SIAM J. OPTIM.
Vol. 35, No. 3, pp. 2066–2097
© 2025 Society for Industrial and Applied Mathematics

Downloaded 10/22/25 to 128.62.19.54 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

GENERALIZED OPTIMISTIC METHODS FOR CONVEX-CONCAVE SADDLE POINT PROBLEMS*

RUICHEN JIANG† AND ARYAN MOKHTARI†

Abstract. The optimistic gradient method has seen increasing popularity as an efficient first-order method for solving convex-concave saddle point problems. To analyze its iteration complexity, a recent work [A. Mokhtari, A. E. Ozdaglar, and S. Pattathil, *SIAM J. Optim.*, 30 (2020), pp. 3230–3251] proposed an interesting perspective that interprets the optimistic gradient method as an approximation to the proximal point method. In this paper, we follow this approach and distill the underlying idea of optimism to propose a generalized optimistic method, which encompasses the optimistic gradient method as a special case. Our general framework can handle constrained saddle point problems with composite objective functions. Moreover, we also develop a backtracking line search scheme to select the step sizes without knowledge of the smoothness coefficients. By instantiating our general framework with a second-order oracle, we propose a second-order optimistic method and prove a complexity bound of $\mathcal{O}(\epsilon^{-2/3})$ in terms of the primal-dual gap in the convex-concave setting and a complexity bound of $\mathcal{O}((L_2 D/\mu)^{2/3} + \log\log(1/\epsilon))$ in terms of the distance to the optimal solution in the strongly-convex-strongly-concave setting, where L_2 is the Lipschitz constant of the Jacobian, μ is the strong convexity parameter, and D is the initial Euclidean distance to the saddle point. We also establish convergence rates in terms of the tangent residual, which generalizes the operator norm as a metric in the unconstrained setting. Moreover, our line search scheme provably only requires a constant number of calls to a subproblem solver per iteration on average.

Key words. convex-concave saddle point problems, optimistic methods, second-order methods

MSC codes. 90C25, 90C33, 90C47

DOI. 10.1137/24M1630475

1. Introduction.

In this paper, we study convex-concave saddle point problems, also known as minimax optimization problems, where the objective function for both minimization and maximization has a composite structure. Specifically, we consider

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \ \ell(\mathbf{x},\mathbf{y}) := f(\mathbf{x},\mathbf{y}) + h_1(\mathbf{x}) - h_2(\mathbf{y}), \tag{1.1}$$

where $\mathcal{X} \subset \mathbb{R}^m$ and $\mathcal{Y} \subset \mathbb{R}^n$ are nonempty closed convex sets, $h_1 : \mathbb{R}^m \to (-\infty, +\infty]$ and $h_2 : \mathbb{R}^n \to (-\infty, +\infty]$ are proper closed convex functions, and f is a smooth function defined on some open set $\operatorname{dom} f \subset \mathbb{R}^m \times \mathbb{R}^n$ containing $\mathcal{X} \times \mathcal{Y}$. Throughout this paper, we assume that f is convex with respect to $\mathbf{x}$ and concave with respect to $\mathbf{y}$, i.e., $f(\cdot, \mathbf{y})$ is convex for each $\mathbf{y} \in \mathcal{Y}$ and $f(\mathbf{x}, \cdot)$ is concave for each $\mathbf{x} \in \mathcal{X}$. As a result, the objective function ℓ is also convex-concave.

Problem (1.1), which we refer to as the composite saddle point problem, encompasses several important special cases. For instance, when $h_1 \equiv 0$ on $\mathcal{X}$ and $h_2 \equiv 0$ on $\mathcal{Y}$, it becomes the constrained smooth saddle point problem [24, 48, 41]. If we further have $\mathcal{X} = \mathbb{R}^m$ and $\mathcal{Y} = \mathbb{R}^n$, it reduces to the unconstrained smooth saddle point problem [12, 37, 38]. Problem (1.1) is of central importance in the duality

*Received by the editors January 11, 2024; accepted for publication (in revised form) June 16, 2025; published electronically September 16, 2025.
https://doi.org/10.1137/24M1630475
Funding: This work was supported by the National Science Foundation under grants 2338846 and 2007668.
†Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712 USA (rjiang@utexas.edu, mokhtari@austin.utexas.edu).

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 10/22/25 to 128.62.19.54 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

theory of constrained optimization and appears as the primal-dual formulation of conic programming [17] and regularized empirical risk minimization [55].

Various iterative methods—mostly first-order—have been proposed in the literature for solving saddle point problems. In particular, we will focus on optimistic methods, whose idea was first introduced by Popov [48] and have gained much attention recently in the machine learning community [11, 49, 12, 28]. Specifically, we follow and extend the approach in [37, 38] by interpreting the optimistic method as an approximation of the proximal point method (PPM) [35, 51]. In particular, we propose a generalized optimistic method (see Algorithm 4.1), where the future gradient required in the update of the PPM is replaced by the combination of a prediction term and a correction term. The prediction term serves as a local approximation of the future gradient, while the correction term is given by the prediction error at the previous iteration. Moreover, we develop a simple backtracking line search scheme (Algorithm 4.2) to select the step sizes in our generalized optimistic method adaptively without knowledge of the smoothness coefficients of the objective. By instantiating our method with a second-order oracle, we obtain an adaptive second-order optimistic method for the saddle point problem in (1.1) and prove the best-known complexity bounds in both convex-concave and strongly-convex-strongly-concave settings. The complexity of our line search scheme in terms of the number of calls to a subproblem solver is also presented. Our theoretical findings are summarized as follows:

(a) In the convex-concave case, we show that our proposed second-order optimistic method achieves a complexity of $\mathcal{O}(\epsilon^{-\frac{2}{3}})$ for the averaged iterate in terms of the primal-dual gap and a complexity of $\mathcal{O}(\epsilon^{-1})$ for the best iterate in terms of the tangent residual (see section 2 for precise definitions).

(b) In the strongly-convex-strongly-concave case, we prove global convergence and a local R-superlinear convergence rate of the order $\frac{3}{2}$ for our method, leading to an overall complexity of $\mathcal{O}((\frac{L_2 D}{\mu})^{\frac{2}{3}} + \log\log\frac{1}{\epsilon})$ in terms of the distance to the optimal solution. Here, L_2 is the Lipschitz constant of the Jacobian, μ is the strong convexity parameter, and D is the initial Euclidean distance to the optimal solution. We also establish a similar complexity bound in terms of the tangent residual.

(c) In both settings, we prove that the line search procedure on average only requires a constant number of calls to a subproblem solver per iteration. Additionally, we analyze the total computational cost of our subproblem solver under specific constraints and nonsmooth terms h_1 and h_2.

Remark 1.1. For ease of presentation, we restrict ourselves to Euclidean norms, but our framework can also handle non-Euclidean norms with Bregman distances; see our technical report [22] for more details. Further, our convergence analysis is not tied to saddle point structures and easily extends to general variational inequality problems. Moreover, our generalized optimistic method can be instantiated with first- or higher-order oracles, yielding respective optimistic methods. Here, we focus on the second-order case, deferring first- and higher-order results to [22].

1.1. Related work. Next, we briefly recap related work on optimistic methods and second-order methods for minimax optimization.

Optimistic methods. The idea of optimism dates back to Popov [48], who proposed a modification of the Arrow–Hurwicz method [3] for solving constrained saddle point problems. It was later rediscovered in the context of online learning [11, 49, 50] and has been used in training of Generative Adversarial Networks (GANs) [12, 16, 36, 28], sometimes under the name of optimistic gradient descent

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 10/22/25 to 128.62.19.54 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

ascent (OGDA). When the objective is L_1-smooth and μ-strongly-convex-strongly-concave, it is shown to have a complexity of $\mathcal{O}(\kappa_1 \log(\frac{1}{\epsilon}))$, $\kappa_1 := L_1/\mu$, for constrained saddle point problems [16, 4, 37]; see also a similar result in [34] for a variant of Popov's method on monotone inclusion problems. Moreover, in [38], OGDA was viewed as an approximation of the PPM and proved an ergodic complexity of $\mathcal{O}(1/\epsilon)$ when the objective is smooth and convex-concave. Several other variants of Popov's method have also appeared in the literature [32, 33, 25] with similar convergence guarantees; see [19, 34] for more details. Notably, all the above works focus on gradient-based first-order methods, while this paper pioneers extending the optimistic idea to second- and higher-order methods.

Second-order methods. Unlike first-order methods, only a few methods have been proposed for incorporating second-order information in saddle point problems, mainly focusing on generalizing cubic regularized Newton's method [45, 44] and generalizing the first-order extragradient method [24, 41]. In this paper, we offer a third alternative with comparable and sometimes better convergence properties.

In [42], the cubic regularized Newton method [45] was extended to variational inequalities, achieving $\mathcal{O}(1/\epsilon)$ complexity for monotone operators. For μ-strongly-monotone operators, a local quadratic convergence was shown with an overall complexity of $\mathcal{O}(\kappa_2 + \log\log(\epsilon^{-1}))$. Here, $\kappa_2 := \frac{L_2 D}{\mu}$, where L_2 is the Lipschitz constant of the Hessian and D is the initial Euclidean distance to the optimal solution. A new cubic regularization scheme for unconstrained saddle point problems was presented in [20], with an iteration complexity of $\mathcal{O}(\kappa_1^2 + \kappa_1\kappa_2 + \log\log(\epsilon^{-1}))$, when the objective is strongly-convex-strongly-concave and its gradient and Hessian are Lipschitz continuous. Note that these bounds are worse than the ones presented in this paper.

Under the hybrid proximal extragradient (HPE) framework [52], a Newton-type extragradient method was proposed in [39, 40] for solving monotone inclusion problems and was shown to achieve a complexity of $\mathcal{O}(\epsilon^{-\frac{2}{3}})$. Recently, a similar second-order method was proposed in [6] by directly extending the classical mirror-prox method [41], achieving the same complexity. Later, [46] built upon this work and proposed a restarted variant in the strongly-convex-strongly-concave setting, achieving a complexity of $\mathcal{O}(\kappa_2^{2/3} \log(\epsilon^{-1}))$. Under the additional assumption that the gradient of f is Lipschitz, the authors further combined the restart scheme with cubic regularized Newton in [20] to obtain a complexity of $\mathcal{O}(\kappa_2^{2/3} \log(\kappa_1\kappa_2) + \log\log(\epsilon^{-1}))$. While these works achieve comparable iteration complexities as ours, their total complexity is worse due to the additional cost caused by their line search. Specifically, the two-stage line search subroutine in [40] requires $O(\log\log(\epsilon^{-1}))$ calls to a subproblem solver per iteration. A simpler binary search subroutine is considered in [6] with a higher per-iteration cost of $\mathcal{O}(\log(\epsilon^{-1}))$. In comparison, our line search scheme on average requires a constant number of calls to a subproblem solver per iteration, thus shaving the additional factor of $\mathcal{O}(\log\log(\epsilon^{-1}))$. Moreover, our second-order optimistic method is conceptually different from the extragradient approach in [39, 40, 6, 46].

Since the initial release of this paper, new developments have emerged in second-order methods for saddle point problems. Specifically, to eliminate the need of line search, the methods proposed in [1, 30, 31, 21] combine the idea of cubic regularization [45] with either mirror-prox [41] or dual extrapolation [43]. In the convex-concave setting, they all achieve a complexity of $\mathcal{O}(\epsilon^{-\frac{2}{3}})$, which matches the lower bound in [1, 30] under certain assumptions (see [30] for details). For smooth and uniformly monotone variational inequalities (i.e., F satisfies $\langle F(\mathbf{z}) - F(\mathbf{z}'), \mathbf{z} - \mathbf{z}'\rangle \geq \mu\|\mathbf{z} - \mathbf{z}'\|^3$ for all $\mathbf{z}, \mathbf{z}' \in \mathbb{R}^d$ for some constant $\mu > 0$), the authors in [30] further proposed a

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 10/22/25 to 128.62.19.54 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

restart scheme to obtain a complexity of $\mathcal{O}(\kappa_2^{2/3}\log(\epsilon^{-1}))$. Later, a similar restart scheme was studied in [21] in the strongly-convex-strongly-concave setting, leading to a complexity of $\mathcal{O}(\kappa_2^{2/3}\log\log(\epsilon^{-1}))$. However, although these methods avoid an explicit line search subroutine, each iteration requires solving a cubic regularization subproblem, which may be costlier than ours. For instance, in the unconstrained setting, [1, 31] showed that the subproblem can be reformulated as a one-dimensional nonlinear equation and they proposed solving it with a bisection scheme or Newton's iteration, respectively, requiring $\mathcal{O}(\log(\epsilon^{-1}))$ or $\mathcal{O}(\log\log(\epsilon^{-1}))$ calls to a linear system solver. In comparison, on average each iteration of our method requires solving a constant number of subproblems, which are in the form of affine variational inequalities and reduce to linear systems of equations in unconstrained settings. Hence, the overall complexity of our method can be advantageous at least in unconstrained cases.

Finally, the concurrent work [2], based on the large-step HPE framework [40, 53], introduces a line-search-free second-order method. Using the same subproblem solver as ours, it achieves a complexity of $\mathcal{O}(\epsilon^{-\frac{2}{3}} + \log(\epsilon^{-1}))$ for monotone variational inequalities. However, unlike ours, their method requires knowledge of the Hessian's Lipschitz constant. Further, our method addresses both convex-concave and strongly-convex-strongly-concave cases, whereas their extension to the latter is unclear.

2. Preliminaries.

2.1. Monotonicity and smoothness.

Since our convergence analysis relies heavily on the monotone operator theory, we review some of its basic concepts in this section. A *set-valued operator* $T : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ maps a point $\mathbf{z} \in \mathbb{R}^d$ to a subset of $\mathbb{R}^d$, and its domain is given by $\operatorname{dom} T = \{\mathbf{z} \in \mathbb{R}^d : T(\mathbf{z}) \neq \emptyset\}$. Alternatively, we can identify T with its graph defined as $\operatorname{Gr}(T) = \{(\mathbf{z}, \mathbf{v}) : \mathbf{v} \in T(\mathbf{z})\} \subset \mathbb{R}^d \times \mathbb{R}^d$. With these notations, we introduce the class of monotone operators.

DEFINITION 2.1. *A set-valued operator* $T : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ *is* μ-strongly monotone *if for all* $(\mathbf{z}, \mathbf{v}), (\mathbf{z}', \mathbf{v}') \in \operatorname{Gr}(T)$*, we have* $\langle \mathbf{v} - \mathbf{v}', \mathbf{z} - \mathbf{z}' \rangle \geq \mu \|\mathbf{z} - \mathbf{z}'\|^2$*. Also, a monotone operator* T *is* maximal *if there is no other monotone operator* S *that* $\operatorname{Gr}(T) \subsetneq \operatorname{Gr}(S)$.

Note that by setting $\mu = 0$ above, we recover the definition for a monotone operator.

For a smooth operator $F : \operatorname{dom} F \to \mathbb{R}^d$ defined on an open domain $\operatorname{dom} F \subset \mathbb{R}^d$, we use $DF(\mathbf{z}) \in \mathbb{R}^{d \times d}$ to denote its Jacobian at $\mathbf{z}$. In this paper, we focus on smooth operators with Lipschitz Jacobian, defined as follows.

DEFINITION 2.2. *We state that* F *is* second-order L_2-smooth *if for any* $\mathbf{z}, \mathbf{z}' \in \operatorname{dom} F$ *we have* $\|DF(\mathbf{z}) - DF(\mathbf{z}')\|_{\mathrm{op}} \leq L_2 \|\mathbf{z} - \mathbf{z}'\|$.

We define the *first-order Taylor expansion* of F at point $\mathbf{z}$ by $T^{(1)}(\mathbf{z}'; \mathbf{z}) = F(\mathbf{z}) + DF(\mathbf{z})(\mathbf{z}' - \mathbf{z})$. For a second-order smooth operator F, its first-order Taylor expansion can be considered as its approximation. Specifically, if F is second-order L_2-smooth, we have $\|F(\mathbf{z}') - T^{(1)}(\mathbf{z}'; \mathbf{z})\| \leq \frac{L_2}{2}\|\mathbf{z} - \mathbf{z}'\|^2$ for all $\mathbf{z}, \mathbf{z}' \in \operatorname{dom} F$.

2.2. Saddle point problems.

An optimal solution $(\mathbf{x}^*, \mathbf{y}^*)$ of (1.1) is also called a *saddle point* of the objective function ℓ, as it satisfies $\ell(\mathbf{x}^*, \mathbf{y}) \leq \ell(\mathbf{x}^*, \mathbf{y}^*) \leq \ell(\mathbf{x}, \mathbf{y}^*)$ for any $\mathbf{x} \in \mathcal{X}$ and any $\mathbf{y} \in \mathcal{Y}$. Throughout this paper, we assume that problem (1.1) has at least one saddle point in $\mathcal{X} \times \mathcal{Y}$. There are several ways to reformulate the saddle point problem. To simplify the notation, let $d := m + n$, $\mathbf{z} := (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^d$ and $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$. Define the operators $F : \mathcal{Z} \to \mathbb{R}^d$ and $H : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ by

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

$$F(\mathbf{z}) := \begin{bmatrix} \nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y}) \\ -\nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y}) \end{bmatrix} \quad \text{and} \quad H(\mathbf{z}) := \begin{bmatrix} \partial h_1(\mathbf{x}) \\ \partial h_2(\mathbf{y}) \end{bmatrix} + \begin{bmatrix} N_{\mathcal{X}}(\mathbf{x}) \\ N_{\mathcal{Y}}(\mathbf{y}) \end{bmatrix}, \tag{2.1}$$

where ∂h_1 and ∂h_2 are the subdifferentials of h_1 and h_2, respectively, and $N_{\mathcal{X}}$ and $N_{\mathcal{Y}}$ are the normal cone operators of $\mathcal{X}$ and $\mathcal{Y}$, respectively. In the convex-concave setting, finding a saddle point of (1.1) is equivalent to the monotone inclusion problem:

$$\text{find } \mathbf{z}^* \in \mathbb{R}^d \text{ such that } 0 \in F(\mathbf{z}^*) + H(\mathbf{z}^*). \tag{2.2}$$

Moreover, if $h(\mathbf{z}) := h_1(\mathbf{x}) + h_2(\mathbf{y})$, (1.1) is equivalent to the variational inequality:

$$\text{find } \mathbf{z}^* \in \mathcal{Z} \text{ such that } \langle F(\mathbf{z}^*), \mathbf{z} - \mathbf{z}^* \rangle + h(\mathbf{z}) - h(\mathbf{z}^*) \geq 0 \ \forall \mathbf{z} \in \mathcal{Z}. \tag{2.3}$$

In this paper, we consider two specific problem classes for our theoretical results: (i) convex-concave saddle point problems (Assumption 2.3) and (ii) strongly-convex-strongly-concave saddle point problems (Assumption 2.4).

Assumption 2.3. The functions h_1 and h_2 are proper closed convex, and the function f is convex in $\mathbf{x}$ and concave in $\mathbf{y}$ on $\mathcal{X} \times \mathcal{Y}$.

Assumption 2.4. The functions h_1 and h_2 are proper closed convex, and the function f is μ-strongly convex in $\mathbf{x}$ and μ-strongly concave in $\mathbf{y}$ on $\mathcal{X} \times \mathcal{Y}$.

Note that Assumption 2.4 reduces to Assumption 2.3 when $\mu = 0$, and sometimes this allows us to unify our results in both settings by considering $\mu \geq 0$. Here, we mainly work with the operator F to develop our algorithms and convergence proof. The following lemma serves the purpose of translating the properties of the function f into the properties of the operator F and is well-known in the literature (see, e.g., [15, section 1.4.1]). For completeness, we present the result below.

LEMMA 2.5. *The operator F defined in* (2.1) *is monotone on $\mathcal{Z}$ under Assumption* 2.3*, and is μ-strongly monotone on $\mathcal{Z}$ under Assumption* 2.4*.*

In the convex-concave case, we measure the suboptimality of a solution $(\mathbf{x},\mathbf{y})$ by $\Delta(\mathbf{x},\mathbf{y}) := \max_{\mathbf{y}' \in \mathcal{Y}} \ell(\mathbf{x},\mathbf{y}') - \min_{\mathbf{x}' \in \mathcal{X}} \ell(\mathbf{x}',\mathbf{y})$, which is known as the *primal-dual gap*. We note that $\Delta(\mathbf{x},\mathbf{y}) \geq 0$ for any $(\mathbf{x},\mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$ and the equality holds if and only if $(\mathbf{x},\mathbf{y})$ is a saddle point. However, in some cases where the feasible sets are unbounded, the primal-dual gap can be always infinite except at the saddle points, rendering it useless. One remedy is to use the restricted primal-dual gap function [43, 10]. Given two bounded sets $\mathcal{B}_1 \subset \mathcal{X}$ and $\mathcal{B}_2 \subset \mathcal{Y}$, we define the *restricted primal-dual gap* as

$$\Delta_{\mathcal{B}_1 \times \mathcal{B}_2}(\mathbf{x},\mathbf{y}) := \max_{\mathbf{y}' \in \mathcal{B}_2} \ell(\mathbf{x},\mathbf{y}') - \min_{\mathbf{x}' \in \mathcal{B}_1} \ell(\mathbf{x}',\mathbf{y}). \tag{2.4}$$

As discussed in [10], the gap function has two properties: (i) If $(\mathbf{x}^*,\mathbf{y}^*) \in \mathcal{B}_1 \times \mathcal{B}_2$, then we have $\Delta_{\mathcal{B}_1 \times \mathcal{B}_2}(\mathbf{x},\mathbf{y}) \geq 0$ for any $(\mathbf{x},\mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$. (ii) If $(\mathbf{x},\mathbf{y})$ is in the interior of $\mathcal{B}_1 \times \mathcal{B}_2$, then $\Delta_{\mathcal{B}_1 \times \mathcal{B}_2}(\mathbf{x},\mathbf{y}) = 0$ if and only if $(\mathbf{x},\mathbf{y})$ is a saddle point. As we shall show later, the iterates of our algorithm stay in a bounded set centered at $(\mathbf{x}^*,\mathbf{y}^*)$. Hence, the restricted primal-dual gap function will serve as a good measure of suboptimality if we choose $\mathcal{B}_1 \times \mathcal{B}_2$ large enough such that it contains a saddle point $(\mathbf{x}^*,\mathbf{y}^*)$.

The following result is the first step of our analysis, providing an upper bound on the primal-dual gap at the averaged iterate (see, e.g., [41, Proposition 2.2]).

LEMMA 2.6. *Suppose Assumption* 2.3 *holds. For any $\mathbf{z}, \mathbf{z}_1, \ldots, \mathbf{z}_N \in \mathcal{Z}$ and $\theta_1, \ldots, \theta_N \geq 0$ with $\sum_{k=1}^N \theta_k = 1$, we have $\ell(\bar{\mathbf{x}}_N, \mathbf{y}) - \ell(\mathbf{x}, \bar{\mathbf{y}}_N) \leq \sum_{k=1}^N \theta_k (\langle F(\mathbf{z}_k), \mathbf{z}_k - \mathbf{z} \rangle + h(\mathbf{z}_k) - h(\mathbf{z}))$, where $\bar{\mathbf{x}}_N = \sum_{k=1}^N \theta_k \mathbf{x}_k$ and $\bar{\mathbf{y}}_N = \sum_{k=1}^N \theta_k \mathbf{y}_k$.*

Downloaded 10/22/25 to 128.62.19.54 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 10/22/25 to 128.62.19.54 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Remark 2.7. For brevity, we report our convergence results in the form of an upper bound on $\ell(\bar{\mathbf{x}}_N, \mathbf{y}) - \ell(\mathbf{x}, \bar{\mathbf{y}}_N)$ for *any* $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$ (e.g., see Proposition 3.1(a)). Then taking the supremum of both sides over $\mathbf{x} \in \mathcal{B}_1$ and $\mathbf{y} \in \mathcal{B}_2$ results in an upper bound on the restricted primal-dual gap in (2.4).

In the strongly-convex-strongly-concave case, problem (1.1) has a unique saddle point $\mathbf{z}^* \in \mathcal{Z}$, and we measure the suboptimality by $\|\mathbf{z} - \mathbf{z}^*\|^2$. Next, we state a known result that will be used later, which follows immediately from Assumption 2.4.

LEMMA 2.8. *If Assumption* 2.4 *holds with* $\mu \geq 0$ *and* $\mathbf{z}^*$ *is a saddle point of problem* (1.1)*, then for any* $\mathbf{z} \in \mathcal{Z}$*, we have* $\langle F(\mathbf{z}), \mathbf{z} - \mathbf{z}^* \rangle + h(\mathbf{z}) - h(\mathbf{z}^*) \geq \mu \|\mathbf{z} - \mathbf{z}^*\|^2$.

Finally, we introduce an additional metric to define the inexactness in solving our subproblem, referred to as the tangent residual [15, 8], which we define at $\mathbf{z}$ as

$$\operatorname{res}(\mathbf{z}) := \min_{\mathbf{w} \in H(\mathbf{z})} \|F(\mathbf{z}) + \mathbf{w}\|. \tag{2.5}$$

Note that $\operatorname{res}(\mathbf{z}) \leq \rho$ if and only if there exists $\mathbf{r} \in \mathbb{R}^d$ such that $\mathbf{r} \in F(\mathbf{z}) + H(\mathbf{z})$ and $\|\mathbf{r}\| \leq \rho$. In particular, $\mathbf{z} = (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^d$ is a saddle point of problem (1.1) if and only if $\operatorname{res}(\mathbf{z}) = 0$. In the unconstrained case, $\operatorname{res}(\mathbf{z}) = \|F(\mathbf{z})\|$ captures the residual at $\mathbf{z}$.

3. Proximal point method and its approximation. As mentioned in the introduction, the generalized optimistic method we propose in this paper can be interpreted as a systematic approach to approximating the classical proximal point method (PPM). Hence, to better motivate our method, we first review the basics of the PPM [35, 51]. In the kth iteration of the PPM for solving problem (1.1), the new point $\mathbf{z}_{k+1}$ is given by the unique solution of the monotone inclusion subproblem

$$0 \in \eta_k F(\mathbf{z}) + \eta_k H(\mathbf{z}) + \mathbf{z} - \mathbf{z}_k, \tag{3.1}$$

where $\eta_k > 0$ is the "step size". Equivalently, we can also write

$$\mathbf{z}_{k+1} = \arg\min_{\mathbf{z} \in \mathcal{Z}} \left\{ \eta_k \langle F(\mathbf{z}_{k+1}), \mathbf{z} - \mathbf{z}_k \rangle + \eta_k h(\mathbf{z}) + \frac{1}{2} \|\mathbf{z} - \mathbf{z}_k\|^2 \right\}. \tag{3.2}$$

Note, however, that (3.2) is not an explicit update. Regarding its convergence property, we have the following results. The proofs are presented in Appendix A.

PROPOSITION 3.1. *Let* $\{\mathbf{z}_k\}_{k \geq 0}$ *be the iterates generated by the PPM in* (3.1).

(a) *Under Assumption* 2.3*, we have* $\|\mathbf{z}_{k+1} - \mathbf{z}^*\| \leq \|\mathbf{z}_k - \mathbf{z}^*\|$ *for any* $k \geq 0$. *Moreover, we have* $\ell(\bar{\mathbf{x}}_N, \mathbf{y}) - \ell(\mathbf{x}, \bar{\mathbf{y}}_N) \leq \frac{\|\mathbf{z}_0 - \mathbf{z}\|^2}{2} \left(\sum_{k=0}^{N-1} \eta_k\right)^{-1}$ *for any* $\mathbf{z} = (\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$*, where* $\bar{\mathbf{z}}_N = (\bar{\mathbf{x}}_N, \bar{\mathbf{y}}_N)$ *is given by* $\bar{\mathbf{z}}_N = \frac{1}{\sum_{k=0}^{N-1} \eta_k} \sum_{k=0}^{N-1} \eta_k \mathbf{z}_{k+1}$.

(b) *Under Assumption* 2.4*, we have* $\|\mathbf{z}_N - \mathbf{z}^*\|^2 \leq \|\mathbf{z}_0 - \mathbf{z}^*\|^2 \prod_{k=0}^{N-1} (1 + 2\eta_k \mu)^{-1}$.

Proposition 3.1 characterizes the convergence rates of the PPM in terms of $\{\eta_k\}_{k \geq 0}$, without any condition on the step sizes η_k. Hence, in theory, the PPM can converge arbitrarily fast with sufficiently large step sizes. On the other hand, each step of the PPM involves solving a highly nontrivial monotone inclusion problem in (3.1) and can be computationally intractable. Hence, our goal is to introduce a general class of optimistic methods that approximates the PPM by replacing (3.1) with more tractable subproblems. As we shall see, our generalized optimistic method in Algorithm 4.1 can achieve the same convergence guarantees (up to constant) as in Proposition 3.1, but unlike the PPM it introduces additional restrictions on the allowable step sizes.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 10/22/25 to 128.62.19.54 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

To better illustrate the idea behind our generalized optimistic method, we first discuss how the PPM can be approximated by first-order methods with explicit and efficient update rules, as shown in [37, 38]. In particular, we focus on the optimistic gradient method, which can be dated back to Popov's method for constrained saddle point problems [48]. The key insight of Popov's method is that when F is Lipschitz continuous, we can approximately solve (3.2) in two steps per iteration, with a properly chosen step size.

Specifically, the update rule consists of these two steps:

$$\begin{aligned} \mathbf{z}_k &= \arg\min_{\mathbf{z}\in\mathcal{Z}} \left\{ \eta_k \langle F(\mathbf{z}_{k-1}), \mathbf{z} - \mathbf{z}_{k-\frac{1}{2}} \rangle + \frac{1}{2}\|\mathbf{z} - \mathbf{z}_{k-\frac{1}{2}}\|^2 \right\}, \\ \mathbf{z}_{k+\frac{1}{2}} &= \arg\min_{\mathbf{z}\in\mathcal{Z}} \left\{ \eta_k \langle F(\mathbf{z}_k), \mathbf{z} - \mathbf{z}_{k-\frac{1}{2}} \rangle + \frac{1}{2}\|\mathbf{z} - \mathbf{z}_{k-\frac{1}{2}}\|^2 \right\}. \end{aligned} \tag{3.3}$$

In other words, we first take a proximal gradient step to compute $\mathbf{z}_k$. Then, the second step can be viewed as an approximation of the PPM, where we use $F(\mathbf{z}_k)$ as a surrogate for $F(\mathbf{z}_{k+\frac{1}{2}})$. Also note that in the first step, we reuse the gradient $F(\mathbf{z}_{k-1})$ from the previous iteration instead of computing a new gradient $F(\mathbf{z}_{k-\frac{1}{2}})$, and thus each iteration of this algorithm requires one gradient computation.

The discussions above view the "midpoints" $\{\mathbf{z}_{k+\frac{1}{2}}\}_{k\geq 0}$ as the approximate iterates of the PPM. Alternatively, we can interpret $\{\mathbf{z}_k\}_{k\geq 0}$ as the approximate iterates in the special setting of unconstrained saddle point problems. Such a viewpoint was first discussed in [37]. In this case, the PPM in (3.1) can be written as

$$\mathbf{z}_{k+1} = \mathbf{z}_k - \eta_k F(\mathbf{z}_{k+1}), \tag{3.4}$$

while (3.3) leads to the simple update rule

$$\mathbf{z}_{k+1} = \mathbf{z}_k - 2\eta_k F(\mathbf{z}_k) + \eta_k F(\mathbf{z}_{k-1}) = \mathbf{z}_k - \eta_k F(\mathbf{z}_k) - \eta_k (F(\mathbf{z}_k) - F(\mathbf{z}_{k-1})), \tag{3.5}$$

also known as optimistic gradient descent-ascent (OGDA) [37, 38]. To better illustrate its main idea, in (3.5) we write the OGDA update as a combination of two terms: a "prediction term" $F(\mathbf{z}_k)$ that serves as a surrogate for $F(\mathbf{z}_{k+1})$, and a "correction term" $F(\mathbf{z}_k) - F(\mathbf{z}_{k-1})$ given by the deviation between $F(\mathbf{z}_k)$ and its prediction in the previous iteration. If we make the optimistic assumption that the prediction error between two consecutive iterations does not change significantly, i.e., $F(\mathbf{z}_{k+1}) - F(\mathbf{z}_k) \approx F(\mathbf{z}_k) - F(\mathbf{z}_{k-1})$, then it holds that $F(\mathbf{z}_{k+1}) \approx 2F(\mathbf{z}_k) - F(\mathbf{z}_{k-1})$ and we can expect (3.5) to be a good approximation for (3.4).

Extragradient and optimistic gradient methods use first-order information to approximate the PPM, prompting the question of whether higher-order information could improve accuracy and speed. While prior works [39, 40, 6] explored this for the extragradient method, no such efforts exist for the optimistic gradient method. Additionally, extending the convergence analysis of the optimistic gradient method to general composite saddle point problems remains unresolved. In this paper, we address these gaps with a general framework for optimistic methods.

4. Main algorithm: The generalized optimistic method. In this section, we present the generalized optimistic method, which extends the "optimistic" idea for approximating the PPM. Our proposed scheme can handle the composite structure in (1.1) and can utilize second-order information. To better explain the idea behind our generalized optimistic method, let us first focus on the unconstrained smooth saddle point problem. In the OGDA method defined in (3.5),

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Algorithm 4.1 Generalized optimistic method for monotone inclusion problems.

Input: $\mathbf{z}_0 \in \mathcal{Z}$, approximation function P, and $0 < \alpha \leq 1$
Initialize: set $\mathbf{z}_{-1} \leftarrow \mathbf{z}_0$, $P(\mathbf{z}_0; \mathcal{I}_{-1}) \leftarrow F(\mathbf{z}_0)$, and $\mathbf{r}_0 \leftarrow 0$
for iteration $k = 0, \ldots, N-1$ **do**
    Choose $\eta_k, \hat{\eta}_k > 0$ and compute $\mathbf{z}_{k+1}$ and $\mathbf{r}_{k+1}$ such that

$$\mathbf{r}_{k+1} \in \eta_k P(\mathbf{z}_{k+1}; \mathcal{I}_k) + \hat{\eta}_k \left(F(\mathbf{z}_k) - P(\mathbf{z}_k; \mathcal{I}_{k-1}) + \frac{\mathbf{r}_k}{\eta_{k-1}} \right) + \eta_k H(\mathbf{z}_{k+1}) + \mathbf{z}_{k+1} - \mathbf{z}_k,$$

$$\|\mathbf{r}_{k+1}\| \leq \frac{\alpha}{4} \|\mathbf{z}_{k+1} - \mathbf{z}_k\|, \quad \eta_k \|F(\mathbf{z}_{k+1}) - P(\mathbf{z}_{k+1}; \mathcal{I}_k)\| \leq \frac{\alpha}{4} \|\mathbf{z}_{k+1} - \mathbf{z}_k\|.$$

end for

the term $F(\mathbf{z}_{k+1})$ is approximated by a linear combination of the prediction term $F(\mathbf{z}_k)$ and the correction term $F(\mathbf{z}_k) - F(\mathbf{z}_{k-1})$. Now, instead of using $F(\mathbf{z}_k)$ as the prediction for $F(\mathbf{z}_{k+1})$, suppose that we have access to a general approximation function denoted by $P(\mathbf{z}; \mathcal{I}_k)$, which is constructed from the available information $\mathcal{I}_k$ up to the kth iteration. For instance, with a second-order oracle we have $\mathcal{I}_k = \{F(\mathbf{z}_0), DF(\mathbf{z}_0), \ldots, F(\mathbf{z}_k), DF(\mathbf{z}_k)\}$, and a possible candidate for P is the first-order Taylor expansion $T^{(1)}(\mathbf{z}; \mathbf{z}_k) = F(\mathbf{z}_k) + DF(\mathbf{z}_k)(\mathbf{z} - \mathbf{z}_k)$. Similar to the logic in OGDA, the correction term is then given by the difference between $F(\mathbf{z}_k)$ and our prediction in the previous step, i.e., $F(\mathbf{z}_k) - P(\mathbf{z}_k; \mathcal{I}_{k-1})$. Indeed, if we are "optimistic" that the change in the prediction error between consecutive iterations is small, i.e., $F(\mathbf{z}_{k+1}) - P(\mathbf{z}_{k+1}; \mathcal{I}_k) \approx F(\mathbf{z}_k) - P(\mathbf{z}_k; \mathcal{I}_{k-1})$, then this implies $F(\mathbf{z}_{k+1}) \approx P(\mathbf{z}_{k+1}; \mathcal{I}_k) + (F(\mathbf{z}_k) - P(\mathbf{z}_k; \mathcal{I}_{k-1}))$ and adding the correction term can help reducing the approximation error. Considering these generalizations, we propose to compute the new iterate as

$$\mathbf{z}_{k+1} = \mathbf{z}_k - \left(\eta_k P(\mathbf{z}_{k+1}; \mathcal{I}_k) + \hat{\eta}_k (F(\mathbf{z}_k) - P(\mathbf{z}_k; \mathcal{I}_{k-1})) \right).$$

Unlike the OGDA method, this is not an explicit update in general and requires us to solve the equation with respect to $\mathbf{z}_{k+1}$. Also, we use different step size parameters for the approximation term and correction term to make our algorithm as general as possible. In fact, our analysis suggests that the best choice of $\hat{\eta}_k$ is not always equal to η_k. As we shall see, we will set $\hat{\eta}_k = \eta_{k-1}$ in the convex-concave setting while setting $\hat{\eta}_k = \eta_{k-1}/(1 + 2\eta_{k-1}\mu)$ in the strongly-convex-strongly-concave setting.

The same methodology can be readily generalized to composite constrained saddle point problems and to accommodate inexact subproblem solutions. Specifically, in the most general form, the new point $\mathbf{z}_{k+1}$ is obtained by solving the inclusion problem

$$0 \in \eta_k P(\mathbf{z}; \mathcal{I}_k) + \hat{\eta}_k (F(\mathbf{z}_k) - P(\mathbf{z}_k; \mathcal{I}_{k-1})) + \eta_k H(\mathbf{z}) + \mathbf{z} - \mathbf{z}_k. \tag{4.1}$$

Comparing the update rule (4.1) with the one for the PPM in (3.1), we can see that the only modification is replacing the term $\eta_k F(\mathbf{z}_{k+1})$ with its optimistic approximation $\eta_k P(\mathbf{z}_{k+1}; \mathcal{I}_k) + \hat{\eta}_k (F(\mathbf{z}_k) - P(\mathbf{z}_k; \mathcal{I}_{k-1}))$. Moreover, it suffices to find an approximate solution of (4.1) in the following sense: we seek $\mathbf{z}_{k+1}$ and $\mathbf{r}_{k+1}$ such that

Downloaded 10/22/25 to 128.62.19.54 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

(4.2)
$$\mathbf{r}_{k+1} \in \eta_k P(\mathbf{z}_{k+1};\mathcal{I}_k) + \hat{\eta}_k \left(F(\mathbf{z}_k) - P(\mathbf{z}_k;\mathcal{I}_{k-1}) + \frac{\mathbf{r}_k}{\eta_{k-1}} \right) + \eta_k H(\mathbf{z}_{k+1}) + \mathbf{z}_{k+1} - \mathbf{z}_k,$$

(4.3)
$$\|\mathbf{r}_{k+1}\| \leq \frac{\alpha}{4} \|\mathbf{z}_{k+1} - \mathbf{z}_k\|,$$

where $\alpha \in (0,1]$ is a user-specified constant. Note that the correction term in (4.2) is modified to incorporate the subproblem residual $\mathbf{r}_k$ from the previous iteration. If the residual vectors $\mathbf{r}_k$ and $\mathbf{r}_{k+1}$ are zero, then $\mathbf{z}_{k+1}$ is the exact solution of (4.1). Finally, to ensure $P(\mathbf{z}_{k+1};\mathcal{I}_k)$ remains a valid approximation of $F(\mathbf{z}_{k+1})$, we impose the condition

$$\eta_k \|F(\mathbf{z}_{k+1}) - P(\mathbf{z}_{k+1};\mathcal{I}_k)\| \leq \frac{\alpha}{4} \|\mathbf{z}_{k+1} - \mathbf{z}_k\|. \tag{4.4}$$

Our generalized optimistic method serves as a framework rather than a directly applicable algorithm. The selection of η_k and $\mathbf{z}_{k+1}$ to meet (4.4) depends on F and the choice of P. A generic line search scheme is detailed in section 4.1, with section 5 focusing on implementing our second-order optimistic method.

The following proposition underpins the convergence results in this paper, showing that Algorithm 4.1 offers similar guarantees to the PPM under specified conditions, even when solving an approximate version of (3.1). The proof is in Appendix B.1.

PROPOSITION 4.1. *Let $\{\mathbf{z}_k\}_{k\geq 0}$ be the iterates generated by Algorithm* 4.1.

(a) *Under Assumption* 2.3*, if we set $\hat{\eta}_k = \eta_{k-1}$, we have $\|\mathbf{z}_k - \mathbf{z}^*\|^2 \leq \frac{2}{2-\alpha}\|\mathbf{z}_0 - \mathbf{z}^*\|^2$ for any $k \geq 0$. Moreover, if we define $\bar{\mathbf{z}}_N = (\bar{\mathbf{x}}_N, \bar{\mathbf{y}}_N)$ as $\bar{\mathbf{z}}_N = \frac{1}{\sum_{k=0}^{N-1}\eta_k} \sum_{k=0}^{N-1} \eta_k \mathbf{z}_{k+1}$, then for any $\mathbf{z} = (\mathbf{x},\mathbf{y}) \in \mathcal{X}\times\mathcal{Y}$, we have $\ell(\bar{\mathbf{x}}_N,\mathbf{y}) - \ell(\mathbf{x},\bar{\mathbf{y}}_N) \leq \frac{1}{2}\|\mathbf{z} - \mathbf{z}_0\|^2 \big(\sum_{k=0}^{N-1}\eta_k\big)^{-1}$.*

(b) *Under Assumption* 2.4*, if we set $\hat{\eta}_k = \eta_{k-1}/(1+2\eta_{k-1}\mu)$, we have $\|\mathbf{z}_N - \mathbf{z}^*\|^2 \leq \frac{2}{2-\alpha}\|\mathbf{z}_0 - \mathbf{z}^*\|^2 \prod_{k=0}^{N-1}(1+2\eta_k\mu)^{-1}$.*

As in Proposition 3.1 for the PPM, the convergence rate in Proposition 4.1 for the generalized optimistic method depends on the choice of $\{\eta_k\}_{k\geq 0}$. On the other hand, the step sizes here are not arbitrary but constrained by (4.4). Depending on our choice of the approximation function P, the condition on η_k varies, and, as a result, we obtain different rates. In fact, the step size η_k is closely related to the displacement $\|\mathbf{z}_{k+1} - \mathbf{z}_k\|$ and more specifically it is approximately on the order of $\frac{1}{\|\mathbf{z}_{k+1}-\mathbf{z}_k\|}$ in our second-order method. Thus, next we provide an upper bound on a (weighted) sum of $\|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2$, which will later translate into bounds on $\sum_{k=0}^{N-1}\eta_k$ and $\prod_{k=0}^{N-1}(1+2\eta_k\mu)$.

LEMMA 4.2. *Let $\{\mathbf{z}_k\}_{k\geq 0}$ be generated by Algorithm* 4.1 *with $\hat{\eta}_k$ chosen as in Proposition* 4.1*. Then, $\sum_{k=0}^{N-1}\|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2 \leq \frac{1}{1-\alpha}\|\mathbf{z}_0 - \mathbf{z}^*\|^2$ under Assumption* 2.3 *and $\sum_{k=0}^{N-1}(\|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2 \prod_{l=0}^{k-1}(1+2\eta_l\mu)) \leq \frac{1}{1-\alpha}\|\mathbf{z}_0 - \mathbf{z}^*\|^2$ under Assumption* 2.4.

Proof. See Appendix B.2 for the proof. ∎

Finally, Lemma 4.2 also allows us to characterize the convergence rate of the generalized optimistic method in terms of the tangent residual as defined in (2.5). Specifically, if we define $\mathcal{E}_k := F(\mathbf{z}_k) - P(\mathbf{z}_k;\mathcal{I}_{k-1}) + \frac{\mathbf{r}_k}{\eta_{k-1}}$, then by rearranging the update rule in (4.2), it can be written as $\frac{\mathbf{z}_k - \mathbf{z}_{k+1}}{\eta_k} + \mathcal{E}_{k+1} - \frac{\hat{\eta}_k\mathcal{E}_k}{\eta_k} \in F(\mathbf{z}_{k+1}) + H(\mathbf{z}_{k+1})$, which implies $\mathrm{res}(\mathbf{z}_{k+1}) \leq \|\frac{\mathbf{z}_k - \mathbf{z}_{k+1}}{\eta_k} + \mathcal{E}_{k+1} - \frac{\hat{\eta}_k\mathcal{E}_k}{\eta_k}\| \leq \frac{\|\mathbf{z}_k - \mathbf{z}_{k+1}\|}{\eta_k} + \|\mathcal{E}_{k+1}\| + \frac{\hat{\eta}_k\|\mathcal{E}_k\|}{\eta_k}$.

Downloaded 10/22/25 to 128.62.19.54 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 10/22/25 to 128.62.19.54 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Moreover, using (4.3) and (4.4), we can relate $\mathcal{E}_{k+1}$ and $\mathcal{E}_k$ to $\|\mathbf{z}_{k+1}-\mathbf{z}_k\|$ and $\|\mathbf{z}_k-\mathbf{z}_{k-1}\|$, which, in turn, will be controlled by Lemma 4.2. The above argument is formalized in the next proposition.

PROPOSITION 4.3. *Let* $\{\mathbf{z}_k\}_{k\geq 0}$ *be generated by Algorithm* 4.1 *with* $\hat{\eta}_k$ *chosen as in Proposition* 4.1.

(a) *Under Assumption* 2.3*, we have* $\min_{i\in\{1,\dots,N\}} \mathrm{res}(\mathbf{z}_i) \leq \frac{(\alpha+1)\|\mathbf{z}_0-\mathbf{z}^*\|}{\sqrt{1-\alpha}\sqrt{\sum_{k=0}^{N-1}\eta_k^2}}$.

(b) *Under Assumption* 2.4*, we have* $\mathrm{res}(\mathbf{z}_N) \leq \frac{\frac{1+\alpha}{\sqrt{1-\alpha}}\|\mathbf{z}_0-\mathbf{z}^*\|}{\eta_{N-1}\prod_{k=0}^{N-2}\sqrt{1+2\eta_k\mu}}$.

Proof. See Appendix B.3 for the proof. ∎

Proposition 4.3 characterizes the convergence rate of the tangent residual in terms of the step sizes $\{\eta_k\}_{k\geq 0}$. In section 5, we will translate this result into an explicit convergence rate for our second-order optimistic method.

4.1. Backtracking line search. As mentioned earlier, our general scheme in Algorithm 4.1 does not directly lead to a practical algorithm, as it requires a proper policy to ensure that (4.4) holds. As the first step towards practical implementations, in this section, we propose a backtracking line search scheme to select the step size η_k adaptively in Algorithm 4.1. To simplify the notation, let $\mathbf{z}^- = \mathbf{z}_k$, $\eta = \eta_k$, $P(\mathbf{z}) = P(\mathbf{z};\mathcal{I}_k)$, and $\mathbf{v}^- = \mathbf{v}_k := \hat{\eta}_k(F(\mathbf{z}_k) - P(\mathbf{z}_k;\mathcal{I}_{k-1}) + \frac{1}{\eta_{k-1}}\mathbf{r}_k)$. In light of (4.2), (4.3), and (4.4), at the kth iteration our goal is to find a step size $\eta > 0$, a point $\mathbf{z}\in\mathcal{Z}$, and the residual vector $\mathbf{r}\in\mathbb{R}^d$ such that

$$\mathbf{r} \in \eta P(\mathbf{z}) + \mathbf{v}^- + \eta H(\mathbf{z}) + \mathbf{z} - \mathbf{z}^-, \quad \|\mathbf{r}\| \leq \frac{\alpha}{4}\|\mathbf{z}-\mathbf{z}^-\|, \tag{4.5}$$

$$\eta\|F(\mathbf{z}) - P(\mathbf{z})\| \leq \frac{\alpha}{4}\|\mathbf{z}-\mathbf{z}^-\|. \tag{4.6}$$

Note that the conditions in (4.5) and (4.6) depend on both η and $\mathbf{z}$. Hence, in general, we need to choose η and $\mathbf{z}$ simultaneously. We assume that the approximation function P is maximal monotone and continuous on $\mathcal{Z}$. Also, we assume access to a black-box solver of the subproblem in (4.5), which we formally define as follows for later reference.

DEFINITION 4.4 (optimistic subsolver). *For any given* η, $\mathbf{v}^-$ *and* $\mathbf{z}^-$, *the* optimistic subsolver *returns* $\mathbf{z}\in\mathcal{Z}$ *and* $\mathbf{r}\in\mathbb{R}^d$ *that satisfies* (4.5).

Now we are ready to describe our line search scheme, which takes an initial trial step size σ and a parameter $\beta\in(0,1)$ as inputs. At a high level, it repeatedly solves (4.5) for different step sizes in $\{\sigma\beta^i : i\geq 0\}$ and then checks the condition in (4.6), each of which involves one call to the optimistic subsolver. Specifically, we first assign the value of σ to the step size η and compute $\mathbf{z}$ and $\mathbf{r}$ as the solution of (4.5). If the condition in (4.6) is satisfied with respect to η and $\mathbf{z}$, we accept η as the final step size and choose $\mathbf{z}$ as the next iterate. Otherwise, we decrease the step size by $\eta\leftarrow\beta\eta$ and recompute $\mathbf{z}$ and $\mathbf{r}$ from (4.5), and this backtracking process is repeated until $(\eta,\mathbf{z})$ satisfies (4.6). The whole procedure is formally described in Algorithm 4.2.

Under suitable conditions, the following lemma ensures that the line search scheme will terminate properly and find an admissible step size.

LEMMA 4.5. *Assume there exist* $\delta, L>0$ *such that* $\|F(\mathbf{z})-P(\mathbf{z})\|\leq L\|\mathbf{z}-\mathbf{z}^-\|$ *for any* $\mathbf{z}$ *satisfying* $\|\mathbf{z}-\mathbf{z}^-\|\leq\delta$. *Then Algorithm* 4.2 *terminates in finite steps.*

Proof. Our proof is adapted from [34, Lemma 3.2]. Given $\eta\in[0,\sigma]$, let $\mathbf{z}$ and $\mathbf{r}$ be any pair of solutions that satisfies the condition in (4.5). First, we show that

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Algorithm 4.2 Backtracking line search scheme.

Input: current iterate $\mathbf{z}^- \in \mathcal{Z}$, $\mathbf{v}^- \in \mathbb{R}^d$, initial trial step size $\sigma > 0$, approximation function P, $\alpha \in (0,1]$, $\beta \in (0,1)$
Output: step size η, the next iterate $\mathbf{z} \in \mathcal{Z}$, and the residual vector $\mathbf{r}$
Set $\eta \leftarrow \sigma$ and let $(\mathbf{z}, \mathbf{r})$ be the outputs of optimistic subsolver for $0 \in \eta P(\mathbf{z}) + \mathbf{v}^- + \eta H(\mathbf{z}) + \mathbf{z} - \mathbf{z}^-$
while $\eta \|F(\mathbf{z}) - P(\mathbf{z})\| > \frac{\alpha}{4}\|\mathbf{z} - \mathbf{z}^-\|$ **do**
    Set $\eta \leftarrow \eta\beta$ and let $(\mathbf{z}, \mathbf{r})$ be the outputs of optimistic subsolver for $0 \in \eta P(\mathbf{z}) + \mathbf{v}^- + \eta H(\mathbf{z}) + \mathbf{z} - \mathbf{z}^-$
end while
Return η, $\mathbf{z}$, and $\mathbf{r}$

$\mathbf{z}$ lies in a compact set $\mathcal{C}_\sigma \subset \mathcal{Z}$. By (4.5), there exists $\mathbf{w} \in H(\mathbf{z})$ such that $\mathbf{r} = \eta P(\mathbf{z}) + \mathbf{v}^- + \eta\mathbf{w} + \mathbf{z} - \mathbf{z}^-$, which implies that

$$\|\mathbf{z} - \mathbf{z}^-\|^2 = \langle \mathbf{r} - \mathbf{v}^-, \mathbf{z} - \mathbf{z}^- \rangle - \eta \langle P(\mathbf{z}) + \mathbf{w}, \mathbf{z} - \mathbf{z}^- \rangle. \tag{4.7}$$

Since $\mathbf{z}^-$ is feasible, we can find $\mathbf{w}^- \in \mathbb{R}^d$ such that $\mathbf{w}^- \in H(\mathbf{z}^-)$. By the monotonicity of P and H, it holds that $\langle P(\mathbf{z}) + \mathbf{w} - (P(\mathbf{z}^-) + \mathbf{w}^-), \mathbf{z} - \mathbf{z}^- \rangle \geq 0$ and hence $-\eta\langle P(\mathbf{z}) + \mathbf{w}, \mathbf{z} - \mathbf{z}^- \rangle \leq -\eta\langle P(\mathbf{z}^-) + \mathbf{w}^-, \mathbf{z} - \mathbf{z}^- \rangle$. Moreover, we have $\langle \mathbf{r}, \mathbf{z} - \mathbf{z}^- \rangle \leq \|\mathbf{r}\|\|\mathbf{z} - \mathbf{z}^-\| \leq \frac{\alpha}{4}\|\mathbf{z} - \mathbf{z}^-\|^2$, where we used the error condition in (4.5). Hence, from (4.7) we further have $(1 - \frac{\alpha}{4})\|\mathbf{z} - \mathbf{z}^-\|^2 \leq -\langle \mathbf{v}^-, \mathbf{z} - \mathbf{z}^- \rangle - \eta\langle P(\mathbf{z}^-) + \mathbf{w}^-, \mathbf{z} - \mathbf{z}^- \rangle$. Together with Cauchy–Schwarz inequality and the fact that $\eta \leq \sigma$, this leads to $(1 - \frac{\alpha}{4})\|\mathbf{z} - \mathbf{z}^-\| \leq \|\mathbf{v}^-\| + \eta\|P(\mathbf{z}^-) + \mathbf{w}^-\| \leq \|\mathbf{v}^-\| + \sigma\|P(\mathbf{z}^-) + \mathbf{w}^-\|$. As the upper bound in this inequality is independent of η, this proves that $\mathbf{z}$ belongs to the compact set $\mathcal{C}_\sigma := \{\mathbf{z} \in \mathbb{R}^d : \|\mathbf{z} - \mathbf{z}^-\| \leq \frac{4}{4-\alpha}(\|\mathbf{v}^-\| + \sigma\|P(\mathbf{z}^-) + \mathbf{w}^-\|)\} \cap \mathcal{Z}$ for all $0 \leq \eta \leq \sigma$.

Now we prove the lemma by contradiction. Denote the trial step size after i steps in Algorithm 4.2 by $\eta^{(i)} := \sigma\beta^i$ and the corresponding approximate solution of (4.5) by $\mathbf{z}^{(i)}$. Since $\eta^{(i)} \leq \sigma$, it follows from the arguments above that $\mathbf{z}^{(i)} \in \mathcal{C}_\sigma$ for all $i \geq 0$. If Algorithm 4.2 does not terminate in finite steps, then for any integer $i \geq 0$, the trial step size $\eta^{(i)}$ and the solution $\mathbf{z}^{(i)}$ do not satisfy condition (4.6), i.e.,

$$\eta^{(i)}\|F(\mathbf{z}^{(i)}) - P(\mathbf{z}^{(i)})\| > \frac{1}{2}\alpha\|\mathbf{z}^{(i)} - \mathbf{z}^-\| \quad \forall i \geq 0. \tag{4.8}$$

Since both F and P are continuous on $\mathcal{Z}$ and $\mathbf{z}^{(i)}$ remains in a compact set $\mathcal{C}_\sigma$ for all $i \geq 0$, we can upper bound $\|F(\mathbf{z}^{(i)}) - P(\mathbf{z}^{(i)})\|$ by a constant independent of i. Thus, we obtain $\mathbf{z}^{(i)} \to \mathbf{z}^-$ by taking i to infinity in (4.8). Hence, for i large enough, we have $\|\mathbf{z}^{(i)} - \mathbf{z}^-\| \leq \delta$, which further implies that $\|F(\mathbf{z}^{(i)}) - P(\mathbf{z}^{(i)})\| \leq L\|\mathbf{z}^{(i)} - \mathbf{z}^-\|$ by our assumption. Combining this with (4.8), we conclude that $\eta^{(i)}L > \frac{1}{2}\alpha$ for any large enough i. Since $\eta^{(i)} \to 0$ as $i \to \infty$, this leads to a contradiction. ∎

While Lemma 4.5 guarantees Algorithm 4.2 terminates in finite steps, we need an explicit upper bound on subsolver calls to characterize the line search scheme's complexity. The following proposition provides this bound in terms of the returned step size η and the initial trial step size σ.

PROPOSITION 4.6. *Let η be the step size returned by Algorithm* 4.2. *Then the procedure takes* $\log_{\frac{1}{\beta}} \frac{\sigma}{\eta} + 1$ *calls to the optimistic subsolver in total.*

Proof. The line search scheme iteratively tests the step size from the set $\{\sigma\beta^i : i \geq 0\}$ in a decreasing order, each of which corresponds to one call to the optimistic

Downloaded 10/22/25 to 128.62.19.54 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

subsolver. Hence, if N_{ls} denotes the total number of subsolver calls in Algorithm 4.2, the returned step size satisfies $\eta = \sigma\beta^{N_{\text{ls}}-1}$, which implies $N_{\text{ls}} = \log_{\frac{1}{\beta}} \frac{\sigma}{\eta} + 1$. ∎

This result does not provide a complexity bound for our line search scheme, as it depends on the value of the returned step size η and our choice of the initial trial step size σ. In the following section, we establish lower bounds on η and use them to derive an explicit upper bound on the total number of calls to the optimistic subsolver.

Remark 4.7. Our line search scheme is distinct from those in [40, 6]. Specifically, their line search conditions impose both a lower bound and an upper bound on the step size, requiring a more complex two-stage line search procedure and incurring an additional $\log\log\frac{1}{\epsilon}$ factor in the complexity. In comparison, our scheme implicitly ensures a lower bound on the step size through the careful selection of the initial trial step size σ and the backtracking mechanism (see Lemma 5.3). This simplifies the line search condition and eliminates the need for the Hessian Lipschitz constant L_2.

5. Second-order generalized optimistic method. In this section, we instantiate Algorithm 4.1 with a second-order oracle to derive a novel *second-order optimistic* method for solving the saddle point problem in (1.1). We require the following assumption on the smooth component of the objective in (2.8) denoted by f.

Assumption 5.1. The operator F defined in (2.1) is second-order L_2-smooth on $\mathcal{Z}$. Also, we have access to an oracle that returns $F(\mathbf{z})$ and $DF(\mathbf{z})$ for any given $\mathbf{z}$.

Under Assumption 5.1, it follows from our discussions after Definition 2.2 that

$$\|F(\mathbf{z}) - F(\mathbf{z}_k) - DF(\mathbf{z}_k)(\mathbf{z}-\mathbf{z}_k)\| \le \frac{L_2}{2}\|\mathbf{z}-\mathbf{z}_k\|^2 \quad \forall \mathbf{z}\in\mathcal{Z}, \tag{5.1}$$

which suggests we can choose $P(\mathbf{z};\mathcal{I}_k) := T^{(1)}(\mathbf{z};\mathbf{z}_k) = F(\mathbf{z}_k) + DF(\mathbf{z}_k)(\mathbf{z}-\mathbf{z}_k)$. Note that computing this approximation function requires access to the operator $DF(\cdot)$, which involves the second-order derivative of f. Therefore, we refer to the resulting algorithm as the second-order generalized optimistic method. According to (4.2) and (4.3), the update of the proposed second-order optimistic method can be written as

$$\begin{aligned} \mathbf{r}_{k+1} \in{}& \eta_k T^{(1)}(\mathbf{z}_{k+1};\mathbf{z}_k) + \hat{\eta}_k\left(F(\mathbf{z}_k) - T^{(1)}(\mathbf{z}_k;\mathbf{z}_{k-1}) + \frac{\mathbf{r}_k}{\eta_{k-1}}\right) \\ &+ \eta_k H(\mathbf{z}_{k+1}) + \mathbf{z}_{k+1} - \mathbf{z}_k, \\ \|\mathbf{r}_{k+1}\| \le{}& \frac{\alpha}{4}\|\mathbf{z}_{k+1}-\mathbf{z}_k\|, \end{aligned} \tag{5.2}$$

where $\mathbf{r}_{k+1}$ is the residual vector, and the condition in (4.4) on η_k becomes

$$\eta_k\|F(\mathbf{z}_{k+1}) - T^{(1)}(\mathbf{z}_{k+1};\mathbf{z}_k)\| \le \frac{\alpha}{4}\|\mathbf{z}_{k+1}-\mathbf{z}_k\|. \tag{5.3}$$

Next, we discuss how second-order information accelerates convergence. Intuitively, in light of (5.3), choosing a more accurate approximation $T^{(1)}(\cdot;\mathbf{z}_k)$ allows us to take a larger η_k than the optimistic gradient method. Moreover, as Proposition 4.1 shows, the convergence speed depends on $\{\eta_k\}_{k\ge 0}$, and larger step sizes lead to faster convergence. More precisely, by (5.1), the left-hand side of (5.3) can be bounded above by $\frac{1}{2}\eta_k L_2\|\mathbf{z}_{k+1}-\mathbf{z}_k\|^2$, suggesting that we can pick step size η_k as large as

$$\eta_k \approx \frac{\alpha}{2L_2\|\mathbf{z}_{k+1}-\mathbf{z}_k\|}. \tag{5.4}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 10/22/25 to 128.62.19.54 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Algorithm 5.1 Second-order optimistic method.

Input: initial point $\mathbf{z}_0 \in \mathcal{Z}$, initial trial step size σ_0, strong convexity parameter $\mu \geq 0$, line search parameters $\alpha, \beta \in (0,1)$
Initialize: set $\mathbf{z}_{-1} \leftarrow \mathbf{z}_0$, $P(\mathbf{z}_0; \mathcal{I}_{-1}) \leftarrow F(\mathbf{z}_0)$, and $\mathbf{r}_0 \leftarrow 0$
for iteration $k = 0, \ldots, N-1$ **do**
  Set $\hat{\eta}_k = \eta_{k-1}/(1 + 2\eta_{k-1}\mu)$
  Obtain η_k, $\mathbf{z}_{k+1}$, and $\mathbf{r}_{k+1}$ by Algorithm 4.2 with $\sigma = \sigma_k$, $\mathbf{z}^- = \mathbf{z}_k$, $P(\mathbf{z}) = T^{(1)}(\mathbf{z}; \mathbf{z}_k)$, and $\mathbf{v}^- = \hat{\eta}_k(F(\mathbf{z}_k) - T^{(1)}(\mathbf{z}_k; \mathbf{z}_{k-1}) + \frac{1}{\eta_{k-1}}\mathbf{r}_k)$
  Set $\sigma_{k+1} \leftarrow \eta_k\sqrt{1 + 2\eta_k\mu}/\beta$
end for

Hence, as $\{\mathbf{z}_k\}_{k\geq 0}$ approach the optimal solution and the displacement $\|\mathbf{z}_{k+1} - \mathbf{z}_k\|$ becomes smaller, we can afford to select a larger step size and accelerate convergence.

The above argument for faster convergence relies on the premise that η_k is adaptive to the displacement $\|\mathbf{z}_{k+1} - \mathbf{z}_k\|$, i.e., at least on the order of $\frac{1}{\|\mathbf{z}_{k+1} - \mathbf{z}_k\|}$. As mentioned, a major challenge is the interdependence between η_k and $\mathbf{z}_{k+1}$: the iterate $\mathbf{z}_{k+1}$ depends on η_k via (5.2), while the step size η_k depends on $\mathbf{z}_{k+1}$ as illustrated in (5.3) and (5.4). Hence, we need to simultaneously select η_k and $\mathbf{z}_{k+1}$ to satisfy the above requirements. This goal can be achieved by using the line search scheme discussed in section 4.1. In addition, we use a warm-start strategy and choose the initial trial step size σ_{k+1} based on the previously accepted η_k. Specifically, we set $\sigma_{k+1} = \eta_k/\beta$ in the convex-concave setting and set $\sigma_{k+1} = \eta_k\sqrt{1 + 2\eta_k\mu}/\beta$ in the strongly-convex-strongly-concave setting. As shown in section 5.4, our choice of η_{k+1} is crucial to show the average number of line search steps per iteration is bounded by a constant. Our second-order optimistic method is described in Algorithm 5.1.

Two remarks on the line search scheme in Algorithm 5.1 follow. First, by using the result in Lemma 4.5, we can guarantee that the line search scheme always terminates in finite steps, and hence the step sizes $\{\eta_k\}_{k\geq 0}$ are well-defined. To see why Lemma 4.5 applies, note that for any $\|\mathbf{z} - \mathbf{z}_k\| \leq \delta$, by (5.1) we have $\|F(\mathbf{z}) - T^{(1)}(\mathbf{z}; \mathbf{z}_k)\| \leq (L_2/2)\|\mathbf{z} - \mathbf{z}_k\|^2 \leq (L_2\delta/2)\|\mathbf{z} - \mathbf{z}_k\|$. Hence, the condition in Lemma 4.5 is indeed satisfied. Second, at the kth iteration, the optimistic subsolver in this case is required to find $\mathbf{z}$ and $\mathbf{r}$ such that

$$\mathbf{r} \in \eta T^{(1)}(\mathbf{z}; \mathbf{z}_k) + \mathbf{v}_k + \eta H(\mathbf{z}) + \mathbf{z} - \mathbf{z}_k,\ \|\mathbf{r}\| \leq \frac{\alpha}{4}\|\mathbf{z} - \mathbf{z}_k\|, \tag{5.5}$$

for a given vector $\mathbf{v}_k = \hat{\eta}_k(F(\mathbf{z}_k) - T^{(1)}(\mathbf{z}_k; \mathbf{z}_{k-1}) + \frac{\mathbf{r}_k}{\eta_{k-1}})$ and step size $\eta > 0$ (cf. Definition 4.4). Note that compared to our original problem in (2.2), there are two key structures in our subproblem: (i) the operator $T^{(1)}(\mathbf{z}; \mathbf{z}_k)$ in our subproblem is affine, (ii) the operator is strongly monotone due to the proximal term $\mathbf{z} - \mathbf{z}_k$. In the following, we further discuss how one may exploit these structures depending on the functions h_1 and h_2 and the constraint sets $\mathcal{X}$ and $\mathcal{Y}$.

- **Unconstrained setting:** $h_1 = h_2 = 0$, $\mathcal{X} = \mathbb{R}^m$, and $\mathcal{Y} = \mathbb{R}^n$. Then the subproblem reduces to $0 = \eta_k(F(\mathbf{z}_k) + DF(\mathbf{z}_k)(\mathbf{z} - \mathbf{z}_k)) + \mathbf{v}_k + \mathbf{z} - \mathbf{z}_k$, which is a linear system of equations. Hence, one can either choose a direct method such as LU factorization to obtain the exact solution or employ an iterative method such as the CGLS method [18, 47] to obtain an approximate solution.
- **Proximal setting:** The resolvent of H defined in (2.2) is easy to compute. In this case, the subproblem takes the form $0 \in A(\mathbf{z}) + B(\mathbf{z})$, where A is an

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 10/22/25 to 128.62.19.54 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

affine operator and B has a resolvent that is easy to compute. We can employ an operator-splitting method such as the Douglas–Rachford splitting method [14], which requires solving a linear system in each iteration. Alternatively, we can use the recently developed composite fast extragradient (FEG) method in [7] that extends the FEG method in [27], only requiring two matrix-vector products per iteration.

Later in section 5.5, we will further analyze both cases separately and characterize the total computational cost of the optimistic subsolver.

Remark 5.2. All existing second-order and higher-order methods require solving specific subproblems as part of their update rules. The second-order methods proposed in [40, 6, 46] solve subproblems with the same structure as ours. In contrast, the methods in [42, 1, 30, 31, 20, 21] solve subproblems of the form $0 \in \mathbf{H}_k\mathbf{z} + \mathbf{g}_k + M\|\mathbf{z} - \mathbf{z}_k\|(\mathbf{z} - \mathbf{z}_k) + H(\mathbf{z})$, where M is a constant, $\mathbf{H}_k \in \mathbb{R}^{d\times d}$, $\mathbf{g}_k \in \mathbb{R}^d$, and H is a maximal monotone operator. The key difference between their subproblem and ours lies in the nonlinear term $\|\mathbf{z} - \mathbf{z}_k\|(\mathbf{z} - \mathbf{z}_k)$. While specialized solvers are proposed in the unconstrained setting [1, 31, 20], it remains unclear if they can handle constrained settings or a general maximal monotone operator H.

5.1. Intermediate results. To establish the convergence properties of the proposed second-order optimistic method, we first state several intermediate results. Throughout the rest of the analysis, we define $\mathcal{B} = \{k : \eta_k < \sigma_k\}$, i.e., the set of iteration indices where the line search scheme backtracks the step size. To begin with, in the following lemma, we establish a lower bound on η_k when $k \in \mathcal{B}$.

LEMMA 5.3. *Let* $\{\eta_k\}_{k\geq 0}$ *be the step sizes in* (5.2) *generated by Algorithm* 5.1 *and suppose that* F *is second-order* L_2*-smooth on* $\mathcal{Z}$*. Further define* $\mathbf{v}_k := \hat{\eta}_k(F(\mathbf{z}_k) - T^{(1)}(\mathbf{z}_k;\mathbf{z}_{k-1}) + \frac{\mathbf{r}_k}{\eta_{k-1}})$*. If* $k \in \mathcal{B}$*, then* $\eta_k \geq \frac{(4-\alpha)\alpha\beta^2/L_2}{(4+\alpha)\|\mathbf{z}_{k+1}-\mathbf{z}_k\|+((4-\alpha)\beta+4+\alpha)\|\mathbf{v}_k\|}$.

Proof. See Appendix C for the proof. ∎

By combining Lemmas 5.3 and 4.2, we further present the following result.

LEMMA 5.4. *We have* $\sum_{k\in\mathcal{B}} \frac{\prod_{l=0}^{k-1}(1+2\eta_l\mu)}{\eta_k^2} \leq \gamma_2^2 L_2^2\|\mathbf{z}_0 - \mathbf{z}^*\|^2$*, where*

$$\gamma_2 = \frac{1}{\sqrt{1-\alpha}}\left(\frac{4+\alpha}{(4-\alpha)\alpha\beta^2} + \frac{(4-\alpha)\beta+4+\alpha}{2(4-\alpha)\beta^2}\right). \tag{5.6}$$

Proof. We first apply Lemma 5.3 for $k \in \mathcal{B}$. Note that by the choice of $\hat{\eta}_k$, (5.2), and (5.3), we can upper bound $\|\mathbf{v}_k\| = \frac{\eta_{k-1}}{1+2\eta_{k-1}\mu}\|F(\mathbf{z}_k) - T^{(1)}(\mathbf{z}_k;\mathbf{z}_{k-1}) + \frac{1}{\eta_{k-1}}\mathbf{r}_k\| \leq \frac{\eta_{k-1}\|F(\mathbf{z}_k)-T^{(1)}(\mathbf{z}_k;\mathbf{z}_{k-1})\|+\|\mathbf{r}_k\|}{1+2\eta_{k-1}\mu} \leq \frac{\alpha\|\mathbf{z}_k-\mathbf{z}_{k-1}\|}{2(1+2\eta_{k-1}\mu)}$. By applying Lemma 5.3, this leads to $\frac{1}{\eta_k} \leq \frac{L_2}{(4-\alpha)\alpha\beta^2}((4+\alpha)\|\mathbf{z}_{k+1} - \mathbf{z}_k\| + ((4-\alpha)\beta+4+\alpha)\|\mathbf{v}_k\|) \leq c_1L_2\|\mathbf{z}_{k+1} - \mathbf{z}_k\| + c_2L_2\frac{\|\mathbf{z}_k-\mathbf{z}_{k-1}\|}{1+2\eta_{k-1}\mu}$, where we define $c_1 := \frac{4+\alpha}{(4-\alpha)\alpha\beta^2}$ and $c_2 := \frac{(4-\alpha)\beta+4+\alpha}{2(4-\alpha)\beta^2}$. We square both sides of the above inequality and use Young's inequality to get

$$\frac{1}{\eta_k^2} \leq (c_1^2 + c_1c_2)L_2^2\|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2 + (c_2^2 + c_1c_2)L_2^2\frac{\|\mathbf{z}_k - \mathbf{z}_{k-1}\|^2}{(1+2\eta_{k-1}\mu)^2}. \tag{5.7}$$

Multiplying both sides of (5.7) by $\prod_{l=0}^{k-1}(1+2\eta_l\mu)$, we further get $\frac{1}{\eta_k^2}\prod_{l=0}^{k-1}(1+2\eta_l\mu) \leq (c_1^2+c_1c_2)L_2^2\|\mathbf{z}_{k+1}-\mathbf{z}_k\|^2\prod_{l=0}^{k-1}(1+2\eta_l\mu)+(c_2^2+c_1c_2)L_2^2\|\mathbf{z}_k-\mathbf{z}_{k-1}\|^2\prod_{l=0}^{k-2}(1+2\eta_l\mu)$. By summing both sides of the inequality above over $k \in \mathcal{B}$ and applying Lemma 4.2, we get the desired result in Lemma 5.4. ∎

Finally, we build on Lemma 5.4 to control step sizes η_k for all $k \geq 0$.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 10/22/25 to 128.62.19.54 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Downloaded 10/22/25 to 128.62.19.54 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

LEMMA 5.5. *For any* $k \geq 0$, $\frac{1}{\eta_k^2}\prod_{l=0}^{k-1}(1+2\eta_l\mu) \leq \max\{\gamma_2^2 L_2^2\|\mathbf{z}_0-\mathbf{z}^*\|^2, \frac{\beta^{2k}}{\sigma_0^2}\}$.

Proof. We shall prove Lemma 5.5 by induction. For the base case where $k=0$, we consider two subcases. If $0 \notin \mathcal{B}$, then we have $\eta_0 = \sigma_0$, which directly implies the inequality. Otherwise, since $0 \in \mathcal{B}$, we can use Lemma 5.4 to obtain that $\frac{1}{\eta_0^2} \leq \sum_{k\in\mathcal{B}}\left(\frac{1}{\eta_k^2}\prod_{l=0}^{k-1}(1+2\eta_l\mu)\right) \leq \gamma_2^2 L_2^2\|\mathbf{z}_0-\mathbf{z}^*\|^2$. This proves the inequality in Lemma 5.5 for $k=0$.

Now assume that the inequality in Lemma 5.5 holds for $k=s$ where $s \geq 0$. If $s+1 \in \mathcal{B}$, then similarly we have $\frac{1}{\eta_{s+1}^2}\prod_{l=0}^{s}(1+2\eta_l\mu) \leq \sum_{k\in\mathcal{B}}(\frac{1}{\eta_k^2}\prod_{l=0}^{k-1}(1+2\eta_l\mu)) \leq \gamma_2^2 L_2^2\|\mathbf{z}_0-\mathbf{z}^*\|^2$ by Lemma 5.4, which implies the desired inequality. Otherwise, if $s+1 \notin \mathcal{B}$, then by definition we have $\eta_{s+1} = \sigma_{s+1} = \eta_s\sqrt{1+2\eta_s\mu}/\beta$, and furthermore, $\frac{1}{\eta_{s+1}^2}\prod_{l=0}^{s}(1+2\eta_l\mu) = \frac{\beta^2}{\eta_s^2}\prod_{l=0}^{s-1}(1+2\eta_l\mu) \leq \beta^2\max\{\gamma_2^2 L_2^2\|\mathbf{z}_0-\mathbf{z}^*\|^2, \frac{\beta^{2s}}{\sigma_0^2}\}$, where we used the induction hypothesis in the last inequality. In both cases, we prove that the inequality in Lemma 5.5 is satisfied for $k=s+1$, and hence the claim follows by induction. ∎

5.2. Convergence analysis: Convex-concave case. Now we turn to the convergence analysis of Algorithm 5.1 and start with the case where the smooth component f of the objective in (1.1) is merely convex-concave.

LEMMA 5.6. *Let* $\{\eta_k\}_{k\geq 0}$ *be the step sizes in* (5.2) *generated by Algorithm* 5.1. *Then,* $\sum_{k=0}^{N-1}\frac{1}{\eta_k} \leq \frac{1}{1-\beta}(\frac{1}{\sigma_0}+\gamma_2 L_2\|\mathbf{z}^*-\mathbf{z}_0\|\sqrt{N})$, *where* γ_2 *is defined in* (5.6).

Proof. Recall that $\mathcal{B} = \{k : \eta_k < \sigma_k\}$, i.e., the set of iteration indices where the line search backtracks the step size. We divide the sum $\sum_{k=0}^{N-1}\frac{1}{\eta_k}$ into two parts, i.e., $\sum_{k\in\mathcal{B},0\leq k\leq N-1}\frac{1}{\eta_k}$ and $\sum_{k\notin\mathcal{B},0\leq k\leq N-1}\frac{1}{\eta_k}$. Moreover, in Algorithm 4.2, we have $\eta_k = \sigma_k$ when $k \notin \mathcal{B}$ and $\sigma_k = \eta_{k-1}/\beta$ for $k \geq 1$. Hence, for the second part, we have $\sum_{k\notin\mathcal{B},0\leq k\leq N-1}\frac{1}{\eta_k} \leq \frac{1}{\sigma_0}+\sum_{k\notin\mathcal{B},1\leq k\leq N-1}\frac{\beta}{\eta_{k-1}} \leq \frac{1}{\sigma_0}+\sum_{k=0}^{N-1}\frac{\beta}{\eta_k}$. This leads to $\sum_{k=0}^{N-1}\frac{1}{\eta_k} \leq \sum_{k\in\mathcal{B},0\leq k\leq N-1}\frac{1}{\eta_k}+\frac{1}{\sigma_0}+\sum_{k=0}^{N-1}\frac{\beta}{\eta_k}$. Combining the left-hand side with the last term on the right-hand side, we obtain $\sum_{k=0}^{N-1}\frac{1}{\eta_k} \leq \frac{1}{1-\beta}\frac{1}{\sigma_0}+\frac{1}{1-\beta}\sum_{k\in\mathcal{B},0\leq k\leq N-1}\frac{1}{\eta_k} \leq \frac{1}{1-\beta}\frac{1}{\sigma_0}+\frac{1}{1-\beta}\sqrt{N\sum_{k\in\mathcal{B},0\leq k\leq N-1}\frac{1}{\eta_k^2}}$. Finally, Lemma 5.6 follows from the above inequality and Lemma 5.4. ∎

By instantiating Proposition 4.1, we obtain the following convergence result.

THEOREM 5.7. *Suppose Assumptions* 2.3 *and* 5.1 *hold. Let* $\{\mathbf{z}_k\}_{k\geq 0}$ *be the iterates generated by Algorithm* 5.1, *where* $\hat{\eta}_k = \eta_{k-1}$, $\sigma_0 > 0$, $\alpha \in (0,1)$, *and* $\beta \in (0,1)$. *Then* $\|\mathbf{z}_k-\mathbf{z}^*\|^2 \leq \frac{2}{2-\alpha}\|\mathbf{z}_0-\mathbf{z}^*\|^2$ *for* $k \geq 0$. *Moreover, for any* $\mathbf{z} = (\mathbf{x},\mathbf{y}) \in \mathcal{X}\times\mathcal{Y}$, $\ell(\bar{\mathbf{x}}_N,\mathbf{y}) - \ell(\mathbf{x},\bar{\mathbf{y}}_N) \leq \frac{\gamma_2 L_2\|\mathbf{z}_0-\mathbf{z}^*\|\|\mathbf{z}_0-\mathbf{z}\|^2}{2(1-\beta)N^{3/2}} + \frac{\|\mathbf{z}_0-\mathbf{z}\|^2}{2(1-\beta)\sigma_0 N^2}$, *where* $\bar{\mathbf{z}}_N = (\bar{\mathbf{x}}_N,\bar{\mathbf{y}}_N)$ *is given by* $\bar{\mathbf{z}}_N = \frac{1}{\sum_{k=0}^{N-1}\eta_k}\sum_{k=0}^{N-1}\eta_k\mathbf{z}_{k+1}$ *and* γ_2 *is an absolute constant defined in* (5.6).

Proof. By Cauchy–Schwarz inequality, we have $\sum_{k=0}^{N-1}\eta_k\sum_{k=0}^{N-1}\frac{1}{\eta_k} \geq N^2$. Together with Lemma 5.6, this implies that $(\sum_{k=0}^{N-1}\eta_k)^{-1} \leq \frac{1}{N^2}\sum_{k=0}^{N-1}\frac{1}{\eta_k} \leq \frac{1}{(1-\beta)\sigma_0 N^2}+\frac{\gamma_2 L_2\|\mathbf{z}^*-\mathbf{z}_0\|}{(1-\beta)N^{\frac{3}{2}}}$. The rest follows from Proposition 4.1. ∎

Theorem 5.7 shows that the proposed method converges at a rate of $\mathcal{O}(N^{-3/2})$, faster than the rate of $\mathcal{O}(N^{-1})$ for first-order methods. As a corollary, to obtain a solution with a primal-dual gap of ϵ, the proposed method requires $\mathcal{O}(1/\epsilon^{2/3})$ iterations. Note that the total number of subsolver calls in the line search procedure after N iterations can also be explicitly controlled, as we discuss later in Theorem 5.15.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 10/22/25 to 128.62.19.54 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Comparison with [40] and [6]. Similar iteration complexity bounds have been reported in [40, 6] for extragradient-type second-order methods. In comparison, our method is based on a different algorithmic framework. Moreover, as we discuss in section 5.4, our line search scheme requires fewer calls to the subproblem solver in total compared to the approaches in [40, 6].

Similarly, using Proposition 4.3, we can also derive the following convergence result in terms of the tangent residual.

THEOREM 5.8. *Under the same conditions as in Theorem* 5.7*, it also holds that* $\min_{i\in\{1,\dots,N\}} \mathrm{res}(\mathbf{z}_i) \le \frac{(\alpha+1)\|\mathbf{z}_0-\mathbf{z}^*\|}{\sqrt{1-\alpha}(1-\beta)\sigma_0 N^{3/2}} + \frac{(\alpha+1)\gamma_2 L_2\|\mathbf{z}^*-\mathbf{z}_0\|^2}{\sqrt{1-\alpha}(1-\beta)N}$.

Proof. Recall from Lemma 5.6 that $\sum_{k=0}^{N-1}\frac{1}{\eta_k} \le \frac{1}{1-\beta}(\frac{1}{\sigma_0}+\gamma_2 L_2\|\mathbf{z}^*-\mathbf{z}_0\|\sqrt{N})$. By Hölder's inequality, $(\sum_{k=0}^{N-1}\eta_k^2)^{\frac{1}{3}}(\sum_{k=0}^{N-1}\frac{1}{\eta_k})^{\frac{2}{3}} \ge N$, which implies $\frac{1}{\sqrt{\sum_{k=0}^{N-1}\eta_k^2}} \le N^{-\frac{3}{2}}\sum_{k=0}^{N-1}\frac{1}{\eta_k} \le \frac{1}{(1-\beta)\sigma_0 N^{3/2}} + \frac{\gamma_2 L_2\|\mathbf{z}^*-\mathbf{z}_0\|}{(1-\beta)N}$. The rest follows from Proposition 4.3. ∎

Theorem 5.8 shows that, after N iterations, the best iterate among $\{\mathbf{z}_1,\dots,\mathbf{z}_N\}$ converges at a rate of $\mathcal{O}(\frac{1}{N})$ in terms of the tangent residual. This matches the pointwise convergence rate of the Newton proximal extragradient method in [40, Theorem 2.5]. Moreover, in the unconstrained setting where $H=\{0\}$, the tangent residual $\mathrm{res}(\mathbf{z}_i)$ reduces to the operator norm $\|F(\mathbf{z}_i)\|$. Thus, in this special case, our method achieves the same convergence rate of $\mathcal{O}(\frac{1}{N})$ as the second-order method in [29]. However, as noted in [29], it may be possible to achieve a faster rate using the anchoring scheme explored for first-order methods [54, 27]. We leave this as a direction for future research.

5.3. Convergence analysis: Strongly-convex-strongly-concave case. We proceed to the case where the smooth component of the objective in (1.1) is μ-strongly-convex-strongly-concave. We define a positive decreasing sequence $\{\zeta_k\}_{k\ge 0}$ by

$$\zeta_k = \begin{cases} 1 & \text{if } k=0, \\ \prod_{l=0}^{k-1}(1+2\eta_l\mu)^{-1} & \text{if } k\ge 1. \end{cases} \tag{5.8}$$

According to the result in part (b) of Proposition 4.1, the squared distance to the optimal solution $\|\mathbf{z}_k-\mathbf{z}^*\|^2$ can be bounded above by

$$\|\mathbf{z}_k-\mathbf{z}^*\|^2 \le \frac{2\|\mathbf{z}_0-\mathbf{z}^*\|^2}{2-\alpha}\,\zeta_k \le 2\|\mathbf{z}_0-\mathbf{z}^*\|^2\zeta_k, \tag{5.9}$$

where $\alpha\in(0,1)$ is a line search parameter. As a result, $\|\mathbf{z}_k-\mathbf{z}^*\|^2\le\epsilon$ when we have $\zeta_k \le \frac{\epsilon}{2\|\mathbf{z}_0-\mathbf{z}^*\|^2}$. Hence, in the following we will characterize the convergence behavior of the sequence $\{\zeta_k\}_{k\ge0}$, which immediately implies that of $\|\mathbf{z}_k-\mathbf{z}^*\|^2$. To begin with, we present the following lemma, which is a corollary of Lemma 5.4.

LEMMA 5.9. *Let* $\{\eta_k\}_{k\ge0}$ *be the step sizes in* (5.2) *generated by Algorithm* 5.1*. Then, we have* $\sum_{k=0}^{N-1}\frac{1}{\zeta_k\eta_k^2} \le \frac{1}{(1-\beta^2)\sigma_0^2} + \frac{\gamma_2^2L_2^2}{1-\beta^2}\|\mathbf{z}_0-\mathbf{z}^*\|^2$.

Proof. Recall $\mathcal{B}=\{k:\eta_k<\sigma_k\}$, i.e., the set of iteration indices where the line search backtracks. Since $\{\eta_k\}_{k\ge0}$ is determined by Algorithm 4.2, we have $\eta_k=\sigma_k$ when $k\notin\mathcal{B}$, where $\sigma_k=\eta_{k-1}\sqrt{1+2\eta_{k-1}\mu}/\beta$ for $k\ge1$. Hence, we can upper bound $\sum_{k\notin\mathcal{B},0\le k\le N-1}\frac{1}{\zeta_k\eta_k^2} \le \frac{1}{\sigma_0^2}+\sum_{k\notin\mathcal{B},1\le k\le N-1}\frac{\beta^2}{\zeta_k\eta_{k-1}^2(1+2\eta_{k-1}\mu)} = \frac{1}{\sigma_0^2}+\sum_{k\notin\mathcal{B},1\le k\le N-1}\frac{\beta^2}{\zeta_{k-1}\eta_{k-1}^2}$, where we used $\zeta_{k-1}=\zeta_k(1+2\eta_{k-1}\mu)$ in the last equality. This leads to $\sum_{k=0}^{N-1}\frac{1}{\zeta_k\eta_k^2} \le \sum_{k\in\mathcal{B},0\le k\le N-1}\frac{1}{\zeta_k\eta_k^2}+\frac{1}{\sigma_0^2}+\sum_{k\notin\mathcal{B},1\le k\le N-1}\frac{\beta^2}{\zeta_{k-1}\eta_{k-1}^2} \le$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

$\sum_{k\in\mathcal{B},0\le k\le N-1}\frac{1}{\zeta_k\eta_k^2}+\frac{1}{\sigma_0^2}+\sum_{k=0}^{N-1}\frac{\beta^2}{\zeta_k\eta_k^2}$. Combining the left-hand side with the last term on the right-hand side, we obtain $\sum_{k=0}^{N-1}\frac{1}{\zeta_k\eta_k^2}\le\frac{1}{1-\beta^2}\frac{1}{\sigma_0^2}+\frac{1}{1-\beta^2}\sum_{k\in\mathcal{B}}\frac{1}{\zeta_k\eta_k^2}$.

The rest follows from Lemma 5.4. ∎

By using Lemma 5.9, next we characterize the dynamic of the sequence $\{\zeta_k\}_{k\ge0}$.

LEMMA 5.10. *Let $\{\eta_k\}_{k\ge0}$ be the step sizes in* (5.2) *generated by Algorithm* 5.1*, and let $\{\zeta_k\}_{k\ge0}$ be defined in* (5.8)*. Further, recall the definition of γ_2 in* (5.6)*. Then for any $0\le k_1<k_2\le N$, we have $\frac{1}{\zeta_{k_2}}\ge\frac{1}{\zeta_{k_1}}+C_2^{-\frac{1}{2}}(\sum_{k=k_1}^{k_2-1}\frac{1}{\zeta_k})^{\frac{3}{2}}$, where*

$$\kappa_2(\mathbf{z}_0):=\frac{L_2\|\mathbf{z}_0-\mathbf{z}^*\|}{\mu}\quad\text{and}\quad C_2:=\frac{1}{4(1-\beta^2)}\left(\frac{1}{\sigma_0^2\mu^2}+\gamma_2^2\kappa_2^2(\mathbf{z}_0)\right). \tag{5.10}$$

Proof. By the definition of ζ_k in (5.8), we have $\eta_k=(\zeta_k/\zeta_{k+1}-1)/(2\mu)$. We apply Lemma 5.9 and rewrite the inequality in terms of $\{\zeta_k\}_{k=0}^{N-1}$ as

$$\begin{aligned}&\sum_{k=0}^{N-1}\left(\frac{2\mu\zeta_{k+1}}{\zeta_k-\zeta_{k+1}}\right)^2\frac{1}{\zeta_k}\le\frac{1}{(1-\beta^2)\sigma_0^2}+\frac{\gamma_2^2L_2^2}{1-\beta^2}\|\mathbf{z}_0-\mathbf{z}^*\|^2\\ \Leftrightarrow\ &\sum_{k=0}^{N-1}\left(\frac{1}{\zeta_{k+1}^{-1}-\zeta_k^{-1}}\right)^2\frac{1}{\zeta_k^3}\le\frac{1}{4(1-\beta^2)\sigma_0^2\mu^2}+\frac{\gamma_2^2\kappa_2^2(\mathbf{z}_0)}{4(1-\beta^2)}=C_2,\end{aligned} \tag{5.11}$$

where in (5.11) we used the definition of $\kappa_2(\mathbf{z}_0)$ and C_2. Since each summand in (5.11) is nonnegative, $\sum_{k=k_1}^{k_2-1}(\frac{1}{\zeta_{k+1}^{-1}-\zeta_k^{-1}})^2\frac{1}{\zeta_k^3}\le C_2$ for any $k_2>k_1\ge0$. Furthermore, by Hölder's inequality we get $[\sum_{k=k_1}^{k_2-1}(\frac{1}{\zeta_{k+1}}-\frac{1}{\zeta_k})]^{\frac{2}{3}}[\sum_{k=k_1}^{k_2-1}(\frac{1}{1/\zeta_{k+1}-1/\zeta_k})^2\frac{1}{\zeta_k^3}]^{\frac{1}{3}}\ge\sum_{k=k_1}^{k_2-1}\frac{1}{\zeta_k}$, and Lemma 5.10 follows from the above two inequalities. ∎

By leveraging Lemma 5.10, we establish a global complexity bound for our proposed method. Specifically, we show that the sequence $\{\zeta_k\}_{k\ge0}$ halves its value after a geometrically decreasing number of iterations. Our argument is inspired by the restarting strategy in [44, 42], but we do not need to restart our algorithm here.

THEOREM 5.11 (global convergence). *Suppose Assumptions* 2.4 *and* 5.1 *hold. Let $\{\mathbf{z}_k\}_{k\ge0}$ be the iterates generated by Algorithm* 5.1*, where $\hat{\eta}_k=\frac{\eta_{k-1}}{1+2\eta_{k-1}\mu}$, $\alpha\in(0,1)$, $\beta\in(0,1)$, and $\sigma_0>0$. Recall the definition of C_2 in* (5.10)*. Then for any $\epsilon>0$, we have $\zeta_N\le\frac{\epsilon}{\|\mathbf{z}_0-\mathbf{z}^*\|^2}$ after at most $\max\{5C_2^{1/3}+\log_2(\frac{\|\mathbf{z}_0-\mathbf{z}^*\|^2}{\epsilon})+1,1\}$ iterations.*

Proof. Without loss of generality, we assume $\frac{\epsilon}{\|\mathbf{z}_0-\mathbf{z}^*\|^2}\le1$; otherwise, the result becomes trivial as $\zeta_0=1$. Based on (5.8), $\{\zeta_k\}_{k\ge0}$ is nonincreasing in k. Hence, from Lemma 5.10 we have $\frac{1}{\zeta_{k_2}}\ge\frac{1}{\zeta_{k_1}}+C_2^{-\frac{1}{2}}(\sum_{k=k_1}^{k_2-1}\frac{1}{\zeta_k})^{\frac{3}{2}}\ge\frac{1}{\zeta_{k_1}}+C_2^{-\frac{1}{2}}(\frac{k_2-k_1}{\zeta_{k_1}})^{\frac{3}{2}}$. In particular, this implies $\zeta_{k_2}\le\frac{1}{2}\zeta_{k_1}$ when we have $k_2-k_1\ge C_2^{\frac{1}{3}}\zeta_{k_1}^{\frac{1}{3}}$. Hence, for any integer $l\ge0$, we can prove by induction that the number of iterations N required to achieve $\zeta_N\le2^{-l}$ does not exceed $\sum_{k=0}^{l-1}\lceil C_2^{\frac{1}{3}}2^{-\frac{1}{3}k}\rceil\le\sum_{k=0}^{l-1}(C_2^{\frac{1}{3}}2^{-\frac{1}{3}k}+1)\le\frac{1}{1-2^{-\frac{1}{3}}}C_2^{\frac{1}{3}}+l$. Theorem 5.11 follows by setting $l=\lceil\log_2(\frac{\|\mathbf{z}_0-\mathbf{z}^*\|^2}{\epsilon})\rceil\le\log_2(\frac{\|\mathbf{z}_0-\mathbf{z}^*\|^2}{\epsilon})+1$. ∎

Theorem 5.11 ensures global convergence of our second-order optimistic method, achieving any accuracy ϵ within $\mathcal{O}(C_2^{1/3}+\log(\frac{\|\mathbf{z}_0-\mathbf{z}^*\|^2}{\epsilon}))$ iterations. Furthermore, we show it attains fast local R-superlinear convergence, as stated in the following theorem.

Downloaded 10/22/25 to 128.62.19.54 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 10/22/25 to 128.62.19.54 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

THEOREM 5.12 (local convergence). *Under the same conditions as in Theorem* 5.11*, for any* $k \geq 0$*, we have* $\zeta_{k+1} \leq \frac{1}{2}\max\{\gamma_2\kappa_2(\mathbf{z}_0), \frac{\beta^k}{\sigma_0\mu}\}\zeta_k^{\frac{3}{2}}$*, where* γ_2 *and* $\kappa_2(\mathbf{z}_0)$ *are defined in* (5.6) *and* (5.10)*, respectively.*

Proof. The definition in (5.8) implies $\zeta_{k+1} = \frac{\zeta_k}{1+2\eta_k\mu} \leq \frac{\zeta_k}{2\eta_k\mu}$, and Lemma 5.5 implies $\frac{1}{\eta_k} \leq \max\{\gamma_2 L_2\|\mathbf{z}_0 - \mathbf{z}^*\|, \frac{\beta^k}{\sigma_0}\}\prod_{l=0}^{k-1}(1+2\eta_l\mu)^{-\frac{1}{2}} = \max\{\gamma_2 L_2\|\mathbf{z}_0 - \mathbf{z}^*\|, \frac{\beta^k}{\sigma_0}\}\zeta_k^{-\frac{1}{2}}$. The claim follows from these inequalities and the definition of $\kappa_2(\mathbf{z}_0)$ in (5.10). ∎

Theorem 5.12 implies that the sequence $\{\zeta_k\}_{k\geq 0}$ converges to 0 at a superlinear convergence rate of order $\frac{3}{2}$ once we have $\zeta_k < 4\min\{\frac{1}{\gamma_2^2\kappa_2^2(\mathbf{z}_0)}, \frac{\sigma_0^2\mu^2}{\beta^{2k}}\}$. Due to the result in Theorem 5.11, the sequence indeed eventually falls below this threshold after a finite number of iterations. Hence, in light of the bound in (5.9), this, in turn, implies the local R-superlinear convergence of $\|\mathbf{z}_k - \mathbf{z}^*\|^2$.

By combining the global convergence result in Theorem 5.11 and the local convergence result in Theorem 5.12, we characterize the overall iteration complexity of the second-order optimistic method in Algorithm 5.1. Specifically, if the required accuracy ϵ is moderate, we simply use the global complexity result in Theorem 5.11. On the other hand, if ϵ is small, we upper bound the number of iterations needed to reach the local convergence neighborhood using Theorem 5.11. Subsequently, we require at most $\mathcal{O}(\log\log(1/\epsilon))$ additional iterations to achieve the desired accuracy. We summarize the corresponding complexity bounds in Corollary 5.13.

COROLLARY 5.13. *Suppose* ϵ *is the required accuracy that we aim for, i.e.,* $\|\mathbf{z} - \mathbf{z}^*\|^2 \leq \epsilon$*. Under the same conditions as in Theorem* 5.11*, we have the following complexity bound for Algorithm* 5.1*.*

- *If* $\epsilon \geq \frac{4\mu^2}{\gamma_2^2 L_2^2}$*, then* $\|\mathbf{z}_N - \mathbf{z}^*\|^2 \leq \epsilon$ *when the number of iterations* N *exceeds* $\max\{5C_2^{1/3} + \log_2(\frac{\|\mathbf{z}_0 - \mathbf{z}^*\|^2}{\epsilon}) + 1, 1\}$*, where* C_2 *is defined in* (5.10)*.*
- *If* $\epsilon < \frac{4\mu^2}{\gamma_2^2 L_2^2}$*, then* $\|\mathbf{z}_N - \mathbf{z}^*\|^2 \leq \epsilon$ *when the number of iterations* N *exceeds* $\max\{5C_2^{\frac{1}{3}} + 2\log_2(\gamma_2\kappa_2(\mathbf{z}_0)), 1 - \log_{\frac{1}{\beta}}(\gamma_2\sigma_0\mu\kappa_2(\mathbf{z}_0)), 1\} + \log_{\frac{3}{2}}\log_2(\frac{8\mu^2}{\gamma_2^2 L_2^2 \epsilon}) + 1$.

Proof. Recall that $\|\mathbf{z}_k - \mathbf{z}^*\|^2 \leq \epsilon$ if $\zeta_k \leq \frac{\epsilon}{2\|\mathbf{z}_0 - \mathbf{z}^*\|^2}$ by (5.9). Hence, it suffices to upper bound the number of iterations required such that the latter condition holds. Also, we need only prove the second case where $\epsilon < \frac{4\mu^2}{\gamma_2^2 L_2^2}$, as the first case directly follows from Theorem 5.11. Let N_1 be the smallest integer such that $\zeta_{N_1} \leq 2/(\gamma_2^2\kappa_2^2(\mathbf{z}_0))$ and $\frac{\beta^{N_1}}{\sigma_0} \leq \gamma_2 L_2\|\mathbf{z}_0 - \mathbf{z}^*\|$. By setting $\epsilon = \frac{4\|\mathbf{z}_0 - \mathbf{z}^*\|^2}{\gamma_2^2\kappa_2^2(\mathbf{z}_0)} = \frac{4\mu^2}{\gamma_2^2 L_2^2}$ in Theorem 5.11, we obtain that the first condition is satisfied when N_1 exceeds

$$\max\left\{5C_2^{\frac{1}{3}} + \log_2\left(\frac{\gamma_2^2 L_2^2\|\mathbf{z}_0 - \mathbf{z}^*\|^2}{4\mu^2}\right) + 1, 1\right\} \leq \max\left\{5C_2^{\frac{1}{3}} + 2\log_2\left(\gamma_2\kappa_2(\mathbf{z}_0)\right), 1\right\}.$$

Moreover, the second condition holds when $N_1 \geq -\log_{1/\beta}(\gamma_2\sigma_0 L_2\|\mathbf{z}_0 - \mathbf{z}^*\|) = -\log_{1/\beta}(\gamma_2\sigma_0\mu\kappa_2(\mathbf{z}_0))$. As N_1 is the smallest integer meeting both conditions, we have

$$N_1 \leq \max\left\{5C_2^{\frac{1}{3}} + 2\log_2\left(\gamma_2\kappa_2(\mathbf{z}_0)\right), 1 - \log_{\frac{1}{\beta}}(\gamma_2\sigma_0\mu\kappa_2(\mathbf{z}_0)), 1\right\}. \tag{5.12}$$

Furthermore, since $\frac{\beta^{N_1}}{\sigma_0} \leq \gamma_2 L_2\|\mathbf{z}_0 - \mathbf{z}^*\|$, we have $\zeta_{k+1} \leq \frac{1}{2}\gamma_2\kappa_2(\mathbf{z}_0)\zeta_k^{\frac{3}{2}}$ by Theorem 5.12 for all $k \geq N_1$, which implies that $\frac{1}{4}\gamma_2^2\kappa_2^2(\mathbf{z}_0)\zeta_{k+1} \leq (\frac{1}{4}\gamma_2^2\kappa_2^2(\mathbf{z}_0)\zeta_k)^{\frac{3}{2}}$. By induction, we can prove that $\frac{1}{4}\gamma_2^2\kappa_2^2(\mathbf{z}_0)\zeta_k \leq 2^{-(\frac{3}{2})^{k-N_1}}$ for all $k \geq N_1$. Hence, the additional number of iterations required to achieve $\zeta_k \leq \frac{\epsilon}{2\|\mathbf{z}_0 - \mathbf{z}^*\|^2}$ does not exceed

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

$\lceil \log_{\frac{3}{2}} \log_2 \left(\frac{8\|\mathbf{z}_0-\mathbf{z}^*\|^2}{\gamma_2^2\kappa_2^2(\mathbf{z}_0)\epsilon}\right)\rceil \leq \log_{\frac{3}{2}} \log_2 \left(\frac{8\mu^2}{\gamma_2^2 L_2^2 \epsilon}\right)+1$, where we used the definition of $\kappa_2(\mathbf{z}_0)$. Together with (5.12), the second complexity bound follows. ∎

Corollary 5.13 characterizes the overall iteration complexity of our proposed method. To summarize if the required accuracy ϵ is larger than $\frac{4\mu^2}{\gamma_2^2 L_2^2}$, then the iteration complexity is bounded by

$$\mathcal{O}\left(\left(\frac{L_2\|\mathbf{z}_0-\mathbf{z}^*\|}{\mu}\right)^{\frac{2}{3}}+\log\left(\frac{\|\mathbf{z}_0-\mathbf{z}^*\|^2}{\epsilon}\right)\right). \tag{5.13}$$

Otherwise, for ϵ smaller than $\frac{4\mu^2}{\gamma_2^2 L_2^2}$, the overall iteration complexity is

$$\mathcal{O}\left(\left(\frac{L_2\|\mathbf{z}_0-\mathbf{z}^*\|}{\mu}\right)^{\frac{2}{3}}+\log\log\left(\frac{\mu^2}{L_2^2\epsilon}\right)\right). \tag{5.14}$$

In (5.14), the first term captures the number of required iterations to reach the local neighborhood and the second term corresponds to the number of iterations in the local neighborhood to achieve accuracy ϵ.

Finally, by using Proposition 4.3, we can derive a similar complexity bound for the case of finding a point with small tangent residual.

COROLLARY 5.14. *If the goal is to find a point* $\mathbf{z}$ *such that* $\mathrm{res}(\mathbf{z}) \leq \rho$*, then under the conditions of Theorem* 5.11*, the complexity of Algorithm* 5.1 *is bounded as follows:*

- *If* $\rho = \Omega(\frac{\mu^2}{L_2})$*, then* $\mathrm{res}(\mathbf{z}_N) \leq \rho$ *after* $\mathcal{O}((\frac{L_2\|\mathbf{z}_0-\mathbf{z}^*\|}{\mu})^{2/3}+\log(\frac{L_2\|\mathbf{z}_0-\mathbf{z}^*\|^2}{\rho}))$ *iterations.*
- *If* $\rho = \mathcal{O}(\frac{\mu^2}{L_2})$*, then* $\mathrm{res}(\mathbf{z}_N) \leq \rho$ *after* $\mathcal{O}((\frac{L_2\|\mathbf{z}_0-\mathbf{z}^*\|}{\mu})^{2/3}+\log\log(\frac{\mu^2}{L_2\rho}))$ *iterations.*

Proof. By Lemma 5.5, it holds that $\frac{1}{\eta_{N-1}} \leq \max\{\gamma_2 L_2\|\mathbf{z}_0-\mathbf{z}^*\|, \frac{\beta^{N-1}}{\sigma_0}\}\prod_{k=0}^{N-2}(1+2\eta_k\mu)^{-\frac{1}{2}}$ for any $k \geq 0$. Hence, it follows from Proposition 4.3 that $\mathrm{res}(\mathbf{z}_N) \leq \frac{\frac{1+\alpha}{\sqrt{1-\alpha}}\|\mathbf{z}_0-\mathbf{z}^*\|\max\{\gamma_2 L_2\|\mathbf{z}_0-\mathbf{z}^*\|, \frac{\beta^{N-1}}{\sigma_0}\}}{\prod_{k=0}^{N-2}(1+2\eta_k\mu)}$. Thus, the rest of the proof follows by applying the same arguments as in Corollary 5.13. ∎

Comparison with [46]. We close the discussion on our iteration complexity by mentioning that similar complexity bounds to (5.13) and (5.14) were also established in [46] but via a different approach. Specifically, in [46] the authors applied a restarting technique on the extragradient-type method in [6]. Under Assumptions 2.4 and 5.1, they proved a complexity bound of $\mathcal{O}((\frac{L_2 D}{\mu})^{\frac{2}{3}}\log\frac{\mu D^2}{\epsilon})$ in terms of the primal-dual gap, where D is an upper bound on $\|\mathbf{z}_0-\mathbf{z}^*\|$. Moreover, with the additional assumptions that F is L_1-Lipschitz and L_2-smooth, they further combined it with the cubic-regularization-based method in [20] and improved the complexity bound to $\mathcal{O}((\frac{L_2 D}{\mu})^{\frac{2}{3}}\log\frac{L_1 L_2 D}{\mu^2}+\log\log\frac{L_1^3}{\mu^2\epsilon})$. Compared with ours, their method is restricted to unconstrained problems, and can only achieve local superlinear convergence under stronger assumptions. Also, our method is simpler as it eliminates the need to alternate between multiple update rules. Most importantly, they did not provide any detail on solving the subproblems nor identify the procedure and cost of finding an admissible step size, while we propose a line search scheme and identify the number of subsolver calls that our proposed method requires. This is the topic of the next subsection.

5.4. Complexity bound of the line search scheme. So far, we have analyzed the iterations needed to solve (1.1) using Algorithm 5.1 for both convex-concave and

Downloaded 10/22/25 to 128.62.19.54 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 10/22/25 to 128.62.19.54 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

strongly-convex strongly-concave settings. What remains is to determine the total number of optimistic subsolver calls, each solving an instance of (5.5).

THEOREM 5.15. *Consider Algorithm* 5.1 *with* $\alpha \in (0,1)$, $\beta \in (0,1)$, *and* $\sigma_0 > 0$, *and recall the definition of* γ_2 *in* (5.6). *For both convex-concave (Theorem* 5.7*) and strongly-convex-strongly-concave (Theorem* 5.11*) cases, the total number of calls to the optimistic subsolver after* N *iterations is* $\max\{2N - 1 + \log_{\frac{1}{\beta}}(\sigma_0\gamma_2 L_2\|\mathbf{z}_0 - \mathbf{z}^*\|), N\}$.

Proof. By Proposition 4.6, the number of calls to the optimistic subsolver at the kth iteration is given by $\log_{\frac{1}{\beta}} \frac{\sigma_k}{\eta_k} + 1$. Since $\sigma_k = \eta_{k-1}\sqrt{1 + 2\eta_{k-1}\mu}/\beta$ for $k \geq 1$, the total number of calls after N iterations is $N_{\text{total}} = \sum_{k=0}^{N-1}(\log_{\frac{1}{\beta}} \frac{\sigma_k}{\eta_k} + 1) = N + \log_{\frac{1}{\beta}} \frac{\sigma_0}{\eta_0} + \sum_{k=1}^{N-1} \log_{\frac{1}{\beta}} \frac{\eta_{k-1}\sqrt{1+2\eta_{k-1}\mu}}{\beta\eta_k} = 2N - 1 + \log_{\frac{1}{\beta}} \frac{\sigma_0}{\eta_0} + \sum_{k=1}^{N-1} \log_{\frac{1}{\beta}} \frac{\eta_{k-1}\sqrt{1+2\eta_{k-1}\mu}}{\eta_k} = 2N - 1 + \log_{\frac{1}{\beta}}(\frac{\sigma_0}{\eta_{N-1}} \prod_{k=0}^{N-2} \sqrt{1 + 2\eta_k\mu})$. Further, Lemma 5.5 shows that $\frac{1}{\eta_{N-1}^2} \prod_{k=0}^{N-2}(1 + 2\eta_k\mu) \leq \max\{\gamma_2^2 L_2^2\|\mathbf{z}_0 - \mathbf{z}^*\|^2, \frac{\beta^{2(N-1)}}{\sigma_0^2}\}$. Hence, combining these two inequalities, we have $N_{\text{total}} \leq 2N - 1 + \max\{\log_{\frac{1}{\beta}}(\sigma_0\gamma_2 L_2\|\mathbf{z}_0 - \mathbf{z}^*\|), -(N-1)\} = \max\{2N - 1 + \log_{\frac{1}{\beta}}(\sigma_0\gamma_2 L_2\|\mathbf{z}_0 - \mathbf{z}^*\|), N\}$. ∎

This result shows that if we run Algorithm 5.1 for N iterations, the total number of calls to a subsolver of (5.5) is of the order $2N + \mathcal{O}(\log(\sigma_0 L_2\|\mathbf{z}_0 - \mathbf{z}^*\|))$. Hence, the average number of calls per iteration can be bounded by a constant close to 2 for large N. In comparison, the line search subroutines presented in [40, 6] require a total number of subsolver calls of $\mathcal{O}(N \log\log(1/\epsilon))$ and $\mathcal{O}(N \log(1/\epsilon))$, respectively.

5.5. Complexity of the optimistic subsolver. This section aims to characterize the computational cost of solving (5.5) in our second-order optimistic method. To simplify the notation, at iteration k, we define $\mathbf{A}_k(\eta) := \eta DF(\mathbf{z}_k) + \mathbf{I}$, $\mathbf{b}_k(\eta) := \eta F(\mathbf{z}_k) + \mathbf{v}_k$, and $\mathbf{s} := \mathbf{z} - \mathbf{z}_k$, where we make the dependence on η in $\mathbf{A}_k$ and $\mathbf{b}_k$ explicit. The subproblem in (5.5) can be cast as finding $\mathbf{s} \in \mathbb{R}^d$ and $\mathbf{r} \in \mathbb{R}^d$ such that

$$\mathbf{r} \in \mathbf{A}_k(\eta)\mathbf{s} + \mathbf{b}_k(\eta) + \eta H(\mathbf{s} + \mathbf{x}_k), \quad \|\mathbf{r}\| \leq \frac{\alpha}{4}\|\mathbf{s}\|. \tag{5.15}$$

Due to the line search procedure, at the kth iteration, we solve a series of subproblems of the form in (5.15), where the step size η takes values in the sequence $\eta = \sigma_k, \sigma_k\beta, \ldots, \sigma_k\beta^{l_k - 1}$. Here, l_k is the number of calls to the optimistic subsolver at the kth iteration and is given by $l_k = \log_{\frac{1}{\beta}} \frac{\sigma_k}{\eta_k} + 1$ as established in Proposition 4.6. We now proceed to discuss the two special cases outlined earlier, prior to section 5.1.

Unconstrained setting. In this case, $H(\mathbf{z}) = \{0\}$ for all $\mathbf{z}$ and thus $\mathbf{r} = \mathbf{A}_k(\eta)\mathbf{s} + \mathbf{b}_k(\eta)$. As a result, (5.15) reduces to finding $\mathbf{s} \in \mathbb{R}^d$ such that $\|\mathbf{A}_k(\eta)\mathbf{s} + \mathbf{b}_k(\eta)\| \leq \frac{\alpha}{4}\|\mathbf{s}\|$. This can be achieved by solving the linear system $\mathbf{A}_k(\eta)\mathbf{s} + \mathbf{b}_k(\eta) = 0$ exactly, which has an arithmetic complexity of $\mathcal{O}(d^\omega)$, where ω is the matrix multiplication exponent. Alternatively, we can employ an iterative linear solver relying solely on matrix-vector products. Specifically, we consider the CGLS method [18, 47], which is equivalent to running the conjugate gradient method on the normal equation under the exact arithmetic but is numerically more stable. Our analysis adopts a similar approach to that in [23] and builds on the following key lemma, whose proof is in Appendix D.1.

LEMMA 5.16. *For inputs* $\mathbf{A} \in \mathbb{R}^{d\times d}$, $\mathbf{b} \in \mathbb{R}^d$, $\alpha \in (0,1]$, *if* $\frac{1}{2}(\mathbf{A} + \mathbf{A}^\top) \succeq \mathbf{I}$ *holds and* $\kappa(\mathbf{A}) := \frac{\|\mathbf{A}\|_{\text{op}}}{\lambda_{\min}(\frac{1}{2}(\mathbf{A}+\mathbf{A}^\top))}$, *then CGLS returns* $(\mathbf{s}, \mathbf{r})$ *with* $\mathbf{r} = \mathbf{A}\mathbf{s} + \mathbf{b}$ *and* $\|\mathbf{r}\| \leq \frac{\alpha}{4}\|\mathbf{s}\|$ *after at most* $\min\{\frac{2\sqrt{2}(\alpha+4)}{\alpha}\|\mathbf{A}\|_{\text{op}}, 2\kappa(\mathbf{A})\log(\frac{2(\alpha+4)}{\alpha}\|\mathbf{A}\|_{\text{op}})\}$ *matrix-vector products.*

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 10/22/25 to 128.62.19.54 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Lemma 5.16 provides an upper bound on the cost of solving a single linear system in terms of the operator norm and the condition number of the input matrix $\mathbf{A}$. Using this result, we can derive an upper bound on the total number of matrix-vector products, as stated in the following theorem. The proof is in Appendix D.2.

THEOREM 5.17. *Let $\{\mathbf{z}_k\}_{k\geq 0}$ be generated by Algorithm 5.1 with parameters $\alpha \in (0,1)$, $\beta \in (0,1)$, and $\sigma_0 > 0$. Suppose that $\|DF(\mathbf{z}_k)\|_{\text{op}} \leq L_1$ for all $k = 0, \ldots, N-1$.*

(a) *For the convex-concave setting (Theorem 5.7), the total number of matrix-vector products required to find $(\mathbf{x}, \mathbf{y})$ with $\Delta_{\mathcal{B}_1 \times \mathcal{B}_2}(\mathbf{x}, \mathbf{y}) \leq \epsilon$ can be bounded by $\mathcal{O}(\frac{L_1 D^2}{\epsilon} + (\frac{L_2 \|\mathbf{z}_0 - \mathbf{z}^*\| D^2}{\epsilon})^{2/3} + (\frac{D^2}{\sigma_0 \epsilon})^{1/2})$, where $D = \max_{\mathbf{z} \in \mathcal{B}_1 \times \mathcal{B}_2} \|\mathbf{z}_0 - \mathbf{z}\|$.*

(b) *For the strongly-convex-strongly-concave setting (Theorem 5.11), let N_ϵ be the smallest integer N to satisfy $\|\mathbf{z}_N - \mathbf{z}^*\|^2 \leq \epsilon$. Then the total number of matrix-vector products can be bounded by $\mathcal{O}(\frac{L_1 N_\epsilon}{\mu} \log(\frac{L_1 \|\mathbf{z}_0 - \mathbf{z}^*\|^2}{\mu \epsilon}))$.*

The above result requires the operator norm of the Jacobian to be bounded by L_1. This assumption is without loss of generality, as Theorem 5.7 establishes that the iterates remain in a bounded set. Consequently, Theorem 5.17 shows that the total number of matrix-vector products is bounded by $\mathcal{O}(\frac{L_1 D^2}{\epsilon})$ in the convex-concave setting, and $\mathcal{O}(\frac{L_1 N_\epsilon}{\mu} \log(\frac{L_1 \|\mathbf{z}_0 - \mathbf{z}^*\|^2}{\mu \epsilon}))$ in the strongly-convex-strongly-concave setting.

Proximal setting. In this case, the resolvent of the operator H is easy to compute. This enables us to use the composite fast extragradient (FEG) method proposed in [7], which is specifically designed to minimize the tangent residual in the composite setting. In this setting, we can establish a result analogous to Lemma 5.16.

LEMMA 5.18. *Consider the same setting as in Lemma 5.16. Then the composite FEG algorithm returns $(\mathbf{s}, \mathbf{r})$ with $\mathbf{r} \in \mathbf{A}\mathbf{s} + \mathbf{b} + \eta H(\mathbf{s} + \mathbf{x}_k)$ and $\|\mathbf{r}\| \leq \frac{\alpha}{4}\|\mathbf{s}\|$ after at most $\min\{\frac{4(\alpha+4)}{\alpha}\|\mathbf{A}\|_{\text{op}}, 8\kappa(\mathbf{A}) \log(\frac{4(\alpha+4)}{\alpha}\|\mathbf{A}\|_{\text{op}})\}$ matrix-vector products and resolvent computations, where $\kappa(\mathbf{A}) := \frac{\|\mathbf{A}\|_{\text{op}}}{\lambda_{\min}(\frac{1}{2}(\mathbf{A} + \mathbf{A}^\top))}$.*

The complexity bound in Lemma 5.18 matches that in Lemma 5.16 up to a constant factor. Thus, using the same argument, we can bound the total complexity as in Theorem 5.17, with details omitted for brevity.

6. Numerical experiments. To test our second-order optimistic method, we consider the unconstrained saddle point problem

$$\min_{\mathbf{x} \in \mathbb{R}^n} \max_{\mathbf{y} \in \mathbb{R}^n} \; f(\mathbf{x}, \mathbf{y}) = \frac{L_2}{6}\|\mathbf{x}\|^3 + \langle \mathbf{A}\mathbf{x} - \mathbf{b}, \mathbf{y} \rangle + \frac{\mu}{2}\|\mathbf{x}\|^2 - \frac{\mu}{2}\|\mathbf{y}\|^2, \tag{6.1}$$

where $L_2 > 0$, $\mu \geq 0$, and the matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is chosen as $\mathbf{A} = \begin{bmatrix} 1 & -1 & & & \\ & 1 & -1 & & \\ & & \ddots & \ddots & \\ & & & 1 & -1 \\ & & & & 1 \end{bmatrix}$. This is a special case of problem (1.1) with $\mathcal{X} = \mathbb{R}^m$, $\mathcal{Y} = \mathbb{R}^n$, $h_1 = 0$, and $h_2 = 0$. Moreover, we generate the entries of the vector $\mathbf{b} \in \mathbb{R}^n$ independently and randomly from the interval $[-1, 1]$ and then normalize the vector such that $\|\mathbf{b}\| = 1$. We can verify that problem (6.1) is convex-concave when $\mu = 0$ and strongly-convex-strongly-concave when $\mu > 0$. Moreover, F defined in (2.1) is L_2-second-order smooth. The subproblem in (5.5) for the optimistic subsolver is equivalent to solving a system of linear equations. Specifically, the exact update rule in (5.2) can be instantiated as

$$\mathbf{z}_{k+1} = \mathbf{z}_k - (\mathbf{I} + \eta_k DF(\mathbf{z}_k))^{-1}(\eta_k F(\mathbf{z}_k) + \mathbf{v}_k), \tag{6.2}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

where $\mathbf{v}_k = \hat{\eta}_k(F(\mathbf{z}_k) - F(\mathbf{z}_{k-1}) - DF(\mathbf{z}_{k-1})(\mathbf{z}_k - \mathbf{z}_{k-1}))$ and η_k is determined by the adaptive line search scheme (see Algorithm 5.1). In our experiments, we set the initial point as $(\mathbf{x}_0, \mathbf{y}_0) = (\mathbf{0}, \mathbf{0})$.

Convex-concave setting. We consider problem (6.1) with $\mu = 0$. It can be verified that it has a unique saddle point $\mathbf{z}^* = (\mathbf{x}^*, \mathbf{y}^*)$ given by $\mathbf{x}^* = \mathbf{A}^{-1}\mathbf{b}$ and $\mathbf{y}^* = -\frac{L_2}{2}\|\mathbf{x}^*\|(\mathbf{A}^\mathsf{T})^{-1}\mathbf{x}^*$. Since the feasible set is unbounded, we use the restricted primal-dual gap defined in (2.4) as the performance metric. With $\mathcal{B}_1 = \mathbb{R}^n$ and $\mathcal{B}_2 = \{\mathbf{y} : \|\mathbf{y}\| \le R\}$, the gap can be computed by $\Delta_{\mathcal{B}_1 \times \mathcal{B}_2}(\mathbf{x}, \mathbf{y}) = \frac{L_2}{6}\|\mathbf{x}\|^3 + R\|\mathbf{A}\mathbf{x} - \mathbf{b}\| + \frac{2}{3}\sqrt{\frac{2}{L_2}}\|\mathbf{A}^\mathsf{T}\mathbf{y}\|^{\frac{3}{2}} + \langle \mathbf{b}, \mathbf{y}\rangle$. By taking the supremum over $(\mathbf{x}, \mathbf{y}) \in \mathcal{B}_1 \times \mathcal{B}_2$, we obtain $\mathbf{x} = -\sqrt{\frac{2}{L_2}}\frac{\mathbf{A}^\mathsf{T}\bar{\mathbf{y}}_N}{\sqrt{\|\mathbf{A}^\mathsf{T}\bar{\mathbf{y}}_N\|}}$ and $\mathbf{y} = \frac{R(\mathbf{A}\bar{\mathbf{x}}_N - \mathbf{b})}{\|\mathbf{A}\bar{\mathbf{x}}_N - \mathbf{b}\|}$, leading to the following upper bound:

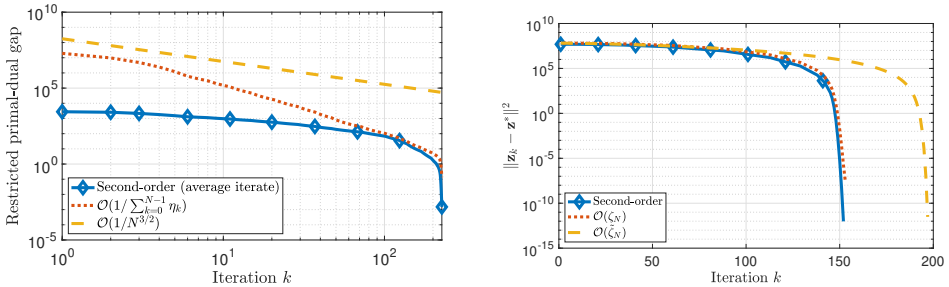
$$\Delta_{\mathcal{B}_1 \times \mathcal{B}_2}(\bar{\mathbf{x}}_N, \bar{\mathbf{y}}_N) \le \frac{1}{2}\left(\frac{2}{L_2}\|\mathbf{A}^\mathsf{T}\bar{\mathbf{y}}_N\| + R^2\right)\left(\sum_{k=0}^{N-1}\eta_k\right)^{-1}, \tag{6.3}$$

where $\bar{\mathbf{x}}_N = \frac{1}{\sum_{k=0}^{N-1}\eta_k}\sum_{k=0}^{N-1}\eta_k\mathbf{x}_{k+1}$ and $\bar{\mathbf{y}}_N = \frac{1}{\sum_{k=0}^{N-1}\eta_k}\sum_{k=0}^{N-1}\eta_k\mathbf{y}_{k+1}$.

In our experiment, the problem parameters are chosen as $n = 200$ and $L_2 = 10$, and the line search parameters are $\alpha = 0.5$, $\beta = 0.5$, $\sigma_0 = 1$, and $\epsilon = 10^{-10}$. In Figure 1a, we plot the restricted primal-dual gap at the averaged iterate $(\bar{\mathbf{x}}_N, \bar{\mathbf{y}}_N)$ and the convergence bounds in (6.3) for comparison. We observe that the second-order method exhibits a sublinear convergence rate in the early stage, which is comparable to our theoretical rate of $\mathcal{O}(N^{-\frac{3}{2}})$. Moreover, the bound in (6.3) provides a tight upper bound for the second-order method especially in the later stage.

Strongly-convex-strongly-concave setting. We consider problem (6.1) with $\mu > 0$. In this case, it has a unique saddle point $\mathbf{z}^* = (\mathbf{x}^*, \mathbf{y}^*)$, which we compute numerically using MATLAB's built-in nonlinear equation solver. By Proposition 4.1, for the second-order optimistic method in Algorithm 5.1, we have $\|\mathbf{z}_k - \mathbf{z}^*\|^2 \le \frac{2\|\mathbf{z}_0 - \mathbf{z}^*\|^2}{2-\alpha}\zeta_k$, where ζ_k is given in (5.8). To compare the empirical results with our theoretical findings, we consider a "simulated" sequence $\{\tilde{\zeta}_k\}_{k\ge 0}$ with $\tilde{\zeta}_0 = 1$ following the dynamic $\tilde{\zeta}_k^{-1} = 1 + C(\sum_{l=0}^{k-1}\tilde{\zeta}_l^{-1})^{\frac{3}{2}}$ for any $k \ge 1$ for some $C > 0$. In light of Lemma 5.10, we can see that $\zeta_k \le \tilde{\zeta}_k$ for all $k \ge 0$ if the parameter C is chosen as $\gamma_2\kappa_2(\mathbf{z}_0)$.

In our experiment, the problem parameters are chosen as $n = 200$, $L_2 = 10,000$, $\mu = 10^{-3}$, and the line search parameters are the same as in the convex-concave setting. In Figure 1b we plot the distance to the saddle point $\mathbf{z}^*$ of the second-order



(a) Convex-concave setting (b) Strongly-convex-strongly-concave setting

FIG. 1. *The performance of the second-order optimistic method on solving the saddle point problems in* (6.1) *with* $\mu = 0$ *and* $\mu = 10^{-3}$*, respectively.*

Downloaded 10/22/25 to 128.62.19.54 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

TABLE 1
The maximum average number of subsolver calls per iteration in the second-order optimistic method among 50 random instances.

Parameters	Max average number of subsolver calls per iteration	
	Convex-concave setting	Strongly-convex-strongly-concave setting
$\sigma_0 = 1, \beta = 0.5$	1.9780	2.0174
$\sigma_0 = 10, \beta = 0.5$	1.9860	2.0492
$\sigma_0 = 100, \beta = 0.5$	1.9920	2.0964
$\sigma_0 = 1, \beta = 0.9$	1.8580	2.1504
$\sigma_0 = 10, \beta = 0.9$	1.9020	2.1681
$\sigma_0 = 100, \beta = 0.9$	1.9440	2.4609

optimistic methods versus the number of iterations. For comparison, we also plot the bound $\frac{2\|\mathbf{z}_0-\mathbf{z}^*\|^2}{2-\alpha}\zeta_k$ using the actual sequence $\{\zeta_k\}_{k\geq 0}$ as well as the bound $\frac{2\|\mathbf{z}_0-\mathbf{z}^*\|^2}{2-\alpha}\tilde{\zeta}_k$ the simulated sequence $\{\tilde{\zeta}_k\}_{k\geq 0}$. In particular, our second-order method exhibits local superlinear convergence in the neighborhood of the saddle point, doubling the accuracy of the solution within a few iterations. Also, the bound $\frac{2\|\mathbf{z}_0-\mathbf{z}^*\|^2}{2-\alpha}\zeta_k$ provides a tight upper bound for the second-order method. Moreover, if the parameter C in definition of $\tilde{\zeta}_k^{-1}$ is carefully chosen, we observe that the bound $\frac{2\|\mathbf{z}_0-\mathbf{z}^*\|^2}{2-\alpha}\tilde{\zeta}_k$ has a qualitatively similar convergence behavior as the empirical result.

The complexity of line search. We also empirically evaluate the number of subsolver calls made in our second-order optimistic method. We vary the line search parameters σ_0 and β and run our method on problem (6.1) until it finds a solution with accuracy 10^{-10} or completes 500 iterations. The maximum is reported in Table 1. We observe that the line search complexity is unaffected by the initial trial step size σ_0, but a larger β may increase subsolver calls in the strongly-convex-strongly-concave setting. In our tests, setting $\beta = 0.5$ keeps the average subsolver calls per iteration around 2, a reasonable trade-off for improved convergence.

7. Conclusions. We proposed the generalized optimistic method for composite saddle point problems by approximating the proximal point method. We also designed a novel line search scheme to select the step size in our method adaptively and without knowledge of the smoothness coefficients of the objective. Under our unified framework, our second-order optimistic method achieves a complexity of $\mathcal{O}(\epsilon^{-\frac{2}{3}})$ in terms of the primal-dual gap in the convex-concave setting, and a complexity of $\mathcal{O}((L_2\|\mathbf{z}_0-\mathbf{z}^*\|/\mu)^{\frac{2}{3}} + \log\log 1/\epsilon)$ in terms of the distance to the optimal solution in the strongly-convex-strongly-concave setting. We also showed our line search scheme requires a constant number of calls to the optimistic subsolvers per iteration on average.

Two concurrent papers [9, 26] enhance the HPE scheme with acceleration, eliminating binary search from prior methods and achieving optimal convergence for second-order convex optimization. Extending our second-order optimistic method to this setting is a promising direction for future research.

Appendix A. Proof of Proposition 3.1. The update rule of PPM in (3.1) implies $\eta_k F(\mathbf{z}_{k+1}) + \mathbf{z}_{k+1} - \mathbf{z}_k \in -\eta_k H(\mathbf{z}_{k+1})$. Hence, using the definitions of H in (2.1), we have $\eta_k\langle F(\mathbf{z}_{k+1}), \mathbf{z}_{k+1} - \mathbf{z}\rangle + \langle \mathbf{z}_{k+1} - \mathbf{z}_k, \mathbf{z}_{k+1} - \mathbf{z}\rangle \leq \eta_k(h(\mathbf{z}) - h(\mathbf{z}_{k+1}))$ for any $\mathbf{z} \in \mathcal{Z}$. Since $\langle \mathbf{z}_{k+1} - \mathbf{z}_k, \mathbf{z}_{k+1} - \mathbf{z}\rangle = \frac{1}{2}\|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2 + \frac{1}{2}\|\mathbf{z}_{k+1} - \mathbf{z}\|^2 - \frac{1}{2}\|\mathbf{z}_k - \mathbf{z}\|^2$, we can rearrange the terms to get

Downloaded 10/22/25 to 128.62.19.54 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 10/22/25 to 128.62.19.54 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

$$\begin{aligned}(\text{A.1})\qquad &\eta_k(\langle F(\mathbf{z}_{k+1}),\mathbf{z}_{k+1}-\mathbf{z}\rangle + h(\mathbf{z}_{k+1})-h(\mathbf{z}))\\ &\le \frac{\|\mathbf{z}_k-\mathbf{z}\|^2}{2}-\frac{\|\mathbf{z}_{k+1}-\mathbf{z}\|^2}{2}-\frac{\|\mathbf{z}_{k+1}-\mathbf{z}_k\|^2}{2}\le \frac{\|\mathbf{z}_k-\mathbf{z}\|^2}{2}-\frac{\|\mathbf{z}_{k+1}-\mathbf{z}\|^2}{2}.\end{aligned}$$

Proof of part (a). In this part, Assumption 2.3 holds. We first set $\mathbf{z}=\mathbf{z}^*$ in (A.1). Invoking Lemma 2.8 with $\mathbf{z}=\mathbf{z}_{k+1}$ and $\mu=0$, we get $\|\mathbf{z}_{k+1}-\mathbf{z}^*\|\le\|\mathbf{z}_k-\mathbf{z}^*\|$ for any $k\ge 0$. Next, we sum both sides of (A.1) over $k=0,1,\dots,N-1$, which yields $\sum_{k=0}^{N-1}\eta_k(\langle F(\mathbf{z}_{k+1}),\mathbf{z}_{k+1}-\mathbf{z}\rangle+h(\mathbf{z}_{k+1})-h(\mathbf{z}))\le\frac{1}{2}\|\mathbf{z}_0-\mathbf{z}\|^2-\frac{1}{2}\|\mathbf{z}_N-\mathbf{z}\|^2\le\frac{1}{2}\|\mathbf{z}_0-\mathbf{z}\|^2$. Finally, dividing both sides by $\sum_{k=0}^{N-1}\eta_k$ and using Lemma 2.6 complete the proof.

Proof of part (b). In this part, we assume that Assumption 2.4 holds. We set $\mathbf{z}=\mathbf{z}^*$ in (A.1) and apply Lemma 2.8 with $\mathbf{z}=\mathbf{z}_{k+1}$ to get $(1+2\eta_k\mu)\|\mathbf{z}_{k+1}-\mathbf{z}^*\|^2\le\|\mathbf{z}_k-\mathbf{z}^*\|^2$ for any $k\ge 0$. The result follows by induction on the iteration counter k.

Appendix B. Proofs for generalized optimistic method. Our proof relies on a carefully designed Lyapunov function. We define the function and discuss its properties in the following lemma, which is the cornerstone of proving Proposition 4.1.

LEMMA B.1. *Let $\{\mathbf{z}_k\}_{k\ge0}$ be the iterates generated by Algorithm 4.1 with $\mu\ge0$, $0<\alpha\le1$, and $\hat\eta_k$ chosen as in Proposition 4.1. Define a Lyapunov function*

$$(\text{B.1})\qquad V(\mathbf{z}_k,\mathbf{z}_{k-1};\mathbf{z})=-\frac{\eta_{k-1}\langle\mathcal{E}_k,\mathbf{z}_k-\mathbf{z}\rangle}{1+2\eta_{k-1}\mu}+\frac{\|\mathbf{z}-\mathbf{z}_k\|^2}{2}+\frac{\alpha\|\mathbf{z}_k-\mathbf{z}_{k-1}\|^2}{4(1+2\eta_{k-1}\mu)^2},$$

where $\mathcal{E}_k:=F(\mathbf{z}_k)-P(\mathbf{z}_k;\mathcal{I}_{k-1})+\frac{1}{\eta_{k-1}}\mathbf{r}_k$ and $\mathbf{z}\in\mathcal{Z}$ is an arbitrary point. Then for any $\mathbf{z}\in\mathcal{Z}$ and $k\ge0$, we have the following:

(a) $V(\mathbf{z}_0,\mathbf{z}_{-1};\mathbf{z})=\frac{1}{2}\|\mathbf{z}-\mathbf{z}_0\|^2$ *and* $V(\mathbf{z}_k,\mathbf{z}_{k-1};\mathbf{z})\ge\frac{2-\alpha}{4}\|\mathbf{z}-\mathbf{z}_k\|^2$;

(b) $\eta_k(\langle F(\mathbf{z}_{k+1}),\mathbf{z}_{k+1}-\mathbf{z}\rangle+h(\mathbf{z}_{k+1})-h(\mathbf{z})-\mu\|\mathbf{z}-\mathbf{z}_{k+1}\|^2)\le V(\mathbf{z}_k,\mathbf{z}_{k-1};\mathbf{z})-(1+2\eta_k\mu)V(\mathbf{z}_{k+1},\mathbf{z}_k;\mathbf{z})-\frac{1-\alpha}{2}\|\mathbf{z}_{k+1}-\mathbf{z}_k\|^2$.

Proof. By the triangle inequality, we obtain from (4.3) and (4.4) that $\eta_k\|\mathcal{E}_{k+1}\|\le\eta_k\|F(\mathbf{z}_{k+1})-P(\mathbf{z}_{k+1};\mathcal{I}_k)\|+\|\mathbf{r}_{k+1}\|\le\frac{\alpha}{4}\|\mathbf{z}_{k+1}-\mathbf{z}_k\|+\frac{\alpha}{4}\|\mathbf{z}_{k+1}-\mathbf{z}_k\|=\frac{\alpha}{2}\|\mathbf{z}_{k+1}-\mathbf{z}_k\|$. For part (a), note that we initialize Algorithm 4.1 with $\mathbf{z}_{-1}=\mathbf{z}_0$ and $P(\mathbf{z}_0;\mathcal{I}_{-1})=F(\mathbf{z}_0)$. Hence, $V(\mathbf{z}_0,\mathbf{z}_{-1};\mathbf{z})=\frac{1}{2}\|\mathbf{z}-\mathbf{z}_0\|^2$. Moreover, the first term in $V(\mathbf{z}_k,\mathbf{z}_{k-1};\mathbf{z})$ can be lower bounded by $-\frac{\eta_{k-1}\langle\mathcal{E}_k,\mathbf{z}_k-\mathbf{z}\rangle}{1+2\eta_{k-1}\mu}\ge-\frac{\eta_{k-1}\|\mathcal{E}_k\|\|\mathbf{z}_k-\mathbf{z}\|}{1+2\eta_{k-1}\mu}\ge-\frac{\alpha\|\mathbf{z}_k-\mathbf{z}_{k-1}\|\|\mathbf{z}_k-\mathbf{z}\|}{2(1+2\eta_{k-1}\mu)}\ge-\frac{\alpha}{4}\|\mathbf{z}_k-\mathbf{z}\|^2-\frac{\alpha\|\mathbf{z}_k-\mathbf{z}_{k-1}\|^2}{4(1+2\eta_{k-1}\mu)^2}$, where we used Cauchy–Schwarz inequality in the first inequality, $\eta_{k-1}\|\mathcal{E}_k\|\le\frac{\alpha}{2}\|\mathbf{z}_k-\mathbf{z}_{k-1}\|$ in the second inequality, and Young's inequality in the last inequality. Thus, we have $V(\mathbf{z}_k,\mathbf{z}_{k-1};\mathbf{z})=-\frac{\eta_{k-1}\langle\mathcal{E}_{k+1},\mathbf{z}_k-\mathbf{z}\rangle}{1+2\eta_{k-1}\mu}+\frac{\|\mathbf{z}-\mathbf{z}_k\|^2}{2}+\frac{\alpha\|\mathbf{z}_k-\mathbf{z}_{k-1}\|^2}{4(1+2\eta_{k-1}\mu)^2}\ge\frac{2-\alpha}{4}\|\mathbf{z}_k-\mathbf{z}\|^2$. This completes the proof for part (a).

For part (b), by the definition of H in (2.1), the update rule (4.2) implies that $\langle\eta_kP(\mathbf{z}_{k+1};\mathcal{I}_k)-\mathbf{r}_{k+1}+\hat\eta_k\mathcal{E}_k+\mathbf{z}_{k+1}-\mathbf{z}_k,\mathbf{z}_{k+1}-\mathbf{z}\rangle\le\eta_k(h(\mathbf{z})-h(\mathbf{z}_{k+1}))$ for any $\mathbf{z}\in\mathcal{Z}$. Moreover, $\hat\eta_k$ in Proposition 4.1 can be written as $\hat\eta_k=\frac{\eta_{k-1}}{1+2\eta_{k-1}\mu}$ for $\mu\ge0$. By using $\mathcal{E}_{k+1}=F(\mathbf{z}_{k+1})-P(\mathbf{z}_{k+1};\mathcal{I}_k)+\frac{1}{\eta_k}\mathbf{r}_{k+1}$ and rearranging the terms, we obtain

$$\begin{aligned}(\text{B.2})\qquad &\eta_k\Big(\langle F(\mathbf{z}_{k+1}),\mathbf{z}_{k+1}-\mathbf{z}\rangle+h(\mathbf{z}_{k+1})-h(\mathbf{z})-\mu\|\mathbf{z}-\mathbf{z}_{k+1}\|^2\Big)\\ &\le(1+2\eta_k\mu)\left[\frac{\eta_k\langle\mathcal{E}_{k+1},\mathbf{z}_{k+1}-\mathbf{z}\rangle}{1+2\eta_k\mu}-\frac{\|\mathbf{z}-\mathbf{z}_{k+1}\|^2}{2}\right]-\frac{1}{2}\|\mathbf{z}_{k+1}-\mathbf{z}_k\|^2\\ &\quad-\left[\frac{\eta_{k-1}\langle\mathcal{E}_k,\mathbf{z}_k-\mathbf{z}\rangle}{1+2\eta_{k-1}\mu}-\frac{\|\mathbf{z}-\mathbf{z}_k\|^2}{2}\right]+\frac{\eta_{k-1}}{1+2\eta_{k-1}\mu}\langle\mathcal{E}_k,\mathbf{z}_k-\mathbf{z}_{k+1}\rangle.\end{aligned}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 10/22/25 to 128.62.19.54 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

The first two bracketed terms in (B.2) resemble the Lyapunov function in (B.1), and now we upper bound the remaining terms. By Cauchy–Schwarz inequality, the inequality $\eta_{k-1}\|\mathcal{E}_k\| \le \frac{\alpha}{2}\|\mathbf{z}_k - \mathbf{z}_{k-1}\|$, and Young's inequality, we have $\frac{\eta_{k-1}\langle \mathcal{E}_k, \mathbf{z}_k - \mathbf{z}_{k+1}\rangle}{1+2\eta_{k-1}\mu} \le \frac{\eta_{k-1}\|\mathcal{E}_k\|\|\mathbf{z}_{k+1}-\mathbf{z}_k\|}{1+2\eta_{k-1}\mu} \le \frac{\alpha\|\mathbf{z}_k-\mathbf{z}_{k-1}\|^2}{4(1+2\eta_{k-1}\mu)^2} + \frac{\alpha\|\mathbf{z}_{k+1}-\mathbf{z}_k\|^2}{4}$. Thus, we have $\frac{\eta_{k-1}}{1+2\eta_{k-1}\mu}\langle \mathcal{E}_k, \mathbf{z}_k - \mathbf{z}_{k+1}\rangle - \frac{1}{2}\|\mathbf{z}_{k+1}-\mathbf{z}_k\|^2 \le \frac{\alpha\|\mathbf{z}_k-\mathbf{z}_{k-1}\|^2}{4(1+2\eta_{k-1}\mu)^2} - \frac{2-\alpha}{4}\|\mathbf{z}_{k+1}-\mathbf{z}_k\|^2 \le \frac{\alpha\|\mathbf{z}_k-\mathbf{z}_{k-1}\|^2}{4(1+2\eta_{k-1}\mu)^2} - \frac{\alpha\|\mathbf{z}_{k+1}-\mathbf{z}_k\|^2}{4(1+2\eta_k\mu)} - \frac{1-\alpha}{2}\|\mathbf{z}_{k+1}-\mathbf{z}_k\|^2$. Part (b) follows from combining (B.2) and the above inequality. ∎

B.1. Proof of Proposition 4.1. Proof of part (a). In this part, we assume that Assumption 2.3 holds and hence $\mu = 0$. By Lemma 2.8, for any $k \ge 0$ we have $0 \le \eta_k(\langle F(\mathbf{z}_{k+1}), \mathbf{z}_{k+1} - \mathbf{z}^*\rangle + h(\mathbf{z}_{k+1}) - h(\mathbf{z}^*))$. Moreover, by using Lemma B.1(b) with $\mathbf{z} = \mathbf{z}^*$, we further have $0 \le \eta_k(\langle F(\mathbf{z}_{k+1}), \mathbf{z}_{k+1} - \mathbf{z}^*\rangle + h(\mathbf{z}_{k+1}) - h(\mathbf{z}^*)) \le V(\mathbf{z}_k, \mathbf{z}_{k-1}; \mathbf{z}^*) - V(\mathbf{z}_{k+1}, \mathbf{z}_k; \mathbf{z}^*) - \frac{1-\alpha}{2}\|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2 \le V(\mathbf{z}_k, \mathbf{z}_{k-1}; \mathbf{z}^*) - V(\mathbf{z}_{k+1}, \mathbf{z}_k; \mathbf{z}^*)$.

Thus, the Lyapunov function $V(\mathbf{z}_k, \mathbf{z}_{k-1}; \mathbf{z}^*)$ is nonincreasing with respect to k, which implies that $V(\mathbf{z}_N, \mathbf{z}_{N-1}; \mathbf{z}^*) \le V(\mathbf{z}_0, \mathbf{z}_{-1}; \mathbf{z}^*)$. Then by Lemma B.1(a), we get $\frac{2-\alpha}{4}\|\mathbf{z}_N - \mathbf{z}^*\|^2 \le V(\mathbf{z}_N, \mathbf{z}_{N-1}; \mathbf{z}^*) \le V(\mathbf{z}_0, \mathbf{z}_{-1}; \mathbf{z}^*) = \frac{1}{2}\|\mathbf{z}_0 - \mathbf{z}^*\|^2$, which proves that $\|\mathbf{z}_N - \mathbf{z}^*\|^2 \le \frac{2}{2-\alpha}\|\mathbf{z}_0 - \mathbf{z}^*\|^2$ for any $N \ge 0$.

Next, note that $\frac{1-\alpha}{2}\|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2 \ge 0$ and hence from Lemma B.1(b) we get $\eta_k(\langle F(\mathbf{z}_{k+1}), \mathbf{z}_{k+1} - \mathbf{z}\rangle + h(\mathbf{z}_{k+1}) - h(\mathbf{z})) \le V(\mathbf{z}_k, \mathbf{z}_{k-1}; \mathbf{z}) - V(\mathbf{z}_{k+1}, \mathbf{z}_k; \mathbf{z})$ for any $\mathbf{z} \in \mathcal{Z}$ and $k \ge 0$. Summing this inequality over $k = 0, \dots, N-1$, we obtain $\sum_{k=0}^{N-1} \eta_k(\langle F(\mathbf{z}_{k+1}), \mathbf{z}_{k+1} - \mathbf{z}\rangle + h(\mathbf{z}_{k+1}) - h(\mathbf{z})) \le V(\mathbf{z}_0, \mathbf{z}_{-1}; \mathbf{z}) - V(\mathbf{z}_N, \mathbf{z}_{N-1}; \mathbf{z}) \le \frac{1}{2}\|\mathbf{z}_0 - \mathbf{z}\|^2 - \frac{2-\alpha}{4}\|\mathbf{z}_{N-1} - \mathbf{z}\|^2 \le \frac{1}{2}\|\mathbf{z}_0 - \mathbf{z}\|^2$, where we used Lemma B.1(a) in the second inequality. Part (a) follows by similar arguments as in the proof of Proposition 3.1(a).

Proof of part (b). In this part, we assume that Assumption 2.4 holds. We invoke Lemma B.1(b) with $\mathbf{z} = \mathbf{z}^*$ and Lemma 2.8 to get $0 \le \eta_k(\langle F(\mathbf{z}_{k+1}), \mathbf{z}_{k+1} - \mathbf{z}^*\rangle + h(\mathbf{z}_{k+1}) - h(\mathbf{z}^*) - \mu\|\mathbf{z}_{k+1} - \mathbf{z}^*\|^2) \le V(\mathbf{z}_k, \mathbf{z}_{k-1}; \mathbf{z}^*) - (1 + 2\eta_k\mu)V(\mathbf{z}_{k+1}, \mathbf{z}_k; \mathbf{z}^*)$. Hence, $V(\mathbf{z}_{k+1}, \mathbf{z}_k; \mathbf{z}^*) \le V(\mathbf{z}_k, \mathbf{z}_{k-1}; \mathbf{z}^*)(1 + 2\eta_k\mu)^{-1}$ for any $k \ge 0$, and by induction $V(\mathbf{z}_N, \mathbf{z}_{N-1}; \mathbf{z}^*) \le V(\mathbf{z}_0, \mathbf{z}_{-1}; \mathbf{z}^*)\prod_{k=0}^{N-1}(1 + 2\eta_k\mu)^{-1}$. Finally, the claim follows via $V(\mathbf{z}_0, \mathbf{z}_{-1}; \mathbf{z}^*) = \frac{1}{2}\|\mathbf{z}_0 - \mathbf{z}^*\|^2$ and $V(\mathbf{z}_N, \mathbf{z}_{N-1}; \mathbf{z}^*) \ge \frac{2-\alpha}{4}\|\mathbf{z}_N - \mathbf{z}^*\|^2$ (cf. Lemma B.1(a)).

B.2. Proof of Lemma 4.2. As the convex-concave case is a special instance of the strongly-convex-strongly-concave case with $\mu = 0$, we prove Lemma 4.2 under both assumptions uniformly. To simplify the notation, we define the sequence $\{\Gamma_k\}_{k\ge 0}$ by $\Gamma_0 = 1$ and $\Gamma_k = \prod_{l=0}^{k-1}(1 + 2\eta_l\mu)$ for $k \ge 1$. We first apply Lemma B.1(b) with $\mathbf{z} = \mathbf{z}^*$. Together with Lemma 2.8, it implies that $\frac{1-\alpha}{2}\|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2 \le V(\mathbf{z}_k, \mathbf{z}_{k-1}; \mathbf{z}^*) - (1 + 2\eta_k\mu)V(\mathbf{z}_{k+1}, \mathbf{z}_k; \mathbf{z}^*)$. We then multiply both sides of the inequality by Γ_k and note that $\Gamma_k(1 + 2\eta_k\mu) = \Gamma_{k+1}$. Summing over $k = 0, 1, \dots, N-1$, we obtain that $\frac{1-\alpha}{2}\sum_{k=0}^{N-1}\big(\|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2\Gamma_k\big) \le \sum_{k=0}^{N-1}\big(V(\mathbf{z}_k, \mathbf{z}_{k-1}; \mathbf{z}^*)\Gamma_k - V(\mathbf{z}_{k+1}, \mathbf{z}_k; \mathbf{z}^*)\Gamma_{k+1}\big) = V(\mathbf{z}_0, \mathbf{z}_{-1}; \mathbf{z}^*) - V(\mathbf{z}_N, \mathbf{z}_{N-1}; \mathbf{z}^*)\Gamma_N$. Then Lemma 4.2 follows from the facts that $V(\mathbf{z}_0, \mathbf{z}_{-1}; \mathbf{z}^*) = \frac{1}{2}\|\mathbf{z}_0 - \mathbf{z}^*\|^2$ and $V(\mathbf{z}_N, \mathbf{z}_{N-1}; \mathbf{z}^*) \ge 0$ (cf. Lemma B.1(a)).

B.3. Proof of Proposition 4.3. Recall that we define $\mathcal{E}_k := F(\mathbf{z}_k) - P(\mathbf{z}_k; \mathcal{I}_k) + \frac{1}{\eta_{k-1}}\mathbf{r}_k$. By rearranging, the update rule in (4.2) can be equivalently written as $\frac{\mathbf{z}_k - \mathbf{z}_{k+1}}{\eta_k} + \mathcal{E}_{k+1} - \frac{\hat{\eta}_k}{\eta_k}\mathcal{E}_k \in F(\mathbf{z}_{k+1}) + H(\mathbf{z}_{k+1})$. Thus, this implies that $\mathrm{res}(\mathbf{z}_{k+1}) \le \|\frac{\mathbf{z}_k - \mathbf{z}_{k+1}}{\eta_k} + \mathcal{E}_{k+1} - \frac{\hat{\eta}_k}{\eta_k}\mathcal{E}_k\| \le \frac{\|\mathbf{z}_k - \mathbf{z}_{k+1}\|}{\eta_k} + \|\mathcal{E}_{k+1}\| + \frac{\hat{\eta}_k}{\eta_k}\|\mathcal{E}_k\|$. Moreover, we proved in Lemma B.1 that $\eta_k\|\mathcal{E}_{k+1}\| \le \frac{\alpha}{2}\|\mathbf{z}_{k+1} - \mathbf{z}_k\|$ using (4.3) and (4.4), leading to

$$\mathrm{res}(\mathbf{z}_{k+1}) \le \frac{\|\mathbf{z}_k - \mathbf{z}_{k+1}\|}{\eta_k} + \frac{\alpha}{2}\frac{\|\mathbf{z}_{k+1} - \mathbf{z}_k\|}{\eta_k} + \frac{\alpha}{2}\frac{\hat{\eta}_k\|\mathbf{z}_k - \mathbf{z}_{k-1}\|}{\eta_k\eta_{k-1}}. \tag{B.3}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 10/22/25 to 128.62.19.54 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

We first consider the convex-concave setting under Assumption 2.3, where we set $\hat{\eta}_k = \eta_{k-1}$. Hence, we obtain from (B.3) that $\eta_k \mathrm{res}(\mathbf{z}_{k+1}) \leq \frac{\alpha+2}{2}\|\mathbf{z}_k - \mathbf{z}_{k+1}\| + \frac{\alpha}{2}\|\mathbf{z}_k - \mathbf{z}_{k-1}\|$, which further implies that $\eta_k^2 \mathrm{res}(\mathbf{z}_{k+1})^2 \leq \frac{(\alpha+2)(\alpha+1)}{2}\|\mathbf{z}_k - \mathbf{z}_{k+1}\|^2 + \frac{\alpha(\alpha+1)}{2}\|\mathbf{z}_k - \mathbf{z}_{k-1}\|^2$. Summing the above inequality from $k = 0$ to $N-1$ and applying Lemma 4.2, we obtain $\sum_{k=0}^{N-1} \eta_k^2 \mathrm{res}(\mathbf{z}_{k+1})^2 \leq (\alpha+1)\sum_{k=0}^{N-1}(\frac{\alpha+2}{2}\|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2 + \frac{\alpha}{2}\|\mathbf{z}_k - \mathbf{z}_{k-1}\|^2) \leq \frac{(\alpha+1)^2\|\mathbf{z}_0 - \mathbf{z}^*\|^2}{1-\alpha}$. Hence, by lower bounding the left-hand side by $\sum_{k=0}^{N-1} \eta_k^2 \min_{i\in\{1,\ldots,N\}} \mathrm{res}(\mathbf{z}_i)^2$, the statement in part (a) follows.

Next, we consider the strongly-convex-strongly-concave setting under Assumption 2.4, where we set $\hat{\eta}_k = \frac{\eta_{k-1}}{1+2\eta_{k-1}\mu}$. In this case, we obtain from (B.3) that $\mathrm{res}(\mathbf{z}_{k+1}) \leq \frac{\|\mathbf{z}_k - \mathbf{z}_{k+1}\|}{\eta_k} + \frac{\alpha}{2}\frac{\|\mathbf{z}_{k+1} - \mathbf{z}_k\|}{\eta_k} + \frac{\alpha\|\mathbf{z}_k - \mathbf{z}_{k-1}\|}{2\eta_k(1+2\eta_{k-1}\mu)}$. Furthermore, by Lemma 4.2, we have $\|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2 \prod_{l=0}^{k-1}(1+2\eta_l\mu) \leq \frac{1}{1-\alpha}\|\mathbf{z}_0 - \mathbf{z}^*\|^2$, which implies that $\|\mathbf{z}_{k+1} - \mathbf{z}_k\| \leq \frac{\|\mathbf{z}_0 - \mathbf{z}^*\|}{\sqrt{1-\alpha}}\prod_{l=0}^{k-1}(1+2\eta_l\mu)^{-\frac{1}{2}}$. Combining these two bounds, we obtain that $\mathrm{res}(\mathbf{z}_{k+1}) \leq (1+\frac{\alpha}{2})\frac{\|\mathbf{z}_0 - \mathbf{z}^*\|}{\sqrt{1-\alpha}\eta_k}\prod_{l=0}^{k-1}(1+2\eta_l\mu)^{-\frac{1}{2}} + \frac{\alpha\|\mathbf{z}_0 - \mathbf{z}^*\|}{2\sqrt{1-\alpha}\eta_k(1+2\eta_{k-1}\mu)}\prod_{l=0}^{k-2}(1+2\eta_l\mu)^{-\frac{1}{2}} \leq \frac{(1+\alpha)\|\mathbf{z}_0 - \mathbf{z}^*\|}{\sqrt{1-\alpha}\eta_k}\prod_{l=0}^{k-1}(1+2\eta_l\mu)^{-\frac{1}{2}}$. This completes the proof of part (b).

Appendix C. Proof of Lemma 5.3.

We need the following lemma from [40].

LEMMA C.1 (see [40, Lemma 4.3]). *Suppose A is a maximal monotone operator, and let $\mathbf{z}(\eta;\mathbf{w})$ denote the unique solution of the monotone inclusion problem $0 \in \eta A(\mathbf{z}) + \mathbf{z} - \mathbf{w}$. If $0 < \eta < \eta'$, then $\|\mathbf{z}(\eta';\mathbf{w}) - \mathbf{w}\| \leq \frac{\eta'}{\eta}\|\mathbf{z}(\eta;\mathbf{w}) - \mathbf{w}\|$.*

In addition, we use the following lemma to relate an approximate solution of a monotone inclusion problem to its exact solution.

LEMMA C.2. *Suppose A is a maximal monotone operator and let $\mathbf{z}(\eta;\mathbf{w})$ denote the unique solution of the monotone inclusion problem $0 \in \eta A(\mathbf{z}) + \mathbf{z} - \mathbf{w}$. Moreover, suppose $\mathbf{z}$ and $\mathbf{r}$ satisfy $\mathbf{r} \in \eta A(\mathbf{z}) + \mathbf{z} - \mathbf{w}$ with $\|\mathbf{r}\| \leq \frac{\alpha}{4}\|\mathbf{z} - \mathbf{w}\|$. Then we have $\frac{4}{4+\alpha}\|\mathbf{z}(\eta;\mathbf{w}) - \mathbf{w}\| \leq \|\mathbf{z} - \mathbf{w}\| \leq \frac{4}{4-\alpha}\|\mathbf{z}(\eta;\mathbf{w}) - \mathbf{w}\|$.*

Proof. We have $\frac{1}{\eta}(\mathbf{w} - \mathbf{z}(\eta;\mathbf{w})) \in A(\mathbf{z}(\eta;\mathbf{w}))$ and $\frac{1}{\eta}(\mathbf{w} - \mathbf{z} + \mathbf{r}) \in A(\mathbf{z})$. Since $\mathbf{A}$ is monotone, $\langle \frac{1}{\eta}(\mathbf{z} - \mathbf{z}(\eta;\mathbf{w}) - \mathbf{r}), \mathbf{z}(\eta;\mathbf{w}) - \mathbf{z}\rangle \geq 0$, which leads to $\|\mathbf{z}(\eta;\mathbf{w}) - \mathbf{z}\|^2 \leq -\langle \mathbf{r}, \mathbf{z}(\eta;\mathbf{w}) - \mathbf{z}\rangle$ by rearranging. Moreover, using Cauchy–Schwarz inequality and our bound on $\mathbf{r}$, we obtain $-\langle \mathbf{r}, \mathbf{z}(\eta;\mathbf{w}) - \mathbf{z}\rangle \leq \|\mathbf{r}\|\|\mathbf{z}(\eta;\mathbf{w}) - \mathbf{z}\| \leq \frac{\alpha}{4}\|\mathbf{z} - \mathbf{w}\|\|\mathbf{z}(\eta;\mathbf{w}) - \mathbf{z}\|$, and hence this leads to $\|\mathbf{z}(\eta;\mathbf{w}) - \mathbf{z}\| \leq \frac{\alpha}{4}\|\mathbf{z} - \mathbf{w}\|$. Finally, by the triangle inequality, we get $\|\mathbf{z}(\eta;\mathbf{w}) - \mathbf{w}\| \leq \|\mathbf{z}(\eta;\mathbf{w}) - \mathbf{z}\| + \|\mathbf{z} - \mathbf{w}\| \leq (1+\frac{\alpha}{4})\|\mathbf{z} - \mathbf{w}\|$ and $\|\mathbf{z}(\eta;\mathbf{w}) - \mathbf{w}\| \geq \|\mathbf{z} - \mathbf{w}\| - \|\mathbf{z}(\eta;\mathbf{w}) - \mathbf{z}\| \geq (1-\frac{\alpha}{4})\|\mathbf{z} - \mathbf{w}\|$. This completes the proof. ∎

As a corollary of Lemma C.1, we have the following result.

LEMMA C.3. *Suppose $0 < \eta < \eta'$. Let $(\mathbf{z}, \mathbf{r})$ and $(\mathbf{z}', \mathbf{r}')$ be the solution of (4.5) with η and η', respectively. Then, $\|\mathbf{z}' - \mathbf{z}^-\| \leq \frac{(4+\alpha)\eta'}{(4-\alpha)\eta}\|\mathbf{z} - \mathbf{z}^-\| + \big(1 + \frac{(4+\alpha)\eta'}{(4-\alpha)\eta}\big)\|\mathbf{v}^-\|$.*

Proof. The inclusion problem in (4.5) can be cast as $\mathbf{r} \in \eta A(\mathbf{z}) + \mathbf{z} - \mathbf{z}^- + \mathbf{v}^-$, where $A := P + H$ is a maximal monotone operator. Also, recall $\mathbf{z}(\eta;\mathbf{w})$ as the unique solution of $0 \in \eta A(\mathbf{z}) + \mathbf{z} - \mathbf{w}$. Hence, by Lemmas C.1 and C.2, we have $\|\mathbf{z}' - \mathbf{z}^- + \mathbf{v}^-\| \leq \frac{4}{4-\alpha}\|\mathbf{z}(\eta';\mathbf{z}^- - \mathbf{v}^-) - \mathbf{z}^- + \mathbf{v}^-\| \leq \frac{4\eta'}{(4-\alpha)\eta}\|\mathbf{z}(\eta;\mathbf{z}^- - \mathbf{v}^-) - \mathbf{z}^- + \mathbf{v}^-\| \leq \frac{(4+\alpha)\eta'}{(4-\alpha)\eta}\|\mathbf{z} - \mathbf{z}^- + \mathbf{v}^-\|$. Thus, using the triangle inequality we have $\|\mathbf{z}' - \mathbf{z}^-\| \leq \|\mathbf{z}' - \mathbf{z}^- + \mathbf{v}^-\| + \|\mathbf{v}^-\| \leq \frac{(4+\alpha)\eta'}{(4-\alpha)\eta}\|\mathbf{z} - \mathbf{z}^- + \mathbf{v}^-\| + \|\mathbf{v}^-\| \leq \frac{(4+\alpha)\eta'}{(4-\alpha)\eta}\|\mathbf{z} - \mathbf{z}^-\| + (1 + \frac{(4+\alpha)\eta'}{(4-\alpha)\eta})\|\mathbf{v}^-\|$. ∎

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 10/22/25 to 128.62.19.54 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Next, we prove Lemma 5.3. To simplify the notation, we denote $\mathbf{z}_k$ by $\mathbf{z}^-$. Since $k \in \mathcal{B}$, by Algorithm 4.2, (4.6) is violated for $\eta' = \eta/\beta$ and $\mathbf{z}'$ returned by the optimistic subsolver for solving (4.5) with η' (otherwise, the line search scheme would choose η' instead of η). Hence, $\frac{1}{2}\alpha\|\mathbf{z}' - \mathbf{z}^-\| < \eta'\|F(\mathbf{z}') - T^{(1)}(\mathbf{z}'; \mathbf{z}^-)\| \le \frac{\eta' L_2}{2}\|\mathbf{z}' - \mathbf{z}^-\|^2$, where we used (5.1). Hence, $\eta' > \frac{\alpha}{L_2\|\mathbf{z}' - \mathbf{z}^-\|}$, and $\eta = \beta\eta' > \frac{\alpha\beta}{L_2\|\mathbf{z}' - \mathbf{z}^-\|}$. Together with Lemma C.3 and $\eta'/\eta = 1/\beta$, this leads to Lemma 5.3.

Appendix D. Proofs for optimistic subsolver complexity.

D.1. Proof of Lemma 5.16. In the unconstrained setting, our implementation of the optimistic subsolver relies on the CGLS method. Specifically, starting with $\mathbf{s}^{(0)} = 0$, we run CGLS on the linear system $\mathbf{A}\mathbf{s} + \mathbf{b} = 0$. Each iteration requires two matrix-vector products, and the process continues until the termination criterion $\|\mathbf{A}\mathbf{s}^{(k)} + \mathbf{b}\| \le \frac{\alpha}{4}\|\mathbf{s}^{(k)}\|$ is satisfied.

Before proving Lemma 5.16, we first present two useful lemmas. The first lemma characterizes the convergence rate of CGLS, while the second provides a sufficient condition for satisfying the termination criterion.

LEMMA D.1. *Let $\mathbf{s}^{(k)}$ be the kth iterate generated by CGLS, and let $\mathbf{r}^{(k)} = \mathbf{A}\mathbf{s}^{(k)} + \mathbf{b}$. Let $\mathbf{s}^*$ denote the solution of $\mathbf{A}\mathbf{s} + \mathbf{b} = 0$, and recall that $\kappa(\mathbf{A}) = \frac{\|\mathbf{A}\|_{\mathrm{op}}}{\lambda_{\min}(\frac{1}{2}(\mathbf{A}+\mathbf{A}^\top))}$. Then $\|\mathbf{A}\mathbf{s}^{(k)} + \mathbf{b}\| \le \min\{\frac{\sqrt{2}\|\mathbf{A}\|_{\mathrm{op}}\|\mathbf{s}^*\|}{k+1}, 2(\frac{\kappa(\mathbf{A})-1}{\kappa(\mathbf{A})+1})^k\|\mathbf{A}\|_{\mathrm{op}}\|\mathbf{s}^*\|\}$.*

Proof. Under exact arithmetic, CGLS is equivalent to running the conjugate gradient (CG) method on the normal equation $\mathbf{A}^\top\mathbf{A}\mathbf{s} + \mathbf{A}^\top\mathbf{b} = 0$. From the standard analysis of CG (see, e.g., [13, section B.2] and [5, Chapter 7.4.2]), we have these two bounds: $\|\mathbf{A}\mathbf{s}^{(k)} + \mathbf{b}\|^2 \le \frac{2\|\mathbf{A}^\top\mathbf{A}\|_{\mathrm{op}}\|\mathbf{s}^{(0)} - \mathbf{s}^*\|^2}{(k+1)^2}$ and $\|\mathbf{A}\mathbf{s}^{(k)} + \mathbf{b}\| \le 2(\frac{\tilde{\kappa}(\mathbf{A})-1}{\tilde{\kappa}(\mathbf{A})+1})^k\|\mathbf{A}\mathbf{s}^{(0)} + \mathbf{b}\|$, where $\tilde{\kappa}(\mathbf{A}) = \sqrt{\frac{\lambda_{\max}(\mathbf{A}^\top\mathbf{A})}{\lambda_{\min}(\mathbf{A}^\top\mathbf{A})}}$. Note that $\sqrt{\lambda_{\max}(\mathbf{A}^\top\mathbf{A})} = \|\mathbf{A}\|_{\mathrm{op}}$ and $\sqrt{\lambda_{\min}(\mathbf{A}^\top\mathbf{A})} \ge \lambda_{\min}(\frac{1}{2}(\mathbf{A} + \mathbf{A}^\top))$ (see, e.g., [23, Lemma B.1]), which implies $\tilde{\kappa}(\mathbf{A}) \le \kappa(\mathbf{A})$. Since we initialize $\mathbf{s}^{(0)} = 0$ and $\|\mathbf{A}^\top\mathbf{A}\|_{\mathrm{op}} = \|\mathbf{A}\|_{\mathrm{op}}^2$, the first bound simplifies to $\|\mathbf{A}\mathbf{s}^{(k)} + \mathbf{b}\| \le \frac{\sqrt{2}\|\mathbf{A}\|_{\mathrm{op}}\|\mathbf{s}^*\|}{k+1}$. Similarly, with $\mathbf{s}^{(0)} = 0$ and $\|\mathbf{b}\| = \|\mathbf{A}\mathbf{s}^*\| \le \|\mathbf{A}\|_{\mathrm{op}}\|\mathbf{s}^*\|$, the second bound becomes $\|\mathbf{A}\mathbf{s}^{(k)} + \mathbf{b}\| \le 2(\frac{\kappa(\mathbf{A})-1}{\kappa(\mathbf{A})+1})^k\|\mathbf{A}\|_{\mathrm{op}}\|\mathbf{s}^*\|$. This completes the proof. ∎

LEMMA D.2. *Suppose that $\frac{1}{2}(\mathbf{A} + \mathbf{A}^\top) \succeq \mathbf{I}$. Then for a given vector $\mathbf{s} \in \mathbb{R}^d$, $\|\mathbf{A}\mathbf{s} + \mathbf{b}\| \le \frac{\alpha}{\alpha+4}\|\mathbf{s}^*\|$ implies that $\|\mathbf{A}\mathbf{s} + \mathbf{b}\| \le \frac{\alpha}{4}\|\mathbf{s}\|$.*

Proof. Note that $(\mathbf{s} - \mathbf{s}^*)^\top\mathbf{A}(\mathbf{s} - \mathbf{s}^*) = \frac{1}{2}(\mathbf{s} - \mathbf{s}^*)^\top(\mathbf{A} + \mathbf{A}^\top)(\mathbf{s} - \mathbf{s}^*) \ge \|\mathbf{s} - \mathbf{s}^*\|^2$, where the inequality follows from our assumption that $\frac{1}{2}(\mathbf{A} + \mathbf{A}^\top) \succeq \mathbf{I}$. Hence, by Cauchy–Schwarz inequality $\|\mathbf{A}(\mathbf{s} - \mathbf{s}^*)\| \ge \|\mathbf{s} - \mathbf{s}^*\|$, which implies $\|\mathbf{A}\mathbf{s} + \mathbf{b}\| \ge \|\mathbf{s} - \mathbf{s}^*\|$. Combining this with the inequality $\|\mathbf{s}\| \ge \|\mathbf{s}^*\| - \|\mathbf{s} - \mathbf{s}^*\|$ implies $\|\mathbf{s}\| \ge \|\mathbf{s}^*\| - \|\mathbf{A}\mathbf{s} + \mathbf{b}\|$. Now suppose that $\|\mathbf{A}\mathbf{s} + \mathbf{b}\| \le \frac{\alpha}{\alpha+4}\|\mathbf{s}^*\|$. Then we further have $\|\mathbf{s}\| \ge (1 + \frac{4}{\alpha})\|\mathbf{A}\mathbf{s} + \mathbf{b}\| - \|\mathbf{A}\mathbf{s} + \mathbf{b}\| = \frac{4}{\alpha}\|\mathbf{A}\mathbf{s} + \mathbf{b}\|$. Hence, we conclude that $\|\mathbf{A}\mathbf{s} + \mathbf{b}\| \le \frac{\alpha}{4}\|\mathbf{s}\|$. ∎

Next, we prove Lemma 5.16 using Lemmas D.1 and D.2. By Lemma D.1, $\|\mathbf{A}\mathbf{s}^{(k)} + \mathbf{b}\| \le \frac{\alpha}{\alpha+4}\|\mathbf{s}^*\|$ when $k \ge \min\{\frac{\sqrt{2}(\alpha+4)\|\mathbf{A}\|_{\mathrm{op}}}{\alpha} - 1, \kappa(\mathbf{A})\log(\frac{2(\alpha+4)}{\alpha}\|\mathbf{A}\|_{\mathrm{op}})\}$. Since each iteration of CGLS requires two matrix-vector products, the claim follows.

D.2. Proof of Theorem 5.17. First, consider the convex-concave setting. We start by bounding the total number of matrix-vector products at each iteration. Recall that at kth iteration, we solve l_k linear systems using CGLS, where l_k is the number of calls to the optimistic subsolver at the kth iteration and the input matrix is given by $\mathbf{A}_k(\eta) = \eta DF(\mathbf{z}_k) + \mathbf{I}$ for $\eta = \sigma_k, \sigma_k\beta, \ldots, \sigma_k\beta^{l_k - 1}$. Since $\|DF(\mathbf{z}_k)\|_{\mathrm{op}} \le L_1$ by

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 10/22/25 to 128.62.19.54 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

assumption, we have $\|\mathbf{A}_k(\eta)\|_{\text{op}} \le 1 + \eta L_1$. Moreover, since f is convex-concave by Assumption 2.3, the Jacobian matrix $DF(\mathbf{z}_k)$ is postive semidefinite, i.e., $\frac{1}{2}(DF(\mathbf{z}_k) + DF(\mathbf{z}_k)^\top) \succeq 0$. This implies that $\frac{1}{2}(\mathbf{A}_k(\eta) + \mathbf{A}_k(\eta)^\top) \succeq \mathbf{I}$. Hence, applying the first bound in Lemma 5.16, the total number of matrix-vector products at the kth iteration is bounded by $\mathsf{MV}_k \le \sum_{i=1}^{l_k} \frac{2\sqrt{2}(\alpha+4)}{\alpha}(1+\sigma_k\beta^{i-1}L_1) \le \frac{2\sqrt{2}(\alpha+4)}{\alpha} l_k + \frac{2\sqrt{2}(\alpha+4)L_1\sigma_k}{\alpha(1-\beta)}$. Let N_ϵ be the smallest integer N with $\Delta_{\mathcal{B}_1\times\mathcal{B}_2}(\bar{\mathbf{x}}_N, \bar{\mathbf{y}}_N) \le \epsilon$. It follows from Theorem 5.7 that $N_\epsilon = \mathcal{O}((\frac{L_2\|\mathbf{z}_0-\mathbf{z}^*\|D^2}{\epsilon})^{2/3} + (\frac{D^2}{\sigma_0\epsilon})^{1/2})$, where $D = \max_{\mathbf{z}\in\mathcal{B}_1\times\mathcal{B}_2}\|\mathbf{z}_0 - \mathbf{z}\|$. Hence, the total number of matrix-vector products over N_ϵ iterations is bounded by

$$\text{(D.1)} \qquad \mathsf{MV}_{\text{total}} = \sum_{k=0}^{N_\epsilon-1} \mathsf{MV}_k \le \frac{2\sqrt{2}(\alpha+4)}{\alpha} \sum_{k=0}^{N_\epsilon-1} l_k + \frac{2\sqrt{2}(\alpha+4)L_1}{\alpha(1-\beta)} \sum_{k=0}^{N_\epsilon-1} \sigma_k.$$

We now bound the two sums in the above expression separately. For the first sum, we obtain from Theorem 5.15 that $\sum_{k=0}^{N_\epsilon-1} l_k = \mathcal{O}(N_\epsilon)$. For the second sum, recall that $\sigma_k = \frac{\eta_{k-1}}{\beta}$ for $k \ge 1$. By the definition of N_ϵ, we must have $\Delta_{\mathcal{B}_1\times\mathcal{B}_2}(\bar{\mathbf{x}}_{N_\epsilon-1}, \bar{\mathbf{y}}_{N_\epsilon-1}) > \epsilon$. Hence, together with Proposition 4.1(a), we obtain $\epsilon < \Delta_{\mathcal{B}_1\times\mathcal{B}_2}(\bar{\mathbf{x}}_{N_\epsilon-1}, \bar{\mathbf{y}}_{N_\epsilon-1}) \le \frac{D^2}{2\sum_{k=0}^{N_\epsilon-2}\eta_k}$, which leads to $\sum_{k=0}^{N_\epsilon-2}\eta_k \le \frac{D^2}{2\epsilon}$. Hence, we can bound $\sum_{k=0}^{N_\epsilon-1}\sigma_k = \sigma_0 + \sum_{k=0}^{N_\epsilon-2}\frac{\eta_k}{\beta} \le \sigma_0 + \frac{D^2}{2\beta\epsilon}$. Combining both inequalities wtih (D.1), we conclude that $\mathsf{MV}_{\text{total}} = \mathcal{O}(\frac{L_1D^2}{\epsilon} + (\frac{L_2\|\mathbf{z}_0-\mathbf{z}^*\|D^2}{\epsilon})^{2/3} + (\frac{D^2}{\sigma_0\epsilon})^{1/2})$, which completes the proof for part (a).

Next, we move to the strongly-convex-strongly-concave setting. Similar to the arguments above, we first bound the number of matrix-vector products at each iteration. In this setting, we will use the second bound from Lemma 5.16 and hence we first bound the condition number of the matrix $\mathbf{A}_k(\eta) = \eta DF(\mathbf{z}_k) + \mathbf{I}$. Again, by assumption we have $\|\mathbf{A}_k(\eta)\|_{\text{op}} \le 1 + \eta L_1$. Moreover, since f is μ-strongly-convex-strongly-concave by Assumption 2.4, the Jacobian matrix $DF(\mathbf{z}_k)$ satisfies $\frac{1}{2}(DF(\mathbf{z}_k) + DF(\mathbf{z}_k)^\top) \succeq \mu\mathbf{I}$, which implies that $\frac{1}{2}(\mathbf{A}_k(\eta) + \mathbf{A}_k(\eta)^\top) \succeq (1+\eta\mu)\mathbf{I}$. Hence, we obtain $\kappa(\mathbf{A}_k(\eta)) \le \frac{1+\eta L_1}{1+\eta\mu} \le \frac{L_1}{\mu}$. Using the second bound in Lemma 5.16, the number of matrix-vector products at the kth iteration is bounded by $\mathsf{MV}_k \le \sum_{i=1}^{l_k}\frac{2L_1}{\mu}\log(\frac{2(\alpha+4)}{\alpha}(1+\sigma_k\beta^{i-1}L_1)) \le \frac{2L_1l_k}{\mu}\log(\frac{2(\alpha+4)}{\alpha}(1+\sigma_kL_1))$. Moreover, we have $\log(\frac{2(\alpha+4)}{\alpha}(1+\sigma_kL_1)) \le \log(\frac{2(\alpha+4)}{\alpha}(\frac{L_1}{\beta\mu}+\sigma_kL_1)) = \log(\frac{2(\alpha+4)L_1}{\alpha\beta\mu}) + \log(1+\frac{\sigma_k}{\beta}\mu)$. Let N_ϵ be the smallest integer N such that $\|\mathbf{z}_N - \mathbf{z}^*\|^2 \le \epsilon$. Then the total number of matrix-vector products needed to find an ϵ-optimal solution is bounded by

$$\text{(D.2)} \qquad \text{MV}_{\text{total}} = \sum_{k=0}^{N_\epsilon-1} \mathsf{MV}_k \le \frac{2L_1}{\mu}\log\left(\frac{2(\alpha+4)L_1}{\alpha\beta\mu}\right)\sum_{k=0}^{N_\epsilon-1} l_k + \sum_{k=0}^{N_\epsilon-1}\frac{2L_1l_k}{\mu}\log(1+\frac{\sigma_k}{\beta}\mu).$$

By the definition of N_ϵ, we must have $\|\mathbf{z}_{N_\epsilon-1} - \mathbf{z}^*\|^2 > \epsilon$. Hence, together with Proposition 4.1(b), we obtain that $\epsilon < \|\mathbf{z}_{N_\epsilon-1} - \mathbf{z}^*\|^2 < \frac{2}{2-\alpha}\|\mathbf{z}_0 - \mathbf{z}^*\|^2\prod_{k=0}^{N_\epsilon-2}(1+2\eta_k\mu)^{-1}$, which further leads to $\prod_{k=0}^{N_\epsilon-2}(1+2\eta_k\mu) < \frac{2\|\mathbf{z}_0-\mathbf{z}^*\|^2}{(2-\alpha)\epsilon}$. Hence, $\sum_{k=0}^{N_\epsilon-1}\log(1+\frac{\sigma_k}{\beta}\mu) = \log(1+\frac{\sigma_0}{\beta}\mu) + \log\prod_{k=1}^{N_\epsilon-1}(1+\eta_{k-1}\mu) \le \log(1+\frac{\sigma_0}{\beta}\mu) + \log(\frac{2\|\mathbf{z}_0-\mathbf{z}^*\|^2}{(2-\alpha)\epsilon})$. Applying this inequality, we can bound the second sum in (D.2) by $\sum_{k=0}^{N_\epsilon-1}\frac{2L_1l_k}{\mu}\log(1+\frac{\sigma_k}{\beta}\mu) \le \frac{2L_1}{\mu}\sum_{k=0}^{N_\epsilon-1}l_k\sum_{k=0}^{N_\epsilon-1}\log(1+\frac{\sigma_k}{\beta}\mu) \le \frac{2L_1}{\mu}\log\big(\frac{2(\beta+\sigma_0\mu)\|\mathbf{z}_0-\mathbf{z}^*\|^2}{(2-\alpha)\beta\epsilon^2}\big)\sum_{k=0}^{N_\epsilon-1}l_k$. Hence, we conclude that $\text{MV}_{\text{total}} \le \frac{2L_1}{\mu}\log(\frac{4(\beta+\sigma_0\mu)(\alpha+4)L_1\|\mathbf{z}_0-\mathbf{z}^*\|^2}{(2-\alpha)\alpha\beta^2\mu\epsilon^2})\sum_{k=0}^{N_\epsilon-1}l_k$. The result in part (b) follows by bounding $\sum_{k=0}^{N_\epsilon-1}l_k = \mathcal{O}(N_\epsilon)$ using Theorem 5.15.

D.3. Proof of Lemma 5.18. We first consider the convex-concave setting and our implementation is based on the composite FEG method in [7]. Specifically,

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 10/22/25 to 128.62.19.54 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

starting from $\mathbf{s}^{(0)} = 0$, we run the composite FEG method to solve the monotone inclusion problem $0 \in \mathbf{A}\mathbf{s} + \mathbf{b} + \tilde{H}(\mathbf{s})$ until we find $\mathbf{r}^{(k)} \in \mathbf{A}\mathbf{s}^{(k)} + \mathbf{b} + \tilde{H}(\mathbf{s}^{(k)})$ with $\|\mathbf{r}^{(k)}\| \leq \frac{\alpha}{4}\|\mathbf{s}^{(k)}\|$. Note that each iteration of the composite FEG method requires two matrix-vector products with $\mathbf{A}$ and one evaluation of the resolvent of $\tilde{H}$. Similar to the CGLS method in section D.1, we first present a convergence rate result from [7].

LEMMA D.3 (see [7, Theorem 4.1]). *Let $\mathbf{s}^{(k)}$ and $\mathbf{r}^{(k)}$ be the kth iterates generated by composite FEG, and let $\mathbf{s}^*$ denote the solution of $0 \in \mathbf{A}\mathbf{s}+\mathbf{b}+\tilde{H}(\mathbf{s})$. Then we have $\mathbf{r}^{(k)} \in \mathbf{A}\mathbf{s}^{(k)} + \mathbf{b} + \tilde{H}(\mathbf{s}^{(k)})$ for all $k \geq 0$ and $\|\mathbf{r}^{(k)}\| \leq \frac{2\|\mathbf{A}\|_{\mathrm{op}}\|\mathbf{s}^{(0)}-\mathbf{s}^*\|}{k}$.*

Moreover, similar to Lemma D.2, we have the following sufficient condition for satisfying the termination criterion in the composite setting.

LEMMA D.4. *Suppose that $\frac{1}{2}(\mathbf{A}+\mathbf{A}^\top) \succeq \mathbf{I}$ and $\tilde{H}$ is maximal montone. Then for a given vector $\mathbf{s} \in \mathbb{R}^d$ and $\mathbf{r} \in \mathbf{A}\mathbf{s}+\mathbf{b}+\tilde{H}(\mathbf{s})$, $\|\mathbf{r}\| \leq \frac{\alpha}{\alpha+4}\|\mathbf{s}^*\|$ implies that $\|\mathbf{r}\| \leq \frac{\alpha}{4}\|\mathbf{s}\|$.*

Proof. By definition, we have $\mathbf{r} - \mathbf{A}\mathbf{s} - \mathbf{b} \in \tilde{H}(\mathbf{s})$. Since $0 \in \mathbf{A}\mathbf{s}^* + \mathbf{b} + \tilde{H}(\mathbf{s}^*)$, we also have $-\mathbf{A}\mathbf{s}^* - \mathbf{b} \in \tilde{H}(\mathbf{s}^*)$. As $\tilde{H}$ is maximal montone, it follows that $\langle(\mathbf{r} - \mathbf{A}\mathbf{s} - \mathbf{b}) - (-\mathbf{A}\mathbf{s}^* - \mathbf{b}), \mathbf{s} - \mathbf{s}^*\rangle \geq 0$, which leads to $\langle \mathbf{r}, \mathbf{s} - \mathbf{s}^*\rangle \geq \langle \mathbf{s} - \mathbf{s}^*, \mathbf{A}(\mathbf{s} - \mathbf{s}^*)\rangle$. Moreover, note that $(\mathbf{s}-\mathbf{s}^*)^\top\mathbf{A}(\mathbf{s}-\mathbf{s}^*) \geq \|\mathbf{s}-\mathbf{s}^*\|^2$, where we use our assumption that $\frac{1}{2}(\mathbf{A}+\mathbf{A}^\top) \succeq \mathbf{I}$. Hence, by Cauchy–Schwarz inequality, we have $\|\mathbf{r}\| \geq \|\mathbf{s} - \mathbf{s}^*\|$. The rest of the argument follows similarly to the proof of Lemma D.2. ∎

By Lemma D.3, $\|\mathbf{r}^{(k)}\| \leq \frac{\alpha}{\alpha+4}\|\mathbf{s}^*\|$ for $k \geq \frac{2(\alpha+4)\|\mathbf{A}\|_{\mathrm{op}}}{\alpha}$. Combining this with the above lemma and noting that each composite FEG iteration requires two matrix-vector products and one resolvent evaluation of $\tilde{H}$, the claim follows.

Next, we describe our implementation for the strongly-convex-strongly-concave case, based on a restarted variant of composite FEG. We first run the composite FEG initialized with $\mathbf{s}^{(0)} = 0$ for one iteration to compute $\mathbf{s}^{(1)}$ and $\mathbf{r}^{(1)}$. Then the method proceeds with a sequence of epochs. At the kth epoch ($k \geq 1$), the composite FEG method is initialized with $\mathbf{s}^{(k)}$ and $\mathbf{r}^{(k)}$, and it is run until the tangent residual is reduced by half. Moreover, the iterates obtained at the end of the epoch are set to $\mathbf{s}^{(k+1)}$ and $\mathbf{r}^{(k+1)}$ and then used to restart the composite FEG method for the next epoch. We repeat this process until we find $\mathbf{r}^{(k)} \in \mathbf{A}\mathbf{s}^{(k)} + \mathbf{b} + \tilde{H}(\mathbf{s}^{(k)})$ with $\|\mathbf{r}^{(k)}\| \leq \frac{\alpha}{4}\|\mathbf{s}^{(k)}\|$. The following result characterizes the convergence of the restarted composite FEG method.

LEMMA D.5. *Let $\mathbf{s}^{(k)}$ and $\mathbf{r}^{(k)}$ be generated by the restarted composite FEG after k iterations, and let $\mathbf{s}^*$ denote the solution of $0 \in \mathbf{A}\mathbf{s} + \mathbf{b} + \tilde{H}(\mathbf{s})$. Then:* (i) *$\mathbf{r}^{(k)} \in \mathbf{A}\mathbf{s}^{(k)}+\mathbf{b}+\tilde{H}(\mathbf{s}^{(k)})$ for all $k \geq 1$ and $\|\mathbf{r}^{(k)}\| \leq 2^{2-k}\|\mathbf{A}\|_{\mathrm{op}}\|\mathbf{s}^*\|$.* (ii) *Each epoch requires no more than $4\kappa(\mathbf{A})$ inner iterations, where $\kappa(\mathbf{A}) := \frac{\|\mathbf{A}\|_{\mathrm{op}}}{\lambda_{\min}(\frac{1}{2}(\mathbf{A}+\mathbf{A}^\top))}$.*

Proof. First, note that $\mathbf{s}^{(1)}$ and $\mathbf{r}^{(1)}$ correspond to the first iteration of the composite FEG method. By Lemma D.3, we have $\|\mathbf{r}^{(1)}\| \leq \frac{2\|\mathbf{A}\|_{\mathrm{op}}\|\mathbf{s}^{(0)}-\mathbf{s}^*\|}{1} = 2\|\mathbf{A}\|_{\mathrm{op}}\|\mathbf{s}^*\|$. Moreover, based on our restarting condition $\|\mathbf{r}^{(k)}\| \leq \frac{1}{2}\|\mathbf{r}^{(k-1)}\|$ for $k \geq 2$. Hence, the first part of the lemma follows. For the second part, consider the kth epoch, where we run the composite FEG method initialized by $\mathbf{s}_0$ and $\mathbf{r}_0$. By Lemma D.3, $\|\mathbf{r}_t\| \leq \frac{2\|\mathbf{A}\|_{\mathrm{op}}\|\mathbf{s}_0-\mathbf{s}^*\|}{t}$, where t denotes the inner iteration index. Following the same arguments as in Lemma D.4, we have $\|\mathbf{r}_0\| \geq \frac{\langle \mathbf{s}_0-\mathbf{s}^*, \mathbf{A}(\mathbf{s}_0-\mathbf{s}^*)\rangle}{\|\mathbf{s}_0-\mathbf{s}^*\|} \geq \lambda_{\min}(\frac{1}{2}(\mathbf{A}+\mathbf{A}^\top))\|\mathbf{s}_0-\mathbf{s}^*\|$. This implies that $\|\mathbf{r}_t\| \leq \frac{2\kappa(\mathbf{A})}{t}\|\mathbf{r}_0\|$. Hence, after at most $4\kappa(\mathbf{A})$ iterations, we have $\|\mathbf{r}_t\| \leq \frac{1}{2}\|\mathbf{r}_0\|$ and thus the epoch ends. This proves the second part of the lemma. ∎

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

The above result implies $\|\mathbf{r}^{(k)}\| \leq \frac{\alpha}{\alpha+4}\|\mathbf{s}^*\|$ when $k \geq \log_2(\frac{\alpha+4}{\alpha}\|\mathbf{A}\|_{\text{op}})+2$. As each epoch requires at most $4\kappa(\mathbf{A})$ inner iterations, the total number of inner iterations is bounded by $4\kappa(\mathbf{A})\log_2(\frac{4(\alpha+4)}{\alpha}\|\mathbf{A}\|_{\text{op}})$. The proof concludes as each inner iteration requires two matrix-vector products with $\mathbf{A}$ and one resolvent evaluation of $\tilde{H}$.

REFERENCES

[1] D. Adil, B. Bullins, A. Jambulapati, and S. Sachdeva, *Optimal Methods for Higher-Order Smooth Monotone Variational Inequalities*, preprint, https://arxiv.org/abs/2205.06167, 2022.

[2] M. M. Alves and B. F. Svaiter, *A Search-Free $O(1/k^{3/2})$ Homotopy Inexact Proximal-Newton Extragradient Algorithm for Monotone Variational Inequalities*, preprint, https://arxiv.org/abs/2308.05887, 2023.

[3] K. J. Arrow, L. Hurwicz, and H. Uzawa, *Studies in Linear and Non-Linear Programming*, Stanford Mathematical Studies in the Social Sciences, Stanford University Press, Stanford, CA, 1958.

[4] W. Azizian, I. Mitliagkas, S. Lacoste-Julien, and G. Gidel, *A tight and unified analysis of gradient-based methods for a whole spectrum of differentiable games*, in Proceedings of Machine Learning Research (PMLR), 2020, pp. 2863–2873.

[5] Å. Björck, *Numerical Methods for Least Squares Problems*, SIAM, Philadelphia, 1996, https://doi.org/10.1137/1.9781611971484.

[6] B. Bullins and K. A. Lai, *Higher-order methods for convex-concave min-max optimization and monotone variational inequalities*, SIAM J. Optim., 32 (2022), pp. 2208–2229, https://doi.org/10.1137/21M1395764.

[7] Y. Cai, A. Oikonomou, and W. Zheng, *Accelerated algorithms for constrained nonconvex-nonconcave min-max optimization and comonotone inclusion*, in Proceedings of the 41st International Conference on Machine Learning, 2024, pp. 5312–5347.

[8] Y. Cai and W. Zheng, *Accelerated Single-Call Methods for Constrained Min-Max Optimization*, preprint, https://arxiv.org/abs/2210.03096, 2022.

[9] Y. Carmon, D. Hausler, A. Jambulapati, Y. Jin, and A. Sidford, *Optimal and adaptive Monteiro-Svaiter acceleration*, in Proceedings of the 36th International Conference on Neural Information Processing Systems, 2022, pp. 20338–20350.

[10] A. Chambolle and T. Pock, *A first-order primal-dual algorithm for convex problems with applications to imaging*, J. Math. Imaging Vision, 40 (2011), pp. 120–145, https://doi.org/10.1007/s10851-010-0251-1.

[11] C.-K. Chiang, T. Yang, C.-J. Lee, M. Mahdavi, C.-J. Lu, R. Jin, and S. Zhu, *Online optimization with gradual variations*, in Proceedings of the 25th Annual Conference on Learning Theory, 2012, 23.

[12] C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng, *Training GANs with optimism*, in Proc. ICLR 2018.

[13] A. d'Aspremont, D. Scieur, and A. Taylor, *Acceleration methods*, Found. Trends Optim., 5 (2021), pp. 1–245, https://doi.org/10.1561/2400000036.

[14] J. Eckstein and M. C. Ferris, *Operator-splitting methods for monotone affine variational inequalities, with a parallel application to optimal control*, INFORMS J. Comput., 10 (1998), pp. 218–235, https://doi.org/10.1287/ijoc.10.2.218.

[15] F. Facchinei and J.-S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems*, Springer-Verlag, New York, 2003.

[16] G. Gidel, H. Berard, G. Vignoud, P. Vincent, and S. Lacoste-Julien, *A variational inequality perspective on generative adversarial networks*, in Proc. ICLR 2019.

[17] E. Y. Hamedani and N. S. Aybat, *A primal-dual algorithm with line search for general convex-concave saddle point problems*, SIAM J. Optim., 31 (2021), pp. 1299–1329, https://doi.org/10.1137/18M1213488.

[18] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Res. Natl. Bur. Stand., 49 (1952), pp. 409–436, https://doi.org/10.6028/jres.049.044.

[19] Y.-G. Hsieh, F. Iutzeler, J. Malick, and P. Mertikopoulos, *On the convergence of single-call stochastic extra-gradient methods*, in Proc. NeurIPS, 2019, pp. 6938–6948.

[20] K. Huang, J. Zhang, and S. Zhang, *Cubic regularized Newton method for the saddle point models: A global and local convergence analysis*, J. Sci. Comput., 91 (2022), pp. 1–31, https://doi.org/10.1007/s10915-022-01819-6.

[21] K. Huang and S. Zhang, *An Approximation-Based Regularized Extra-Gradient Method for Monotone Variational Inequalities*, preprint, https://arxiv.org/abs/2210.04440, 2022.

Downloaded 10/22/25 to 128.62.19.54 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

[22] R. Jiang and A. Mokhtari, *Generalized Optimistic Methods for Convex-Concave Saddle Point Problems*, preprint, https://arxiv.org/abs/2202.09674, 2022.
[23] R. Jiang and A. Mokhtari, *Online Learning Guided Quasi-Newton Methods with Global Non-Asymptotic Convergence*, preprint, https://arxiv.org/abs/2410.02626, 2024.
[24] G. Korpelevich, *The extragradient method for finding saddle points and other problems*, Ekon. Mat. Metody, 12 (1976), pp. 747–756, https://cs.uwaterloo.ca/∼y328yu/classics/extragrad.pdf.
[25] G. Kotsalis, G. Lan, and T. Li, *Simple and optimal methods for stochastic variational inequalities,* I*: Operator extrapolation*, SIAM J. Optim., 32 (2022), pp. 2041–2073, https://doi.org/10.1137/20M1381678.
[26] D. Kovalev and A. Gasnikov, *The first optimal acceleration of high-order methods in smooth convex optimization*, in Proceedings of the 36th International Conference on Neural Information Processing System, 2022, pp. 35339–35351.
[27] S. Lee and D. Kim, *Fast extra gradient methods for smooth structured nonconvex-nonconcave minimax problems*, in Proceedings of the 35th International Conference on Neural Information Processing System, 2021, pp. 22588–22600.
[28] T. Liang and J. Stokes, *Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks*, in Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS), 2019, pp. 907–915.
[29] T. Lin and M. Jordan, *A Continuous-Time Perspective on Global Acceleration for Monotone Equation Problems*, preprint, https://arxiv.org/abs/2206.04770, 2022.
[30] T. Lin and M. I. Jordan, *Perseus: A simple and optimal high-order method for variational inequalities*, Math. Program., 209 (2025), pp. 609–650, https://doi.org/10.1007/s10107-024-02075-2.
[31] T. Lin, P. Mertikopoulos, and M. Jordan, *Explicit second-Order Min-Max Optimization Methods with Optimal Convergence Guarantee*, preprint, https://arxiv.org/abs/2210.12860, 2022.
[32] Y. Malitsky, *Projected reflected gradient methods for monotone variational inequalities*, SIAM J. Optim., 25 (2015), pp. 502–520, https://doi.org/10.1137/14097238X.
[33] Y. Malitsky, *Proximal extrapolated gradient methods for variational inequalities*, Optim. Methods Softw., 33 (2017), pp. 140–164, https://doi.org/10.1080/10556788.2017.1300899.
[34] Y. Malitsky and M. K. Tam, *A forward-backward splitting method for monotone inclusions without cocoercivity*, SIAM J. Optim., 30 (2020), pp. 1451–1472, https://doi.org/10.1137/18M1207260.
[35] B. Martinet, *Brève communication. régularisation d'inéquations variationnelles par approximations successives*, ESAIM: Math. Model. Numer. Anal., 4 (1970), pp. 154–158, https://doi.org/10.1051/m2an/197004R301541.
[36] P. Mertikopoulos, B. Lecouat, H. Zenati, C.-S. Foo, V. Chandrasekhar, and G. Piliouras, *Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile*, in Proc. ICLR 2018.
[37] A. Mokhtari, A. Ozdaglar, and S. Pattathil, *A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems*, in Proximal Point Approach, in Proc. AISTATS, 2020, pp. 1497–1507..
[38] A. Mokhtari, A. E. Ozdaglar, and S. Pattathil, *Convergence rate of $\mathcal{O}(1/k)$ for optimistic gradient and extragradient methods in smooth convex-concave saddle point problems*, SIAM J. Optim., 30 (2020), pp. 3230–3251, https://doi.org/10.1137/19M127375X.
[39] R. D. C. Monteiro and B. F. Svaiter, *On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean*, SIAM J. Optim., 20 (2010), pp. 2755–2787, https://doi.org/10.1137/090753127.
[40] R. D. C. Monteiro and B. F. Svaiter, *Iteration-complexity of a Newton proximal extragradient method for monotone variational inequalities and inclusion problems*, SIAM J. Optim., 22 (2012), pp. 914–935, https://doi.org/10.1137/11083085X.
[41] A. Nemirovski, *Prox-method with rate of convergence O(1/t) for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems*, SIAM J. Optim., 15 (2004), pp. 229–251, https://doi.org/10.1137/S1052623403425629.
[42] Y. Nesterov, *Cubic Regularization of Newton's Method for Convex Problems with Constraints*, CORE Discussion Paper 2006/39, 2006, https://doi.org/10.2139/ssrn.921825.
[43] Y. Nesterov, *Dual extrapolation and its applications to solving variational inequalities and related problems*, Math. Program., 109 (2007), pp. 319–344, https://doi.org/10.1007/s10107-006-0034-z.

Downloaded 10/22/25 to 128.62.19.54 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

[44] Y. Nesterov, *Accelerating the cubic regularization of Newton's method on convex problems*, Math. Program., 112 (2008), pp. 159–181, https://doi.org/10.1007/s10107-006-0089-x.

[45] Y. Nesterov and B. Polyak, *Cubic regularization of Newton method and its global performance*, Math. Program., 108 (2006), pp. 177–205, https://doi.org/10.1007/s10107-006-0706-8.

[46] P. Ostroukhov, R. Kamalov, P. Dvurechensky, and A. Gasnikov, *Tensor Methods for Strongly Convex Strongly Concave Saddle Point Problems and Strongly Monotone Variational Inequalities*, preprint, https://arxiv.org/abs/2012.15595, 2020.

[47] C. C. Paige and M. A. Saunders, *LSQR: An algorithm for sparse linear equations and sparse least squares*, ACM Trans. Math. Softw., 8 (1982), pp. 43–71, https://doi.org/10.1145/355984.355989.

[48] L. D. Popov, *A modification of the Arrow-Hurwicz method for search of saddle points*, Math. Notes Acad. Sci. USSR, 28 (1980), pp. 845–848, https://doi.org/10.1007/BF01141092.

[49] A. Rakhlin and K. Sridharan, *Online learning with predictable sequences*, in Proceedings of COLT, 2013, pp. 993–1019.

[50] S. Rakhlin and K. Sridharan, *Optimization, learning, and games with predictable sequences*, in Proc. NIPS, 2013, pp. 3066–3074.

[51] R. T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877–898, https://doi.org/10.1137/0314056.

[52] M. V. Solodov and B. F. Svaiter, *A hybrid approximate extragradient – proximal point algorithm using the enlargement of a maximal monotone operator*, Set-Valued Anal., 7 (1999), pp. 323–345, https://doi.org/10.1023/A:1008777829180.

[53] B. F. Svaiter, *Complexity of the Relaxed Hybrid Proximal-Extragradient Method under the Large-Step Condition*, preprint, https://arxiv.org/abs/2303.04972, 2023.

[54] T. Yoon and E. K. Ryu, *Accelerated algorithms for smooth convex-concave minimax problems with $O(1/k^2)$ rate on squared gradient norm*, in Proceedings of the 38th International Conference on Machine Learning, 2021, pp. 12098–12109.

[55] Y. Zhang and X. Lin, *Stochastic primal-dual coordinate method for regularized empirical risk minimization*, in Proceedings of the 32nd International Conference on Machine Learning, 2015, pp. 353–361.

Downloaded 10/22/25 to 128.62.19.54 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.