xi)\right)
\end{aligned}$$

where the third equality follows from the definition of $\gamma$. The claim now follows from (3.5). □

The following lemma provides an upper bound for the merit function $V_{\Lambda,\xi}$.

**Lemma 3.3.** *Under the assumptions of Lemma* 3.2*, we have*

$$V_{\Lambda,\xi}(y,z,x) \le Q_{\alpha(f+h)}(y;x) + p(y) + g(y)$$

*for all* $(y,z) \in \operatorname{dom}(g)\times\mathbb{R}^n$*,* $x \in \operatorname{prox}_{\alpha f}(z)$*,* $\xi \in \partial p(y)$*,* $y^+ \in \operatorname*{arg\,min}_{w\in\mathbb{R}^n} \Phi_{\Lambda,\xi}(w;y,z,x)$*.*

*Proof.* We immediately get the result by noting from (3.7) that $V_{\Lambda,\xi}(y,z,x) \le \Phi_{\Lambda,\xi}(y;y,z,x)$. □

We now compute a lower bound for $V_{\Lambda,\xi}$. We make use of the estimate provided in Lemma 2.1(b) in the following lemma, inspired by Themelis and Patrinos (2020) when they studied the Douglas-Rachford algorithm.

**Lemma 3.4.** *Suppose that Assumption* 1.1 *holds. Let* $\rho_{\mathrm{f}} \ge 0$ *be such that* $f$ *is* $\rho_{\mathrm{f}}$*-weakly convex and* $L \ge L_{\mathrm{f}}$ *be such that* $L-\rho_{\mathrm{f}} > 0$*.*[3] *Then*

$$\begin{aligned}
&V_{\Lambda,\xi}(y,z,x)\\
\ge\ &Q_{\alpha(f+h)}(y^+;\hat{x}) + p(y^+) + g(y^+) + \left\langle \nabla f(\hat{x}) - \nabla f(x) - \frac{1}{\alpha}(\hat{x}-x), x-y^+\right\rangle\\
&+\langle \nabla h(x)-\nabla h(\hat{x}), y^+-\hat{x}\rangle + \frac{1}{2(L-\rho_{\mathrm{f}})}\|\nabla f(x)-\nabla f(\hat{x})\|^2\\
&-\left(\frac{L_{\mathrm{h}}}{2}+\frac{1}{2\alpha}+\frac{\rho_{\mathrm{f}}L}{2(L-\rho_{\mathrm{f}})}\right)\|\hat{x}-x\|^2 + \frac{1-\beta\rho_{\mathrm{p}}}{2\beta}\|y^+-y\|^2
\end{aligned} \tag{3.10}$$

*for all* $(y,z) \in \operatorname{dom}(g)\times\mathbb{R}^n$*,* $x \in \operatorname{prox}_{\alpha f}(z)$*,* $\hat{x} \in \mathbb{R}^n$*,* $\xi \in \partial p(y)$*, and* $y^+ \in \arg\min_{w\in\mathbb{R}^n} \Phi_{\Lambda,\xi}(w;y,z,x)$*.*

[3] By Assumption 1.1(a) and Remark 2.2, we see that there indeed exists $\rho_{\mathrm{f}} \ge 0$ such that $f + \frac{\rho_{\mathrm{f}}}{2}\|\cdot\|^2$ is convex.

*Proof.* We have from Lemma 2.1(b) that for all $\hat{x} \in \mathbb{R}^n$,

$$\begin{aligned} Q_{\alpha f}(y^+; x) \geq & \left( f(\hat{x}) + \langle \nabla f(\hat{x}), x - \hat{x} \rangle + \frac{1}{2(L - \rho_{\mathrm{f}})} \|\nabla f(x) - \nabla f(\hat{x})\|^2 \right. \\ & \left. - \frac{\rho_{\mathrm{f}} L}{2(L - \rho_{\mathrm{f}})} \|x - \hat{x}\|^2 \right) + \langle \nabla f(x), y^+ - x \rangle + \frac{1}{2\alpha} \|y^+ - x\|^2 \\ = & f(\hat{x}) + \langle \nabla f(\hat{x}) - \nabla f(x), x - y^+ \rangle + \langle \nabla f(\hat{x}), y^+ - \hat{x} \rangle \\ & + \frac{1}{2(L - \rho_{\mathrm{f}})} \|\nabla f(x) - \nabla f(\hat{x})\|^2 - \frac{\rho_{\mathrm{f}} L}{2(L - \rho_{\mathrm{f}})} \|x - \hat{x}\|^2 + \frac{1}{2\alpha} \|y^+ - x\|^2, \end{aligned}$$

for any $L \geq L_{\mathrm{f}}$ such that $L > \rho_{\mathrm{f}}$. Using the identity $\|a - b\|^2 - \|a - c\|^2 = -\|b - c\|^2 + 2\langle b - c, b - a \rangle$ and after some routine calculations, we further obtain

$$\begin{aligned} Q_{\alpha f}(y^+; x) \geq & Q_{\alpha f}(y^+; \hat{x}) + \left\langle \nabla f(\hat{x}) - \nabla f(x) - \frac{1}{\alpha}(\hat{x} - x), x - y^+ \right\rangle \\ & - \left( \frac{1}{2\alpha} + \frac{\rho_{\mathrm{f}} L}{2(L - \rho_{\mathrm{f}})} \right) \|x - \hat{x}\|^2 + \frac{1}{2(L - \rho_{\mathrm{f}})} \|\nabla f(x) - \nabla f(\hat{x})\|^2. \end{aligned} \tag{3.11}$$

On the other hand, we have from Lemma 2.1(a) that

$$\begin{aligned} & h(x) + \langle \nabla h(x), y^+ - x \rangle \\ \geq & \left( h(\hat{x}) - \langle \nabla h(x), \hat{x} - x \rangle - \frac{L_{\mathrm{h}}}{2} \|\hat{x} - x\|^2 \right) + \langle \nabla h(x), y^+ - x \rangle \\ = & h(\hat{x}) + \langle \nabla h(\hat{x}), y^+ - \hat{x} \rangle + \langle \nabla h(x) - \nabla h(\hat{x}), y^+ - \hat{x} \rangle - \frac{L_{\mathrm{h}}}{2} \|\hat{x} - x\|^2 \end{aligned} \tag{3.12}$$

for all $\hat{x} \in \mathbb{R}^n$. By Assumption 1.1(d) and the fact that

$$\partial \left( \frac{\rho_{\mathrm{p}}}{2} \|\cdot\|^2 - p \right)(y) = \rho_{\mathrm{p}} y - \partial p(y), \tag{3.13}$$

which holds by (Rockafellar and Wets, 1998, Exercise 8.8(c)), we have

$$\frac{\rho_{\mathrm{p}}}{2} \|y^+\|^2 - p(y^+) \geq \frac{\rho_{\mathrm{p}}}{2} \|y\|^2 - p(y) + \left\langle \rho_{\mathrm{p}} y - \xi, y^+ - y \right\rangle,$$

for any $\xi \in \partial p(y)$. This implies that

$$p(y) \geq p(y^+) - \langle \xi, y^+ - y \rangle - \frac{\rho_{\mathrm{p}}}{2} \|y^+ - y\|^2 \tag{3.14}$$

and hence

$$Q_{\beta p}(y^+; y, \xi) \geq \left( p(y^+) + \frac{1 - \beta \rho_{\mathrm{p}}}{2\beta} \|y^+ - y\|^2 \right). \tag{3.15}$$

Using the fact that $V_{\Lambda,\xi}(y, z, x) = \Phi_{\Lambda,\xi}(y^+; y, z, x)$ together with the bounds (3.11), (3.12) and (3.15), we obtain the desired conclusion. □

We now show that $\{V_{\Lambda,\xi^k}(y^k, z^k, x^k)\}$ is a nonincreasing sequence for appropriately chosen stepsizes. To simplify the notations, we denote

$$V_k \coloneqq V_{\Lambda,\xi^k}(y^k, z^k, x^k) = \Phi_{\Lambda,\xi^k}(y^{k+1}; y^k, z^k, x^k).$$

In what follows, we discuss the cases $\tau \in (0,1]$, $\tau \in (1,2)$ and $\tau \in [2,\infty)$ separately. For each case, we will show that when $L_{\mathrm{f}} + L_{\mathrm{h}} > 0$, there exist a function $c(\alpha)$ and a finite interval $I \subseteq (0,\infty)$ such that (i) $c(\alpha) \le 0$ on $I$, (ii) $c(\alpha) < 0$ on the interior of $I$, and (iii) the inequality

$$V_{k-1} - V_k \ge -\frac{c(\alpha)}{2\tau\alpha}\left\|x^k - x^{k-1}\right\|^2 + \frac{1-\beta\rho_{\mathrm{p}}}{2\beta}\left\|y^k - y^{k-1}\right\|^2 \tag{3.16}$$

holds for all $k$. For the case $L_{\mathrm{f}} + L_{\mathrm{h}} = 0$, the above inequality also holds but with $\alpha = \infty$ and $c(\alpha)$ a negative constant, so that the first term on the right-hand side vanishes; see Remark 3.8.

**Theorem 3.5** (Stepsize for $\tau \in (0,1]$)**.** *Suppose Assumption* 1.1 *holds and* $L_{\mathrm{f}} + L_{\mathrm{h}} > 0$. *If* $\{(x^k, y^k, z^k)\}$ *is generated by Algorithm* 1 *with* $\tau \in (0,1]$ *and* $\alpha \in (0, \bar{\alpha}]$,

$$\bar{\alpha} := \begin{cases} \frac{1}{L_{\mathrm{f}}+L_{\mathrm{h}}} & \text{if } (2-\tau)L_{\mathrm{f}} - 2\rho_{\mathrm{f}} \ge \tau L_{\mathrm{h}}, \\ \frac{\tau}{2\eta^*} & \text{otherwise,} \end{cases} \tag{3.17}$$

*and* $\eta^*$ *is the positive root of*

$$q(\eta) := 2(2-\tau)\eta^2 - \tau((2-\tau)L_{\mathrm{h}} + \rho_{\mathrm{f}}\tau)\eta - \tau(\rho_{\mathrm{f}}^2 + L_{\mathrm{f}}L_{\mathrm{h}}), \tag{3.18}$$

*then* (3.16) *holds with*

$$c(\alpha) := \begin{cases} 2L_{\mathrm{f}}(L_{\mathrm{f}}+L_{\mathrm{h}})\alpha^2 + ((2-\tau)L_{\mathrm{h}} - \tau L_{\mathrm{f}})\,\alpha - (2-\tau) & \text{if } (2-\tau)L_{\mathrm{f}} - 2\rho_{\mathrm{f}} \ge \tau L_{\mathrm{h}}, \\ 2L_{\mathrm{f}}L_{\mathrm{h}}\alpha^2 + \left((2-\tau)L_{\mathrm{h}} + \frac{\rho_{\mathrm{f}}\tau(\eta^*+\rho_{\mathrm{f}})}{\eta^*}\right)\alpha - (2-\tau) & \text{otherwise.} \end{cases} \tag{3.19}$$

*In particular,* $\{V_k\}$ *is nonincreasing if* $\alpha \le \bar{\alpha}$ *and* $\beta \le \rho_{\mathrm{p}}^{-1}$*, and strictly decreasing if at least one holds with strict inequality.*

*Proof.* From Lemma 3.2, we know that for each $k$, there exists $\xi^k \in \partial p(y^k)$ such that $y^{k+1} \in \arg\min_{w\in\mathbb{R}^n} \Phi_{\Lambda,\xi^k}(w; y^k, z^k, x^k)$. By Lemma 3.3,

$$V_k \le Q_{\alpha(f+h)}(y^k; x^k) + p(y^k) + g(y^k). \tag{3.20}$$

On the other hand, setting $(y, z, x) = (y^{k-1}, z^{k-1}, x^{k-1})$ and $\hat{x} = x^k$ in Lemma 3.4, we have

$$\begin{aligned} V_{k-1} \ge\ & Q_{\alpha(f+h)}(y^k; x^k) + p(y^k) + g(y^k) \\ & + \left\langle \nabla f(x^k) - \nabla f(x^{k-1}) - \frac{1}{\alpha}(x^k - x^{k-1}), x^{k-1} - y^k \right\rangle \\ & + \left\langle \nabla h(x^{k-1}) - \nabla h(x^k), y^k - x^k \right\rangle + \frac{1}{2(L-\rho_{\mathrm{f}})}\left\|\nabla f(x^{k-1}) - \nabla f(x^k)\right\|^2 \\ & - \left(\frac{L_{\mathrm{h}}}{2} + \frac{1}{2\alpha} + \frac{\rho_{\mathrm{f}}L}{2(L-\rho_{\mathrm{f}})}\right)\left\|x^k - x^{k-1}\right\|^2 + \frac{1-\beta\rho_{\mathrm{p}}}{2\beta}\left\|y^k - y^{k-1}\right\|^2. \end{aligned} \tag{3.21}$$

Subtracting (3.20) from (3.21), we get

$$\begin{aligned} V_{k-1} - V_k \ge\ & \left\langle \nabla f(x^k) - \nabla f(x^{k-1}) - \frac{1}{\alpha}(x^k - x^{k-1}), x^{k-1} - y^k \right\rangle \\ & + \left\langle \nabla h(x^{k-1}) - \nabla h(x^k), y^k - x^k \right\rangle + \frac{1}{2(L-\rho_{\mathrm{f}})}\left\|\nabla f(x^{k-1}) - \nabla f(x^k)\right\|^2 \\ & - \left(\frac{L_{\mathrm{h}}}{2} + \frac{1}{2\alpha} + \frac{\rho_{\mathrm{f}}L}{2(L-\rho_{\mathrm{f}})}\right)\left\|x^k - x^{k-1}\right\|^2 + \frac{1-\beta\rho_{\mathrm{p}}}{2\beta}\left\|y^k - y^{k-1}\right\|^2 \end{aligned} \tag{3.22}$$

Meanwhile, we have from (3.3) that $y^k = \frac{1}{\tau}(z^k - z^{k-1}) + x^{k-1}$ and from (3.1) and (2.4) that

$$z^k - z^{k-1} = (x^k + \alpha\nabla f(x^k)) - (x^{k-1} + \alpha\nabla f(x^{k-1})). \tag{3.23}$$

Thus,

$$y^k - x^{k-1} = \frac{1}{\tau}(x^k - x^{k-1}) + \frac{\alpha}{\tau}(\nabla f(x^k) - \nabla f(x^{k-1})), \tag{3.24}$$

which can also be rewritten as

$$y^k - x^k = \left(\frac{1}{\tau} - 1\right)(x^k - x^{k-1}) + \frac{\alpha}{\tau}(\nabla f(x^k) - \nabla f(x^{k-1})). \tag{3.25}$$

With these, algebraic calculations lead to

$$\begin{aligned} &\left\langle \nabla f(x^k) - \nabla f(x^{k-1}) - \frac{1}{\alpha}(x^k - x^{k-1}), x^{k-1} - y^k \right\rangle \\ \overset{(3.24)}{=}\ &-\frac{\alpha}{\tau}\left\|\nabla f(x^k) - \nabla f(x^{k-1})\right\|^2 + \frac{1}{\tau\alpha}\left\|x^k - x^{k-1}\right\|^2 \end{aligned} \tag{3.26}$$

and

$$\begin{aligned} \left\langle \nabla h(x^{k-1}) - \nabla h(x^k), y^k - x^k \right\rangle &\overset{(3.25)}{=} \left(1 - \frac{1}{\tau}\right)\left\langle \nabla h(x^k) - \nabla h(x^{k-1}), x^k - x^{k-1} \right\rangle \\ &\quad - \frac{\alpha}{\tau}\left\langle \nabla h(x^k) - \nabla h(x^{k-1}), \nabla f(x^k) - \nabla f(x^{k-1}) \right\rangle. \qquad (3.27) \\ &\geq \frac{\tau - 1}{\tau}\left\|\nabla h(x^k) - \nabla h(x^{k-1})\right\| \cdot \left\|x^k - x^{k-1}\right\| \\ &\quad - \frac{\alpha}{\tau}\left\|\nabla h(x^k) - \nabla h(x^{k-1})\right\| \cdot \left\|\nabla f(x^k) - \nabla f(x^{k-1})\right\| \\ &\geq \frac{\tau - 1}{\tau}L_{\mathrm{h}}\left\|x^k - x^{k-1}\right\|^2 - \frac{\alpha}{\tau}L_{\mathrm{f}}L_{\mathrm{h}}\left\|x^k - x^{k-1}\right\|^2, \qquad (3.28) \end{aligned}$$

where the first inequality is by the Cauchy-Schwarz inequality and noting that $\tau \in (0, 1]$, while the last inequality holds by the Lipschitz continuity of the gradients of $f$ and $h$. Combining (3.22), (3.26) and (3.28), we obtain

$$\begin{aligned} V_{k-1} - V_k \geq &\left(-\frac{\alpha}{\tau} + \frac{1}{2(L - \rho_{\mathrm{f}})}\right)\left\|\nabla f(x^{k-1}) - \nabla f(x^k)\right\|^2 \\ &+ \left[\left(\frac{1}{\tau} - \frac{1}{2}\right)\frac{1}{\alpha} - \frac{L_{\mathrm{h}}}{2} - \frac{\rho_{\mathrm{f}}L}{2(L - \rho_{\mathrm{f}})} + \frac{\tau - 1}{\tau}L_{\mathrm{h}} - \frac{\alpha}{\tau}L_{\mathrm{h}}L_{\mathrm{f}}\right]\left\|x^k - x^{k-1}\right\|^2 \\ &+ \frac{1 - \beta\rho_{\mathrm{p}}}{2\beta}\left\|y^k - y^{k-1}\right\|^2. \end{aligned} \tag{3.29}$$

Now, we discuss two disjoint cases.

**Case 1.** Suppose that $(2 - \tau)L_{\mathrm{f}} - 2\rho_{\mathrm{f}} \geq \tau L_{\mathrm{h}}$. Then $L_{\mathrm{f}} - \rho_{\mathrm{f}} \geq \frac{\tau}{2}(L_{\mathrm{f}} + L_{\mathrm{h}}) > 0$ and we may take $L = L_{\mathrm{f}}$ in (3.29). Let $\hat{\alpha} > 0$ be such that $\hat{\alpha} \geq \frac{\tau}{2(L_{\mathrm{f}} - \rho_{\mathrm{f}})}$, and suppose that $0 < \alpha \leq \hat{\alpha}$. Then

$$-\frac{\alpha}{\tau} + \frac{1}{2(L_{\mathrm{f}} - \rho_{\mathrm{f}})} \geq -\frac{\hat{\alpha}}{\tau} + \frac{1}{2(L_{\mathrm{f}} - \rho_{\mathrm{f}})},$$

where the quantity on the right-hand side is at most zero. Together with the Lipschitz continuity of $\nabla f$ and (3.29) with $L = L_{\mathrm{f}}$, for any $\alpha \in (0, \hat{\alpha}]$ we have

$$\begin{aligned} V_{k-1} - V_k &\geq \left( \frac{2-\tau}{2\tau\hat{\alpha}} - \hat{\alpha}\frac{L_{\mathrm{f}}(L_{\mathrm{f}}+L_{\mathrm{h}})}{\tau} + \frac{L_{\mathrm{f}}}{2} - \frac{L_{\mathrm{h}}}{2} + \frac{\tau-1}{\tau}L_{\mathrm{h}} \right) \left\| x^k - x^{k-1} \right\|^2 \\ &\quad + \frac{1-\beta\rho_{\mathrm{p}}}{2\beta} \left\| y^k - y^{k-1} \right\|^2 \\ &= -\frac{c(\hat{\alpha})}{2\tau\hat{\alpha}} \left\| x^k - x^{k-1} \right\|^2 + \frac{1-\beta\rho_{\mathrm{p}}}{2\beta} \left\| y^k - y^{k-1} \right\|^2 \end{aligned} \tag{3.30}$$

where

$$\begin{aligned} c(\hat{\alpha}) &\coloneqq 2L_{\mathrm{f}}(L_{\mathrm{f}}+L_{\mathrm{h}})\hat{\alpha}^2 + ((2-\tau)L_{\mathrm{h}} - \tau L_{\mathrm{f}})\,\hat{\alpha} - (2-\tau) \\ &= \left( \hat{\alpha} - \frac{1}{L_{\mathrm{f}}+L_{\mathrm{h}}} \right) (2L_{\mathrm{f}}(L_{\mathrm{f}}+L_{\mathrm{h}})\hat{\alpha} + (2-\tau)(L_{\mathrm{f}}+L_{\mathrm{h}})), \end{aligned} \tag{3.31}$$

which is nonpositive when $\hat{\alpha} \leq \frac{1}{L_{\mathrm{f}}+L_{\mathrm{h}}}$. Hence, (3.30) holds with nonpositive $c(\hat{\alpha})$ when $\hat{\alpha} \in \left[ \frac{\tau}{2(L_{\mathrm{f}}-\rho_{\mathrm{f}})}, \frac{1}{L_{\mathrm{f}}+L_{\mathrm{h}}} \right]$, which is a nonempty interval due to our hypothesis that $(2-\tau)L_{\mathrm{f}} - 2\rho_{\mathrm{f}} \geq \tau L_{\mathrm{h}}$. It is clear that $c(\alpha) < 0$ when $\alpha < \bar{\alpha}$.

**Case 2.** Suppose now that

$$(2-\tau)L_{\mathrm{f}} - 2\rho_{\mathrm{f}} < \tau L_{\mathrm{h}}. \tag{3.32}$$

Given $L \geq L_{\mathrm{f}}$ with $L - \rho_{\mathrm{f}} > 0$, we define $\hat{\alpha}(L) \coloneqq \frac{\tau}{2(L-\rho_{\mathrm{f}})}$ and select $\alpha \in (0, \hat{\alpha}(L)]$. Then

$$-\frac{\alpha}{\tau} + \frac{1}{2(L-\rho_{\mathrm{f}})} \geq -\frac{\hat{\alpha}(L)}{\tau} + \frac{1}{2(L-\rho_{\mathrm{f}})} = 0.$$

Hence, we have from (3.29) that

$$V_{k-1} - V_k \geq -\frac{c(\hat{\alpha}(L))}{2\tau\hat{\alpha}(L)} \left\| x^k - x^{k-1} \right\|^2 + \frac{1-\beta\rho_{\mathrm{p}}}{2\beta} \left\| y^k - y^{k-1} \right\|^2 \tag{3.33}$$

where

$$c(\hat{\alpha}(L)) \coloneqq 2L_{\mathrm{f}}L_{\mathrm{h}}\hat{\alpha}(L)^2 + \left( (2-\tau)L_{\mathrm{h}} + \frac{\rho_{\mathrm{f}}\tau L}{L-\rho_{\mathrm{f}}} \right) \hat{\alpha}(L) - (2-\tau). \tag{3.34}$$

To determine the largest allowable stepsize $\hat{\alpha}(L)$ so that $c(\hat{\alpha}) \leq 0$, we calculate

$$L^* \coloneqq \min\{L : c(\hat{\alpha}(L)) \leq 0\ , L \geq L_{\mathrm{f}}, L > \rho_{\mathrm{f}}\}, \tag{3.35}$$

so that $\hat{\alpha}(L^*)$ is the desired stepsize. By some routine calculations, it can be shown that $c(\hat{\alpha}(L)) = -\frac{1}{2(L-\rho_{\mathrm{f}})^2} q(L-\rho_{\mathrm{f}}) = -\frac{1}{2\eta^2} q(\eta)$, where $q$ is given by the polynomial (3.18) and $\eta \coloneqq L - \rho_{\mathrm{f}}$. Hence, if $\eta^*$ is the (strictly) positive root of $q$, then $L^* \coloneqq \max\{\eta^* + \rho_{\mathrm{f}}, L_{\mathrm{f}}\}$. We now claim that $L^* = \eta^* + \rho_{\mathrm{f}}$. If $L_{\mathrm{f}} = \rho_{\mathrm{f}}$, this immediately holds since $\eta^* > 0$. Suppose now that $L_{\mathrm{f}} > \rho_{\mathrm{f}}$. By the definition of $\hat{\alpha}(L)$, note that we may write $c$ as

$$c(\hat{\alpha}(L)) = c(\hat{\alpha}(L)) + 2LL_{\mathrm{f}}\hat{\alpha}(L)^2 - \frac{\tau L L_{\mathrm{f}}}{L-\rho_{\mathrm{f}}}\hat{\alpha}(L)$$

for any $L > \rho_{\mathrm{f}}$. Simplifying this expression, we obtain

$$c(\hat{\alpha}(L)) = 2L_{\mathrm{f}}(L_{\mathrm{h}}+L)\hat{\alpha}(L)^2 + \left( (2-\tau)L_{\mathrm{h}} - \frac{\tau L(L_{\mathrm{f}}-\rho_{\mathrm{f}})}{L-\rho_{\mathrm{f}}} \right) \hat{\alpha}(L) - (2-\tau). \tag{3.36}$$

Since $L_{\mathrm{f}} > \rho_{\mathrm{f}}$, $c(\hat{\alpha}(L_{\mathrm{f}}))$ is well-defined and can be calculated as

$$\begin{aligned} c(\hat{\alpha}(L_{\mathrm{f}})) &= 2L_{\mathrm{f}}(L_{\mathrm{f}} + L_{\mathrm{h}})\hat{\alpha}(L_{\mathrm{f}})^2 + ((2-\tau)L_{\mathrm{h}} - \tau L_{\mathrm{f}})\,\hat{\alpha}(L_{\mathrm{f}}) - (2-\tau) \\ &= \left(\hat{\alpha}(L_{\mathrm{f}}) - \frac{1}{L_{\mathrm{f}} + L_{\mathrm{h}}}\right)(2L_{\mathrm{f}}(L_{\mathrm{f}} + L_{\mathrm{h}})\hat{\alpha}(L_{\mathrm{f}}) + (2-\tau)(L_{\mathrm{f}} + L_{\mathrm{h}}))\,. \end{aligned}$$

Since (3.32) implies that $\hat{\alpha}(L_{\mathrm{f}}) > \frac{1}{L_{\mathrm{f}}+L_{\mathrm{h}}}$, it follows that $c(\hat{\alpha}(L_{\mathrm{f}})) > 0$. Hence, $L^* \neq L_{\mathrm{f}}$ by the definition of $L^*$ in (3.35). The claim that $L^* = \eta^* + \rho_{\mathrm{f}}$ now follows. We also note that since $c(\hat{\alpha}(L^*)) \leq 0$, it follows from (3.34) that $c(\alpha) < 0$ for any $\alpha < \hat{\alpha}(L^*)$, where $c(\alpha)$ is as defined in (3.19). □

**Remark 3.6.** To gain insight on the magnitude of the stepsize upper bound $\bar{\alpha} = \frac{\tau}{2\eta^*}$ in the second case of the above proof, consider $L \coloneqq \frac{\tau L_{\mathrm{h}} + 2\rho_{\mathrm{f}}}{2-\tau}$. By (3.32), $L > L_{\mathrm{f}}$ and $L > \rho_{\mathrm{f}}$. In addition, for this choice of $L$, $c(\hat{\alpha}(L)) \leq 0$. By (3.35), it holds that $L \geq L^*$, ensuring $\frac{\tau}{2\eta^*} \geq \frac{\tau}{2(L-\rho_{\mathrm{f}})} = \frac{2-\tau}{2(L_{\mathrm{h}}+\rho_{\mathrm{f}})}$. To obtain an upper bound, we note that from (3.36), it can be verified that an alternative way to express $c(\hat{\alpha}(L))$ is

$$\begin{aligned} c(\hat{\alpha}(L)) &= \left[2L_{\mathrm{f}}(L_{\mathrm{f}} + L_{\mathrm{h}})\hat{\alpha}(L)^2 + ((2-\tau)L_{\mathrm{h}} - \tau L_{\mathrm{f}})\hat{\alpha}(L) - (2-\tau)\right] \\ &\quad + 2L_{\mathrm{f}}(L - L_{\mathrm{f}})\hat{\alpha}(L)^2 + \left(\tau L_{\mathrm{f}} - \frac{\tau L(L_{\mathrm{f}} - \rho_{\mathrm{f}})}{L_{\mathrm{f}} - \rho_{\mathrm{f}}}\right)\hat{\alpha}(L) \\ &= \left(\hat{\alpha}(L) - \frac{1}{L_{\mathrm{f}} + L_{\mathrm{h}}}\right)(2L_{\mathrm{f}}(L_{\mathrm{f}} + L_{\mathrm{h}})\hat{\alpha}(L) + (2-\tau)(L_{\mathrm{f}} + L_{\mathrm{h}})) \\ &\quad + 2L_{\mathrm{f}}(L - L_{\mathrm{f}})\hat{\alpha}(L)^2 + \frac{\tau\rho_{\mathrm{f}}(L - L_{\mathrm{f}})}{L - \rho_{\mathrm{f}}}\hat{\alpha}(L). \end{aligned}$$

Note that the last two terms are nonnegative when $L = L^*$ since $L^* \geq L_{\mathrm{f}}$ and $L^* > \rho_{\mathrm{f}}$. Since $c(\hat{\alpha}(L^*)) = 0$, it follows that $\hat{\alpha}(L^*) - \frac{1}{L_{\mathrm{f}}+L_{\mathrm{h}}} \leq 0$. That is, $\frac{\tau}{2\eta^*} \leq \frac{1}{L_{\mathrm{f}}+L_{\mathrm{h}}}$. In summary, when $\tau \in (0, 1]$, we have $\frac{2-\tau}{2(L_{\mathrm{h}}+\rho_{\mathrm{f}})} \leq \frac{\tau}{2\eta^*} \leq \frac{1}{L_{\mathrm{f}}+L_{\mathrm{h}}}$ and therefore

$$\bar{\alpha} \leq \frac{1}{L_f + L_h}. \tag{3.37}$$

**Remark 3.7** (Stepsize comparison with Bian and Zhang (2021))**.** For the case $\tau = 1$, the above theorem implies that strict monotonicity of $\{V_k\}$ holds when $\alpha < \bar{\alpha}$, where

$$\bar{\alpha} = \begin{cases} \frac{1}{L_{\mathrm{f}}+L_{\mathrm{h}}} & \text{if } L_{\mathrm{f}} - 2\rho_{\mathrm{f}} \geq L_{\mathrm{h}}, \\ \frac{2}{L_{\mathrm{h}}+\rho_{\mathrm{f}}+\sqrt{(L_{\mathrm{h}}+\rho_{\mathrm{f}})^2+8(\rho_{\mathrm{f}}^2+L_{\mathrm{f}}L_{\mathrm{h}})}} & \text{otherwise.} \end{cases}$$

On the other hand, the bound derived in (Bian and Zhang, 2021, Lemma 3.3) for the DYS algorithm (*i.e.*, $p \equiv 0$) indicates that the stepsize $\alpha > 0$ should satisfy

$$\frac{1}{2}\left(\frac{1}{\alpha} - \rho_{\mathrm{f}}\right) - L_{\mathrm{h}} - \left(\frac{1}{\alpha} + \frac{L_{\mathrm{h}}}{2}\right)(2\alpha\rho_{\mathrm{f}} + 2\alpha L_{\mathrm{f}} + \alpha^2 L_{\mathrm{f}}^2) > 0,$$

or equivalently, under the constraint $\alpha > 0$,

$$d(\alpha) \coloneqq L_{\mathrm{f}}^2 L_{\mathrm{h}}\alpha^3 + 2(L_{\mathrm{f}}^2 + L_{\mathrm{h}}L_{\mathrm{f}} + \rho_{\mathrm{f}}L_{\mathrm{h}})\alpha^2 + (5\rho_{\mathrm{f}} + 2L_{\mathrm{h}} + 4L_{\mathrm{f}})\alpha - 1 < 0.$$

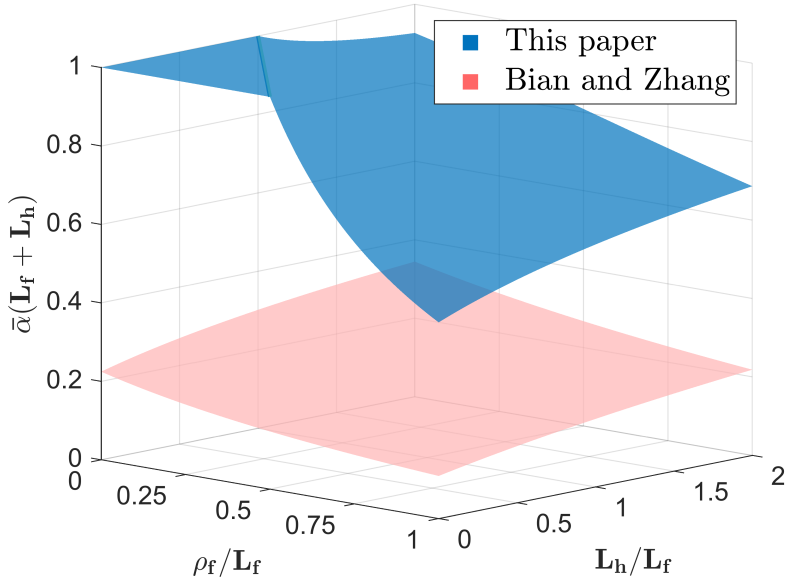



Figure 1: Comparison of stepsize upper bounds (denoted by $\bar{\alpha}$) for the DYS algorithm in this paper (with $\tau = 1$) and in Bian and Zhang (2021).

Hence, the upper bound for $\alpha$ is the unique positive root $\hat{\alpha}$ of the polynomial $d$ given above. Consider $c(\alpha)$ given in (3.19). In the first case, that is, when $L_{\mathrm{f}} - 2\rho_{\mathrm{f}} \geq L_{\mathrm{h}}$, as long as $L_{\mathrm{f}} + L_{\mathrm{h}} > 0$ and $\alpha > 0$, we have

$$c(\alpha) - d(\alpha) = -L_{\mathrm{f}}^2 L_{\mathrm{h}} \alpha^3 - 2\rho_{\mathrm{f}} L_{\mathrm{h}} \alpha^2 - (5\rho_{\mathrm{f}} + L_{\mathrm{h}} + 5L_{\mathrm{f}}) < 0.$$

Hence, $-d(\bar{\alpha}) = c(\bar{\alpha}) - d(\bar{\alpha}) < 0$, and since $\hat{\alpha}$ is the unique positive root of $d$, it follows that $\hat{\alpha} < \bar{\alpha}$. In the second case, recall that $\bar{\alpha} = \hat{\alpha}(L^*) = \frac{1}{2\eta^*} = \frac{1}{2(L^* - \rho_{\mathrm{f}})}$, so that

$$\begin{aligned} c(\bar{\alpha}) &= c(\bar{\alpha}) + 2\rho_{\mathrm{f}} L_{\mathrm{f}} \bar{\alpha}^2 - \frac{\rho_{\mathrm{f}} L_{\mathrm{f}}}{L^* - \rho_{\mathrm{f}}} \bar{\alpha} \\ &= 2(L_{\mathrm{f}} L_{\mathrm{h}} + \rho_{\mathrm{f}} L_{\mathrm{f}}) \bar{\alpha}^2 + \left( L_{\mathrm{h}} + \frac{\rho_{\mathrm{f}}(L^* - L_{\mathrm{f}})}{L^* - \rho_{\mathrm{f}}} \right) \bar{\alpha} - 1 \\ &\leq 2(L_{\mathrm{f}} L_{\mathrm{h}} + L_{\mathrm{f}}^2) \bar{\alpha}^2 + (L_{\mathrm{h}} + \rho_{\mathrm{f}})\, \bar{\alpha} - 1, \end{aligned}$$

where the last inequality holds since $\rho_{\mathrm{f}} \leq L_{\mathrm{f}}$ and $L^* > \rho_{\mathrm{f}}$. Then, provided that $L_{\mathrm{f}} + L_{\mathrm{h}} > 0$, we have

$$c(\bar{\alpha}) - d(\bar{\alpha}) \leq -L_{\mathrm{f}}^2 L_{\mathrm{h}} \bar{\alpha}^3 - 2\rho_{\mathrm{f}} L_{\mathrm{h}} \bar{\alpha}^2 - (4\rho_{\mathrm{f}} + L_{\mathrm{h}} + 4L_{\mathrm{f}}) < 0,$$

and similar to the previous case, we get that $d(\bar{\alpha}) > 0$ and therefore $\hat{\alpha} < \bar{\alpha}$. This shows that our stepsize upper bound is always larger than that in Bian and Zhang (2021). The significant gap between the computed stepsizes is evident in Fig. 1. As we will see in the experiment in Section 4.1.1, a larger $\alpha$ leads to faster convergence. Our improved upper bound for $\alpha$ therefore also provides numerical acceleration for DYS.

**Remark 3.8** (The case $L_{\mathrm{f}} + L_{\mathrm{h}} = 0$)**.** Suppose that $L_{\mathrm{f}} + L_{\mathrm{h}} = 0$, in which case $\alpha = \infty$ and $\gamma = \beta$. From (3.22), we immediately obtain

$$V_{k-1} - V_k \geq \frac{1 - \beta\rho_{\mathrm{p}}}{2\beta}\left\|y^k - y^{k-1}\right\|^2,$$

for any $\tau > 0$ (Note that in this case, the $x$ and $z$ sequences generated by Algorithm 1 are irrelevant). Thus, we still obtain the desired inequality (3.16) by setting $c(\alpha)$ to be any negative number.

The following provides a result similar to the previous theorem but for parameters $\tau$ in $(1, 2)$. Note that the parameter $\sigma_{\mathrm{h}}$ in the theorem below exists; for instance, we can take $\sigma_{\mathrm{h}} = -L_{\mathrm{h}}$ by Lemma 2.1(a).

**Theorem 3.9** (Stepsize for $\tau \in (1, 2)$)**.** *Suppose that Assumption* 1.1 *holds and* $L_{\mathrm{f}} + L_{\mathrm{h}} > 0$. *Let* $\sigma_{\mathrm{h}} \in \mathbb{R}$ *be such that* $h - \frac{\sigma_{\mathrm{h}}}{2}\|\cdot\|^2$ *is convex.*[4] *If* $\{(x^k, y^k, z^k)\}$ *is generated by Algorithm* 1 *with* $\tau \in (1, 2)$ *and* $\alpha \in (0, \bar{\alpha}]$*, where*

$$\bar{\alpha} := \begin{cases} \bar{\alpha}_1 & \text{if } \tau \leq 2\bar{\alpha}_1(L_{\mathrm{f}} - \rho_{\mathrm{f}}) \\ \frac{\tau}{2\eta^*} & \text{otherwise} \end{cases},$$

$\bar{\alpha}_1$ *is the positive root of*

$$c(\alpha) := 2L_{\mathrm{f}}(L_{\mathrm{f}} + L_{\mathrm{h}})\alpha^2 + (\tau L_{\mathrm{h}} - 2(\tau - 1)\sigma_{\mathrm{h}} - \tau L_{\mathrm{f}})\,\alpha - (2 - \tau), \tag{3.38}$$

*and* $\eta^*$ *is the positive root of*

$$q(\eta) := 2(2 - \tau)\eta^2 - \tau(\tau L_{\mathrm{h}} - 2(\tau - 1)\sigma_{\mathrm{h}} + \rho_{\mathrm{f}}\tau)\eta - \tau^2(\rho_{\mathrm{f}}^2 + L_{\mathrm{f}}L_{\mathrm{h}}),$$

*then* (3.16) *holds with* $c(\alpha)$ *given by* (3.38) *if* $\tau \leq 2\bar{\alpha}_1(L_{\mathrm{f}} - \rho_{\mathrm{f}})$*, and*

$$c(\alpha) = 2L_{\mathrm{f}}L_{\mathrm{h}}\alpha^2 + \left(\tau L_{\mathrm{h}} - 2(\tau - 1)\sigma_{\mathrm{h}} + \frac{\rho_{\mathrm{f}}\tau(\eta^* + \rho_{\mathrm{f}})}{\eta^*}\right)\alpha - (2 - \tau))$$

*otherwise. In particular,* $\{V_k\}$ *is nonincreasing if* $\alpha \leq \bar{\alpha}$ *and* $\beta \leq \rho_{\mathrm{p}}^{-1}$*, and strictly decreasing if at least one holds with strict inequality.*

*Proof.* Given $\tau \geq 1$, we have from (3.27) and the monotonicity of $\nabla h - \sigma_{\mathrm{h}} Id$ that

$$\left\langle \nabla h(x^{k-1}) - \nabla h(x^k), y^k - x^k \right\rangle \geq \left(\frac{\tau - 1}{\tau}\right)\sigma_{\mathrm{h}}\left\|x^k - x^{k-1}\right\|^2 - \frac{\alpha}{\tau}L_{\mathrm{f}}L_{\mathrm{h}}\left\|x^k - x^{k-1}\right\|^2. \tag{3.39}$$

The rest of the proof follows arguments similar to those in the proof of Theorem 3.5. □

**Remark 3.10.** From Theorem 3.9, it can be shown that the stepsize bounds for $\tau \in (1, 2)$ are also given by (3.17) if $\sigma_{\mathrm{h}} = L_{\mathrm{h}}$. In particular, when $h = 0$, we obtain that $\bar{\alpha} = \min\left\{\frac{1}{L_{\mathrm{f}}}, \frac{2-\tau}{2\rho_{\mathrm{f}}}\right\}$ for any $\tau \in (0, 2)$, which is the same as the one obtained in (Themelis and Patrinos, 2020, Theorem 4.1), where the case $h = p = 0$ is considered. Another scenario for $\sigma_{\mathrm{h}} = L_{\mathrm{h}}$ is when $h$ is a quadratic function with its Hessian being a positive multiple of the identity matrix.

[4] We note that there indeed exists $\sigma_{\mathrm{h}} \in \mathbb{R}$ such that $h - \frac{\sigma_{\mathrm{h}}}{2}\|\cdot\|^2$ is convex; for instance, we can take $\sigma_{\mathrm{h}} = -L_{\mathrm{h}}$ by Lemma 2.1(a).

**Remark 3.11.** Similar to Remark 3.6, the bound (3.37) also holds for $\tau \in (1,2)$. To see this, note that $c(\alpha)$ in (3.38) can also be written as

$$c(\alpha) = \hat{c}(\alpha) + 2(\tau-1)(L_{\mathrm{h}} - \sigma_{\mathrm{h}})\alpha,$$

where $\hat{c}(\alpha) \coloneqq 2L_{\mathrm{f}}(L_{\mathrm{f}}+L_{\mathrm{h}})\alpha^2 + ((2-\tau)L_{\mathrm{h}} - \tau L_{\mathrm{f}})\,\alpha - (2-\tau)$. Since $\frac{1}{L_{\mathrm{f}}+L_{\mathrm{h}}}$ is the positive root of $\hat{c}(\alpha)$ (see (3.31)) and $L_{\mathrm{h}} - \sigma_{\mathrm{h}} \geq 0$ by Remark 2.2, it holds that $c\left(\frac{1}{L_{\mathrm{f}}+L_{\mathrm{h}}}\right) = \frac{2(\tau-1)(L_{\mathrm{h}}-\sigma_{\mathrm{h}})}{L_{\mathrm{f}}+L_{\mathrm{h}}} \geq 0$. Since $c(0) = \tau - 2 < 0$ and $\bar{\alpha}_1$ is the positive root of $c(\alpha)$, it follows that $\bar{\alpha}_1 \leq \frac{1}{L_{\mathrm{f}}+L_{\mathrm{h}}}$. On the other hand, following the same arguments in Remark 3.6, it can be shown that $\tau/(2\eta^*) \leq \bar{\alpha}_1$, so (3.37) holds as claimed.

Lastly, we derive stepsize upper bounds for $\tau \geq 2$.

**Theorem 3.12** (Stepsize for $\tau \in [2,\infty)$)**.** *Suppose that Assumption* 1.1 *holds and $\rho_{\mathrm{h}} \in [0, L_{\mathrm{h}}]$ such that $h + \frac{\rho_{\mathrm{h}}}{2}\|\cdot\|^2$ is convex.*[5] *In addition, suppose that $f$ is $\sigma_{\mathrm{f}}$-strongly convex for some $\sigma_{\mathrm{f}} > 0$. Let $\tau \geq 2$ be such that*

$$\delta(\tau) \coloneqq (\tau\nu - \tau\theta_1 - 2(\tau-1)\theta_2)^2 - 8(\theta_0+\nu)(\tau-2) > 0, \tag{3.40}$$

*and*

$$\tau\nu - \tau\theta_1 - 2(\tau-1)\theta_2 > 0, \tag{3.41}$$

*where $\nu \coloneqq \frac{\sigma_{\mathrm{f}}}{L_{\mathrm{f}}+L_{\mathrm{h}}}$, $\theta_0 \coloneqq \frac{L_{\mathrm{h}}(L_{\mathrm{f}}^2-\sigma_{\mathrm{f}}^2)}{L_{\mathrm{f}}(L_{\mathrm{f}}+L_{\mathrm{h}})^2}$, $\theta_1 \coloneqq \frac{L_{\mathrm{h}}}{L_{\mathrm{f}}+L_{\mathrm{h}}}$ and $\theta_2 \coloneqq \frac{\rho_{\mathrm{h}}}{L_{\mathrm{f}}+L_{\mathrm{h}}}$, so that*

$$r(\mu) \coloneqq \tau^2(\theta_0+\nu)\mu^2 - \tau^2\left(\nu - \theta_1 - \frac{2(\tau-1)}{\tau}\theta_2\right)\mu + 2(\tau-2) \tag{3.42}$$

*has two distinct roots $\mu_*$ and $\mu^*$ with $0 \leq \mu_* < \mu^* \leq 1$. If $\{(x^k, y^k, z^k)\}$ is generated by Algorithm* 1 *with stepsize $\alpha = \frac{\tau\mu}{2(L_{\mathrm{f}}+L_{\mathrm{h}})}$, where $\mu > 0$ satisfies $\mu \in [\mu_*, \mu^*]$, then* (3.16) *holds with $c(\alpha)$ given by*

$$c(\alpha) = 2L_{\mathrm{h}}L_{\mathrm{f}}\alpha^2 + \left(\tau\mu\sigma_{\mathrm{f}} + \tau L_{\mathrm{h}} + 2(\tau-1)\rho_{\mathrm{h}} - \tau\sigma_{\mathrm{f}} - \frac{-\tau L_{\mathrm{h}}\sigma_{\mathrm{f}}^2\mu}{L_{\mathrm{f}}(L_{\mathrm{f}}+L_{\mathrm{h}})}\right)\alpha - (2-\tau). \tag{3.43}$$

*In particular, $\{V_k\}$ is nonincreasing if $\alpha \leq \bar{\alpha}$ and $\beta \leq \rho_{\mathrm{p}}^{-1}$, and strictly decreasing if at least one holds with strict inequality.*

*Proof.* $\sigma_{\mathrm{f}}$-strong convexity implies that

$$f(x) \geq f(\hat{x}) + \langle \nabla f(\hat{x}), x - \hat{x}\rangle + \frac{\sigma_{\mathrm{f}}}{2}\|x-\hat{x}\|^2, \quad \forall x, \hat{x} \in \mathbb{R}^n.$$

By the above inequality together with the $L_{\mathrm{f}}$-smoothness of $f$ and (2.5), we have

$$f(x) \geq f(\hat{x}) + \langle \nabla f(\hat{x}), x - \hat{x}\rangle + \frac{\mu}{2L_{\mathrm{f}}}\|\nabla f(x) - \nabla f(\hat{x})\|^2 + \frac{(1-\mu)\sigma_{\mathrm{f}}}{2}\|x-\hat{x}\|^2,$$

[5] By Assumption 1.1(a) and Remark 2.2, we see that there indeed exists $\rho_{\mathrm{h}} \geq 0$ such that $h + \frac{\rho_{\mathrm{h}}}{2}\|\cdot\|^2$ is convex.

for any $x, \hat{x} \in \mathbb{R}^n$ and $\mu \in [0,1]$. Following the arguments in Lemma 3.4, we obtain

$$\begin{aligned}
&V_{\Lambda,\xi}(y,z,x) \geq \\
&\quad Q_{\alpha(f+h)}(y^+;\hat{x}) + p(y^+) + g(y^+) + \left\langle \nabla f(\hat{x}) - \nabla f(x) - \frac{1}{\alpha}(\hat{x}-x), x - y^+ \right\rangle \\
&\quad + \langle \nabla h(x) - \nabla h(\hat{x}), y^+ - \hat{x}\rangle + \frac{\mu}{2L_{\mathrm{f}}}\|\nabla f(x) - \nabla f(\hat{x})\|^2 \\
&\quad - \left(\frac{L_{\mathrm{h}}}{2} + \frac{1}{2\alpha} - \frac{(1-\mu)\sigma_{\mathrm{f}}}{2}\right)\|\hat{x} - x\|^2 + \frac{1-\beta\rho_{\mathrm{p}}}{2\beta}\|y^+ - y\|^2.
\end{aligned}$$

By using this bound and inequality (3.39) with $\sigma_{\mathrm{h}} = -\rho_{\mathrm{h}}$, we obtain by following the same arguments and calculations in the proof of Theorem 3.5 that

$$\begin{aligned}
V_{k-1} - V_k \geq &\left(-\frac{\alpha}{\tau} + \frac{\mu}{2L_{\mathrm{f}}}\right)\left\|\nabla f(x^{k-1}) - \nabla f(x^k)\right\|^2 \\
&+ \left[\frac{2-\tau}{2\tau\alpha} - \frac{L_{\mathrm{h}}}{2} + \frac{(1-\mu)\sigma_{\mathrm{f}}}{2} - \frac{\tau-1}{\tau}\rho_{\mathrm{h}} - \frac{\alpha}{\tau}L_{\mathrm{h}}L_{\mathrm{f}}\right]\left\|x^k - x^{k-1}\right\|^2 \\
&+ \frac{1-\beta\rho_{\mathrm{p}}}{2\beta}\left\|y^k - y^{k-1}\right\|^2
\end{aligned} \tag{3.44}$$

By setting $\alpha = \frac{\tau\mu}{2(L_{\mathrm{f}}+L_{\mathrm{h}})}$ with $\mu \in (0,1]$, we have

$$\left(-\frac{\alpha}{\tau} + \frac{\mu}{2L_{\mathrm{f}}}\right)\left\|\nabla f(x^{k-1}) - \nabla f(x^k)\right\|^2 \geq \frac{\mu L_{\mathrm{h}}\sigma_{\mathrm{f}}^2}{2L_{\mathrm{f}}(L_{\mathrm{f}}+L_{\mathrm{h}})}\left\|x^{k-1} - x^k\right\|^2,$$

where the inequality follows from the strong convexity of $f$. Therefore, (3.16) holds with $c(\alpha)$ given by (3.43). To obtain a nonincreasing sequence $\{V_k\}$, we need to find the range of $\mu$ that makes $c(\alpha) \leq 0$. Plugging in $\alpha = \frac{\tau\mu}{2(L_{\mathrm{f}}+L_{\mathrm{h}})}$ in (3.43), some algebraic calculations lead to

$$c(\alpha) = c\left(\frac{\tau\mu}{2(L_{\mathrm{f}}+L_{\mathrm{h}})}\right) = \frac{1}{2}r(\mu),$$

with $r(\mu)$ given by (3.42). We note that $r(\mu)$ has two distinct roots $\mu_*$ and $\mu^*$ with $\mu_* < \mu^*$ if and only if its discriminant is positive. This condition is equivalent to (3.40). Meanwhile, since (3.41) holds and $\rho_{\mathrm{h}} \geq 0$, we see that

$$0 < \frac{\tau\nu - \tau\theta_1 - 2(\tau-1)\theta_2}{2\tau(\theta_0+\nu)} \leq \frac{\tau\nu}{2\tau\nu} = \frac{1}{2}.$$

That is to say, the first coordinate of the vertex of the parabola defined by $r(\mu)$ lies in $(0, \frac{1}{2}]$. Since $\tau \geq 2$, it follows that $0 \leq \mu_* < \mu^* \leq 1$. Thus, $c(\alpha) \leq 0$ for $\alpha = \frac{\tau\mu}{2(L_{\mathrm{f}}+L_{\mathrm{h}})}$ with $\mu \in [\mu_*, \mu^*]$, and $c(\alpha) < 0$ if $\mu \in (\mu_*, \mu^*)$. This completes the proof. □

**Remark 3.13** (Stepsize for $\tau = 2$)**.** Suppose that $\tau = 2$. Then (3.40) and (3.41) are equivalent to having $\sigma_{\mathrm{f}} > L_{\mathrm{h}} + \rho_{\mathrm{h}}$, and the roots of (3.42) are $\mu_* = 0$ and $\mu^* = \frac{L_{\mathrm{f}}(L_{\mathrm{f}}+L_{\mathrm{h}})(\sigma_{\mathrm{f}}-L_{\mathrm{h}}-\rho_{\mathrm{h}})}{L_{\mathrm{h}}(L_{\mathrm{f}}^2-\sigma_{\mathrm{f}}^2)+\sigma_{\mathrm{f}}L_{\mathrm{f}}(L_{\mathrm{f}}+L_{\mathrm{h}})}$. Hence, $c(\alpha)$ is strictly negative for all $\alpha$ such that

$$0 < \alpha < \frac{L_{\mathrm{f}}(\sigma_{\mathrm{f}} - L_{\mathrm{h}} - \rho_{\mathrm{h}})}{L_{\mathrm{h}}(L_{\mathrm{f}}^2 - \sigma_{\mathrm{f}}^2) + \sigma_{\mathrm{f}}L_{\mathrm{f}}(L_{\mathrm{f}}+L_{\mathrm{h}})}.$$

**Remark 3.14.** If $h \equiv 0$, the condition (3.41) automatically holds since $\sigma_{\mathrm{f}} > 0$. Moreover, the condition (3.40) with the constraint $\tau \geq 2$ is equivalent to having

$$2 \leq \tau < \frac{4}{1+\sqrt{1-\nu}} \quad \text{or} \quad \tau > \frac{4}{1-\sqrt{1-\nu}}. \tag{3.45}$$

On the other hand, the roots of (3.42) are given by

$$\mu_* = \frac{1}{2} - \frac{\sqrt{\nu(\nu\tau^2 - 8\tau + 16)}}{2\tau\nu} \quad \text{and} \quad \mu^* = \frac{1}{2} + \frac{\sqrt{\nu(\nu\tau^2 - 8\tau + 16)}}{2\tau\nu}.$$

Theorem 3.12 asserts that if we choose $\tau$ that satisfies (3.45), then $\{V_k\}$ is strictly decreasing provided that the stepsize $\alpha$ satisfies

$$\frac{\tau\nu - \sqrt{\nu(\nu\tau^2 - 8\tau + 16)}}{4\sigma_{\mathrm{f}}} < \alpha < \frac{\tau\nu + \sqrt{\nu(\nu\tau^2 - 8\tau + 16)}}{4\sigma_{\mathrm{f}}}, \tag{3.46}$$

which are the bounds obtained in (Themelis and Patrinos, 2020, Theorem 4.1), where the case $h \equiv p \equiv 0$ is considered. However, in the said result, the analysis restricts $\tau$ to satisfy only the first condition in (3.45), due to their imposed constraint that $\alpha L_{\mathrm{f}}$ must be at most 1. Meanwhile, the analysis we provide in the proof of Theorem 3.12 does not require this condition to establish the nonincreasing property, and therefore we have shown that the range of $\tau$ can be widened to include those that satisfy $\tau \geq \frac{4}{1-\sqrt{1-\nu}}$. For instance, if we are given $\nu = 3/4$, following (3.45) we can choose $\tau = 12$. By (3.46), for this value we may allow any stepsize $\alpha$ that satisfies $3 - \frac{\sqrt{21}}{3} < \alpha L_{\mathrm{f}} < 3 + \frac{\sqrt{21}}{3}$, and the lower bound is strictly greater than 1. Nevertheless, we point out that this wider range of $\tau$ provided above is only sufficient to guarantee monotonicity of $\{V_k\}$. If we want to establish boundedness of a sequence generated by Algorithm 1, the restriction that $\alpha(L_{\mathrm{f}} + L_{\mathrm{h}}) < 1$ will inevitably be required (see Proposition 3.15).

### 3.3 Subsequential convergence

After obtaining the stepsize upper bounds, our next goal is to show subsequential convergence and convergence rates of Algorithm 1. We first establish the boundedness of its iterate sequence.

**Proposition 3.15.** *Suppose that Assumption* 1.1 *holds,* $\Psi$ *has bounded level sets and* $\beta \in (0, \rho_{\mathrm{p}}^{-1}]$. *In addition, suppose that* $L_{\mathrm{f}} + L_{\mathrm{h}} > 0$, *and* $\alpha, \tau > 0$ *are chosen such that*

*(a)* $\tau \in (0, 1]$ *and* $\alpha \in (0, \bar{\alpha})$, *where* $\bar{\alpha}$ *is given in Theorem* 3.5*;*

*(b)* $\tau \in (1, 2)$ *and* $\alpha \in (0, \bar{\alpha})$, *where* $\bar{\alpha}$ *is given in Theorem* 3.9*; or*

*(c)* $\tau \geq 2$ *satisfies* (3.40) *and* (3.41), *and* $\alpha \in \left(\frac{\tau\mu_*}{2(L_{\mathrm{f}}+L_{\mathrm{h}})}, \frac{\tau\mu^*}{2(L_{\mathrm{f}}+L_{\mathrm{h}})}\right) \cap \left(0, \frac{1}{L_{\mathrm{f}}+L_{\mathrm{h}}}\right)$.

*If* $\{(x^k, y^k, z^k)\}$ *is generated by Algorithm* 1, *then*

*(i)* $\{(x^k, y^k, z^k)\}$ *is bounded; and*

*(ii)* $\left\|(x^k, y^k, z^k) - (x^{k-1}, y^{k-1}, z^{k-1})\right\| \to 0$.

*Proof.* We recall from (3.7) that $V_{\Lambda,\xi}(y,z,x) = \Phi_{\Lambda,\xi}(y^+;y,z,x)$ where $y^+ \in \arg\min_{w\in\mathbb{R}^n} \Phi_{\Lambda,\xi}(w;y,z,x)$. We then have from (3.8) that

$$\begin{aligned} V_{\Lambda,\xi}(y,z,x) &= Q_{\alpha(f+h)}(y^+;x) + Q_{\beta p}(y^+;y,\xi) + g(y^+) \\ &\geq f(y^+) + h(y^+) + \frac{1-\alpha(L_{\mathrm{f}}+L_{\mathrm{h}})}{2\alpha}\|y^+ - x\|^2 + Q_{\beta p}(y^+;y,\xi) + g(y^+) \\ &\overset{(3.15)}{\geq} \Psi(y^+) + \frac{1-\alpha(L_{\mathrm{f}}+L_{\mathrm{h}})}{2\alpha}\|y^+ - x\|^2 + \frac{1-\beta\rho_{\mathrm{p}}}{2\beta}\|y^+ - y\|^2, \end{aligned} \tag{3.47}$$

where the first inequality holds by $(L_{\mathrm{f}} + L_{\mathrm{h}})$-smoothness of $f + h$ and Lemma 2.1(a). Setting $(y,z,x) = (y^{k-1}, z^{k-1}, x^{k-1})$, we have from (3.47) that

$$V_{k-1} \geq \Psi(y^k) + \frac{1-\alpha(L_{\mathrm{f}}+L_{\mathrm{h}})}{2\alpha}\left\|y^k - x^{k-1}\right\|^2 + \frac{1-\beta\rho_{\mathrm{p}}}{2\beta}\left\|y^k - y^{k-1}\right\|^2. \tag{3.48}$$

By our choice of $(\alpha,\beta,\gamma)$, it follows from Theorems 3.5, 3.9 and 3.12 that $\{V_k\}$ is strictly decreasing. Meanwhile, we have from Remark 3.6 and Remark 3.11 that under conditions (a) and (b), respectively, it holds that $\alpha < \frac{1}{L_{\mathrm{f}}+L_{\mathrm{h}}}$. In condition (c), it is explicitly assumed that $\alpha < \frac{1}{L_{\mathrm{f}}+L_{\mathrm{h}}}$. Hence, the second term in (3.48) is always bounded below by zero. By our choice of $\beta$, the last term is also nonnegative. With these, we conclude that $\{\Psi(y^k)\}$ is bounded above, and therefore $\{y^k\}$ is a bounded sequence by the level-boundedness of $\Psi$. On the other hand, since $\Psi$ is closed, $\Psi$ is also bounded below. It thus follows from (3.48) that $\{\|y^k - x^{k-1}\|\}$ is bounded above. Since $\{y^k\}$ is bounded, it then follows that $\{x^{k-1}\}$ is also bounded. Finally, since $z^k = x^k + \alpha\nabla f(x^k)$ by (3.1) and $\nabla f$ is continuous, we obtain the boundedness of $\{z^k\}$. This completes the proof of part (i).

To prove part (ii), note that since $\{V_k\}$ is a bounded decreasing sequence, it follows that $V_{k-1} - V_k \to 0$. Meanwhile, under our hypotheses, the inequality (3.16) holds and the coefficient of $\|x^k - x^{k-1}\|^2$ is strictly positive since $\alpha < \frac{1}{L_{\mathrm{f}}+L_{\mathrm{h}}}$ as mentioned above. Using these facts, we have that $\|x^k - x^{k-1}\| \to 0$. From (3.23) and the $L_{\mathrm{f}}$-smoothness of $f$, we have

$$\left\|z^k - z^{k-1}\right\| = \left\|x^k - x^{k-1} + \alpha\left(\nabla f(x^k) - \nabla f(x^{k-1})\right)\right\| \leq (1+\alpha L_{\mathrm{f}})\left\|x^k - x^{k-1}\right\|, \tag{3.49}$$

and so $\|z^k - z^{k-1}\| \to 0$. From (3.3), we have $\|y^k - x^{k-1}\| \to 0$, and therefore $\|y^k - y^{k-1}\| \to 0$ by the triangle inequality. □

**Remark 3.16.** When $L_{\mathrm{f}} + L_{\mathrm{h}} = 0$, $x$ and $z$ play no role in the algorithm. Similar to (3.47), we can obtain the following inequality:

$$V_{\Lambda,\xi}(y,z,x) \geq \Psi(y^+) + \frac{1-\beta\rho_{\mathrm{p}}}{2\beta}\|y^+ - y\|^2.$$

Hence, if $\Psi$ has bounded level sets and $\beta < \frac{1}{\rho_{\mathrm{p}}}$, we obtain the boundedness of $\{y^k\}$ and that $\|y^k - y^{k-1}\| \to 0$.

We will next show that accumulation points of the $x$ and $y$ sequence generated from Algorithm 1 are *stationary points* of $\Psi$. We say that a point $x^*$ is a stationary point of $\Psi$ (see (Wen et al., 2018, Definition 4.1)) if

$$0 \in \nabla f(x^*) + \partial g(x^*) + \nabla h(x^*) + \partial p(x^*). \tag{3.50}$$

Note that any local minimizer of $\Psi$ is a stationary point of $\Psi$ (see (Pham and Thi, 1997, Theorem 2(i))). To establish stationarity of accumulation points, we will prove that these points

are fixed points of an operator associated with Algorithm 1, which we derive as follows. Observe that Algorithm 1 can be concisely written as fixed-point iterations of the operator $T_\Lambda : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n \times \mathbb{R}^n$, given by

$$(y^{k+1}, z^{k+1}) \in T_\Lambda(y^k, z^k), \quad k = 0, 1, \ldots,$$

where $T_\Lambda$ is defined as

$$T_\Lambda(y, z) \coloneqq \left\{ \begin{pmatrix} y^+ \\ z + \tau(y^+ - \operatorname{prox}_{\alpha f}(z)) \end{pmatrix} : y^+ \in P_\Lambda(y, z) \right\}. \tag{3.51}$$

We say that $(y, z)$ is a *fixed point* of $T_\Lambda$ if $(y^*, z^*) \in T_\Lambda(y^*, z^*)$. The set of fixed points of $T_\Lambda$ is denoted by $\operatorname{Fix}(T_\Lambda)$. The following proposition shows that fixed points of $T_\Lambda$ are stationary points of $\Phi$.

**Proposition 3.17.** *Suppose that $f$ is $L_\mathrm{f}$-smooth, $g$ is proper and closed, $h$ is continuously differentiable, and $p$ is continuous on an open set containing* $\operatorname{dom}(g)$. *In addition, let $\alpha, \beta, \gamma > 0$ satisfy $\frac{1}{\gamma} = \frac{1}{\alpha} + \frac{1}{\beta}$ with $\alpha < L_\mathrm{f}^{-1}$. If $(y^*, z^*) \in \operatorname{Fix}(T_\Lambda)$, then $y^*$ is a stationary point of $\Psi$.*

*Proof.* If $(y^*, z^*) \in \operatorname{Fix}(T_\Lambda)$, then $y^* \in P_\Lambda(y^*, z^*)$, $z^* \in z^* + \tau(y^* - \operatorname{prox}_{\alpha f}(z^*))$. Hence, by noting that $\operatorname{prox}_{\alpha f}$ is single-valued since $\alpha < L_\mathrm{f}^{-1}$, we have

$$y^* = \operatorname{prox}_{\alpha f}(z^*) \tag{3.52}$$

and thus

$$\frac{1}{\alpha}(z^* - y^*) = \nabla f(y^*). \tag{3.53}$$

From (3.52) and the optimality condition of (3.4), we also have that there exists $\xi^* \in \partial p(y^*)$ such that

$$\frac{1}{\alpha}(2y^* - z^* - \alpha\nabla h(y^*)) + \frac{1}{\beta}(y^* - \beta\xi^*) - \frac{1}{\gamma}y^* \in \partial g(y^*). \tag{3.54}$$

Adding (3.53) and (3.54) then leads to (3.50). □

Thus, we define the *residual function* for Algorithm 1 as

$$R(y, z) \coloneqq \operatorname{dist}(0, (y, z) - T_\Lambda(y, z)), \tag{3.55}$$

and use it as a measure of stationarity from the view that $R(y, z) = 0$ if and only if $(y, z) \in \operatorname{Fix}(T_\Lambda)$. This residual function reduces to that in Lee and Wright (2019) when $f \equiv 0$ and $p \equiv 0$, and to that in Liu and Takeda (2022) when $f \equiv 0$. The next theorem shows that accumulation points of the $(y, z)$-sequence generated by Algorithm 1 are fixed points of $T_\Lambda$.

**Theorem 3.18.** *Suppose that the hypotheses of Proposition* 3.15 *hold, and let $(x^*, y^*, z^*)$ be an accumulation point of $\{(x^k, y^k, z^k)\}$. Then $x^* = y^*$ and $(y^*, z^*) \in \operatorname{Fix}(T_\Lambda)$.*

*Proof.* Let $\{(x^{k_j}, y^{k_j}, z^{k_j})\}$ be such that $(x^{k_j}, y^{k_j}, z^{k_j}) \to (x^*, y^*, z^*)$. From the proof of Proposition 3.15, we have that $\left\|y^k - x^{k-1}\right\| \to 0$ and $\left\|y^k - y^{k-1}\right\| \to 0$. It then follows that $y^{k_j} \to x^*$ by the triangle inequality, and therefore $x^* = y^*$. To show that $(y^*, z^*) \in \operatorname{Fix}(T_\Lambda)$, we first note that since $x^* = y^*$ and $z^k = x^k + \alpha\nabla f(x^k)$ by (2.4), it follows that $z^* = y^* + \alpha\nabla f(y^*)$. Since $\alpha < \frac{1}{L_\mathrm{f}+L_\mathrm{h}} \le \frac{1}{L_\mathrm{f}}$, then $f + \frac{1}{2\alpha}\|\cdot\|^2$ is strongly convex by Remark 2.2 and therefore $y^* \in \operatorname{prox}_{\alpha f}(z^*)$. Thus, $z^* = z^* + \tau(y^* - \operatorname{prox}_{\alpha f}(z^*))$. Hence, to show that $(y^*, z^*) \in \operatorname{Fix}(T_\Lambda)$, it remains to prove that $y^* \in P_\Lambda(y^*, z^*)$. To this end, note that by the convexity and continuity of $\frac{\rho_\mathrm{p}}{2}\|\cdot\|^2 - p$ on $\operatorname{dom}(g)$ (by Assumption 1.1(d)) and the formula (3.13), we have from (Beck, 2017, Theorem 3.16) and (3.13)

that $\{\xi^{k_j}\}$ is bounded. Using again the continuity of $p$ on $\mathrm{dom}(g)$ together with (2.2) and the fact that $\{y^k\} \subseteq \mathrm{dom}(g)$, we see that accumulation points of $\{\xi^{k_j}\}$ belong to $\partial p(y^*)$. Without loss of generality, we may assume that $\xi^{k_j} \to \xi^*$ where $\xi^* \in \partial p(y^*)$.

Meanwhile, we have from (3.2) that

$$y^{k_j+1} \in \mathrm{prox}_{\gamma g}(u^{k_j}) \quad \text{where } u^{k_j} := \frac{\gamma}{\alpha}(2x^{k_j} - z^{k_j} - \alpha\nabla h(x^{k_j}) + \frac{\gamma}{\beta}(y^{k_j} - \beta\xi^{k_j}). \tag{3.56}$$

Note that $y^{k_j+1} \to y^*$ and $u^{k_j} \to u^* := \frac{\gamma}{\alpha}(2y^* - z^* - \alpha\nabla h(y^*)) + \frac{\gamma}{\beta}(y^* - \beta\xi^*)$, where we have used the fact that $\xi^{k_j} \to \xi^*$ and $x^* = y^*$. To prove the claim that $y^* \in P_\Lambda(y^*, z^*)$, we only need to show that $y^* \in \mathrm{prox}_{\gamma g}(u^*)$, noting that $y^* \in \mathrm{prox}_{\alpha f}(z^*)$ .

From Lemma 2.1(a) and (3.14), we have that for all $x \in \mathbb{R}^n$,

$$\begin{aligned}\Psi^* &\le f(x) + h(x) + p(x) + g(x)\\ &\le (f+h+p)(\bar{x}) + \left\langle \nabla(f+h)(\bar{x}) + \bar{\xi}, x - \bar{x}\right\rangle + \frac{L_{\mathrm{f}} + L_{\mathrm{h}} + \rho_{\mathrm{p}}}{2}\|x - \bar{x}\|^2 + g(x),\end{aligned}$$

where $\bar{x}$ is a fixed but arbitrary element of $\mathrm{dom}(g)$ and $\bar{\xi} \in \partial p(\bar{x})$. We may write the above inequality as

$$0 \le a + \langle v, x\rangle + \frac{L_{\mathrm{f}} + L_{\mathrm{h}} + \rho_{\mathrm{p}}}{2}\|x\|^2 + g(x) \tag{3.57}$$

for some $a \in \mathbb{R}$ and $v \in \mathbb{R}^n$. It then follows that $\liminf_{\|x\|\to\infty} \frac{g(x)}{\|x\|^2} \ge -\frac{L_{\mathrm{f}}+L_{\mathrm{h}}+\rho_{\mathrm{p}}}{2}$ by dividing both sides of (3.57) by $\|x\|^2$ and using Cauchy-Schwarz inequality. By (Rockafellar and Wets, 1998, Exercise 1.24), the fact that $\frac{1}{\gamma} > L_{\mathrm{f}} + L_{\mathrm{h}} + \frac{1}{\beta} \ge L_{\mathrm{f}} + L_{\mathrm{h}} + \rho_{\mathrm{p}}$ (that is, $\gamma < \frac{1}{L_{\mathrm{f}}+L_{\mathrm{h}}+\rho_{\mathrm{p}}}$), and (Rockafellar and Wets, 1998, Theorem 1.25), we know that $\{y^{k_j+1}\}$ given in (3.56) is bounded, with accumulation points lying in $\mathrm{prox}_{\gamma g}(u^*)$. Since $y^{k_j+1} \to y^*$, it follows that $y^* \in \mathrm{prox}_{\gamma g}(u^*)$, as desired. □

**Remark 3.19.** It is not difficult to verify that when $\alpha < L_{\mathrm{f}}^{-1}$, $(y^*, z^*) \in \mathrm{Fix}(T_\Lambda)$ is equivalent to the conditions

$$\begin{cases} y^* \in \mathrm{prox}_{\gamma g}\left(\frac{\gamma}{\alpha}(y^* - \alpha\nabla f(y^*) - \alpha\nabla h(y^*)) + \frac{\gamma}{\beta}T_{\beta p}(y^*)\right),\\ z^* = y^* + \alpha\nabla f(y^*). \end{cases} \tag{3.58}$$

Meanwhile, in certain applications Lu and Zhou (2019); Lu et al. (2019), it is of interest to consider cases in which $p$ takes the form

$$p(x) = -\max_{y\in\mathcal{Y}} \phi(x, y), \tag{3.59}$$

where $\phi(\cdot, y)$ is continuously differentiable and convex for all $y \in \mathcal{Y}$, and $\mathcal{Y}$ is a compact set. This satisfies Assumption 1.1(d) with $\rho_{\mathrm{p}} = 0$. If $p$ takes this form and $f$, $h$, $g$ are convex and satisfy Assumption 1.1(a)-(c), then from (Lu et al., 2019, Proposition 1) and (3.58), it follows that any point $(y^*, z^*) \in \mathrm{Fix}(T_\Lambda)$ ensures that $y^*$ corresponds to a D-stationary point of (1.1), which, as discussed in Pang et al. (2017), is generally a stronger notion than the stationarity we used in (3.50). Together with Theorem 3.18, this shows that under these conditions on $f, g, h$, and $p$, our proposed algorithm is globally subsequentially convergent to a point $(y^*, z^*)$ where $y^*$ is a D-stationary point of (1.1).

The following theorem provides the $O(1/T)$ iteration complexity of Algorithm 1.

**Theorem 3.20.** *Suppose that the hypotheses of Proposition* 3.15 *hold. There exists* $\omega > 0$ *such that* $\min_{k=1,2,\ldots,T} R(y^k, z^k)^2 \le \frac{\omega}{T}$, *for all* $T > 1$.

*Proof.* From (3.16), we have for any $N \geq 0$ that

$$\sum_{k=1}^{N+1} -\frac{c(\alpha)}{2\tau\alpha}\left\|x^k - x^{k-1}\right\|^2 \leq V_0 - \Psi^*, \tag{3.60}$$

where $\Psi^* \coloneqq \min_{x\in\mathbb{R}^n} \Psi(x)$. On the other hand, from (3.3), we get $y^{k+1} - y^k = x^k - x^{k-1} + \tau^{-1}(z^{k+1}-z^k) - \tau^{-1}(z^k - z^{k-1})$. It follows from the triangle inequality and the estimate $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$ (for $a, b, c \in \mathbb{R}$) that

$$\left\|y^{k+1} - y^k\right\|^2 \leq 3\left\|x^k - x^{k-1}\right\|^2 + 3\tau^{-2}\left\|z^{k+1} - z^k\right\|^2 + 3\tau^{-2}\left\|z^k - z^{k-1}\right\|^2.$$

Then

$$\begin{aligned}
\sum_{k=1}^{N}\left\|y^{k+1} - y^k\right\|^2 &\leq 3\sum_{k=1}^{N}\left\|x^k - x^{k-1}\right\|^2 + 3\tau^{-2}\sum_{k=1}^{N}\left\|z^{k+1} - z^k\right\|^2 + 3\tau^{-2}\sum_{k=1}^{N}\left\|z^k - z^{k-1}\right\|^2 \\
&\overset{(3.49)}{\leq} 3\sum_{k=1}^{N}\left\|x^k - x^{k-1}\right\|^2 + 3\tau^{-2}(1+\alpha L_{\mathrm{f}})^2\sum_{k=1}^{N}\left\|x^{k+1} - x^k\right\|^2 \\
&\quad + 3\tau^{-2}(1+\alpha L_{\mathrm{f}})^2\sum_{k=1}^{N}\left\|x^k - x^{k-1}\right\|^2 \\
&= \left(3 + 3\tau^{-2}(1+\alpha L_{\mathrm{f}})^2\right)\sum_{k=1}^{N}\left\|x^k - x^{k-1}\right\|^2 + 3\tau^{-2}(1+\alpha L_{\mathrm{f}})^2\sum_{k=1}^{N}\left\|x^{k+1} - x^k\right\|^2 \\
&\overset{(3.60)}{\leq} \left(3 + 6\tau^{-2}(1+\alpha L_{\mathrm{f}})^2\right)\left(-\frac{2\tau\alpha(V_0 - \Psi^*)}{c(\alpha)}\right).
\end{aligned}$$

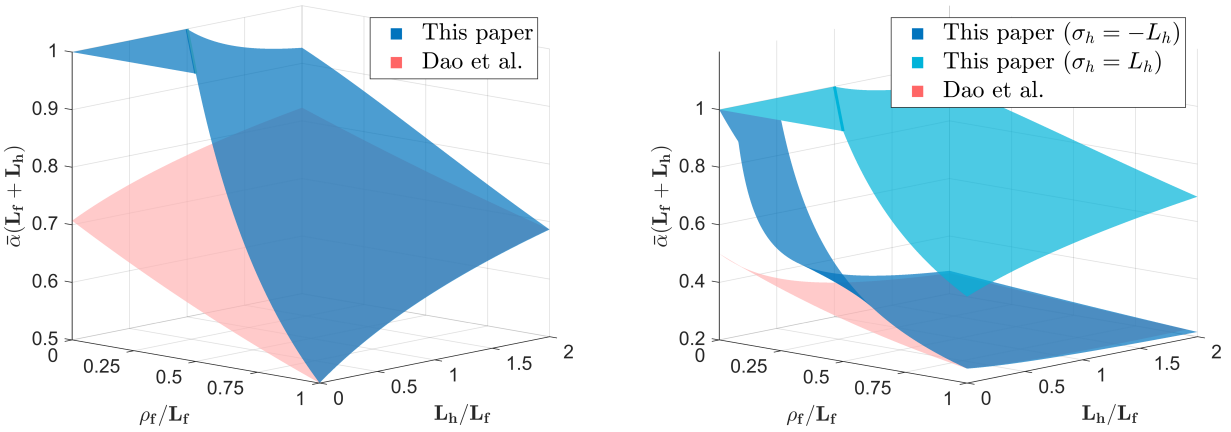
Combining this with (3.60) yields

$$\sum_{k=1}^{N}\left\|x^{k+1} - x^k\right\|^2 + \sum_{k=1}^{N}\left\|y^{k+1} - y^k\right\|^2 \leq \left(4 + 6\tau^{-2}(1+\alpha L_{\mathrm{f}})^2\right)\left(-\frac{2\tau\alpha(V_0 - \Psi^*)}{c(\alpha)}\right),$$

from where the result follows. □

While finalizing this paper, we became aware of a recent preprint by Dao et al. (2024) that proposed to solve (1.1) with $\rho_{\mathrm{p}} = 0$ using an algorithm similar to ours.[6] The main advantage of our analysis for this coincided special case is that our upper bounds for stepsizes are significantly larger than those in Dao et al. (2024); see Figs. 2 and 3 for geometric illustrations. Numerical results in Section 4.1.1 indicate that such a larger $\alpha$ is meaningful as it provides faster practical convergence. Meanwhile, subsequential convergence to stationary points is proven in both works, but ours allows for a wider range of stepsizes. Dao et al. (2024) also proved global convergence of the full sequence under a Kurdyka-Łojasiewicz hypothesis, wherein they used the techniques by Liu et al. (2019b). We note that a similar strategy can be employed to establish the same global convergence property for our algorithm with a wider range of stepsizes.

[6] Specifically, setting $\theta = 1$ in (Dao et al., 2024, Algorithm 1) coincides with our algorithm for the case of $\rho_{\mathrm{p}} = 0$.




(a) $\tau = 1$ (b) $\tau = 1.5$ and $\sigma_{\mathrm{h}} = \pm L_{\mathrm{h}}$

Figure 2: Comparison of stepsize upper bounds (denoted by $\bar{\alpha}$) computed in the present work and in Dao et al. (2024) for $\tau = 1$ and $\tau = 1.5$. We can see that the stepsize upper bounds in this work are always larger than those in Dao et al. (2024). Figure (a) demonstrates that the upper bound we have derived for the stepsize of the Davis-Yin splitting algorithm is $\frac{1}{L_{\mathrm{f}}+L_{\mathrm{h}}}$ provided that $\frac{L_{\mathrm{h}}}{L_{\mathrm{f}}} + \frac{2\rho_{\mathrm{f}}}{L_{\mathrm{f}}} \leq 1$ (see Theorem 3.5). Figure (b) demonstrates Theorem 3.9 with $\tau = 1.5$ using a rough estimate $-\sigma_{\mathrm{h}} = L_{\mathrm{h}}$, but this can be further improved when $-\sigma_{\mathrm{h}} < L_{\mathrm{h}}$. For instance, the case of $h(x) = \lambda\|x\|_2^2$ ($\sigma_{\mathrm{h}} = L_{\mathrm{h}}$) is also shown in figure (b) (see Remark 3.10).

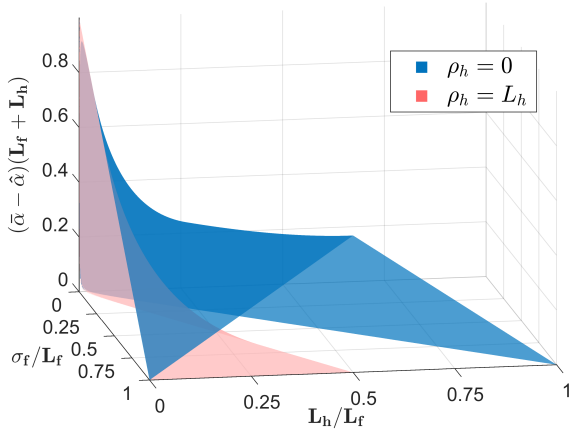



Figure 3: Gaps between the stepsize upper bounds in the present work and in Dao et al. (2024), denoted respectively by $\bar{\alpha}$ and $\hat{\alpha}$, when $\tau = 2$. In this case, Dao et al. (2024) always requires that $\sigma_{\mathrm{f}} > 2L_{\mathrm{h}}$. Using the rough estimate $\rho_{\mathrm{h}} = L_{\mathrm{h}}$ in Theorem 3.9, the red graph shows that the difference $\bar{\alpha} - \hat{\alpha}$ between our and their stepsizes is always nonnegative. For a convex $h$ (namely when $\rho_{\mathrm{h}} = 0$), the upper bound $\hat{\alpha}$ provided in Dao et al. (2024) does not change. As depicted in the blue graph, ours, on the other hand, provides a larger stepsize upper bound for the convex case, resulting in a bigger gap between $\bar{\alpha}$ and $\hat{\alpha}$. Ours also has a wider region in which $\tau = 2$ is applicable, since our condition on the parameters (namely $\sigma_{\mathrm{f}} > L_{\mathrm{h}} + \rho_{\mathrm{h}}$) becomes weaker than requiring $\sigma_{\mathrm{f}} > 2L_{\mathrm{h}}$.

# 4 Numerical experiments

In this section, we present experiments to corroborate the theoretical findings in the previous section. All experiments were conducted in MATLAB. For our method, we experiment with various values of $\tau$. For each $\tau$, we use the corresponding $\bar{\alpha}$ in Theorems 3.5, 3.9 and 3.12, and we calculate $\alpha$ by multiplying $\bar{\alpha}$ by 0.99 to ensure global subsequential convergence as suggested by Theorem 3.18. We also set $\beta = 1/\rho_{\text{p}}$ according to our theoretical results when the $p$ term is present, and set it to $\infty$ otherwise as described in Section 3.1. The value of $\gamma$ is then decided by $\alpha$ and $\beta$ through Step 0 of Algorithm 1.

For a fair comparison of different methods that can fit in the framework of Algorithm 1, instead of (3.55), we report their respective required number of iterations and running time to reach an approximate first-order optimality measure independent of the algorithm. In particular, in view of (3.58), for any algorithm compared, regardless of their specific reformulation/transformation, we stop it when its main iterate $y^k$ satisfies

$$\operatorname{dist}\left(0, y^k - \operatorname{prox}_{\hat{\gamma} g}\left(\frac{\hat{\gamma}}{\hat{\alpha}}\left(y^k - \hat{\alpha}\nabla f(y^k) - \hat{\alpha}\nabla h(y^k)\right) + \frac{\hat{\gamma}}{\hat{\beta}} T_{\hat{\beta} p}(y^k)\right)\right) \leq \epsilon \tag{4.1}$$

for some $\epsilon > 0$, where the choice of $f, g, h$, and $p$ are those used in our algorithm, and $(\hat{\alpha}, \hat{\beta}, \hat{\gamma}) = (\alpha, \beta, \gamma)$ are calculated following the above procedure for our algorithm with $\tau = 1$. From Remark 3.19, it is clear that at any stationary point $y^*$ of (1.1), the left-hand side of (4.1) vanishes, and thus it is a valid measure for the approximate first-order optimality measure. In all experiments in this section, we set $\epsilon = 10^{-6}$.

## 4.1 Nonnegative and low-rank matrix completion

The first experiment we consider is nonnegative matrix completion/factorization Lee and Seung (1999), whose goal is to recover missing entries of a partially observed matrix using nonnegative entries, and we impose the additional structure requirement that the recovered matrix should be low-rank. Given a data matrix $M \in \mathbb{R}^{m\times n}$, we let $\Omega$ denote the entries $(i, j)$ observed and $P_\Omega$ denote the associated projection operation such that $P_\Omega(X)_{i,j}$ outputs $X_{i,j}$ if $(i, j) \in \Omega$ and 0 otherwise. We consider (1.1) with $p(X) \equiv 0$ and

$$f(X) = \frac{\lambda_1}{2} \min_{Y \in \mathbb{R}^{m\times n}_+} \|X - Y\|_F^2, \quad g(X) = \lambda_2 \|X\|_*, \quad h(X) = \frac{1}{2}\|P_\Omega(X - M)\|_F^2, \tag{4.2}$$

where $\lambda_1, \lambda_2 \geq 0$ are weights for the respective terms, $\|\cdot\|_*$ is the nuclear norm, $\|\cdot\|_F$ is the Frobenius norm, and $\mathbb{R}^{m\times n}_+$ is the nonnegative orthant of $\mathbb{R}^{m\times n}$. We can easily see that (4.2) satisfies our assumptions with $L_f = \lambda_1, L_h = 1$, and $\rho_p = \rho_f = \rho_h = \sigma_h = 0$.

We follow Toh and Yun (2010) to generate $M \in \mathbb{R}^{m\times n}$ with a specific rank $r$ as the product of an $m \times r$ and an $r \times n$ matrix whose entries are all identically and independently distributed as standard Gaussian, and then randomly select $s$ entries in uniform random to form $\Omega$. We in particular use $m = n$ and consider $(n, s) \in \{(100, 1000), (300, 10000)\}$, $r \in \{10, 30\}$, $\lambda_1 = 10$ and $\lambda_2 = 5$. We set $\beta = \infty$ for all applicable methods due to the absence of the $p$ term. All algorithms are run with a cap of $30,000$ iterations.

### 4.1.1 Comparison of different bounds for $\alpha$

We first use our method with $\tau = 1$ to recover the DYS algorithm, and compare our obtained bound for $\bar{\alpha}$ with those in Bian and Zhang (2021); Dao et al. (2024) for the same setting. We use the same

Table 1: Comparison between different bounds for $\alpha$ in DYS using nonnegative and low-rank matrix completion. $*$ denotes that the maximum iteration of 30000 is reached. The respective values of $\bar{\alpha}$ in this experiment are: Bian and Zhang (2021): 0.0212, Dao et al. (2024): 0.0645, ours: 0.09.

| Bound | $n = 100, r = 10$ | | $n = 100, r = 30$ | | $n = 300, r = 10$ | | $n = 300, r = 30$ | |
|---|---|---|---|---|---|---|---|---|
| | iter | time | iter | time | iter | time | iter | time |
| Bian and Zhang (2021) | 14371 | 17.37 | 27335 | 32.64 | 14937 | 154.11 | *30000 | *322.91 |
| Dao et al. (2024) | 9087 | 11.89 | 17283 | 21.70 | 9445 | 100.31 | 21929 | 229.87 |
| Ours | 6516 | 8.71 | 12394 | 16.47 | 6773 | 73.02 | 15725 | 167.23 |

Table 2: Comparison between methods for nonnegative and low-rank matrix completion. $*$ denotes that the maximum iteration of 30000 is reached. $\tau = t$ denotes our method with the specified $\tau$.

| Method | $n = 100, r = 10$ | | $n = 100, r = 30$ | | $n = 300, r = 10$ | | $n = 300, r = 30$ | |
|---|---|---|---|---|---|---|---|---|
| | iter | time | iter | time | iter | time | iter | time |
| PG | 6516 | 8.48 | 12393 | 16.45 | 6773 | 72.24 | 15725 | 168.03 |
| DYS | 6516 | 8.71 | 12394 | 16.47 | 6773 | 73.02 | 15725 | 167.23 |
| $\tau = 1.1$ | 5963 | 8.03 | 11342 | 14.61 | 6198 | 65.64 | 14391 | 152.51 |
| $\tau = 1.2$ | 5505 | 7.33 | 10472 | 14.10 | 5723 | 60.82 | 13287 | 142.07 |
| $\tau = 1.3$ | 5122 | 6.47 | 9742 | 13.16 | 5324 | 56.25 | 12362 | 136.20 |
| $\tau = 1.4$ | 4796 | 6.07 | 9124 | 12.02 | 4985 | 52.89 | 11577 | 123.30 |
| $\tau = 1.5$ | 4518 | 5.76 | 8595 | 11.35 | 4697 | 49.69 | 10906 | 116.72 |
| $\tau = 1.6$ | 4279 | 5.47 | 8141 | 10.74 | 4448 | 47.04 | 10330 | 109.19 |
| $\tau = 1.7$ | **4074** | **5.28** | **7750** | **10.22** | **4235** | **44.96** | **9834** | **104.33** |
| $\tau = 1.8$ | 5087 | 6.57 | 9678 | 12.78 | 5288 | 56.09 | 12280 | 131.24 |
| $\tau = 1.9$ | 8270 | 10.70 | 15732 | 20.78 | 8595 | 91.02 | 19961 | 213.87 |

factor of 0.99 for all the obtained upper bounds to ensure that strict inequality holds, which is essential to guarantee theoretical convergence. The result in Table 1 shows that a larger $\alpha$ leads to faster convergence, and therefore our new upper bound for $\alpha$ provides not only theoretical but also practical improvements.

#### 4.1.2 Comparison with other methods

We next compare our method with different $\tau$'s. Since $f$ is not strongly convex, our theoretical results indicate that we should use $\tau < 2$. We thus run our method with $\tau \in \{1, 1.1, \cdots, 1.9\}$.[7]

We compare our method with the proximal gradient method (PG) by treating $f + h$ together as the smooth term, and we set the stepsize to be $0.99/(L_{\mathrm{f}} + L_{\mathrm{h}})$ following standard practice. In addition, as noticed in the previous part, for this problem, our method with $\tau = 1$ recovers the Davis-Yin splitting (DYS). According to the result of the previous experiment, for DYS we use the new upper bound for $\alpha$ obtained in this work. We can clearly see from the results in Table 2 that our method with $\tau \in (1, 1.9)$ constantly outperforms DYS (with our improved upper bound for $\alpha$) and PG, and DYS and PG are performing very similarly in this experiment.

[7] We have also experimented with $\tau \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$, but those values give worse performance, with smaller $\tau$ leading to slower convergence. We therefore omit these results in our experiments.

### 4.2 Cardinality-constrained problems

Our next experiment considers a least-squares problem with a penalized form of the cardinality constraint $\|x\|_0 \leq k$ for some given positive integer $k$. As argued in Gotoh et al. (2018), this constraint is equivalent to $\|x\|_1 - \|x\|_{(k)} = 0$, where $\|x\|_{(k)}$ is the Ky Fan $k$-norm that computes the sum of the top $k$ largest elements of the element-wise absolute value of $x$, and we follow Gotoh et al. (2018) to take $\|x\|_1 - \|x\|_{(k)}$ as a penalty in the objective function instead of enforcing the hard constraint. We also follow a common convention in machine learning to introduce a Tikhonov regularization of the form $\|x\|_2^2$ to improve the problem condition. This leads to the following setting for (1.1).

$$f(x) = \frac{1}{2}\|Ax - b\|_2^2, \quad g(x) = \lambda_2\|x\|_1, \quad p(x) = -\lambda_2\|x\|_{(k)}, \quad h(x) = \frac{\lambda_1}{2}\|x\|_2^2. \tag{4.3}$$

For (4.3), we have the following parameter estimates: $L_f = \|A^\top A\|_2, L_h = \lambda_1, \rho_f = \rho_h = \rho_p = 0$, and $\sigma_f$ equals to the smallest eigenvalue of $A^\top A$.

Given the existence of the $p$ term in this experiment, PG is not applicable, but by again treating $(f + h)$ as the smooth term, we can use the proximal DC algorithm described in (3.6). In this experiment, we set $\lambda_1 = 5$, $\lambda_2 = 10$, and $k = \lfloor n/10 \rfloor$. As (3.41) does not hold for any $\tau \geq 2$, for our method we use the same range of $\tau$ as the previous experiment. All algorithms are run with a cap of $100,000$ iterations. Results on publicly available real-world data sets are shown in Table 3.[8]

In terms of the number of iterations, we can see that our method with $\tau > 1$ is more efficient than PDC, and before reaching $\tau = 2$, the number of iterations of our method is roughly proportional to $1/\tau$. For those datasets on which $\tau = 2$ is feasible, we see that the per-iteration decrease for $\tau = 2$ calculated in the proof of Theorem 3.12 is very close to 0, which already suggested a slow convergence. We therefore see that the value $c(\alpha)$ could be a good indicator for the speed of practical convergence as well.

On the other hand, in terms of the total running time, we observe that by utilizing the problem structure wisely, our method also achieves significantly lower time cost per iteration. This shows that although it is always possible to solve the problem (1.1) by existing methods by considering less problem structures (in the worst case, one can always just consider it as a nonsmooth optimization problem and use a subgradient method to solve it), wisely exploiting problem structures can lead to much higher practical efficiency.

## 5 Conclusion and Future Work

In this work, we presented a new splitting algorithm for the four-term optimization problem (1.1). Our algorithm generalizes the Davis-Yin splitting algorithm for three-term optimization to allow for an additional nonsmooth term. We derived stepsize estimates for the proposed algorithm that ensure global subsequential convergence to stationary points of the objective function. A notable implication of our results is the significant improvement in the upper bound estimates for the stepsize of the DYS algorithm that guarantee subsequential convergence.

Our numerical experiments demonstrate that larger stepsizes can significantly accelerate the convergence of our proposed algorithm, emphasizing the critical role of stepsize selection in practical performance. Previous works, such as Wright et al. (2009); Lu and Zhou (2019), have shown that incorporating adaptive stepsizes safeguarded by line search techniques can improve the efficiency of PG and PDC, which are special cases of our framework. Our preliminary experiments suggest that

[8] Downloaded from `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`.

Table 3: Comparison between proximal DC (denoted by PDC) and our method for (4.3). $*$ denotes that the maximum iteration of 100000 is reached, $\tau = t$ denotes our method with the specified $\tau$, and $-$ denotes that this value of $\tau$ is not applicable for this problem.

| | a9a | | cpusmall_scale | | ijcnn1 | | phishing | | heart | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $m =$ | $n =$ | $m =$ | $n =$ | $m =$ | $n =$ | $m =$ | $n =$ | $m =$ | $n =$ |
| | 32561 | 123 | 8192 | 12 | 49900 | 22 | 11055 | 68 | 270 | 13 |
| Method | iter | time | iter | time | iter | time | iter | time | iter | time |
| PDC | 91859 | 67.21 | 37132 | 5.64 | 2942 | 1.92 | 7283 | 3.27 | *100000 | *2.42 |
| $\tau = 1$ | 91865 | 6.80 | 37133 | 1.34 | 2948 | 0.16 | 7282 | 0.40 | *100000 | *1.77 |
| $\tau = 1.1$ | 83514 | 6.02 | 33757 | 1.23 | 2680 | 0.14 | 6620 | 0.36 | 91675 | 1.56 |
| $\tau = 1.2$ | 76555 | 5.74 | 30944 | 1.18 | 2456 | 0.14 | 6069 | 0.34 | 84035 | 1.48 |
| $\tau = 1.3$ | 70664 | 5.05 | 28564 | 1.03 | 2267 | 0.13 | 5602 | 0.32 | 77571 | 1.42 |
| $\tau = 1.4$ | 65618 | 4.85 | 26524 | 1.10 | 2105 | 0.13 | 5202 | 0.30 | 72030 | 1.36 |
| $\tau = 1.5$ | 61242 | 4.64 | 24756 | 0.97 | 1965 | 0.12 | 4856 | 0.28 | 67228 | 1.20 |
| $\tau = 1.6$ | 57415 | 4.13 | 23209 | 0.86 | 1842 | 0.11 | 4553 | 0.26 | 63027 | 1.13 |
| $\tau = 1.7$ | 54038 | 4.09 | 21844 | 0.76 | 1733 | 0.11 | 4285 | 0.24 | 59319 | 1.17 |
| $\tau = 1.8$ | 51037 | 3.88 | 20630 | 0.74 | 1637 | 0.10 | 4047 | 0.23 | 56024 | 1.01 |
| $\tau = 1.9$ | **48350** | **3.63** | **19545** | **0.68** | **1550** | **0.09** | **3835** | **0.22** | **53075** | **0.94** |
| $\tau = 2$ | - | - | - | - | 1746 | 0.11 | - | - | 97789 | 1.83 |

such techniques, while not preserving iteration complexity guarantees, can substantially enhance empirical convergence speed. Motivated by these findings, we are actively exploring extensions of adaptive stepsize strategies to our more general algorithm, aiming to further enhance efficiency while maintaining solution quality and the convergence rate guarantees.

## References


N. T. An and N. M. Nam. Convergence analysis of a proximal point algorithm for minimizing differences of functions. *Optimization*, 66(1):129–147, 2016. 3, 7

Amir Beck. *First-Order Methods in Optimization*. SIAM - Society for Industrial and Applied Mathematics, Philadelphia, PA, United States, 2017. ISBN 978-1-611974-98-0. 20

Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific Optimization and Computation Series. Athena Scientific, 2016. ISBN 9781886529052. 5

Fengmiao Bian and Xiaoqun Zhang. A three-operator splitting algorithm for nonconvex sparsity regularization. *SIAM Journal on Scientific Computing*, 43(4):2809–2839, 2021. 2, 3, 4, 13, 14, 24, 25

Minh N. Dao, Tan Nhat Pham, and Phan Thanh Tung. Doubly relaxed forward-Douglas–Rachford splitting for the sum of two nonconvex and a DC function, 2024. arXiv:2405.08485. 4, 22, 23, 24, 25

Damek Davis and Wotao Yin. A three-operator splitting scheme and its optimization applications. *Set-Valued and Variational Analysis*, 25:829–858, 2017. 2, 3, 6

Jim Douglas and Henry H. Rachford. On the numerical solution of heat conduction problems in two and three space variables. *Transactions of the American Mathematical Society*, 82(2):421–439, 1956. 6

Jun-ya Gotoh, Akiko Takeda, and Katsuya Tono. DC formulations and algorithms for sparse optimization problems. *Mathematical Programming*, 169:141–176, 2018. 26

Ching-pei Lee and Stephen J. Wright. Inexact successive quadratic approximation for regularized optimization. *Computational Optimization and Applications*, 72:641–674, 2019. 20

Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999. 24

Guoyin Li and Ting Kei Pong. Douglas–Rachford splitting for nonconvex optimization with application to nonconvex feasibility problems. *Mathematical Programming*, 159(1-2):371–401, 2016. 2

Tianxiang Liu and Akiko Takeda. An inexact successive quadratic approximation method for a class of difference-of-convex optimization problems. *Computational Optimization and Applications*, 82: 141–173, 2022. 20

Tianxiang Liu, Ting-Kei Pong, and Akiko Takeda. A successive difference-of-convex approximation method for a class of nonconvex nonsmooth optimization problems. *Mathematical Programming Series B*, 176:339–367, 2019a. 2

Tianxiang Liu, Ting Kei Pong, and Akiko Takeda. A refined convergence analysis of pDCAe with applications to simultaneous sparse recovery and outlier detection. *Computational Optimization and Applications*, 73(1):69–100, 2019b. 22

Yanli Liu and Wotao Yin. An envelope for Davis-Yin splitting and strict saddle-point avoidance. *Journal of Optimization Theory and Applications*, 181:567–587, 2019. 7

Zhaosong Lu and Zirui Zhou. Nonmonotone enhanced proximal DC algorithms for a class of structured nonsmooth DC programming. *SIAM Journal on Optimization*, 29(4):2725–2752, 2019. 21, 26

Zhaosong Lu, Zirui Zhou, and Zhe Sun. Enhanced proximal DC algorithms with extrapolation for a class of structured nonsmooth DC minimization. *Mathematical Programming*, 176:369–401, 2019. 21

Yves Lucet. Fast Moreau envelope computation I: Numerical algorithms. *Numerical Algorithms*, 43:235–249, 2006. 2

Jong-Shi Pang, Meisam Razaviyayn, and Alberth Alvarado. Computing B-stationary points of nonsmooth DC programs. *Mathematics of Operations Research*, 42(1):95–118, 2017. 21

Ankit Parekh and Ivan W. Selesnick. Convex fused lasso denoising with non-convex regularization and its use for pulse detection. In *Proceedings of IEEE Signal Processing in Medicine and Biology Symposium*, pages 1–6, 2015. 2

Panagiotis Patrinos, Lorenzo Stella, and Alberto Bemporad. Douglas-Rachford splitting: Complexity estimates and accelerated variants. In *53rd IEEE Conference on Decision and Control*, pages 4234–4239, 2014. 7

Dinh Tao Pham and Hoai An Le Thi. Convex analysis approach to D.C. programming: Theory, algorithms and applications. *Acta Mathematica Vietnamica*, 22:289–355, 1997. 19

Emile Richard, Pierre-André Savalle, and Nicolas Vayatis. Estimation of simultaneously sparse and low rank matrices. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pages 1351–1358, 2012. 2

R. Tyrrell Rockafellar and Roger J-B Wets. *Variational Analysis*, volume 317 of *Grundlehren der Mathematischen Wissenschaften*. Springer, Berlin, 1998. ISBN 978-3-540-62772-2. 9, 21

Andreas Themelis and Panagiotis Patrinos. Douglas–Rachford splitting and ADMM for nonconvex optimization: Tight convergence results. *SIAM Journal on Optimization*, 30(1):149–181, 2020. 5, 8, 15, 18

Andreas Themelis, Lorenzo Stella, and Panagiotis Patrinos. Forward-backward envelope for the sum of two nonconvex functions: Further properties and nonmonotone linesearch algorithms. *SIAM Journal on Optimization*, 28(3):2274–2303, 2018. 7

Kim-Chuan Toh and Sangwoon Yun. An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. *Pacific Journal of optimization*, 6(615-640):15, 2010. 24

Bo Wen, Xiaojun Chen, and Ting Kei Pong. A proximal difference-of-convex algorithm with extrapolation. *Computational Optimization and Applications*, 69:297–324, 2018. 3, 7, 19

Stephen J. Wright, Robert D. Nowak, and Mário A. T. Figueiredo. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57:2479–2493, August 2009. 26

Penghang Yin, Yifei Lou, Qing He, and Jack Xin. Minimization of $\ell_{1-2}$ for compressed sensing. *SIAM Journal on Scientific Computing*, 37:A536–A563, 2015. 2