Generalized Optimistic Methods for Convex-Concave Saddle Point Problems

Ruichen Jiang*  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 [MOP20a] 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 and can work with arbitrary norms with compatible Bregman distances. Moreover, we also develop an adaptive line search scheme to select the stepsizes without knowledge of the smoothness coefficients. We instantiate our method with first-order, second-order and higher-order oracles and give sharp global iteration complexity bounds. When the objective function is convex-concave, we show that the averaged iterates of our $p$-th-order method ($p \geq 1$) converge at a rate of $O(1/N^{p+1})$.

When the objective function is further strongly-convex-strongly-concave, we prove a complexity bound of $O(L_1/\mu \log \frac{1}{\epsilon})$ for our first-order method and a bound of $O((L_p D/p + \log \log \frac{1}{\epsilon})^{\frac{1}{p+1}})$ for our $p$-th-order method ($p \geq 2$) respectively, where $L_p$ ($p \geq 1$) is the Lipschitz constant of the $p$-th-order derivative, $\mu$ is the strongly-convex parameter, and $D$ is the initial Bregman distance to the saddle point. Moreover, our line search scheme provably only requires an almost constant number of calls to a subproblem solver per iteration on average, making our first-order and second-order methods particularly amenable to implementation.

*Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX, USA
{rjiang@utexas.edu, mokhtari@austin.utexas.edu}
# Contents

1 Introduction .................. 4
  1.1 Related Work .................. 6
  1.2 Outline .................. 9

2 Preliminaries .................. 9
  2.1 Bregman Distances ................. 10
  2.2 Monotone Operators ................. 10
  2.3 Lipschitz and Smooth Operators ................. 11
  2.4 Saddle Point Problems ................. 12

3 Bregman Proximal Point Method and Its Approximations .......... 14
  3.1 Approximations of the BPP Method ................. 15

4 Main algorithm: the Generalized Optimistic Method .......... 17
  4.1 Adaptive Line Search ................. 19

5 First-order Generalized Optimistic Method .......... 25
  5.1 Fixed Stepsize Scheme ................. 26
  5.2 Adaptive Line Search Scheme ................. 27

6 Second-Order Generalized Optimistic Method .......... 30
  6.1 Intermediate Results ................. 31
  6.2 Convergence Analysis: Convex-Concave Case ................. 33
  6.3 Convergence Analysis: Strongly-Convex-Strongly-Concave Case ................. 34
  6.4 Complexity Bound of the Line Search Scheme ................. 38

7 Higher-Order Generalized Optimistic Method .......... 40
  7.1 Convergence Analysis: Convex-Concave Case ................. 42
  7.2 Convergence Analysis: Strongly-Convex-Strongly-Concave Case ................. 43
  7.3 Complexity Bound of the Line Search Scheme ................. 45

8 Numerical Experiments .......... 46
  8.1 First-Order Optimistic Method ................. 46
  8.2 Second-Order Optimistic Method ................. 49

9 Conclusions and Final Remarks .......... 52
# A Proofs for Preliminary Lemmas

A.1 Proof of Lemma 2.2 .......................................................... 58  
A.2 Proof of Lemma 2.3 .......................................................... 58  
A.3 Proof of Lemma 2.4 .......................................................... 59

# B Proof of Theorem 3.1

# C Proofs for Generalized Optimistic Method

C.1 Proof of Proposition 4.1 ..................................................... 61  
C.2 Proof of Lemma 4.2 .......................................................... 62

# D Proofs for the Second-Order Optimistic Method

D.1 Proof of Lemma 6.1 .......................................................... 63

# E Proofs for the Higher-Order Optimistic Method

E.1 Proof of Lemma 7.1 .......................................................... 65  
E.2 Intermediate Results ......................................................... 66  
E.3 Proof of Theorem 7.2 .......................................................... 68  
E.4 Proof of Lemma 7.3 .......................................................... 68  
E.5 Proof of Theorem 7.4 .......................................................... 69  
E.6 Proof of Theorem 7.5 .......................................................... 69  
E.7 Proof of Corollary 7.6 ......................................................... 69  
E.8 Proof of Theorem 7.7 .......................................................... 70
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 aim to solve

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

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 \( \text{dom} f \subset \mathbb{R}^m \times \mathbb{R}^n \) containing \( \mathcal{X} \times \mathcal{Y} \). Throughout the 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 \( \ell \) is also convex-concave.

The general form in (1), which we refer to as the composite saddle point problem, encompasses several important special cases studied in the literature. For instance, when \( h_1 \equiv 0 \) on \( \mathcal{X} \) and \( h_2 \equiv 0 \) on \( \mathcal{Y} \), Problem (1) becomes the constrained smooth saddle point problem [Kor76; Pop80; Nem04]. If we further have \( \mathcal{X} = \mathbb{R}^m \) and \( \mathcal{Y} = \mathbb{R}^n \), then it reduces to the unconstrained smooth saddle point problem [DISZ18; MOP20a; MOP20b]. Problem (1) is of central importance in the duality theory of constrained optimization. It also naturally arises in a game-theoretic context, with applications including zero-sum games [BO98], robust optimization [BGN09], and generative adversarial networks (GANs) [GPMXWOCB14].

Over the last few decades, various iterative methods—mostly first-order methods—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 [Pop80] and have gained much attention recently in the machine learning community [CYLMLJZ12; RS13; DISZ18; LS19; GBVVL19; PDZC20]. For constrained smooth strongly-convex strongly-concave problems (where \( f \) is smooth and strongly-convex-strongly-concave), the result in [GBVVL19] showed that the Popov’s method converges linearly and finds an \( \epsilon \)-accurate solution with an iteration complexity bound of \( O(\kappa_1 \log(1/\epsilon)) \), where \( \kappa_1 \) is the first-order condition number we shall define in Section 1.1. Moreover, a variant of the Popov’s method proposed in [MT20] for monotone inclusion problems is proved to achieve the same complexity bound in the more general composite setting. For unconstrained smooth convex-concave problems, the Popov’s method, also more commonly known as the optimistic gradient descent-ascent (OGDA) method, is shown to achieve the complexity bound of \( O(1/\epsilon) \) in terms of the primal-dual gap [MOP20b]. This makes the first-order optimistic method an attractive alternative to the classical extragradient method [Kor76; Nem04], since it only requires one gradient computation (instead of two) per iteration while enjoying the same convergence rates.

In this paper, we follow and extend the proximal point approach in [MOP20a; MOP20b] by interpreting the optimistic method as an approximation of the proximal point method [Mar70; Roc76]. In particular, we propose a generalized optimistic method (see Algorithm 1), where the future gradient required in the update of the proximal point method 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. Under suitable conditions on the stepsizes, we prove that our method indeed enjoys similar convergence guarantees as the proximal point method in both the convex-concave and strongly-convex-strongly-concave settings (see Proposition 4.1). In our framework, the existing first-order optimistic method corresponds to using the current gradient as the prediction term, whereas our
theory allows general prediction terms in the setup of arbitrary norms and compatible Bregman distances.

Moreover, to find proper stepsizes that satisfy the condition specified by the generalized optimistic method, we further develop an adaptive line search scheme (see Algorithm 2) to choose the stepsizes according to the local geometry of the current iterate and without knowledge of the smoothness coefficients of the objective. It first quickly identifies a coarse interval containing the desired stepsizes, and then locates the stepsizes by performing a logarithm-type bisection search. We note that our line search scheme in general can be viewed as a faster alternative to the classical backtracking line search scheme [NW06], which may be of independent interest.

By instantiating our method with different oracles, we obtain the first-order, second-order and higher-order optimistic methods for the saddle point problem in (1). We give sharp complexity bounds of these methods in both the convex-concave and strongly-convex-strongly-concave settings, all matching the best existing upper bounds in their corresponding problem classes (see Section 1.1 for detailed comparisons). The complexity of our line search scheme is also rigorously analyzed in terms of the number of calls to a subproblem solver. Our theoretical findings are summarized as follows:

(a) When the smooth component \( f \) of the objective function in (1) has Lipschitz continuous gradient, our first-order optimistic method generalizes the OGDA method in [MOP20a; MOP20b] which only focuses on unconstrained smooth saddle point problems. We prove a complexity bound of \( O(1/\epsilon) \) in terms of the primal-dual gap in the convex-concave setting, and prove linear convergence with a complexity bound of \( O(\kappa_1 \log(1/\epsilon)) \) in terms of the distance to the optimal solution in the strongly-convex-strongly-concave setting. In addition, by incorporating our line search scheme, we obtain an adaptive first-order optimistic method with the same convergence rates while only requiring a constant number of calls to a subproblem solver per iteration on average.

(b) When \( f \) has Lipschitz continuous Hessian, we obtain an adaptive second-order optimistic method where the stepsizes are chosen by our line search scheme. In the convex-concave setting, we show that it improves upon the first-order method and achieves a complexity bound of \( O(\epsilon^{-\frac{2}{3}}) \) in terms of the primal-dual gap. In the strongly-convex-strongly-concave setting, we prove the global convergence and a local R-superlinear convergence rate of the order \( \frac{3}{2} \) for our method, leading to an overall complexity bound of \( O((\kappa_2(z_0))^{\frac{3}{2}} + \log \log \frac{1}{\epsilon}) \) in terms of the distance to the optimal solution (here \( \kappa_2(z_0) \) is the second-order condition number we shall define in (2)). Also, we prove that the line search procedure on average only requires an almost constant number of calls to a subproblem solver per iteration.

(c) When \( f \) has Lipschitz \( p \)-th-order derivative with \( p \geq 3 \), we further extend the results above and propose an adaptive \( p \)-th-order optimistic method. In the convex-concave setting, the complexity bound is improved to \( O(\epsilon^{-\frac{p}{p+1}}) \). In the strongly-convex-strongly-concave setting, we prove the global convergence and a local R-superlinear convergence rate of the order \( \frac{p+1}{2} \), leading to an overall complexity bound of \( O((\kappa_p(z_0))^{\frac{p}{p+1}} + \log \log \frac{1}{\epsilon}) \) (here \( \kappa_p(z_0) \) is the \( p \)-th-order condition number we shall define in (2)). Similarly, the line search procedure on average only requires an almost constant number of calls to a subproblem solver per iteration.
1.1 Related Work

In the following, we review some iterative methods for convex-concave saddle point problems, with an emphasis on their corresponding iteration complexities to find an $\epsilon$-accurate saddle point. It is worth noting that most existing methods are developed by reformulating the saddle point problem in (1) as a monotone variational inequality and/or inclusion problem (see Section 2.4), which is also the approach we take to derive our generalized optimistic method. Therefore, in the following, we will also include methods for solving this broader class of problems.

For ease of exposition, in the strongly-convex-strongly-concave setting we define the first-order condition number by

$$\kappa_1 := \frac{L_1}{\mu},$$

where $L_1$ is the Lipschitz constant of the gradient of $f$ (cf. Assumption 5.1) and $\mu$ is the strongly-convex parameter (cf. Assumption 2.2). More generally, for $p \geq 2$, we define the $p$-th-order condition number by

$$\kappa_p(z_0) = \frac{L_p(D_\Phi(z^*, z_0))^{p-1}}{\mu},$$

where $L_p$ is the Lipschitz constant of the $p$-th-order derivative of $f$ (cf. Assumptions 6.1 and 7.1) and $D_\Phi(z^*, z_0)$ is the Bregman distance between the optimal solution $z^* = (x^*, y^*)$ and the initial point $z_0 = (x_0, y_0)$ (see Section 2 for formal definitions). Intuitively, these quantities measure the hardness of solving the saddle point problems and they naturally appear in the complexity bounds of the first-order, second-order and higher-order methods respectively.

**First-order methods.** One of the earliest and simplest methods for solving constrained saddle point problems is the Arrow-Hurwicz method [AHU58], also known as gradient descent ascent. It is a natural generalization of the projected gradient method for minimax optimization problems and only requires one gradient computation per iteration. However, this method could fail to converge when the function $f$ in (1) is smooth and only convex-concave, and only achieves a suboptimal iteration complexity of $O(\kappa_1^2 \log(1/\epsilon))$ when $f$ is smooth and strongly-convex-strongly-concave [NS11; FOP20]. To remedy these issues, Korplevich [Kor76] and Popov [Pop80] proposed two different modifications by introducing “extrapolation” and “optimism” into the Arrow-Hurwicz method, respectively.

The Korplevich’s extragradient method [Kor76] solves constrained saddle point problems by modifying the Arrow-Hurwicz method with an extrapolation step. For smooth and strongly-convex-strongly-concave objective functions, it is proved to converge linearly and find an $\epsilon$-accurate saddle point within $O(\kappa_1 \log(1/\epsilon))$ iterations [Tse95; AMLG20; MOP20a]. The mirror-prox method [Nem04] generalizes the extragradient method and works with a general Bregman distance, achieving the ergodic iteration complexity of $O(1/\epsilon)$ for smooth monotone inequalities and smooth convex-concave saddle point problems with compact feasible sets. Subsequently, it was extended to a class of distance-generating functions by [AT05], to unbounded feasible sets by [MS12], and to composite objectives by [Tse08; MS11; HJN15], all with the same rate of convergence. Along another line of research, the dual extrapolation method [Nes07] was developed for variational inequalities and saddle point problems, where the extrapolation step is performed in the dual space. It is also shown to achieve the $O(1/\epsilon)$ complexity bound for smooth monotone inequalities and smooth convex-concave saddle point problems. Moreover, for smooth strongly monotone variational inequalities, it is proved

---

Footnote 1: We may regard composite convex-concave saddle point problems as a subclass of monotone inclusion problems, and similarly constrained convex-concave saddle point problems as a subclass of variational inequality problems.

Footnote 2: It corresponds to the projection method for variational inequalities, and the forward-backward method for monotone inclusion problems (see, e.g., [FP03]).
to achieve the optimal iteration complexity of $O(\kappa_1 \log(1/\epsilon))$ when combined with the technique of estimate functions [NS11]. Finally, Tseng [Tse00] utilized the similar extrapolation idea to propose a splitting method for monotone inclusion problems and proved that the method converges linearly and achieves the iteration complexity of $O(\kappa_1 \log(1/\epsilon))$ for smooth strongly monotone inclusion problems. Later, a variant of Tseng’s splitting method is shown to also achieve the iteration complexity of $O(1/\epsilon)$ for smooth monotone inclusion problems and smooth convex-concave saddle point problems by using the hybrid proximal extragradient (HPE) framework [MS11].

One defining characteristic (or drawback) of the above methods is that they need to maintain two intertwined sequences of updates at which the gradients are computed and solve two projection subproblems per iteration. In contrast, the Popov’s method [Pop80], whose special case is also known as the optimistic gradient descent ascent (OGDA) [DISZ18; MOP20a; MOP20b], only requires one gradient computation for each iteration. It is shown to have a complexity of $O(\kappa_1 \log(1/\epsilon))$ for constrained saddle point problems when the objective is smooth and strongly-convex-strongly-concave [GBVVL19; AMLG20; MOP20a]; see also a similar result in [MT20] for a variant of Popov’s method on monotone inclusion problems. Moreover, by viewing the OGDA method as an approximation of the proximal point method, the authors in [MOP20b] proved an ergodic complexity of $O(1/\epsilon)$ for unconstrained saddle point problems when the objective is smooth and only convex-concave, matching the convergence rate of the extragradient method. Several other variants of the Popov’s method have also appeared in the literature [Mal15; Mal17], and we refer the readers to [HIMM19] for more detailed discussions. In a recent concurrent work [KLL20] on variational inequalities, the authors essentially extended the method in [MT20] to a general Bregman distance function and proved the same convergence guarantees as above for both the convex-concave and strongly-convex-strongly-concave settings. We note that their method is similar to our first-order optimistic method in Section 5, though proposed from a different perspective. In this paper, we recover and further extend their results to composite objectives under our unified framework.

There is also a popular family of algorithms in the literature, known as the primal-dual method [CP11; Con13; CLO14; CP16; MP18; HA21], that is more tailored to the structure of the saddle point problem by updating the $x$ and $y$ variables alternatively. They are mostly designed for bilinear saddle point problems where the function $f$ in (1) is given by $f(x, y) = f_1(x) + \langle Ax, y \rangle$, with $f_1$ being a smooth convex function and $A$ being an $n$-by-$m$ matrix. A notable exception is the recent work in [HA21], which generalized the original primal-dual method in [CP16] for solving the general composite saddle point problem as in (1). It was shown to achieve the iteration complexity of $O(1/\epsilon)$ in terms of the primal-dual gap when $f$ is smooth and convex-concave, and a better complexity of $O(1/\sqrt{\epsilon})$ when $f$ is smooth and strongly-convex-linear\footnote{A function $f(x, y)$ is called strongly-convex-linear if $f(\cdot, y)$ is strongly convex for each fixed $y$ and $f(x, \cdot)$ is linear for each fixed $x$.}. Some of the existing results as well as ours are summarized in Table 1.

**Second-order and higher-order methods.** Unlike first-order methods, relatively few methods have been proposed to leverage second-order and higher-order information in saddle point problems. So far, there have been two separate approaches in the literature: one is to generalize the well-studied second-order methods for optimization problems, in particular the cubic regularized Newton’s method [NP06; Nes08]; and the other is to generalize the first-order extragradient method [Kor76; Nem04]. In this paper, we offer a third alternative with comparable and sometimes better convergence properties.

A celebrated technique of globalizing Newton’s method for unconstrained optimization problems
Table 1: Summary of first-order methods for saddle point problems. The third column indicates whether the method can handle general convex functions \(h_1\) and \(h_2\) in (1) or requires \(h_1 \equiv 0\) and \(h_2 \equiv 0\). In the fourth column, “C-C” stands for “convex-concave” and “SC-SC” stands for “strongly-convex-strongly-concave”. In the fifth column, “\(L_1\)” means \(\nabla f\) is \(L_1\)-Lipschitz. The sixth column indicates whether the saddle point problem is constrained or unconstrained. The seventh column shows whether the method can work with a general Bregman distance or only the Euclidean norm.

| Frameworks     | Methods                      | Composite objective? | Assumptions       | Constrained? | Distance | Complexity               |
|---------------|------------------------------|----------------------|-------------------|--------------|----------|--------------------------|
| Extragradient | Mirror-prox [Tse08]          | ✓                    | C-C               | \(L_1\)      | ✓        | Bregman \(O(e^{-1})\)    |
|               | Extragradient [Tse95]        | ×                    | SC-SC             | \(L_1\)      | ✓        | Euclidean \(O(\kappa_1 \log e^{-1})\) |
|               | Dual extrapolation [Nes07]   | ×                    | C-C               | \(L_1\)      | ✓        | Bregman \(O(e^{-1})\)    |
|               | Dual extrapolation [NS11]    | ×                    | SC-SC             | \(L_1\)      | ✓        | Euclidean \(O(\kappa_1 \log e^{-1})\) |
|               | Tseng’s splitting [MS11]     | ✓                    | C-C               | \(L_1\)      | ✓        | Euclidean \(O(e^{-1})\)   |
|               | Tseng’s splitting [Tse00]    | ✓                    | SC-SC             | \(L_1\)      | ✓        | Euclidean \(O(\kappa_1 \log e^{-1})\) |
| Primal-Dual   | Accelerated primal-dual method [HA21] | ✓        | C-C               | \(L_1\)      | ✓        | Bregman \(O(e^{-1})\)    |
| Optimistic    | OGDA [MOP20b]                | ×                    | C-C               | \(L_1\)      | ✓        | Euclidean \(O(e^{-1})\)   |
|               | OGDA [MOP20a]                | ×                    | SC-SC             | \(L_1\)      | ✓        | Euclidean \(O(\kappa_1 \log e^{-1})\) |
|               | Popov’s method [GBVVL19]     | ×                    | SC-SC             | \(L_1\)      | ✓        | Euclidean \(O(\kappa_1 \log e^{-1})\) |
|               | Forward-reflected-backward method [MT20] | ✓        | SC-SC             | \(L_1\)      | ✓        | Euclidean \(O(\kappa_1 \log e^{-1})\) |
|               | Operator extrapolation method [KLL20] | ✓        | C-C               | \(L_1\)      | ✓        | Bregman \(O(e^{-1})\)    |
|               | Operator extrapolation method [KLL20] | ×        | SC-SC             | \(L_1\)      | ✓        | Bregman \(O(\kappa_1 \log e^{-1})\) |
| Theorems 5.1 & 5.3 | Theorems 5.1 & 5.3         | ✓                    | C-C               | \(L_1\)      | ✓        | Bregman \(O(e^{-1})\)    |

is the cubic regularization scheme proposed in [NP06], which enjoys superior global complexity bounds and numerical performance [Nes08; CGT11]. In [Nes06], Nesterov further extended this methodology to solving variational inequalities. For monotone operators with Lipschitz Jacobian, the proposed second-order method achieves a complexity bound of \(O(1/e)\). For strongly-monotone operator with Lipschitz Jacobian, a local quadratic convergence is established for the regularized Newton method, and an overall complexity of \(O(\kappa_2 + \log \log(\epsilon^{-1}))\) is achieved by combining several techniques. More recently, a new cubic regularization scheme was proposed in [HZZ20] for unconstrained smooth saddle point problems. When the objective is strongly-convex-strongly-concave and both its gradient and Hessian are Lipschitz continuous, the authors showed a global linear convergence and a local quadratic convergence with an overall complexity of \(O(\kappa_2^2 + \kappa_1 \kappa_2 + \log \log(\epsilon^{-1}))\). For the convex-concave setting, they used a homotopy continuation approach to derive a complexity of \(O(\log(\epsilon^{-1}))\) under a Lipschitz-type error bound condition, and a complexity of \(O(\epsilon^{-1/2})\) under a Hölderian-type error bound condition with parameter \(\theta \in (0, 1)\).

Along another line of research, a Newton-type extrapolation method was analyzed in [MS10] for finding a zero of a smooth monotone operator under the HPE framework [SS99], and was shown to achieve an iteration complexity of \(O(\epsilon^{-3/4})\). Later, it was further extended to the composite objective with the same convergence rate [MS12]. More recently, the authors in [BL20] proposed second-order and higher-order methods by directly extending the analysis of the classical mirror-prox method [Nem04]. They proved that the proposed \(p\)-th-order method enjoys a complexity bound of \(O(\epsilon^{-1/p+1})\) for smooth convex-concave saddle point problems with Lipschitz \(p\)-th-order
Table 2: Summary of second-order and higher-order methods for saddle point problems. We adopt similar notations as in Table 1. In addition, “$L_2$” means the Hessian of $f$ is $L_2$-Lipschitz and more generally “$L_p$” means the $p$-th order derivative of $f$ is $L_p$-Lipschitz for some $p$. We note that the hc-CRN-SPP method in [HZZ20] also requires an additional error bound assumption, which we discuss in the main text.

| Frameworks          | Methods                              | Composite objective? | Assumptions | Constrained? | Distance | Complexity                          |
|---------------------|--------------------------------------|----------------------|-------------|--------------|----------|--------------------------------------|
| Cubic Regularization| Dual Newton’s method [Nes06]         | ✗                    | C-C         |              | Euclidean | $O(\epsilon^{-1})$                   |
|                     | Restarted dual Newton’s method [Nes06]| ✗                    | SC-SC       | $L_2$        | Euclidean | $O(L_2 \log \log (\epsilon^{-1}))$ |
|                     | hc-CRN-SPP [HZZ20]                  | ✗                    | C-C         | $L_1, L_2$   | Euclidean | $O(\log(\epsilon^{-1}))$ or $O(\epsilon^{-1/2})$ |
|                     | CRN-SPP [HZZ20]                     | ✗                    | SC-SC       | $L_1, L_2$   | Euclidean | $O(\sqrt{\kappa_1 \kappa_2} + \log \log (\epsilon^{-1}))$ |
| Extragradient       | Newton proximal extragradient [MS12]| ✓                    | C-C         | $L_2$        | Euclidean | $O(\epsilon^{-1})$                   |
|                     | High-order Mirror-Prox [BL20]       | ✗                    | C-C         | $L_p$ ($p \geq 2$) | Bregman | $O(\epsilon^{-1/p})$                |
|                     | Restarted high-order Mirror-Prox [OKDG20] | ✗ | SC-SC | $L_p$ ($p \geq 2$) | Euclidean | $O(\epsilon^{-1/p} \log(\epsilon^{-1}))$ |
| Optimistic          | Theorems 6.3 & 7.2                  | ✓                    | C-C         | $L_p$ ($p \geq 2$) | Bregman | $O(\epsilon^{-1/p})$                |
|                     | Corollaries 6.7 & 7.6               | ✓                    | SC-SC       | $L_p$ ($p \geq 2$) | Bregman | $O(\epsilon^{-1/p} \log(\epsilon^{-1}))$ |

We should point out that unlike the first-order methods mentioned earlier, the second-order and higher-order methods often require an expensive subroutine in each iteration. Specifically, the cubic regularization approach requires solving a nontrivial subproblem, while the Newton-type extragradient approach requires a line search scheme to select a stepsize with certain properties. To our best knowledge, only the works in [MS12; BL20] analyzed the complexity of the line search subroutine used in their second-order methods. In this paper, we will also give explicit complexity bounds for our line search scheme in terms of the number of calls to a subproblem solver.

1.2 Outline

The rest of the paper is organized as follows. In section 2, we present some definitions and preliminaries on monotone operator theory and saddle point problems. In section 3, we review the classical Bregman proximal point method and show how some popular first-order methods can be viewed as its approximations. We propose our generalized optimistic method and the adaptive line search scheme in Section 4. The first-order, second-order and higher-order optimistic method are discussed in Sections 5-7, respectively. We report our numerical results in Section 8. Finally, we conclude the paper with some additional remarks in Section 9.

2 Preliminaries

In this section, we introduce our notations and some preliminaries required for presenting our proposed algorithms and their convergence results.
2.1 Bregman Distances

For an arbitrary norm \( \| \cdot \| \), we denote its dual norm by \( \| \cdot \|_* \). To deal with general non-Euclidean norms, we use the Bregman distance to measure the proximity between two points, defined as

\[
D_\Phi(x', x) := \Phi(x') - \Phi(x) - \langle \nabla \Phi(x), x' - x \rangle,
\]

where the “mirror map” \( \Phi \) is a differentiable strictly convex function. A common choice for \( \Phi \) is the squared Euclidean norm \( \frac{1}{2} \| \cdot \|_2^2 \) on \( \mathbb{R}^m \), for which the Bregman distance is given by \( D_\Phi(x', x) = \frac{1}{2} \| x' - x \|^2 \). In the following, we refer to this important special case as the “Euclidean setup”. Note, however, that the Bregman distance is in general not symmetric, i.e., \( D_\Phi(x', x) \neq D_\Phi(x, x') \).

An important property of the Bregman distance is the following three-point identity [CT93]: for any \( u, v, w \in \text{int dom } \Phi \), we have

\[
\langle \nabla \Phi(u) - \nabla \Phi(v), u - w \rangle = D_\Phi(u, v) + D_\Phi(w, u) - D_\Phi(w, v).
\]

Using Bregman distance, one can define a more general notion of strong convexity (e.g., [BBT17; LFN18]).

**Definition 2.1.** A differentiable function \( g \) is \( \mu \)-strongly convex w.r.t. a convex function \( \Phi \) on a convex set \( \mathcal{X} \) if

\[
g(x') \geq g(x) + \langle \nabla g(x), x' - x \rangle + \mu D_\Phi(x', x), \quad \forall x, x' \in \mathcal{X}.
\]

Furthermore, \( g \) is \( \mu \)-strongly concave if \(-g\) is \( \mu \)-strongly convex.

In the Euclidean setup, the strong convexity defined in Definition 2.1 coincides with the standard definition of strong convexity. Also note that by setting \( \mu = 0 \) in the above inequality, we recover the definition for a convex function \( g \). Similarly, we state that \( g \) is concave if \(-g\) is convex.

2.2 Monotone Operators

Our convergence analysis relies heavily on the monotone operator theory, and hence, next we review some of its basic concepts. A set-valued operator \( T : \mathbb{R}^d \rightrightarrows \mathbb{R}^d \) maps a point \( z \in \mathbb{R}^d \) to a (possibly empty) subset of \( \mathbb{R}^d \), and its domain is given by \( \text{dom } T = \{ z \in \mathbb{R}^d : T(z) \neq \emptyset \} \). Alternatively, we can identify \( T \) with its graph defined as \( \text{Gr}(T) = \{ (z, v) : v \in T(z) \} \subset \mathbb{R}^d \times \mathbb{R}^d \). With these notations we introduce the class of monotone operators, which is closely related to convex functions.

**Definition 2.2.** A set-valued operator \( T : \mathbb{R}^d \rightrightarrows \mathbb{R}^d \) is monotone if

\[
\langle v - v', z - z' \rangle \geq 0, \quad \forall (z, v), (z', v') \in \text{Gr}(T).
\]

Also, a monotone operator \( T \) is maximal if there is no other monotone operator \( S \) that \( \text{Gr}(T) \subsetneq \text{Gr}(S) \).
Similar to Definition 2.1, we introduce a stronger notion of monotonicity using Bregman distance.

**Definition 2.3.** A set-valued operator $T : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is $\mu$-strongly monotone w.r.t. a convex function $\Phi$ if for all $(z, v), (z', v') \in \text{Gr}(T)$ we have

$$\langle v - v', z - z' \rangle \geq \mu \langle \nabla \Phi(z) - \nabla \Phi(z') , z - z' \rangle = \mu(D\Phi(z', z) + D\Phi(z, z')).$$

To see the connection between monotone operators and convex functions, we define the subdifferential of a closed proper convex function $g : \mathbb{R}^d \to (-\infty, +\infty]$ at $x$ as the set

$$\partial g(x) := \{ g \in \mathbb{R}^d : g(x') \geq g(x) + \langle g, x' - x \rangle , \forall x' \in \mathbb{R}^d \}.$$

In particular, when $g$ is the indicator function of a closed convex set $X \subset \mathbb{R}^d$, $\partial g$ is the normal cone operator defined as

$$N_X(x) = \begin{cases} \emptyset, & x \not\in X, \\ \{ v \in \mathbb{R}^d | \langle x' - x, v \rangle \leq 0, \forall x' \in X \}, & x \in X. \end{cases}$$

It is known that the subdifferential operator $\partial g$ is maximal monotone if $g$ is closed proper convex.

Moreover, if $g$ is $\mu$-strongly convex w.r.t. $\Phi$, then the operator $\partial g$ is $\mu$-strongly monotone w.r.t. $\Phi$.

Next, we summarize some useful well-known properties of monotone operators; see, e.g., [FP03].

**Proposition 2.1.** Let $T : \mathbb{R}^d \to \mathbb{R}^d$ be a single-valued operator.

(a) A differentiable operator $T$ is monotone if and only if $\langle DT(z)w, w \rangle \geq 0$ for all $w \in \mathbb{R}^d$, $z \in \text{dom} T$, where $DT(z)$ is the Jacobian matrix of $T$ at $z$.

(b) If $T$ is continuous, monotone, and defined on the whole space $\mathbb{R}^d$, then $T$ is maximal monotone.

### 2.3 Lipschitz and Smooth Operators

In this paper, we deal with higher-order derivatives of monotone operators. Specifically, for a smooth operator $F : \text{dom} F \to \mathbb{R}^d$ defined on an open domain $\text{dom} F \subset \mathbb{R}^d$, we use $D^p F(z)[h_1, h_2, \ldots, h_p] \in \mathbb{R}^d$ to denote the $p$-th-order directional derivative of $F$ at point $z$ along the directions $h_1, \ldots, h_p \in \mathbb{R}^d$. Moreover, when $h_1, \ldots, h_p$ are all identical, we abbreviate the directional derivative as $D^p F(z)[h]^p$.

Note that the $p$-th-order derivative tensor $D^p F(z)[\cdot]$ is a multilinear map, and it coincides with the Jacobian of $F$ when $p = 1$. Recalling that $\| \cdot \|$ is the dual norm of $\| \cdot \|$, we define the operator norm of the tensor induced by a norm $\| \cdot \|$ on $\mathbb{R}^d$ by

$$\|D^p F(z)\|_{\text{op}} = \max_{\|h_i\| = 1, i = 1, \ldots, p} \|D^p F(z)[h_1, \ldots, h_p]\|_\ast.$$

In this paper, the problem classes of interest consist of Lipschitz continuous and smooth operators, defined as follows.

**Definition 2.4.** An operator $F : \text{dom} F \to \mathbb{R}^d$ is $L_1$-Lipschitz if

$$\|F(z) - F(z')\|_\ast \leq L_1 \|z - z'\|, \quad \forall z, z' \in \text{dom} F.$$ 

Moreover, for $p \geq 2$, $F$ is $p$-th-order $L_p$-smooth if

$$\|D^{p-1} F(z) - D^{p-1} F(z')\|_{\text{op}} \leq L_p \|z - z'\|, \quad \forall z, z' \in \text{dom} F.$$
We define the $p$-th-order Taylor expansion of $F$ at point $z$ by

$$T^{(p)}(z'; z) = F(z) + \sum_{i=1}^{p} \frac{1}{i!} D^i F(z)(z' - z)^i.$$  

(5)

Note that for a $p$-th-order smooth operator $F$, its $(p-1)$-th-order Taylor expansion can be considered as its natural approximation. Specifically, if $F$ is $p$-th-order $L_p$-smooth, the residual of the Taylor expansion is bounded above by

$$\|F(z') - T^{(p-1)}(z'; z)\|_\ast \leq \frac{L_p}{p!} \|z - z'\|^p, \quad \forall z, z' \in \text{dom } F.$$  

(6)

Similarly, we also have

$$\|DF(z') - DT^{(p-1)}(z'; z)\|_{op} \leq \frac{L_p}{(p-1)!} \|z - z'\|^{p-1}, \quad \forall z, z' \in \text{dom } F.$$  

(7)

### 2.4 Saddle Point Problems

Recall the convex-concave minimax problem defined in (1). An optimal solution $(x^*, y^*)$ of the problem in (1) is also called a saddle point of the objective function $\ell$, as it satisfies the following condition

$$\ell(x^*, y) \leq \ell(x^*, y^*) \leq \ell(x, y^*), \quad \forall x \in \mathcal{X}, \forall y \in \mathcal{Y}.$$  

Throughout the paper, we assume that Problem (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 $z := (x, y) \in \mathbb{R}^{m+n}$ and $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$. Define the operators $F : \mathcal{Z} \rightarrow \mathbb{R}^{m+n}$ and $H : \mathbb{R}^{m+n} \rightarrow \mathbb{R}^{m+n}$ by

$$F(z) := \begin{bmatrix} \nabla_x f(x, y) \\ -\nabla_y f(x, y) \end{bmatrix} \quad \text{and} \quad H(z) := \begin{bmatrix} \partial h_1(x) \\ \partial h_2(y) \end{bmatrix} + \begin{bmatrix} N_X(x) \\ N_Y(y) \end{bmatrix}.$$  

(8)

Then by the first-order optimality condition, $z^*$ is a saddle point of (1) if and only if it solves the following monotone inclusion problem:

$$\text{find } z^* \in \mathbb{R}^{m+n} \text{ such that } 0 \in F(z^*) + H(z^*).$$  

(9)

Moreover, if we define $h(z) := h_1(x) + h_2(y)$, (1) is also equivalent to the variational inequality:

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

(10)

We assume that $\mathcal{X}$ and $\mathcal{Y}$ are equipped with the norms $\| \cdot \|_{\mathcal{X}}$ and $\| \cdot \|_{\mathcal{Y}}$, respectively. Moreover, we have two mirror maps — a differentiable function $\Phi_\mathcal{X} : \mathcal{X} \rightarrow \mathbb{R}$ that is 1-strongly convex w.r.t. $\| \cdot \|_{\mathcal{X}}$ and a differentiable function $\Phi_\mathcal{Y} : \mathcal{Y} \rightarrow \mathbb{R}$ that is 1-strongly convex w.r.t. $\| \cdot \|_{\mathcal{Y}}$. We endow $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ with the norm $\| \cdot \|_{\mathcal{Z}}$ defined by $\|z\|_{\mathcal{Z}} := \sqrt{\|x\|^2_{\mathcal{X}} + \|y\|^2_{\mathcal{Y}}}$ for all $z = (x, y) \in \mathcal{Z}$, and with the mirror map $\Phi_{\mathcal{Z}} : \mathcal{Z} \rightarrow \mathbb{R}$ defined by $\Phi_{\mathcal{Z}}(z) := \Phi_\mathcal{X}(x) + \Phi_\mathcal{Y}(y)$. It can be verified that

$$D_{\Phi_{\mathcal{Z}}}(z', z) = D_{\Phi_\mathcal{X}}(x', x) + D_{\Phi_\mathcal{Y}}(y', y),$$  

(11)

and $\Phi_{\mathcal{Z}}$ is 1-strongly convex w.r.t. $\| \cdot \|_{\mathcal{Z}}$. To simplify the notation, we omit $\mathcal{Z}$ in the subscripts of $\| \cdot \|_{\mathcal{Z}}$ and $\Phi_{\mathcal{Z}}$ when there is no ambiguity.

In this paper, we consider two specific problem classes for our theoretical results: (i) convex-concave saddle point problems (Assumption 2.1) and (ii) strongly-convex-strongly-concave saddle point problems (Assumption 2.2). Note that Assumption 2.2 reduces to Assumption 2.1 when $\mu = 0$, and sometimes this allows us to unify our results in both settings by letting $\mu \geq 0$. 

12
Assumption 2.1. The functions $h_1$ and $h_2$ are proper closed convex, and the function $f$ is convex in $x$ and concave in $y$ on $\mathcal{X} \times \mathcal{Y}$.

Assumption 2.2. The functions $h_1$ and $h_2$ are proper closed convex, and the function $f$ is $\mu$-strongly convex in $x$ w.r.t. $\Phi_{\mathcal{X}}$ and $\mu$-strongly concave in $y$ w.r.t. $\Phi_{\mathcal{Y}}$ on $\mathcal{X} \times \mathcal{Y}$.

As mentioned in Section 2.2, we will mainly work with the operator $F$ to develop our algorithms and convergence proofs in this paper. The following lemma serves the purpose of translating the properties of the function $f$ into the properties of the operator $F$ (see Appendix A.1 for the proof).

**Lemma 2.2.** The operator $F$ defined in (8) is monotone on $\mathcal{Z}$ under Assumption 2.1, and is $\mu$-strongly monotone w.r.t. $\Phi_{\mathcal{Z}}$ on $\mathcal{Z}$ under Assumption 2.2.

In the convex-concave setting, the suboptimality of a solution $(x,y)$ is measured by

$$
\Delta(x,y) := \max_{y' \in \mathcal{Y}} \ell(x,y') - \min_{x' \in \mathcal{X}} \ell(x',y),
$$

(12)

which is known as the primal-dual gap. Since we can write $\Delta(x,y) = (\max_{y' \in \mathcal{Y}} \ell(x,y') - \ell(x^*,y^*)) + \left(\ell(x^*,y^*) - \min_{x' \in \mathcal{X}} \ell(x',y)\right)$, we obtain that $\Delta(x,y) \geq 0$ for any $(x,y) \in \mathcal{X} \times \mathcal{Y}$ and the equality holds if and only if $(x,y)$ is a saddle point. However, in some cases where the feasible sets are unbounded, the primal-dual gap defined in (12) can be always infinite except at the saddle points\(^4\), rendering it useless. One remedy is to use the restricted primal-dual gap function [Nes07; CP11].

Given two bounded sets $B_1 \subset \mathcal{X}$ and $B_2 \subset \mathcal{Y}$, we define the restricted primal-dual gap as

$$
\Delta_{B_1 \times B_2}(x,y) := \max_{y' \in B_2} \ell(x,y') - \min_{x' \in B_1} \ell(x',y).
$$

(13)

As discussed in [CP11], this gap function has two major properties:

1. If $(x^*,y^*) \in B_1 \times B_2$, then we have $\Delta_{B_1 \times B_2}(x,y) \geq 0$ for any $(x,y) \in \mathcal{X} \times \mathcal{Y}$.

2. If $(x,y)$ is in the interior of $B_1 \times B_2$, then $\Delta_{B_1 \times B_2}(x,y) = 0$ if and only if $(x,y)$ is a saddle point of $\ell$.

As we discuss in upcoming sections, the iterates of our algorithm will always stay in a bounded set centered at $(x^*,y^*)$. Hence, the restricted primal-dual gap function will serve as a good measure of suboptimality if we choose $B_1 \times B_2$ large enough such that it contains a saddle point $(x^*,y^*)$.

The following well-known result is the first step of our convergence analysis, which provides an upper bound for the primal-dual gap at the averaged iterate. The proof is available in Appendix A.2.

**Lemma 2.3.** Suppose that Assumption 2.1 holds. Then for any $z,z_1,\ldots,z_N \in \mathcal{Z}$ and $\theta_1,\ldots,\theta_N \geq 0$ with $\sum_{k=1}^{N} \theta_k = 1$, we have

$$
\ell(\bar{x}_N,y) - \ell(x,\bar{y}_N) \leq \sum_{k=1}^{N} \theta_k \left(\langle F(z_k), z_k - z \rangle + h(z_k) - h(z)\right),
$$

where $\bar{x}_N$ and $\bar{y}_N$ are given by $\bar{x}_N = \sum_{k=1}^{N} \theta_k x_k$ and $\bar{y}_N = \sum_{k=1}^{N} \theta_k y_k$.

---

\(^4\)For such an example, consider the unconstrained saddle point problem $\min_{x \in \mathbb{R}^{m}} \max_{y \in \mathbb{R}^{m}} \langle x, y \rangle$. Then the primal-dual gap is infinite except at the unique saddle point $(0,0)$.
Remark 2.1. For brevity, we report our convergence results in the form of an upper bound on \( \ell(\bar{x}_N, y) - \ell(x, y_N) \) for any \((x, y) \in X \times Y\) (e.g., see Theorem 3.1(a)). Then taking the supremum of both sides over \( x \in B_1 \) and \( y \in B_2 \) results in an upper bound on the restricted primal-dual gap in (13).

In the strongly-convex-strongly-concave setting, Problem (1) has a unique saddle point \( z^* \in Z \). Hence, we measure the suboptimality of \( z \) by the Bregman distance \( D_\Phi(z^*, z) \). The following lemma presents the key property we shall use in our convergence analysis. The proof is available in Appendix A.3.

Lemma 2.4. If Assumption 2.2 holds with \( \mu \geq 0 \) and \( z^* \) is a saddle point of Problem (1), then for any \( z \in Z \)

\[
\langle F(z), z - z^* \rangle + h(z) - h(z^*) \geq \mu D_\Phi(z^*, z). \tag{14}
\]

Remark 2.2. In fact, our convergence results still hold if we replace Assumption 2.2 with the weaker assumption that the inequality in (14) is always satisfied, which is referred to as the generalized monotonicity condition in [KLL20].

Remark 2.3. It is not hard to see from the proof that (14) can be improved to

\[
\langle F(z), z - z^* \rangle + h(z) - h(z^*) \geq (1 + s(\Phi))\mu D_\Phi(z^*, z),
\]

where \( s(\Phi) \) is the symmetry coefficient defined in (3). For simplicity, we omit the constant \( 1 + s(\Phi) \) when reporting our main convergence results, but we will take it into consideration in Section 8 for a more accurate comparison between our theoretical analysis and numerical results.

Finally, we introduce another metric that will serve as the termination criterion in our line search scheme. In light of the monotone inclusion formulation (9), define the residual at \( z \) as

\[
\text{res}(z) := \min_{w \in H(z)} \| F(z) + w \|_*.
\tag{15}
\]

It can be verified that \( z = (x, y) \in \mathbb{R}^{m+n} \) is a saddle point of problem (1) if and only if \( \text{res}(z) = 0 \). Also note that in the unconstrained setting where \( H(z) \equiv 0 \), the saddle point problem (1) is equivalent to the nonlinear equation \( F(z^*) = 0 \), and \( \text{res}(z) = \| F(z) \|_* \) captures the residual at the solution \( z \).

With these optimality metrics in place, we can formulate our goal as finding an \( \epsilon \)-accurate solution of problem (1) defined below in Definition 2.5.

Definition 2.5. A point \( z = (x, y) \in X \times Y \) is \( \epsilon \)-accurate for the saddle point problem (1) if

\[
\min\{ \Delta_{B_1 \times B_2}(x, y), D_\Phi(z, z^*), \text{res}(z) \} \leq \epsilon,
\]

i.e., at least one of the optimality metrics defined above does not exceed \( \epsilon \) at point \( z \).

3 Bregman Proximal Point Method and Its Approximations

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. Hence, to better motivate our method, we first review the basics of the Bregman proximal point
(BPP) method \cite{CT93;Eck93}. In the $k$-th iteration of the BPP method for solving Problem (1), the new point $z_{k+1}$ is given by the unique solution of the monotone inclusion subproblem

$$0 \in \eta_k F(z) + \eta_k H(z) + \nabla \Phi(z) - \nabla \Phi(z_k),$$

(16)

where $\eta_k > 0$ is the “stepsize”. Equivalently, we can also write

$$z_{k+1} = \arg\min_{z \in Z} \{ \eta_k \langle F(z_{k+1}), z - z_k \rangle + \eta_k h(z) + D_\Phi(z, z_k) \}.$$  

(17)

Note, however, that (17) is not an explicit update rule since its right-hand side also depends on $z_{k+1}$. Regarding its convergence property, we have the following results. The proofs are presented in Appendix B. While the result for the convex-concave setting is well-known in the literature \cite{Nem04}, we note that the result for the strongly-convex-strongly-concave setting appears to be new\textsuperscript{5}.

**Theorem 3.1.** Let $\{z_k\}_{k \geq 0}$ be the iterates generated by the BPP method in (16).

(a) Under Assumption 2.1, we have $D_\Phi(z^*, z_{k+1}) \leq D_\Phi(z^*, z_k)$ for any $k \geq 0$. Moreover, for any $z = (x, y) \in X \times Y$, we have

$$\ell(\tilde{x}_N, y) - \ell(x, \tilde{y}_N) \leq D_\Phi(z, z_0) \left( \sum_{k=0}^{N-1} \eta_k \right)^{-1},$$

where $\tilde{z}_N = (\tilde{x}_N, \tilde{y}_N)$ is given by $\tilde{z}_N = \frac{1}{\sum_{k=0}^{N-1} \eta_k} \sum_{k=0}^{N-1} \eta_k z_{k+1}$.

(b) Under Assumption 2.2, we have

$$D_\Phi(z^*, z_N) \leq D_\Phi(z^*, z_0) \prod_{k=0}^{N-1} (1 + \eta_k \mu)^{-1}.$$  

Theorem 3.1 characterizes the convergence rates of the BPP method in terms of the stepsizes $\{\eta_k\}_{k \geq 0}$. We note that it requires no condition on the stepsizes $\eta_k$, and hence in theory the BPP method can converge arbitrarily fast with sufficiently large stepsizes. On the other hand, in each step it involves solving a highly nontrivial monotone inclusion problem in (16) and can be computationally intractable. Hence, our goal is to introduce a general class of optimistic methods that approximates the BPP method by replacing (16) with more tractable subproblems. As we shall see, our method can achieve the same convergence guarantees (up to constant) as in Theorem 3.1, but at the same time introduces additional restrictions on the stepsizes.

### 3.1 Approximations of the BPP Method

To leverage the superior convergence performance of the BPP method while maintaining computational tractability, several iterative algorithms have been developed that aim at approximating BPP with explicit and efficient update rules. In particular, we describe two first-order methods, which represent the two different frameworks we reviewed in Section 1.1: (i) The mirror-prox method and (ii) The optimistic mirror descent method.

\textsuperscript{5}Rockafellar proved a similar (and better) result for the proximal point method in \cite{Roc76} in the Euclidean setup, while our result also applies to a general Bergman distance.
**Mirror-Prox method.** The key insight of the mirror-prox method proposed in [Nem04] is that when $F$ is Lipschitz continuous, we can approximately solve (17) with two steps per iteration, with a properly chosen stepsize. Specifically, for constrained saddle point problems, the mirror-prox update is

$$z_{k+1} = \arg \min_{z \in \mathcal{Z}} \left\{ \eta_k \langle F(z), z - z_k \rangle + D_\Phi(z, z_k) \right\},$$

$$z_{k+1} = \arg \min_{z \in \mathcal{Z}} \left\{ \eta_k \langle F(z_{k+1/2}), z - z_k \rangle + D_\Phi(z, z_k) \right\},$$

In words, we first take a step of mirror descent to the middle point $z_{k+1/2}$, and then use $F(z_{k+1/2})$ as a surrogate for $F(z_k)$ (c.f. (17)). In the Euclidean setup, it is the same as the extragradient method proposed in [Kor76]. Also note that the more general version of the mirror-prox method for composite saddle point problems has also been studied in [Tse08], and its update is formally defined as

$$z_{k+1/2} = \arg \min_{z \in \mathcal{Z}} \left\{ \eta_k \langle F(z), z - z_k \rangle + \eta_k h(z) + D_\Phi(z, z_k) \right\},$$

$$z_{k+1} = \arg \min_{z \in \mathcal{Z}} \left\{ \eta_k \langle F(z_{k+1/2}), z - z_k \rangle + \eta_k h(z) + D_\Phi(z, z_k) \right\},$$

where $h(z) := h_1(x) + h_2(y)$.

**Optimistic mirror descent method.** Another approach for properly approximating proximal point updates is the optimistic mirror descent method, which dates back to [Pop80]. Later on, it also gained attention in the context of online learning [CYLMLJZ12; RS13] and Generative Adversarial Networks (GANs) [DISZ18; GBVVL19; LS19; PDZC20]. Specifically, the update rule of this method for constrained saddle point problems is

$$z_k = \arg \min_{z \in \mathcal{Z}} \left\{ \eta_k \langle F(z_{k-1}), z - z_{k-1} \rangle + D_\Phi(z, z_{k-1/2}) \right\},$$

$$z_{k+1/2} = \arg \min_{z \in \mathcal{Z}} \left\{ \eta_k \langle F(z_k), z - z_k \rangle + D_\Phi(z, z_{k-1/2}) \right\}.$$ 

Comparing (19) with (18), we observe that optimistic mirror descent operates in a similar way as the mirror-prox method, but instead of computing $F(z_{k-1/2})$, we reuse the gradient $F(z_{k-1})$ from the previous iteration to obtain $z_k$. Hence, each iteration of this algorithm requires one gradient computation, unlike the mirror-prox update which requires two gradient evaluations per iteration.

The discussions above view the “midpoints” $\{z_{k+1/2}\}_{k \geq 0}$ as the approximate iterates of the BPP method. Alternatively, we can also interpret $\{z_k\}_{k \geq 0}$ as the approximate iterates in the special setting of unconstrained saddle point problems in the Euclidean setup. Such viewpoint was first discussed in [MOP20a]. In this case, the BPP method in (16) can be written as the implicit update

$$z_{k+1} = z_k - \eta_k F(z_{k+1}),$$

while (19) leads to the simple update rule

$$z_{k+1} = z_k - 2\eta_k F(z_k) + \eta_k F(z_{k-1}) = z_k - \eta_k F(z_k) - \eta_k (F(z_k) - F(z_{k-1})),$$

which is also known as the optimistic gradient descent-ascent (OGDA) update [MOP20a; MOP20b]. To better illustrate the main idea, in (21) we write the OGDA update as a combination of two terms: a “prediction term” $F(z_k)$ that serves as a surrogate for $F(z_{k+1})$, and a “correction term” $F(z_k) - F(z_{k-1})$ given by the deviation between $F(z_k)$ 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., $F(z_{k+1}) - F(z_k) \approx F(z_k) - F(z_{k-1})$, then it holds that $F(z_{k+1}) \approx 2F(z_k) - F(z_{k-1})$ and we can expect (21) to be a good approximation for (20).

Finally, we note that both mirror-prox and optimistic mirror descent methods only use first-order information (i.e., gradients) to approximate the BPP method. It is natural to ask whether one can exploit higher-order information to improve the accuracy of approximation and hence achieve faster convergence. While previous works [MS10; MS12; BL20] have explored such possibility for the mirror-prox method, to our best knowledge, no similar effort has been made for the optimistic mirror descent method. Moreover, it remains unclear whether the current convergence analysis for the optimistic mirror decent method can be extended to more general composite saddle point problem. In the rest of the paper, we close these gaps and present a general framework for optimistic methods, following the viewpoint in [MOP20a].

4 Main algorithm: the Generalized Optimistic Method

In this section, we present the generalized optimistic method, which extends the “optimistic” idea described above in the OGDA method for approximating the BPP method. Our proposed scheme is able to handle the composite structure in (1) and can utilize second-order and higher-order information as we shall discuss in the upcoming sections.

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 (21), the term $F(z_{k+1})$ is approximated by a linear combination of the prediction term $F(z_k)$ and the correction term $F(z_k) - F(z_{k-1})$. Now instead of using $F(z_k)$ as the prediction for $F(z_{k+1})$, suppose that we have access to a general approximation function denoted by

$$
\text{prediction term:} \quad P(z; I_k),
$$

which is constructed from the available information $I_k$ up to the $k$-th iteration. For instance, with a second-order oracle we have $I_k = \{F(z_0), DF(z_0), F(z_1), DF(z_1), \ldots, F(z_k), DF(z_k)\}$, and a possible candidate for $P$ is the second-order Taylor expansion $T^{(1)}(z; z_k) = F(z_k) + DF(z_k)(z - z_k)$.

And similar to the logic in OGDA, the correction term is then given by the difference between $F(z_k)$ and our prediction in the previous step, i.e.,

$$
\text{correction term:} \quad F(z_k) - P(z_k; I_{k-1}).
$$

Indeed, if we are “optimistic” that the change in the prediction error between consecutive iterations is small, i.e., $F(z_{k+1}) - P(z_{k+1}; I_k) \approx F(z_k) - P(z_k; I_{k-1})$, then this implies $F(z_{k+1}) \approx P(z_{k+1}; I_k) + (F(z_k) - P(z_k; I_{k-1}))$ and adding the correction term can help reducing the approximation error. Considering these generalizations, we propose to compute the new iterate by following the update

$$
\begin{align*}
\hat{z}_{k+1} &= z_k - (\eta_k P(z_{k+1}; I_k) + \eta_k (F(z_k) - P(z_k; I_{k-1}))).
\end{align*}
$$

Unlike the OGDA method, this is not an explicit update in general and requires us to solve the equation with respect to $z_{k+1}$. Also, we use different stepsize 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 $\eta_k$. As we shall see, we will set $\hat{\eta}_k = \eta_{k-1}/(1 + \mu\eta_{k-1})$ in the strongly-convex-strongly-concave setting.
Algorithm 1 Generalized optimistic method for monotone problems

1: Input: initial point $z_0 \in \mathcal{Z}$, approximation function $P$, strongly-convex parameter $\mu \geq 0$, and $0 < \alpha \leq 1$

2: Initialize: set $z_{-1} \leftarrow z_0$ and $P(z_0; I_{-1}) \leftarrow F(z_0)$

3: for iteration $k = 0, \ldots, N-1$ do

4: Choose $\eta_k, \hat{\eta}_k > 0$ and compute $z_{k+1}$ such that

\[
0 \in \eta_k P(z_{k+1}; I_k) + \hat{\eta}_k (F(z_k) - P(z_{k}; I_{k-1})) + \eta_k H(z_{k+1}) + \nabla \Phi(z_{k+1}) - \nabla \Phi(z_k),
\]

\[
\eta_k \| F(z_{k+1}) - P(z_{k+1}; I_k) \|_* \leq \frac{\alpha}{2} \| z_{k+1} - z_k \|.
\]

5: end for

The same methodology can be readily generalized for the composite setting with a non-Euclidean norm. Specifically, in the most general form of our proposed generalized optimistic method for solving problem (1), the new point $z_{k+1}$ is given by the unique solution of the following inclusion problem

\[
0 \in \eta_k P(z; I_k) + \hat{\eta}_k (F(z_k) - P(z_{k}; I_{k-1})) + \eta_k H(z) + \nabla \Phi(z) - \nabla \Phi(z_k),
\]

which can also be equivalently written as

\[
z_{k+1} = \arg \min_{z \in \mathcal{Z}} \{ \eta_k P(z_{k+1}; I_k) + \hat{\eta}_k (F(z_k) - P(z_{k}; I_{k-1})), z - z_k \} + \eta_k h(z) + D_\Phi(z, z_k) \}. \tag{23}
\]

Comparing the update rules (22) and (23) with the ones for the BPP method in (16) and (17), we can see that the only modification is replacing the term $\eta_k F(z_{k+1})$ with its optimistic approximation $\eta_k P(z_{k+1}; I_k) + \hat{\eta}_k (F(z_k) - P(z_{k}; I_{k-1}))$. Moreover, to ensure that $P(z_{k+1}; I_k)$ remains a valid approximation of $F(z_{k+1})$, we impose the following condition

\[
\eta_k \| F(z_{k+1}) - P(z_{k+1}; I_k) \|_* \leq \frac{\alpha}{2} \| z_{k+1} - z_k \|, \tag{24}
\]

where $\alpha \in (0, 1]$ is a user-specified constant. The steps of the generalized optimistic method are summarized in Algorithm 1.

We remark that our generalized optimistic method should be regarded as a general framework rather than a directly applicable algorithm. At this point, we do not specify how to choose $\eta_k$ and $z_{k+1}$ to satisfy the condition in (24), which will depend on the properties of $F$ and the particular choice of $P$. A generic line search scheme will be described in Section 4.1, and we will further devote Sections 5, 6, and 7 to discuss the particular implementations in different settings.

The following proposition will form the basis for all the convergence results derived later in this paper. Essentially, it proves that Algorithm 1 enjoys similar convergence guarantees to the BPP method under the specified conditions, even if it only solves an approximated version of the sub-problem (16).

Proposition 4.1. Let $\{z_k\}_{k \geq 0}$ be the iterates generated by Algorithm 1.

(a) Under Assumption 2.1, if we set $\hat{\eta}_k = \eta_{k-1}$, we have

\[
D_\Phi(z^*, z_k) \leq \frac{2}{2 - \alpha} D_\Phi(z^*, z_0), \quad \forall k \geq 0. \tag{25}
\]

Moreover, for any $z = (x, y) \in X \times Y$, we have
\[
\ell(\mathbf{x}_N, \mathbf{y}) - \ell(\mathbf{x}, \mathbf{y}_N) \leq D_{\Phi}(\mathbf{z}, \mathbf{z}_0) \left( \sum_{k=0}^{N-1} \eta_k \right)^{-1},
\]

where \( \mathbf{z}_N = (\mathbf{x}_N, \mathbf{y}_N) \) is given by \( \mathbf{z}_N = \frac{1}{\sum_{k=0}^{N-1} \eta_k} \sum_{k=0}^{N-1} \eta_k \mathbf{z}_{k+1} \).

(b) Under Assumption 2.2, if we set \( \tilde{\eta}_k = \eta_{k-1}/(1 + \mu \eta_{k-1}) \), we have
\[
D_{\Phi}(\mathbf{z}^*, \mathbf{z}_N) \leq \frac{2}{2 - \alpha} D_{\Phi}(\mathbf{z}^*, \mathbf{z}_0) \prod_{k=0}^{N-1} (1 + \eta_k \mu)^{-1}.
\]

Proof. See Appendix C.1.

As in Theorem 3.1 for the BPP method, the convergence rate in Proposition 4.1 for the generalized optimistic method depends on the choice of \( \{\eta_k\}_{k \geq 0} \). Unlike the BPP method though, the stepsizes here are not arbitrary but constrained by condition (24). Depending on our choice of the approximation function \( P \), the condition on \( \eta_k \) varies and as a result we obtain different convergence rates. In fact, as we shall see, the stepsize \( \eta_k \) is closely related to the displacement \( \|\mathbf{z}_{k+1} - \mathbf{z}_k\| \): it is approximately on the order of \( \|\mathbf{z}_{k+1} - \mathbf{z}_k\|^{p-1} \) in our \( p \)-th order method (\( p \geq 2 \)). In the following lemma, 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 on \( \prod_{k=0}^{N-1} (1 + \eta_k \mu) \).

Lemma 4.2. Let \( \{\mathbf{z}_k\}_{k \geq 0} \) be the iterates generated by Algorithm 1 with \( \tilde{\eta}_k \) chosen as in Proposition 4.1. Then under Assumption 2.1, we have
\[
\sum_{k=0}^{N-1} \|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2 \leq \frac{2}{1 - 1/\alpha} D_{\Phi}(\mathbf{z}^*, \mathbf{z}_0),
\]

and under Assumption 2.2, we have
\[
\sum_{k=0}^{N-1} \left( \|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2 \prod_{l=0}^{k-1} (1 + \eta l \mu) \right) \leq \frac{2}{1 - 1/\alpha} D_{\Phi}(\mathbf{z}^*, \mathbf{z}_0).
\]

Proof. See Appendix C.2.

4.1 Adaptive Line Search

As mentioned earlier, our general method proposed in Algorithm 1 does not directly lead to a practical algorithm as it requires a proper policy for ensuring the condition in (24). As the first step towards practical implementations, in this section we propose a generic line search scheme to select the stepsize \( \eta_k \) adaptively in Algorithm 1. To simplify the notation, let \( \mathbf{z}^- = \mathbf{z}_k, \eta = \eta_k, P(\mathbf{z}) = P(\mathbf{z}; I_k), \) and \( \mathbf{v}^- = \mathbf{v}_k := \tilde{\eta}_k (F(\mathbf{z}_k) - P(\mathbf{z}_k; I_{k-1})) \). In light of (22) and (24), at the \( k \)-th iteration our goal is to find a stepsize \( \eta > 0 \) and a point \( \mathbf{z} \in \mathcal{Z} \) such that
\[
0 \in \eta P(\mathbf{z}) + \mathbf{v}^- + \eta H(\mathbf{z}) + \nabla \Phi(\mathbf{z}) - \nabla \Phi(\mathbf{z}^-),
\]
\[
\eta \|F(\mathbf{z}) - P(\mathbf{z})\|_* \leq \frac{\alpha}{2} \|\mathbf{z} - \mathbf{z}^-\|.
\]

Note that the conditions in (27) and (28) depend on both \( \eta \) and \( \mathbf{z} \). Hence, in general we need to choose \( \eta \) and \( \mathbf{z} \) simultaneously, except in the first-order setting (see Section 5).
We assume that the approximation function \( P \) is maximal monotone and continuous on \( Z \). As a corollary, for any fixed \( \eta > 0 \), the monotone inclusion subproblem with respect to \( z \) in (27) has a unique solution (see, e.g., [Eck93]), which we denote by \( z(\eta; z^-) \). Also, we assume access to a black-box solver of the subproblem in (27), which we formally define as follows for later reference.

**Definition 4.1 (Optimistic Subsolver).** For any given \( \eta, v^- \) and \( z^- \), the optimistic subsolver returns the solution of (27).

Hence, the conditions in (27) and (28) can be rewritten as

\[
\eta \| F(z(\eta; z^-)) - P(z(\eta; z^-)) \|_* \leq \frac{\alpha}{2} \| z(\eta) - z^- \|.
\] (29)

Our line search scheme works by repeatedly checking the condition (29) for different stepsizes, each of which involves one call to the optimistic subsolver. More specifically, in each step we pick a stepsize \( \eta \) and find its corresponding solution \( z(\eta; z^-) \) in (27) to see if the condition in (29) is satisfied. If not, then we update the stepsize \( \eta \) and repeat the process.

**Remark 4.1.** The major computational cost of our line search schemes comes from solving the subproblem (27) with different stepsizes, namely, multiple calls to the optimistic subsolver. On the other hand, the computational cost of the subsolver depends on the specific saddle point problem and our choice of the approximation function \( P \). Hence, we use one call to the optimistic subsolver as the unit and measure the complexity of our line search scheme in terms of the number of calls to the solver.

We not only search for an admissible stepsize \( \eta \) satisfying (29), but also hope to make it as large as possible to achieve faster convergence. Formally, we aim to find a \( \beta \)-optimal stepsize defined below.

**Definition 4.2.** Given \( 0 < \beta < 1 \), the stepsize \( \eta \) is \( \beta \)-optimal for condition (29) if:

(a) it is admissible, i.e., it satisfies the condition in (29);

(b) there exists \( \eta' \in (\eta, \eta/\beta] \) that is inadmissible, i.e., it does not satisfy the condition in (29).

According to this definition, the stepsize \( \eta \) is \( \beta \)-optimal if it satisfies condition (29) but cannot be further increased by a factor of \( 1/\beta \), where \( \beta \in (0, 1) \).

On a high level, our line search scheme follows the bracket-then-bisect approach similar to the one in [MS12], consisting of a bracketing stage that outputs a lower bound and an upper bound on the target stepsizes, and a bisection stage that further refines the bounds to locate the final stepsizes. Specifically, suppose that we start by the initial trial stepsize \( \sigma \). In the bracketing stage, we aim to find an admissible stepsize \( \eta_{\text{lo}} \) as the lower bound and an inadmissible stepsize \( \eta_{\text{up}} \) as the upper bound, while keeping the two as close as possible. If \( \sigma \) does not satisfy (29), we will call the subroutine \textsc{Backtrack} to initialize \( \eta_{\text{up}} \) as \( \sigma \) and then keep decreasing the trial stepsize \( \eta \) and updating the upper bound \( \eta_{\text{up}} \) until we meet an admissible stepsize \( \eta_{\text{lo}} \). In the same spirit, if \( \sigma \) satisfies (29), we will call the subroutine \textsc{Advance} to initialize \( \eta_{\text{lo}} \) as \( \sigma \) and then keep increasing \( \eta \) and updating \( \eta_{\text{lo}} \) until we meet an inadmissible stepsize \( \eta_{\text{up}} \) (we also need to take care of a technical issue to ensure finite termination, which we shall detail later). Subsequently, in the bisection stage the subroutine \textsc{Bisection} performs a logarithm-type bisection on the interval \([\eta_{\text{lo}}, \eta_{\text{up}}]\) until a \( \beta \)-optimal stepsizes is found. Note that in the case where the initial trial stepsize \( \sigma \) is admissible, we may also skip the subroutines \textsc{Advance} and \textsc{Bisection} and simply accept \( \sigma \) as the final stepsize,
Algorithm 2 Generalized optimistic method with line search

1: **Input:** current point $z^- \in Z$, $v^- \in \mathbb{R}^d$, initial trial stepsize $\sigma > 0$, approximation function $P$, strongly-convex parameter $\mu \geq 0$, $\alpha \in (0, 1]$, $\beta \in (0, 1)$, $\epsilon > 0$
2: **Output:** stepsize $\eta$
3: if $\sigma$ does not satisfy (29) then
   4: Set $\eta_{lo}, \eta_{up} \leftarrow$ Backtrack($z^-, \sigma, \alpha, \beta$)
   5: Set $\eta \leftarrow$ Bisection($z^-, \eta_{lo}, \eta_{up}, \alpha, \beta$) and return
   6: else
     7: **Option I (without advancing):**
      8: Set $\eta \leftarrow \sigma$ and return
     9: **Option II (with advancing):**
       10: Set $\eta_{lo}, \eta_{up} \leftarrow$ Advance($z^-, \sigma, \alpha, \beta, \epsilon$)
       11: if $\text{res}(z(\eta_{lo}; z_k)) \leq \epsilon$ then
          12: Set $\eta \leftarrow \eta_{lo}$ and return
       13: else
          14: Set $\eta \leftarrow$ Bisection($z^-, \eta_{lo}, \eta_{up}, \alpha, \beta$) and return
       15: end if

Subroutine 1 Backtrack($z^-, \sigma, \alpha, \beta$)

1: **Input:** current point $z^- \in Z$, initial stepsize $\sigma$, $\alpha \in (0, 1]$, $\beta \in (0, 1)$
2: **Initialize:** set $\eta_{up} \leftarrow \sigma$, $\eta \leftarrow \beta \sigma$
3: while $\eta$ does not satisfy (29) do
   4: Update $\eta_{up} \leftarrow \eta$ and decrease the trial stepsize $\eta \leftarrow \beta \eta^2 / \sigma$
5: end while
6: Set $\eta_{lo} \leftarrow \eta$
7: **Return:** $\eta_{lo}, \eta_{up}$

which we refer to as the line search scheme without advancing. The whole procedure is formally described in Algorithm 2.

We now specify the subroutines used in Algorithm 2. If the initial trial stepsize $\sigma$ is inadmissible, we call Backtrack (Subroutine 1) and set $\eta_{up}$ as $\sigma$. In each step of Backtrack, we maintain a trial stepsize $\eta$ and a stepsize $\eta_{up}$ that is the smallest among all inadmissible stepsizes generated so far. If the trial stepsize $\eta$ violates condition (29), we assign its value to $\eta_{up}$ and further decrease it as

$$\eta \leftarrow \frac{\beta \eta^2}{\sigma}. \quad (30)$$

Equivalently, the trial stepsize after $i$ steps is given by $\eta^{(i)} = \sigma \beta^{2^i - 1}$. Once we find an admissible stepsize $\eta_{lo}$, the subroutine stops and outputs the interval $[\eta_{lo}, \eta_{up}]$. Note that compared with the conventional backtracking line search scheme, our scheme shrinks the stepsize much faster, as at the $i$-th step it is multiplied by $\beta^{2^i}$ instead of a constant factor $\beta$. On the other hand, the gap $\eta_{up}/\eta_{lo}$ could be larger than $1/\beta$, so we need the subroutine Bisection (to be discussed later) to further narrow down the interval. The following lemma ensures that the subroutine will eventually find an admissible stepsize under suitable conditions.

**Lemma 4.3.** Assume that there exist $\delta > 0$ and $L > 0$ such that $\|F(z) - P(z)\|_* \leq L \|z - z^-\|$ for any $z$ satisfying $\|z - z^-\| \leq \delta$. Then Subroutine 1 will always terminate in finite steps.

**Proof.** The proof is adapted from [MT20, Lemma 3.2]. We first show that $z(\eta; z^-)$ lies in a compact
set $C_\sigma \subset \mathcal{Z}$ for any $\eta \in [0, \sigma]$. For simplicity, in the following we denote $z(\eta; z^-)$ by $z$. By (27), there exists $w \in H(z)$ such that

$$
\eta P(z) + v^- + \eta w + \nabla \Phi(z) - \nabla \Phi(z^-) = 0,
$$

which implies that

$$
\langle \eta P(z) + v^- + \eta w + \nabla \Phi(z) - \nabla \Phi(z^-), z - z^- \rangle = 0.
$$

Since $z^-$ is feasible, we can also find $w^- \in \mathbb{R}^d$ such that $w^- \in H(z^-)$. By the strong convexity of $\Phi$ and the monotonicity of $P$ and $H$, from (31) we have

$$
-\langle v^-, z - z^- \rangle = \langle \nabla \Phi(z) - \nabla \Phi(z^-), z - z^- \rangle + \eta \langle P(z) + w, z - z^- \rangle
$$

$$
\geq \|z - z^-\|^2 + \eta \langle P(z^-) + w^-, z - z^- \rangle,
$$

which, together with the Cauchy-Schwarz inequality and the fact that $\eta \leq \sigma$, leads to

$$
\|z - z^-\| \leq \|v^-\|_* + \eta \|P(z^-) + w^-\|_* \leq \|v^-\|_* + \sigma \|P(z^-) + w^-\|_*.
$$

Since the upper bound in (32) is independent of $\eta$, this proves that $z(\eta; z^-)$ belongs to the compact set $C_\sigma := \{z \in \mathbb{R}^d : \|z - z^-\| \leq \|v^-\|_* + \sigma \|P(z^-) + w^-\|_*\} \cap \mathcal{Z}$ for all $0 \leq \eta \leq \sigma$.

Now we prove the lemma by contradiction. Denote the trial stepsize after $i$ steps in Subroutine 1 by $\eta^{(i)} := \sigma \beta^{2i-1}$. If Subroutine 1 does not terminate in finite steps, then for any integer $i \geq 0$ the trial stepsize $\eta^{(i)}$ does not satisfy condition (29), i.e.,

$$
\eta^{(i)} \|F(z(\eta^{(i)}; z^-)) - P(z(\eta^{(i)}; z^-))\|_* > \frac{1}{2} \alpha \|z(\eta^{(i)}; z^-) - z^-\|.
$$

Since both $F$ and $P$ are continuous on $\mathcal{Z}$ and $z(\eta^{(i)}; z^-)$ remains in a compact set $C_\sigma$ for all $i \geq 0$, we can upper bound $\|F(z(\eta^{(i)}; z^-)) - P(z(\eta^{(i)}; z^-))\|_*$ by a constant independent of $i$. Thus, we obtain $z(\eta^{(i)}; z^-) \to z^-$ by taking $i$ to infinity in (33). Hence, for $i$ large enough, we have $\|z(\eta^{(i)}; z^-) - z^-\| \leq \delta$, which by our assumption further implies that

$$
\|F(z(\eta^{(i)}; z^-)) - P(z(\eta^{(i)}; z^-))\|_* \leq L \|z(\eta^{(i)}; z^-) - z^-\|.
$$

Combining this with (33), 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.

If the initial trial stepsize $\sigma$ is admissible, we instead call ADVANCE (Subroutine 2) and set $\eta_{ho}$ as $\sigma$. In each step, we maintain a trial stepsize $\eta$ and a stepsize $\eta_{ho}$ that is the largest admissible stepsize generated so far. If the trial stepsize $\eta$ satisfies (29), we assign its value to $\eta_{ho}$ and increase it by

$$
\eta \leftarrow \frac{\eta^2}{\beta \sigma}.
$$

which is also equivalent to setting $\eta^{(i)} := \sigma \beta^{-2i+1}$ after $i$ steps. However, there is one technical issue we need to take care of. In theory, the iterative process (34) may never terminate, which is the case when (29) is satisfied for any $\eta > 0$. Hence, to avoid the endless loop, we will exit the line search when the residual defined in (15) at the point $z(\eta_{ho}; z^-)$ is smaller than a prescribed accuracy $\epsilon$. To see why this safeguard works, we first establish the following intermediate lemma that provides an upper bound on the the residual $\text{res}(z(\eta; z^-))$ in terms of the stepsize $\eta$.  

22
Lemma 4.4. Assume that $\nabla \Phi$ is $L_\Phi$-Lipschitz w.r.t. the norm $\| \cdot \|$ on $\mathcal{Z}$. Let $z = z(\eta; z^-)$ and suppose that $\eta$ satisfies (29). Then we have
\[
\text{res}(z) \leq \frac{1}{\eta} \left( \frac{\alpha}{2} + L_\Phi \right) \| z - z^- \| + \| v^- \|_v .
\]

Proof. Since $z$ is the solution of (27) at $\eta$, there exists $w \in H(z)$ such that
\[
0 = \eta P(z) + v^- + \eta w + \nabla \Phi(z) - \nabla \Phi(z^-)
\]
\[
\iff F(z) + w = \frac{1}{\eta} \left( \eta(F(z) - P(z)) - v^- - (\nabla \Phi(z) - \nabla \Phi(z^-)) \right) .
\]

Hence, by using the triangle inequality we obtain
\[
\text{res}(z) \leq \| F(z) + w \|_v \leq \frac{1}{\eta} \| F(z) - P(z) \|_v + \| v^- \|_v + \| \nabla \Phi(z) - \nabla \Phi(z^-) \|_v .
\]

The lemma now follows directly from (29) and the fact that $\nabla \Phi$ is $L_\Phi$-Lipschitz. \hfill \Box

we can use Lemma 4.4 to show that Subroutine 2 always terminates in a finite number of steps.

Lemma 4.5. Assume that $\nabla \Phi$ is $L_\Phi$-Lipschitz w.r.t. the norm $\| \cdot \|$ on $\mathcal{Z}$. Then Subroutine 2 will always terminate in finite steps.

Proof. Denote by $\eta^{(i)} := \sigma \beta^{-2^{i+1}}$ the trial stepsize used in Subroutine 2 after $i$ steps. By the method of contradiction, assume that the subroutine does not terminate in finite steps at the $k$-th iteration. We can see that this can only happen if $\eta^{(i)}$ satisfies (29) and $\text{res}(z(\eta^{(i)}; z_k)) \geq \epsilon$ for all $i$. Together with Lemma 4.4, this implies that
\[
\epsilon \leq \frac{1}{\eta^{(i)}} \left( \frac{\alpha}{2} + L_\Phi \right) \| z(\eta^{(i)}; z_k) - z_k \| + \| v_k \|_v \]

for all integer $i \geq 0$. On the other hand, since $\eta^{(i)}$ satisfies (29), we can choose $z_{k+1} = z(\eta^{(i)}; z_k)$ in Algorithm 1 and then apply Lemma 4.2 to get $\| z(\eta^{(i)}; z_k) - z_k \| = \| z_{k+1} - z_k \| \leq \sqrt{\frac{2}{1-\alpha}} \Phi(z^*, z_0)$. Hence, as $i \to +\infty$ the right-hand side of (35) tends to 0, which leads to a contradiction. \hfill \Box

Due to Lemmas 4.3 and 4.5, both BACKTRACK and ADVANCE are guaranteed to terminate in finite steps and return either an $\epsilon$-accurate solution or a well-defined interval $[\eta_0, \eta_{\text{up}}]$. In the latter case, we subsequently use the subroutine BISECTION to further refine this interval. To be specific, at each step we test the condition (29) at the geometric mean $\eta = \sqrt{\eta_0 \eta_{\text{up}}}$ and update the interval accordingly. This process is repeated until we have $\eta_{\text{up}}/\eta_0 \leq 1/\beta$. We can see that the stepsize $\eta_0$ will always be admissible while $\eta_{\text{up}}$ will always be inadmissible throughout the procedure. Hence, by definition $\eta_0$ will be a $\beta$-optimal stepsize when BISECTION terminates.

From the discussions above, we make the following observations:

- If we run Algorithm 2 without advancing (Option I), then either the output stepsize $\eta$ is $\beta$-optimal (we exit at line 5), or we have $\eta = \sigma$ (we exit at line 8);
- If we run Algorithm 2 with advancing (Option II), then either the output stepsize $\eta$ is $\beta$-optimal (we exit at line 5 or line 14), or we find an $\epsilon$-accurate solution $z$ (we exit at line 12).
Subroutine 2 \textsc{Advance}(z^-, \sigma, \alpha, \beta, \epsilon)

1: \textbf{Input:} current point $z^-$ $\in \mathbb{Z}$, initial stepsize $\sigma$, $\alpha \in (0, 1)$, $\beta \in (0, 1)$, $\epsilon > 0$
2: \textbf{Initialize:} set $\eta_l \leftarrow \sigma$, $\eta \leftarrow \sigma/\beta$
3: \textbf{while} $\text{res}(z(\eta_l; z^-)) > \epsilon$ and $\eta$ satisfies \eqref{cond} \textbf{do}
4: \quad Update $\eta_l \leftarrow \eta$ and increase the trial stepsize $\eta \leftarrow \eta^2/(\beta\sigma)$
5: \textbf{end while}
6: Set $\eta_{up} \leftarrow \eta$
7: \textbf{Return:} $\eta_l$, $\eta_{up}$

Subroutine 3 \textsc{Bisection}(z^-, \eta_l, \eta_{up}, \alpha, \beta)

1: \textbf{Input:} current point $z^- \in \mathbb{Z}$, admissible stepsize $\eta_l$, inadmissible stepsize $\eta_{up}$, $\alpha, \beta \in (0, 1)$
2: \textbf{while} $\eta_{up}/\eta_l > 1/\beta$ do
3: \quad $\eta \leftarrow \sqrt{\eta_l\eta_{up}}$
4: \quad if $\eta$ satisfy \eqref{cond} then
5: \quad \quad Update $\eta_l \leftarrow \eta$
6: \quad \quad else
7: \quad \quad Update $\eta_{up} \leftarrow \eta$
8: \quad \textbf{end if}
9: \textbf{end while}
10: \textbf{Return:} $\eta_l$

These facts are crucial for our complexity analysis and will be used repeatedly in later sections. Finally, while Lemmas 4.3 and 4.5 ensure that the number of steps taken in Subroutines 1 and 2 must be finite, we still need an explicit upper bound on the total number of subsolver calls to characterize the complexity of our line search scheme. The following proposition partially addresses this issue by giving such an upper bound in terms of the returned stepsize $\eta$ and the initial trial stepsize $\sigma$.

**Proposition 4.6.** Let $\eta$ be the stepsize returned by Algorithm 2. Then the procedure takes at most $2 \log_2 \log_{1/\beta} \left( \frac{\sigma^2}{\eta_l \eta_{up}} \right)$ calls to the optimistic subsolver in total if it is run without advancing, or at most $2 \log_2 \log_{1/\beta} \left( \frac{1}{\beta} \max \left\{ \frac{\sigma^2}{\eta_l^2}, \frac{\eta_{up}^2}{\sigma^2} \right\} \right)$ calls to the optimistic subsolver in total if it is run with advancing.

**Proof.** First, note that each step in Subroutines 1-3 consists of exactly one call to the optimistic subsolver. Hence, in the following we will focus on the total number of steps taken in Algorithm 2.

We first consider the line search scheme without advancing. Let $\tilde{\eta}_l$ and $\tilde{\eta}_{up}$ be the step sizes returned by Subroutine 1 and denote the total number of steps taken in Subroutine 1 by $k_B$. From the update rule \eqref{update}, we obtain

$$\tilde{\eta}_l = \sigma \beta^{2k_B-1} \quad \iff \quad k_B = \log_2 \log_{1/\beta} \frac{\sigma}{\beta \tilde{\eta}_l}.$$ 

Moreover, we have

$$\tilde{\eta}_l = \frac{\beta \tilde{\eta}_{up}^2}{\sigma} \quad \iff \quad \log_{1/\beta} \frac{\tilde{\eta}_{up}}{\tilde{\eta}_l} = \frac{1}{2} \log_{1/\beta} \frac{\sigma}{\beta \tilde{\eta}_l} = 2^{k_B-1}.$$ 

The interval $[\tilde{\eta}_l, \tilde{\eta}_{up}]$ then serves as the input of Subroutine 3. In this second phase, note that the quantity $\log_{1/\beta}(\tilde{\eta}_{up}/\tilde{\eta}_l)$ halves after each step, and the bisection phase terminates once $\log_{1/\beta}(\tilde{\eta}_{up}/\tilde{\eta}_l) \leq 1$. Hence, we conclude that Subroutine 3 takes exactly $(k_B - 1)$ steps. Taking the initial trial with $\sigma$ into account, we obtain that Algorithm 2 takes exactly $2k_B$ steps.
Finally, note that the final stepsize \( \eta \) returned by Subroutine 3 lies between \( \tilde{\eta}_0 \) and \( \tilde{\eta}_{up} \). This implies

\[
\tilde{\eta}_0 = \frac{\beta \tilde{\eta}_{up}^2}{\sigma} \geq \frac{\beta \eta^2}{\sigma}.
\]

Hence, we can bound

\[
2k_B = 2 \log_2 \log_1/\beta \frac{\sigma}{\beta \tilde{\eta}_0} \leq 2 \log_2 \log_1/\beta \frac{\sigma^2}{\beta^2 \eta^2},
\]

which completes the first part of the proof.

Now we turn to the line search scheme with advancing. If the final stepsize \( \eta \) is no larger than the initial stepsize \( \sigma \), then Algorithm 2 must have gone through the backtracking phase and the bisection phase, so the same analysis applies as above. If \( \eta > \sigma \), then Algorithm 2 must have gone through the advancing phase and the bisection phase. Let \( \tilde{\eta}_0 \) and \( \tilde{\eta}_{up} \) be the stepsizes returned by Subroutine 2 and denote the number of total steps taken in Subroutine 2 by \( k_A \). From (30), we have

\[
\tilde{\eta}_{up} = \sigma \left( \frac{1}{\beta} \right)^{2k_A - 1} \quad \Leftrightarrow \quad k_A = \log_2 \log_1/\beta \frac{\tilde{\eta}_{up}}{\beta \sigma}.
\]

Moreover, we have

\[
\tilde{\eta}_{up} = \frac{\tilde{\eta}_0^2}{\beta \sigma} \quad \Leftrightarrow \quad \log_1/\beta \frac{\tilde{\eta}_{up}}{\tilde{\eta}_0} = \frac{1}{2} \log_1/\beta \frac{\tilde{\eta}_{up}}{\beta \sigma} = 2k_A - 1.
\]

The interval \([\tilde{\eta}_0, \tilde{\eta}_{up}]\) serves as the input of Subroutine 3. Following the same reasoning as in the first part of the proof, we can show that Subroutine 3 takes exactly \((k_A - 1)\) steps. Taking the initial trial into account, we obtain that Algorithm 2 takes exactly \(2k_A\) steps.

Finally, we observe that the final stepsize \( \eta \) returned by Subroutine 3 lies between \( \tilde{\eta}_0 \) and \( \tilde{\eta}_{up} \). Hence,

\[
\tilde{\eta}_{up} = \frac{\eta^2}{\beta \sigma} \leq \frac{\eta^2}{\beta \sigma}.
\]

Therefore, we can write

\[
2k_A = 2 \log_2 \log_1/\beta \frac{\tilde{\eta}_{up}}{\beta \sigma} \leq 2 \log_2 \log_1/\beta \frac{\eta^2}{\beta^2 \sigma^2},
\]

and the proof is complete.  

\[\square\]

**Remark 4.2.** Note that the above result does not directly provide a complexity bound for our line search scheme, as it depends on the value of the returned stepsize \( \eta \). In the following sections, for each specific algorithm that we develop, we will establish lower and upper bounds on \( \eta \) and use them to derive an explicit upper bound on the number of calls to the optimistic subsolver during the whole process.

## 5 First-order Generalized Optimistic Method

In this section, we focus on the case where only first-order information of the smooth component \( f \) of the objective function in (1) is available. We also make the following assumption on \( f \).

**Assumption 5.1.** The operator \( F \) defined in (8) is \( L_1 \)-Lipschitz on \( Z \). Also, we have access to an oracle that returns \( F(z) \) for any given \( z \).
Algorithm 3 First-order optimistic method

1: Input: initial point $z_0 \in \mathcal{Z}$, strongly-convex parameter $\mu \geq 0$
2: Initialize: set $z_{-1} \leftarrow z_0$ and $P(z_0; I_{-1}) \leftarrow F(z_0)$
3: Option I (fixed stepsize scheme):
   4: Choose $M > 0$
   5: for iteration $k = 0, \ldots, N-1$ do
   6: Compute $z_{k+1} = \arg\min_{z \in \mathcal{Z}} \left\{ \left( \frac{1}{M} F(z_k) + \frac{1}{M+\mu} (F(z_k) - F(z_{k-1})) , z \right) + \frac{1}{M} h(z) + D_\Phi(z, z_k) \right\}$
   7: end for
8: Option II (line search scheme):
   9: Choose initial trial stepsize $\sigma_0 > 0$, line search parameters $\alpha \in (0,1]$ and $\beta \in (0,1)$
10: for iteration $k = 0, \ldots, N-1$ do
11: Set $\hat{\eta}_k = \eta_{k-1}/(1 + \eta_{k-1} \mu)$
12: Select $\eta_k$ by Algorithm 2 without advancing, where $\sigma = \sigma_k, z = z_k, P(z) = F(z_k)$,
13: and $\nu = \hat{\eta}_k (F(z_k) - F(z_{k-1}))$
14: Compute $z_{k+1} = \arg\min_{z \in \mathcal{Z}} \left\{ (\eta_k F(z_k) + \hat{\eta}_k (F(z_k) - F(z_{k-1}))) , z \right\}$
15: Set $\sigma_{k+1} \leftarrow \eta_k / \beta$
16: end for

Under the above assumption, we have from Definition 2.4 that
\[
\| F(z) - F(z_k) \|_* \leq L_1 \| z - z_k \|, \quad \forall z \in \mathcal{Z}.
\] (36)

In this case, a natural choice for the approximation function is $P(z; I_k) := F(z_k)$, and accordingly at iteration $k$, our proposed optimistic method in Algorithm 1 aims to find $z \in \mathcal{Z}$ and $\eta > 0$ such that
\[
z = \arg\min_{w \in \mathcal{Z}} \left\{ (\eta F(z_k) + \hat{\eta}_k (F(z_k) - F(z_{k-1}))) , w - z_k \right\} + \eta h(w) + D_\Phi(w, z_k) \right\},
\] (37)
\[
\eta \| F(z) - F(z_k) \|_* \leq \frac{\alpha}{2} \| z - z_k \|.
\] (38)

In particular, the optimistic subsolver defined in Definition 4.1 solves the subproblem (37), which is a standard update in first-order methods.

As described in Algorithm 3, we will study two different approaches for selecting the stepsize $\eta_k$ in the first-order optimistic method: (i) fixed stepsize scheme; (ii) adaptive line search scheme. As the name suggests, in the first approach a fixed stepsize $\eta_k \equiv \eta$ is properly selected to ensure that the condition in (38) is always satisfied. In the second approach, we instead select the stepsize according to the line search scheme described in Algorithm 2. One major advantage of the second approach is that it does not require any prior knowledge of the Lipschitz constant of $F$, while in the former we at least need to know an upper bound on the Lipschitz constant.

5.1 Fixed Stepsize Scheme

In light of the inequality in (36), if we select the stepsize such that $\eta_k \leq \frac{1}{2L_1}$, then the condition in (38) required for the convergence of the optimistic method is always satisfied with $\alpha = 1$. Hence, we can simply select a fixed stepsize of $\eta_k \equiv 1/M$ with $M \geq 2L_1$ for all $k \geq 0$. In this case, the update of the first-order optimistic method becomes
\[
z_{k+1} = \arg\min_{z \in \mathcal{Z}} \left\{ \left( \frac{1}{M} F(z_k) + \hat{\eta}_k (F(z_k) - F(z_{k-1})) , z - z_k \right) + \frac{1}{M} h(z) + D_\Phi(z, z_k) \right\},
\] (39)
where the correction coefficient $\hat{\eta}_k$ will be chosen as in Proposition 4.1 based on the strongly-convex parameter $\mu$. By viewing (39) as an instance of our generalized optimistic method, the convergence guarantees immediately follow from Proposition 4.1, which are summarized in the theorem below.

**Theorem 5.1.** Assume that the operator $F$ is $L_1$-Lipschitz on $\mathcal{Z}$. Let $\{z_k\}_{k \geq 0}$ be the iterates generated by (39). Given $M \geq 2L_1$, the following statements hold:

(a) Under Assumption 2.1, set $\hat{\eta}_k = 1/M$. Then we have $D_{\phi}(z^*, z_k) \leq 2D_{\phi}(z^*, z_0)$ for any $k \geq 0$. Moreover, for any $z = (x, y) \in \mathcal{X} \times \mathcal{Y}$, we have

$$
\ell(\bar{x}_N, y) - \ell(x, \bar{y}_N) \leq \frac{MD_{\phi}(z, z_0)}{N},
$$

where $\bar{z}_N = (\bar{x}_N, \bar{y}_N)$ is given by $\bar{z}_N = \frac{1}{N} \sum_{k=0}^{N-1} z_{k+1}$.

(b) Under Assumption 2.2, set $\hat{\eta}_k = 1/(M + \mu)$. Then we have

$$
D_{\phi}(z^*, z_N) \leq 2D_{\phi}(z^*, z_0) \left(\frac{M}{\mu + M}\right)^N.
$$

The result in Theorem 5.1 shows that the proposed first-order generalized optimistic method converges at a sublinear rate of $O(1/N)$ for convex-concave saddle point problems, and converges at a linear rate of $O((M/(\mu + M))^N)$ for strongly-convex-strongly-concave saddle point problems. Asymptotically, these bounds are the same as those derived in [MOP20a; MOP20b; GBVVL19] for the OGDA method. However, these prior works are restricted to the unconstrained saddle point problems in the Euclidean setup. In this regard, our results can be viewed as a natural extension of the OGDA method and its analysis to the more general composite saddle point problems equipped with an arbitrary Bregman distance.

We should also point out that the forward-reflected-backward splitting method in [MT20] for monotone inclusion problems and the operator extrapolation method in [KLL20] for constrained variational inequalities share similar update rule and convergence analysis as the above first-order optimistic method. Moreover, all differ from the classical Popov’s method in (19) except in the unconstrained setting. In particular, a crucial advantage of our method and those in [MT20; KLL20] is that they require one projection onto the feasible set per iteration, while the update in (19) requires two projections per iteration. On the other hand, our setting is more general than the mentioned works in the sense that they only considered either the Euclidean norm (i.e., $\Phi = \frac{1}{2}\|\cdot\|_2^2$) or smooth objectives (i.e., $h_1 = h_2 \equiv 0$).

### 5.2 Adaptive Line Search Scheme

While simple in concept, the fixed stepsize scheme discussed in the previous section requires knowledge of the Lipschitz constant of $F$, which may be difficult to compute in practice. Also, choosing the same stepsize for all iterations could fail to exploit the local geometry and be overly conservative when it is possible to take larger steps. Next, we overcome these issues by choosing the stepsizes adaptively with the line search scheme proposed in Algorithm 2. Our proposed method is shown as Option II in Algorithm 3. Specifically, we use the version in Algorithm 2 that does not execute the advancing subroutine. In addition, at the $k$-th iteration ($k \geq 1$), we let our line search procedure start from $\sigma_k = \eta_{k-1}/\beta$, where $\eta_{k-1}$ is the stepsize we choose at the last iteration.
Similar initialization strategy is also used in [Mal17; MT20]. We make two observations regarding the line search procedure. To begin with, note that Assumption 5.1 ensures that the condition in Lemma 4.3 is satisfied, and hence the line search scheme is guaranteed to terminate in finite steps and hence the stepsizes \( \{\eta_k\}_{k \geq 0} \) are well defined. Moreover, as discussed in Section 4.1, the stepsize \( \eta_k \) is either \( \beta \)-optimal or satisfies \( \eta_k = \sigma_k \). Note that to obtain convergence rates and bound the number of subsolver calls required for the line search scheme, we need to bound the stepsize \( \eta_k \) away from zero. The following lemma serves this purpose and gives a lower bound on \( \eta_k \) when it is \( \beta \)-optimal.

**Lemma 5.2.** Suppose that \( F \) is \( L_1 \)-Lipschitz on \( Z \). If the stepsize \( \eta_k \) at the \( k \)-th iteration is \( \beta \)-optimal, then \( \eta_k > \alpha \beta/(2L_1) \).

**Proof.** To simplify the notation, we drop the subscript \( k \) and denote \( z_k \) by \( z^- \). By definition, there exists a stepsize \( \eta' \) such that \( \eta < \eta' \leq \eta/\beta \) and \((\eta', z')\) violates (38) with \( z' \) computed via (37). Hence,

\[
\frac{\alpha}{2} \|z' - z^-\| < \eta' \|F(z') - F(z^-)\|_* \leq \eta' L_1 \|z' - z^-\|,
\]

where we used the fact that \( F \) is \( L_1 \)-Lipschitz in the last inequality. This implies that \( \eta' > \alpha/(2L_1) \), and we immediately obtain that \( \eta \geq \beta \eta' > \alpha \beta/(2L_1) \).

Combining all pieces above, we arrive at the following theorem by specializing Propositions 4.1 and 4.6.

**Theorem 5.3.** Assume that the operator \( F \) is \( L_1 \)-Lipschitz on \( Z \). Let \( \{z_k\}_{k \geq 0} \) be the iterates generated by (37), where the stepsizes are determined by Algorithm 2 without the advancing subroutine and with parameters \( \alpha \in (0, 1) \), \( \beta \in (0, 1) \), and an initial trial stepsize \( \sigma_0 \). Then the following statements hold:

(a) Under Assumption 2.1, set \( \hat{\eta}_k = \eta_{k-1} \). Then we have \( D_\Phi(z^*, z_k) \leq \frac{2}{2-\alpha} D_\Phi(z^*, z_0) \) for any \( k \geq 0 \). Moreover, for any \( z = (x, y) \in X \times Y \), we have

\[
\ell(x_N, y) - \ell(x, y_N) \leq \frac{2L_1 D_\Phi(z, z_0)}{\alpha \beta N} + \frac{D_\Phi(z, z_0)}{(1 - \beta) \sigma_0 N^2},
\]

where \( z_N = (x_N, y_N) \) is given by \( z_N = \frac{1}{\sum_{k=0}^{N-1} \eta_k} \sum_{k=0}^{N-1} \eta_k z_{k+1} \).

(b) Under Assumption 2.2, set \( \hat{\eta}_k = \eta_{k-1}/(1 + \mu \eta_{k-1}) \). Then we have

\[
D_\Phi(z^*, z_N) \leq \frac{2C}{2 - \alpha} D_\Phi(z^*, z_0) \left(1 + \frac{\mu \alpha \beta}{2L_1}\right)^{-N},
\]

where \( C = \exp \left( \frac{\alpha \beta}{2(1 - \beta) \sigma_0 L_1} \frac{\alpha \beta \mu/(2L_1)}{1 + \alpha \beta \mu/(2L_1)} \right) \) is a constant that only depends on \( \alpha, \beta, \sigma_0, L_1 \) and \( \mu \).

(c) In both cases of (a) and (b) above, the total number of calls to the optimistic subsolver after \( N \) iterations can be bounded by

\[
\max \left\{ 2N, 2N \log_2 \left( 4 + \frac{2}{N} \log_{1/\beta} \frac{2\sigma_0 L_1}{\alpha} \right) \right\}.
\]

28
Proof. Since \( \{\eta_k\}_{k \geq 0} \) is determined by Algorithm 2 without advancing, we recall that either \( \eta_k \) is \( \beta \)-optimal or \( \eta_k = \sigma_k \). In the former case, we also have \( \eta_k \geq (\alpha \beta)/(2L_1) \) by Lemma 5.2. From these observations, we shall prove by induction that for all \( k \geq 0 \),

\[
\eta_k \geq \min \left\{ \frac{\sigma_0}{\beta^k}, \frac{\alpha \beta}{2L_1} \right\}. \tag{41}
\]

It is easy to verify that (41) holds for \( k = 0 \). Now assume that (41) holds for all \( 0 \leq k \leq l - 1 \). Then we either have \( \eta_l \geq (\alpha \beta)/(2L_1) \) if \( \eta_l \) is \( \beta \)-optimal, or \( \eta_l = \sigma_l = \eta_{l-1} \beta \geq \frac{1}{\beta} \min \left\{ \frac{\sigma_0}{\beta^{l-1}}, \frac{\alpha \beta}{2L_1} \right\} = \min \left\{ \frac{\sigma_0}{\beta^l}, \frac{\alpha}{2L_1} \right\}. \)

In both cases, we can see that (41) is satisfied for \( k = l \), and hence the claim is proved by induction.

**Proof of Part (a).** From (41), we have

\[
\frac{1}{\eta_k} \leq \max \left\{ \frac{\beta^k}{\sigma_0}, \frac{2L_1}{\alpha \beta} \right\} \leq \frac{\beta^k}{\sigma_0} + \frac{2L_1}{\alpha \beta}. \tag{42}
\]

By summing (42) over \( k = 0, \ldots, N - 1 \) and using the fact that \( \beta < 1 \), we obtain

\[
\sum_{k=0}^{N-1} \frac{1}{\eta_k} \leq \frac{1}{(1 - \beta)\sigma_0} + \frac{2L_1}{\alpha \beta} N. \tag{43}
\]

Now combining (43) with the Cauchy-Schwarz inequality leads to

\[
\sum_{k=0}^{N-1} \eta_k \geq N^2 \left( \sum_{k=0}^{N-1} \frac{1}{\eta_k} \right)^{-1} \geq \left( \frac{1}{(1 - \beta)\sigma_0 N^2} + \frac{2L_1}{\alpha \beta N} \right)^{-1}. \]

The rest follows from Proposition 4.1.

**Proof of Part (b).** We can write

\[
\log \left( \prod_{k=0}^{N-1} (1 + \eta_k \mu) \right) = \sum_{k=0}^{N-1} \log (1 + \eta_k \mu) = \sum_{k=0}^{N-1} \log \left( 1 + \frac{1}{1/(\eta_k \mu)} \right).
\]

Note that \( \log(1 + \frac{t}{1}) \) is a convex and monotonically decreasing function of \( t \) on \( \mathbb{R}_{++} \). Hence, by Jensen’s inequality and (43) we have

\[
\log \left( \prod_{k=0}^{N-1} (1 + \eta_k \mu) \right) \geq N \log \left( 1 + \mu \left( \frac{1}{N} \sum_{k=0}^{N-1} \frac{1}{\eta_k} \right)^{-1} \right) \geq N \log \left( 1 + \frac{\alpha \beta \mu}{2L_1} \right) + N \log \left( 1 - \frac{c_1}{1 + c_1} \frac{1}{1 + c_2 N} \right), \tag{44}
\]

where \( c_1 := \alpha \beta \mu/2L_1 \) and \( c_2 := 2(1 - \beta)\sigma_0 L_1/\alpha \beta \). Moreover, by using the elementary inequality \( \log(1 + x) \geq x/(1 + x) \) for all \( x \geq -1 \), we have

\[
N \log \left( 1 - \frac{c_1}{1 + c_1} \frac{1}{1 + c_2 N} \right) \geq -\frac{c_1 N}{1 + (1 + c_1)c_2 N} \geq -\frac{c_1}{(1 + c_1)c_2} = -\frac{\alpha \beta}{2(1 - \beta)\sigma_0 L_1} \frac{\alpha \beta \mu/(2L_1)}{1 + \alpha \beta \mu/(2L_1)}. \tag{45}
\]
Combining (44) and (45), we arrive at
\[
\prod_{k=0}^{N-1} (1 + \eta_k \mu) \geq \left(1 + \frac{\alpha \beta \mu}{2L_1}\right)^N \exp \left(-\frac{1}{2} \frac{\alpha \beta}{2(1 - \beta) \sigma_0 L_1} \frac{\alpha \beta \mu}{2L_1}\right).
\]
The rest follows from Proposition 4.1.

**Proof of Part (c).** By Proposition 4.6, we can bound the number of calls to the optimistic subsolver at the k-th iteration by \(2 \log_2 \log_{1/\beta} \frac{\alpha}{\beta^2 \eta_0^2} \). Also note that \(\sigma_k = \eta_{k-1}/\beta\) for \(k \geq 1\). Hence, the total number of calls after \(N\) iterations can be bounded by
\[
2 \log_2 \log_{1/\beta} \frac{\alpha}{\beta^2 \eta_0^2} + \sum_{k=1}^{N-1} 2 \log_2 \log_{1/\beta} \frac{\eta_{k-1}}{\beta^2 \eta_k^2} \leq 2N \log_2 \left(\frac{1}{N} \left(\log_1 \frac{\alpha}{\beta} \frac{\sigma_0^2}{\eta_0^2} + \sum_{k=1}^{N-1} \log_1 \frac{\eta_{k-1}}{\eta_k^2}\right)\right)
\]
\[
= 2N \log_2 \left(4 + \frac{2}{N} \log_{1/\beta} \frac{\beta \sigma_0}{\eta_{N-1}}\right),
\] (46)
where the inequality holds due to Jensen’s inequality. Since \(\eta_{N-1} \geq \min \left\{ \frac{\sigma_0}{\beta N}, \frac{\alpha \beta}{2L_1}\right\}\), the bound in (46) proves the claim. \(\square\)

Asymptotically, Theorem 5.3 shows that the line search scheme achieves the same convergence rates as the fixed stepsize scheme where the stepsize in all iterations are chosen as \((\alpha/\beta)/(2L_1)\), although it does not require any prior information of the Lipschitz constant \(L_1\). Moreover, it shows the total number of calls to a subsolver of (37) required by the line search scheme is \(O(2N \log_2(4 + 1/\beta))\), and hence, the average number of calls per iteration can be bounded by a constant close to 4 when \(N\) is large.

## 6 Second-Order Generalized Optimistic Method

In this section, we instantiate Algorithm 1 with a second-order oracle to derive a novel second-order optimistic method for solving the saddle point problem in (1). For technical reasons, throughout the section, we restrict ourselves to the case where the Bregman function \(\Phi\) is \(L_\Phi\)-smooth on \(Z\), i.e., \(\nabla \Phi\) is \(L_\Phi\)-Lipschitz continuous. We also require the following assumption on the smooth component of the objective in (2.4) denoted by \(f\).

**Assumption 6.1.** The operator \(F\) defined in (8) is second-order \(L_2\)-smooth on \(Z\). Also, we have access to an oracle that returns \(F(z)\) and \(DF(z)\) for any given \(z\).

Under Assumption 6.1, it follows from (6) that
\[
\|F(z) - F(z_k) - DF(z_k)(z - z_k)\|_* \leq \frac{L_2}{2} \|z - z_k\|^2, \quad \forall z \in Z,
\] (47)
which suggests we can choose the approximation function as \(P(z; I_k) := T^{(1)}(z; z_k) = F(z_k) + DF(z_k)(z - z_k)\). Note that computing this approximation function requires access to the operator \(DF(\cdot)\), which involves the second-order derivative of the function \(f\). Therefore, we refer to the resulting algorithm as the second-order generalized optimistic method. The update for the proposed second-order optimistic method in Algorithm 1 can be written as
\[
0 \in \eta_k T^{(1)}(z_{k+1}; z_k) + \tilde{\eta}_k(F(z_k) - T^{(1)}(z_k; z_{k-1})) + \eta_k H(z_{k+1}) + \nabla \Phi(z_{k+1}) - \nabla \Phi(z_k),
\] (48)
where the condition in (24) on \( \eta_k \) becomes

\[
\eta_k \| F(z_{k+1}) - T^{(1)}(z_{k+1}; z_k) \| \leq \frac{\alpha}{2} \| z_{k+1} - z_k \|. \tag{49}
\]

To begin with, we explain how the second-order information could help accelerate convergence. Intuitively, in light of (49), choosing a more accurate approximation \( T^{(1)}(\cdot; z_k) \) allows us to take a larger stepsize \( \eta_k \) than the first-order method in (37). Moreover, as Proposition 4.1 shows, the speed of convergence for our generalized optimistic method depends on the choice of stepsizes \( \{ \eta_k \}_{k \geq 0} \), and larger stepsizes lead to faster convergence. More precisely, by (47) the left hand side of (49) can be bounded above by \( \frac{1}{2} \eta_k L_2 \| z_{k+1} - z_k \|^2 \), which suggests that we can pick our stepsize \( \eta_k \) as large as

\[
\eta_k \approx \frac{\alpha}{L_2 \| z_{k+1} - z_k \|}. \tag{50}
\]

Hence, as the iterates \( \{ z_k \}_{k \geq 0} \) approach the optimal solution and the displacement \( \| z_{k+1} - z_k \| \) becomes smaller, we can afford selecting a larger stepsize and accelerate convergence.

The above argument for faster convergence relies crucially on the premise that the stepsize \( \eta_k \) is adaptive to the displacement \( \| z_{k+1} - z_k \| \), i.e., at least on the order of \( \frac{1}{\| z_{k+1} - z_k \|} \). As mentioned in Section 4, a major challenge here is the interdependence between \( \eta_k \) and \( z_{k+1} \): the iterate \( z_{k+1} \) depends on \( \eta_k \) via (48), while the stepsize \( \eta_k \) depends on \( z_{k+1} \) as illustrated in (49) and (50). Hence, we need to simultaneously select \( \eta_k \) and \( z_{k+1} \) to satisfy the above requirements. This goal can be achieved by using the line search scheme discussed in Section 4.1. Moreover, to reach the full potential of the second-order method and pick the largest possible stepsize, we propose to select the stepsize by the more aggressive version of Algorithm 2 with the advancing subroutine, at the cost of a potential increase in the number of calls to the optimistic subsolver. In addition, similar to the line search scheme for the first-order optimistic method, we choose the initial trial stepsize as \( \sigma_k = \eta_{k-1}/\beta \) for \( k \geq 1 \). The second-order optimistic method is formally described in Algorithm 4.

Two remarks on the line search scheme in Algorithm 4 follow. First, by using the results in Lemmas 4.3 and 4.5 we can guarantee that the line search scheme always terminates in finite steps, and hence the stepsize \( \{ \eta_k \}_{k \geq 0} \) are well-defined. To see why Lemma 4.3 applies, note that for any \( \| z - z_k \| \leq \delta \), by (47) we have \( \| F(z) - T^{(1)}(z; z_k) \| \leq (L_2/2) \| z - z_k \|^2 \leq (L_2/\delta) \| z - z_k \| \). Hence, the condition in Lemma 4.3 is indeed satisfied. Second, at the \( k \)-th iteration, the optimistic subsolver in this case is required to solve a subproblem of \( z \) in the form of

\[
0 \in \eta T^{(1)}(z; z_k) + v_k + \eta H(z) + \nabla \Phi(z) - \nabla \Phi(z_k) \tag{51}
\]

for given \( v_k = \hat{\eta}_k (F(z_k) - T^{(1)}(z_k; z_k-1)) \) and \( \eta > 0 \) (cf. Definition 4.1). Note that \( T^{(1)}(z; z_k) \) is an affine operator, and hence this subproblem can be solved with respect to \( z \) efficiently by standard solvers in some special cases. For instance, in the Euclidean setup where \( \Phi(z) = \frac{1}{2} \| z \|^2 \), the subproblem (51) is equivalent to an affine variational inequality [FP03] for constrained saddle point problems (where \( h_1 \equiv 0 \) on \( \mathcal{X} \) and \( h_2 \equiv 0 \) on \( \mathcal{Y} \)), and further reduces to a system of linear equations for unconstrained saddle point problems (where \( \mathcal{X} = \mathbb{R}^n \) and \( \mathcal{Y} = \mathbb{R}^m \)).

### 6.1 Intermediate Results

To establish the convergence properties of the proposed second-order optimistic method, we first state a few intermediate results that will be used in the following sections. To begin with, recall
Algorithm 4 Second-order optimistic method

1: **Input:** initial point $z_0 \in \mathcal{Z}$, initial trial stepsize $\sigma_0$, strongly-convex parameter $\mu \geq 0$, line search parameters $\alpha, \beta \in (0, 1)$, and $\epsilon > 0$
2: **Initialize:** set $z_{-1} \leftarrow z_0$ and $P(z_0; I_{-1}) \leftarrow F(z_0)$
3: **for** iteration $k = 0, \ldots, N - 1$ **do**
4:   Set $\eta_k = \eta_{k-1}/(1 + \eta_{k-1}\mu)$
5:   Select $\eta_k$ by Algorithm 2 with advancing, where $\sigma = \sigma_k$, $z^- = z_k$, $P(z) = T(1)(z; z_k)$, and $v^- = \eta_k(F(z_k) - T(1)(z_k; z_{k-1}))$
6:   Compute $z_{k+1}$ by solving the monotone inclusion problem

$$0 \in \eta_kT(1)(z; z_k) + \hat{\eta}_k(F(z_k) - T(1)(z_k; z_{k-1})) + \eta_k H(z) + \nabla \Phi(z) - \nabla \Phi(z_k)$$

7: **if** $\text{res}(z_{k+1}) \leq \epsilon$ **then**
8:   Return $z_{k+1}$ as an $\epsilon$-accurate solution
9: **else**
10:   Set $\sigma_{k+1} \leftarrow \eta_k/\beta$
11: **end if**
12: **end for**

that at the $k$-th iteration our goal is to find a pair $(\eta, z)$ such that

$$\eta\|F(z) - T(1)(z; z_k)\|_* \leq \frac{\alpha}{2}\|z - z_k\|, \quad (52)$$

where $z$ is computed via (51). And by using Algorithm 2 with advancing, at the end of the line search scheme either the stepsize $\eta_k$ is $\beta$-optimal for condition (52) (cf. Definition 4.2), or we succeed in finding an $\epsilon$-accurate point with $\text{res}(z_k) \leq \epsilon$. In the former case, the following lemma provides a lower bound on $\eta_k$ analogous to Lemma 5.2.

**Lemma 6.1.** Suppose that $F$ is second-order $L_2$-smooth and $\Phi$ is $L_\Phi$-smooth on $\mathcal{Z}$. Further define $v_k := \eta_k(F(z_k) - T(1)(z_k; z_{k-1}))$. If the stepsize $\eta_k$ at the $k$-th iteration is $\beta$-optimal, then

$$\eta_k \geq \frac{\alpha \beta^2/L_2}{L_\Phi^{3/2}\|z_{k+1} - z_k\| + (\beta + L_\Phi^{1/2})\|v_k\|_*}.$$

**Proof.** See Appendix D.1. \qed

**Lemma 6.2.** Let $\{\eta_k\}_{k \geq 0}$ be the stepizes in (48) generated by Algorithm 4. If $\text{res}(x_N) > \epsilon$, then in the convex-concave setting we have

$$\sum_{k=0}^{N-1} \frac{1}{\eta_k^2} \leq \gamma_2^2 L_2^2 D_\Phi(z^*, z_0),$$

and in the strongly-convex-strongly-concave setting we have

$$\sum_{k=0}^{N-1} \left( \frac{1}{\eta_k^2} \prod_{l=0}^{k-1} (1 + \eta_l\mu) \right) \leq \gamma_2^2 L_2^2 D_\Phi(z^*, z_0), \quad (53)$$

where $\gamma_2$ is defined as

$$\gamma_2 = \sqrt{\frac{2}{1 - \alpha}} \left( \frac{L_\Phi^{3/2}}{\alpha \beta^2} + \frac{\beta + L_\Phi^{1/2}}{2\beta^2} \right). \quad (54)$$
Proof. Note that if \( \text{res}(x_N) > \epsilon \), then all the stepsizes \( \eta_0, \eta_1, \ldots, \eta_{N-1} \) returned by Algorithm 2 with advancing are \( \beta \)-optimal. Similar to the proof of Lemma 4.2, we will prove both results in a unified way by regarding the convex-concave setting as a special case of the strongly-convex-strongly-concave setting with \( \mu = 0 \).

By the choice of \( \hat{\eta}_k \) and (49), we have
\[
\|v_k\|_* = \frac{\eta_{k-1}}{1 + \eta_{k-1}\mu} \|F(z_k) - T^{(1)}(z_k; z_{k-1})\|_* \leq \frac{\alpha}{2(1 + \eta_{k-1}\mu)} \|z_k - z_{k-1}\|.
\]

Together with Lemma 6.1, this leads to
\[
\frac{1}{\eta_k} \leq \frac{L_2}{\alpha\beta^2} \left(L_\Phi^{3/2}\|z_{k+1} - z_k\| + (\beta + L_\Phi^{1/2})\|v_k\|_*\right) \leq c_1 L_2\|z_{k+1} - z_k\| + c_2 L_2 \|z_k - z_{k-1}\| \frac{1}{1 + \eta_{k-1}\mu},
\]

where we let \( c_1 := L_\Phi^{3/2}/(\alpha\beta^2) \) and \( c_2 := (\beta + L_\Phi^{1/2})/(2\beta^2) \) to simplify the notation. Moreover, by using Young’s inequality we obtain
\[
\frac{1}{\eta_k} \leq (c_1^2 + c_1 c_2) L_2^2\|z_{k+1} - z_k\|^2 + \frac{(c_2^2 + c_1 c_2) L_2^2}{(1 + \eta_{k-1}\mu)^2} \|z_k - z_{k-1}\|^2.
\]

Multiplying both sides of the above inequality by \( \prod_{l=0}^{k-1} (1 + \eta_l\mu) \), we further have
\[
\frac{1}{\eta_k^\prime} \prod_{l=0}^{k-1} (1 + \eta_l\mu) \leq (c_1^2 + c_1 c_2) L_2^2\|z_{k+1} - z_k\|^2 \prod_{l=0}^{k-1} (1 + \eta_l\mu) + \frac{(c_2^2 + c_1 c_2) L_2^2}{(1 + \eta_{k-1}\mu)^2} \|z_k - z_{k-1}\|^2 \prod_{l=0}^{k-2} (1 + \eta_l\mu)
\]
\[
\leq (c_1^2 + c_1 c_2) L_2^2\|z_{k+1} - z_k\|^2 \prod_{l=0}^{k-1} (1 + \eta_l\mu) + (c_2^2 + c_1 c_2) L_2^2\|z_k - z_{k-1}\|^2 \prod_{l=0}^{k-2} (1 + \eta_l\mu).
\]

By summing both sides of (55) over \( k = 0, 1, \ldots, N-1 \) and applying Lemma 4.2, we get the desired result in Lemma 6.2. \( \square \)

### 6.2 Convergence Analysis: Convex-Concave Case

Now we turn to the convergence analysis of Algorithm 4 and start with the case where the smooth component \( f \) of the objective in (1) is merely convex-concave. By instantiating Proposition 4.1, we obtain the following ergodic convergence rate result.

**Theorem 6.3.** Suppose that Assumptions 2.1 and 6.1 hold and the Bregman function \( \Phi \) is \( L_\Phi \)-smooth. Let \( \{z_k\}_{k\geq0} \) be the iterates generated by Algorithm 4, where \( \hat{\eta}_k = \eta_k - 1, \sigma_0 > 0, \alpha \in (0, 1), \beta \in (0, 1), \) and \( \epsilon > 0 \). Then, we have \( D_\Phi(z^*, z_k) \leq \frac{2}{\alpha} D_\Phi(z^*, z_0) \) for any \( k \geq 0 \). Moreover, we either have \( \text{res}(z_N) \leq \epsilon \), or
\[
\ell(\bar{x}_N, y) - \ell(x, \bar{y}_N) \leq \gamma_2 L_2 D_\Phi(z, z_0) \sqrt{D_\Phi(z^*, z_0)} N^{-\frac{1}{2}}
\]
for any \( z = (x, y) \in X \times Y \), where \( \bar{z}_N = (\bar{x}_N, \bar{y}_N) \) is given by \( \bar{z}_N = \frac{1}{\sum_{k=0}^{N-1} \eta_k^\prime} \sum_{k=0}^{N-1} \eta_k z_{k+1} \) and \( \gamma_2 \) is defined in (54).
Proof. If \( \text{res}(z_N) \leq \epsilon \), then we are done. Otherwise, we can apply Lemma 6.2 to derive a bound on \( \sum_{k=0}^{N-1} \eta_k \). By Hölder’s inequality, we have \( \left( \sum_{k=0}^{N-1} \eta_k \right)^\frac{2}{3} \left( \sum_{k=0}^{N-1} \frac{1}{\eta_k} \right)^\frac{1}{3} \geq N \). Together with Lemma 6.2, this further implies

\[
\sum_{k=0}^{N-1} \eta_k \geq \left( \sum_{k=0}^{N-1} \frac{1}{\eta_k} \right)^{-\frac{1}{2}} N^\frac{2}{3} \geq \frac{1}{\gamma_2 L_2 \sqrt{D_\Phi(z^*, z_0)}} N^\frac{2}{3}.
\]

The rest follows from Proposition 4.1.

Remark 6.1. Note that the constant \( \gamma_2 \) is independent of the problem parameters and solely depends on the implementation parameters. For instance, if we select the line search parameters as \( \alpha = 0.5 \) and \( \beta = 0.9 \) and focus on the Euclidean setup where \( L_\phi = 1 \), then we have \( \gamma_2 \approx 7.3 \).

Theorem 6.3 shows that the second-order optimistic method converges at a rate of \( \mathcal{O}(N^{-3/2}) \) in terms of the primal-dual gap, which is 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 \( \epsilon \), the proposed second-order optimistic method requires at most \( \mathcal{O}(1/\epsilon^{2/3}) \) 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 6.9.

Comparison with [MS12] and [BL20]. Similar complexity bounds are also reported in [MS12; BL20] for extragradient-type second-order methods. In comparison, our method is based on a different algorithmic idea and has a lower per-iteration computational cost, since it only requires the gradient information at one point and solves one single subproblem instead of two in each iteration. Moreover, in [MS12] the authors only considered the Euclidean setup, while [BL20] only discussed implementation details for the special case of unconstrained problems under stronger assumptions (\( F \) is \( L_1 \)-Lipschitz and strongly monotone).

6.3 Convergence Analysis: Strongly-Convex-Strongly-Concave Case

Next, we proceed to the setting where the smooth component \( f \) of the objective in (1) is \( \mu \)-strongly-convex-strongly-concave. We first define a positive decreasing sequence \( \{\zeta_k\}_{k \geq 0} \) by

\[
\zeta_k = \begin{cases} 1, & \text{if } k = 0; \\ \prod_{l=0}^{k-1} (1 + \eta l)^{-1}, & \text{if } k \geq 1. 
\end{cases}
\]

(57)

According to the result in part (b) of Proposition 4.1, the Bregman distance to the optimal solution \( D_\Phi(z^*, z_k) \) can be bounded above by

\[
D_\Phi(z^*, z_k) \leq \frac{2D_\Phi(z^*, z_0)}{2 - \alpha} \zeta_k,
\]

(58)

where \( \alpha \in (0, 1) \) is a line search parameter. In particular, the iterate \( z_k \) achieves \( \epsilon \)-accuracy with \( D_\Phi(z^*, z_k) \leq \epsilon \) when we have \( \zeta_k \leq \frac{(2-\alpha)\epsilon}{2D_\Phi(z^*, z_0)} \). Hence, in the following we will characterize the convergence behavior of the sequence \( \{\zeta_k\}_{k \geq 0} \), which immediately implies that of \( D_\Phi(z^*, z_k) \).

To begin with, we present the following lemma, which is a direct corollary of Lemma 6.2.
Lemma 6.4. Let \( \{\eta_k\}_{k \geq 0} \) be the stepsizes in (48) generated by Algorithm 4 and \( \{\zeta_k\}_{k \geq 0} \) be defined in (57). If \( \text{res}(x_N) > \epsilon \), then for any \( 0 \leq k_1 < k_2 \leq N \) we have

\[
\frac{1}{\zeta_{k_2}} \geq \frac{1}{\zeta_{k_1}} + \frac{1}{\gamma_2 \kappa_2(z_0)} \left( \sum_{k=k_1}^{k_2-1} \frac{1}{\zeta_k} \right)^{\frac{2}{3}},
\]

where \( \kappa_2(z_0) \) and \( \gamma_2 \) are defined in (2) and (54), respectively.

Proof. By the definition of \( \zeta_k \) in (57), we have \( \eta_k = (\zeta_k / \zeta_{k+1} - 1) / \mu \). Since \( \text{res}(x_N) > \epsilon \), we can apply Lemma 6.2 and rewrite the bound in (53) in terms of \( \{\zeta_k\}_{k=0}^{N-1} \) as

\[
\sum_{k=0}^{N-1} \left( \frac{\mu \zeta_{k+1}}{\zeta_k - \zeta_{k+1}} \right)^2 = \frac{1}{\zeta_k} \leq \frac{\gamma_2^2 L_2^2 D_\Phi(z^*, z_0)}{\mu^2} = \frac{\gamma_2^2 \kappa_2(z_0)}{\mu^2}. \tag{59}
\]

Since each summand in (59) is nonnegative, it follows that for any \( k_2 > k_1 \geq 0 \),

\[
\sum_{k=k_1}^{k_2-1} \left( \frac{1}{\zeta_{k+1} - 1/\zeta_k} \right)^2 \frac{1}{\zeta_k^3} \leq \gamma_2^2 \kappa_2(z_0).
\]

Furthermore, by applying Hölder’s inequality we get

\[
\left[ \sum_{k=k_1}^{k_2-1} \left( \frac{1}{\zeta_{k+1} - 1/\zeta_k} \right) \right]^{\frac{2}{3}} \left[ \sum_{k=k_1}^{k_2-1} \left( \frac{1}{\zeta_{k+1} - 1/\zeta_k} \right)^2 \right]^{\frac{1}{3}} \geq \sum_{k=k_1}^{k_2-1} \frac{1}{\zeta_k},
\]

and Lemma 6.4 follows from the above two inequalities.

By leveraging Lemma 6.4, we can establish a global complexity bound for our proposed second-order method. Specifically, we show that the sequence \( \{\zeta_k\}_{k \geq 0} \) halves its value after a geometrically decreasing number of iterations. Our argument is inspired by the restarting strategy used in [Nes08; Nes06], although here we do not need to actually restart our algorithm.

Theorem 6.5 (Global convergence). Suppose that Assumptions 2.2 and 6.1 hold and the Bregman function \( \Phi \) is \( L_\Phi \)-smooth. Let \( \{z_k\}_{k \geq 0} \) be the iterates generated by Algorithm 4, where \( \tilde{\eta}_k = \eta_{k-1} / (1 + \mu \eta_{k-1}) \), \( \alpha \in (0, 1) \), \( \beta \in (0, 1) \), \( \sigma_0 > 0 \), and \( \epsilon > 0 \). Then for any \( \epsilon > 0 \), we either have \( \text{res}(z_N) \leq \epsilon \) or \( \zeta_N \leq (2 - \alpha) \kappa(z_0) / \left( 2D_\Phi(z^*, z_0) \right) \) after at most the following number of iterations

\[
N \leq \max \left\{ \frac{1}{1 - 2^{-\frac{1}{4}}} \left[ \gamma_2 \kappa_2(z_0) \right]^{\frac{1}{2}} + \log_2 \left( \frac{2D_\Phi(z^*, z_0)}{(2 - \alpha) \epsilon} \right) + 1, 1 \right\}, \tag{60}
\]

where \( \kappa_2(z_0) \) and \( \gamma_2 \) are defined in (2) and (54), respectively.
Proof. Without loss of generality, we can assume that \(\frac{(2-\alpha)\epsilon}{2D_F(z^*, z_0)} \leq 1\); otherwise the result becomes trivial as \(\zeta_0 = 1\). It is easy to see from the definition (57) that \(\{\zeta_k\}_{k \geq 0}\) is non-increasing in \(k\). Hence, from Lemma 6.4 we get

\[
\frac{1}{\zeta_{k+1}} \geq \frac{1}{\zeta_k} + \frac{1}{\gamma_2 \kappa_2(z_0)} \left( \sum_{k=1}^{k-1} \frac{1}{\zeta_k} \right)^{\frac{2}{3}} \geq \frac{1}{\zeta_k} + \frac{1}{\gamma_2 \kappa_2(z_0)} \left( \frac{k_2 - k_1}{\zeta_k} \right)^{\frac{2}{3}}.
\]

In particular, this implies that \(\zeta_{k_2} \leq \frac{1}{3} \zeta_{k_1}\) when we have \(k_2 - k_1 \geq \lceil (\gamma_2 \kappa_2(z_0))^\frac{1}{2} \rceil \). Hence, for any integer \(l \geq 0\), we can prove by induction that the number of iterations \(N\) required to achieve \(\zeta_N \geq 2^{-l}\) does not exceed

\[
\sum_{k=0}^{l-1} \left( \lceil (\gamma_2 \kappa_2(z_0))^\frac{1}{2} \rceil - \frac{1}{3} \right) 2^{-\frac{k}{3}} \leq \sum_{k=0}^{l-1} \left( \lceil (\gamma_2 \kappa_2(z_0))^\frac{1}{2} \rceil - 1 \right) 2^{-\frac{k}{3}} + 1 \leq \frac{1}{1 - 2^{-\frac{1}{3}}} \left( \lceil (\gamma_2 \kappa_2(z_0))^\frac{1}{2} \rceil + l \right).
\]

The bound in (60) immediately follows by setting \(l = \lceil \log_2 \left( \frac{2D_F(z^*, z_0)}{2D_F(z^*, z_0)} \right) \rceil \leq \log_2 \left( \frac{2D_F(z^*, z_0)}{2D_F(z^*, z_0)} \right) + 1\). \(\Box\)

Theorem 6.5 guarantees the global convergence of our proposed second-order optimistic method, since it shows that we can achieve an arbitrary accuracy \(\epsilon\) after running the number of iterations given in (60). Better yet, we proceed to show that our method eventually achieves a fast local R-superlinear convergence rate. This result is formally stated in the following theorem.

**Theorem 6.6 (Local convergence).** Suppose that Assumptions 2.2 and 6.1 hold and the Bregman function \(\Phi\) is \(L_k\)-smooth. Let \(\{z_k\}_{k \geq 0}\) be the iterates generated by Algorithm 4, where \(\hat{\eta}_k = \eta_{k-1}/(1 + \mu \eta_{k-1})\), \(\alpha \in (0, 1)\), \(\beta \in (0, 1)\), \(\sigma_0 > 0\), and \(\epsilon > 0\). Then for any \(k \geq 0\), we either obtain \(\text{res}(z_{k+1}) \leq \epsilon\), or we have

\[
\zeta_{k+1} \leq \gamma_2 \kappa_2(z_0) \zeta_k^\frac{3}{2},
\]

where \(\kappa_2(z_0)\) and \(\gamma_2\) are defined in (2) and (54), respectively.

Proof. If \(\text{res}(z_{k+1}) \leq \epsilon\), then we are done. Otherwise, by letting \(k_1 = k\) and \(k_2 = k+1\) in Lemma 6.4, we obtain

\[
\frac{1}{\zeta_{k+1}} \geq \frac{1}{\zeta_k} + \frac{1}{\gamma_2 \kappa_2(z_0)} \zeta_k^2 \geq \frac{1}{\gamma_2 \kappa_2(z_0)} \zeta_k^\frac{3}{2},
\]

which immediately leads to (61). \(\Box\)

Theorem 6.6 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 \(\zeta_k < (\gamma_2 \kappa_2(z_0))^{-\frac{3}{2}}\). Due to the result in Theorem 6.5, the sequence indeed eventually falls below this threshold after a finite number of iterations. Hence, in light of the bound in (58), this in turn implies the local R-superlinear convergence of \(D_F(z^*, z_k)\).

By combining the global convergence result in Theorem 6.5 and the local convergence result in Theorem 6.6, we can characterize the overall iteration complexity of the second-order optimistic method in Algorithm 4. Specifically, if the required accuracy \(\epsilon\) is moderate, we simply use the global complexity result in Theorem 6.5. On the other hand, if the required accuracy \(\epsilon\) is sufficiently small, we will first upper bound the number of iterations required to reach the local convergence neighborhood via Theorem 6.5. And once the local convergence rate takes over, we need at most additional \(O(\log \log(1/\epsilon))\) iterations to reach the desired accuracy. We summarize the corresponding complexity bounds in Corollary 6.7.
Corollary 6.7. Suppose that Assumptions 2.2 and 6.1 hold and the Bregman function $\Phi$ is $L_\Phi$-smooth. Let $\{z_k\}_{k \geq 0}$ be the iterates generated by Algorithm 4, where $\bar{\eta}_k = \eta_{k-1}/(1 + \mu \eta_{k-1})$, $\alpha \in (0, 1)$, $\beta \in (0, 1)$, $\epsilon > 0$, and $\sigma_0 > 0$. Further, suppose $\epsilon$ is the required accuracy that we aim for, i.e., $D_\Phi(z^*, z) \leq \epsilon$. Then we have the following complexity bound for Algorithm 4.

- If $\epsilon \geq \frac{\mu^2}{(2-\alpha)^2 L_\Phi^2}$, then the second-order optimistic method finds an $\epsilon$-accurate solution after at most the following number of iterations

$$N \leq \max \left\{ \frac{1}{1 - 2^{-\frac{3}{2}}} \left[ \gamma_2 \kappa_2(z_0) \right]^{\frac{3}{2}} + \log_2 \left( \frac{2 D_\Phi(z^*, z_0)}{(2 \alpha) \epsilon} \right) + 1, 1 \right\},$$

where $\kappa_2(z_0)$ and $\gamma_2$ are defined in (2) and (54), respectively.

- If $\epsilon < \frac{\mu^2}{(2-\alpha)^2 L_\Phi^2}$, then the second-order optimistic method finds an $\epsilon$-accurate solution after at most the following number of iterations

$$N \leq \max \left\{ \frac{1}{1 - 2^{-\frac{3}{2}}} \left[ \gamma_2 \kappa_2(z_0) \right]^{\frac{3}{2}} + 2 \log_2 (\gamma_2 \kappa_2(z_0)) + 2, 1 \right\} + \log_{3/2} \log_2 \left( \frac{2 \mu^2}{(2 - \alpha) \gamma_2^2 L_\Phi^2 \epsilon} \right) + 1. \tag{62}$$

Proof. To begin with, recall that we have $D_\Phi(z^*, z) \leq \epsilon$ if $\zeta_k \leq \frac{\mu^2}{(2-\alpha)^2 L_\Phi^2}$ by (58). Hence, it suffices to upper bound the number of iterations required such that the latter condition holds. Also, we only need to prove the case where $\epsilon < \frac{\mu^2}{(2-\alpha)^2 L_\Phi^2}$, as the other case directly follows from Theorem 6.5.

Let $N_1$ be the smallest integer such that $\zeta_{N_1} \leq 1/(2 \gamma_2^2 \kappa_2^2(z_0))$. By setting $\epsilon = \frac{D_\Phi(z^*, z_0)}{(2-\alpha) \gamma_2^2 \kappa_2^2(z_0)} = \frac{\mu^2}{(2-\alpha)^2 L_\Phi^2}$ in Theorem 6.5, we obtain that

$$N_1 \leq \max \left\{ \frac{1}{1 - 2^{-\frac{3}{2}}} \left[ \gamma_2 \kappa_2(z_0) \right]^{\frac{3}{2}} + \log_2 \left( \frac{2 \gamma_2 L_\Phi^2 D_\Phi(z^*, z_0)}{\mu^2} \right) + 1, 1 \right\}$$

$$= \max \left\{ \frac{1}{1 - 2^{-\frac{3}{2}}} \left[ \gamma_2 \kappa_2(z_0) \right]^{\frac{3}{2}} + 2 \log_2 (\gamma_2 \kappa_2(z_0)) + 2, 1 \right\}. \tag{63}$$

Furthermore, we can rewrite (61) in Theorem 6.6 as $\gamma_2^2 \kappa_2^2(z_k) \zeta_{k+1} \leq \left( \gamma_2^2 \kappa_2^2(z_0) \right)^{\frac{3}{2}} \zeta_k^2$. By induction, we can prove that $\gamma_2^2 \kappa_2^2(z_k) \zeta_k \leq 2^{-\left(\frac{3}{2}\right) k + N_1}$ for all $k \geq N_1$. Hence, the additional number of iterations required to achieve $\zeta_k \leq \frac{\mu^2}{2 D_\Phi(z^*, z_0)}$ does not exceed

$$\left[ \log_{3/2} \log_2 \left( \frac{2 D_\Phi(z^*, z_0)}{(2 - \alpha) \gamma_2^2 L_\Phi^2 \epsilon} \right) \right] \leq \log_{3/2} \log_2 \left( \frac{2 \mu^2}{(2 - \alpha) \gamma_2^2 L_\Phi^2 \epsilon} \right) + 1. \tag{64}$$

The result in (62) now follows from (63) and (64). \qed

The result in Corollary 6.7 characterizes the overall iteration complexity of our proposed method. To be brief, if the required accuracy $\epsilon$ is larger than $\frac{\mu^2}{(2-\alpha)^2 L_\Phi^2}$, then the iteration complexity is bounded by

$$O \left( \frac{L_\Phi \sqrt{D_\Phi(z^*, z_0)}}{\mu} + \log \left( \frac{D_\Phi(z^*, z_0)}{\epsilon} \right) \right). \tag{65}$$
Otherwise, for $\epsilon$ smaller than $\frac{\mu^2}{(2-\alpha)^2 L^2}$, the overall iteration complexity is bounded by

$$O\left(\left(\frac{L_2 \sqrt{D_\Phi(z^*, z_0)}}{\mu}\right)^{\frac{2}{3}} + \log \log \left(\frac{\mu L^2}{\mu^2 \epsilon}\right)\right).$$  \hfill (66)

Note that the first term in (66) captures the number of required iterations to reach the local neighborhood, while the second term corresponds to the number of iterations in the local neighborhood to achieve accuracy $\epsilon$.

Comparison with [OKDG20]. We close the discussion on our iteration complexity by mentioning that similar complexity bounds to (65) and (66) were also established in [OKDG20] but via a very different approach. Specifically, the authors of that paper proposed to apply a restarting technique on the extragradient-type method in [BL20]. Under Assumptions 2.2 and 6.1, they proved a complexity bound of $O\left((\frac{L^2 R}{\mu})^{\frac{2}{3}} \log \frac{L^2 R}{\mu^2 \epsilon} + \log \log \frac{L^1 L^2 R}{\mu^2 \epsilon}\right)$ in terms of the primal-dual gap, where $R$ is an upper bound on $\|z_0 - z^*\|$. 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 [HZZ20] and improved the complexity bound to $O\left((\frac{L^2 R}{\mu})^{\frac{2}{3}} \log \frac{L^1 L^2 R}{\mu^2 \epsilon} + \log \log \frac{L^1 \mu}{\epsilon}\right)$. Compared with ours, their method is restricted to the unconstrained problems in the Euclidean setup, and can only achieve local superlinear convergence under stronger assumptions. Also, our method appears to be simpler as we do not need to switch between different update rules. Most importantly, they did not provide any detail on solving the subproblems nor identify the procedure and cost of finding an admissible stepsize, while we formally propose a line search scheme and identify the number of subsolver calls that our proposed second-order method requires. This is the topic of next subsection.

6.4 Complexity Bound of the Line Search Scheme

So far, we have characterized the number of iterations required to solve the saddle problem in (1) using the second-order method in Algorithm 4 for both convex-concave and strongly-convex strongly-concave settings. The only missing piece of our analysis is characterizing the total number of calls to the optimistic subsolver during the whole process of our proposed method, where each call solves an instance of (51). Before stating our result, we prove the following intermediate lemma which is useful for bounding the complexity of line search schemes.

Lemma 6.8. Let $\{\eta_k\}_{k \geq 0}$ be the stepsizes in (48) generated by Algorithm 2 with advancing. For any $N \geq 0$, if $\text{res}(z_N) > \epsilon$, then we have

$$\sum_{k=0}^{N-1} \eta_k^2 \leq \frac{2(\alpha + L_\Phi)^2 D_\Phi(z^*, z_0)}{(1 - \alpha)^2 \epsilon^2}.$$  \hfill (67)

Proof. From Lemma 4.4, for any $k = 0, \ldots, N - 1$ we have

$$\epsilon < \text{res}(z_{k+1}) \leq \frac{1}{\eta_k} \left(\left(\frac{\alpha}{2} + L_\Phi\right) \|z_{k+1} - z_k\| + \frac{\alpha}{2} \|z_k - z_{k-1}\|\right).$$

Hence, we obtain that

$$\eta_k^2 \leq \frac{1}{\epsilon^2} \left(\frac{\alpha^2 + 3L_\Phi \alpha + 2L_\Phi^2}{2} \|z_{k+1} - z_k\|^2 + \frac{\alpha^2 + L_\Phi \alpha}{2} \|z_k - z_{k-1}\|^2\right).$$  \hfill (68)

Now (67) follows from summing (68) over $k = 0, \ldots, N - 1$ and applying Lemma 4.2. \hfill $\square$

38
In the following theorem, we leverage the result in Lemma 6.8 to derive an upper bound on the total number of subsolver calls required after \( N \) iterations. This result holds for both convex-concave and strongly-convex strongly-concave settings.

**Theorem 6.9.** Consider Algorithm 4 with parameters \( \alpha \in (0,1) \), \( \beta \in (0,1) \), \( \epsilon > 0 \), and \( \sigma_0 > 0 \). For both convex-concave (Theorem 6.3) and strongly-convex-strongly-concave (Theorem 6.5) settings, the total number of calls to the optimistic subsolver after \( N \) iterations does not exceed

\[
2N \log_2 \left( 4 + 2 \log_\frac{1}{\beta} \left( 1 + \frac{\sigma_0^2 \gamma_2^2 L_0^2}{N} \right) \right) + 2 \log_\frac{1}{\beta} \left( 1 + \frac{2(\alpha + 1)^2 D_F(z^*, z_0)}{\sigma_0^2 (1 - \alpha) N c^2} \right),
\]

where \( \gamma_2 \) is a constant defined in (54).

**Proof.** Denote \( \eta_{-1} := \beta \sigma_0 \). By Proposition 4.6, the total number of calls to the optimistic subsolver after \( N \) iterations can be bounded by

\[
\sum_{k=0}^{N-1} 2 \log_2 \log_\frac{1}{\beta} \left( \max \left\{ \frac{\eta_{k-1}^2}{\beta^4 \eta_k^2}, \frac{\eta_k^2}{\eta_{k-1}^2} \right\} \right) \leq 2N \log_2 \left( \frac{1}{N} \sum_{k=0}^{N-1} \log_\frac{1}{\beta} \left( \max \left\{ \frac{\eta_{k-1}^2}{\beta^4 \eta_k^2}, \frac{\eta_k^2}{\eta_{k-1}^2} \right\} \right) \right),
\]

where we used the fact that \( \frac{1}{N} \sum_{k=0}^{N-1} \log a_k \leq \log \left( \frac{1}{N} \sum_{k=0}^{N-1} a_k \right) \). We make a simple observation that for any \( k = 1, \ldots, N - 1 \),

\[
\log_\frac{1}{\beta} \left( \max \left\{ \frac{\eta_{k-1}^2}{\beta^4 \eta_k^2}, \frac{\eta_k^2}{\eta_{k-1}^2} \right\} \right) = \max \left\{ 4 + \log_\frac{1}{\beta} \left( \frac{\eta_{k-1}^2}{\eta_k^2} \right), \log_\frac{1}{\beta} \left( \frac{\eta_k^2}{\eta_{k-1}^2} \right) \right\} \\
= \max \left\{ 4 + \log_\frac{1}{\beta} \left( \frac{\eta_{k-1}^2}{\sigma_0^2} + 1 \right), \log_\frac{1}{\beta} \left( \frac{\eta_k^2}{\sigma_0^2} + 1 \right) + \log_\frac{1}{\beta} \left( \frac{\sigma_0^2}{\eta_{k-1}^2} + 1 \right) \right\} \\
\leq 4 + \log_\frac{1}{\beta} \left( \frac{\eta_{k-1}^2}{\sigma_0^2} + 1 \right) + \log_\frac{1}{\beta} \left( \frac{\eta_k^2}{\sigma_0^2} + 1 \right) + \log_\frac{1}{\beta} \left( \frac{\sigma_0^2}{\eta_{k-1}^2} + 1 \right).
\]

Similarly, using the fact that \( \eta_{-1} = \beta \sigma_0 \), for the first summand in (69) we have

\[
\log_\frac{1}{\beta} \left( \max \left\{ \frac{\eta_{-1}^2}{\beta^4 \eta_0^2}, \frac{\eta_0^2}{\eta_{-1}^2} \right\} \right) = \max \left\{ 2 + \log_\frac{1}{\beta} \left( \frac{\sigma_0^2}{\eta_0^2} \right), 2 + \log_\frac{1}{\beta} \left( \frac{\eta_0^2}{\sigma_0^2} \right) \right\} \\
\leq 4 + \log_\frac{1}{\beta} \left( \frac{\sigma_0^2}{\eta_0^2} + 1 \right) + \log_\frac{1}{\beta} \left( \frac{\eta_0^2}{\sigma_0^2} + 1 \right).
\]

Hence, we can show that

\[
\sum_{k=0}^{N-1} \log_\frac{1}{\beta} \left( \max \left\{ \frac{\eta_{k-1}^2}{\beta^4 \eta_k^2}, \frac{\eta_k^2}{\eta_{k-1}^2} \right\} \right) \\
\leq 4N + 2 \sum_{k=0}^{N-1} \left( \log_\frac{1}{\beta} \left( \frac{\sigma_0^2}{\eta_k^2} + 1 \right) + \log_\frac{1}{\beta} \left( \frac{\eta_k^2}{\sigma_0^2} + 1 \right) \right) \\
\leq 4N + 2N \log_\frac{1}{\beta} \left( 1 + \frac{\sigma_0^2}{N} \sum_{k=0}^{N-1} \frac{1}{\eta_k^2} \right) + 2N \log_\frac{1}{\beta} \left( 1 + \frac{1}{\sigma_0^2 N} \sum_{k=0}^{N-1} \eta_k^2 \right),
\]

39
where the first inequality simply follows from (70) and (71), and the second inequality holds since \( \frac{1}{N} \sum_{k=0}^{N-1} \log a_k \leq \log \left( \frac{1}{N} \sum_{k=0}^{N-1} a_k \right) \). Now we only need to establish bounds on \( \sum_{k=0}^{N-1} \frac{1}{\eta_k} \) and \( \sum_{k=0}^{N-1} \eta_k^2 \). Since the term \( (1 + \eta k \mu) \) is always larger than 1, by Lemma 6.2 we have

\[
\sum_{k=0}^{N-1} \frac{1}{\eta_k} \leq \gamma \frac{L_2^2 D_\Phi(z^*, z_0)}{N} \]

for both convex-concave and strongly-convex-strongly-concave cases. Also, Lemma 6.8 provides an upper bound on \( \sum_{k=0}^{N-1} \eta_k^2 \). Applying these bounds in (72) implies that

\[
\sum_{k=0}^{N-1} \frac{1}{\eta_k} \left( \max \left\{ \frac{\eta_k^2}{\eta_k^2}, \frac{\eta_k^2}{\eta_{k-1}^2} \right\} \right) 
\leq 4N + 2N \log_\frac{1}{\eta} \left( 1 + \frac{\sigma_0^2 \gamma^2 L_2^2 D_\Phi(z^*, z_0)}{N} \right) + 2N \log_\frac{1}{\eta} \left( 1 + \frac{2(\alpha + L_\Phi)^2 D_\Phi(z^*, z_0)}{\sigma_0^2 (1 - \alpha) N \epsilon^2} \right). \tag{73}
\]

Now the claim follows by applying the upper bound in (73) in the right-hand side of (69).

This result shows that if we run our proposed second-order optimistic method in Algorithm 4 for \( N \) iterations to achieve a specific accuracy \( \epsilon \), the total number of calls to a subsolver of (51) is of the order \( O(N \log \log(D_\Phi(z^*, z_0)/N \epsilon^2)) \). Hence, on average we only need to make \( O(\log \log(D_\Phi(z^*, z_0)/N \epsilon^2)) \) subsolver calls per iteration, which is almost a constant number. For instance, in the convex-concave setting, to find an \( \epsilon \)-accurate solution we require \( N = O(1/\epsilon^{2/3}) \) iterations by Theorem 6.3, which implies that the average number of subsolver calls per iteration is bounded by \( O(\log \log(1/\epsilon^{1/3})) \).

### 7 Higher-Order Generalized Optimistic Method

In this section, we explore the possibility of designing methods for saddle point problems using higher-order derivatives. As in Section 6, we require that the Bregman function \( \Phi \) is \( L_\Phi \)-smooth on \( Z \). Also, we make the following assumption on \( f \).

**Assumption 7.1.** The operator \( F \) defined in (8) is \( p \)-th order \( L_\Phi \)-smooth (\( p \geq 3 \)) on \( Z \). Also, we have access to an oracle that returns \( F(z), DF(z), \ldots, D^{(p-1)}(z) \) for any given \( z \).

Under Assumption 7.1, it follows from (6) that

\[
\|F(z) - T^{(p-1)}(z; z_k)\|_\epsilon \leq \frac{L_\Phi}{p!} \|z - z_k\|^p, \quad \forall z \in Z, \tag{74}
\]

where \( T^{(p-1)}(z; z_k) \) is the \((p - 1)\)-th order Taylor expansion of \( F \) at point \( z_k \) (cf. (5)). A natural generalization of the second-order optimistic method in Section 6 is to choose the approximation function as \( P(z; I_k) := T^{(p-1)}(z; z_k) \). Alas, a subtle technical issue arises on closer inspection of the resulting subproblem. Specifically, with this choice of \( P(z; I_k) \) we need to solve the inclusion problem:

\[
0 \in \eta_k T^{(p-1)}(z; z_k) + \eta_k \left( F(z_k) - T^{(p-1)}(z_k; z_{k-1}) \right) + \eta_k H(z) + \nabla \Phi(z) - \nabla \Phi(z_k). \tag{75}
\]

We observe that this is a non-monotone inclusion problem, since the operator \( T^{(p-1)}(z; z_k) \), being the \((p - 1)\)-th Taylor approximation of a monotone operator \( F \), is not monotone for \( p \geq 3 \) in
Therefore, to instantiate the generalized optimistic method (Algorithm 1) with a $p$-th-order oracle, we choose the approximation function as the regularized Taylor expansion $T^{(p-1)}_\lambda(z; z_k)$. Consider the operator $T^{(p-1)}_\lambda(z; z) : \mathbb{R}^d \rightarrow \mathbb{R}^d$ defined as

$$T^{(p-1)}_\lambda(z'; z) := T^{(p-1)}(z'; z) + \frac{\lambda}{(p-1)!} (2D\Phi(z'; z))^{\frac{p-1}{2}} (\nabla\Phi(z') - \nabla\Phi(z)),$$

where $\lambda > 0$ is the regularization parameter. Note that the regularization term is the gradient of the convex function $\frac{\lambda^p}{(p+1)!} (2D\Phi(z'; z))^{\frac{p+1}{2}}$, and its norm is on the order of $O(\|z' - z\|^p)$ since $\Phi$ is $L_\Phi$-smooth. Hence, for sufficiently large $\lambda$, we expect $T^{(p-1)}_\lambda(z; z_k)$ to be a monotone operator of $z$, while maintaining a similar approximation error as in (74). The following lemma confirms this intuition.

**Lemma 7.1.** The following statements hold:

(a) If $F$ is $p$-th-order $L_p$-smooth, then for any $\lambda \geq 0$ we have

$$\|F(z') - T^{(p-1)}_\lambda(z'; z)\|_* \leq \frac{L_p(\Phi, \lambda)}{p!} \|z' - z\|^p, \quad \forall z' \in \text{dom } F,$$

where we denote

$$L_p(\Phi, \lambda) := L_p + pL_\Phi^\lambda \lambda. \quad (76)$$

(b) If $F$ is monotone and $p$-th-order $L_p$-smooth, then for $\lambda \geq L_p$, the operator $T^{(p-1)}_\lambda(\cdot; z)$ is maximal monotone with domain $\mathbb{R}^d$.

**Proof.** See Appendix E.1. \hfill \Box

Therefore, to instantiate the generalized optimistic method (Algorithm 1) with a $p$-th-order oracle, we choose the approximation function as the regularized Taylor expansion $T^{(p-1)}_\lambda(z; z_k)$. Accordingly, the update rule for the proposed $p$-th-order optimistic method can be written as

$$0 \in \eta_k T^{(p-1)}_\lambda(z_{k+1}; z_k) + \tilde{\eta}_k \left(F(z_k) - T^{(p-1)}_\lambda(z_k; z_{k-1})\right) + \eta_k H(z_{k+1}) + \nabla\Phi(z_{k+1}) - \nabla\Phi(z_k), \quad (77)$$

where the condition in (24) on $\eta_k$ can be written as

$$\eta_k \|F(z_{k+1}) - T^{(p-1)}_\lambda(z_{k+1}; z_k)\|_* \leq \frac{\alpha}{2} \|z_{k+1} - z_k\|. \quad (78)$$

Apart from the regularization modification, the intuition and analysis for the $p$-th-order optimistic method are very similar to those for the second-order algorithm in Section 6. To see how the $p$-th-order information can lead to further speedup, note that the left hand side of (78) is upper bounded by $\frac{\eta_k L_p(\Phi, \lambda)}{p!} \|z_{k+1} - z_k\|^p$ by Lemma 7.1(a). This suggests that we can pick our stepsize $\eta_k$ as

$$\eta_k \approx \frac{p!\alpha}{2L_p(\Phi, \lambda) \|z_{k+1} - z_k\|^{p-1}}. \quad (79)$$

---

For a counterexample, note that any multivariate polynomial of odd degree in $\mathbb{R}^d$ is not convex. Hence, when $p$ is odd, the $p$-th-order Taylor expansion of a convex function $f$ is not convex, which implies that the $(p-1)$-th-order Taylor expansion of the monotone gradient operator $\nabla f$ is not monotone.
Algorithm 5 \( p \)-th-order optimistic method

1: **Input:** initial point \( z_0 \in Z \), initial trial stepsize \( \sigma_0 \), strongly-convex parameter \( \mu \geq 0 \), regularization parameter \( \lambda \geq L_p \), line search parameters \( \alpha, \beta \in (0, 1) \), and \( \epsilon > 0 \)
2: **Initialize:** set \( z_{-1} \leftarrow z_0 \) and \( P(z_0) \leftarrow F(z_0) \)
3: for iteration \( k = 0, \ldots, N-1 \) do
   4:   Set \( \eta_k = \eta_{k-1}/(1 + \eta_{k-1}\mu) \)
   5:   Select \( \eta_k \) by Algorithm 2 with advancing, where \( \sigma = \sigma_k \), \( z^- = z_k \), \( P(z) = T^{(p-1)}\lambda(z; z_k) \), and \( v^- = \tilde{\eta}_k (F(z_k) - T^{(p-1)}\lambda(z_k; z_{k-1})) \)
   6:   Compute \( z_{k+1} \) by solving \( 0 \in \eta_k T^{(p-1)}\lambda(z_k; z_{k-1}) + \tilde{\eta}_k (F(z_k) - T^{(p-1)}\lambda(z_k; z_{k-1})) + \eta_k H(z) + \nabla \Phi(z) - \nabla \Phi(z_k) \)
   7:   if \( \text{res}(z_{k+1}) \leq \epsilon \) then
      8:      Return \( z_{k+1} \) as an \( \epsilon \)-accurate solution
   9:   else
   10:      Set \( \sigma_{k+1} \leftarrow \eta_k/\beta \)
   11:   end if
12: end for

Compared with the stepsize in (50) for the second-order optimistic method, we can see that the stepsize in (79) will eventually be larger as the iterates approach the optimal solution and the displacement \( \|z_{k+1} - z_k\| \) tends to zero. Hence, according to Proposition 4.1, we shall expect better convergence rates for the \( p \)-th-order optimistic method.

Also, to address the interdependence between the stepsize \( \eta_k \) and the iterate \( z_{k+1} \), we take the same approach as in Section 6: we use Algorithm 2 with the advancing subroutine to select the stepsize in each iteration. The resulting \( p \)-th-order optimistic method is formally described in Algorithm 5.

Two remarks on the line search scheme in Algorithm 5 follow. First, similar to the arguments in Section 6, we can guarantee that the line search scheme always terminates in finite steps by using the results in Lemmas 4.3 and 4.5. Hence, the stepsizes \( \{\eta_k\}_{k \geq 0} \) are well-defined. Second, at iteration \( k \), the optimistic subsolver in this case is required to solve a subproblem of \( z \) in the form of

\[
0 \in \eta T^{(p-1)}\lambda(z; z_k) + v_k + \eta H(z) + \nabla \Phi(z) - \nabla \Phi(z_k) \tag{80}
\]

for given \( v_k = \tilde{\eta}_k (F(z_k) - T^{(p-1)}\lambda(z_k; z_{k-1})) \) and \( \eta > 0 \) (cf. Definition 4.1). As shown in Lemma 7.1(b), the operator \( T^{(p-1)}\lambda(z; z_k) \) is monotone in \( z \) and hence the problem in (80) can be solved in general by many methods of monotone inclusion problems.

In the following sections, we derive convergence rates of the higher-order optimistic method and upper bound its total number of calls to the optimistic subsolver. Since the analysis largely mirrors the one for the second-order case in Section 6, we relegate the proofs to Appendix E.

### 7.1 Convergence Analysis: Convex-Concave Case

We first consider the case where \( f \) is merely convex-concave. The ergodic convergence rate result below follows from Proposition 4.1.

**Theorem 7.2.** Suppose that Assumption 2.1 and Assumption 7.1 hold and the Bregman function \( \Phi \) is \( L_\Phi \)-smooth. Let \( \{z_k\}_{k \geq 0} \) be the iterates generated by Algorithm 5, where \( \tilde{\eta}_k = \eta_{k-1} \), \( \sigma_0 > 0 \), \( \lambda \geq L_p \), \( \alpha \in (0, 1) \), \( \beta \in (0, 1) \), and \( \epsilon > 0 \). Then we have \( D_\Phi(z^*, z_k) \leq \frac{\epsilon^2}{2\alpha L_\Phi} D_\Phi(z^*, z_0) \) for any \( k \geq 0 \).
Moreover, we either have \( \text{res}(z_N) \leq \epsilon \), or
\[
\ell(\tilde{x}_N, y) - \ell(x, y_N) \leq \gamma_p^{p-1} L_p(\Phi, \lambda) D_\Phi(z, z_0)(D_\Phi(z^*, z_0)) \frac{\epsilon}{2N^{p/2}}
\]
for any \( z = (x, y) \in X \times Y \), where \( \tilde{z}_N = (\tilde{x}_N, \tilde{y}_N) \) is given by \( \tilde{z}_N = \frac{1}{\sum_{k=0}^{N-1} \eta_k} \eta_k z_{k+1} \) and \( \gamma_p \) is defined by
\[
\gamma_p = \sqrt{\frac{2}{1 - \alpha}} L_\Phi^{3/2} \left( \frac{2}{p! \alpha^p} \right)^{\frac{1}{p-1}} + \frac{1}{2(1 - \alpha)} \left( \frac{2}{p! \alpha^p} \right)^{\frac{1}{p-1}} \alpha (\beta + L_\Phi^{1/2}).
\]  

**Proof.** See Appendix E.3.

**Remark 7.1.** Note that the constant \( \gamma_p \) is independent of the problem parameters and solely depends on the implementation parameters. For instance, if we select the line search parameters as \( \alpha = 0.5 \) and \( \beta = 0.9 \) and focus on the Euclidean case where \( L_\Phi = 1 \), then we have \( \gamma_3 \approx 2.8 \).

Theorem 7.2 shows that the \( p \)-th-order optimistic method converges at a rate of \( \mathcal{O}(N^{-(p+1)/2}) \) in terms of the primal-dual gap, which is faster than the rate of \( \mathcal{O}(N^{-1}) \) for first-order methods and the rate of \( \mathcal{O}(N^{-3/2}) \) for our second-order optimistic methods. As a corollary, to obtain a solution with a primal-dual gap of \( \epsilon \), the proposed \( p \)-th-order optimistic method requires at most \( \mathcal{O}(1/\epsilon^{2/(p+1)}) \) iterations.

**Comparison with [BL20].** We note that similar iteration complexity bounds are also reported in [BL20] for extragradient-type higher-order methods. However, the authors in [BL20] did not discuss how to solve the subproblem (which can be non-monotone) or how to select a stepsize that satisfies their specified conditions. In contrast, we propose and analyze our line search scheme in detail, and the total number of calls to the optimistic subsolver during the whole process after \( N \) iterations can be also explicitly upper bounded, as we discuss later in Theorem 7.7.

### 7.2 Convergence Analysis: Strongly-Convex-Strongly-Concave Case

Next, we proceed to the setting where the smooth component \( f \) of the objective in (1) is \( \mu \)-strongly-convex-strongly-concave. As discussed in Section 6, in this case the distance to the optimal solution \( D_\Phi(z^*, z_k) \) is bounded above by the decreasing sequence \( \{\zeta_k\}_{k \geq 0} \) defined in (57). Hence, it suffices to characterize the convergence behavior of the sequence \( \zeta_k \), which immediately implies that of \( D_\Phi(z^*, z_k) \) from (58).

To simplify the notation, we define
\[
\bar{\zeta}_p(z_0) := \frac{\gamma_p^{p-1} L_p(\Phi, \lambda)(D_\Phi(z^*, z_0))^{\frac{p-1}{2}}}{\mu},
\]  

where \( L_p(\Phi, \lambda) \) and \( \gamma_p \) are defined in (76) and (81), respectively. Similar to Lemma 6.4, the following lemma is the key to our convergence results.

**Lemma 7.3.** Let \( \{\eta_k\}_{k \geq 0} \) be the stepsizes in (77) generated by Algorithm 5 and \( \{\zeta_k\}_{k \geq 0} \) be defined in (57). If \( \text{res}(x_N) > \epsilon \), then for any \( 0 \leq k_1 < k_2 \leq N \) we have
\[
\frac{1}{\zeta_{k_2}} \geq \frac{1}{\zeta_{k_1}} + \frac{1}{\bar{\zeta}_p(z_0)} \left( \sum_{k=k_1}^{k_2-1} \zeta_k \right)^{\frac{p-1}{2}}.
\]
Proof. See Appendix E.4. □

By leveraging Lemma 7.3, we can establish a global complexity bound for the proposed p-th-order method as in Theorem 6.5.

**Theorem 7.4** (Global convergence). Suppose that Assumptions 2.2 and 7.1 hold and the Bregman function \( \Phi \) is \( L_\Phi \)-smooth. Let \( \{z_k\}_{k \geq 0} \) be the iterates generated by Algorithm 5, where \( \tilde{\eta}_k = \eta_{k-1} / (1 + \mu \eta_{k-1}) \), \( \alpha \in (0, 1) \), \( \beta \in (0, 1) \), \( \epsilon > 0 \), \( \sigma_0 > 0 \) and \( \lambda \geq L_p \). Then for any \( \epsilon > 0 \), we either have \( \text{res}(z_N) \leq \epsilon \) or \( \zeta_N \leq \frac{(2 - \alpha) \epsilon}{2(D_\Phi(z^*, z_0)} \) after at most the following number of iterations

\[
N \leq \max \left\{ \frac{1}{1 - 2^{\frac{1}{p+1}}} \left[ \tilde{\kappa}_p(z_0) \right]^{\frac{2}{p+1}} + \log_2 \left( \frac{2D_\Phi(z^*, z_0)}{(2 - \alpha) \epsilon} \right) + 1, 1 \right\},
\]

(83)

where \( \tilde{\kappa}_p(z_0) \) is defined in (82).

Proof. See Appendix E.5. □

Theorem 7.4 already guarantees the global convergence of our p-th-order optimistic method. Moreover, as in the case of our second-order optimistic method, the proposed higher-order optimistic method eventually achieves a fast local R-superlinear convergence rate. This result is formally stated in the following theorem.

**Theorem 7.5** (Local convergence rate). Suppose that Assumptions 2.2 and 7.1 hold and the Bregman function \( \Phi \) is \( L_\Phi \)-smooth. Let \( \{z_k\}_{k \geq 0} \) be the iterates generated by Algorithm 5, where \( \tilde{\eta}_k = \eta_{k-1} / (1 + \mu \eta_{k-1}) \), \( \alpha \in (0, 1) \), \( \beta \in (0, 1) \), \( \epsilon > 0 \), \( \sigma_0 > 0 \) and \( \lambda \geq L_p \). Further, recall the definition of \( \gamma_p \) in (81). Then for any \( k \geq 0 \), we either obtain \( \text{res}(z_{k+1}) \leq \epsilon \), or we have

\[
\zeta_{k+1} \leq \tilde{\kappa}_p(z_0)^{\frac{p+1}{\gamma_p}},
\]

(84)

where \( \tilde{\kappa}_p(z_0) \) is defined in (82).

Proof. See Appendix E.6. □

Theorem 7.5 implies that the sequence \( \{\zeta_k\}_{k \geq 0} \) converges to 0 at a superlinear convergence rate of order \( \frac{p+1}{\gamma_p} \) once we have \( \zeta_k < (\tilde{\kappa}_p(z_0))^{-\frac{p+1}{\gamma_p}} \). Due to the result in Theorem 7.4, the sequence indeed eventually falls below this threshold after a finite number of iterations. Hence, in light of the bound in (58), this in turn implies the local R-superlinear convergence of \( D_\Phi(z^*, z_k) \).

By combining the global convergence result in Theorem 7.4 and the local convergence result in Theorem 7.5, we can characterize the overall computational complexity of the p-th-order optimistic method in Algorithm 5. As in Corollary 6.7 for the second-order method, the complexity result below is stated in two cases depending on the required accuracy \( \epsilon \).

**Corollary 7.6**. Suppose that Assumptions 2.2 and 7.1 hold and the Bregman function \( \Phi \) is \( L_\Phi \)-smooth. Let \( \{z_k\}_{k \geq 0} \) be the iterates generated by Algorithm 5, where \( \tilde{\eta}_k = \eta_{k-1} / (1 + \mu \eta_{k-1}) \), \( \alpha \in (0, 1) \), \( \beta \in (0, 1) \), \( \epsilon > 0 \), \( \sigma_0 > 0 \) and \( \lambda \geq L_p \). Further, suppose \( \epsilon \) is the required accuracy that we aim for, i.e., \( D_\Phi(z^*, z) \leq \epsilon \). Then we have the following complexity bound for Algorithm 5.
• If $\epsilon \geq \frac{1}{(2-\alpha)\gamma_0^p} \left( \frac{\mu}{L_p(\Phi, \lambda)} \right)^{\frac{2}{p-1}}$, then the $p$-th-order optimistic method finds an $\epsilon$-accurate solution after at most the following number of iterations

$$N \leq \max \left\{ \frac{1}{1 - 2^{-\frac{1}{p-1}}} \tilde{\kappa}_p(z_0)^{\frac{2}{p+1}} + \log_2 \left( \frac{2D_\Phi(z^*, z_0)}{(2 - \alpha)\epsilon} \right) + 1, 1 \right\},$$

where $\tilde{\kappa}_p(z_0)$ is defined in (82).

• If $\epsilon < \frac{1}{(2-\alpha)\gamma_0^p} \left( \frac{\mu}{L_p(\Phi, \lambda)} \right)^{\frac{2}{p-1}}$, then the $p$-th-order optimistic method finds an $\epsilon$-accurate solution after at most the following number of iterations

$$N \leq \max \left\{ \frac{\tilde{\kappa}_p(z_0)^{\frac{2}{p+1}}}{1 - 2^{-\frac{1}{p-1}}} + \frac{2}{p-1} \log_2 \frac{\tilde{\kappa}_p(z_0)^{\frac{2}{p+1}}}{\gamma_0^p(L_p(\Phi, \lambda))^{\frac{2}{p-1}}} + 2, 1 \right\} + \log_{\frac{2}{p-1}} \log_2 \left( \frac{2\mu^{\frac{2}{p-1}}}{(2 - \alpha)\gamma_0^p(L_p(\Phi, \lambda))^{\frac{2}{p-1}}\epsilon} \right) + 1. \tag{85}$$

**Proof.** See Appendix E.7. \qed

The result in Corollary 7.6 characterizes the overall iteration complexity of our proposed method. To simplify the discussions, assume that the regularization parameter is chosen as $\lambda = O(L_p)$ and the smooth parameter $L_\Phi$ of the Bregman function $\Phi$ can be bounded by some absolute constant. As a result, $L_p(\Phi, \lambda)$ defined in (76) is on the same order of $L_p$ and further $\tilde{\kappa}_p(z_0)$ in (82) is on the same order of $\kappa_p(z_0)$ defined in (2). If the required accuracy $\epsilon$ is larger than $\frac{1}{(2-\alpha)\gamma_0^p} \left( \frac{\mu}{L_p(\Phi, \lambda)} \right)^{\frac{2}{p-1}}$, then the iteration complexity can be bounded by

$$O \left( \left( \frac{L_p(D_\Phi(z^*, z_0))^{\frac{p-1}{2}}}{\mu} \right)^{\frac{2}{p+1}} + \log \left( \frac{D_\Phi(z^*, z_0)}{\epsilon} \right) \right). \tag{86}$$

Otherwise, for $\epsilon$ smaller than $\frac{1}{(2-\alpha)\gamma_0^p} \left( \frac{\mu}{L_p(\Phi, \lambda)} \right)^{\frac{2}{p}},$ the overall iteration complexity is bounded by

$$O \left( \left( \frac{L_p(D_\Phi(z^*, z_0))^{\frac{p-1}{2}}}{\mu} \right)^{\frac{2}{p+1}} + \log \left( \frac{\mu^{\frac{2}{p-1}}}{L_p^{\frac{2}{p-1}} \epsilon} \right) \right), \tag{87}$$

Note that the first term in (87) captures the number of iterations required to reach the local neighborhood, while the second term corresponds to the number of iterations in the local neighborhood to achieve accuracy $\epsilon$.

It is worth noting that the work [OKDG20] also reported complexity bounds similar to (86) and (87). The limitations of their results we discussed at the end of Section 6.3 also apply for the higher-order setting, and we refer the reader to the discussions therein.

### 7.3 Complexity Bound of the Line Search Scheme

As in Section 6, the final piece of our analysis is to characterize the complexity of our line search scheme used in Algorithm 5. In the following theorem, we derive an upper bound on the total number of calls to the optimistic subsolver required after $N$ iterations, where each call solves an instance of (80). This result holds for both convex-concave and strongly-convex strongly-concave settings.
Theorem 7.7. Consider Algorithm 5 with parameters \( \alpha \in (0, 1) \), \( \beta \in (0, 1) \), \( \epsilon > 0 \), \( \sigma_0 > 0 \) and \( \lambda \geq L_p \). For both convex-concave (Theorem 7.2) and strongly-convex-strongly-concave (Theorem 7.4) settings, the total number of calls to the optimistic sub solver after \( N \) iterations does not exceed
\[
2N\log_2 \left[ 4 + 2(p-1)\log_\frac{1}{\epsilon} \left( 1 + \frac{\sigma_0^2 \gamma_p^2 (L_p(\alpha, \lambda))^{p-1} D_\Phi(z^*, z_0)}{N} \right) + 2 \log_\frac{1}{\epsilon} \left( 1 + \frac{2(\alpha + L_p)^2 D_\Phi(z^*, z_0)}{\sigma_0 (1 - \alpha) N \epsilon^2} \right) \right]
\]
where \( \gamma_p \) is a constant defined in (81).

Proof. See Appendix E.8.

This result shows that if we run our proposed \( p \)-th-order optimistic method in Algorithm 5 for \( N \) iterations to achieve a specific accuracy \( \epsilon \), the total number of calls to a sub solver of (80) is of the order \( \mathcal{O}(N \log \log(D_\Phi(z^*, z_0)/N \epsilon^2)) \). Hence, on average we only need to make \( \mathcal{O}(\log \log(D_\Phi(z^*, z_0)/N \epsilon^2)) \) subroutine calls per iteration, which is almost a constant number. For instance, in the convex-concave setting, to find an \( \epsilon \)-accurate solution we require \( N = \mathcal{O}(1/\epsilon^{2/(p+1)}) \) iterations by Theorem 7.2, which implies that the average number of sub solver calls per iteration is bounded by \( \mathcal{O}(\log(1/\epsilon^{2p/(p+1)})). \)

8 Numerical Experiments

In this section, we demonstrate the numerical performance of our first-order and second-order optimistic methods on some synthetic saddle point problems.

8.1 First-Order Optimistic Method

Consider the following composite saddle point problem with box constraints:

\[
\min_{\mathbf{x} \in \mathcal{X}} \max_{\mathbf{y} \in \mathcal{Y}} \quad \langle \mathbf{A} \mathbf{x} - \mathbf{b}, \mathbf{y} \rangle + \lambda \|\mathbf{x}\|_1 + \frac{\mu}{2} \|\mathbf{x}\|_2^2 - \lambda \|\mathbf{y}\|_1 - \frac{\mu}{2} \|\mathbf{y}\|_2^2, \tag{88}
\]

where \( \mathbf{A} \in \mathbb{R}^{n \times m} \), \( \mathbf{b} \in \mathbb{R}^n \), \( \mathcal{X} = \{\mathbf{x} \in \mathbb{R}^m : \|\mathbf{x}\|_\infty \leq R\} \) and \( \mathcal{Y} = \{\mathbf{y} \in \mathbb{R}^n : \|\mathbf{y}\|_\infty \leq R\} \). We can see that this is an instance of Problem (1) with \( f(\mathbf{x}, \mathbf{y}) = \langle \mathbf{A} \mathbf{x} - \mathbf{b}, \mathbf{y} \rangle + \frac{\mu}{2} \|\mathbf{x}\|_2^2 - \frac{\mu}{2} \|\mathbf{y}\|_2^2 \), \( h_1(\mathbf{x}) = \lambda \|\mathbf{x}\|_1 \) and \( h_2(\mathbf{y}) = \lambda \|\mathbf{y}\|_1 \). In all of the numerical instances, we generate the entries of \( \mathbf{A} \) and \( \mathbf{b} \) independently and uniformly from the interval \([-1, 1]\).

We consider the Euclidean setup where \( \| \cdot \|_X = \| \cdot \|_2 \) and \( \| \cdot \|_Y = \| \cdot \|_2 \) and \( D_\Phi(\mathbf{z}, \mathbf{z}') = \frac{1}{2} \| \mathbf{z} - \mathbf{z}' \|_2^2 \). It is easy to verify that Assumption 5.1 holds: the operator \( F \) defined in (8) is \( L_1 \)-Lipschitz with \( L_1 \) being the operator norm of the matrix \( \begin{bmatrix} \mu I_m & \mathbf{A}^T \\ -\mathbf{A} & \mu I_n \end{bmatrix} \). Moreover, our first-order optimistic method in (37) can be instantiated as

\[
\begin{align*}
\mathbf{x}_{k+1} &= T_{\eta_k, \lambda, R}(\mathbf{x}_k - \eta_k \nabla_X f(\mathbf{x}_k, \mathbf{y}_k) - \hat{\eta}_k (\nabla_X f(\mathbf{x}_k, \mathbf{y}_k) - \nabla_X f(\mathbf{x}_{k-1}, \mathbf{y}_{k-1}))), \\
\mathbf{y}_{k+1} &= T_{\eta_k, \lambda, R}(\mathbf{y}_k - \eta_k \nabla_Y f(\mathbf{x}_k, \mathbf{y}_k) - \hat{\eta}_k (\nabla_Y f(\mathbf{x}_k, \mathbf{y}_k) - \nabla_Y f(\mathbf{x}_{k-1}, \mathbf{y}_{k-1}))),
\end{align*}
\]

where the operator \( T_{t, R}(\cdot) : \mathbb{R} \rightarrow \mathbb{R} \) is defined by

\[
T_{t, R}(z) = \begin{cases} 
0, & \text{if } |z| \leq t; \\
(z - t) \cdot \text{sgn}(z), & \text{if } t < |z| \leq t + R; \\
R \cdot \text{sgn}(z), & \text{Otherwise},
\end{cases}
\]

46
10
0
10
1
10
2
10
3
10
4
10
-4
10
-2
10
0
10
2
72x522 Figure 1: The primal-dual gap versus the number of iterations for the first-order optimistic methods on solving a convex-concave saddle point problem in (88).

and it is applied elementwise. The stepsize $\eta_k$ is fixed as $1/M$ in our fixed stepsize scheme (Option I in Algorithm 3), while it is chosen adaptively in our line search scheme (Option II in Algorithm 3). The initial point $(x_0, y_0)$ is chosen as the origin in $\mathbb{R}^{m+n}$.

**Convex-concave setting.** When $\mu$ is set to 0, the function $f(x, y)$ is convex-concave. In this case, we measure the quality of a solution $(x, y)$ by the primal-dual gap function defined in (12), which can be computed in closed form as

$$
\Delta(x, y) = [R||Ax - b - \lambda||_1 + \lambda||x||_1] - [R|||A^Ty| - \lambda||_1 - \langle b, y \rangle - \lambda||y||_1].
$$

Here, the operators $|\cdot|$ and $(z)_+ = \max\{z, 0\}$ are applied elementwise. By taking the supremum on both sides of (40), we obtain the following upper bound for the fixed stepsize scheme via Theorem 5.1:

$$
\Delta(\bar{x}_N, \bar{y}_N) \leq \frac{M}{2N} \cdot \arg \max_{z \in \mathbb{R}^{m+n}} \|z - z_0\|^2 = \frac{M(m + n)R^2}{2N},
$$

(89)

where $\bar{x}_N = \frac{1}{N} \sum_{k=0}^{N-1} x_{k+1}$ and $\bar{y}_N = \frac{1}{N} \sum_{k=0}^{N-1} y_{k+1}$. A similar bound for the gap at the average iterate of the line search scheme can also be derived from Theorem 5.3.

In our experiment, the problem parameters are chosen as $m = 600$, $n = 300$, $\lambda = 0.1$, $\mu = 0$, and $R = 0.05$. For the fixed stepsize scheme, we choose $M = 2L_1$, while for the adaptive line search scheme we choose $\alpha = 1, \beta = 0.8$, and $\sigma_0 = 1$. In Fig. 1, we plot the primal-dual gap at the averaged iterate $(\bar{x}_N, \bar{y}_N)$ as well as at the last iterate $(x_N, y_N)$ of both the fixed stepsize scheme and the line search scheme. The $\mathcal{O}(1/N)$ upper bound in (89) is also shown for comparison. We can see that the averaged iterates in both schemes converge exactly at the rate of $\mathcal{O}(1/N)$, as predicted by our convergence analysis. Moreover, the line search scheme also slightly outperforms the fixed stepsize scheme, and we note that it is able to achieve so without any prior knowledge of the Lipschitz constant of $F$. Interestingly, the gap at the last iterate appears to enjoy faster convergence in this particular example, which is similar to the phenomenon observed in [CP16] for a different first-order method.

**Strongly-convex-strongly-concave setting.** When $\mu$ is set to be positive, the function $f(x, y)$ becomes $\mu$-strongly-convex-$\mu$-strongly-concave. In this case, the problem in (88) has a unique saddle
Figure 2: The distance to the saddle point versus the number of iterations for the first-order optimistic methods on solving a strongly-convex-strongly-concave saddle point problem in (88).

point \((x^*, y^*)\) and we measure the quality of a solution \((x, y)\) by its Bregman distance to \((x^*, y^*)\). Theorem 5.1 provides the following linear convergence rate for the fixed stepsize scheme:

\[
\|z_N - z^*\|^2 \leq \|z_0 - z^*\|^2 \left( \frac{M}{(1 + s(\Phi))\mu + M} \right)^N = \|z_0 - z^*\|^2 \left( \frac{M}{2\mu + M} \right)^N, \tag{90}
\]

where we also take the symmetry coefficient of \(\Phi\) into account (cf. Remark 2.3). A similar bound can also be derived for the line search scheme from Theorem 5.3. Since the saddle point of the problem in (88) does not admit a closed form, in the experiment we approximate it by running the first-order optimistic method for a very long time (more than \(10^5\) iterations).

The problem parameters are the same as those in the convex-concave setting, except that \(\mu\) is set to 0.1. In Fig. 2, we plot the distance to the saddle point \(z^* = (x^*, y^*)\) of the fixed stepsize scheme and the line search scheme versus the number of iterations, along with the convergence bound in (90). We observe that both schemes exhibit linear convergence, and the rate of the fixed stepsize scheme agrees well with our theory. Moreover, as in the convex-concave setting, we can see that the line search scheme converges faster.

The complexity of line search. We also empirically evaluate the number of subsolver calls made in our line search scheme. To see the impacts of the line search parameters, we vary \(\sigma_0\) and \(\beta\) and run our method on both the convex-concave and strongly-convex-strongly-concave saddle point problems of different sizes. For each configuration, we generate 50 random instances and run our method until it finds a solution with accuracy \(10^{-9}\) or completes 1,000 iterations. Then we compute the average number of subsolver calls per iteration in each run and report the maximum among the 50 runs in Table 3. From the table, we can see that the line search complexity is insensitive to the choice of the line search parameters. In fact, in all test instances our method requires no more than 3 calls to the subsolver per iteration on average, which verifies that our line search scheme is highly practical.
Table 3: The maximum average number of subsolver calls per iteration in the first-order optimistic method among 50 random instances.

| Line search parameters | Convex-concave ($\mu = 0$) | Strongly-convex-strongly-concave ($\mu = 0.1$) |
|-------------------------|-------------------------------|-----------------------------------------------|
|                         | (m, n) = (200, 100)          | (m, n) = (600, 300)                           |
| $\sigma_0 = 1, \beta = 0.5$ | 2.007                        | 2.011                                         |
| $\sigma_0 = 100, \beta = 0.5$ | 2.037                        | 2.039                                         |
| $\sigma_0 = 10,000, \beta = 0.5$ | 2.240                        | 2.256                                         |
| $\sigma_0 = 1, \beta = 0.9$ | 2.256                        | 2.279                                         |
| $\sigma_0 = 100, \beta = 0.9$ | 2.260                        | 2.290                                         |
| $\sigma_0 = 10,000, \beta = 0.9$ | 2.265                        | 2.297                                         |

8.2 Second-Order Optimistic Method

To test our second-order optimistic method, we consider unconstrained saddle point problems:

$$\min_{x \in \mathbb{R}^n} \max_{y \in \mathbb{R}^n} f(x, y).$$

In the Euclidean setup where $\| \cdot \|_x = \| \cdot \|_2$, $\| \cdot \|_y = \| \cdot \|_2$ and $D_\Phi(z, z') = \frac{1}{2} \| z - z' \|_2^2$, the subproblem in (51) for the optimistic subsolver is equivalent to solving a system of linear equations. Specifically, the update rule (48) can be instantiated as

$$z_{k+1} = z_k - (I + \eta_k DF(z_k))^{-1}(\eta_k F(z_k) + v_k),$$

where $v_k = \tilde{\eta}_k (F(z_k) - F(z_{k-1}) - DF(z_{k-1})(z_k - z_{k-1}))$ and $\eta_k$ is determined by the adaptive line search scheme (see Algorithm 4). In our experiments, we set the initial point as $(x_0, y_0) = (0, 0)$.

**Convex-concave setting.** Consider the saddle point problem

$$\min_{x \in \mathbb{R}^n} \max_{y \in \mathbb{R}^n} f(x, y) = \frac{L_2}{6} \|x\|^3 + \langle Ax - b, y \rangle,$$  \hspace{1cm} (91)

where $L_2 > 0$, the entries of the vector $b \in \mathbb{R}^n$ are generated independently and randomly from the interval $[-1, 1]$ and the matrix $A \in \mathbb{R}^{n \times n}$ is given by

$$A = \begin{bmatrix} 1 & -1 & & \\ 1 & -1 & & \\ & \ddots & \ddots & \\ & & 1 & -1 \\ & & & 1 \end{bmatrix}.$$  

We can verify that the problem in (91) is convex-concave and $F$ defined in (8) is $L_2$-second-order smooth. Moreover, it has a unique saddle point $z^* = (x^*, y^*)$ given by $x^* = A^{-1}b$ and $y^* = -\frac{L_2}{2} \|x^*\|_2(A^T)^{-1}x^*$. Since the feasible set is unbounded, we will use the restricted primal-dual gap defined in (13) as the performance metric. With $B_1 = \mathbb{R}^m$ and $B_2 = \{ y : \|y\|_2 \leq R \}$, the gap can be given as

$$\Delta_{B_1 \times B_2}(x, y) = \frac{L_2}{6} \|x\|^3 + R \|Ax - b\|_2 + \frac{2}{3} \sqrt{\frac{2}{L_2} \|A^Ty\|_2^3 + \langle b, y \rangle}.$$  

49
Figure 3: The restricted primal-dual gap versus the number of iterations for the first-order and second-order optimistic methods on solving a convex-concave saddle point problem in (91).

By taking the supremum over \((x, y) \in B_1 \times B_2\) on both sides of (26) in Proposition 4.1, we obtain that 
\[
x = -\frac{2}{L_2} \sqrt{\frac{A^T \bar{y}}{\|A\|}} \quad \text{and} \quad y = \frac{R(A\bar{x} - b)}{\|A\|}
\]
leading to the following upper bound on the gap in terms of the stepsizes:
\[
\Delta_{B_1 \times B_2}(\bar{x}, \bar{y}) \leq \frac{1}{2} \left( \frac{2}{L_2} \|A^T \bar{y}\| + R^2 \|A\bar{x} - b\|^2 \right) \left( \sum_{k=0}^{N-1} \eta_k \right)^{-1},
\]
(92)
where 
\[
\bar{x}_N = \frac{1}{\sum_{k=0}^{N-1} \eta_k} \sum_{k=0}^{N-1} \eta_k x_{k+1}
\]
and 
\[
\bar{y}_N = \frac{1}{\sum_{k=0}^{N-1} \eta_k} \sum_{k=0}^{N-1} \eta_k y_{k+1}.
\]
Similarly, from (56) in Theorem 6.3, we can further obtain the following upper bound:
\[
\Delta_{B_1 \times B_2}(\bar{x}, \bar{y}) \leq \frac{\gamma L_2}{2\sqrt{2}} \left( \frac{2}{L_2} \|A^T \bar{y}\| + R^2 \|A\bar{x} - b\|^2 \right) \|z^*\|_2 N^{-\frac{3}{2}},
\]
(93)
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 Fig. 3, we plot the restricted primal-dual gap at the averaged iterate \((\bar{x}_N, \bar{y}_N)\) and at the last iterate \((x_N, y_N)\) of the second-order method. For comparison, we also plot the performance of the first-order optimistic method as well as the two convergence bounds in (92) and (93). We can see that the second-order method is far superior to the first-order method in terms of the iteration complexity: the first-order method barely makes any progress when the second-order method converges successfully. We should note that the poor performance of the first-order method is not contradictory to our theory, as the objective function in the test problem in (91) does not have Lipschitz gradient and hence our convergence bound is not directly applicable. Also, the bound in (92) provides a tight upper bound for the second-order method especially in the later stage. Moreover, while the \(O(N^{-\frac{3}{2}})\) bound in (93) does not fit the empirical results as perfectly as we have seen for the first-order method, it captures the convergence behavior of the bound in (92) well.

**Strongly-convex-strongly-concave setting.** Consider the unconstrained saddle point problem
\[
\min_{x \in \mathbb{R}^n} \max_{y \in \mathbb{R}^n} \frac{L_2}{36} \left( \sum_{k=1}^{m-1} |x_k - x_{k+1}|^3 - 3cx_1 \right) + \frac{\mu}{2} \|x\|^2 + \langle Ax, y \rangle - \frac{\mu}{2} \|y\|^2,
\]
(94)
Figure 4: The distance to the saddle point versus the number of iterations for the first-order and second-order optimistic methods on solving a strongly-convex-strongly-concave saddle point problem in (94).

where $L_2, \mu > 0$ and the entries of $A \in \mathbb{R}^{n \times m}$ are chosen independently and uniformly from the interval $[-1, 1]$. We can verify that the problem in (94) is $\mu$-strongly-convex-strongly-concave and $F$ defined in (8) is $L_2$-second-order smooth. It has a unique saddle point $z^* = (x^*, y^*)$, which we compute numerically using MATLAB’s built-in nonlinear equation solver.

As in the first-order case, we use the distance to the saddle point $z^*$ as the performance metric. By Proposition 4.1, we have

$$\|z_k - z^*\|^2 \leq \frac{2\|z_0 - z^*\|^2}{2 - \alpha} \zeta_k,$$

(95)

where $\zeta_k$ is given in (57) and can be computed from the stepsizes $\{\eta_k\}_{k \geq 0}$. To compare the empirical results with our theoretical analysis, we also consider a “simulated” sequence $\{\tilde{\zeta}_k\}_{k \geq 0}$ with $\tilde{\zeta}_0 = 1$ that follows the dynamic

$$\frac{1}{\tilde{\zeta}_k} = 1 + \frac{1}{C} \left( \sum_{l=0}^{k-1} \frac{1}{\tilde{\zeta}_l} \right)^{\frac{3}{2}},$$

(96)

for any $k \geq 1$ for some $C > 0$. In light of Lemma 6.4, we can see that $\zeta_k \leq \tilde{\zeta}_k$ for all $k \geq 0$ if the parameter $C$ is chosen as $\gamma_2 \kappa_2(z_0)$.

In our experiment, the problem parameters are chosen as $m = 400$, $n = 200$, $L_2 = 10,000$, $c = 100$, $\mu = 1$, and the line search parameters are $\alpha = 0.5$, $\beta = 0.5$, $\sigma_0 = 1$, and $\epsilon = 10^{-10}$. In Fig. 4 we plot the distance to the saddle point $z^*$ of the second-order optimistic method versus the number of iterations. For comparison, we also plot the performance of the first-order optimistic method and the bound in (95) using the actual sequence $\{\zeta_k\}_{k \geq 0}$ as well as the simulated sequence $\{\tilde{\zeta}_k\}_{k \geq 0}$. As in the convex-concave case, we can see that the second-order method converges much faster than the first-order method. In particular, it exhibits local superlinear convergence in the neighborhood of the saddle point, doubling the accuracy of the solution within a few iterations. Again, we note that our convergence theory for the first-order method is not directly applicable to (94) as it does not have Lipschitz gradient. Also, the bound in (95) using $\{\zeta_k\}_{k \geq 0}$ provides a tight upper bound for the second-order method. Moreover, if the parameter $C$ in (96) is carefully chosen, we observe that the bound in (95) using $\{\tilde{\zeta}_k\}_{k \geq 0}$ 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. As in the first-order case, we vary the line search parameters \( \sigma_0 \) and \( \beta \) and run our method on Problems (91) and (94) of different sizes until it finds a solution with accuracy \( 10^{-10} \) or completes 500 iterations. The maximum is reported in Table 4. We can see that the line search complexity is insensitive to the choice of the initial trial stepsize \( \sigma_0 \), while a larger \( \beta \) could lead to more calls to the subsolver. Still, in our test instances the average number of subsolver calls per iteration can be controlled under 4 when \( \beta \) is chosen as 0.5, which is a reasonable price to pay given its superior convergence performance.

| Line search parameters | Max average number of subsolver calls per iteration | Problem (91) | Problem (94) |
|------------------------|-----------------------------------------------|--------------|--------------|
| \( \sigma_0 = 1, \beta = 0.5 \) | \( \sigma_0 = 100, \beta = 0.5 \) | 2.010 | 2.004 |
| \( \sigma_0 = 10,000, \beta = 0.5 \) | \( \sigma_0 = 100, \beta = 0.5 \) | 2.025 | 2.012 |
| \( \sigma_0 = 1, \beta = 0.9 \) | \( \sigma_0 = 100, \beta = 0.9 \) | 2.295 | 2.088 |
| \( \sigma_0 = 10,000, \beta = 0.9 \) | \( \sigma_0 = 100, \beta = 0.9 \) | 2.181 | 2.102 |

9 Conclusions and Final Remarks

In this paper, we proposed the generalized optimistic method for composite convex-concave saddle point problems by approximating the Bregman proximal point method. We also designed a novel line search scheme to select the stepsize in our method adaptively and without knowledge of the smoothness coefficients of the objective. Under our unified framework, our first-order optimistic method provably achieves a complexity bound of \( O(1/\epsilon) \) in terms of the restricted primal-dual gap in the convex-concave setting, and a complexity bound of \( O(k_1 \log \frac{1}{\epsilon}) \) in terms of the Bregman distance to the saddle point in the strongly-convex-strongly-concave setting. Furthermore, our \( p \)-th-order optimistic method \( (p \geq 2) \) provably achieves a complexity bound of \( O(1/\epsilon^{\frac{2}{p+1}}) \) in the convex-concave setting, and a complexity bound of \( O((k_p(z_0))^{\frac{2}{p+1}} + \log \log \frac{1}{\epsilon}) \) in the strongly-convex-strongly-concave setting. We also proved that our line search scheme only requires an almost constant number of calls to the optimistic subsolvers per iteration on average, which is supported by our numerical experiments.

While we focus on solving the saddle point problems in this paper, we mention that it is straightforward to apply our results to the convex minimization problems as well as the more general monotone inclusion problems. Moreover, we note that the worst-case iteration complexities of second-order and higher-order methods for the saddle point problems are less understood, and the only known results come from the convex optimization literature. Specifically, for the problem of minimizing a convex function or a strongly convex function with Lipschitz \( p \)-th-order derivative \( (p \geq 2) \), recent works [ASS19; Nes19; KS20] have established a worst-case complexity bound of

\[
\Omega \left( \left( \frac{1}{\epsilon} \right)^{\frac{2}{p+1}} \right) \quad \text{and} \quad \Omega \left( \left( \frac{L_p \|z_0 - z^*\|_2^{p-1}}{\mu} \right)^{\frac{2}{p+1}} + \log \log \frac{1}{\epsilon} \right)
\]  

(97)
respectively, where $L_p$ is the Lipschitz constant of the $p$-th-order derivative and $\mu$ is the strongly-convex coefficient. Interestingly, our corresponding results bear a similar structure as (97), except that the number on the power is $2/(p + 1)$ instead of $2/(3p + 1)$. An interesting future direction will be improving our complexity bounds of second-order and higher-order methods for the saddle point problems, or showing that this is impossible by establishing a lower bound.

### Acknowledgements

This research of R. Jiang and A. Mokhtari is supported in part by NSF Grants 2007668, 2019844, and 2112471, ARO Grant W911NF2110226, the Machine Learning Lab (MLL) at UT Austin, and the Wireless Networking and Communications Group (WNCG) Industrial Affiliates Program.

### References

[ASS19] Yossi Arjevani, Ohad Shamir, and Ron Shiff. “Oracle Complexity of Second-Order Methods for Smooth Convex Optimization”. In: Mathematical Programming 178.1-2 (2019), pp. 327–360 (page 52).

[AHU58] K. J. Arrow, L. Hurwicz, and H. Uzawa. Studies in Linear and Non-Linear Programming. Stanford Mathematical Studies in the Social Sciences. Stanford, CA: Stanford University Press, 1958 (page 6).

[AT05] Alfred Auslender and Marc Teboulle. “Interior Projection-Like Methods for Monotone Variational Inequalities”. In: Mathematical Programming 104.1 (2005), pp. 39–68 (page 6).

[AMLG20] Waïss Azizian, Ioannis Mitliagkas, Simon Lacoste-Julien, and Gauthier Gidel. “A Tight and Unified Analysis of Gradient-Based Methods for a Whole Spectrum of Differentiable Games”. In: Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS). Vol. 108. Proceedings of Machine Learning Research. PMLR, 2020, pp. 2863–2873 (pages 6, 7).

[BO98] Tamer Başar and Geert Jan Olsder. Dynamic Noncooperative Game Theory, 2nd Edition. SIAM, 1998 (page 4).

[BBT17] Heinz H. Bauschke, Jérôme Bolte, and Marc Teboulle. “A Descent Lemma beyond Lipschitz Gradient Continuity: First-Order Methods Revisited and Applications”. In: Mathematics of Operations Research 42.2 (2017), pp. 330–348 (page 10).

[Bec17] Amir Beck. First-Order Methods in Optimization. SIAM, 2017 (page 63).

[BGN09] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. Robust Optimization. Princeton Ser. Appl. Math 28. Princeton, NJ: Princeton University Press, 2009 (page 4).

[BL20] Brian Bullins and Kevin A. Lai. Higher-Order Methods for Convex-Concave Min-max Optimization and Monotone Variational Inequalities. 2020. arXiv: 2007.04528 [math.OC] (pages 8, 9, 17, 34, 38, 43).
[CGT11] Coralia Cartis, Nicholas I. M. Gould, and Philippe L. Toint. “Adaptive Cubic Regularisation Methods for Unconstrained Optimization. Part I: Motivation, Convergence and Numerical Results”. In: Mathematical Programming 127.2 (2011), pp. 245–295 (page 8).

[CP11] Antonin Chambolle and Thomas Pock. “A First-Order Primal-Dual Algorithm for Convex Problems With Applications to Imaging”. In: Journal of Mathematical Imaging and Vision 40.1 (2011), pp. 120–145 (pages 7, 13).

[CP16] Antonin Chambolle and Thomas Pock. “On the Ergodic Convergence Rates of a First-Order Primal–Dual Algorithm”. In: Mathematical Programming 159.1-2 (2016), pp. 253–287 (pages 7, 47).

[CT93] Gong Chen and Marc Teboulle. “Convergence Analysis of a Proximal-Like Minimization Algorithm Using Bregman Functions”. In: SIAM Journal on Optimization 3.3 (1993), pp. 538–543 (pages 10, 15).

[CLO14] Yunmei Chen, Guanghui Lan, and Yuyuan Ouyang. “Optimal Primal-Dual Methods for a Class of Saddle Point Problems”. In: SIAM Journal on Optimization 24.4 (2014), pp. 1779–1814 (page 7).

[CYLMLJZ12] Chao-Kai Chiang, Tianbao Yang, Chia-Jung Lee, Mehrdad Mahdavi, Chi-Jen Lu, Rong Jin, and Shenghuo Zhu. “Online Optimization with Gradual Variations”. In: Proceedings of the 25th Annual Conference on Learning Theory (COLT). Vol. 23. Proceedings of Machine Learning Research. JMLR Workshop and Conference Proceedings, 2012, pp. 61–620 (pages 4, 16).

[Con13] Laurent Condat. “A Primal-Dual Splitting Method for Convex Optimization Involving Lipschitzian, Proximable and Linear Composite Terms”. In: Journal of Optimization Theory and Applications 158.2 (2013), pp. 460–479. ISSN: 1573-2878 (page 7).

[DISZ18] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. “Training GANs with Optimism”. In: Proceedings of International Conference on Learning Representations (ICLR). 2018 (pages 4, 7, 16).

[Eck93] Jonathan Eckstein. “Nonlinear Proximal Point Algorithms Using Bregman Functions, with Applications to Convex Programming”. In: Mathematics of Operations Research 18.1 (1993), pp. 202–226 (pages 15, 20).

[FP03] Francisco Facchinei and Jong-Shi Pang. Finite-Dimensional Variational Inequalities and Complementarity Problems. New York: Springer-Verlag, 2003 (pages 6, 11, 31).

[FOP20] Alireza Fallah, Asuman Ozdaglar, and Sarath Pattathil. “An Optimal Multistage Stochastic Gradient Method for Minimax Problems”. In: Proceedings of the 59th IEEE Conference on Decision and Control (CDC). IEEE, 2020 (page 6).

[GBVVL19] Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. “A Variational Inequality Perspective on Generative Adversarial Networks”. In: Proceedings of International Conference on Learning Representations (ICLR). 2019 (pages 4, 7, 8, 16, 27).
Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. “Generative Adversarial Nets”. In: Advances in Neural Information Processing Systems. Vol. 27. Curran Associates, Inc., 2014 (page 4).

Erfan Yazdandoost Hamedani and Necdet Serhat Aybat. “A Primal-Dual Algorithm with Line Search for General Convex-Concave Saddle Point Problems”. In: SIAM Journal on Optimization 31.2 (2021), pp. 1299–1329 (pages 7, 8).

Niao He, Anatoli Juditsky, and Arkadi Nemirovski. “Mirror Prox Algorithm for Multi-Term Composite Minimization and Semi-Separable Problems”. In: Computational Optimization and Applications 61.2 (2015), pp. 275–319 (page 6).

Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. “On the Convergence of Single-Call Stochastic Extra-Gradient Methods”. In: Advances in Neural Information Processing Systems. Vol. 32. Curran Associates, Inc., 2019 (page 7).

Kevin Huang, Junyu Zhang, and Shuzhong Zhang. Cubic Regularized Newton Method for Saddle Point Models: A Global and Local Convergence Analysis. 2020. arXiv: 2008.09919 [math.OC] (pages 8, 9, 38).

Guy Kornowski and Ohad Shamir. High-Order Oracle Complexity of Smooth and Strongly Convex Optimization. 2020. arXiv: 2010.06642 [math.OC] (page 52).

G. Korpelevich. “The Extragradient Method for Finding Saddle Points and Other Problems”. In: Ekonomika i Matematicheskie Metody 12 (1976). In Russian; English translation in Matekon, pp. 747–756 (pages 4, 6, 7, 16).

Georgios Kotsalis, Guanghui Lan, and Tianjiao Li. Simple and Optimal Methods for Stochastic Variational Inequalities, I: Operator Extrapolation. 2020. arXiv: 2011.02987 [math.OC] (pages 7, 8, 14, 27).

Tengyuan Liang and James Stokes. “Interaction Matters: A Note on Non-Asymptotic Local Convergence of Generative Adversarial Networks”. In: Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS). Vol. 89. Proceedings of Machine Learning Research. PMLR, 2019, pp. 907–915 (pages 4, 16).

Haihao Lu, Robert M. Freund, and Yurii Nesterov. “Relatively Smooth Convex Optimization by First-Order Methods, and Applications”. In: SIAM Journal on Optimization 28.1 (2018), pp. 333–354 (page 10).

Yu Malitsky. “Proximal Extrapolated Gradient Methods for Variational Inequalities”. In: Optimization Methods and Software 33.1 (2017), pp. 140–164 (pages 7, 28).

Yu. Malitsky. “Projected Reflected Gradient Methods for Monotone Variational Inequalities”. In: SIAM Journal on Optimization 25.1 (2015), pp. 502–520 (page 7).

Yura Malitsky and Thomas Pock. “A First-Order Primal-Dual Algorithm with Linesearch”. In: SIAM Journal on Optimization 28.1 (2018), pp. 411–432 (page 7).

Yura Malitsky and Matthew K. Tam. “A Forward-Backward Splitting Method for Monotone Inclusions without Cocoercivity”. In: SIAM Journal on Optimization 30.2 (2020), pp. 1451–1472 (pages 4, 7, 8, 21, 27, 28).
B. Martinet. “Brève Communication. Régularisation D’inéquations Variationnelles Par Approximations Successives”. In: *ESIAM Mathematical Modelling and Numerical Analysis* 4.R3 (1970), pp. 154–158 (page 4).

Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. “A Unified Analysis of Extra-Gradient and Optimistic Gradient Methods for Saddle Point Problems: Proximal Point Approach”. In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics (AISTATS)*. Vol. 108. Proceedings of Machine Learning Research. PMLR, 2020, pp. 1497–1507 (pages 1, 4–8, 16, 17, 27).

Aryan Mokhtari, Asuman E. Ozdaglar, and Sarath Pattathil. “Convergence Rate of $O(1/k)$ for Optimistic Gradient and Extragradient Methods in Smooth Convex-Concave Saddle Point Problems”. In: *SIAM Journal on Optimization* 30.4 (Jan. 2020), pp. 3230–3251 (pages 4, 5, 7, 8, 16, 27).

Renato D. C. Monteiro and B. F. Svaiter. “Complexity of Variants of Tseng’s Modified F-B Splitting and Korpelevich’s Methods for Hemivariational Inequalities with Applications to Saddle-Point and Convex Optimization Problems”. In: *SIAM Journal on Optimization* 21.4 (2011), pp. 1688–1720 (pages 6–8).

Renato D. C. Monteiro and B. F. Svaiter. “On the Complexity of the Hybrid Proximal Extragradient Method for the Iterates and the Ergodic Mean”. In: *SIAM Journal on Optimization* 20.6 (2010), pp. 2755–2787 (pages 8, 17).

Renato D. C. Monteiro and Benar F. Svaiter. “Iteration-Complexity of a Newton Proximal Extragradient Method for Monotone Variational Inequalities and Inclusion Problems”. In: *SIAM Journal on Optimization* 22.3 (2012), pp. 914–935 (pages 6, 8, 9, 17, 20, 34, 63).

Arkadi 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”. In: *SIAM Journal on Optimization* 15.1 (2004), pp. 229–251 (pages 4, 6–8, 15, 16).

Yu. Nesterov. “Accelerating the Cubic Regularization of Newton’s Method on Convex Problems”. In: *Mathematical Programming* 112.1 (2008), pp. 159–181 (pages 7, 8, 35).

Yuuri Nesterov. “Cubic Regularization of Newton’s Method for Convex Problems with Constraints”. In: *CORE Discussion Paper No. 2006/39* (2006) (pages 8, 9, 35).

Yuuri Nesterov. “Dual Extrapolation and Its Applications to Solving Variational Inequalities and Related Problems”. In: *Mathematical Programming* 109.2-3 (2007), pp. 319–344 (pages 6, 8, 13).

Yuuri Nesterov. “Implementable Tensor Methods in Unconstrained Convex Optimization”. In: *Mathematical Programming* (2019) (pages 41, 52).

Yuuri Nesterov and B.T. Polyak. “Cubic Regularization of Newton Method and Its Global Performance”. In: *Mathematical Programming* 108.1 (2006), pp. 177–205 (pages 7, 8).
A Proofs for Preliminary Lemmas

A.1 Proof of Lemma 2.2

We prove the lemma in a united way by regarding Assumption 2.1 as a special case of Assumption 2.2 with $\mu = 0$. Let $z = (x, y)$ and $z' = (x', y')$ be any two points in $\mathcal{X} \times \mathcal{Y}$. Since $f$ is $\mu$-strongly convex in $x$ w.r.t. $\Phi_\mathcal{X}$, we have

$$f(x', y) \geq f(x, y) + \langle \nabla_x f(x, y), x' - x \rangle + \mu D_{\Phi_\mathcal{X}}(x', x),$$

$$f(x, y') \geq f(x', y') + \langle \nabla_x f(x', y'), x - x' \rangle + \mu D_{\Phi_\mathcal{X}}(x, x').$$

Similarly, since $f$ is $\mu$-strongly concave in $y$ w.r.t. $\Phi_\mathcal{Y}$, we have

$$-f(x, y') \geq -f(x, y) - \langle \nabla_y f(x, y), y' - y \rangle + \mu D_{\Phi_\mathcal{Y}}(y', y),$$

$$-f(x', y) \geq -f(x', y') - \langle \nabla_y f(x', y'), y - y' \rangle + \mu D_{\Phi_\mathcal{Y}}(y, y').$$

By summing all the inequalities above, we get

$$\langle F(z) - F(z'), z - z' \rangle \geq \mu (D_{\Phi_\mathcal{X}}(z', z) + D_{\Phi_\mathcal{Y}}(z, z')),$$

where we used the definition of $F$ in (8) and the equality in (11). This, by definition, shows that $F$ is $\mu$-strongly monotone w.r.t. $\Phi_\mathcal{Z}$ on $\mathcal{Z}$.

A.2 Proof of Lemma 2.3

For any $z_k = (x_k, y_k), z = (x, y) \in \mathcal{X} \times \mathcal{Y}$, by the definitions of $\ell$ and $h$ we have

$$\ell(x_k, y) - \ell(x, y_k) = f(x_k, y) - f(x, y) + h_1(x_k) + h_2(y_k) - h_1(x) - h_2(y)$$

$$= f(x_k, y) - f(x, y) + h(z_k) - h(z).$$

(98)

Since $f$ is convex in $x$ and concave in $y$, by Assumption 2.1 we have

$$f(x_k, y_k) - f(x, y_k) \leq \langle \nabla_f(x_k, y_k), x_k - x \rangle, \quad f(x, y) - f(x_k, y_k) \leq \langle -\nabla_f(x_k, y_k), y_k - y \rangle.$$

Adding these two inequalities gives us

$$f(x_k, y) - f(x, y_k) \leq \langle F(z_k), z_k - z \rangle,$$

(99)

where we used the definition of $F$ in (8). Hence, combining (98) and (99) leads to

$$\sum_{k=1}^N \theta_k \ell(x_k, y) - \sum_{k=1}^N \theta_k \ell(x, y_k) \leq \sum_{k=1}^N \theta_k (\langle F(z_k), z_k - z \rangle + h(z_k) - h(z)).$$

Finally, since $\ell$ is convex in $x$ and concave in $y$, by Jensen’s inequality we have

$$\sum_{k=1}^N \theta_k \ell(x_k, y) \geq \ell(\bar{x}_N, y) \quad \text{and} \quad \sum_{k=1}^N \theta_k \ell(x, y_k) \leq \ell(x, \bar{y}_N),$$

from which the result follows.

58
A.3 Proof of Lemma 2.4

Since \( z^* \) is a saddle point of Problem (1), it also solves the variational inequality in (10). Hence, for any \( z \in \mathcal{Z} \),
\[
\langle F(z^*), z - z^* \rangle + h(z) - h(z^*) \geq 0. \tag{100}
\]
Moreover, since \( F \) is \( \mu \)-strongly monotone w.r.t. \( \Phi \), we have
\[
\langle F(z) - F(z^*), z - z^* \rangle \geq \mu D_\Phi(z, z^*) + \mu D_\Phi(z^*, z) \geq \mu D_\Phi(z^*, z). \tag{101}
\]
Adding (100) and (101) gives us the desired result.

B Proof of Theorem 3.1

The update rule for the BPP method in (16) implies that
\[
\eta_k F(z_{k+1}) + \nabla \Phi(z_{k+1}) - \nabla \Phi(z_k) \in -\eta_k H(z_{k+1}).
\]
Hence, by using the definitions of \( H \) in (8), for any \( z \in \mathcal{Z} \) we have
\[
\eta_k \langle F(z_{k+1}), z_{k+1} - z \rangle + \langle \nabla \Phi(z_{k+1}) - \nabla \Phi(z_k), z_{k+1} - z \rangle \leq \eta_k (h(z) - h(z_{k+1})).
\]
We apply the three-point identity (4) and rearrange the terms to get
\[
\eta_k \langle (F(z_{k+1}), z_{k+1} - z) + h(z_{k+1}) - h(z) \rangle \leq D_\Phi(z, z_k) - D_\Phi(z, z_{k+1}) - D_\Phi(z_{k+1}, z_k)
\leq D_\Phi(z, z_k) - D_\Phi(z, z_{k+1}). \tag{102}
\]

**Proof of Part (a).** In this part, we assume that Assumption 2.1 holds. We first set \( z = z^* \) in (102). By invoking Lemma 2.4 with \( z = z_{k+1} \) and \( \mu = 0 \), we get \( D_\Phi(z^*, z_{k+1}) \leq D_\Phi(z^*, z_k) \) for any \( k \geq 0 \).

Next, we sum both sides of (102) over \( k = 0, 1, \ldots, N - 1 \) to get
\[
\sum_{k=0}^{N-1} \eta_k \langle (F(z_{k+1}), z_{k+1} - z) + h(z_{k+1}) - h(z) \rangle \leq D_\Phi(z, z_0) - D_\Phi(z, z_N) \leq D_\Phi(z, z_0). \tag{103}
\]
Now the result follows from dividing both sides of (103) by \( \sum_{k=0}^{N-1} \eta_k \) and applying Lemma 2.3.

**Proof of Part (b).** In this part, we assume that Assumption 2.2 holds. We set \( z = z^* \) in (102) and apply Lemma 2.4 with \( z = z_{k+1} \) to get
\[
(1 + \mu \eta_k) D_\Phi(z^*, z_{k+1}) \leq D_\Phi(z^*, z_k), \quad \forall k \geq 0.
\]
The result now follows by induction on the iteration counter \( k \).
C Proofs for Generalized Optimistic Method

Like many convergence proofs in optimization literature, our proof relies on a carefully-designed Lyapunov function. We define the function and discuss its properties in the following lemma, which will be the cornerstone of proving Proposition 4.1.

Lemma C.1. Let \( \{z_k\}_{k \geq 0} \) be the iterates generated by Algorithm 1 with \( \mu \geq 0, 0 < \alpha \leq 1 \), and \( \eta_k \) chosen as in Proposition 4.1. Define a Lyapunov function

\[
V(z_k, z_{k-1}; z) = -\frac{\eta_k}{1 + \eta_k \mu} \langle F(z_k) - P(z_k; I_{k-1}), z_k - z \rangle + D_\Phi(z, z_k) + \frac{\alpha \|z_k - z_{k-1}\|^2}{4(1 + \eta_k \mu)^2},
\]

(104)

where \( z \in Z \) is an arbitrary point. Then the following statements hold:

(a) For any \( z \in Z \) and \( k \geq 0 \), we have

\[
V(z_0, z_{-1}; z) = D_\Phi(z, z_0) \quad \text{and} \quad V(z_k, z_{k-1}; z) \geq \frac{2 - \alpha}{2} D_\Phi(z, z_k).
\]

(b) For any \( z \in Z \) and \( k \geq 0 \), we have

\[
\eta_k \langle F(z_{k+1}), z_{k+1} - z \rangle + h(z_{k+1}) - h(z) - \mu D_\Phi(z, z_{k+1}) \\
\leq V(z_k, z_{k-1}; z) - (1 + \eta_k \mu) V(z_{k+1}, z_k; z) + \frac{1 - \alpha}{2} \|z_{k+1} - z_k\|^2.
\]

(105)

Proof. For Part (a), note that we initialize Algorithm 1 with \( z_{-1} = z_0 \) and \( P(z_0; I_{-1}) = F(z_0) \). Hence, it is straightforward to verify that \( V(z_0, z_{-1}; z) = D_\Phi(z, z_0) \). Moreover, we can lower bound \( V(z_k, z_{k-1}; z) \) by

\[
V(z_k, z_{k-1}; z) \geq -\frac{\eta_k}{1 + \eta_k \mu} \|F(z_k) - P(z_k; I_{k-1})\|_* \|z_k - z\| + D_\Phi(z, z_k) + \frac{\alpha \|z_k - z_{k-1}\|^2}{4(1 + \eta_k \mu)^2}
\]

(106)

\[
\geq -\frac{\alpha}{2(1 + \eta_k \mu)} \|z_k - z_{k-1}\| \|z_k - z\| + D_\Phi(z, z_k) + \frac{\alpha \|z_k - z_{k-1}\|^2}{4(1 + \eta_k \mu)^2}
\]

(107)

\[
\geq \frac{\alpha}{4} \|z_k - z\|^2 + D_\Phi(z, z_k)
\]

(108)

\[
\geq \frac{2 - \alpha}{2} D_\Phi(z, z_k),
\]

(109)

where we used the generalized Cauchy-Schwarz inequality in (106), the condition (24) in (107), Young’s inequality\(^\text{7}\) in (108), and the strong convexity of \( \Phi \) in (109). This completes the proof for Part (a).

For Part (b), by the definition of \( H \) in (8), the update rule (22) implies that

\[
\left\langle \eta_k P(z_{k+1}; I_{k}) + \dot{\eta}_k (F(z_k) - P(z_k; I_{k-1})) + \nabla \Phi(z_{k+1}) - \nabla \Phi(z_k), z_{k+1} - z \right\rangle \leq \eta_k (h(z) - h(z_{k+1}))
\]

for any \( z \in Z \). Moreover, note that the stepsize \( \dot{\eta}_k \) in Proposition 4.1 can be written as \( \dot{\eta}_k = \eta_{k-1}/(1 + \eta_{k-1} \mu) \) for \( \mu \geq 0 \). Hence, by applying the three-point identity (4) and rearranging the

\(^7\)In this paper, it refers to the elementary inequality that \( ab \leq \frac{a^2}{2} + \frac{b^2}{2} \) for any \( a, b \in \mathbb{R} \).
The first two bracketed terms in (110) resemble the Lyapunov function defined in (104), and now we upper bound the remaining terms. By the generalized Cauchy-Schwarz inequality, the condition (24), Young’s inequality and the strong convexity of $\Phi$, we have

\[
\frac{\eta_{k-1}}{1 + \eta_{k-1} \mu} (F(z_k) - P(z_k; \mathcal{I}_{k-1}), z_k - z_k) + D_\Phi(z_k, z_{k+1}) \\
\leq \frac{\eta_{k-1}}{1 + \eta_{k-1} \mu} \|F(z_k) - P(z_k; \mathcal{I}_{k-1})\|_{\Phi} \|z_{k+1} - z_k\| - \frac{1}{2} \|z_{k+1} - z_k\|^2 \\
\leq \frac{\alpha}{2(1 + \eta_{k-1} \mu)} \|z_k - z_{k-1}\| \|z_{k+1} - z_k\| - \frac{1}{2} \|z_{k+1} - z_k\|^2 \\
\leq \frac{\alpha}{4(1 + \eta_{k-1} \mu)^2} \|z_k - z_{k-1}\|^2 + \frac{\alpha}{4} \|z_{k+1} - z_k\|^2 - \frac{1}{2} \|z_{k+1} - z_k\|^2 \\
\leq \frac{\alpha}{4(1 + \eta_{k-1} \mu)^2} - \frac{\alpha}{4(1 + \eta_{k-1} \mu)} - \frac{1 - \alpha}{2} \|z_{k+1} - z_k\|^2. \tag{111}
\]

Then (105) follows directly from (110) and (111). This completes the proof for Part (b). \hfill \Box

C.1 Proof of Proposition 4.1

Proof of Part (a). In this part, we assume that Assumption 2.1 holds and hence $\mu = 0$. We first apply Lemma C.1(b) with $z = z^*$. Combining this with Lemma 2.4, we get

\[
0 \leq \eta_k (\langle F(z_{k+1}), z_{k+1} - z^* \rangle + h(z_{k+1}) - h(z^*)) \\
\leq V(z_k, z_{k-1}; z^*) - V(z_{k+1}, z_{k}; z^*) - \frac{1 - \alpha}{2} \|z_{k+1} - z_k\|^2 \leq V(z_k, z_{k-1}; z^*) - V(z_{k+1}, z_{k}; z^*)
\]

to any $k \geq 0$. This implies that $V(z_N, z_{N-1}; z^*) \leq V(z_0, z_{-1}; z^*)$. Then by Lemma C.1(a), we get

\[
\frac{2 - \alpha}{2} D_\Phi(z^*, z_N) \leq V(z_N, z_{N-1}; z^*) \leq V(z_0, z_{-1}; z^*) = D_\Phi(z^*, z_0),
\]

which proves (25).

Next, note that $\frac{1 - \alpha}{2} \|z_{k+1} - z_k\|^2 \geq 0$ and hence from Lemma C.1(b) we get

\[
\eta_k (\langle F(z_{k+1}), z_{k+1} - z \rangle + h(z_{k+1}) - h(z)) \leq V(z_k, z_{k-1}; z) - V(z_{k+1}, z_{k}; z)
\]
for any \( z \in \mathcal{Z} \) and \( k \geq 0 \). Summing the above inequality over \( k = 0, 1, \ldots, N - 1 \) leads to
\[
\sum_{k=0}^{N-1} \eta_k ((F(z_{k+1}), z_{k+1} - z) + h(z_{k+1}) - h(z)) \leq V(z_0, z_{-1}; z) - V(z_N, z_{N-1}; z)
\]
\[
\leq D_\Phi(z, z_0) - \frac{2 - \alpha}{2} D_\Phi(z, z_{N-1}) \leq D_\Phi(z, z_0),
\]
where we used Lemma C.1(a) in (112). Now Part (a) follows by similar arguments as in the proof of Theorem 3.1(a).

**Proof of Part (b).** In this part, we assume that Assumption 2.2 holds. We invoke Lemma C.1(b) with \( z = z^* \) and Lemma 2.4 to get
\[
0 \leq \eta_k ((F(z_{k+1}), z_{k+1} - z^*) + h(z_{k+1}) - h(z^*) - \mu D_\Phi(z^*, z_{k+1})) \leq V(z_k, z_{k-1}; z^*) - (1 + \eta_k \mu) V(z_{k+1}, z_k; z^*).
\]
By using induction and Lemma C.1(a), we arrive at
\[
\frac{2 - \alpha}{2} D_\Phi(z^*, z_N) \leq V(z_N, z_{N-1}; z^*) \leq V(z_0, z_{-1}; z^*) \prod_{k=0}^{N-1} (1 + \eta_k \mu)^{-1} \leq D_\Phi(z^*, z_0) \prod_{k=0}^{N-1} (1 + \eta_k \mu)^{-1}.
\]
This completes the proof.

**C.2 Proof of Lemma 4.2**

Since we can regard the convex-concave setting as a special case of the strongly-convex-strongly-concave setting with \( \mu = 0 \), we will prove Lemma 4.2 under both assumptions in a united way.

We apply Lemma C.1(b) with \( z = z^* \). Together with Lemma 2.4, (105) implies that
\[
\frac{1 - \alpha}{2} \| z_{k+1} - z_k \|^2 \leq V(z_k, z_{k-1}; z^*) - (1 + \eta_k \mu) V(z_{k+1}, z_k; z^*).
\]
Multiplying both sides of (113) by \( \prod_{l=0}^{k-1} (1 + \eta_l \mu) \) and summing the inequality over \( k = 0, 1, \ldots, N - 1 \) give us
\[
\frac{1 - \alpha}{2} \sum_{k=0}^{N-1} \left( \| z_{k+1} - z_k \|^2 \prod_{l=0}^{k-1} (1 + \eta_l \mu) \right) \leq \sum_{k=0}^{N-1} \left( V(z_k, z_{k-1}; z^*) \prod_{l=0}^{k-1} (1 + \eta_l \mu) - V(z_{k+1}, z_k; z^*) \prod_{l=0}^{k} (1 + \eta_l \mu) \right)
\]
\[
= V(z_0, z_{-1}; z^*) - V(z_N, z_{N-1}; z^*) \prod_{l=0}^{N-1} (1 + \eta_l \mu).
\]
Then the lemma follows from the facts that \( V(z_0, z_{-1}; z^*) = D_\Phi(z^*, z_0) \) and \( V(z_N, z_{N-1}; z^*) \geq 0 \) (cf. Lemma C.1(b)).
D Proofs for the Second-Order Optimistic Method

D.1 Proof of Lemma 6.1

We will need the following lemma adapted from [MS12, Lemma 4.3].

**Lemma D.1.** Suppose that $A$ is a maximal monotone operator and $\Phi$ is $1$-strongly convex and $L_\Phi$-smooth w.r.t. $\| \cdot \|$ on $\mathcal{Z}$. Define $\phi_A(\eta; w) = \| \nabla \Phi((\eta A + \nabla \Phi)^{-1} w) - w \|_s$, where $(\eta A + \nabla \Phi)^{-1} w$ denotes the unique solution of the monotone inclusion problem of $z$:

$$
0 \in \eta A(z) + \nabla \Phi(z) - w. \quad (114)
$$

If $0 < \eta < \eta'$, then

$$
\phi_A(\eta'; w) \leq \frac{\eta'}{\eta} \sqrt{L_\Phi} \phi_A(\eta; w). \quad (115)
$$

**Proof.** Let $z := (\eta A + \nabla \Phi)^{-1} w$, $u := (w - \nabla \Phi(z))/\eta$ and $z' := (\eta' A + \nabla \Phi)^{-1} w$, $u' := (w - \nabla \Phi(z'))/\eta'$. In these notations, our goal in (115) is equivalent to $\|u'\|_s \leq \sqrt{L_\Phi}\|u\|_s$.

By definition, we can write

$$
u = \frac{w - \nabla \Phi(z)}{\eta} \Leftrightarrow z = \nabla \Phi^*(w - \eta u) \quad \text{and} \quad z := (\eta A + \nabla \Phi)^{-1} w \Leftrightarrow z \in A^{-1}(u),$$

where $\nabla \Phi^*$ denotes the Fenchel conjugate of $\Phi$. Moreover, by the assumptions on $\Phi$, we note that $\Phi^*$ is $1/L_\Phi$-strongly convex and $1$-smooth w.r.t. $\| \cdot \|$ (see [Bec17, Theorem 5.26]). From (114) we can write $u \in A(z)$, which implies that

$$z = \nabla \Phi^*(w - \eta u) \in A^{-1}(u).$$

Similarly, we have

$$\nabla \Phi^*(w - \eta' u') \in A^{-1}(u').$$

Since $A$ is maximal monotone, so is the operator $A^{-1}$. Thus, we get

$$\langle \nabla \Phi^*(w - \eta u) - \nabla \Phi^*(w - \eta' u'), u - u' \rangle \geq 0$$

$$\Leftrightarrow \langle \nabla \Phi^*(w - \eta u') - \nabla \Phi^*(w - \eta' u'), u - u' \rangle \geq \langle \nabla \Phi^*(w - \eta u') - \nabla \Phi^*(w - \eta u), u - u' \rangle. \quad (116)$$

For the right-hand side of (116), we have from the convexity of $\Phi^*$ that

$$\langle \nabla \Phi^*(w - \eta u') - \nabla \Phi^*(w - \eta u), u - u' \rangle = \frac{1}{\eta} \langle \nabla \Phi^*(w - \eta u') - \nabla \Phi^*(w - \eta u), w - \eta u' - (w - \eta u) \rangle \geq 0. \quad (117)$$

For the left-hand side of (116), we can use the three-point identity (4) to get

$$\langle \nabla \Phi^*(w - \eta u') - \nabla \Phi^*(w - \eta' u'), u - u' \rangle$$

$$= \frac{1}{\eta' - \eta} \langle \nabla \Phi^*(w - \eta u') - \nabla \Phi^*(w - \eta' u'), [w - \eta u' + (\eta' - \eta)u] - (w - \eta u') \rangle$$

$$= - \frac{1}{\eta' - \eta}[D_{\Phi^*}(w - \eta u', w - \eta' u') + D_{\Phi^*}(w - \eta' u' + (\eta' - \eta)u, w - \eta u')$$

$$- D_{\Phi^*}(w - \eta' u' + (\eta' - \eta)u, w - \eta u')]. \quad (118)$$
Since \( \eta' > \eta \), (116), (117), and (118) together imply that
\[
D_{\Phi^*}(w - \eta u', w - \eta' u') \leq D_{\Phi^*}(w - \eta u' + (\eta' - \eta)u, w - \eta' u').
\] (119)

Furthermore, since \( \Phi^* \) is \( 1/L_\Phi \)-strongly convex and \( 1 \)-smooth w.r.t. \( \| \cdot \| \), we have
\[
D_{\Phi^*}(w - \eta u', w - \eta' u') \geq \frac{1}{2L_\Phi} \| (\eta' - \eta)u' \|_*^2
\]
\[
D_{\Phi^*}(w - \eta u' + (\eta' - \eta)u, w - \eta' u') \leq \frac{1}{2} \| (\eta' - \eta)u \|_*^2.
\] (120)

Combining (119) and (120) gives us \( \| u' \|_* \leq \sqrt{L_\Phi} \| u \|_* \). The proof is complete. \( \square \)

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

**Lemma D.2.** Suppose that \( 0 < \eta < \eta' \). Let \( z = z(\eta; z^-) \) and \( z' = z(\eta'; z^-) \) be the solution of (27) with stepsize \( \eta \) and \( \eta' \), respectively. Then we have
\[
\| z' - z^- \| \leq \frac{L_\Phi^{3/2} \eta'}{\eta} \| z - z^- \| + \left( 1 + \frac{\sqrt{L_\Phi \eta'}}{\eta} \right) \| v^- \|_*.
\] (121)

**Proof.** First note that the solution of (27) can be written as \( z = (\nabla \Phi + \eta A)^{-1}(\nabla \Phi(z^-) - v^-) \), where \( A = P + H \) is a maximal monotone operator. Hence, by Lemma D.1 we have
\[
\| \nabla \Phi(z') - \nabla \Phi(z^-) + v^- \|_* \leq \frac{\eta'}{\eta} \sqrt{L_\Phi} \| \nabla \Phi(z) - \nabla \Phi(z^-) + v^- \|_*.
\] (122)

Furthermore, we can bound
\[
\| z' - z^- \| \leq \| \nabla \Phi(z') - \nabla \Phi(z^-) \|_*
\]
\[
\leq \| \nabla \Phi(z') - \nabla \Phi(z^-) + v^- \|_* + \| v^- \|_*
\]
\[
\leq \frac{\eta'}{\eta} \sqrt{L_\Phi} \| \nabla \Phi(z) - \nabla \Phi(z^-) + v^- \|_* + \| v^- \|_*
\]
\[
\leq \frac{\eta'}{\eta} \sqrt{L_\Phi} \| \nabla \Phi(z) - \nabla \Phi(z^-) \|_* + \left( 1 + \frac{\eta'}{\eta} \sqrt{L_\Phi} \right) \| v^- \|_*
\]
\[
\leq \frac{\eta'}{\eta} L_\Phi^{3/2} \| z - z^- \| + \left( 1 + \frac{\eta'}{\eta} \sqrt{L_\Phi} \right) \| v^- \|_*.
\] (127)

where in (123) we used the fact that \( \Phi \) is \( 1 \)-strongly convex w.r.t. \( \| \cdot \| \), in (124) and (126) we used the triangle inequality, (125) follows from (122) and (127) follows from the fact that \( \Phi \) is \( L_\Phi \)-smooth. This completes the proof. \( \square \)

Now we are ready to prove Lemma 6.1. To simplify the notation, we drop the subscript \( k \) and denote \( z_k \) by \( z^- \). Since the stepsize \( \eta \) is \( \beta \)-optimal for condition (52), by definition, there exists a stepsize \( \eta' \) such that \( \eta < \eta' \leq \eta/\beta \) and \( (\eta', z') \) violates (52) with \( z' \) computed via (51). Hence, we have
\[
\frac{1}{2} \alpha \| z' - z^- \| < \eta' \| F(z') - T^{(1)}(z'; z^-) \|_* \leq \frac{\eta' L_2}{2} \| z' - z^- \|^2,
\]
where we used (47) in the last inequality. This implies that \( \eta' > \frac{\alpha \beta}{L_2 \| z' - z^- \|} \), and we further have \( \eta \geq \beta \eta' > \frac{\alpha \beta}{L_2 \| z' - z^- \|} \). Together with (121) and the fact that \( \eta' / \eta \leq 1 / \beta \), this leads to Lemma 6.1.
E Proofs for the Higher-Order Optimistic Method

E.1 Proof of Lemma 7.1

Proof of Part (a). By using the triangle inequality and (6), we have
\[
\|F'(z') - T^{(p-1)}_\lambda(z'; z)\|_s \leq \|F'(z') - T^{(p-1)}(z'; z)\|_s + \frac{\lambda}{(p - 1)!} (2D_\Phi(z', z))^{\frac{p-1}{2}} \|\nabla \Phi(z') - \nabla \Phi(z)\|_s
\]
\[
\leq \frac{L_p}{p!} \|z' - z\|^p + \frac{\lambda}{(p - 1)!} (L_\Phi \|z' - z\|^2)^{\frac{p-1}{2}} L_\Phi \|z' - z\|
\]
\[
= \frac{L_p + p\lambda L_\Phi^{\frac{p-1}{2}}}{p!} \|z' - z\|^p.
\]
where we used the fact that $\Phi(z)$ is $L_\Phi$-smooth in (128). This proves Part (a).

Proof of Part (b). It is evident that $T^{(p-1)}_\lambda(\cdot; z)$ is single-valued and continuous with full domain. Hence, it suffices to show that it is monotone as the maximality follows from Proposition 2.1(b). By definition, we only need to verify that
\[
\langle T^{(p-1)}_\lambda(z_2; z) - T^{(p-1)}_\lambda(z_1; z), z_2 - z_1 \rangle \geq 0
\]
for any $z_1, z_2 \in \mathbb{R}^{m+n}$.

To simplify the notations, denote $R(z'; z) = (2D_\Phi(z', z))^{\frac{p-1}{2}} (\nabla \Phi(z') - \nabla \Phi(z))$. We first prove that
\[
\langle R(z_2; z) - R(z_1; z), z_2 - z_1 \rangle \geq \frac{1}{2} \left( \|z_1 - z\|^{p-1} + \|z_2 - z\|^{p-1} \right) \|z_2 - z_1\|^2.
\]
By the three-point identity in (4), we can write
\[
D_\Phi(z_2, z) = D_\Phi(z_1, z) + \langle \nabla \Phi(z_1) - \nabla \Phi(z), z_2 - z_1 \rangle + D_\Phi(z_2, z_1).
\]
Furthermore, by Bernoulli’s inequality that $(1 + x)^n \geq 1 + nx$ for all $x \geq -1$ and $n \geq 1$, we have
\[
(2D_\Phi(z_2, z))^{\frac{p+1}{2}} = (2D_\Phi(z_1, z))^{\frac{p+1}{2}} \left( 1 + \frac{\langle \nabla \Phi(z_1) - \nabla \Phi(z), z_2 - z_1 \rangle + D_\Phi(z_2, z_1)}{D_\Phi(z_1, z)} \right)^{\frac{p+1}{2}}
\]
\[
\geq (2D_\Phi(z_1, z))^{\frac{p+1}{2}} \left( 1 + \frac{p + 1}{2} \langle \nabla \Phi(z_1), z_2 - z_1 \rangle + D_\Phi(z_2, z_1) \right)
\]
\[
= (2D_\Phi(z_1, z))^{\frac{p+1}{2}} + (p + 1) \langle R(z_1; z), z_2 - z_1 \rangle + (p + 1)(2D_\Phi(z_1, z))^{\frac{p+1}{2}} D_\Phi(z_2, z_1)
\]
\[
\geq (2D_\Phi(z_1, z))^{\frac{p+1}{2}} + (p + 1) \langle R(z_1; z), z_2 - z_1 \rangle + \frac{p + 1}{2} \|z_1 - z\|^{p-1} \|z_2 - z_1\|^2.
\]
\[
(131)
\]
Similarly, we have
\[
(2D_\Phi(z_1, z))^{\frac{p+1}{2}} \geq (2D_\Phi(z_2, z))^{\frac{p+1}{2}} + (p + 1) \langle R(z_2; z), z_1 - z_2 \rangle + \frac{p + 1}{2} \|z_2 - z\|^{p-1} \|z_1 - z_2\|^2.
\]
Adding (131) and (132) together yields (130).
Next, we will prove that
\[
\langle T^{(p-1)}(z_2; z) - T^{(p-1)}(z_1; z), z_2 - z_1 \rangle \geq - \frac{L_p}{2(p-1)!} \left( \|z_2 - z\|^{p-1} + \|z_1 - z\|^{p-1} - \|z_2 - z_1\|^2 \right). \tag{133}
\]
Indeed, by the fundamental theorem of calculus, we can write
\[
T^{(p-1)}(z_2; z) - T^{(p-1)}(z_1; z) = \int_0^1 DT^{(p-1)}(z_1 + t(z_2 - z_1); z)(z_2 - z_1) dt,
\]
and hence
\[
\langle T^{(p-1)}(z_2; z) - T^{(p-1)}(z_1; z), z_2 - z_1 \rangle = \int_0^1 \langle DT^{(p-1)}(z_1 + t(z_2 - z_1); z)(z_2 - z_1), z_2 - z_1 \rangle dt. \tag{134}
\]
Furthermore, for any \(w \in \mathbb{R}^d\) we have
\[
\langle DT^{(p-1)}(z'; z)w, w \rangle = \langle DF(z')w, w \rangle + \langle (DT^{(p-1)}(z'; z) - DF(z'))w, w \rangle \\
\geq - \|DT^{(p-1)}(z'; z) - DF(z')\|_{op}\|w\|^2 \tag{135}
\]
\[
\geq - \frac{L_p}{(p-1)!}\|z' - z\|^{p-1}\|w\|^2, \tag{136}
\]
where in (135) we used the fact that \(F\) is monotone (cf. Proposition 2.1(a)) and in (136) we used (7). Hence, (134) implies that
\[
\langle T^{(p-1)}(z_2; z) - T^{(p-1)}(z_1; z), z_2 - z_1 \rangle \geq - \frac{L_p\|z_2 - z_1\|^2}{(p-1)!} \int_0^1 \|z_1 + t(z_2 - z_1) - z\|^{p-1} dt \\
\geq - \frac{L_p\|z_2 - z_1\|^2}{(p-1)!} \int_0^1 \left( t\|z_2 - z\|^{p-1} + (1 - t)\|z_1 - z\|^{p-1} \right) dt \\
= - \frac{L_p\|z_2 - z_1\|^2}{2(p-1)!} \left( \|z_2 - z\|^{p-1} + \|z_1 - z\|^{p-1} \right), \tag{137}
\]
where we used the convexity of the function \(\|\cdot\|^{p-1}\) in (137). Since \(T^{(p-1)}_\lambda(z'; z) = T^{(p-1)}(z'; z) + \frac{\lambda}{(p-1)!} R(z'; z)\), combining (130) and (133) we conclude that (129) is satisfied when \(\lambda \geq L_p\). This proves Part (b).

E.2 Intermediate Results

To establish the convergence properties of the proposed \(p\)-th-order optimistic method, we first state a few intermediate results that will be used in the following sections.

At the \(k\)-th iteration of our \(p\)-th-order optimistic method, our goal is to find a pair \((\eta, z)\) such that
\[
\eta\|F(z) - T^{(p-1)}(z; z_k)\|_* \leq \frac{\alpha}{2}\|z - z_k\|, \tag{138}
\]
where \(z\) is computed via (80). The following lemma is the higher-order counterpart of Lemma 6.1 and can be proved in a similar way. In particular, we will reuse Lemma D.2 from Appendix D.1.
Lemma E.1. Suppose that $F$ is $p$-th-order $L_p$-smooth and $\Phi$ is $L_\alpha$-smooth. If the stepsize $\eta_k$ at the $k$-th iteration is $\beta$-optimal for condition (138), then

$$\eta_k \geq \frac{p!\alpha \beta^p}{2L_p(\Phi, \lambda) \left( L_{3/2}^\beta \|z_{k+1} - z_k\| + (\beta + L_\Phi^{1/2}) \|v_k\| \right)^{p-1}},$$

where $v_k = \hat{\eta}_k (F(z_k) - T^{(p-1)}(z_k; z_{k-1})).$

Proof. To simplify the notation, we drop the subscript $k$ and denote $z_k$ by $z^-$. By definition, there exists a stepsize $\eta'$ such that $\eta > \eta' \geq \eta / \beta$ and $(\eta', z')$ violates (138) with $z'$ computed via (80). Hence, we have

$$\frac{1}{2} \alpha \|z' - z^-\| < \eta' \|F(z') - T^{(p-1)}(z'; z_k)\| \leq \frac{\eta' L_p(\Phi, \lambda)}{p!} \|z' - z^-\|^p,$$

where we used Lemma 7.1(a) in the last inequality. This gives us

$$\eta \geq \frac{\beta \eta'}{2L_p(\Phi, \lambda) \|z(\eta'; z^-) - z^-\|^{p-1}}.$$  

(139)

Finally, combining (139) and Lemma D.2 leads to Lemma E.1. \hfill \Box

Lemma E.2. Let $\{\eta_k\}_{k \geq 0}$ be the stepsize in (77) generated by Algorithm 2 with advancing. If $\text{res}(x_N) > \epsilon$, then in the convex-concave setting we have

$$\sum_{k=0}^{N-1} \eta_k \frac{2}{p-1} \leq \gamma_p^2(L_p(\Phi, \lambda)) \frac{2}{p-1} D_\Phi(z^*, z_0),$$

and in the strongly-convex-strongly-concave setting we have

$$\sum_{k=0}^{N-1} \left( \eta_k \frac{2}{p-1} \prod_{l=0}^{k-1} (1 + \eta_l \mu) \right) \leq \gamma_p^2(L_p(\Phi, \lambda)) \frac{2}{p-1} D_\Phi(z^*, z_0),$$

(140)

where $\gamma_p$ is defined in (81).

Proof. Note that if $\text{res}(x_N) > \epsilon$, then all the stepsizes $\eta_0, \eta_1, \ldots, \eta_{N-1}$ returned by Algorithm 2 with advancing are $\beta$-optimal. Similar to the proof of Lemma 4.2, we will prove both results in a united way by regarding the convex-concave setting as a special case of the strongly-convex-strongly-concave setting with $\mu = 0$. By the choice of $\hat{\eta}_k$ and (78), we have

$$\|v_k\| \leq \frac{\eta_{k-1}}{1 + \eta_{k-1} \mu} \|F(z_k) - T(z_k; z_{k-1})\| \leq \frac{\alpha}{2(1 + \eta_{k-1} \mu)} \|z_k - z_{k-1}\|.$$

Together with Lemma E.1, this leads to

$$\eta_k \frac{1}{p-1} \leq \left( \frac{2L_p(\Phi, \lambda)}{p! \alpha \beta^p} \right)^{\frac{1}{p-1}} \left( L_{3/2}^\beta \|z_{k+1} - z_k\| + (\beta + L_\Phi^{1/2}) \|v_k\| \right)^{p-1},$$

$$\leq c_1(L_p(\Phi, \lambda)) \frac{1}{p-1} \|z_{k+1} - z_k\| + c_2(L_p(\Phi, \lambda)) \frac{1}{p-1} \|z_k - z_{k-1}\|,$$
where we let $c_1 := L_\Phi^{3/2}(2/p+l/\alpha^p)^{1/p}$ and $c_2 := L_\Phi^{3/2}(\beta + L_\Phi^{1/2})\alpha c_1/2$ to simplify the notation. Moreover, by using Young’s inequality we obtain
\[
\eta_k \frac{2^{p-1}}{p} \leq (c_1^2 + c_1 c_2)(L_p(\Phi, \lambda))^{\frac{2}{p-1}} \|z_{k+1} - z_k\|^2 + \frac{(c_2^2 + c_1 c_2)(L_p(\Phi, \lambda))^{\frac{2}{p-1}}}{(1 + \eta_{k-1} \mu)^2} \|z_k - z_{k-1}\|^2.
\]

Multiplying both sides of the above inequality by $\prod_{l=0}^{k-1}(1 + \eta \mu)$, we further have
\[
\eta_k \frac{2^{p-1}}{p} \prod_{l=0}^{k-1}(1 + \eta \mu) \leq (c_1^2 + c_1 c_2)(L_p(\Phi, \lambda))^{\frac{2}{p-1}} \|z_{k+1} - z_k\|^2 \prod_{l=0}^{k-1}(1 + \eta \mu)
\]
\[+ (c_2^2 + c_1 c_2)(L_p(\Phi, \lambda))^{\frac{2}{p-1}} \|z_k - z_{k-1}\|^2 \prod_{l=0}^{k-2}(1 + \eta \mu).
\]

By summing both sides of (141) over $k = 0, 1, \ldots, N - 1$ and applying Lemma 4.2, we get the desired result in Lemma E.2.

\section{Proof of Theorem 7.2}

If $\text{res}(z_N) \leq \epsilon$, then we are done. Otherwise, we can apply Lemma E.2 to give a bound on $\sum_{k=0}^{N-1} \eta_k$. By Hölder’s inequality, we have
\[
\left( \sum_{k=0}^{N-1} \eta_k \right)^{\frac{2}{p+1}} \left( \sum_{k=0}^{N-1} \eta_k \frac{2}{p} \right)^{\frac{p-1}{p+1}} \geq N.
\]

Together with Lemma E.2, this further implies
\[
\sum_{k=0}^{N-1} \eta_k \geq \left( \sum_{k=0}^{N-1} \eta_k \frac{2}{p} \right)^{-\frac{p-1}{2}} N^{\frac{p+1}{2}} \geq \frac{1}{\gamma_p^{\frac{p-1}{2}} L_p(\Phi, \lambda)(D_\Phi(z^*, z_0))^{\frac{2}{p-1}}} N^{\frac{p+1}{2}}.
\]

The rest follows from Proposition 4.1.

\section{Proof of Lemma 7.3}

By the definition of $\zeta_k$ in (57), we have $\eta_k = (\zeta_k / \zeta_{k+1} - 1)/\mu$. Since $\text{res}(x_N) > \epsilon$, we can apply Lemma E.2 and rewrite the bound in (140) in terms of $\{\zeta_k\}_{k=0}^{N-1}$ as
\[
\sum_{k=0}^{N-1} \left( \frac{\mu \zeta_{k+1}}{\zeta_k - \zeta_{k+1}} \right)^{\frac{2}{p-1}} \frac{1}{\zeta_k} \leq \gamma_p^2(L_p(\Phi, \lambda))^{\frac{2}{p-1}} D_\Phi(z^*, z_0)
\]
\[\Leftrightarrow \sum_{k=0}^{N-1} \left( \frac{1}{1/\zeta_{k+1} - 1/\zeta_k} \right)^{\frac{2}{p-1}} \zeta_k^{-\frac{p+1}{p-1}} \leq \gamma_p^2(L_p(\Phi, \lambda))^{\frac{2}{p-1}} D_\Phi(z^*, z_0) = [\tilde{\kappa}_p(z_0)]^{\frac{2}{p-1}}.
\]

Since each summand in (142) is nonnegative, it follows that for any $k_2 > k_1 \geq 0$
\[
\sum_{k=k_1}^{k_2-1} \left( \frac{1}{1/\zeta_{k+1} - 1/\zeta_k} \right)^{\frac{2}{p-1}} \zeta_k^{-\frac{p+1}{p-1}} \leq [\tilde{\kappa}_p(z_0)]^{\frac{2}{p-1}}.
\]
Furthermore, by applying Hölder’s inequality we get
\[
\left[\sum_{k=k_1}^{k_2-1} \left(\frac{1}{\zeta_{k+1}} - \frac{1}{\zeta_k}\right)\right]^{\frac{p+1}{p}} \left[\sum_{k=k_1}^{k_2-1} \left(\frac{1}{1/\zeta_{k+1} - 1/\zeta_k}\right)^{\frac{p}{p+1}} \zeta_k^{\frac{p}{p+1}}\right] \geq \sum_{k=k_1}^{k_2-1} \frac{1}{\zeta_k},
\]
and Lemma 7.3 follows from the above two inequalities.

E.5 Proof of Theorem 7.4

Without loss of generality, we can assume that \(\frac{(2-\alpha)\epsilon}{2D_\Phi(z^*,z_0)} \leq 1\); otherwise the result becomes trivial as \(\zeta_0 = 1\). It is easy to see that \(\{\zeta_k\}_{k \geq 0}\) is non-increasing in \(k\). Hence, from Lemma 7.3 we get
\[
\frac{1}{\zeta_2} \geq \frac{1}{\zeta_1} + \frac{1}{\tilde{\kappa}_p(z_0)} \left(\sum_{k=1}^{k_2-1} \frac{1}{\zeta_k}\right)^{\frac{p+1}{p}} \geq \frac{1}{\zeta_1} + \frac{1}{\tilde{\kappa}_p(z_0)} \left(\sum_{k=1}^{k_2-1} \frac{1}{\zeta_k}\right)^{\frac{p+1}{p}}.
\]
In particular, this implies that \(\zeta_{k_2} \leq \frac{1}{2} \zeta_{k_1}\) when we have \(k_2 - k_1 \geq [\tilde{\kappa}_p(z_0)]^{\frac{2}{p+1}}\). Hence, for any integer \(l \geq 0\), we can prove by induction that the number of iterations \(N\) to achieve \(\zeta_N \geq 2^{-l}\) does not exceed
\[
\sum_{k=0}^{l-1} \left[\tilde{\kappa}_p(z_0)\right]^{\frac{p+1}{p}} 2^{-\frac{2}{p+1} k} \leq \sum_{k=0}^{l-1} \left[\tilde{\kappa}_p(z_0)\right]^{\frac{p+1}{p}} 2^{-\frac{2}{p+1} k} + 1 \leq \frac{1}{1 - 2^{-\frac{2}{p+1}}} \left[\tilde{\kappa}_p(z_0)\right]^{\frac{2}{p+1}} + l.
\]
The bound in (83) immediately follows by setting \(l = \left\lfloor \log_2 \left(\frac{2D_\Phi(z^*,z_0)}{(2-\alpha)\epsilon}\right)\right\rfloor \leq \log_2 \left(\frac{2D_\Phi(z^*,z_0)}{(2-\alpha)\epsilon}\right) + 1.

E.6 Proof of Theorem 7.5

If \(\text{res}(z_{k+1}) \leq \epsilon\), then we are done. Otherwise, by letting \(k_1 = k\) and \(k_2 = k + 1\) in Lemma 7.3, we obtain
\[
\frac{1}{\zeta_{k+1}} \geq \frac{1}{\zeta_k} + \frac{1}{\tilde{\kappa}_p(z_0)} \zeta_k^{\frac{p+1}{p}} \geq \frac{1}{\tilde{\kappa}_p(z_0)} \zeta_k^{\frac{p+1}{p}},
\]
which immediately leads to (84).

E.7 Proof of Corollary 7.6

To begin with, recall that we have \(D_\Phi(z^*,z) \leq \epsilon\) if \(\zeta_k \leq \frac{(2-\alpha)\epsilon}{2D_\Phi(z^*,z_0)}\) by (58). Hence, it suffices to upper bound the number of iterations required such that the latter condition holds. Also, we only need to prove the case where \(\epsilon < \frac{\mu^2}{(2-\alpha)\gamma_2 L_2^2}\), as the other case directly follows from Theorem 7.4.

Let \(N_1\) be the smallest integer such that \(\zeta_{N_1} \leq 1/(2[\tilde{\kappa}_p(z_0)]^{p-1})\). By setting \(\epsilon = \frac{1}{2}\left\lfloor \frac{\mu^2}{L_\Phi(\Phi,\lambda)} \right\rfloor^{\frac{2}{p-1}}\) in Theorem 7.4, we obtain that
\[
N_1 \leq \max \left\{ \frac{1}{1 - 2^{-\frac{2}{p+1}}} \left[\tilde{\kappa}_p(z_0)\right]^{\frac{p+1}{p}} + \log_2 \left(\frac{2\gamma_2 L_\Phi(\Phi,\lambda) \left[\tilde{\kappa}_p(z_0)\right]^{\frac{2}{p-1}} D_\Phi(z^*,z_0)}{\mu^{\frac{2}{p-1}}}\right) + 1, 1 \right\}
= \max \left\{ \frac{1}{1 - 2^{-\frac{2}{p+1}}} \left[\tilde{\kappa}_p(z_0)\right]^{\frac{p+1}{p}} + \frac{2}{p-1} \log_2 \tilde{\kappa}_p(z_0) + 2, 1 \right\}.
\]
(143)
Furthermore, we can rewrite (84) in Theorem 7.5 as \((\bar{k}_p(z_0))^{\frac{2}{p+1}} \zeta_{k+1} \leq \left((\bar{k}_p(z_0))^{\frac{2}{p+1}} \zeta_k \right)^{\frac{p+1}{2}}\). By induction, we can prove that \((\bar{k}_p(z_0))^{\frac{2}{p+1}} \zeta_k \leq 2^{-\left(\frac{1}{2^{k-N_1}}\right)}\) for all \(k \geq N_1\). Therefore, the number of iterations in total to reach \(\zeta_k \leq \frac{2^{14.5}}{2D_0(z^*, z_0)}\) does not exceed

\[
\log_{\frac{2}{p+1}} \log_2 \left( \frac{2D_0(z^*, z_0)}{(2 - \alpha) \gamma_0^2 (\bar{k}_p(z_0))^{\frac{2}{p+1}} \epsilon} \right) \leq \log_{\frac{2}{p+1}} \log_2 \left( \frac{2 \mu^{\frac{2}{p+1}}}{(2 - \alpha) \gamma_0^2 (L_0(\Phi, \lambda))^{\frac{2}{p+1}} \epsilon} \right) + 1. \tag{144}
\]

The result in (85) now follows from (143) and (144).

E.8 Proof of Theorem 7.7

Denote \(\eta_{-1} := \beta \sigma_0\). By Proposition 4.6, the total number of calls to the optimistic subsolver after \(N\) iterations can be bounded by

\[
\sum_{k=0}^{N-1} 2 \log_2 \log_\frac{1}{\eta} \left( \max \left\{ \frac{\eta_{k-1}^2}{\beta^4 \eta_k^2}, \frac{n_k}{\eta_{k-1}} \right\} \right) \leq 2N \log_2 \left( \frac{1}{N} \sum_{k=0}^{N-1} \log_\frac{1}{\eta} \left( \max \left\{ \frac{n_{k-1}^2}{\beta^4 n_k^2}, \frac{n_k}{\eta_{k-1}} \right\} \right) \right). \tag{145}
\]

where we used the fact that \(\frac{1}{N} \sum_{k=0}^{N-1} \log a_k \leq \log \left( \frac{1}{N} \sum_{k=0}^{N-1} a_k \right)\). Similar to the proof of Theorem 6.9, we make a simple observation that for any \(k = 0, \ldots, N-1,\)

\[
\log_\frac{1}{\eta} \left( \max \left\{ \frac{n_{k-1}^2}{\sigma_0^2}, \frac{n_k}{\eta_{k-1}} \right\} \right) \leq \max \left\{ 4 + \log_\frac{1}{\eta} \left( \frac{\eta_{k-1}}{\sigma_0^2} \right) + (p - 1) \log_\frac{1}{\eta} \left( \frac{\eta_k^2}{\sigma_0^2} \right), \log_\frac{1}{\eta} \left( \frac{\eta_k^2}{\sigma_0^2} \right) + (p - 1) \log_\frac{1}{\eta} \left( \frac{\eta_k^2}{\sigma_0^2} \right) \right\}
\]

\[
\leq 4 + \log_\frac{1}{\eta} \left( \frac{\eta_{k-1}}{\sigma_0^2} + 1 \right) + (p - 1) \log_\frac{1}{\eta} \left( \frac{\eta_k}{\sigma_0^2} + 1 \right) + (p - 1) \log_\frac{1}{\eta} \left( \frac{\eta_k}{\sigma_0^2} + 1 \right). \tag{146}
\]

Similarly, using the fact that \(\eta_{-1} = \beta \sigma_0\), for the first summand in (145) we have

\[
\log_\frac{1}{\eta} \left( \max \left\{ \frac{n_{k-1}^2}{\beta^4 \eta_0}, \frac{n_k}{\eta_{-1}} \right\} \right) = \max \left\{ 2 + (p - 1) \log_\frac{1}{\eta} \left( \frac{\sigma_0^2}{\eta_{-1}} \right), 2 + \log_\frac{1}{\eta} \left( \frac{\eta_0^2}{\sigma_0^2} \right) \right\}
\]

\[
\leq 4 + (p - 1) \log_\frac{1}{\eta} \left( \frac{\sigma_0^2}{\eta_{-1}} \right) + \log_\frac{1}{\eta} \left( \frac{\eta_0^2}{\sigma_0^2} + 1 \right). \tag{147}
\]
Hence, we can show that

\[
\sum_{k=0}^{N-1} \log \frac{1}{\beta} \left( \max \left\{ \frac{\eta_{k-1}^2}{\beta^4 \eta_k^2}, \frac{\eta_k^2}{\eta_{k-1}^2} \right\} \right)
\leq 4N + 2 \sum_{k=0}^{N-1} \left( (p-1) \log \frac{1}{\beta} \left( \frac{\sigma_0^2}{\eta_k^2} + 1 \right) + \log \frac{1}{\beta} \left( \frac{\eta_k^2}{\sigma_0^2} + 1 \right) \right)
\leq 4N + 2(p-1)N \log \frac{1}{\beta} \left( 1 + \frac{\sigma_0^2}{\sum_{k=0}^{N-1} \eta_k^{p-1}} \right) + 2N \log \frac{1}{\beta} \left( 1 + \frac{1}{\sigma_0^2 N \sum_{k=0}^{N-1} \eta_k^2} \right).
\]

where the first inequality simply follows from (146) and (147), and the second inequality holds since

\[
\frac{1}{N} \sum_{k=0}^{N-1} \log a_k \leq \log \left( \frac{1}{N} \sum_{k=0}^{N-1} a_k \right).
\]

Now we only need to establish bounds on \( \sum_{k=0}^{N-1} \eta_k^{p-1} \) and \( \sum_{k=0}^{N-1} \eta_k^2 \). Since the term \((1 + \eta_k \mu)\) is always larger than 1, by Lemma E.2 we have

\[
\sum_{k=0}^{N-1} \eta_k^{p-1} \leq \gamma_p^2 (L_p(\Phi, \lambda)) \gamma_p^2 D_\Phi(z^*, z_0)
\]

for both convex-concave and strongly-convex-strongly-concave cases. Also, Lemma 6.8 provides an upper bound on \( \sum_{k=0}^{N-1} \eta_k^2 \). Applying these bounds in (72) implies that

\[
\sum_{k=0}^{N-1} \log \frac{1}{\beta} \left( \max \left\{ \frac{\eta_{k-1}^2}{\beta^4 \eta_k^2}, \frac{\eta_k^2}{\eta_{k-1}^2} \right\} \right) \leq 4N + 2(p-1)N \log \frac{1}{\beta} \left( 1 + \frac{\sigma_0^2}{\gamma_p^2 \gamma_p^2 (L_p(\Phi, \lambda)) \gamma_p^2 D_\Phi(z^*, z_0)} \right)
\]

\[+ 2N \log \frac{1}{\beta} \left( 1 + \frac{2(\alpha + L_\Phi)^2 D_\Phi(z^*, z_0)}{\sigma_0^2 (1 - \alpha) N \eta_0^2} \right). \tag{148}
\]

Now the claim follows by applying the upper bound in (148) in the right-hand side of (145).