Numerical Algebra, Control and Optimization

doi:10.3934/naco.2024005

A UNIFIED PROXIMAL GRADIENT METHOD FOR NONCONVEX COMPOSITE OPTIMIZATION WITH EXTRAPOLATION

Miao Zhang[1] and Hongchao Zhang[*1]

[1]Department of Mathematics, Louisiana State University
Baton Rouge, LA 70803-4918, USA

(Communicated by Jinyan Fan)

Abstract. In this paper, we propose a unified proximal gradient method with extrapolation (UPG-E) to solve a class of nonconvex and nonsmooth composite optimization. UPG-E provides a unified treatment to both convex and nonconvex problems, and adaptively estimates the nonconvexity modulus of the possibly nonconvex component function in the objective function. It is shown that without restarting the extrapolation, UPG-E achieves the optimal convergence rate of the first-order methods for solving convex composite optimization. In the case that the problem is nonconvex, the method performs as a proximal gradient method with extrapolation and guaranteed global convergence. Moreover, a linear convergence rate can be achieved by UPG-E under proper additional regularity assumptions. Our numerical experiments show the performance of UPG-E is very promising compared with other well-established proximal gradient methods in the literature.

1. Introduction.

Let us consider the composite optimization problem

$$\min_{\mathbf{x}\in\mathcal{X}} F(\mathbf{x}) := f(\mathbf{x}) + p(\mathbf{x}), \tag{1}$$

where $\mathcal{X} \subset \mathbb{R}^n$ is a closed convex set, f is Lipschitz continuously differentiable on an open set containing $\mathcal{X}$, but possibly nonconvex and $p : \mathcal{X} \to \mathbb{R}$ is a proper closed convex, but possibly nonsmooth, function. Note that the constraint $\mathbf{x} \in \mathcal{X}$ can be also formulated as an indicator function of $\mathcal{X}$ into the function p. In applications, the function f often serves as a model fitting term, while the function p usually plays as a regularization term to promote certain solution structure and/or increase the model stability. The model optimization problem (1) recently has many important applications in machine learning, statistical inference, and image processing (e.g., [9, 8, 7, 5]), especially when the component objective function f is nonconvex.

In theory, problem (1) can be solved by standard Proximal Gradient (PG) methods, which in some sense are natural extensions of the gradient descent methods to

2020 *Mathematics Subject Classification.* Primary: 90C06, 90C26; Secondary: 65Y20.

Key words and phrases. Nonconvex composite optimization, proximal gradient method, optimal convergence rate, accelerated gradient method, global and local linear convergence, nonsmooth optimization.

The authors are supported by the National Science Foundation under grants DMS2110722 and DMS2309549.

*Corresponding author: Hongchao Zhang.

solve the composite optimization. Hence, as the standard gradient descent methods, the PG methods could be quite slow. To accelerate the convergence speed, more efficient proximal gradient algorithms exploiting extrapolation techniques have been developed, that is to let $\mathbf{y}_k = \mathbf{x}_k + \beta_k(\mathbf{x}_k - \mathbf{x}_{k-1})$ with extrapolation factor $\beta_k \in [0, 1]$ and compute the next iteration as

$$\mathbf{x}_{k+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \langle \nabla f(\mathbf{y}_k), \mathbf{x}\rangle + p(\mathbf{x}) + \eta/2\|\mathbf{x} - \mathbf{y}_k\|_2^2 \right\},$$

where $\eta > 0$ is some constant and usually required to be greater than the Lipschitz constant of the gradient of f. When f is convex, one of the most well-known methods using extrapolation techniques to solve (1) is the Fast Iterative Shrinkage-thresholding Algorithm (FISTA) [1], which is in spirit similar to the Nesterov's accelerated gradient method [12, 13]. These methods can be shown to have optimal $\mathcal{O}(1/k^2)$ rate to reduce the function value gap, where k is the iteration number. However, all these traditional optimal methods using extrapolation can be only applied to convex optimization and can not guarantee convergence when f is not convex. There are more recent developments of proximal gradient methods to deal with the case that the component objective function f in (1) is not necessarily convex, such as the methods developed in [5, 17, 18]. However, there is no convergence rate analysis for these methods in either of the cases that the objective function is convex or not. A more recent effective remedy for proximal gradient methods to deal with the nonconvexity of f in (1) is proposed in [16] by the Proximal Gradient method with Extrapolation (PGE), which restricts the extrapolation parameter $0 \le \beta_k \le \overline{\beta}$ for some $\overline{\beta} < 1$. PGE is shown to have much better performance than standard PG methods. But this crucial threshold $\overline{\beta}$ given in [16] depends explicitly on the usually unknown nonconvexity modulus of f and a poor estimate of this parameter could significantly affect the practical performance. Moreover, when f is not explicitly known to be convex but has hidden convex structure, the restricted extrapolation applied in PGE does not automatically reduce to the optimal extrapolation used in the accelerated gradient methods such as FISTA or Nesterov's optimal methods. To overcome this drawback, some uniform proximal gradient methods were proposed in [3, 4, 10]. These methods would automatically reduce to an accelerated proximal gradient method when the objective function is convex, while the global convergence is still guaranteed even when the objective function is nonconvex. However, to ensure global convergence, [3] requires all iterates must belong to a bounded set, which might not be theoretically justified in many applications, and the method in [4] would just reduce to a simple proximal descent method without any momentum acceleration steps for nonconvex optimization. The method in [10] is developed based on modifications of FISTA. However, it is unclear how the extrapolation steps would apply when minimizing a nonconvex function. In addition, all these uniform gradient methods do not provide a linear convergence rate analysis for nonconvex optimization under certain additional proper regularity conditions, which are often practically satisfied in many applications.

Motivated from the extrapolation techniques to accelerate convergence for both convex and nonconvex optimization [1, 3, 16], we propose a new uniform proximal gradient method with momentum extrapolation, called UPG-E, to solve (1). As previous uniform gradient methods, our UPG-E guarantees global convergence for solving the possibly nonconvex problem (1) and will automatically reduce to an optimal proximal gradient method, which ensures optimal iteration complexity, when the objective is convex and no restart step is applied. Unlike the PGE method

Initialization: Given $\mathbf{x}_1 \in \mathcal{X}$, $\rho > 1$, $\lambda \in [0,1]$, integer $\bar{t} \geq 3$ and $L > \mathcal{L}$; Set $\breve{\mathbf{x}}_1 = \mathbf{x}_1$, $t_0 = 0$ and $\mu_0 = 0$.

For $t = 1, 2, 3, \ldots$

1. Set $\overline{\beta}_t = 2/(t + 1 - t_0)$.
2. If $\overline{\beta}_t == 1$, let $\widehat{\mathbf{x}}_t = \mathbf{x}_t$ and $\mu_t = \mu_{t-1}$;
3. Else choose the smallest integer $j \geq 0$ such that $\beta_t = \max\{\overline{\beta}_t, \tau_t\}$, $\widehat{\mathbf{x}}_t = \beta_t \breve{\mathbf{x}}_t + (1-\beta_t)\mathbf{x}_t$ and $\mu_t = \min\{\mu_{t-1} + \rho^j - 1, L\}$ satisfy $f(\mathbf{x}_t) - f(\widehat{\mathbf{x}}_t) - \langle \nabla f(\widehat{\mathbf{x}}_t), \mathbf{x}_t - \widehat{\mathbf{x}}_t \rangle \geq -\frac{\mu_t}{2}\|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2$, where $\tau_t = \lambda \underline{\tau}_t + (1-\lambda)\overline{\tau}_t$, $\underline{\tau}_t = \frac{1}{2}\left(1 - \sqrt{\frac{L-\mu_t}{L+\mu_t}}\right)$ and $\overline{\tau}_t = \frac{\mu_t}{L+\mu_t}$.
4. Set $\eta_t = 2L/(2-\beta_t)$ and $\gamma_t = \beta_t \eta_t$.
5. $\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \langle \nabla f(\widehat{\mathbf{x}}_t), \mathbf{x} \rangle + \frac{\eta_t}{2}\|\mathbf{x} - \widehat{\mathbf{x}}_t\|^2 + p(\mathbf{x}) \right\}$.
6. If $\mathrm{mod}(t, \bar{t}) == 0$, let $t_0 = t$;
7. Else $\breve{\mathbf{x}}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \langle \nabla f(\widehat{\mathbf{x}}_t), \mathbf{x} \rangle + \frac{\gamma_t}{2}\|\mathbf{x} - \breve{\mathbf{x}}_t\|^2 + p(\mathbf{x}) \right\}$.

end

ALG. 1. A unified proximal gradient method for nonconvex composite optimization with extrapolation (UPG-E)

developed in [16], UPG-E adaptively estimates the nonconvexity modulus of the possibly nonconvex function f, which essentially dynamically determines the extent of extrapolation that can be used without losing global convergence. Moreover, unlike the methods given in [3, 4], UPG-E does not require the boundedness of the iterates and the extrapolation techniques are applied even for the nonconvex minimization (see more detail discussions in the next section). Furthermore, under some error bound conditions and strictly separated isocost surface conditions on F, UPG-E achieves linear convergence rate on both the generated iterates and the associated objective function values.

The paper is organized as follows. We first present our unified proximal gradient method with extrapolation (UPG-E), Alg. 1, in Section 2. Then, we show the global convergence of UPG-E in Section 3. The linear convergence rate of UPG-E under additional proper assumptions is presented in Section 4. Numerical experiments evaluating the performance of UPG-E are given in Section 5. We finally draw some conclusions in Section 6.

2. Algorithm description.

Throughout this paper, we use the following assumption.

Assumption 2.1. The gradient of f is Lipschitz continuous, i.e., there exists a constant $\mathcal{L} > 0$ such that for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, we have

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq \mathcal{L}\|\mathbf{x} - \mathbf{y}\|. \tag{2}$$

From Assumption 2.1, there exists a constant $\mu \in [0, \mathcal{L}]$ such that

$$-\frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2 \leq f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \leq \frac{\mathcal{L}}{2}\|\mathbf{x} - \mathbf{y}\|^2, \tag{3}$$

for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$. Clearly, when f is a convex function, then (3) holds with the nonconvexity modulus $\mu = 0$.

Our uniform proximal gradient method with extrapolation (UPG-E) is proposed in Alg. 1, for which we have the following comments.

First, the nonconvex modulus of μ of f is adaptively estimated by a line search procedure in step 3 of Alg. 1. Hence, no prior knowledge of μ is needed. The line search finds a proper μ_t such that the line search condition (the inequality) in step 3 is satisfied. Since the parameter $\rho > 1$, we have ρ^j goes to infinity as j goes to infinity. So, by the Lipschitz continuity (2) of ∇f and the setting of $\mu_t = \min\{\mu_{t-1} + \rho^j - 1, L\}$ with $L > \mathcal{L}$, the line search condition in step 3 will be satisfied for j sufficiently large. So, the line search in step 3 is well-defined.

Second, a momentum extrapolation is used in Alg. 1 with a dynamic momentum factor $\theta_t := 1 - \beta_t \in [0, 1]$. Note that the scalar τ_t in step 3 is dynamically adjusted as a convex combination of $0 \le \underline{\tau}_t \le \overline{\tau}_t \le 1/2$, which depends on the nonconvex modulus estimation μ_t of f. When f is convex, it follows from the initial setting $\mu_0 = 0$ that the line search condition in step 3 will be satisfied with $j = 0$ and $\mu_t = 0$ for all $t \ge 1$. Then, we will have from the definitions of $\overline{\tau}_t$ and $\underline{\tau}_t$ in step 3 that $\overline{\tau}_t = \underline{\tau}_t = 0$, which gives $\tau_t = \lambda\underline{\tau}_t + (1-\lambda)\overline{\tau}_t = 0$ for any $\lambda \in [0, 1]$. So, we have $\beta_t = \max\{\overline{\beta}_t, \tau_t\} = \overline{\beta}_t = 2/(t + 1 - t_0)$ for all $t \ge 1$, where t_0 is a nonnegative integer with initial value zero and is possibly adjusted in step 6. In particular, step 6 resets t_0 to be t whenever $\mathrm{mod}(t, \bar{t}) = 0$ for restarting the extrapolation at $\mathbf{x}_{t+1}$ every $\bar{t}$ iterations. Here, $\bar{t}$ is a fixed positive integer, which is an algorithm parameter, and $\mathrm{mod}(t, \bar{t})$ returns the remainder after division of t by $\bar{t}$. If no restart is applied, i.e., $t_0 = 0$, for example when the restart parameter $\bar{t} = \infty$ or $t \le \bar{t}$, we will have $\beta_t = 2/(t + 1)$ and $\theta_t = 1 - \beta_t = (t - 1)/(t + 1)$, which is just the standard extrapolation factor of Nesterov's optimal gradient method. On the other hand, by restarting the extrapolation after every $\bar{t} \ge 3$ iterations, UPG-E will have a linear convergence under certain proper conditions and often has better practical performance. However, when f is nonconvex, μ_t could be strictly positive, and in fact, would reach a positive limit $\overline{\mu} > 0$ after finite number of iterations. As a result, τ_t would approach a limit $\overline{t} > 0$ (see (26)) after finite number of iterations, which by $\beta_t \ge \tau_t$ ensures that the extrapolation factor $\theta_t = 1 - \beta_t \le 1 - \overline{\tau}$ is strictly less than one. This is a key condition for ensuring global convergence when f is nonconvex. Note that θ_t can be arbitrarily close to one when f is convex. The adaptive updating formula of τ_t given by Alg. 1 is derived from our analysis. On the other hand, observe that even when f is nonconvex, it is still possible that $\mu_t = 0$ for all $t \ge 1$ in practical computations. This essentially implies that the optimal aggressive extrapolation for convex optimization could be even applied to nonconvex optimization in practice without losing convergence. Furthermore, we can see that for the most nonconvex case, i.e. $\mu = L$, the extrapolation factor θ can be still as large as $1/2$.

Third, two proximal gradient steps were performed in steps 5 and 7 of Alg. 1. In our UPG-E algorithm, we assume that the function $p(\cdot)$ and the convex set $\mathcal{X}$ are simple such that these convex proximal subproblems can be solved relatively easily or by closed form solution. When f is convex and $\bar{t} = \infty$, we have from step 4 that the proximal parameter $\gamma_t = \beta_t\eta_t = 2L\beta_t/(2 - \beta_t) = 2L/t$, which corresponds to a more aggressive stepsize $1/\gamma_t = t/(2L)$ as t goes to infinity. And even for the most nonconvex case, i.e. $\mu = L$, we have $\gamma_t = 2L/3$ for all $t \ge 2$ and $\mathrm{mod}(t, \bar{t}) \ne 1$, which is smaller than the gradient Lipschitz constant estimation L often used by proximal gradient descent methods.

3. **Global convergence.** In this section, to discuss the global convergence of UPG-E, Alg. 1, we further need the following assumptions.

Assumption 3.1. Assume p has a strongly convex modulus $\nu \geq 0$, i.e., for any $\mathbf{x} \in \mathcal{X}$, $\mathbf{y} \in \mathcal{X}$ and $\mathbf{p} \in \partial p(\mathbf{x})$, it has

$$p(\mathbf{y}) - p(\mathbf{x}) - \langle \mathbf{p}, \mathbf{y} - \mathbf{x} \rangle \geq \frac{\nu}{2}\|\mathbf{x} - \mathbf{y}\|^2, \tag{4}$$

where $\partial p(\mathbf{x})$ is the subdifferential of the proper closed convex function p at $\mathbf{x}$.

Note that $\nu = 0$ simply means p is a convex function, instead of a strongly convex function.

Assumption 3.2. Assume the function value of F on $\mathcal{X}$ is bounded below, i.e., we have $\overline{F} > -\infty$, where $\overline{F} := \inf_{x \in \mathcal{X}} F(\mathbf{x})$.

We first show that when f is a convex function and no restart extrapolation is used, i.e. by setting $\bar{t} = \infty$, Alg. 1 will be just reduced to an accelerated gradient method for solving convex composite optimization. In this case the convergence properties of Alg. 1 are rather standard and similar convergence results have been established in the literature [3]. Hence, here we just state the following convergence theorem and only provide a sketch of its proof.

Theorem 3.3. *Suppose the Assumptions 2.1 and 3.1 hold, and f is a convex function. If the solution set of problem (1) is not empty, for the iterates generated by Alg. 1 with $\bar{t} = \infty$, we have*

$$F(\mathbf{x}_{t+1}) - F(\mathbf{x}^*) \leq \frac{2L}{t(t+1)}\|\mathbf{x}^* - \mathbf{x}_1\|^2 \tag{5}$$

and

$$\min_{k \in \{1,\ldots,t\}} \|\mathbf{g}(\widehat{\mathbf{x}}_k)\|^2 \leq \frac{24L^3}{(L-\mathcal{L})t^2(t+1)}\|\mathbf{x}^* - \mathbf{x}_1\|^2, \tag{6}$$

where $\mathbf{g}(\widehat{\mathbf{x}}_k) = \eta_k(\widehat{\mathbf{x}}_k - \mathbf{x}_{k+1})$ and $\mathbf{x}^*$ is any optimal solution of (1).

Proof. Since $\mu_0 = 0$ and f is a convex function, we can see from Alg. 1 that $\mu_t = 0$ for all $t \geq 0$, which implies $\underline{\tau}_t = \frac{1}{2}\left(1 - \sqrt{(L-\mu_t)/(L+\mu_t)}\right) = 0$ and $\overline{\tau}_t = \mu_t/(L+\mu_t) = 0$ for all $t \geq 1$. Hence, we have $\tau_t = 0$ and $\beta_t = \overline{\beta}_t$ for all $t \geq 1$. In this case, if $\bar{t} = \infty$, we will have $\mathrm{mod}(t, \bar{t}) = t$ for all $t \geq 1$, which by step 5 gives $t_0 = 0$. Then, $\beta_t = \overline{\beta}_t = 2/(t+1-t_0) = 2/(t+1)$ and Alg. 1 is just reduced to an accelerated gradient method for solving convex composite optimization. Then, following the similar convergence proofs given in [3], it is not difficult to show that the iterates generated by Alg. 1 have the following property: for any $\mathbf{x} \in \mathcal{X}$, we have

$$\begin{aligned} F(\mathbf{x}_{t+1}) - F(\mathbf{x}) \;\leq\; & (1-\beta_t)(F(\mathbf{x}_t) - F(\mathbf{x})) + \frac{\beta_t\gamma_t}{2}\left[\|\mathbf{x} - \breve{\mathbf{x}}_t\|^2 - \|\mathbf{x} - \breve{\mathbf{x}}_{t+1}\|^2\right] \\ & - \frac{\eta_t - \mathcal{L}}{2\eta_t^2}\|\mathbf{g}(\widehat{\mathbf{x}}_t)\|^2 - \frac{\eta_t}{2}\|\mathbf{x}_{t+1} - \widetilde{\mathbf{x}}_{t+1}\|^2, \end{aligned} \tag{7}$$

where $\widetilde{\mathbf{x}}_{t+1} = \beta_t\breve{\mathbf{x}}_{t+1} + (1-\beta_t)\mathbf{x}_t$, $\eta_t - \mathcal{L} = 2L/(2-\overline{\beta}_t) - \mathcal{L} = L(t+1)/t - \mathcal{L} > L - \mathcal{L} > 0$ and $\mathbf{g}(\widehat{\mathbf{x}}_t) = \eta_t(\widehat{\mathbf{x}}_t - \mathbf{x}_{t+1})$. Dividing $\Gamma_t = 2L/(t(t+1))$, $t \geq 1$, on both sides of (7), for $t \geq 2$, we obtain

$$\frac{1}{\Gamma_t}\left(F(\mathbf{x}_{t+1}) - F(\mathbf{x})\right) + \frac{\eta_t - \mathcal{L}}{2\eta_t^2\Gamma_t}\|\mathbf{g}(\widehat{\mathbf{x}}_t)\|^2 + \frac{\eta_t}{2\Gamma_t}\|\mathbf{x}_{t+1} - \widetilde{\mathbf{x}}_{t+1}\|^2$$

$$\leq \frac{1}{\Gamma_{t-1}}(F(\mathbf{x}_t) - F(\mathbf{x})) + \frac{\beta_t\gamma_t}{2\Gamma_t}\left[\|\mathbf{x} - \breve{\mathbf{x}}_t\|^2 - \|\mathbf{x} - \breve{\mathbf{x}}_{t+1}\|^2\right],$$

which by $\eta_t = L(t+1)/t$, $\beta_t = 2/(t+1)$ and $\gamma_t = 2L/t$ can be simplified to

$$\begin{aligned}&\frac{1}{\Gamma_t}\left(F(\mathbf{x}_{t+1}) - F(\mathbf{x})\right) + \frac{L-\mathcal{L}}{4L^3(t+1)/t^3}\|\mathbf{g}(\widehat{\mathbf{x}}_t)\|^2 + \frac{(t+1)^2}{4}\|\mathbf{x}_{t+1} - \widetilde{\mathbf{x}}_{t+1}\|^2\\ &\leq \frac{1}{\Gamma_{t-1}}(F(\mathbf{x}_t) - F(\mathbf{x})) + \|\mathbf{x} - \breve{\mathbf{x}}_t\|^2 - \|\mathbf{x} - \breve{\mathbf{x}}_{t+1}\|^2. \qquad (8)\end{aligned}$$

When $t = 1$, by (7) and $\beta_1 = 1$, we have

$$\begin{aligned}&\frac{1}{\Gamma_1}\left(F(\mathbf{x}_2) - F(\mathbf{x})\right) + \frac{L-\mathcal{L}}{8L^3}\|\mathbf{g}(\widehat{\mathbf{x}}_1)\|^2 + \|\mathbf{x}_2 - \widetilde{\mathbf{x}}_{t+1}\|^2\\ &\leq \|\mathbf{x} - \breve{\mathbf{x}}_1\|^2 - \|\mathbf{x} - \breve{\mathbf{x}}_2\|^2. \qquad (9)\end{aligned}$$

Adding (8) and (9) for $t = 1, \ldots, k$, we have

$$\begin{aligned}&\sum_{t=1}^{k}\left(\frac{L-\mathcal{L}}{4L^3(t+1)/t^3}\|\mathbf{g}(\widehat{\mathbf{x}}_t)\|^2 + \frac{(t+1)^2}{4}\|\mathbf{x}_{t+1} - \widetilde{\mathbf{x}}_{t+1}\|^2\right) + \frac{1}{\Gamma_k}\left(F(\mathbf{x}_{k+1}) - F(\mathbf{x})\right)\\ &\leq \|\mathbf{x} - \breve{\mathbf{x}}_1\|^2 = \|\mathbf{x} - \mathbf{x}_1\|^2, \qquad (10)\end{aligned}$$

for any $\mathbf{x} \in \mathcal{X}$. Then, taking $\mathbf{x} = \mathbf{x}^*$ in (10) and noticing $\Gamma_t = 2L/(t(t+1))$, we can derive (5) and (6) by direct calculations. □

In the following we focus on studying the convergence of Alg. 1 when f is not necessarily a convex function. We first have the following lemma.

Lemma 3.4. *Suppose the Assumptions 2.1 and 3.1 hold. Then, for the iterates generated by Alg. 1, we have*

$$\begin{aligned}F(\mathbf{x}_{t+1}) \quad \leq \quad & F(\mathbf{x}_t) + \frac{\mu_t + \gamma_t/\beta_t}{2}\|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2\\ & - \frac{\gamma_t/\beta_t}{2}\|\widetilde{\mathbf{x}}_{t+1} - \mathbf{x}_t\|^2 - \frac{\eta_t - \mathcal{L}}{2\eta_t^2}\|\mathbf{g}(\widehat{\mathbf{x}}_t)\|^2\\ & - \frac{\beta_t\nu}{2}\|\breve{\mathbf{x}}_{t+1} - \mathbf{x}_t\|^2 - \frac{\eta_t + \nu}{2}\|\mathbf{x}_{t+1} - \widetilde{\mathbf{x}}_{t+1}\|^2, \qquad (11)\end{aligned}$$

where

$$\widetilde{\mathbf{x}}_{t+1} = \beta_t\breve{\mathbf{x}}_{t+1} + (1-\beta_t)\mathbf{x}_t. \qquad (12)$$

Proof. We first observe that all the iterates $\mathbf{x}_t$, $\breve{\mathbf{x}}_t$ and $\widehat{\mathbf{x}}_t$ are contained in $\mathcal{X}$ and $\beta_t \in (0, 1]$ for all $t \geq 1$. Then, by the definition of $\widetilde{\mathbf{x}}_{t+1}$ in (12), we also have $\widetilde{\mathbf{x}}_{t+1} \in \mathcal{X}$, since $\mathcal{X}$ is a convex set. By (3), the following relations hold

$$\begin{aligned}f(\mathbf{x}_{t+1}) \quad \leq \quad & f(\widehat{\mathbf{x}}_t) + \langle\nabla f(\widehat{\mathbf{x}}_t), \mathbf{x}_{t+1} - \widehat{\mathbf{x}}_t\rangle + \frac{\mathcal{L}}{2}\|\mathbf{x}_{t+1} - \widehat{\mathbf{x}}_t\|^2\\ = \quad & f(\widehat{\mathbf{x}}_t) + \langle\nabla f(\widehat{\mathbf{x}}_t), \mathbf{x}_t - \widehat{\mathbf{x}}_t\rangle + \langle\nabla f(\widehat{\mathbf{x}}_t), \mathbf{x}_{t+1} - \mathbf{x}_t\rangle + \frac{\mathcal{L}}{2}\|\mathbf{x}_{t+1} - \widehat{\mathbf{x}}_t\|^2\\ \leq \quad & f(\mathbf{x}_t) + \frac{\mu_t}{2}\|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2 + \langle\nabla f(\widehat{\mathbf{x}}_t), \widetilde{\mathbf{x}}_{t+1} - \mathbf{x}_t\rangle + \frac{\mathcal{L}}{2}\|\mathbf{x}_{t+1} - \widehat{\mathbf{x}}_t\|^2\\ & + \langle\nabla f(\widehat{\mathbf{x}}_t), \mathbf{x}_{t+1} - \widetilde{\mathbf{x}}_{t+1}\rangle. \qquad (13)\end{aligned}$$

Note that $\eta_t = 2L/(2-\beta_t) > L > \mathcal{L}$. Since

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}}\left\{\langle\nabla f(\widehat{\mathbf{x}}_t), \mathbf{x}\rangle + \frac{\eta_t}{2}\|\mathbf{x} - \widehat{\mathbf{x}}_t\|^2 + p(\mathbf{x})\right\} \qquad (14)$$

and $\widetilde{\mathbf{x}}_{t+1} \in \mathcal{X}$, we obtain

$$\begin{aligned}
&\langle \nabla f(\widehat{\mathbf{x}}_t), \mathbf{x}_{t+1} - \widetilde{\mathbf{x}}_{t+1} \rangle + p(\mathbf{x}_{t+1}) \\
&\leq \frac{\eta_t}{2}\left(\|\widetilde{\mathbf{x}}_{t+1} - \widehat{\mathbf{x}}_t\|^2 - \|\mathbf{x}_{t+1} - \widehat{\mathbf{x}}_t\|^2 \right) + p(\widetilde{\mathbf{x}}_{t+1}) - \frac{\eta_t + \nu}{2}\|\mathbf{x}_{t+1} - \widetilde{\mathbf{x}}_{t+1}\|^2 . \quad (15)
\end{aligned}$$

By the definition (12) of $\widetilde{\mathbf{x}}_{t+1}$ and $\widehat{\mathbf{x}}_t = \beta_t \breve{\mathbf{x}}_t + (1-\beta_t)\mathbf{x}_t$, we have

$$\beta_t(\breve{\mathbf{x}}_{t+1} - \widehat{\mathbf{x}}_t) + (1-\beta_t)(\mathbf{x}_t - \widehat{\mathbf{x}}_t) = \widetilde{\mathbf{x}}_{t+1} - \widehat{\mathbf{x}}_t = \beta_t \mathbf{s}_t, \quad (16)$$

where $\mathbf{s}_t = \breve{\mathbf{x}}_{t+1} - \breve{\mathbf{x}}_t$. Let us define

$$\mathbf{g}(\widehat{\mathbf{x}}_t) = \eta_t(\widehat{\mathbf{x}}_t - \mathbf{x}_{t+1}). \quad (17)$$

Then, it follows from (15), (16) and (17) that

$$\begin{aligned}
&\langle \nabla f(\widehat{\mathbf{x}}_t), \mathbf{x}_{t+1} - \widetilde{\mathbf{x}}_{t+1} \rangle \\
\leq\ & \frac{\eta_t \beta_t^2}{2}\|\mathbf{s}_t\|^2 - \frac{1}{2\eta_t}\|\mathbf{g}(\widehat{\mathbf{x}}_t)\|^2 + p(\widetilde{\mathbf{x}}_{t+1}) - p(\mathbf{x}_{t+1}) - \frac{\eta_t + \nu}{2}\|\mathbf{x}_{t+1} - \widetilde{\mathbf{x}}_{t+1}\|^2 .
\end{aligned}$$

So, by (13) and (16), $\widetilde{\mathbf{x}}_{t+1} = \beta_t \breve{\mathbf{x}}_{t+1} + (1-\beta_t)\mathbf{x}_t$, and the convexity of g, we have

$$\begin{aligned}
& F(\mathbf{x}_{t+1}) = f(\mathbf{x}_{t+1}) + p(\mathbf{x}_{t+1}) \\
\leq\ & \beta_t \left[f(\mathbf{x}_t) + \langle \nabla f(\widehat{\mathbf{x}}_t), \breve{\mathbf{x}}_{t+1} - \mathbf{x}_t \rangle + p(\breve{\mathbf{x}}_{t+1}) \right] + (1-\beta_t)\left[f(\mathbf{x}_t) + p(\mathbf{x}_t) \right] \\
& + \frac{\mu_t}{2}\|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2 + \frac{\eta_t \beta_t^2}{2}\|\mathbf{s}_t\|^2 - \frac{\eta_t - \mathcal{L}}{2\eta_t^2}\|\mathbf{g}(\widehat{\mathbf{x}}_t)\|^2 - \frac{\eta_t + \nu}{2}\|\mathbf{x}_{t+1} - \widetilde{\mathbf{x}}_{t+1}\|^2 \\
=\ & \beta_t \left[f(\mathbf{x}_t) + \langle \nabla f(\widehat{\mathbf{x}}_t), \breve{\mathbf{x}}_{t+1} - \mathbf{x}_t \rangle + \frac{\gamma_t}{2}\|\mathbf{s}_t\|^2 + p(\breve{\mathbf{x}}_{t+1}) \right] + (1-\beta_t)F(\mathbf{x}_t) \\
& + \frac{\mu_t}{2}\|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2 + \frac{\eta_t \beta_t^2 - \gamma_t \beta_t}{2}\|\mathbf{s}_t\|^2 - \frac{\eta_t - \mathcal{L}}{2\eta_t^2}\|\mathbf{g}(\widehat{\mathbf{x}}_t)\|^2 \\
& - \frac{\eta_t + \nu}{2}\|\mathbf{x}_{t+1} - \widetilde{\mathbf{x}}_{t+1}\|^2 \\
=\ & \beta_t \left[f(\mathbf{x}_t) + \langle \nabla f(\widehat{\mathbf{x}}_t), \breve{\mathbf{x}}_{t+1} - \mathbf{x}_t \rangle + \frac{\gamma_t}{2}\|\mathbf{s}_t\|^2 + p(\breve{\mathbf{x}}_{t+1}) \right] + (1-\beta_t)F(\mathbf{x}_t) \\
& + \frac{\mu_t}{2}\|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2 - \frac{\eta_t - \mathcal{L}}{2\eta_t^2}\|\mathbf{g}(\widehat{\mathbf{x}}_t)\|^2 - \frac{\eta_t + \nu}{2}\|\mathbf{x}_{t+1} - \widetilde{\mathbf{x}}_{t+1}\|^2 , \quad (18)
\end{aligned}$$

where the last equality follows from $\gamma_t \beta_t - \eta_t \beta_t^2 = 0$. Now, it follows from

$$\breve{\mathbf{x}}_{t+1} = \arg\min_{\mathbf{x} \in \mathcal{X}} \left\{ \langle \nabla f(\widehat{\mathbf{x}}_t), \mathbf{x} \rangle + \frac{\gamma_t}{2}\|\mathbf{x} - \breve{\mathbf{x}}_t\|^2 + p(\mathbf{x}) \right\},$$

$\mathbf{s}_t = \breve{\mathbf{x}}_{t+1} - \breve{\mathbf{x}}_t$, $\mathbf{x}_t \in \mathcal{X}$ and (4) that

$$\begin{aligned}
& \langle \nabla f(\widehat{\mathbf{x}}_t), \breve{\mathbf{x}}_{t+1} - \mathbf{x}_t \rangle + \frac{\gamma_t}{2}\|\mathbf{s}_t\|^2 + p(\breve{\mathbf{x}}_{t+1}) \\
\leq\ & \frac{\gamma_t}{2}\left(\|\mathbf{x}_t - \breve{\mathbf{x}}_t\|^2 - \|\mathbf{x}_t - \breve{\mathbf{x}}_{t+1}\|^2 \right) + p(\mathbf{x}_t) - \frac{\nu}{2}\|\mathbf{x}_t - \breve{\mathbf{x}}_{t+1}\|^2 .
\end{aligned}$$

Hence, by (18), we have

$$\begin{aligned}
F(\mathbf{x}_{t+1}) \leq\ & \beta_t \left[f(\mathbf{x}_t) + \frac{\gamma_t}{2}\left(\|\mathbf{x}_t - \breve{\mathbf{x}}_t\|^2 - \|\mathbf{x}_t - \breve{\mathbf{x}}_{t+1}\|^2 \right) + p(\mathbf{x}_t) \right. \\
& \left. - \frac{\nu}{2}\|\mathbf{x}_t - \breve{\mathbf{x}}_{t+1}\|^2 \right] + (1-\beta_t)F(\mathbf{x}_t) + \frac{\mu_t}{2}\|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2 \\
& - \frac{\eta_t - \mathcal{L}}{2\eta_t^2}\|\mathbf{g}(\widehat{\mathbf{x}}_t)\|^2 - \frac{\eta_t + \nu}{2}\|\mathbf{x}_{t+1} - \widetilde{\mathbf{x}}_{t+1}\|^2 \\
\leq\ & F(\mathbf{x}_t) + \frac{\mu_t}{2}\|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2 + \frac{\beta_t \gamma_t}{2}\left(\|\mathbf{x}_t - \breve{\mathbf{x}}_t\|^2 \right.
\end{aligned}$$

$$- \|\mathbf{x}_t - \breve{\mathbf{x}}_{t+1}\|^2\Big) - \frac{\beta_t \nu}{2}\|\mathbf{x}_t - \breve{\mathbf{x}}_{t+1}\|^2 - \frac{\eta_t - \mathcal{L}}{2\eta_t^2}\|\mathbf{g}(\widehat{\mathbf{x}}_t)\|^2$$
$$- \frac{\eta_t + \nu}{2}\|\mathbf{x}_{t+1} - \widetilde{\mathbf{x}}_{t+1}\|^2. \tag{19}$$

Note that

$$\breve{\mathbf{x}}_t - \mathbf{x}_t = \frac{1}{\beta_t}(\widehat{\mathbf{x}}_t - \mathbf{x}_t) \quad \text{and} \quad \breve{\mathbf{x}}_{t+1} - \mathbf{x}_t = \frac{1}{\beta_t}(\widetilde{\mathbf{x}}_{t+1} - \mathbf{x}_t). \tag{20}$$

Then, we have from (19) that (11) holds. □

Based on Lemma 3.4, when $\bar{t} = \infty$ in Alg. 1, we have the following result on a potential energy reduction, which would play a key role for showing global convergence.

Theorem 3.5. *Suppose the Assumptions 2.1 and 3.1 hold. Then, for the iterates generated by Alg. 1 with $\bar{t} = \infty$, there exists an integer $k_0 \geq 1$ such that*

$$E_{t+1} \leq E_t - \frac{L - \mathcal{L}}{8L^2}\|\mathbf{g}(\widehat{\mathbf{x}}_t)\|^2 - c\eta_t\|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\|^2 - \frac{\beta_t \nu}{2}\|\breve{\mathbf{x}}_{t+1} - \mathbf{x}_t\|^2 \tag{21}$$

for all $t \geq k_0$, where $\widetilde{\mathbf{x}}_t$ is defined in (12), $\mathbf{g}(\widehat{\mathbf{x}}_t) = \eta_t(\widehat{\mathbf{x}}_t - \mathbf{x}_{t+1})$ is defined in (17), $c > 0$ is a constant, and

$$E_t = F(\mathbf{x}_t) + \frac{\eta_{t-1}}{2}\|\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}\|^2 + \frac{\eta_{t-1} + \nu}{2}\|\widetilde{\mathbf{x}}_t - \mathbf{x}_t\|^2. \tag{22}$$

Proof. For $t \geq 2$, by (20) we obtain

$$\begin{aligned}
\widehat{\mathbf{x}}_t - \mathbf{x}_t &= \beta_t(\breve{\mathbf{x}}_t - \mathbf{x}_t) = \beta_t\left((\breve{\mathbf{x}}_t - \mathbf{x}_{t-1}) + (\mathbf{x}_{t-1} - \widetilde{\mathbf{x}}_t)\right) + \beta_t(\widetilde{\mathbf{x}}_t - \mathbf{x}_t) \\
&= \beta_t\left(\frac{1}{\beta_{t-1}}(\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}) + (\mathbf{x}_{t-1} - \widetilde{\mathbf{x}}_t)\right) + \beta_t(\widetilde{\mathbf{x}}_t - \mathbf{x}_t) \\
&= \theta_t(\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}) + \beta_t(\widetilde{\mathbf{x}}_t - \mathbf{x}_t),
\end{aligned} \tag{23}$$

where $\theta_t = \frac{\beta_t}{\beta_{t-1}}(1 - \beta_{t-1})$. By defining $\beta_0 = 1$, $\mathbf{x}_0 = \mathbf{x}_1$ and $\widetilde{\mathbf{x}}_1 = \mathbf{x}_1$, we can see (23) also holds for $t = 1$. Hence, for $t \geq 1$ it follows from (11) and $L \leq \eta_t < 2L$ that

$$\begin{aligned}
F(\mathbf{x}_{t+1}) &\leq F(\mathbf{x}_t) + \frac{\gamma_t/\beta_t + \mu_t}{2}\|\theta_t(\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}) + \beta_t(\widetilde{\mathbf{x}}_t - \mathbf{x}_t)\|^2 \\
&\quad - \frac{\gamma_t/\beta_t}{2}\|\widetilde{\mathbf{x}}_{t+1} - \mathbf{x}_t\|^2 - \frac{\beta_t \nu}{2}\|\breve{\mathbf{x}}_{t+1} - \mathbf{x}_t\|^2 \\
&\quad - \frac{\eta_t - \mathcal{L}}{2\eta_t^2}\|\mathbf{g}(\widehat{\mathbf{x}}_t)\|^2 - \frac{\eta_t + \nu}{2}\|\mathbf{x}_{t+1} - \widetilde{\mathbf{x}}_{t+1}\|^2 \\
&\leq F(\mathbf{x}_t) + \frac{\gamma_t/\beta_t + \mu_t}{2}\|\theta_t(\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}) + \beta_t(\widetilde{\mathbf{x}}_t - \mathbf{x}_t)\|^2 \\
&\quad - \frac{\gamma_t/\beta_t}{2}\|\widetilde{\mathbf{x}}_{t+1} - \mathbf{x}_t\|^2 - \frac{\beta_t \nu}{2}\|\breve{\mathbf{x}}_{t+1} - \mathbf{x}_t\|^2 \\
&\quad - \frac{L - \mathcal{L}}{8L^2}\|\mathbf{g}(\widehat{\mathbf{x}}_t)\|^2 - \frac{\eta_t + \nu}{2}\|\mathbf{x}_{t+1} - \widetilde{\mathbf{x}}_{t+1}\|^2.
\end{aligned} \tag{24}$$

Now, since $\mu_t = \min\{\mu_{t-1} + \rho^j - 1, L\}$ for some $\rho > 1$ and $j \geq 0$, it follows from $L > \mathcal{L} \geq \mu$ and (3) that the sequence $\{\mu_t\}$ is monotonically nondecreasing with upper bound $\mu_{up} = \min\{L, \rho(\mu + 1)\}$. Hence, μ_t can only be increased in finite, in fact at most $\lceil \mu_{up}/(\rho - 1) \rceil$, number of times. So, there exist $\overline{\mu} \geq 0$ and an integer $\overline{k} \geq 0$ such that $\mu_t = \overline{\mu}$ for all $t \geq \overline{k}$.

Since $\overline{t} = \infty$ in Alg. 1, we have $t_0 = 0$ during all iterations, which implies $\overline{\beta}_t = 2/(t+1-t_0) = 2/(t+1)$ for all $t \geq 1$. Moreover, since $\beta_t = \max\{\overline{\beta}_t, \tau_t\}$ and $\mu_t = \overline{\mu}$ for all $t \geq \overline{k}$, defining $\kappa = \overline{\mu}/L \in [0,1]$, we have from Alg. 1 that

$$\beta_t = \max\{\overline{\beta}_t, \overline{\tau}\}, \tag{25}$$

for all $t \geq \overline{k}$, where

$$\overline{\tau} := \frac{\lambda}{2}\left(1 - \sqrt{\frac{1-\kappa}{1+\kappa}}\right) + \frac{(1-\lambda)\kappa}{1+\kappa} \in \left[0, \frac{1}{2}\right]. \tag{26}$$

Hence, for all $t \geq \overline{k}$, we have from (25) and $\overline{\beta}_t = 2/(t+1)$ that $\beta_{t+1} \leq \beta_t$, which gives

$$\eta_t = \gamma_t/\beta_t = 2L/(2-\beta_t) \geq 2L/(2-\beta_{t+1}) = \eta_{t+1} > L. \tag{27}$$

For all $t \geq \overline{k}+1$, it follows from (24) that

$$\begin{aligned} & F(\mathbf{x}_{t+1}) + \frac{\eta_t}{2}\|\widetilde{\mathbf{x}}_{t+1} - \mathbf{x}_t\|^2 + \frac{\eta_t + \nu}{2}\|\mathbf{x}_{t+1} - \widetilde{\mathbf{x}}_{t+1}\|^2 \\ & \leq F(\mathbf{x}_t) + \frac{\eta_{t-1}}{2}\|\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}\|^2 + \frac{\eta_{t-1} + \nu}{2}\|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\|^2 - \frac{L - \mathcal{L}}{8L^2}\|\mathbf{g}(\widehat{\mathbf{x}}_t)\|^2 \\ & -\frac{\beta_t \nu}{2}\|\breve{\mathbf{x}}_{t+1} - \mathbf{x}_t\|^2 - R_t, \end{aligned} \tag{28}$$

where

$$\begin{aligned} R_t &= \frac{\eta_{t-1}}{2}\|\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}\|^2 + \frac{\eta_{t-1} + \nu}{2}\|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\|^2 \\ &\quad -\frac{\eta_t + \overline{\mu}}{2}\|\theta_t(\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}) + \beta_t(\widetilde{\mathbf{x}}_t - \mathbf{x}_t)\|^2 \qquad (29) \\ &\geq \frac{\eta_t}{2}\|\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}\|^2 + \frac{\eta_t}{2}\|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\|^2 \\ &\quad -\frac{\eta_t + \overline{\mu}}{2}\|\theta_t(\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}) + \beta_t(\widetilde{\mathbf{x}}_t - \mathbf{x}_t)\|^2 \\ &= \frac{\eta_t - (\eta_t + \overline{\mu})\theta_t^2}{2}\|\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}\|^2 + \frac{\eta_t - (\eta_t + \overline{\mu})\beta_t^2}{2}\|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\|^2 \\ &\quad -(\eta_t + \overline{\mu})\theta_t\beta_t\|\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}\|\,\|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\| \qquad (30) \end{aligned}$$

and the above second inequality follows from (27) and $\nu \geq 0$. We first show that when $t = 1$ or $t = 2$,

$$R_t \geq c\eta_t\left(\|\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}\|^2 + \|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\|^2\right) \tag{31}$$

for $c = 1/2$. When $t = 1$, (31) holds for any $c > 0$ simply because our definition of $\mathbf{x}_0 = \widetilde{\mathbf{x}}_1 = \mathbf{x}_1$. When $t = 2$, (31) holds with $c = 1/2$ because $\widetilde{\mathbf{x}}_2 = \beta_1\breve{\mathbf{x}}_2 + (1-\beta_1)\mathbf{x}_1 = \mathbf{x}_2$ and $\theta_2 = \beta_2(1-\beta_1)/\beta_1 = 0$. In the following, we divide our analysis into two cases on whether $\overline{\mu} > 0$ or whether $\overline{\mu} = 0$.

Case 1: $\overline{\mu} > 0$. Then, for all $t \geq \overline{k}$, we have from $\kappa = \overline{\mu}/L > 0$ and $\beta_t \geq \overline{\tau} > 0$ by (25) that

$$\kappa_t := \frac{\overline{\mu}}{\eta_t} = \frac{\overline{\mu}}{L}\frac{2-\beta_t}{2} \leq \kappa\frac{2-\overline{\tau}}{2} = \kappa - \frac{\kappa\overline{\tau}}{2}, \tag{32}$$

where $\overline{\tau}$ is defined in (26). In addition, for all $t \geq \widetilde{t} := \max\{\overline{k}+1, 3\}$, by (25), we have $\beta_t \leq 1/2$ and $\beta_t/\beta_{t-1} \geq t/(t+1) \geq 3/4$, which give

$$\theta_t = \frac{\beta_t}{\beta_{t-1}}(1-\beta_{t-1}) \geq \frac{3}{8}. \tag{33}$$

So, it follows from (30), (32) and (33) that

$$\begin{aligned}\frac{R_t}{\eta_t} &\geq \frac{1-(1+\kappa_t)\theta_t^2}{2}\|\widetilde{\mathbf{x}}_t-\mathbf{x}_{t-1}\|^2+\frac{1-(1+\kappa_t)\beta_t^2}{2}\|\mathbf{x}_t-\widetilde{\mathbf{x}}_t\|^2\\ &\quad -(1+\kappa_t)\theta_t\beta_t\|\widetilde{\mathbf{x}}_t-\mathbf{x}_{t-1}\|\,\|\mathbf{x}_t-\widetilde{\mathbf{x}}_t\|\\ &\geq h_t+\frac{\kappa\overline{\tau}}{4}\left(\theta_t^2\|\widetilde{\mathbf{x}}_t-\mathbf{x}_{t-1}\|^2+\beta_t^2\|\mathbf{x}_t-\widetilde{\mathbf{x}}_t\|^2\right)\\ &\geq h_t+c_1\left(\|\widetilde{\mathbf{x}}_t-\mathbf{x}_{t-1}\|^2+\|\mathbf{x}_t-\widetilde{\mathbf{x}}_t\|^2\right), \end{aligned} \tag{34}$$

for all $t\geq\widetilde{t}$, where

$$c_1=\frac{\kappa\overline{\tau}}{4}\min\left\{\frac{9}{64},\overline{\tau}^2\right\}>0 \tag{35}$$

and

$$\begin{aligned}h_t &= \frac{1-(1+\kappa)\theta_t^2}{2}\|\widetilde{\mathbf{x}}_t-\mathbf{x}_{t-1}\|^2+\frac{1-(1+\kappa)\beta_t^2}{2}\|\mathbf{x}_t-\widetilde{\mathbf{x}}_t\|^2\\ &\quad -(1+\kappa)\theta_t\beta_t\|\widetilde{\mathbf{x}}_t-\mathbf{x}_{t-1}\|\,\|\mathbf{x}_t-\widetilde{\mathbf{x}}_t\|. \end{aligned} \tag{36}$$

We now show $h_t\geq 0$ for all $t\geq\widetilde{t}$. By Cauchy-Schwarz inequality and (36), to show $h_t\geq\widetilde{t}$, it is sufficient to show

$$\left[1-(1+\kappa)\,\theta_t^2\right]\left[1-(1+\kappa)\,\beta_t^2\right]\geq(1+\kappa)^2\,\theta_t^2\beta_t^2, \tag{37}$$

which is equivalent to

$$1-(1+\kappa)\left(\theta_t^2+\beta_t^2\right)\geq 0. \tag{38}$$

Notice that for all $t\geq\widetilde{t}$, we have $\beta_t\leq\beta_{t-1}$. Hence, for all $t\geq\widetilde{t}$, we have

$$\theta_t=\frac{\beta_t}{\beta_{t-1}}(1-\beta_{t-1})\leq 1-\beta_t, \tag{39}$$

which gives

$$1-(1+\kappa)\left(\theta_t^2+\beta_t^2\right)\geq 1-(1+\kappa)\left((1-\beta_t)^2+\beta_t^2\right). \tag{40}$$

By the choice of β_t, we have $\frac{1}{2}\geq\beta_t\geq\overline{\tau}\geq\widetilde{\tau}>0$ for all $t\geq\widetilde{t}\geq 3$, where $\overline{\tau}$ is defined in (26) and $\widetilde{\tau}=\frac{1}{2}\left(1-\sqrt{(1-\kappa)/(1+\kappa)}\right)$, which implies

$$(1-\beta_t)^2+\beta_t^2\leq(1-\widetilde{\tau})^2+\widetilde{\tau}^2$$

for all $t\geq\widetilde{t}$. So, for all $t\geq\widetilde{t}$, we have from (40) and (26) that

$$1-(1+\kappa)\left(\theta_t^2+\beta_t^2\right)\geq 1-(1+\kappa)\left((1-\widetilde{\tau})^2+\widetilde{\tau}^2\right)=0.$$

Hence, (38) holds, which shows $h_t\geq 0$ and therefore (31) holds for all $t\geq\widetilde{t}$ with $c=c_1$ defined in (35). Since $c_1<1/2$, by (31), we have in fact (31) holds for all $t\geq\overline{k}+1$ with $c=c_1$. Then, (28) implies (21) holds with $c=c_1$ for all $t\geq\overline{k}+1$.

Case 2: $\overline{\mu}=0$. Then, we have $\mu_t=0$ and $\tau_t=0$ for all $t\geq 1$. So, $\overline{k}=1$ and for all $t\geq 1$, we have $\beta_t=\overline{\beta}_t=2/(t+1)$, $\gamma_t/\beta_t=\eta_t$ and $\eta_t=2L/(2-\beta_t)=L(t+1)/t$. In addition, we have $\theta_t=\frac{\beta_t}{\beta_{t-1}}(1-\beta_{t-1})=\frac{t-2}{t+1}<1-\beta_t$. So, we have

$$\begin{aligned}&\|\theta_t\left(\widetilde{\mathbf{x}}_t-\mathbf{x}_{t-1}\right)+\beta_t\left(\widetilde{\mathbf{x}}_t-\mathbf{x}_t\right)\|^2\\ &\leq\left(\theta_t\|\widetilde{\mathbf{x}}_t-\mathbf{x}_{t-1}\|+\beta_t\|\widetilde{\mathbf{x}}_t-\mathbf{x}_t\|\right)^2\leq\left((1-\beta_t)\|\widetilde{\mathbf{x}}_t-\mathbf{x}_{t-1}\|+\beta_t\|\widetilde{\mathbf{x}}_t-\mathbf{x}_t\|\right)^2\\ &\leq(1-\beta_t)\|\widetilde{\mathbf{x}}_t-\mathbf{x}_{t-1}\|^2+\beta_t\|\widetilde{\mathbf{x}}_t-\mathbf{x}_t\|^2. \end{aligned} \tag{41}$$

So, for all $t \geq 2$, we have from (29) and $\nu \geq 0$ that

$$\begin{aligned}\frac{2R_t}{\eta_t} &= \frac{t^2}{t^2-1}\|\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}\|^2 + \frac{t^2}{t^2-1}\|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\|^2 \\ &\quad -\|\theta_t(\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}) + \beta_t(\widetilde{\mathbf{x}}_t - \mathbf{x}_t)\|^2 \\ &\geq \left(\frac{t^2}{t^2-1} - \frac{t-1}{t+1}\right)\|\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}\|^2 + \left(\frac{t^2}{t^2-1} - \frac{2}{t+1}\right)\|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\|^2 \\ &= \frac{2t-1}{t^2-1}\|\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}\|^2 + \frac{t-1}{t+1}\|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\|^2,\end{aligned}$$

which implies for all $t \geq 3$,

$$R_t \geq \frac{L}{t}\|\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}\|^2 + \frac{\eta_t}{4}\|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\|^2. \tag{42}$$

Then, we have from (28), (31) and (42) that (21) holds with $c = 1/4$ for all $t \geq 1$.

Combing the above two cases, Case 1 and Case 2, we have (21) holds with $c = \min\{c_1, 1/4\} = c_1$ for all $t \geq k_0 := \overline{k} + 1$, where c_1 is defined in (35). □

We now consider the case $\overline{t} < \infty$ in Alg. 1. In this case, when $\mathrm{mod}(t, \overline{t}) = 0$, $\breve{\mathbf{x}}_{t+1}$ is not computed in Alg. 1. To facilitate the convergence proof, when $\mathrm{mod}(t, \overline{t}) = 0$, we still let $\breve{\mathbf{x}}_{t+1}$ be defined as that in step 7 of Alg. 1, although it is not actually calculated in Alg. 1. Then, based on Theorem 3.5, we can easily establish the following properties on potential energy reduction.

Theorem 3.6. *Suppose the Assumptions 2.1 and 3.1 hold. Then, for the iterates generated by Alg. 1 with $\overline{t} < \infty$, there exists an integer $\widehat{k}_0 \geq 1$ such that for all $t \geq \widehat{k}_0$, if mod$(t, \overline{t}) = 1$, we have*

$$\begin{aligned}E_{t+1} &\leq E_t - \frac{L-\mathcal{L}}{8L^2}\|\mathbf{g}(\widehat{\mathbf{x}}_t)\|^2 - \frac{1}{2}\eta_{t-1}\left(\|\mathbf{x}_{t-1} - \widetilde{\mathbf{x}}_t\|^2 + \|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\|^2\right) \\ &\quad -\frac{\beta_t\nu}{2}\|\breve{\mathbf{x}}_{t+1} - \mathbf{x}_t\|^2;\end{aligned} \tag{43}$$

otherwise, i.e., mod$(t, \overline{t}) \neq 1$, we have

$$\begin{aligned}E_{t+1} &\leq E_t - \frac{L-\mathcal{L}}{8L^2}\|\mathbf{g}(\widehat{\mathbf{x}}_t)\|^2 - \widehat{c}\eta_t\left(\|\mathbf{x}_{t-1} - \widetilde{\mathbf{x}}_t\|^2 + \|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\|^2\right) \\ &\quad -\frac{\beta_t\nu}{2}\|\breve{\mathbf{x}}_{t+1} - \mathbf{x}_t\|^2,\end{aligned} \tag{44}$$

where $\widetilde{\mathbf{x}}_t$ is defined in (12), $\mathbf{g}(\widehat{\mathbf{x}}_t) = \eta_t(\widehat{\mathbf{x}}_t - \mathbf{x}_{t+1})$ is defined in (17), E_t is defined in (22) and $\widehat{c} > 0$ is a constant.

Proof. Since $\overline{t} < \infty$, it follows from Alg. 1 that $\overline{\beta}_t \geq 2/(\overline{t}+1) =: \widehat{\tau}$, which implies $\beta_t = \max\{\overline{\beta}_t, \mu_t\} \geq \widehat{\tau}$ for all t. Again, since μ_t is monotonically nondecreasing with upper bound $\mu_{up} = \min\{L, \rho(\mu+1)\}$, by the procedure of updating μ_t in Alg. 1, we can choose $\widehat{k}_0$ sufficiently large such that $\mu_t = \overline{\mu}$ for all $t \geq \widehat{k}_0$ and some $\overline{\mu} \geq 0$.

When $\mathrm{mod}(t, \overline{t}) = 1$, we have from Alg. 1 that $\beta_t = 1$. Then, the first inequality in (43) holds by following from the same arguments for showing the case $t = 1$ in Theorem 3.5. The second inequality in (43) follows from the definition E_t of in (22) and $\nu \geq 0$.

We now consider the case $\mathrm{mod}(t, \overline{t}) \neq 1$. In this case, we have $\overline{\beta}_{t-1} \geq \overline{\beta}_t$, which implies $\beta_{t-1} \geq \beta_t$ and $\eta_{t-1} \geq \eta_t$. Then, for $t \geq \widehat{k}_0$, it follows from (28) that

$$E_{t+1} \leq E_t - \frac{L-\mathcal{L}}{8L^2}\|\mathbf{g}(\widehat{\mathbf{x}}_t)\|^2 - \frac{\beta_t\nu}{2}\|\breve{\mathbf{x}}_{t+1} - \mathbf{x}_t\|^2 - R_t \tag{45}$$

where R_t is defined in (29). If $\overline{\mu} > 0$ or $\mathrm{mod}(t, \bar{t}) = 2$, by the same reasons as (31) and (34), we have

$$R_t \geq \widehat{c}_1 \eta_t \left(\|\mathbf{x}_{t-1} - \widetilde{\mathbf{x}}_t\|^2 + \|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\|^2 \right) \tag{46}$$

where $\widehat{c}_1 = \kappa \widehat{\tau} \min\{9/64, \widehat{\tau}^2\} > 0$ and $\kappa = \overline{\mu}/L$. If $\overline{\mu} = 0$, same as (41), for all $t \geq 2$ we have

$$\|\theta_t (\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}) + \beta_t (\widetilde{\mathbf{x}}_t - \mathbf{x}_t)\|^2 \leq (1-\beta_t)\|\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}\|^2 + \beta_t \|\widetilde{\mathbf{x}}_t - \mathbf{x}_t\|^2.$$

If $\mathrm{mod}(t, \bar{t}) \neq 2$, we have $\widehat{\tau} \leq \beta_t \leq 1/2$. So, if $\overline{\mu} = 0$ and $\mathrm{mod}(t, \bar{t}) \neq 2$, we have from the definition of R_t in (29), $\eta_t \leq \eta_{t-1}$ and $0 < \widehat{\tau} \leq \beta_t \leq 1/2$ that

$$\frac{2R_t}{\eta_t} \geq \beta_t \|\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}\|^2 + (1-\beta_t)\|\widetilde{\mathbf{x}}_t - \mathbf{x}_t\|^2 \geq \widehat{\tau} \left(\|\mathbf{x}_{t-1} - \widetilde{\mathbf{x}}_t\|^2 + \|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\|^2 \right),$$

which together with (46) gives $R_t \geq \widehat{c} \eta_t \left(\|\mathbf{x}_{t-1} - \widetilde{\mathbf{x}}_t\|^2 + \|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\|^2 \right)$, where $\widehat{c} = \min\{\widehat{c}_1, \widehat{\tau}/2\}$. Hence, if $\mathrm{mod}(t, \bar{t}) \neq 1$, we have (44) holds. □

We say $\mathbf{x}^*$ is a stationary point of problem (1) if $-\nabla f(\mathbf{x}^*) \in \partial p(\mathbf{x}^*)$, where $\partial p(\mathbf{x}^*)$ is the subdifferential of p at $\mathbf{x}^*$, which is equivalent to

$$\mathbf{x}^* = \arg\min_{\mathbf{x} \in \mathcal{X}} \left\{ \langle \nabla f(\mathbf{x}^*), \mathbf{x} \rangle + \frac{\eta}{2} \|\mathbf{x} - \mathbf{x}^*\|^2 + p(\mathbf{x}) \right\} \tag{47}$$

for some $\eta > 0$. The following theorem on global convergence and convergence rate can be easily obtained from Theorem 3.5 and Theorem 3.6.

Theorem 3.7. *Suppose the Assumptions 2.1, 3.1 and 3.2 hold. Then, for the iterates generated by Alg. 1, the following properties hold.*
(i) There exists an integer $k_0 > 0$ such that for $T > k_0$ we have

$$\min_{t \in \{k_0, k_0+1, \ldots, T-1\}} \|\mathbf{g}(\widehat{\mathbf{x}}_t)\|^2 \leq \frac{8L^2(E_{k_0} - \overline{F})}{L - \mathcal{L}} \frac{1}{T - k_0} = \mathcal{O}(1/T), \tag{48}$$

where $\mathbf{g}(\widehat{\mathbf{x}}_t) = \eta_t(\widehat{\mathbf{x}}_t - \mathbf{x}_{t+1})$ is defined in (17). Furthermore, we have

$$\lim_{t \to \infty} \|\mathbf{g}(\widehat{\mathbf{x}}_t)\| = 0. \tag{49}$$

(ii) The sequences $\{\widehat{\mathbf{x}}_t\}$, $\{\mathbf{x}_t\}$ and $\{\widetilde{\mathbf{x}}_t\}$ have the same set of cluster points if it is nonempty, which are all stationary points of problem (1).

Proof. First, since $\beta_t \in (0, 1]$, we have from $\eta_t = 2L/(2-\beta_t)$ that $\eta_t \in (L, 2L]$ for all t. Then, by Theorem 3.5, Theorem 3.6 and Assumption 3.2, there exists a $k_0 > 0$ such that for all $T > k_0$ we have

$$\sum_{t=k_0}^{T-1} (\|\mathbf{g}(\widehat{\mathbf{x}}_t)\| + \|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\|) \leq \frac{8L^2(E_{k_0} - \overline{F})}{L - \mathcal{L}}, \tag{50}$$

where $\mathbf{g}(\widehat{\mathbf{x}}_t)$ and $\widetilde{\mathbf{x}}_t$ are defined in (17) and (12), respectively. Then, both (48) and (49) follow from (50).

Now, by (50) and $\mathbf{g}(\widehat{\mathbf{x}}_t) = \eta_t(\widehat{\mathbf{x}}_t - \mathbf{x}_{t+1})$, we have

$$\lim_{t \to \infty} (\|\mathbf{x}_{t+1} - \widehat{\mathbf{x}}_t\| + \|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\|) = 0. \tag{51}$$

Thus, the sequences $\{\widehat{\mathbf{x}}_t\}$, $\{\mathbf{x}_t\}$ and $\{\widetilde{\mathbf{x}}_t\}$ have the same set of cluster points if the set is nonempty. Now, given any cluster point $\widehat{\mathbf{x}}^*$ of $\{\widehat{\mathbf{x}}_t\}$, we can have from (14) (51), $\eta_t \in (L, 2L]$ and the closedness of p that $-\nabla f(\mathbf{x}^*) \in \partial p(\mathbf{x}^*)$. Hence, the statement (ii) holds. □

In terms of the convergence on the objective function value of problem (1), we have the following corollary.

Corollary 3.8. *Suppose the Assumptions 2.1, 3.1 and 3.2 hold.*

(i) If one of the following conditions hold:
- *(a) f is a convex function;*
- *(b) the parameter $\bar{t} < \infty$ in Alg. 1;*
- *(c) $\overline{\mu} > 0$, where $\overline{\mu} = \lim_{t\to\infty} \mu_t$;*
- *(d) $\nu > 0$, where ν is defined in (4);*
- *(e) $\{\mathbf{x}_t\}$ and $\{\breve{\mathbf{x}}_t\}$ are bounded, e.g., when $\mathcal{X}$ is a bounded set,*

we have $\lim_{t\to\infty} F(\mathbf{x}_t)$ exists.

(ii) If $F^* = \lim_{t\to\infty} F(\mathbf{x}_t)$, then for any cluster point $\overline{\mathbf{x}}$ of $\{\mathbf{x}_t\}$, it has $F(\overline{\mathbf{x}}) = F^*$.

Proof. We first show (i) by considering the cases (a) to (e) separately.
Case (a): By Assumption 3.2, there exists an F^* such that $F^* = \liminf_{t\to\infty} F(\mathbf{x}_t)$. Then, when f is a convex function, it follows from (10) that for any $\epsilon > 0$ and $\mathbf{x}_\epsilon$ such that $F(\mathbf{x}_\epsilon) \le F^* + \epsilon$, we have $\lim_{t\to\infty} F(\mathbf{x}_t) \le F^* + \epsilon$, which implies $\lim_{t\to\infty} F(\mathbf{x}_t) \le F^*$. Hence, $\lim_{t\to\infty} F(\mathbf{x}_t) = F^*$.
Case (b): If the parameter $\bar{t} < \infty$ in Alg. 1, we have from Theorem 3.6 that for any integer $\ell > 0$ sufficiently large, we have

$$F(\mathbf{x}_{t_\ell}) \ge E_{t_\ell+1} \ge E_{t_\ell+2} \ge \cdots \ge E_{t_{\ell+1}} \ge F(\mathbf{x}_{t_{\ell+1}}),$$

where $t_\ell = \ell\bar{t}$. Hence, by Assumption 3.2 and $F(\mathbf{x}_t) \le E_t$, there exists a F^* such that $F^* = \lim_{t\to\infty} F(\mathbf{x}_t) = \lim_{t\to\infty} E_t$.
Case (c): By Alg. 1, we have $\mu_t = \overline{\mu}$ for all t sufficiently large and some $\overline{\mu} \ge 0$. Suppose $\bar{t} = \infty$. Otherwise, the claim holds by case (b). Then, if $\overline{\mu} > 0$, we have from (28), (31) and (34) that

$$\begin{aligned} E_{t+1} \;\le\; & E_t - \frac{L-\mathcal{L}}{8L^2}\|\mathbf{g}(\widehat{\mathbf{x}}_t)\|^2 - c_1\eta_t\left(\|\mathbf{x}_{t-1} - \widetilde{\mathbf{x}}_t\|^2 + \|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\|^2\right) \\ & - \frac{\beta_t\nu}{2}\|\breve{\mathbf{x}}_{t+1} - \mathbf{x}_t\|^2 \end{aligned} \tag{52}$$

for all $t \ge \bar{t} + 1$, where E_t is defined in (22) and $c_1 > 0$ is a constant given in (35). Then, by (52), Assumption 3.2 and $\eta_t \in [L, 2L]$, we have

$$\lim_{t\to\infty} \left(\|\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}\| + \|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\|\right) = 0 \tag{53}$$

and there are exists an F^* such that

$$\lim_{t\to\infty} F(\mathbf{x}_t) = \lim_{t\to\infty} E_t = F^*. \tag{54}$$

Case (d): We again suppose $\bar{t} = \infty$. Otherwise, the claim holds by case (b). If $\nu > 0$, it follows from Theorem 3.5 and Assumption 3.2 that

$$\lim_{t\to\infty} \left(\|\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}\| + \|\mathbf{x}_t - \breve{\mathbf{x}}_{t+1}\|\right) = 0.$$

Then, by (20), (53) also holds and hence (54) holds by Theorem 3.5.
Case (e): Suppose the sequences $\{\mathbf{x}_t\}$ and $\{\breve{\mathbf{x}}_t\}$ are bounded. If $\bar{t} < \infty$ or $\overline{\mu} > 0$, the claim follows from Case (b) or Case (c). Hence, we only consider the case that $\bar{t} = \infty$ and $\overline{\mu} = 0$, which gives $\beta_t = 2/(t+1)$ for all $t \ge 1$ and therefore $\lim_{t\to\infty} \beta_t = 0$. By (12), $\widetilde{\mathbf{x}}_{t+1} - \mathbf{x}_t = \beta_t(\breve{\mathbf{x}}_{t+1} - \mathbf{x}_t)$. Hence, we have $\lim_{t\to\infty} \|\widetilde{\mathbf{x}}_{t+1} - \mathbf{x}_t\| = 0$ from the boundedness of $\{\mathbf{x}_t\}$ and $\{\breve{\mathbf{x}}_t\}$, which together with (51) implies (53) holds. Hence, (54) also holds.

Now, we show (ii) holds. By (14), for any $\mathbf{z} \in \mathcal{X}$, we have

$$\begin{aligned}&\langle \nabla f(\widehat{\mathbf{x}}_t), \mathbf{x}_{t+1} - \mathbf{z}\rangle + p(\mathbf{x}_{t+1})\\ &\leq \frac{\eta_t}{2}\left(\|\mathbf{z} - \widehat{\mathbf{x}}_t\|^2 - \|\mathbf{x}_{t+1} - \widehat{\mathbf{x}}_t\|^2\right) + p(\mathbf{z}) - \frac{\eta_t + \nu}{2}\|\mathbf{x}_{t+1} - \mathbf{z}\|^2,\end{aligned}$$

which by $\nu \geq 0$ gives

$$\begin{aligned}&f(\widehat{\mathbf{x}}_t) + \langle \nabla f(\widehat{\mathbf{x}}_t), \mathbf{x}_{t+1} - \widehat{\mathbf{x}}_t\rangle + \frac{\eta_t}{2}\|\mathbf{x}_{t+1} - \widehat{\mathbf{x}}_t\|^2 + p(\mathbf{x}_{t+1})\\ &\leq \frac{\eta_t}{2}\|\mathbf{z} - \widehat{\mathbf{x}}_t\|^2 + f(\widehat{\mathbf{x}}_t) + \langle \nabla f(\widehat{\mathbf{x}}_t), \mathbf{z} - \widehat{\mathbf{x}}_t\rangle + p(\mathbf{z}).\end{aligned} \tag{55}$$

For any $\mathbf{z} \in \mathcal{X}$, it follows from Assumption 2.1 that

$$|f(\mathbf{z}) - f(\widehat{\mathbf{x}}_t) - \langle \nabla f(\widehat{\mathbf{x}}_t), \mathbf{z} - \widehat{\mathbf{x}}_t\rangle| \leq \frac{\mathcal{L}}{2}\|\mathbf{z} - \widehat{\mathbf{x}}_t\|^2.$$

Hence, by (55), $\eta_t \in [L, 2L]$ and $L > \mathcal{L}$, for any $\mathbf{z} \in \mathcal{X}$, we have

$$\begin{aligned}F(\mathbf{x}_{t+1}) &= f(\mathbf{x}_{t+1}) + p(\mathbf{x}_{t+1}) \leq F(\mathbf{z}) + \frac{3L}{2}\|\mathbf{z} - \widehat{\mathbf{x}}_t\|^2\\ &\leq F(\mathbf{z}) + 3L\|\mathbf{z} - \mathbf{x}_{t+1}\|^2 + 3L\|\mathbf{x}_{t+1} - \widehat{\mathbf{x}}_t\|^2.\end{aligned} \tag{56}$$

Then, for any subsequence $\{\mathbf{x}_{t_i+1}\}$ of $\{\mathbf{x}_t\}$ converging to $\overline{\mathbf{x}} \in \mathcal{X}$, we have from (56) that

$$F(\mathbf{x}_{t_i+1}) \leq F(\overline{\mathbf{x}}) + 3L\|\overline{\mathbf{x}} - \mathbf{x}_{t_i+1}\|^2 + 3L\|\mathbf{x}_{t_i+1} - \widehat{\mathbf{x}}_{t_i}\|^2$$

Taking i to infinity in the above inequality, we have from $\lim_{i\to\infty}\mathbf{x}_{t_i+1} = \overline{\mathbf{x}}$, (51) and part (i) that $F^* = \lim_{i\to\infty}F(\mathbf{x}_{t_i+1}) \leq F(\overline{\mathbf{x}})$. In addition, by the lower semicontinuity of F, we have $F(\overline{\mathbf{x}}) \leq \lim_{i\to\infty}F(\mathbf{x}_{t_i+1}) = F^*$. Hence, we have $F(\overline{\mathbf{x}}) = F^*$. □

4. **Linear convergence.** In this section, we would like to discuss the linear convergence of $\{\mathbf{x}_t\}$ and $\{F(\mathbf{x}_t)\}$. Let us define $h(\mathbf{x}) = p(\mathbf{x}) + \delta_{\mathcal{X}}(\mathbf{x})$, where $\delta_{\mathcal{X}}(\mathbf{x})$ is the indicator function on the closed convex set $\mathcal{X}$. Let Ω^* be the set of all stationary points of problem (1), i.e.,

$$\Omega^* = \left\{\mathbf{x}^* \in \mathcal{X} : -\nabla f(\mathbf{x}^*) \in \partial h(\mathbf{x}^*)\right\} = \left\{\mathbf{x}^* \in \mathcal{X} : \mathbf{x}^* \text{ satisfies (47)}\right\}. \tag{57}$$

Note that Ω^* is a closed set. Denoting $\mathrm{Prox}(\mathbf{v})$ be the proximal operator of any closed convex function q at any $\mathbf{v} \in \mathbb{R}^n$, that is

$$\mathrm{Prox}_q(\mathbf{v}) = \arg\min\left\{q(\mathbf{x}) + \frac{1}{2}\|\mathbf{x} - \mathbf{v}\|^2 : \mathbf{x} \in \mathbb{R}^n\right\},$$

then we have $\mathbf{x}^* \in \Omega^*$ if and only if $\mathbf{x}^* = \mathrm{Prox}_{\tau h}\left(\mathbf{x}^* - \tau\nabla f(\mathbf{x}^*)\right)$ for any $\tau > 0$. And, from Alg. 1, we have

$$\mathbf{x}_{t+1} = \mathrm{Prox}_{\frac{1}{\eta_t}h}\left(\widehat{\mathbf{x}}_t - \frac{1}{\eta_t}\nabla f(\widehat{\mathbf{x}}_t)\right) \text{ and } \breve{\mathbf{x}}_{t+1} = \mathrm{Prox}_{\frac{1}{\gamma_t}h}\left(\breve{\mathbf{x}}_t - \frac{1}{\gamma_t}\nabla f(\widehat{\mathbf{x}}_t)\right). \tag{58}$$

For studying linear convergence, we need the following error bound condition and the condition that the isocost surfaces of F are properly separated on the stationary point set Ω^*.

Assumption 4.1. (a) For any $\xi \geq \inf_{x\in\mathcal{X}}F(\mathbf{x})$, there exists $\epsilon > 0$ and $\sigma > 0$ such that

$$\mathrm{dist}(\mathbf{x}, \Omega^*) \leq \sigma\left\|\mathrm{Prox}_{\frac{1}{\eta}h}\left(\mathbf{x} - \frac{1}{\eta}\nabla f(\mathbf{x})\right) - \mathbf{x}\right\|, \tag{59}$$

whenever $\left\|\mathrm{Prox}_{\frac{1}{\eta}h}\left(\mathbf{x}-\frac{1}{\eta}\nabla f(\mathbf{x})\right)-\mathbf{x}\right\|<\epsilon$, $F(\mathbf{x})<\xi$ and $\eta\in[L,2L]$.
(b) Ω^* is nonempty and there exists $\omega>0$ such that $\|\mathbf{x}-\mathbf{y}\|\geq\omega$ whenever $\mathbf{x}$, $\mathbf{y}\in\Omega^*$ and $F(\mathbf{x})\neq F(\mathbf{y})$.

There are many functions of f and p satisfying the above Assumption 4.1 including the case that f is a possibly nonconvex quadratic function or a composition of a Lipschitz continuously differentiable strongly convex function with an affine function and p is a polyhedral function. For more examples and discussions on functions satisfying Assumption 4.1, one may refer to [16, 15, 14, 11] and the references therein. Before establishing linear convergence, we first develop the following lemma.

Lemma 4.2. *Suppose the Assumptions 2.1, 3.1, 3.2 and 4.1 hold. We have*

(i) $\lim_{t\to\infty} dist(\mathbf{x}_t,\Omega^*)=0$;

(ii) in addition, if there exists a constant $\widetilde{c}>0$ *such that for all* t *suffciently large it has*

$$\widetilde{E}_{t+1}\leq\widetilde{E}_t-\widetilde{c}d_t, \tag{60}$$

where

$$\begin{aligned}\widetilde{E}_t &= F(\mathbf{x}_t)+\frac{\eta_{t-1}}{2}\|\widetilde{\mathbf{x}}_t-\mathbf{x}_{t-1}\|^2+\frac{\eta_{t-1}+\nu}{2}\|\widetilde{\mathbf{x}}_t-\mathbf{x}_t\|^2\\ &\quad+\frac{L-\mathcal{L}}{8}\|\mathbf{x}_t-\widehat{\mathbf{x}}_{t-1}\|^2+\frac{\beta_{t-1}\nu}{2}\|\breve{\mathbf{x}}_t-\mathbf{x}_{t-1}\|^2\end{aligned} \tag{61}$$

and

$$\begin{aligned}d_t &= \|\mathbf{x}_t-\widehat{\mathbf{x}}_{t-1}\|^2+\|\mathbf{x}_{t-1}-\widetilde{\mathbf{x}}_t\|^2+\|\mathbf{x}_t-\widetilde{\mathbf{x}}_t\|^2\\ &\quad+\beta_{t-1}\nu\|\breve{\mathbf{x}}_t-\mathbf{x}_{t-1}\|^2,\end{aligned} \tag{62}$$

then for all t *sufficiently large, we have*

$$|F(\mathbf{x}_t)-F^*|\leq\overline{\theta}\|\mathbf{x}_t-\widehat{\mathbf{x}}_{t-1}\|^2 \tag{63}$$

and

$$0\leq\widetilde{E}_{t+1}-F^*\leq\theta(\widetilde{E}_t-F^*), \tag{64}$$

where $\overline{\theta}>0$ *and* $\theta\in(0,1)$ *are constants and* $F^*=\lim_{t\to\infty}F(\mathbf{x}_t)=\lim_{t\to\infty}\widetilde{E}_t$.

Proof. By Theorem 3.5 and Theorem 3.6, there exists a $\xi>0$ such that $E_t\leq\xi$ for all $t\geq1$, which implies $F(\mathbf{x}_t)\leq\xi$ for all $t\geq1$. In addition, by (49) and (58), we have

$$\begin{aligned}0 &= \lim_{t\to\infty}\|\mathbf{g}(\widehat{\mathbf{x}}_t)\|=\lim_{t\to\infty}\|\mathbf{x}_{t+1}-\widehat{\mathbf{x}}_t\|\\ &= \lim_{t\to\infty}\left\|\mathrm{Prox}_{\frac{1}{\eta_t}h}\left(\widehat{\mathbf{x}}_t-\frac{1}{\eta_t}\nabla f(\widehat{\mathbf{x}}_t)\right)-\widehat{\mathbf{x}}_t\right\|.\end{aligned} \tag{65}$$

By the nonexpansion property of the proximal operator, (58), $\eta_t>L>\mathcal{L}$ and Assumption 2.1, we have

$$\begin{aligned}&\left\|\mathrm{Prox}_{\frac{1}{\eta_t}h}\left(\mathbf{x}_{t+1}-\frac{1}{\eta_t}\nabla f(\mathbf{x}_{t+1})\right)-\mathbf{x}_{t+1}\right\|\\ &=\left\|\mathrm{Prox}_{\frac{1}{\eta_t}h}\left(\mathbf{x}_{t+1}-\frac{1}{\eta_t}\nabla f(\mathbf{x}_{t+1})\right)-\mathrm{Prox}_{\frac{1}{\eta_t}h}\left(\widehat{\mathbf{x}}_t-\frac{1}{\eta_t}\nabla f(\widehat{\mathbf{x}}_t)\right)\right\|\\ &\leq\left\|\left(\mathbf{x}_{t+1}-\frac{1}{\eta_t}\nabla f(\mathbf{x}_{t+1})\right)-\left(\widehat{\mathbf{x}}_t-\frac{1}{\eta_t}\nabla f(\widehat{\mathbf{x}}_t)\right)\right\|\end{aligned}$$

$$\leq \left(1+\frac{\mathcal{L}}{\eta_t}\right)\|\mathbf{x}_{t+1}-\widehat{\mathbf{x}}_t\| \leq 2\|\mathbf{x}_{t+1}-\widehat{\mathbf{x}}_t\|.$$

Hence, it follows from $\eta_t \in [L, 2L]$ and Assumption 4.1 (a) and (65) that

$$\begin{aligned}\operatorname{dist}(\mathbf{x}_{t+1}, \Omega^*) &\leq \sigma\left\|\operatorname{Prox}_{\frac{1}{\eta_t}h}\left(\mathbf{x}_{t+1}-\frac{1}{\eta_t}\nabla f(\mathbf{x}_{t+1})\right)-\mathbf{x}_{t+1}\right\| \\ &\leq 2\sigma\|\mathbf{x}_{t+1}-\widehat{\mathbf{x}}_t\|. \end{aligned} \tag{66}$$

for t sufficiently large. So, we have (i) holds by (65).

Now, we prove (ii). Let us define $\overline{\mathbf{x}}_t \in \Omega^*$ such that $\operatorname{dist}(\mathbf{x}_t, \Omega^*) = \|\mathbf{x}_t - \overline{\mathbf{x}}_t\|$. By (60) and Assumption 3.2, we have $\lim_{t\to\infty} d_t = 0$, where d_t is defined in (62), which gives

$$\lim_{t\to\infty}\|\mathbf{x}_t-\mathbf{x}_{t-1}\| \leq \lim_{t\to\infty}\left(\|\mathbf{x}_t-\widetilde{\mathbf{x}}_t\|+\|\widetilde{\mathbf{x}}_t-\mathbf{x}_{t-1}\|\right) \leq \lim_{t\to\infty}\sqrt{2d_t} = 0.$$

Hence, we have from property (i) that

$$\lim_{t\to\infty}\|\overline{\mathbf{x}}_t-\overline{\mathbf{x}}_{t-1}\| \leq \lim_{t\to\infty}\|\overline{\mathbf{x}}_t-\mathbf{x}_t\|+\|\mathbf{x}_t-\mathbf{x}_{t-1}\|+\|\mathbf{x}_{t-1}-\overline{\mathbf{x}}_{t-1}\| = 0.$$

This together with the Assumption 4.1 (b) implies that $F(\overline{\mathbf{x}}_t) = F^*$ for all t sufficiently large, where F^* is some constant. Hence, for t sufficiently large, replacing $t+1$ by t and taking $\mathbf{z} = \overline{\mathbf{x}}_t$ in (56), we have

$$\begin{aligned} F(\mathbf{x}_t) - F^* &\leq 3L\left\|\overline{\mathbf{x}}_t-\mathbf{x}_t\right\|^2 + 3L\left\|\mathbf{x}_t-\widehat{\mathbf{x}}_{t-1}\right\|^2 \\ &= 3L\operatorname{dist}(\mathbf{x}_t,\Omega^*)^2 + 3L\left\|\mathbf{x}_t-\widehat{\mathbf{x}}_{t-1}\right\|^2 \\ &\leq (12\sigma^2+3)L\left\|\mathbf{x}_t-\widehat{\mathbf{x}}_{t-1}\right\|^2, \end{aligned} \tag{67}$$

where the last inequality follows from (66). On the other hand, since $\overline{\mathbf{x}}_t \in \Omega^*$, we have from (47) that

$$\langle\nabla f(\overline{\mathbf{x}}_t), \overline{\mathbf{x}}_t\rangle + p(\overline{\mathbf{x}}_t) \leq \langle\nabla f(\overline{\mathbf{x}}_t), \mathbf{x}_t\rangle + \frac{\eta}{2}\left\|\mathbf{x}_t-\overline{\mathbf{x}}_t\right\|^2 + p(\mathbf{x}_t)$$

for some $\eta > 0$, which by Assumption 2.1 and (66) gives

$$\begin{aligned} F^* &= F(\overline{\mathbf{x}}_t) = f(\overline{\mathbf{x}}_t) + p(\overline{\mathbf{x}}_t) \\ &\leq f(\overline{\mathbf{x}}_t) + \langle\nabla f(\overline{\mathbf{x}}_t), \mathbf{x}_t-\overline{\mathbf{x}}_t\rangle + \frac{\eta}{2}\left\|\mathbf{x}_t-\overline{\mathbf{x}}_t\right\|^2 + p(\mathbf{x}_t) \\ &\leq f(\mathbf{x}_t) + p(\mathbf{x}_t) + \frac{\mathcal{L}+\eta}{2}\left\|\mathbf{x}_t-\overline{\mathbf{x}}_t\right\|^2 \\ &= F(\mathbf{x}_t) + \frac{\mathcal{L}+\eta}{2}\operatorname{dist}(\mathbf{x}_t,\Omega^*)^2 \\ &\leq F(\mathbf{x}_t) + 2(\mathcal{L}+\eta)\sigma^2\left\|\mathbf{x}_t-\widehat{\mathbf{x}}_{t-1}\right\|^2. \end{aligned} \tag{68}$$

So, by (67) and (68), we have (63) holds. In addition, it follows from (60), $\lim_{t\to\infty} d_t = 0$, (65) and (63) that

$$\lim_{t\to\infty}\widetilde{E}_t = \lim_{t\to\infty}F(\mathbf{x}_t) = F^*$$

and $\widetilde{E}_t \geq F^*$ for all t sufficiently large. So, by (67) and the definitions of $\widetilde{E}_t$ and d_t in (61) and (62), respectively, there exists a constant $c > 0$ such that $0 \leq (\widetilde{E}_t - F^*) \leq cd_t$ for t sufficiently large. Therefore, by (60) we have (64) holds with $\theta = (c-1)/c \in (0,1)$. □

Based on the Lemma 4.2, we can have the following linear convergence result.

Theorem 4.3. *Suppose the Assumptions 2.1, 3.1, 3.2 and 4.1 hold. If one of the following conditions hold:*

(a) the parameter $\bar{t} < \infty$ in Alg. 1;
(b) $\overline{\mu} > 0$, where $\overline{\mu} = \lim \mu_t$;
(c) $\nu > 0$, where ν is defined in (4);

we have

(i) the sequence $\{F(\mathbf{x}_t)\}$ converges R-linearly;
(ii) the sequence $\{\mathbf{x}_t\}$ converges R-linearly to a stationary point of problem (1).

Proof. Let us consider each of the following cases.
Case (a) $\bar{t} < \infty$ in Alg. 1. In this case, for all t sufficiently large, we have from Theorem 3.6 that (43) and (44) holds, which together with $\mathbf{g}(\widehat{\mathbf{x}}_t) = \eta_t(\widehat{\mathbf{x}}_t - \mathbf{x}_{t+1})$ and $\eta_t \geq L$ gives

$$\begin{aligned} E_{t+1} \quad \leq \quad & E_t - \frac{L-\mathcal{L}}{8}\|\widehat{\mathbf{x}}_t - \mathbf{x}_{t+1}\|^2 - \widehat{c}L\left(\|\mathbf{x}_{t-1} - \widetilde{\mathbf{x}}_t\|^2 + \|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\|^2\right) \\ & - \frac{\beta_t \nu}{2}\|\breve{\mathbf{x}}_{t+1} - \mathbf{x}_t\|^2, \end{aligned}$$

where $0 < \widehat{c} < 1/2$ is the constant in (44). By rearranging the above inequality with the definition of $\widetilde{E}_t$ in (61), we have

$$\begin{aligned} \widetilde{E}_{t+1} \quad \leq \quad & \widetilde{E}_t - \frac{L-\mathcal{L}}{8}\|\mathbf{x}_t - \widehat{\mathbf{x}}_{t-1}\|^2 - \widehat{c}L\left(\|\mathbf{x}_{t-1} - \widetilde{\mathbf{x}}_t\|^2 + \|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\|^2\right) \\ & - \frac{\beta_{t-1} \nu}{2}\|\breve{\mathbf{x}}_t - \mathbf{x}_{t-1}\|^2, \end{aligned}$$

which implies (60) holds with $\widetilde{c} = \min\{(L-\mathcal{L})/8, \widehat{c}L, 1/2\}$.
Case (b) $\overline{\mu} > 0$. If $\bar{t} = \infty$, since $\overline{\mu} > 0$, we have (52) holds. Then, similarly as the proof of Case (a), for t sufficiently large, we can establish (60) holds with $\widetilde{c} = \min\{(L-\mathcal{L})/8, c_1 L, 1/2\}$, where $c_1 > 0$ is the constant in (35). Hence, combining with Case (a), we have (60) holds with $\widetilde{c} = \min\{(L-\mathcal{L})/8, \widehat{c}L, c_1 L, 1/2\}$.
Case (c) $\nu > 0$. If $\bar{t} = \infty$, we have from Theorem 3.5 that (21) holds, which together with $\mathbf{g}(\widehat{\mathbf{x}}_t) = \eta_t(\widehat{\mathbf{x}}_t - \mathbf{x}_{t+1})$ and $\eta_t \geq L$ gives

$$E_{t+1} \leq E_t - \frac{L-\mathcal{L}}{8}\|\widehat{\mathbf{x}}_t - \mathbf{x}_{t+1}\|^2 - c_1 L\,\|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\|^2 - \frac{\beta_t \nu}{2}\|\breve{\mathbf{x}}_{t+1} - \mathbf{x}_t\|^2$$

for all t sufficiently large. Again, by rearranging this inequality with the definition of $\widetilde{E}_t$ in (61), we have

$$\widetilde{E}_{t+1} \leq \widetilde{E}_t - \frac{L-\mathcal{L}}{8}\|\mathbf{x}_t - \widehat{\mathbf{x}}_{t-1}\|^2 - c_1 L\,\|\mathbf{x}_{t-1} - \widetilde{\mathbf{x}}_t\|^2 - \frac{\beta_{t-1} \nu}{2}\|\breve{\mathbf{x}}_t - \mathbf{x}_{t-1}\|^2,$$

which together with $\beta_t \in (0, 1]$ and $\breve{\mathbf{x}}_t - \mathbf{x}_{t-1} = 1/\beta_{t-1}(\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1})$ implies

$$\begin{aligned} \widetilde{E}_{t+1} \quad \leq \quad & \widetilde{E}_t - \frac{L-\mathcal{L}}{8}\|\mathbf{x}_t - \widehat{\mathbf{x}}_{t-1}\|^2 - c_1 L\,\|\mathbf{x}_{t-1} - \widetilde{\mathbf{x}}_t\|^2 - \frac{\beta_{t-1} \nu}{4}\|\breve{\mathbf{x}}_t - \mathbf{x}_{t-1}\|^2 \\ & - \frac{\nu}{4}\|\widetilde{\mathbf{x}}_t - \mathbf{x}_{t-1}\|^2. \end{aligned}$$

Hence, for t sufficiently large we have from $\nu > 0$ that (60) holds with

$$\widetilde{c} = \min\{(L-\mathcal{L})/8, c_1 L, 1/4, \nu/4\} > 0.$$

By the previous analysis, under either condition (a), (b) or (c), we have (60) holds for sufficiently large t. So, by Lemma 4.2, we have (63) and (64) hold for t

sufficiently large. Hence, by (60), (63), (64) and the definition of d_t in (62), for t sufficiently large, we have

$$|F(\mathbf{x}_t) - F^*| \le \overline{\theta}\|\mathbf{x}_t - \widehat{\mathbf{x}}_{t-1}\|^2 \le \overline{\theta} d_t \le \frac{\overline{\theta}}{\widetilde{c}}(\widetilde{E}_t - \widetilde{E}_{t+1}) \le \frac{\overline{\theta}}{\widetilde{c}}(\widetilde{E}_t - F^*),$$

which together with (64) implies the R-linear convergence of $F(\mathbf{x}_t)$ to F^*, i.e., conclusion (i) holds. By (60), (64) and the definition of d_t in (62), we also have

$$\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 \le 2(\|\mathbf{x}_{t-1} - \widetilde{\mathbf{x}}_t\|^2 + \|\mathbf{x}_t - \widetilde{\mathbf{x}}_t\|^2) \le 2d_t \le \frac{2}{\widetilde{c}}(\widetilde{E}_t - F^*),$$

for t sufficiently large. This inequality and (64) show R-linear convergence of $\|\mathbf{x}_t - \mathbf{x}_{t-1}\|$, which implies there exists an $\mathbf{x}^*$ such that the sequence $\{\mathbf{x}_t\}$ converges to $\mathbf{x}^*$ R-linearly. Finally, the conclusion (i) of Lemma 4.2 shows $\mathbf{x}^*$ is a stationary point of problem (1). Hence, conclusion (ii) holds. □

5. **Numerical experiments.** In this section, we evaluate the performance of our unified gradient method with extrapolation (UPG-E) on solving two nonconvex composite optimization problems: the smoothly clipped absolute deviation (SCAD) penalty problem and the nonconvex quadratic programming with simplex constraint. We compare UPG-E with three other algorithms: the standard proximal gradient method (PG), the fast iterative shrinkage-thresholding algorithm (FISTA) [1] and the proximal gradient algorithm with extrapolation (PGE) [16].

5.1. **SCAD penalty problem.** In this subsection, we apply Alg. 1 to solve the smoothly clipped absolute deviation (SCAD) penalty problem, which is defined as

$$\min_{\mathbf{x}\in\mathbb{R}^n} \ \frac{1}{2}\|A\mathbf{x} - \mathbf{b}\|^2 + \sum_{i=1}^{n} g_\kappa(|\mathbf{x}_i|), \tag{69}$$

where $A \in \mathbb{R}^{m\times n}$, $\mathbf{b} \in \mathbb{R}^m$ and g_κ is the SCAD penalty defined as

$$g_\kappa(\theta) = \begin{cases} \kappa\theta, & \theta \le \kappa, \\ \dfrac{-\theta^2 + 2c\kappa\theta - \kappa^2}{2(c-1)}, & \kappa < \theta \le c\kappa, \\ \dfrac{(c+1)\kappa^2}{2}, & \theta > c\kappa, \end{cases} \tag{70}$$

with parameters $c > 2$ and $\kappa > 0$. The SCAD penalty corresponds to a quadratic spline function with knots at κ and $c\kappa$, and combines the benefits of using l_1 penalty and hard thresholding penalty [2]. The SCAD problem is often used in statistic applications for conducting variable selection when the noise level in the data is low. One may refer to [2] for more details about the SCAD penalty problem.

The SCAD problem (69) is possibly nonconvex due to the SCAD penalty term. However, it was shown that $g_\kappa(\cdot) + \frac{\omega}{2}|\cdot|^2$ with $\omega \ge \frac{1}{c-1}$ is convex [6]. Therefore, we can rewrite problem (69) into the form of (1) with

$$f(\mathbf{x}) := \frac{1}{2}\|A\mathbf{x} - \mathbf{b}\|^2 - \frac{1}{2(c-1)}\|\mathbf{x}\|^2 \text{ and } p(\mathbf{x}) := \sum_{i=1}^{n} g_\kappa(|\mathbf{x}_i|) + \frac{1}{2(c-1)}\|\mathbf{x}\|^2$$

so that f is Lipschitz continuously differentiable but possibly nonconvex and p is a convex function. Then, we can apply UPG-E, PG, PGE and FISTA to solve this reformulated problem. We have to point out that FISTA does not guarantee convergence when the objective function is nonconvex. We apply FISTA here simply

for practical numerical comparison purpose. In this case, the two subproblems in Alg. 1 can be solved as

$$\begin{aligned}\breve{\mathbf{x}}_{t+1} &= \arg\min_{\mathbf{x}\in\mathbb{R}^n}\left\{\langle\nabla f(\widehat{\mathbf{x}}_t),\mathbf{x}\rangle+\frac{\gamma_t}{2}\|\mathbf{x}-\breve{\mathbf{x}}_t\|^2+\sum_{i=1}^n g_\kappa(|\mathbf{x}_i|)+\frac{1}{2(c-1)}\|\mathbf{x}\|^2\right\}\\ &= \arg\min_{\mathbf{x}\in\mathbb{R}^n}\left\{\frac{1}{2\nu_1}\left\|\mathbf{x}-\frac{c-1}{\gamma_t(c-1)+1}\left(\gamma_t\breve{\mathbf{x}}_t-\nabla f(\widehat{\mathbf{x}}_t)\right)\right\|^2+\sum_{i=1}^n g_\kappa(|\mathbf{x}_i|)\right\},\end{aligned}$$

and

$$\begin{aligned}\mathbf{x}_{t+1} &= \arg\min_{\mathbf{x}\in\mathbb{R}^n}\left\{\langle\nabla f(\widehat{\mathbf{x}}_t),\mathbf{x}\rangle+\frac{\eta_t}{2}\|\mathbf{x}-\widehat{\mathbf{x}}_t\|^2+\sum_{i=1}^n g_\kappa(|\mathbf{x}_i|)+\frac{1}{2(c-1)}\|\mathbf{x}\|^2\right\}\\ &= \arg\min_{\mathbf{x}\in\mathbb{R}^n}\left\{\frac{1}{2\nu_2}\left\|\mathbf{x}-\frac{c-1}{\eta_t(c-1)+1}\left(\eta_t\widehat{\mathbf{x}}_t-\nabla f(\widehat{\mathbf{x}}_t)\right)\right\|^2+\sum_{i=1}^n g_\kappa(|\mathbf{x}_i|)\right\},\end{aligned}$$

where $\nu_1=\frac{c-1}{\gamma_t(c-1)+1}$ and $\nu_2=\frac{c-1}{\eta_t(c-1)+1}$. It can be easily verified that $1+\nu_i\le c$ holds for $i=1,2$. In addition, it is known that the following minimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^n}\ \frac{1}{2\nu}\|\mathbf{x}-\mathbf{q}\|^2+\sum_{i=1}^n g_\kappa(|\mathbf{x}_i|) \tag{71}$$

with $1+\nu\le c$ and known $\mathbf{q}$ has a closed form solution [17]. Hence, the subproblems for obtaining $\breve{\mathbf{x}}_{t+1}$ and $\mathbf{x}_{t+1}$ in Alg. 1 can be solved trivially. Also note that the subproblems in PG, PGE and FISTA can be also written in the format as (71).

In the numerical experiments, we vary the problem dimensions (n,m) as those listed in Table 1. The matrix $A\in\mathbb{R}^{m\times n}$ is generated with entries randomly generated from standard normal distribution. The vector $\mathbf{b}$ is obtained as $\mathbf{b}=A\mathbf{b}^*+\boldsymbol{\epsilon}$, where $\mathbf{b}^*\in\mathbb{R}^n$ is a sparse uniformly distributed randomly generated vector with density of 0.02 and $\boldsymbol{\epsilon}\in\mathbb{R}^m$ is a noise vector with entries being generated from normal distribution with mean 0 and standard deviation 0.01. The parameters c and κ could be chosen by cross-validation in practice. Here, we simply choose $c=3.7$ and $\kappa=0.1$. For all the comparison algorithms, we set the Lipschitz constant of ∇f as $\mathcal{L}=\max\{|\lambda_{\max}(H)|,|\lambda_{\min}(H)|\}$, where $H=A^{\mathsf{T}}A-c_1I$ is the Hessian of $f(\mathbf{x})$ with $c_1=1/(2(c-1))$. Here, $\lambda_{\max}(H)$ and $\lambda_{\min}(H)$ are the largest and smallest eigenvalues of H, respectively. We set $l=|\lambda_{\min}(H)|$ and take $\beta_t=0.85\sqrt{\frac{\mathcal{L}}{\mathcal{L}+l}}$ for better performance of PGE. We choose $\rho=1.5$, $\lambda=0.5$ and $\bar{t}=\min\{\lfloor 0.15\min\{n,m\}\rfloor,100\}$ for UPG-E.

The same starting point $\mathbf{x}_0\in\mathbb{R}^n$ for all the comparison algorithms is randomly selected with entries chosen from the uniform distribution in $(0,1)$. For all four comparison algorithms, the algorithm stops when either

$$\frac{\|\mathbf{x}_{t+1}-\mathbf{x}_t\|}{\max\{\|\mathbf{x}_{t+1}\|,1\}}\le 10^{-6}, \tag{72}$$

or the number of iterations exceeds 5000. The computational results are reported in Table 1, where "iter" is the number of iterations where the algorithms stops and "fval" is the minimum function value found by the algorithm. The "F" in Table 1 means the algorithm fails to find a reasonable solution in 5000 iterations. We can see from the experimental results that among all the comparison algorithms, UPG-E is very effective and always converges in much less number of iterations. Moreover, UPG-E is very robust. As the problem dimension n and m increase, the other

TABLE 1. Comparison of UPG-E, PGE, FISTA and PG for solving the SCAD problem (69)

	UPG-E		PGE		PG		FISTA	
(n, m)	iter	fval	iter	fval	iter	fval	iter	fval
(400, 200)	122	9.400	501	9.400	1401	9.400	4163	9.568
(600, 200)	91	1.410e1	180	1.410e1	322	1.410e1	4824	1.410e1
(600, 400)	230	1.410e1	F		F		F	
(800, 200)	72	1.880e1	221	1.880e1	698	1.878e1	4833	1.880e1
(800, 400)	163	1.880e1	366	1.880e1	3312	1.880e1	F	
(800, 600)	362	1.880e1	F		F		F	
(1000, 200)	62	2.350e1	157	2.348e1	1951	2.345e1	3309	2.350e1
(1000, 400)	139	2.350e1	254	2.350e1	514	2.350e1	F	
(1000, 600)	217	2.350e1	2207	2.350e1	F		F	
(1000, 800)	602	2.350e1	F		F		F	

comparison methods start to fail while UPG-E still solves the problem in a good number of iterations.

5.2. **Nonconvex quadratic programming with simplex constraints.** In this subsection, we consider the following possibly nonconvex problem

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^n} \quad & \frac{1}{2}\mathbf{x}^\mathsf{T} H\mathbf{x} - \mathbf{g}^\mathsf{T}\mathbf{x}, \\ \text{s.t.} \quad & \mathbf{e}^\mathsf{T}\mathbf{x} = c,\ \mathbf{x} \geq \mathbf{0}, \end{aligned} \tag{73}$$

where $H \in \mathbb{R}^{n\times n}$ is not necessarily positive semidefinite, $\mathbf{g} \in \mathbb{R}^n$, $\mathbf{e} \in \mathbb{R}^n$ is a vector of ones and c is a positive number. We can easily rewrite (73) in the form of (1) with

$$f(\mathbf{x}) := \frac{1}{2}\mathbf{x}^\mathsf{T} G\mathbf{x} - \mathbf{g}^\mathsf{T}\mathbf{x} \qquad \text{and} \qquad p(\mathbf{x}) := \delta_{\mathcal{C}}(\mathbf{x}),$$

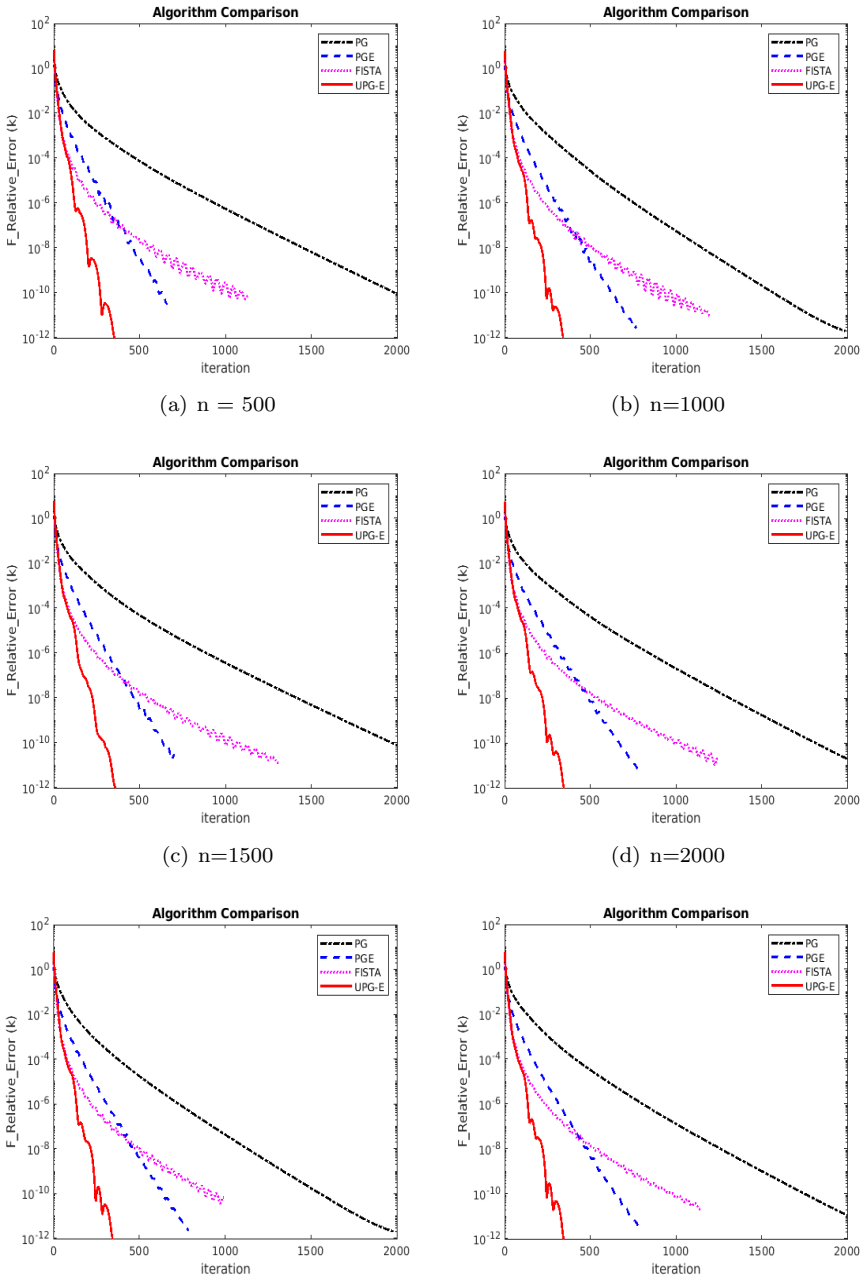
where $\mathcal{C} = \{\mathbf{y} \in \mathbb{R}^n : \mathbf{e}^\mathsf{T}\mathbf{y} = p,\ \mathbf{y} \geq \mathbf{0}\}$ and $\delta_{\mathcal{C}}(\cdot)$ is the indicator function of the simplex $\mathcal{C}$. Note that p is a closed convex function since $\mathcal{C}$ is a closed convex set.

In this experiment, we vary the problem dimensions n as those listed in Table 2. The matrix $G \in \mathbb{R}^{n\times n}$ is generated with entries randomly generated from the normal distribution with mean 0 and standard deviation 10. Then, we set $H = G'DG$, where $D \in \mathbb{R}^{n\times n}$ is a diagonal matrix with diagonal entries $D(i,i) = i - 20$. The vector $\mathbf{g}$ is generated by a standard normal distribution and the constant c is selected as $\max\{1, 10 * t\}$, where t is a random scalar uniformly generated in $[0, 1]$. Again, for all comparison algorithms, the Lipschitz constant of ∇f is set as $\mathcal{L} = \max\{|\lambda_{\max}(H)|, |\lambda_{\min}(H)|\}$. As suggested in [16] for solving (73), we set $l = |\lambda_{\min}(G)|$ and $\beta_t = 0.98\sqrt{\frac{\mathcal{L}}{\mathcal{L}+l}}$ for PGE. We again choose the parameter $\rho = 1.5$, $\lambda = 0.5$ and $\bar{t} = \min\{\lfloor 0.15n \rfloor, 100\}$ for UPG-E.

The same feasible starting point $\mathbf{x}_0 = (c/n)\mathbf{e}$ is used for all comparison algorithms. And for all the comparison algorithms, the algorithm stops when either

$$\frac{|F(\mathbf{x}_k) - F(\mathbf{x}_{k-1})|}{\max\{|F(\mathbf{x}_k)|, 1\}} \leq 10^{-12}$$

or the number of iterations exceeds 5000. The computational results are reported in Table 2, where "iter" is the number of iterations where the method stops and



(a) n = 500 (b) n=1000

(c) n=1500 (d) n=2000

(e) n=2500 (f) n=3000

FIGURE 1. Comparison of UPG-E, PG, PGE and FISTA for the NQP problem: Relative function value gap vs iterations

TABLE 2. Comparison of UPG-E, PGE, FISTA and PG for solving the NQP problem (73)

	UPG-E		PGE		PG		FISTA	
n	iter	fval	iter	fval	iter	fval	iter	fval
500	357	1.0434e5	665	1.0434e5	2497	1.0434e5	1134	1.0434e5
1000	346	3.1052e4	772	3.1052e4	1989	3.1052e4	1199	3.1052e4
1500	362	3.9881e5	704	3.9881e5	2486	3.9881e5	1310	3.9881e5
2000	349	2.7083e5	797	2.7083e5	2309	2.7083e5	1248	2.7083e5
2500	347	1.5133e5	790	2.7083e5	1980	2.7083e5	991	2.7083e5
3000	347	6.6941e5	796	6.6941e5	2229	6.6941e5	1142	6.6941e5

"fval" is the minimum function value found by the algorithm. Moreover, we plot the relative function value gap against the iteration number in Figure 1, where

$$\text{F\_Relative\_Error}(k) := \frac{|F(\mathbf{x}_k) - F_{\min})|}{\max\{|F_{\min}|, 1\}}$$

and $F_{\min}$ is the minimum objective function value obtained by all the comparison algorithms. We can see from both Table 2 and Figure 1 that UPG-E outperforms the other methods for solving this class of randomly generated *nonconvex* quadratic problems (73) with simplex constraints. Furthermore, since the quadratic problem (73) satisfies the Assumption 4.1 required for linear convergence, we can clearly see from Figure 1 that the function value gap generated by UPG-E converges R-linearly to zero for solving this quadratic programming problem.

6. **Conclusion.** In this work, we propose a unified proximal gradient method with extrapolation (UPG-E) to solve a possibly nonconvex and nonsmooth composite optimization problem, where one of the component functions in the objective is s-mooth but possibly nonconvex and the other one is convex but could be nonsmooth. The UPG-E exploits an extrapolation step to accelerate the convergence, where the extrapolation parameter is adaptively adjusted according to the nonconvexity modulus of the smooth component objective function estimated by a line search technique. It is shown that UPG-E automatically maintains optimal convergence rate of proximal gradient methods for minimizing convex composite optimization when extrapolation is not restarted and also ensures global convergence when the objective function is nonconvex. Under further proper regularity assumptions, UPG-E is shown to have a linear convergence rate for both the objective function value gap and the generated iterates. Our numerical experiments show that UPG-E is very robust and could significantly outperform the other well-established methods in the literature for solving the possibly nonconvex and nonsmooth composite optimization (1).

REFERENCES

[1] A. Beck and M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, *SIAM J. Imaging Sci.*, **2** (2009), 183-202.
[2] J. Fan and R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, *J. Am. Stat. Assoc.*, **96** (2001), 1348-1360.
[3] S. Ghadimi and G. Lan, Accelerated gradient methods for nonconvex nonlinear and stochastic programming, *Math. Program.*, **156** (2016), 59-99.

[4] S. Ghadimi, G. Lan and H. Zhang, Generalized uniformly optimal methods for nonlinear programming, *J. Sci. Comput.*, **79** (2019), 1854-1881.
[5] P. Gong, C. Zhang, Z. Lu, J. Huang and J. Ye, A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems, in *International Conference on Machine Learning*, PMLR, (2013), 37-45.
[6] K. Guo, X. Yuan and S. Zeng, Convergence analysis of ISTA and FISTA for "strongly + semi" convex programming, *Optimization Online*, 2016.
[7] T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning*, Springer Series in Statistics, Springer New York Inc., New York, NY, USA, 2001.
[8] N. Ito, A. Takeda and K.-C. Toh, A unified formulation and fast accelerated proximal gradient method for classification, *J. Mach. Learn. Res.*, **18** (2017), 510-558.
[9] G. Lan, Z. Li and Y. Zhou, A unified variance-reduced accelerated gradient method for convex optimizationy, in *Neural Information Processing Systems*, 2019.
[10] H. Li and Z. Lin, Accelerated proximal gradient methods for nonconvex programming, in *Neural Information Processing Systems*, 2015.
[11] Z.-Q. Luo and P. Tseng, Error bounds and convergence analysis of feasible descent methods: A general approach, *Ann. Oper. Res.*, **46** (1993), 157-178.
[12] Y. E. Nesterov, A method for solving the convex programming problem with convergence rate $\mathcal{O}(1/k^2)$ (in russian), *Dokl. Akad. Nauk SSSR*, **269** (1983), 543-547.
[13] Y. E. Nesterov, *Introductory Lectures on Convex Optimization. A Basic Course*, Kluwer, Boston, 2004.
[14] P. Tseng, Approximation accuracy, gradient methods, and error bound for structured convex optimization, *Math. Program. Ser. B*, **125** (2010), 263-295.
[15] P. Tseng and S. Yun, A coordinate gradient descent method for nonsmooth separable minimization, *Mathematical Programming*, **117** (2009), 387-423.
[16] B. Wen, X. Cheng and T. K. Pong, Linear convergence of proximal gradient algorithm with extrapolation for a class of nonconvex nonsmooth minimization problems, *SIAM J. Optim.*, **27** (2017), 124-145.
[17] Z. Wu and M. Li, General inertial proximal gradient method for a class of nonconvex nonsmooth optimization problems, *Comput. Optim. Appl.*, **73** (2019), 129-158.
[18] W. Zhong and J. Kwok, Gradient descent with proximal average for nonconvex and composite regularization, *Proceedings of the AAAI Conference on Artificial Intelligence*, 2014.

Received October 2023; 1^{st} revision January 2024; final revision February 2024; early access February 2024.