A four-operator splitting algorithm for nonconvex and nonsmooth optimization

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

July 3, 2024

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 step sizes, our algorithm exhibits global subsequential convergence to stationary points with a stationarity measure converging at a rate of $1/k$. 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

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

under the following assumptions.

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

(a) $f : \mathbb{R}^n \to (-\infty, \infty)$ is an $L_f$-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 $L_h$-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.
(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 $L_p$-weakly convex with $L_p \geq 0$; that is, $\frac{L_p}{2} \| \cdot \|^2 - p$ is convex.

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 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.

**DC optimization.** The DYS algorithm is not applicable to difference-of-convex (DC) regularized optimization problems, which take the form

$$
\min_{x \in \mathbb{R}^n} f(x) + g(x) - \hat{g}(x), \tag{1.2}
$$

where $g$ and $\hat{g}$ are convex. The limitation arises from the potential nonsmoothness of the concave component $-\hat{g}$. Meanwhile, by setting $h \equiv 0$ and $p := -\hat{g}$, we satisfy Assumption 1.1(c) and (d) with $L_h = L_p = 0$.

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

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

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 $M_{\lambda P_i}$ of each $P_i$ to obtain an approximate problem:

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

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 $P$,

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

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 $h \equiv 0$ and $p := \sum_{i=1}^r M_{\lambda P_i} \circ A_i$ so that Assumption 1.1(d) holds with $L_p = \sum_{i=1}^r \lambda_i^{-1}$.
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 $A, B, C$ are convex, and the $D_i$’s 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 $d_S(x) := \inf_{w \in S} \|w - x\|$ for any set $S$, this problem can be reformulated as the following problem that conforms the form of (1.1):

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} d_A^2(x) + \frac{1}{2} d_B^2(x) + \frac{1}{2} d_C^2(x) + \sum_{i=1}^{r} \frac{1}{2} d_{D_i}^2(x).$$

(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 non-smooth 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 $L_p = r$.

In this paper, we propose a splitting algorithm that generalizes the DYS algorithm to the four-term optimization problem (1.1), which encompasses a broader range of nonconvex and nonsmooth problem classes, including those described above. We prove that our algorithm is globally subsequentially convergent to stationary points for appropriately chosen stepsizes, with a first-order optimality measure converging at a rate of $1/k$. Another major contribution of this work is the derivation of upper bounds for stepsizes for such convergence results. When specialized to $p \equiv 0$, our results yield stepsize estimates for the DYS algorithm that are significantly larger than the existing bounds in Bian and Zhang (2021). Up to our knowledge, for this special case of $p \equiv 0$, our iteration complexity result is also new for the DYS algorithm. During the final stages of preparing this manuscript, we came across a recent preprint by Dao et al. (2024) that examines a specific case of our problem in which $L_p = 0$ in Assumption 1.1(d). Notably, our analysis yields upper bounds for stepsizes that are considerably larger than those reported in their work.

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 $\infty/\infty = 1$.

Let $\phi : \mathbb{R}^n \to (-\infty, \infty]$ be an extended-valued function. The domain of $\phi$ is given by the set $\text{dom}(\phi) = \{x \in \mathbb{R}^n : \phi(x) < \infty\}$. We say that $\phi$ is a proper function if $\text{dom}(\phi) \neq \emptyset$, 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 \( x \in \text{dom}(\phi) \) is defined as

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

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

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

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:

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

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

\[
\text{prox}_{\gamma\phi}(x) := \arg\min_{w \in \mathbb{R}^n} \phi(w) + \frac{1}{2\gamma}\|w - x\|^2, \quad \gamma > 0.
\]

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

\[
y \in \text{prox}_{\gamma\phi}(x) \implies x - y \in \gamma \partial \phi(y),
\]

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

\[
T_{\gamma \phi} := \text{Id} - \gamma \partial \phi,
\]

where \( \text{Id} \) 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

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

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

Let \( \phi : \mathbb{R}^n \rightarrow \mathbb{R} \) be a continuously differentiable function \( \phi : \mathbb{R}^n \rightarrow \mathbb{R} \). 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 \( \phi : \mathbb{R}^n \rightarrow \mathbb{R} \) 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 Patri- nos, 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

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

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

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

**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

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

where $\Lambda := (\alpha, \beta, \gamma)$, $\alpha, \beta \in (0, \infty]$ 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, $x \in \text{prox}_{\alpha f}(z)$ implies that $z - x = \alpha \nabla f(x)$ by (2.4), so we may also write $P_\Lambda$ as

$$
P_\Lambda(y, z) = \bigcup_{x \in \text{prox}_{\alpha f}(z)} \text{prox}_{\gamma g} \left( \frac{\gamma}{\alpha} T_{\alpha (f + h)}(x) + \frac{\gamma}{\beta} T_{\beta p}(y) \right). \tag{3.5}
$$
When \( L_p = 0 \), we in particular set \( \beta = \infty \), so that \( \gamma = \alpha \) by Step 0 of Algorithm 1. On the other hand, when \( L_t = L_n = 0 \), 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:

\[
y^{k+1} = \text{prox}_{\gamma g}(y^k - \gamma \partial p(y^k)),
\]

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

3. (Proximal DC algorithm). When \( f \equiv 0 \), the algorithm simplifies to

\[
y^{k+1} \in \text{prox}_{\gamma g}(y^k - \gamma \nabla h(y^k) - \gamma \partial p(y^k)),
\]

which is the proximal DC algorithm Wen et al. (2018) for solving (1.2).

### 3.2 Stepsize bounds

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

\[
V_{\Lambda,\xi}(y, z, x) := \min_{w \in \mathbb{R}^n} \Phi_{\Lambda,\xi}(w; y, z, x),
\]

where

\[
\Phi_{\Lambda,\xi}(w; y, z, x) := Q_{\alpha(f+h)}(w; x) + Q_{\beta p}(w; y, \xi) + g(w).
\]

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 \( \alpha < L_t^{-1} \) and \( x = \text{prox}_{\alpha f}(z) \) 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

\[
P_{\Lambda}(y, z) = \bigcup_{x \in \text{prox}_{\alpha f}(z)} \arg\min_{w \in \mathbb{R}^n} \Phi_{\Lambda;\xi}(w; y, z, x)
\]

for all \((y, z) \in \text{dom}(g) \times \mathbb{R}^n\), where \( P_{\Lambda} \) is given by (3.4).
Proof. Given \( x \in \text{prox}_\alpha f(z) \) and \( \xi \in \partial p(y) \), we have that

\[
\arg \min_{w \in \mathbb{R}^n} \Phi_{\Lambda, \xi}(w; y, z, x) = \arg \min_{w \in \mathbb{R}^n} f(x) + h(x) + \langle \nabla f(x) + \nabla h(x), w - x \rangle + \frac{1}{2\alpha} \|w - x\|^2 \\
+ p(y) + \langle \xi, w - y \rangle + \frac{1}{2\beta} \|w - y\|^2 + g(w)
\]

\[
= \arg \min_{w \in \mathbb{R}^n} \langle \nabla f(x) + \nabla h(x), w \rangle + \frac{1}{2\alpha} \|w\|^2 - \frac{1}{\alpha} \langle w, x \rangle \\
+ \langle \xi, w \rangle + \frac{1}{2\beta} \|w\|^2 - \frac{1}{\beta} \langle w, y \rangle + g(w)
\]

\[
= \arg \min_{w \in \mathbb{R}^n} \frac{1}{2\gamma} \|w\|^2 - \frac{1}{\alpha} \langle x - \alpha(\nabla f(x) + \nabla h(x)), w \rangle - \frac{1}{\beta} \langle y - \beta \xi, w \rangle + g(w)
\]

\[
= \arg \min_{w \in \mathbb{R}^n} \frac{1}{2\gamma} \|w\|^2 - \frac{1}{\alpha} \langle x - \alpha(\nabla f(x) - \alpha \nabla h(x)) + \frac{\gamma}{\beta}(y - \beta \xi), w \rangle + g(w)
\]

\[
= \text{prox}_{\gamma g} \left( \frac{\gamma}{\alpha} (x - \alpha \nabla f(x) - \alpha \nabla h(x)) + \frac{\gamma}{\beta}(y - \beta \xi) \right)
\]

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

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

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

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

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 \( \rho_i \geq 0 \) be such that \( f \) is \( \rho \)-weakly convex and \( L \geq L_i \) be such that \( L - \rho_i > 0 \). Then

\[
V_{\Lambda, \xi}(y, z, x) \geq Q_{\alpha(f + h)}(y^+; \hat{x}) + p(y^+) + g(y^+) + \langle \nabla f(\hat{x}) - \nabla f(x) - \frac{1}{\alpha}(\hat{x} - x), x - y^+ \rangle \\
+ \langle \nabla h(x) - \nabla h(\hat{x}), y^+ - \hat{x} \rangle + \frac{1}{2(L - \rho_i)} \|\nabla f(x) - \nabla f(\hat{x})\|^2 \\
- \left( \frac{L_i}{2} + \frac{1}{2\alpha} + \frac{\rho_i L}{2(L - \rho_i)} \right) \|\hat{x} - x\|^2 + \frac{1}{2\beta} \|y^+ - y\|^2
\]

for all \((y, z) \in \text{dom}(g) \times \mathbb{R}^n, x \in \text{prox}_\alpha f(z), \hat{x} \in \mathbb{R}^n, \xi \in \partial p(y), \) and \( y^+ \in \arg \min_{w \in \mathbb{R}^n} \Phi_{\Lambda, \xi}(w; y, z, x) \).
Proof. We have from Lemma 2.1(b) that for all \( \hat{x} \in \mathbb{R}^n \),
\[
Q_{\alpha f}(y^+; x) \geq \left( f(\hat{x}) + \langle \nabla f(\hat{x}), x - \hat{x} \rangle + \frac{1}{2(L - \rho_i)} \| \nabla f(x) - \nabla f(\hat{x}) \|^2 \right.
- \frac{\rho_i L}{2(L - \rho_i)} \| x - \hat{x} \|^2)
+ \langle \nabla f(x), y^+ - x \rangle + \frac{1}{2\alpha} \| y^+ - x \|^2
\]
\[
= f(\hat{x}) + \langle \nabla f(\hat{x}) - \nabla f(x), x - y^+ \rangle + \langle \nabla f(\hat{x}), y^+ - \hat{x} \rangle
+ \frac{1}{2(L - \rho_i)} \| \nabla f(x) - \nabla f(\hat{x}) \|^2 - \frac{\rho_i L}{2(L - \rho_i)} \| x - \hat{x} \|^2 + \frac{1}{2\alpha} \| y^+ - x \|^2,
\]
for any \( L \geq \rho_i \) such that \( L > \rho_i \). 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
\[
Q_{\alpha f}(y^+; x) \geq Q_{\alpha f}(y^+; \hat{x}) + \left( \nabla f(\hat{x}) - \nabla f(x) - \frac{1}{\alpha} (\hat{x} - x), x - y^+ \right)
+ \left( \frac{1}{2\alpha} + \frac{\rho_i L}{2(L - \rho_i)} \right) \| x - \hat{x} \|^2 + \frac{1}{2(L - \rho_i)} \| \nabla f(x) - \nabla f(\hat{x}) \|^2.
\] (3.11)

On the other hand, we have from Lemma 2.1(a) that
\[
h(x) + \langle \nabla h(x), y^+ - x \rangle
\geq \left( h(\hat{x}) - \langle \nabla h(\hat{x}), \hat{x} - x \rangle - \frac{L_b}{2} \| \hat{x} - x \|^2 \right) + \langle \nabla h(x), y^+ - x \rangle
\]
\[
= h(\hat{x}) + \langle \nabla h(\hat{x}), y^+ - \hat{x} \rangle + \langle \nabla h(x) - \nabla h(\hat{x}), y^+ - \hat{x} \rangle - \frac{L_b}{2} \| \hat{x} - x \|^2
\] (3.12)
for all \( \hat{x} \in \mathbb{R}^n \). By Assumption 1.1(d) and the fact that
\[
\partial \left( \frac{L_p}{2} \| y \|^2 - p(y) \right)(y) = L_p y - \partial p(y),
\] (3.13)
which holds by Rockafellar and Wets 1998 Exercise 8.8(c), we have
\[
\frac{L_p}{2} \| y^+ \|^2 - p(y^+) \geq \frac{L_p}{2} \| y \|^2 - p(y) + \langle L_p y - \xi, y^+ - y \rangle,
\]
for any \( \xi \in \partial p(y) \). This implies that
\[
p(y) \geq p(y^+) - \langle \xi, y^+ - y \rangle - \frac{L_p}{2} \| y^+ - y \|^2
\] (3.14)
and hence
\[
Q_{\gamma p}(y^+; y, \xi) \geq \left( p(y^+) + \frac{1 - \beta L_p}{2\beta} \| y^+ - y \|^2 \right).
\] (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. \( \square \)

Regarding the requirement of \( \rho_i \) in Lemma 3.4 by Assumption 1.1(a) and Remark 2.2, we see that there indeed exists \( \rho_i \geq 0 \) such that \( f + \frac{\rho_i}{2} \| \cdot \|^2 \) is convex.
We now show that \( \{V_{\tau, \xi_k}(y^k, z^k, x^k)\} \) is a nonincreasing sequence for appropriately chosen step-sizes. To simplify the notations, we denote
\[
V_k := V_{\tau, \xi_k}(y^k, z^k, x^k) = \Phi_{\tau, \xi_k}(y^{k+1}; y^k, z^k, x^k).
\]

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 \( L_i + L_h > 0 \), there exist a function \( c(\alpha) \) and a finite interval \( I \subseteq (0, \infty) \) such that (i) \( c(\alpha) \leq 0 \) on \( I \), (ii) \( c(\alpha) < 0 \) on the interior of \( I \), and (iii) the inequality
\[
V_{k-1} - V_k \geq -\frac{c(\alpha)}{2\tau\alpha} \| x^k - x^{k-1} \|^2 + \frac{1 - \beta L_p}{2\beta} \| y^k - y^{k-1} \|^2
\]
holds for all \( k \). For the case \( L_i + L_h = 0 \), 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 \( L_i + L_h > 0 \). If \( \{(x^k, y^k, z^k)\} \) is generated by Algorithm 1 with \( \tau \in (0, 1] \) and \( \alpha \in (0, \bar{\alpha}] \),
\[
\bar{\alpha} := \begin{cases}
\frac{1}{\bar{\tau}_i + \bar{\tau}_h} & \text{if } (2 - \tau)L_i - 2\rho_i \geq \tau L_h, \\
\frac{\bar{\tau}_i}{2\eta} & \text{otherwise},
\end{cases}
\]
and \( \eta^* \) is the positive root of
\[
q(\eta) := 2(2 - \tau)\eta^2 - \tau((2 - \tau)L_h + \rho_i \tau)\eta - \tau(\rho_i^2 + L_i L_h),
\]
then (3.16) holds with
\[
c(\alpha) := \begin{cases}
2L_i(L_i + L_h)\alpha^2 + ((2 - \tau)L_h - \tau L_i)\alpha - (2 - \tau) & \text{if } (2 - \tau)L_i - 2\rho_i \geq \tau L_h, \\
2L_i L_h \alpha^2 + (2 - \tau)L_h + \frac{\rho_i \tau(\eta^* + \rho_i)}{\eta^*} & \text{otherwise}.
\end{cases}
\]

In particular, \( \{V_k\} \) is nonincreasing if \( \alpha \leq \bar{\alpha} \) and \( \beta \leq L_p^{-1} \), 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 \( y^{k+1} \in \arg\min_{w \in \mathbb{R}^n} \Phi_{\tau, \xi_k}(w; y^{k}; z^k, x^k) \). By Lemma 3.3,
\[
V_k \leq Q_{\alpha(f+h)}(y^{k}; x^k) + p(y^k) + g(y^k).
\]
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
\[
V_{k-1} \geq Q_{\alpha(f+h)}(y^{k}; x^k) + p(y^k) + g(y^k)
+ \left\langle \nabla f(x^k) - \nabla f(x^{k-1}) - \frac{1}{\alpha}(x^k - x^{k-1}), x^{k-1} - y^k \right\rangle
+ \left\langle \nabla h(x^{k-1}) - \nabla h(x^k), y^k - x^k \right\rangle
+ \frac{1}{2(L - \rho_i)} \left\| \nabla f(x^{k-1}) - \nabla f(x^k) \right\|^2
- \left( \frac{L_h}{2} + \frac{1}{2\alpha} + \frac{\rho_i L}{2(L - \rho_i)} \right) \left\| x^k - x^{k-1} \right\|^2 + \frac{1 - \beta L_p}{2\beta} \left\| y^k - y^{k-1} \right\|^2.
\]
Subtracting (3.20) from (3.21), we get
\[
V_{k-1} - V_k \geq \left\langle \nabla f(x^k) - \nabla f(x^{k-1}) - \frac{1}{\alpha} (x^k - x^{k-1}), x^{k-1} - y^k \right\rangle \\
+ \left\langle \nabla h(x^{k-1}) - \nabla h(x^k), y^k - x^k \right\rangle + \frac{1}{2(L - \rho_i)} \left\| \nabla f(x^{k-1}) - \nabla f(x^k) \right\|^2 \\
- \left( \frac{L_h}{2} + \frac{1}{2\alpha} + \frac{\rho_i L}{2(L - \rho_i)} \right) \left\| x^k - x^{k-1} \right\|^2 + \frac{1 - \beta L_e}{2\beta} \left\| y^k - y^{k-1} \right\|^2
\]
(3.22)

Meanwhile, we have from (3.3) that \( y^k = \frac{1}{\tau} (x^k - z^{k-1}) + x^{k-1} \) and from (3.1) and (2.4) that
\[
z^k - z^{k-1} = (x^k + \alpha \nabla f(x^k)) - (x^{k-1} + \alpha \nabla f(x^{k-1})).
\]
(3.23)

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

which can also be rewritten as
\[
y^k - x^k = \left( \frac{1}{\tau} - 1 \right) (x^k - x^{k-1}) + \frac{\alpha}{\tau} (\nabla f(x^k) - \nabla f(x^{k-1}))
\]
(3.25)

With these, algebraic calculations lead to
\[
\left\langle \nabla f(x^k) - \nabla f(x^{k-1}) - \frac{1}{\alpha} (x^k - x^{k-1}), x^{k-1} - y^k \right\rangle \\
\geq \frac{\alpha}{\tau} \left\| \nabla f(x^k) - \nabla f(x^{k-1}) \right\|^2 + \frac{1}{\tau \alpha} \left\| x^k - x^{k-1} \right\|^2
\]
(3.26)

and
\[
\left\langle \nabla h(x^{k-1}) - \nabla h(x^k), y^k - x^k \right\rangle \geq \left( 1 - \frac{1}{\tau} \right) \left\langle \nabla h(x^k) - \nabla h(x^{k-1}), x^k - x^{k-1} \right\rangle \\
- \frac{\alpha}{\tau} \left\langle \nabla h(x^k) - \nabla h(x^{k-1}), \nabla f(x^k) - \nabla f(x^{k-1}) \right\rangle.
\]
(3.27)

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
\[
V_{k-1} - V_k \geq \left( -\frac{\alpha}{\tau} + \frac{1}{2(L - \rho_i)} \right) \left\| \nabla f(x^{k-1}) - \nabla f(x^k) \right\|^2 \\
+ \left[ \left( \frac{1}{\tau} - 1 \right) - \frac{L_h}{2} - \frac{\rho_i L}{2(L - \rho_i)} + \frac{\tau - 1}{\tau} L_h - \frac{\alpha}{\tau} L_h L_i \right] \left\| x^k - x^{k-1} \right\|^2 \\
+ \frac{1 - \beta L_e}{2\beta} \left\| y^k - y^{k-1} \right\|^2
\]
(3.29)

Now, we discuss two disjoint cases.
Case 1. Suppose that \((2 - \tau) L_i - 2 \rho_t \geq \tau L_h\). Then \(L_i - \rho_t \geq \frac{\tau}{2} (L_i + L_h) > 0\) and we may take \(L = L_i\) in (3.29). Let \(\hat{\alpha} > 0\) be such that \(\hat{\alpha} \geq \frac{\tau}{2(L_i - \rho_t)}\), and suppose that \(0 < \alpha \leq \hat{\alpha}\). Then

\[
-\frac{\alpha}{\tau} + \frac{1}{2(L_i - \rho_t)} \geq -\frac{\hat{\alpha}}{\tau} + \frac{1}{2(L_i - \rho_t)},
\]

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

\[
V_{k-1} - V_k \geq \left(2 - \frac{\tau}{2\hat{\alpha}} - \frac{\alpha}{\tau} \frac{L_i (L_i + L_h)}{\tau} + \frac{L_i}{2} - \frac{L_h}{2} + \frac{\tau - 1}{\tau} L_h\right) \|x^k - x^{k-1}\|^2 + \frac{1 - \beta L_p}{2\beta} \|y^k - y^{k-1}\|^2
\]

(3.30)

where

\[
c(\hat{\alpha}) := 2L_i (L_i + L_h) \hat{\alpha}^2 + ((2 - \tau) L_h - \tau L_i) \hat{\alpha} - (2 - \tau)
\]

(3.31)

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

Case 2. Suppose now that

\[
(2 - \tau) L_i - 2 \rho_t < \tau L_h.
\]

(3.32)

Given \(L \geq L_i\) with \(L - \rho_i > 0\), we define \(\hat{\alpha}(L) := \frac{\tau}{2(L_i - \rho_t)}\) and select \(\alpha \in (0, \hat{\alpha}(L)]\). Then

\[
-\frac{\alpha}{\tau} + \frac{1}{2(L - \rho_i)} \geq -\frac{\hat{\alpha}(L)}{\tau} + \frac{1}{2(L - \rho_i)} = 0.
\]

Hence, we have from (3.29) that

\[
V_{k-1} - V_k \geq -\frac{c(\hat{\alpha}(L))}{2\tau \hat{\alpha}(L)} \|x^k - x^{k-1}\|^2 + \frac{1 - \beta L_p}{2\beta} \|y^k - y^{k-1}\|^2
\]

(3.33)

where

\[
c(\hat{\alpha}(L)) := 2L_i L_h \hat{\alpha}(L)^2 + \left(2 - \tau\right) L_h + \frac{\rho_t \tau L}{L - \rho_t}\right) \hat{\alpha}(L) - (2 - \tau).
\]

(3.34)

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

\[
L^* := \min\{L : c(\hat{\alpha}(L)) \leq 0, L \geq L_i, L > \rho_i\},
\]

(3.35)

so that \(\hat{\alpha}(L^*)\) is the desired stepsize. By some routine calculations, it can be shown that \(c(\hat{\alpha}(L)) = -\frac{1}{2(L_i - \rho_t)^2} q(L - \rho_t) = -\frac{1}{4\tau^2} q(\eta),\) where \(q\) is given by the polynomial (3.18) and \(\eta := L - \rho_i\). Hence, if \(\eta^*\) is the (strictly) positive root of \(q\), then \(L^* := \max\{\eta^* + \rho_t, L_i\}\). We
now claim that $L^* = \eta^* + \rho_t$. If $L_t = \rho_t$, this immediately holds since $\eta^* > 0$. Suppose now that $L_t > \rho_t$. By the definition of $\hat{\alpha}(L)$, note that we may write $c$ as

$$c(\hat{\alpha}(L)) = c(\hat{\alpha}(L)) + 2LL_t\hat{\alpha}(L)^2 - \frac{\tau LL_t}{L - \rho_t}\hat{\alpha}(L)$$

for any $L > \rho_t$. Simplifying this expression, we obtain

$$c(\hat{\alpha}(L)) = 2L_t(L_h + L)\hat{\alpha}(L)^2 + \left(2 - \tau\right)L_h - \frac{\tau L(L_i - \rho_i)}{L - \rho_t}\hat{\alpha}(L) - (2 - \tau). \quad (3.36)$$

Since $L_t > \rho_t$, $c(\hat{\alpha}(L_t))$ is well-defined and can be calculated as

$$c(\hat{\alpha}(L_t)) = 2L_t(L_i + L_h)\hat{\alpha}(L_t)^2 + \left(2 - \tau\right)L_h - \frac{\tau L(L_i - \rho_i)}{L - \rho_t}\hat{\alpha}(L_t) - (2 - \tau)$$

$$= \left(\hat{\alpha}(L_t) - \frac{1}{L_i + L_h}\right)\left(2L_i(L_i + L_h)\hat{\alpha}(L_t) + (2 - \tau)(L_i + L_h)\right) + 2L_i(L - L_t)\hat{\alpha}(L)^2 + \frac{\tau\rho_i(L - L_t)}{L - \rho_t}\hat{\alpha}(L).$$

Since (3.32) implies that $\hat{\alpha}(L_t) > \frac{L_h + 2\rho_t}{L_i + L_h}$, it follows that $c(\hat{\alpha}(L_t)) > 0$. Hence, $L^* \neq L_t$ by the definition of $L^*$ in (3.35). The claim that $L^* = \eta^* + \rho_t$ 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 $L := \frac{\tau L_i + 2\rho_t}{2 - \tau}$. By (3.32), $L > L_t$ and $L > \rho_t$. In addition, for this choice of $L$, $c(\hat{\alpha}(L)) \leq 0$. By (3.35), it holds that $L \geq L^*$, ensuring $\frac{\tau}{2\eta^*} \geq \frac{\tau}{2(L - \rho_t)} = \frac{\tau}{2(L_h + \rho_t)}$. 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

$$c(\hat{\alpha}(L)) = \left[2L_t(L_i + L_h)\hat{\alpha}(L)^2 + (2 - \tau)L_hL_t\hat{\alpha}(L) - (2 - \tau)\right]$$

$$+ 2L_t(L - L_t)\hat{\alpha}(L)^2 + \left(\tau L_t - \frac{\tau L(L_i - \rho_i)}{L - \rho_t}\right)\hat{\alpha}(L)$$

$$= \left(\hat{\alpha}(L) - \frac{1}{L_i + L_h}\right)\left(2L_i(L_i + L_h)\hat{\alpha}(L) + (2 - \tau)(L_i + L_h)\right)$$

$$+ 2L_i(L - L_t)\hat{\alpha}(L)^2 + \frac{\tau\rho_i(L - L_t)}{L - \rho_t}\hat{\alpha}(L).$$

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

$$\bar{\alpha} \leq \frac{1}{L_f + L_h}. \quad (3.37)$$

**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

$$\bar{\alpha} = \begin{cases} \frac{1}{L_i + L_h} & \text{if } L_t - 2\rho_t \geq L_h, \\ \frac{2}{L_h + \rho_t + \sqrt{(L_h + \rho_t)^2 + 8(\rho_t^2 + L_tL_h)}} & \text{otherwise.} \end{cases}$$
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

$$
\frac{1}{2} \left( \frac{1}{\alpha} - \rho_t \right) - L_h - \left( \frac{1}{\alpha} + \frac{L_h}{2} \right) (2\alpha\rho_t + 2\alpha L_i + \alpha^2 L_i^2) > 0,
$$
or equivalently, under the constraint $\alpha > 0$,

$$
d(\alpha) := L_t^2 L_h \alpha^3 + 2(L_t^2 + L_h L_i + \rho_t L_h) \alpha^2 + (5\rho_t + 2L_h + 4L_i) \alpha - 1 < 0.
$$

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 $L_t - 2\rho_t \geq L_h$, as long as $L_t + L_h > 0$ and $\alpha > 0$, we have

$$
c(\alpha) - d(\alpha) = -L_t^2 L_h \alpha^3 - 2\rho_t L_h \alpha^2 - (5\rho_t + L_h + 5L_i) < 0.
$$

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

$$
c(\bar{\alpha}) = c(\bar{\alpha}) + 2\rho_t L_t \alpha^2 - \frac{\rho_t L_t}{L^* - \rho_t} \bar{\alpha}
$$

$$
= 2(L_t L_h + \rho_t L_i) \alpha^2 + \left( L_h + \frac{\rho_t (L^* - L_t)}{L^* - \rho_t} \right) \bar{\alpha} - 1
$$

$$
\leq 2(L_t L_h + L_i^2) \bar{\alpha}^2 + (L_h + \rho_t) \bar{\alpha} - 1,
$$

where the last inequality holds since $\rho_t \leq L_t$ and $L^* > \rho_t$. Then, provided that $L_t + L_h > 0$, we have

$$
c(\bar{\alpha}) - d(\bar{\alpha}) \leq -L_t^2 L_h \bar{\alpha}^3 - 2\rho_t L_h \bar{\alpha}^2 - (4\rho_t + L_h + 4L_i) < 0,
$$

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 stepsize is evident in Fig. 1.

**Remark 3.8** (The case $L_i + L_h = 0$). Suppose that $L_i + L_h = 0$, in which case $\alpha = \infty$ and $\gamma = \beta$. From (3.22), we immediately obtain

$$
V_{k-1} - V_k \geq \frac{1 - \beta L_p}{2\beta} \| y^k - y^{k-1} \|^2,
$$

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 $(1, 2)$.

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

$$
\hat{\alpha} := \begin{cases} 
\bar{\alpha}_1 & \text{if } \tau \leq 2\bar{\alpha}_1(L_i - \rho_t) \\
\frac{\tau}{2\bar{\alpha}_1} & \text{otherwise}
\end{cases}
$$
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).

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

$$c(\alpha) := 2 L_t (L_t + L_h) \alpha^2 + (\tau L_h - 2(\tau - 1)\sigma_h - \tau L_t) \alpha - (2 - \tau),$$

(3.38)

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

$$q(\eta) := 2(2 - \tau)\eta^2 - \tau(\tau L_h - 2(\tau - 1)\sigma_h + \rho_f \tau) \eta - \tau^2(\rho_f^2 + L_t L_h),$$

then (3.16) holds with $c(\alpha)$ given by (3.38) if $\tau \leq 2\bar{\alpha}_1(L_t - \rho_f)$, and

$$c(\alpha) = 2 L_t L_h \alpha^2 + \left(\tau L_h - 2(\tau - 1)\sigma_h + \frac{\rho_f \tau (\eta^* + \rho_f)}{\eta^*}\right) \alpha - (2 - \tau)$$

otherwise. In particular, $\{V_k\}$ is nonincreasing if $\alpha \leq \bar{\alpha}$ and $\beta \leq L_p^{-1}$, 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 $\nabla h - \sigma_h \text{Id}$ that

$$\left\langle \nabla h(x^{k-1}) - \nabla h(x^k), y^k - x^k \right\rangle \geq \left(\frac{\tau - 1}{\tau}\right) \sigma_h \|x^k - x^{k-1}\|^2 - \frac{\alpha}{\tau} L_t L_h \|x^k - x^{k-1}\|^2.$$  

(3.39)

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

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

**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 $\sigma_h = L_h$. In particular, when $h = 0$, we obtain that $\bar{\alpha} = \min \left\{\frac{1}{L_t}, \frac{2 - \tau}{2\rho_f}\right\}$ 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 $\sigma_h = L_h$ is when $h$ is a quadratic function with its Hessian being a positive multiple of the identity matrix.
where $\hat{\sigma}$ is the positive root of $c(\alpha)$ (see (3.31)) and $L_h - \sigma_h \geq 0$ by Remark 2.2. It holds that $c\left(\frac{1}{L_h+i}L_h\right) = 2(\tau-1)(L_h-\sigma_h) \geq 0$.

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

$$c(\alpha) = \hat{c}(\alpha) + 2(\tau-1)(L_h - \sigma_h)\alpha,$$

where $\hat{c}(\alpha) := 2L_i(L_i + L_h)\alpha^2 + ((2 - \tau)L_h - \tau L_i)\alpha - (2 - \tau)$. Since $\frac{\hat{c}(\alpha)}{L_i+L_h}$ is the positive root of $\hat{c}(\alpha)$ (see (3.31)) and $L_h - \sigma_h \geq 0$ by Remark 2.2, it holds that $c\left(\frac{1}{L_i+L_h}\right) = 2(\tau-1)(L_h-\sigma_h) \geq 0$.

Remark 3.12 (Stepsize for $\tau \in [2, \infty)$). Suppose that Assumption 1.1 holds, $f$ is $\sigma_f$-strongly convex for some $\sigma_f > 0$, and $\rho_h \in [0, L_h]$ such that $h + \frac{\rho_h}{2} \parallel \cdot \parallel^2$ is convex. Let $\tau \geq 2$ be such that

$$\delta(\tau) := (\tau \nu - \tau \theta_0 - 2(\tau - 1)\theta_0^2) - 8(\theta_0 + \nu)(\tau - 2) > 0,$$

(3.40)

and

$$\tau \nu - \tau \theta_0 - 2(\tau - 1)\theta_0^2 > 0,$$

(3.41)

where $\nu := \frac{\sigma_f}{L_i+L_h}$, $\theta_0 := \frac{L_h(L_i^2 - \sigma_f^2)}{L_i(L_i+L_h)^2}$, $\theta_1 := \frac{L_h}{L_i+L_h}$, and $\theta_2 := \frac{\rho_h}{L_i+L_h}$, so that

$$r(\mu) := \tau^2(\theta_0 + \nu)\mu^2 - \tau^2\left(\nu - \theta_0 - \frac{2(\tau - 1)}{\tau}\theta_0\right)\mu + 2(\tau - 2),$$

(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 $\alpha = \frac{\tau^2}{2(L_i+L_h)}$, where $\mu > 0$ satisfies $\mu \in [\mu_*, \mu^*]$, then (3.16) holds with $c(\alpha)$ given by

$$c(\alpha) = 2L_hL_i\alpha^2 + \left(\tau \mu \sigma_f + \tau L_h + 2(\tau - 1)\rho_h - \tau \sigma_f - \frac{\tau^2 \sigma_f^2}{L_i(L_i+L_h)}\right)\alpha - (2 - \tau).$$

(3.43)

In particular, $\{V_k\}$ is nonincreasing if $\alpha \leq \tilde{\alpha}$ and $\beta \leq \frac{L_i}{\rho_h}$, and strictly decreasing if at least one holds with strict inequality.

Proof. $\sigma_f$-strong convexity implies that

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

By the above inequality together with the $L_i$-smoothness of $f$ and (2.5), we have

$$f(x) \geq f(\hat{x}) + \langle \nabla f(\hat{x}), x - \hat{x} \rangle + \frac{\mu}{2L_i} \parallel \nabla f(x) - \nabla f(\hat{x}) \parallel^2 + \frac{(1 - \mu)\sigma_f}{2} \parallel x - \hat{x} \parallel^2,$$

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

$$V_{\lambda, \xi}(y, z, x) \geq$$

$$Q_{\alpha(f+\lambda)}(y^+; \hat{x}) + p(y^+) + g(y^+) + \left\langle \nabla f(\hat{x}) - \nabla f(x), \frac{1}{\alpha}(\hat{x} - x) - y^+ \right\rangle$$

$$+ \left(\nabla h(x) - \nabla h(\hat{x}), y^+ - \hat{x} \right) + \frac{\mu}{2L_i} \parallel \nabla f(x) - \nabla f(\hat{x}) \parallel^2$$

$$- \left(\frac{L_h}{2} + \frac{1}{2\alpha} - \frac{(1 - \mu)\sigma_f}{2}\right) \parallel \hat{x} - x \parallel^2 + \frac{1 - \beta L_i}{2\beta} \parallel y^+ - y \parallel^2.$$
By using this bound and inequality (3.39) with \( \sigma_b = -\rho_b \), we obtain by following the same arguments and calculations in the proof of Theorem 3.5 that

\[
V_{k-1} - V_k \geq \left( \frac{-\alpha}{\tau} + \frac{\mu}{2L_i} \right) \left\| \nabla f(x^{k-1}) - \nabla f(x^k) \right\|^2 + \frac{2 - \tau - \frac{L_h}{2} + \frac{(1 - \mu)\sigma_i}{2} - \frac{(1 - \mu)\rho_h}{2L_i} - \frac{\alpha}{\tau}L_hL_i}{\tau} \left\| x^k - x^{k-1} \right\|^2
+ \frac{1 - \beta L_p}{2\beta} \left\| y^k - y^{k-1} \right\|^2
\]

(3.44)

By setting \( \alpha = \frac{\tau \mu}{2(L_i + L_h)} \) with \( \mu \in (0, 1) \), we have

\[
\left( \frac{-\alpha}{\tau} + \frac{\mu}{2L_i} \right) \left\| \nabla f(x^{k-1}) - \nabla f(x^k) \right\|^2 \geq \frac{\mu L_h \sigma_i^2}{2L_i(L_i + L_h)} \left\| x^{k-1} - x^k \right\|^2,
\]

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 \( \alpha = \frac{\tau \mu}{2(L_i + L_h)} \) in (3.43), some algebraic calculations lead to

\[
c(\alpha) = c\left( \frac{\tau \mu}{2(L_i + L_h)} \right) = \frac{1}{2} r(\mu),
\]

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 \( \rho_b \geq 0 \), we see that

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

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 \( \alpha = \frac{\tau \mu}{2(L_i + L_h)} \) with \( \mu \in [\mu_*, \mu^*] \), and \( c(\alpha) < 0 \) if \( \mu \in (\mu_*, \mu^*) \). This completes the proof. \( \Box \)

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

\[
0 < \alpha < \frac{L_i(\sigma_i - L_h - \rho_h)}{L_h(L^2 - \sigma_i^2) + \sigma_i L_i(L_i + L_h)}.
\]

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

\[
2 \leq \tau < \frac{4}{1 + \sqrt{1 - \nu}} \quad \text{or} \quad \tau > \frac{4}{1 - \sqrt{1 - \nu}}.
\]

(3.45)

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

\[
\mu_* = \frac{1}{2} - \sqrt{\frac{\nu(\nu \tau^2 - 8\tau + 16)}{2\tau \nu}} \quad \text{and} \quad \mu^* = \frac{1}{2} + \sqrt{\frac{\nu(\nu \tau^2 - 8\tau + 16)}{2\tau \nu}}.
\]
In addition, suppose that \( \tau \) satisfies (3.45), then \( \{V_k\} \) is strictly decreasing provided that the stepsize \( \alpha \) satisfies
\[
\frac{\tau \nu - \sqrt{\nu(\nu \tau^2 - 8 \tau + 16)}}{4 \sigma_t} < \alpha < \frac{\tau \nu + \sqrt{\nu(\nu \tau^2 - 8 \tau + 16)}}{4 \sigma_t},
\]
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 \( \alpha L_\tau \) 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 \( 3 - \frac{\sqrt{37}}{3} < \alpha L_\tau < 3 + \frac{\sqrt{37}}{3} \), 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 \( \alpha (L_\tau + L_h) < 1 \) 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 \( \beta \in (0, L_\nu^{-1}] \). In addition, suppose that \( L_\tau + L_h > 0 \), 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.
- (c) \( \tau \geq 2 \) satisfies (3.40) and (3.41), and \( \alpha \in \left( \frac{\tau \nu^*}{2(L_\tau + L_h)}, \frac{\tau \nu^*}{2(L_\tau + L_h)} \right) \cap \left( 0, \frac{1}{L_\tau + L_h} \right) \).

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

- (i) \( \{x^k, y^k, z^k\} \) is bounded; and
- (ii) \( \| (x^k, y^k, z^k) - (x^{k-1}, y^{k-1}, z^{k-1}) \| \to 0 \).

**Proof.** We recall from (3.7) that \( V_{\Lambda, \xi}(y, z, x) = \Phi_{\Lambda, \xi}(y^\ast; y, z, x) \) where \( y^\ast \in \text{arg min}_{w \in \mathbb{R}^n} \Phi_{\Lambda, \xi}(w; y, z, x) \). We then have from (3.8) that
\[
V_{\Lambda, \xi}(y, z, x) = Q_{\alpha(f + h)}(y^\ast; x) + Q_{\beta p}(y^\ast; y, \xi) + g(y^\ast)
\]
\[
\geq f(y^\ast) + h(y^\ast) + \frac{1 - \alpha L_\tau + L_h}{2\alpha} \| y^\ast - x \|^2 + Q_{\beta p}(y^\ast; y, \xi) + g(y^\ast)
\]
\[
\begin{align*}
&\geq \Psi(y^\ast) + \frac{1 - \alpha L_\tau + L_h}{2\alpha} \| y^\ast - x \|^2 + \frac{1 - \beta L_\nu}{2\beta} \| y^\ast - y \|^2,
\end{align*}
\]
where the first inequality holds by \( (L_\tau + L_h) \)-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
\[
V_{k-1} \geq \Psi(y^k) + \frac{1 - \alpha L_\tau + L_h}{2\alpha} \| y^k - x^{k-1} \|^2 + \frac{1 - \beta L_\nu}{2\beta} \| y^k - y^{k-1} \|^2.
\]
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 \(\alpha < \frac{1}{L_i + L_n}\). In condition (c), it is explicitly assumed that \(\alpha < \frac{1}{L_i + L_n}\). 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 \(\{y^k\}\) 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 \(\alpha < \frac{1}{L_i + L_n}\) as mentioned above. Using these facts, we have that \(\|x^k - x^{k-1}\| \to 0\). From (3.23) and the \(L_i\)-smoothness of \(f\), we have

$$\|z^k - z^{k-1}\| = \|x^k - x^{k-1} + \alpha (\nabla f(x^k) - \nabla f(x^{k-1}))\| \leq (1 + \alpha L_i)\|x^k - x^{k-1}\|, \quad (3.49)$$

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 \(L_i + L_n = 0\), \(x\) and \(z\) play no role in the algorithm. We thus obtain the following inequality similar to (3.47):

$$V_{\lambda, \zeta}(y, z, x) \geq \Psi(y^+) + \frac{1 - \beta L_p}{2\beta} \|y^+ - y\|^2.$$ 

Hence, if \(\Psi\) has bounded level sets and \(\beta < \frac{1}{L_p}\), 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 \(w^*\) is a stationary point of \(\Psi\) (see Wen et al., 2018, Definition 4.1)) if

$$0 \in \nabla f(w^*) + \partial g(w^*) + \nabla h(w^*) + \partial p(w^*). \quad (3.50)$$

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 the defining operator of Algorithm 1. Observe that Algorithm 1 can be concisely written as fixed-point iterations of a certain map; in particular,

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

where \(T_{\Lambda} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n \times \mathbb{R}^n\) is given by

$$T_{\Lambda}(y, z) := \left\{ \left( z + \tau (y^+ - \text{prox}_{\alpha f}(z)) \right) : y^+ \in P_{\Lambda}(y, z) \right\}. \quad (3.51)$$

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

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

\[ y^* = \text{prox}_{\alpha f}(z^*) \tag{3.52} \]

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

\[ R(y, z) := \text{dist}(0, (y, z) - T_\Lambda(y, z)), \tag{3.55} \]

and use it as a measure of stationarity from the view that $R(y, z) = 0$ if and only if $(y, z) \in \text{Fix}(T_\Lambda)$. 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 $(y^*, z^*) \in \text{Fix}(T_\Lambda)$.

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 $\|y^k - x^{k-1}\| \to 0$ and $\|y^k - y^{k-1}\| \to 0$. It then follows that $y^k_j \to x^*$ by the triangle inequality, and therefore $x^* = y^*$. To show that $(y^*, z^*) \in \text{Fix}(T_\Lambda)$, 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 $\alpha < L_t^{-1} + \frac{1}{\alpha}$, then $\text{dom}(\frac{1}{\alpha})$ and hence $y^* \in \text{prox}_{\alpha f}(z^*)$.

Thus, $z^* = z^* + \tau(y^* - \text{prox}_{\alpha f}(z^*))$. Hence, to show that $(y^*, z^*) \in \text{Fix}(T_\Lambda)$, it remains to prove that $y^* \in P_\Lambda(y^*, z^*)$. To this end, note that by the convexity and continuity of $\frac{L_p}{2}\|\cdot\|^2 - p$ on $\text{dom}(g)$ (by Assumption 1.1 d)) and the formula (3.13), we have from Beck (2017) Theorem 3.16 and (3.13) that $\{\xi^k\}$ is bounded. Using again the continuity of $p$ on $\text{dom}(g)$ together with (2.2) and the fact that $\{y^k\} \subseteq \text{dom}(g)$, we see that accumulation points of $\{\xi^k\}$ 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

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

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 $y^* \in \text{prox}_{\gamma g}(u^*)$, noting that $y^* \in \text{prox}_{\alpha f}(z^*)$. 

19
From Lemma 2.1(a) and (3.14), we have that for all \( x \in \mathbb{R}^n \),
\[
\Psi^* \leq f(x) + h(x) + p(x) + g(x)
\leq (f + h + p)(\bar{x}) + \langle \nabla (f + h)(\bar{x}) + \bar{\xi}, x - \bar{x} \rangle + \frac{L_t + L_h + L_p}{2} \| x - \bar{x} \|^2 + g(x),
\]
where \( \bar{x} \) is a fixed but arbitrary element of \( \text{dom}(g) \) and \( \bar{\xi} \in \partial p(\bar{x}) \). We may write the above inequality as
\[
0 \leq a + \langle v, x \rangle + \frac{L_t + L_h + L_p}{2} \| x \|^2 + g(x) \tag{3.57}
\]
for some \( a \in \mathbb{R} \) and \( v \in \mathbb{R}^n \). It then follows that \( \liminf_{\| x \| \to \infty} \frac{g(x)}{\| x \|^2} \geq -\frac{L_t + L_h + L_p}{2} \) 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 \( \frac{1}{\gamma} > L_t + L_h + \frac{1}{\beta} \geq L_t + L_h + L_p \) (that is, \( \gamma < \frac{1}{L_t + L_h + L_p} \)), and (Rockafellar and Wets 1998 Theorem 1.25), we know that \{\( y^{k+1} \)\} given in (3.56) is bounded, with accumulation points lying in \( \text{prox}_{\gamma g}(u^*) \). Since \( y^{k+1} \to y^* \), it follows that \( y^* \in \text{prox}_{\gamma g}(u^*) \), as desired.

The following theorem provides the iteration complexity of Algorithm 1.

**Theorem 3.19.** Suppose that the hypotheses of Proposition 3.15 hold. There exists \( \omega > 0 \) such that \( \min_{k=1,2,\ldots,N} R(y^k, z^k)^2 \leq \frac{\omega}{N} \), for all \( N > 1 \).

**Proof.** From (3.16), we have for any \( N \geq 0 \) that
\[
\sum_{k=1}^{N+1} -\frac{c(\alpha)}{2\alpha} \| x^k - x^{k-1} \|^2 \leq V_0 - \Psi^*, \tag{3.58}
\]
where \( \Psi^* := \min_{w \in \mathbb{R}^n} \Psi(w) \). Note that \( c(\alpha) < 0 \) by the choice of \( \alpha \) in Proposition 3.15. On the other hand, from (3.3), we get \( y^{k+1} - y^k = x^k - x^{k-1} + \tau^{-1}(z^k - z^{k-1}) - \tau^{-1}(z^{k+1} - z^k) \). 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
\[
\| y^{k+1} - y^k \|^2 \leq 3 \| x^k - x^{k-1} \|^2 + 3\tau^{-2} \| z^{k+1} - z^k \|^2 + 3\tau^{-2} \| z^k - z^{k-1} \|^2.
\]
Then
\[
\sum_{k=1}^{N} \| y^{k+1} - y^k \|^2 \leq 3 \sum_{k=1}^{N} \| x^k - x^{k-1} \|^2 + 3\tau^{-2} \sum_{k=1}^{N} \| z^{k+1} - z^k \|^2 + 3\tau^{-2} \sum_{k=1}^{N} \| z^k - z^{k-1} \|^2
\leq 3 \sum_{k=1}^{N} \| x^k - x^{k-1} \|^2 + 3\tau^{-2}(1 + \alpha L_t)^2 \sum_{k=1}^{N} \| x^{k+1} - x^k \|^2
+ 3\tau^{-2}(1 + \alpha L_t)^2 \sum_{k=1}^{N} \| x^k - x^{k-1} \|^2
\leq (3 + 3\tau^{-2}(1 + \alpha L_t)^2) \sum_{k=1}^{N} \| x^k - x^{k-1} \|^2 + 3\tau^{-2}(1 + \alpha L_t)^2 \sum_{k=1}^{N} \| x^{k+1} - x^k \|^2
\leq (3 + 6\tau^{-2}(1 + \alpha L_t)^2) \left( -\frac{2\alpha (V_0 - \Psi^*)}{c(\alpha)} \right).\tag{3.58}
\]
Combining this with (3.58) yields
\[
\sum_{k=1}^{N} \left\| x^{k+1} - x^k \right\|^2 + \sum_{k=1}^{N} \left\| y^{k+1} - y^k \right\|^2 \leq \left( 4 + 6\tau^{-2}(1 + \alpha L_f) \right) \left( -\frac{2\tau\alpha(V_0 - \Psi^*)}{c(\alpha)} \right),
\]
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 \( L_p = 0 \) using an algorithm similar to ours. 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. 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.

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 times \( \frac{9}{10} \) to ensure global subsequential convergence as suggested by Theorem 3.18. For comparing different methods that can fit in the framework of Algorithm 1, we report their required number of iterations and running time to reach \( R(y, z) \leq 10^{-6} \).
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 \( \sigma_f > 2L_h \). Using the rough estimate \( \rho_h = L_h \) 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 \( \rho_h = 0 \)), 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 \( \sigma_f > L_h + \rho_h \)) becomes weaker than requiring \( \sigma_f > 2L_h \).

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

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

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 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 \((n, s) \in \{(100, 1000), (500, 10000)\} \), \( r \in \{10, 30\} \), \( \lambda_1 = 5 \) and \( \lambda_2 = 10 \). We set \( \beta = \infty \) for our method due to the absence of the \( p \) term. Since \( f \) is not strongly convex, our theoretical results indicate that we should use \( \tau < 2 \). We thus run our method with
Table 1: 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 $\tau$.

<table>
<thead>
<tr>
<th>Method</th>
<th>$n = 100, r = 10$</th>
<th>$n = 100, r = 30$</th>
<th>$n = 500, r = 10$</th>
<th>$n = 500, r = 30$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>iter</td>
<td>time</td>
<td>iter</td>
<td>time</td>
</tr>
<tr>
<td>PG</td>
<td>6269</td>
<td>36.57</td>
<td>12044</td>
<td>70.16</td>
</tr>
<tr>
<td>DYS</td>
<td>6892</td>
<td>49.13</td>
<td>13217</td>
<td>94.71</td>
</tr>
<tr>
<td>$\tau = 1.1$</td>
<td>6364</td>
<td>45.05</td>
<td>12223</td>
<td>87.18</td>
</tr>
<tr>
<td>$\tau = 1.2$</td>
<td>5922</td>
<td>42.04</td>
<td>11391</td>
<td>81.11</td>
</tr>
<tr>
<td>$\tau = 1.3$</td>
<td>5549</td>
<td>40.00</td>
<td>10686</td>
<td>76.25</td>
</tr>
<tr>
<td>$\tau = 1.4$</td>
<td>5231</td>
<td>37.35</td>
<td>10083</td>
<td>72.04</td>
</tr>
<tr>
<td>$\tau = 1.5$</td>
<td>4956</td>
<td>34.99</td>
<td>9564</td>
<td>68.71</td>
</tr>
<tr>
<td>$\tau = 1.6$</td>
<td>4720</td>
<td>33.73</td>
<td>9115</td>
<td>64.84</td>
</tr>
<tr>
<td>$\tau = 1.7$</td>
<td>4514</td>
<td>31.78</td>
<td>8725</td>
<td>62.30</td>
</tr>
<tr>
<td>$\tau = 1.8$</td>
<td>5516</td>
<td>39.46</td>
<td>10624</td>
<td>75.42</td>
</tr>
<tr>
<td>$\tau = 1.9$</td>
<td>8537</td>
<td>61.16</td>
<td>16306</td>
<td>116.10</td>
</tr>
</tbody>
</table>

$\tau \in \{1, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9\}$ and compare with the proximal gradient method (PG) by treating $f + h$ together as the smooth term. In addition, for this problem, our method with $\tau = 1$ recovers the Davis-Yin splitting (DYS). All algorithms are run with a cap of 30,000 iterations. We can clearly see from the results shown in Table 1 that our method constantly outperforms DYS and PG.

4.2 Cardinality-constrained problems

Our next experiment considers a least-square 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 add in a small Tikhonov regularization of the form $\|x\|_2^2$ to improve the problem condition. This leads to the following setting for (1.1)

$$f(x) = \frac{\lambda_1}{2} \|x\|_2^2, \quad g(x) = \lambda_2 \|x\|_1, \quad p(x) = -\lambda_2 \|x\|_{(k)} , \quad h(x) = \frac{1}{2} \|Ax - b\|_2^2. \quad (4.1)$$

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). Our experiment uses publicly available real-world data listed in Table 2. In this experiment, we set $\lambda_1 = 0.01$, $\lambda_2 = 0.005$, 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. The results are shown in Table 3. We can see that with appropriately selected $\tau$, our method performs similarly to proximal DC on ijcnn1, and is much more efficient than it on all the other datasets.

---

*We have also experimented with $\tau \in (0, 1)$, but those values tend to give worse performance. We therefore omit these results in our experiments.

Downloaded from [https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/](https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/)
<table>
<thead>
<tr>
<th>Data</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>colon-cancer</td>
<td>62</td>
<td>2000</td>
</tr>
<tr>
<td>duke</td>
<td>38</td>
<td>7129</td>
</tr>
<tr>
<td>ijcnn1</td>
<td>49900</td>
<td>22</td>
</tr>
<tr>
<td>phishing</td>
<td>11055</td>
<td>68</td>
</tr>
<tr>
<td>heart</td>
<td>270</td>
<td>13</td>
</tr>
</tbody>
</table>

Table 2: Datasets used in the experiment of least-square problems.

Table 3: Comparison between proximal DC (denoted by PDC) and our method for (4.1). * denotes that the maximum iteration of 100000 is reached. $\tau = t$ denotes our method with the specified $\tau$.

<table>
<thead>
<tr>
<th>Method</th>
<th>colon</th>
<th>duke</th>
<th>ijcnn1</th>
<th>phishing</th>
<th>heart</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDC</td>
<td>59796</td>
<td>63.88</td>
<td>19108</td>
<td>33.09</td>
<td>1367</td>
</tr>
<tr>
<td>$\tau = 1$</td>
<td>49934</td>
<td>53.09</td>
<td>13600</td>
<td>26.41</td>
<td>1502</td>
</tr>
<tr>
<td>$\tau = 1.1$</td>
<td>37384</td>
<td>39.76</td>
<td>8447</td>
<td>17.08</td>
<td>1650</td>
</tr>
<tr>
<td>$\tau = 1.2$</td>
<td>20352</td>
<td>21.71</td>
<td>5538</td>
<td>10.30</td>
<td>1834</td>
</tr>
<tr>
<td>$\tau = 1.3$</td>
<td>4772</td>
<td>5.07</td>
<td>5500</td>
<td>10.06</td>
<td>2066</td>
</tr>
<tr>
<td>$\tau = 1.4$</td>
<td>4672</td>
<td>4.97</td>
<td>5847</td>
<td>11.64</td>
<td>2370</td>
</tr>
<tr>
<td>$\tau = 1.5$</td>
<td>5148</td>
<td>5.48</td>
<td>6450</td>
<td>13.61</td>
<td>2787</td>
</tr>
<tr>
<td>$\tau = 1.6$</td>
<td>5968</td>
<td>6.39</td>
<td>7365</td>
<td>13.56</td>
<td>3396</td>
</tr>
<tr>
<td>$\tau = 1.7$</td>
<td>7335</td>
<td>7.84</td>
<td>8790</td>
<td>17.05</td>
<td>4380</td>
</tr>
<tr>
<td>$\tau = 1.8$</td>
<td>9881</td>
<td>10.52</td>
<td>11220</td>
<td>21.21</td>
<td>6257</td>
</tr>
<tr>
<td>$\tau = 1.9$</td>
<td>16308</td>
<td>17.54</td>
<td>16169</td>
<td>31.22</td>
<td>11457</td>
</tr>
</tbody>
</table>

5 Conclusion

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 experiments demonstrated the applicability and effectiveness of our proposed algorithm.

References


