Accelerated Minimax Algorithms Flock Together

TaeHo Yoon  
Department of Mathematical Sciences  
Seoul National University  
tetrzim@snu.ac.kr

Ernest K. Ryu  
Department of Mathematical Sciences  
Seoul National University  
ernestyu@snu.ac.kr

Abstract

Several new accelerated methods in minimax optimization and fixed-point iterations have recently been discovered, and, interestingly, they rely on a mechanism distinct from Nesterov’s momentum-based acceleration. In this work, we show that these accelerated algorithms exhibit what we call the merging path (MP) property; the trajectories of these algorithms merge quickly. Using this novel MP property, we establish point convergence of existing accelerated minimax algorithms and derive new state-of-the-art algorithms for the strongly-convex-strongly-concave setup and for the prox-grad setup.

1 Introduction

Minimax optimization problems of the form

\[
\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} L(x, y) \tag{1}
\]

have recently received increased attention in many machine learning applications including robust optimization [86], adversarial training [55], fair machine learning [29], and GANs [32]. Consequently, there has been a large body of research on efficient algorithms for solving minimax optimization. For the deterministic smooth convex-concave setup, the classical extragradient (EG) [47] and the optimistic gradient descent (OG) [78, 80, 21] methods converge with $O(1/k)$ rate on the squared gradient norm [85, 33], and accelerated algorithms [100, 49, 92] with $O(1/k^2)$ rate, based on an anchoring mechanism retracting iterates toward the initial point, were recently discovered.

In a different line of work on the fixed-point problem

\[
\text{find } z \in \mathbb{R}^d \quad z = T(z), \tag{2}
\]

the Halpern iteration [35], which averages the iterates with the initial point, was shown to achieve an accelerated $O(1/k^2)$ rate on the squared fixed-point residual norm when executed with optimal parameters [51, 43]. Interestingly, both the Halpern iteration and the accelerated minimax algorithms rely on the anchoring mechanism, but a precise theoretical explanation of such apparent similarity was, to the best of our knowledge, yet missing.

Contribution. In this work, we identify the merging path (MP) property among anchoring-based algorithms, which state that the trajectories of these algorithms quickly merge, and use this novel property to develop new accelerated algorithms. First, we show that known accelerated minimax algorithms [100, 49, 92] have $O(1/k^2)$-merging paths with the optimized Halpern iteration [35, 51] and thereby identify their point convergence. Second, we present a new accelerated minimax algorithm for smooth strongly-convex-strongly-concave setup designed to approximate, in the MP sense, the OC-Halpern method, which has an optimal accelerated rate. Third, we present a near-optimal accelerated proximal gradient type algorithm, designed to approximate, in the MP sense, the Halpern-accelerated Douglas-Rachford splitting [27, 53].

Preprint. Under review.
Figure 1: Trajectories of (left) Nesterov’s accelerated algorithms (AGM [69, 8, 11]; described in Appendix A.2) with distinct momentum parameters and (right) anchoring-based algorithms. On the right, paths quickly merge and become indistinguishable in 50 iterations. All algorithms are executed on the convex minimization problem with $f(x_1, x_2) = \frac{4x_2^2}{x_2}$, starting at $(x_1, x_2) = (-2, 3)$.

1.1 The merging path (MP) property

Let $\mathfrak{A}$ be a deterministic algorithm. Write $\mathfrak{A}(x_0; \mathcal{P}) = (x_0, x_1, x_2, \ldots)$ to denote that $\mathfrak{A}$ applied to problem $\mathcal{P}$ with starting point $x_0$ produces iterates $x_1, x_2, \ldots$. We say two algorithms $\mathfrak{A}_1$ and $\mathfrak{A}_2$ have $O(r(k))$-merging paths if for any problem $\mathcal{P}$ and $x_0 \in \mathbb{R}^d$, the iterates $\mathfrak{A}_\ell(x_0; \mathcal{P}) = (x_0, x_1^{(\ell)}, x_2^{(\ell)}, \ldots)$ for $\ell = 1, 2$ satisfy $\|x_k^{(1)} - x_k^{(2)}\|^2 = O(r(k))$. More concisely, we say $\mathfrak{A}_1$ and $\mathfrak{A}_2$ are $O(r(k))$-MP if they have $O(r(\ell))$-merging paths.

The MP property precisely formalizes a notion of near-equivalence of algorithms. In particular, it is a stronger notion of approximation compared to the conceptual ones of prior work [63, 1]. Classical first-order algorithms for smooth convex minimization are not $O(r(k))$-MP with $r(k) \to 0$. Figure 1 shows that accelerated gradient methods with different momentum coefficients have divergent paths, even though their function values (of course) converge to the same optimal value. In this paper, we show that the accelerated minimax algorithms for smooth convex-concave minimax optimization are $O(1/k^2)$-MP (which is faster than the $o(1)$-point convergence). Figure 1b shows that the algorithms’ paths indeed merge far before they converge to the limit.

1.2 Related work

An overview of convex-concave minimax optimization. The extragradient (EG) [47] and Popov’s algorithm [78], also known as optimistic gradient (OG) descent [80, 21], have initiated the long stream of research in the optimization literature on the closely interconnected topics of minimax optimization [68, 38, 46, 61, 63, 64, 5, 36], variational inequalities [74, 71, 73, 42, 9, 10, 59, 56, 16, 40, 57], and monotone inclusion problems [87, 93, 65, 66, 54]. A number of advances in minimax optimization have been made in the context of practical machine learning applications, including the multi-player game dynamics [20, 88, 31, 39] and GANs [97, 30, 14, 50, 61, 85, 77]. A recent trend of the field focuses on achieving near-optimal efficiency with respect to primal-dual gap or distance to the optimum for convex-concave minimax problems with distinct strong convexity, strong concavity and smoothness parameters for each variable [91, 52, 95, 2, 99, 41].

Proximal point methods and proximal approximation approaches. The proximal point method (PPM) [67, 60, 83] is an implicit algorithm for solving convex minimization, minimax optimization, and, more generally, monotone inclusion problems. Although PPM is not always directly applicable, it has been used as the conceptual basis for developing many algorithms with approximate proximal evaluations [82, 34, 87, 68, 37, 13, 28, 23, 79]. Several prior works [63, 1] interpreted prominent forward algorithms such as Nesterov’s accelerated gradient method (AGM), EG, and OG as conceptual approximations of PPM.

Acceleration with respect to gradient norm. Since the seminal work of Nesterov [69], acceleration of first-order algorithms has been extensively studied for convex minimization problems
Composite minimax problems and prox-grad-type algorithms. The particular case of bilinearly-coupled minimax problems, \( \mathbf{L}(x, y) = f(x) + (x, Ay) - g(y) \) for some convex functions \( f \) and \( g \) and a linear operator \( A \), have been studied by many researchers. A line of works [70, 96, 90] considered the setups where \( f \) and \( g \) are smooth and one can only access their gradients. Another line of works, more closely connected to the setup we study in Section 5, considered the cases when \( f \) or \( g \) is proximable, and used primal-dual splitting algorithms to solve those problems [12, 18, 94, 15, 38, 46, 98, 84]. Primal-dual problem setups can often be reformulated into a more flexible format of composite monotone inclusion problems and addressed via forward-backward splitting algorithms [93, 17, 18, 56, 19, 58], which also extend the prox-grad-type methods including the iterative shrinkage-thresholding algorithms (ISTA) [22, 8]. Our result of Section 5 can be understood as acceleration of forward-backward algorithms with respect to the forward-backward residual, which generalizes the gradient magnitude. For the special case of the projected-gradient setup, the prior work [24] has also achieved a near-optimal acceleration.

### 2 Background and preliminaries

#### Convex-concave functions and minimax optimization

A function \( \mathbf{L}(\cdot, \cdot) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \) is convex-concave if \( \mathcal{X} \) and \( \mathcal{Y} \) are convex sets, \( \mathbf{L}(x, y) \) is convex as a function of \( x \) for all \( y \), and \( \mathbf{L}(x, y) \) is concave as a function of \( y \) for all \( x \). A point \((x_\star, y_\star) \in \mathcal{X} \times \mathcal{Y}\) is a saddle point, or (minimax) solution of (1) if \( \mathbf{L}(x_\star, y_\star) \leq \mathbf{L}(x, y_\star) \leq \mathbf{L}(x_\star, y) \) for all \((x, y) \in \mathcal{X} \times \mathcal{Y}\).

#### Operators and monotonicity

An operator \( \mathbf{A} \) on \( \mathbb{R}^d \), denoted \( \mathbf{A} : \mathbb{R}^d \rightrightarrows \mathbb{R}^d \), is a set-valued function, i.e., \( \mathbf{A}(z) \subseteq \mathbb{R}^d \) for all \( z \in \mathbb{R}^d \). For simplicity, we often write \( Az = \mathbf{A}(z) \). If \( Az \) contains exactly one element for all \( z \in \mathbb{R}^d \), then we say \( \mathbf{A} \) is single-valued and view it as a function. An operator \( \mathbf{A} : \mathbb{R}^d \rightrightarrows \mathbb{R}^d \) is monotone if

\[
(\mathbf{Az} - \mathbf{A}z' , z - z') \geq 0, \quad \forall z, z' \in \mathbb{R}^d,
\]

where the notation means that \((u,v,z-z') \geq 0\) for any \( u \in Az \) and \( v \in A\mathbf{z'} \). For \( \mu > 0 \), an operator \( \mathbf{A} \) is \( \mu \)-strongly monotone if

\[
(\mathbf{Az} - \mathbf{A}z' , z - z') \geq \mu \| z - z' \|^2, \quad \forall z, z' \in \mathbb{R}^d.
\]

The graph of \( \mathbf{A} \) is denoted \( \text{Gra} \mathbf{A} = \{ (z, u) : u \in Az \} \). A monotone operator \( \mathbf{A} \) is maximally monotone if there is no monotone operator \( \mathbf{A}' \) such that \( \text{Gra} \mathbf{A} \subset \text{Gra} \mathbf{A}' \) strictly. In particular, if \( \mathbf{A} \) is monotone, single-valued, and continuous as a function, then it is maximally monotone [7, Corollary 20.28]. Define

\[
(\mathbf{A} + \mathbf{B})(z) = \{ u + v \in \mathbb{R}^d | u \in Az, v \in Bz \}, \quad (\alpha \mathbf{A})(z) = \{ \alpha u \in \mathbb{R}^d | u \in Az \}
\]

for operators \( \mathbf{A} \) and \( \mathbf{B} \) and scalar \( \alpha \in \mathbb{R} \). The inverse of an operator \( \mathbf{A} \) is the operator \( \mathbf{A}^{-1} : \mathbb{R}^d \rightrightarrows \mathbb{R}^d \) defined by \( \mathbf{A}^{-1}(u) = \{ z \in \mathbb{R}^d | u \in Az \} \).

#### Monotone inclusion and fixed-point problems

In a monotone inclusion problem,

\[
0 \in \mathbf{A}(z),
\]

where \( \mathbf{A} : \mathbb{R}^d \rightrightarrows \mathbb{R}^d \) is monotone, one finds zeros of an operator \( \mathbf{A} \). For any \( \alpha > 0 \), we have

\[
0 \in \mathbf{A}(z) \iff z \in (I + \alpha \mathbf{A})(z) \iff z \in (I + \alpha \mathbf{A})^{-1}(z).
\]

Therefore, (3) is equivalent to the fixed-point problem (2) with \( \mathbf{T} = (I + \alpha \mathbf{A})^{-1} = J_{\alpha \mathbf{A}} \). We call \( J_{\alpha \mathbf{A}} \) the resolvent of \( \alpha \mathbf{A} \).

We say \( \mathbf{T} : \mathbb{R}^d \to \mathbb{R}^d \) is nonexpansive if it is 1-Lipschitz continuous and contractive if it is \( \rho \)-Lipschitz with \( \rho < 1 \). If \( \mathbf{A} \) is maximally monotone, then \( J_{\alpha \mathbf{A}} \) has full domain [62] and is nonexpansive [83] for any \( \alpha > 0 \). Denoting \( \text{Zer} \mathbf{A} = \{ z \in \mathbb{R}^d | 0 \in \mathbf{A}(z) \} \) and \( \text{Fix} \mathbf{T} = \{ z \in \mathbb{R}^d | z = \mathbf{T}(z) \} \), we have \( \text{Zer} \mathbf{A} = \text{Fix} J_{\alpha \mathbf{A}} \). We call the algorithm \( z_{k+1} = J_{\alpha \mathbf{A}}(z_k) \) the proximal point method (PPM) for \( \mathbf{A} \).
Minimax optimization as monotone inclusion problems. For differentiable and convex-concave \( \mathbf{L} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R} \), define \( \nabla_+ \mathbf{L}(x, y) = (\nabla_y \mathbf{L}(x, y), -\nabla_y \mathbf{L}(x, y)) \) and call it the saddle operator of \( \mathbf{L} \). Then \( \nabla_+ \mathbf{L} \) is monotone [82], and \( z_* = (x_*, y_*) \) is a solution of (1) if and only if \( \nabla_+ \mathbf{L}(z_*) = 0 \). Therefore, a convex-concave minimax optimization problem (1) can be reformulated as a monotone inclusion problem (3) with \( \mathbf{A} = \nabla_+ \mathbf{L} \) and as a fixed-point problem (2) with \( \mathbf{T} = \mathbf{J}_{\nabla_+ \mathbf{L}} \).

The fixed-point residual norm \( \|z - \mathbf{T}z\| \) quantifies the rate of convergence of an algorithm for finding fixed points. When \( \mathbf{T} = \mathbf{J}_{\alpha \mathbf{B}} \), we have \( z - \mathbf{T}z = z - \mathbf{J}_{\alpha \mathbf{B}}(z) \in \alpha \mathbf{B}(\mathbf{J}_{\alpha \mathbf{B}}(z)) \), because

\[
\mathbf{u} = \mathbf{J}_{\alpha \mathbf{B}}(z) = (\mathbf{I} + \alpha \mathbf{B})^{-1}(z) \iff z \in (\mathbf{I} + \alpha \mathbf{B})(\mathbf{u}) = \mathbf{u} + \alpha \mathbf{Bu}.
\]

Thus, when \( \mathbf{T} = \mathbf{J}_{\alpha \mathbf{B}} \) and \( \mathbf{B} \) is single-valued, a convergence rate on \( \|z - \mathbf{T}z\| \) is equivalent to a convergence rate on \( \|\mathbf{B}(\cdot)\| \).

3 Accelerated algorithms for smooth minimax optimization are \( \mathcal{O}(1/k^2) \)-MP

In this section, we prove the \( \mathcal{O}(1/k^2) \)-MP property among accelerated minimax algorithms and the optimized Halpern iteration. The problem setup is: minimax problems with convex-concave \( \mathbf{L} \) such that \( \nabla_+ \mathbf{L} \) is \( L \)-Lipschitz. For notational convenience, write \( \mathbf{B} = \nabla_+ \mathbf{L} \). Our MP result provides new insight into these acceleration mechanisms and allows us to establish their point convergence.

Preliminaries: Accelerated minimax and proximal algorithms. Recently, several accelerated minimax algorithms with rate \( \|\mathbf{B}z_k\|^2 = \mathcal{O}(1/k^2) \) have been proposed. The list includes the extra anchored gradient (EAG) algorithm [100]

\[
z_{k+1/2} = \beta_k z_0 + (1 - \beta_k)z_k - \alpha \mathbf{B}z_k
\]

\[
z_{k+1} = \beta_k z_0 + (1 - \beta_k)z_k - \alpha \mathbf{B}z_{k+1/2},
\]

the fast extragradient (FEG) algorithm [49]

\[
z_{k+1/2} = \beta_k z_0 + (1 - \beta_k) (z_k - \alpha \mathbf{B}z_k)
\]

\[
z_{k+1} = \beta_k z_0 + (1 - \beta_k)z_k - \alpha \mathbf{B}z_{k+1/2},
\]

and the anchored Popov’s scheme (APS) [92], with \( v_0 = z_0 \)

\[
v_{k+1} = \beta_k z_0 + (1 - \beta_k)z_k - \alpha \mathbf{B}v_k
\]

\[
z_{k+1} = \beta_k z_0 + (1 - \beta_k)z_k - \alpha \mathbf{B}v_{k+1}.
\]

Prior to the development of these minimax algorithms, the optimized Halpern’s method (OHM) [51, 43] was presented for solving the fixed-point problem (2) with nonexpansive operator \( \mathbf{T} : \mathbb{R}^d \to \mathbb{R}^d \):

\[
w_{k+1/2} = \beta_k w_0 + (1 - \beta_k)w_k
\]

\[
w_{k+1} = \mathbf{T}(w_{k+1/2}),
\]

where \( \beta_k = \frac{1}{k+1} \). OHM converges with rate \( \|w_{k+1/2} - \mathbf{T}w_{k+1/2}\|^2 \leq 4\|w_0 - w_*\|^2/(k + 1)^2 \), if a fixed point \( w_* \) exists.

To clarify, the prior works [100, 92] study further general forms of EAG and APS with iteration-dependent step-sizes. For the sake of simplicity, we consider the case of constant step-sizes and defer the discussion on varying step-sizes to Appendix A and F.

MP property among accelerated algorithms. The accelerated forward algorithms, EAG, FEG, and APS, do not merely resemble OHM in their form. The algorithms are, in fact, near-equivalent in the sense that they have quickly merging paths.

Theorem 1 (EAG≈FEG≈APS≈OHM). Let \( \mathbf{B} : \mathbb{R}^d \to \mathbb{R}^d \) be monotone and \( L \)-Lipschitz and assume \( z_* \in \text{Zer} \mathbf{B} = \text{Fix} \mathbf{J}_{\alpha \mathbf{B}} \) exists. Let \( \mathbf{T} = \mathbf{J}_{\alpha \mathbf{B}} \) and \( \beta_k = \frac{1}{k+1} \). Then FEG and OHM are \( \mathcal{O}(\|z_0 - z_*\|^2/k^2) \)-MP for \( \alpha \in (0, \frac{1}{2}) \), and there exists \( \eta > 0 \) such that EAG, APS, and OHM are \( \mathcal{O}(\|z_0 - z_*\|^2/k^2) \)-MP for \( \alpha \in (0, \frac{\eta}{2}) \).
Proof. We provide the proof for the FEG case and defer the proofs for EAG and APS to Appendix B.2. Let \( \{w_k\}_{k=0,1,...} \) be the OHM iterates and \( \{z_k\}_{k=0,1,...} \) the FEG iterates, and assume \( z_0 = w_0 \). Then

\[
\begin{align*}
  z_{k+1} - w_{k+1} &= (\beta_k z_k + (1 - \beta_k) w_k - \alpha B z_{k+1/2}) - (\beta_k w_0 + (1 - \beta_k) w_k - \alpha B w_{k+1}) \\
  &= (1 - \beta_k) (z_k - w_k) + \alpha (B w_{k+1} - B z_{k+1/2})
\end{align*}
\]

where we used \((I + \alpha B)(w_{k+1}) = w_{k+1/2}\) in the first inequality and \(z_0 = w_0\) in the second. Thus,

\[
\begin{align*}
  \|z_{k+1} - w_{k+1}\|^2 &= (1 - \beta_k)^2 \|z_k - w_k\|^2 + 2 \alpha (1 - \beta_k) (z_k - w_k) \langle z_k - w_k, B w_{k+1} - B z_{k+1/2} \rangle \\
  &\quad + \alpha^2 \|B w_{k+1} - B z_{k+1/2}\|^2.
\end{align*}
\]

(5)

Now use the similar identity

\[
\begin{align*}
  z_{k+1/2} - w_{k+1} &= (1 - \beta_k) (z_k - w_k) - \alpha (1 - \beta_k) B z_k + \alpha B w_{k+1}
\end{align*}
\]

to rewrite the inner product term in (5)

\[
2 \langle \alpha (1 - \beta_k) (z_k - w_k), B w_{k+1} - B z_{k+1/2} \rangle
\]

\[
= 2 \langle \alpha (z_k - w_{k+1}) + \alpha^2 (1 - \beta_k) B z_k - \alpha^2 B w_{k+1}, B w_{k+1} - B z_{k+1/2} \rangle
\]

\[
\leq 2 \alpha^2 \langle (1 - \beta_k) B z_k - B w_{k+1}, B w_{k+1} - B z_{k+1/2} \rangle,
\]

where the last inequality follows from monotonicity of \(B\). Combining this with (5), we obtain

\[
\begin{align*}
  \|z_{k+1} - w_{k+1}\|^2 &\leq (1 - \beta_k)^2 \|z_k - w_k\|^2 + 2 \alpha^2 \langle (1 - \beta_k) B z_k - B w_{k+1}, B w_{k+1} - B z_{k+1/2} \rangle \\
  &\quad + \alpha^2 \|B w_{k+1} - B z_{k+1/2}\|^2
\end{align*}
\]

\[
= (1 - \beta_k)^2 \|z_k - w_k\|^2 - 2 \alpha^2 (1 - \beta_k) \langle B z_k, B z_{k+1/2} \rangle + \alpha^2 \|B z_{k+1/2}\|^2.
\]

Plugging in \(\beta_k = \frac{1}{k+1}\) and multiplying both sides by \((k + 1)^2\), we obtain

\[
(k + 1)^2 \|z_{k+1} - w_{k+1}\|^2 \leq k^2 \|z_k - w_k\|^2 + \alpha^2 \|B z_k - (k + 1) B z_{k+1/2}\|^2.
\]

The conclusion then follows from the following Lemma 2. 

Lemma 2. Let \(B : \mathbb{R}^d \to \mathbb{R}^d\) be monotone and \(L\)-Lipschitz and assume \(z_\ast \in \text{Zer} B\) exists. Let \(\{z_k\}_{k=0,1,...}\) be the iterates of FEG with \(\beta_k = \frac{1}{k+1}\). For \(\alpha \in (0, \frac{1}{L})\),

\[
\sum_{k=0}^{\infty} \|k B z_k - (k + 1) B z_{k+1/2}\|^2 \leq \frac{1}{\alpha^2 (1 - \alpha^2 L^2)} \|z_0 - z_\ast\|^2 < \infty.
\]

Proof outline of Lemma 2. We show \(V_k = \frac{\alpha (1 - \alpha^2 L^2)}{2} \|B z_k\|^2 + k \langle B z_k, z_k - z_\ast \rangle + \frac{1}{2 \alpha^2} \|z_0 - z_\ast\|^2\) satisfy

\[
V_k - V_{k+1} \geq \frac{\alpha (1 - \alpha^2 L^2)}{2} \|k B z_k - (k + 1) B z_{k+1/2}\|^2
\]

and \(V_k \geq 0\) for \(k = 0, 1, \ldots\) (details presented in Appendix B.1). The conclusion follows by taking \(N \to \infty\) in \(\frac{\alpha (1 - \alpha^2 L^2)}{2} \sum_{k=0}^{N} \|k B z_k - (k + 1) B z_{k+1/2}\|^2 \leq V_0 - V_{N+1} \leq \frac{1}{2 \alpha^2} \|z_0 - z_\ast\|^2\). 

As it is well known [7, Theorem 30.1] that OHM converges to

\[
\Pi_{\text{Fix}T}(z_0) = \arg\min_{z \in \text{Fix} T} \|z - z_0\|,
\]

Theorem 1 immediately implies point convergence of EAG, FEG, and APS.

Corollary 3. In the setup of Theorem 1, iterates of EAG, FEG, and APS converge to \(\Pi_{\text{Zer} B}(z_0)\).
4 Fastest algorithm for smooth strongly-convex-strongly-concave minimax optimization with respect to gradient norm

In Section 3, we identified that prior accelerated minimax algorithms approximate, in the MP sense, the optimal proximal algorithm OHM. In this section, we design a novel algorithm to approximate, in the MP sense, the optimal proximal algorithm OC-Halpern [76] and thereby achieve the state-of-the-art rate for minimax problems with $L$-smooth $\mu$-strongly-convex-strongly-concave $L$ (so $\nabla_2 L$ is $L$-Lipschitz and $\mu$-strongly monotone). Later in Section 5, we repeat the strategy of designing an algorithm to approximate, in the MP sense, a known accelerated proximal algorithm.

We design the novel algorithm **Strongly Monotone Extra Anchored Gradient**+\(^1\)

\[
\begin{align*}
z_{k+1/2} & = \beta_k z_0 + (1 - \beta_k) z_k - \eta_k \alpha \mathbf{B} z_k, \\
z_{k+1} & = \beta_k z_0 + (1 - \beta_k) z_k - \alpha \mathbf{B} z_{k+1/2},
\end{align*}
\]

(SM-EAG+)

where $\mathbf{B} = \nabla_\pm L$, $\beta_k = \frac{1}{\sum_{j=0}^{k} (1 + 2\alpha \mu)^j}$, $\eta_k = \frac{1 - \beta_k}{1 + 2\alpha \mu}$, and $0 < \alpha \leq \frac{\sqrt{L^2 + \mu^2 + \mu}}{L^2}$, to approximate OC-Halpern in the MP sense. SM-EAG+ inherits the accelerated rate of OC-Halpern and achieves the fastest known rate in the setup with respect to the gradient norm, improving upon the prior rates of EG and OG by factors of 8 and 4, respectively. (OC-Halpern is not a forward algorithm as it uses the proximal operator $J_{\alpha \mathbf{B}}$, while SM-EAG+ is a forward algorithm using evaluations of $\mathbf{B}$.)

**Preliminaries: Proximal algorithm with exact optimal complexity.** For the fixed-point problem (2) with a $\gamma^{-1}$-contractive $T$, the recently presented **OC-Halpern** [76, Corollary 3.3, Theorem 4.1], which has the same form (4) as OHM but with $\beta_k = \left(\sum_{j=0}^{k} \gamma^{2j}\right)^{-1}$, achieves the exact optimal accelerated rate of $\|w_{k+1/2} - Tw_{k+1/2}\|^2 \leq (1 + \gamma^{-1})\left(\sum_{j=0}^{k} \gamma^{2j}\right)^{-2} \|w_0 - w_*\|^2$, where $w_* \in \text{Fix } \mathbf{B}$. We consider OC-Halpern with $T = J_{\alpha \mathbf{B}}$, which is contractive if $\mathbf{B}$ is maximally monotone and $\mu$-strongly monotone [7, Proposition 23.8, 23.13].

**MP property and convergence analysis of SM-EAG+.** SM-EAG+ approximates OC-Halpern in the MP sense.

**Lemma 4 (SM-EAG+≈OC-Halpern).** Let $\mathbf{B} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ be $\mu$-strongly monotone and $L$-Lipschitz with $0 < \mu \leq L$. Let $z_*$ be the zero of $\mathbf{B}$. If $\alpha \in (0, \frac{\sqrt{L^2 + \mu^2 + \mu}}{L^2})$ and $\epsilon \in (0, 1)$, then for any $\epsilon \in (0, 1)$, SM-EAG+ and OC-Halpern with $T = J_{\alpha \mathbf{B}}$ and $\beta_k = \frac{1}{\sum_{j=0}^{k} (1 + 2\alpha \mu)^j}$ are \(O\left(\frac{\|z_0 - z_*\|^2}{(1 + 2\alpha \mu (1 - \epsilon)^{\epsilon})^2}\right)\)-MP.

To clarify, in Lemma 4, we view $J_{\alpha \mathbf{B}}$ as a $\gamma^{-1}$-contractive operator with $\gamma^{-1} = \frac{1}{\sqrt{1 + 2\alpha \mu}} > \frac{1}{1 + \alpha \mu}$, where $\frac{1}{1 + \alpha \mu}$ is the tight contraction factor for $J_{\alpha \mathbf{B}}$ [7, Proposition 23.13], and then apply OC-Halpern. This slack, which is negligible in the regime $L/\mu \rightarrow \infty$, is necessary for proving the MP result.

SM-EAG+ inherits the convergence rate of OC-Halpern, since the paths of OC-Halpern and SM-EAG+ merge at rate \(O\left(\frac{\|z_0 - z_*\|^2}{(1 + 2\alpha \mu (1 - \epsilon)^{\epsilon})^2}\right)\) for any $\epsilon > 0$, arbitrarily close to the order of convergence $\|\alpha \mathbf{B} w_{k+1}\|^2 = O\left(\frac{\|w_0 - w_*\|^2}{(1 + 2\alpha \mu)^{\epsilon}}\right)$ of OC-Halpern [76]. On the other hand, EG and OG can be viewed as approximations of the proximal point method (PPM) [63], which is slower than OC-Halpern. Additionally, it is unclear whether EG and OG exhibit an MP property to PPM fast enough to preserve PPM’s convergence rate since EG and OG have no anchoring mechanism inducing the MP property.

**Theorem 5** (Fast rate of SM-EAG+). Let $\mathbf{B} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ be $\mu$-strongly monotone and $L$-Lipschitz with $0 < \mu \leq L$. Let $z_*$ be the zero of $\mathbf{B}$. For $\alpha \in \left(0, \frac{\sqrt{L^2 + \mu^2 + \mu}}{L^2}\right]$, then SM-EAG+ exhibits the rate

\[
\|\alpha \mathbf{B} z_k\|^2 \leq \frac{(\sqrt{1 + 2\alpha \mu} + 1)^2}{\alpha^2 \left(\sum_{j=0}^{k-1} (1 + 2\alpha \mu)^{j/2}\right)} \|z_0 - z_*\|^2, \quad \text{for } k = 1, 2, \ldots
\]

\(^1\)The “+” follows the nomenclature of [25], where the two-time-scale EG was named EG+. 

6
We reformulate the minimax optimization problem into the monotone inclusion problem

\[ \text{APG} \]

As a final remark, when a broader range of problems, such as constrained minimax optimization problems, it is necessary the convex subdifferential with respect to the \( x \) (saddle) subdifferential

\[ L \]

If

\[ 5.1 \text{ Preliminaries: Subdifferential operators and forward-backward residual} \]

Consider the setup with minimax objective of the form

\[ L(x, y) = L_p(x, y) + L_s(x, y), \]

where \( L_s \) is \( L \)-smooth. Informally assume that \( L_p \) is nonsmooth and proximable (we clarify this notion below). We reformulate the minimax optimization problem into the monotone inclusion problem

\[ \text{find } z \in \mathbb{R}^d : \quad 0 \in (\partial_{\pm} L_p + \nabla_{\pm} L_s)(z), \quad (6) \]

where \( \partial_{\pm} \) is the analog of \( \nabla_{\pm} \) for nonsmooth convex-concave functions (precisely defined in Section 5.1). We design the Anchored Proximal Gradient” (APG)

\[ \text{SM-EAG+} \quad (I + \alpha B - \xi_k, \epsilon_k) \]

\[ \text{ξ}_k + 1 = \beta_k z_0 + (1 - \beta_k) (J_{\alpha A}(z_k - \alpha B z_k) + \alpha B z_k), \quad (APG) \]

where \( A = \partial_{\pm} L_p, B = \nabla_{\pm} L_s, z_0 \in \mathbb{R}^d \) is the initial point, and \( \text{SM-EAG+} \quad (I + \alpha B - \xi_k, \epsilon_k, \epsilon_k) \) denotes the execution of SM-EAG+ with the 1-strongly monotone \( (1 + \alpha L) \)-smooth operator \( z \mapsto (I + \alpha B)(z) - \xi_k \) with the initial point using the step-size \( \sqrt{(1 + \alpha L)^2 + 1 + \epsilon_k} \) until reaching the tolerance \( \| \xi_k \| \leq \epsilon_k \).

APG is a proximal-gradient-type algorithm; each iteration, APG uses a single proximal operation \( J_{\alpha_{\pm} L_p} \) and multiple (logarithmically many) gradient operations \( \nabla_{\pm} L_s \). (The superscript * in the name APG indicates the multiple gradient evaluations per iteration.) Thus, APG is useful when \( L_p \) is proximable in the sense that \( J_{\alpha_{\pm} L_p} \) can be computed efficiently. APG is designed to approximate the accelerated proximal algorithm ÖHM-DRS, described soon, in the MP sense and thereby achieves a near-optimal accelerated rate.

5 Near-optimal prox-grad-type accelerated minimax algorithm

Consider the setup with minimax objective of the form \( L(x, y) = L_p(x, y) + L_s(x, y) \), where \( L_s \) is \( L \)-smooth. Informally assume that \( L_p \) is nonsmooth and proximable (we clarify this notion below). We reformulate the minimax optimization problem into the monotone inclusion problem

\[ \text{find } z \in \mathbb{R}^d : \quad 0 \in (\partial_{\pm} L_p + \nabla_{\pm} L_s)(z), \quad (6) \]

where \( \partial_{\pm} \) is the analog of \( \nabla_{\pm} \) for nonsmooth convex-concave functions (precisely defined in Section 5.1). We design the Anchored Proximal Gradient” (APG):

\[ z_k = \text{SM-EAG+} \quad (I + \alpha B - \xi_k, \epsilon_k) \]

\[ \text{ξ}_k + 1 = \beta_k z_0 + (1 - \beta_k) (J_{\alpha A}(z_k - \alpha B z_k) + \alpha B z_k), \quad (APG) \]

where \( A = \partial_{\pm} L_p, B = \nabla_{\pm} L_s, z_0 \in \mathbb{R}^d \) is the initial point, and \( \text{SM-EAG+} \quad (I + \alpha B - \xi_k, \epsilon_k, \epsilon_k) \) denotes the execution of SM-EAG+ with the 1-strongly monotone \( (1 + \alpha L) \)-smooth operator \( z \mapsto (I + \alpha B)(z) - \xi_k \) with the initial point using the step-size \( \sqrt{(1 + \alpha L)^2 + 1 + \epsilon_k} \) until reaching the tolerance \( \| \xi_k \| \leq \epsilon_k \).

APG is a proximal-gradient-type algorithm; each iteration, APG uses a single proximal operation \( J_{\alpha_{\pm} L_p} \) and multiple (logarithmically many) gradient operations \( \nabla_{\pm} L_s \). (The superscript * in the name APG indicates the multiple gradient evaluations per iteration.) Thus, APG is useful when \( L_p \) is proximable in the sense that \( J_{\alpha_{\pm} L_p} \) can be computed efficiently. APG is designed to approximate the accelerated proximal algorithm ÖHM-DRS, described soon, in the MP sense and thereby achieves a near-optimal accelerated rate.

5.1 Preliminaries: Subdifferential operators and forward-backward residual

If \( L_p : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R} \) is a finite-valued, not necessarily differentiable convex-concave function, then its (saddle) subdifferential operator \( \partial_{\pm} L_p : \mathbb{R}^n \times \mathbb{R}^m \rightrightarrows \mathbb{R}^n \times \mathbb{R}^m \) defined by

\[ \partial_{\pm} L_p(x, y) = \{ (v, w) \in \mathbb{R}^n \times \mathbb{R}^m : v \in \partial_\pm L_p(x, y), w \in \partial_\mp (-L_p)(x, y) \} \]

is maximal monotone [7, Theorem 35.1, Corollary 37.5.2], where \( \partial_x \) and \( \partial_y \) respectively denote the convex subdifferential with respect to the \( x \)- and \( y \)-variables. However, in order to encompass a broader range of problems, such as constrained minimax optimization problems, it is necessary
We now describe the design of APG with $\beta$ (APG).

The second lines are identical. In the first lines, the resolvent computation of OHM-DRS, and if $O$ is an instance of the problem class of Section 4, and SM-EAG+ can efficiently solve it to accuracy and $\varepsilon$.

Of particular interest are problems with $L(x, y) = f(x) + \langle Ax, y \rangle - g(y)$, where some $A \in \mathbb{R}^{m \times n}$, or, more generally, $L(x, y) = f(x) + L_* (x, y) - g(y)$, where some $L$-smooth convex-concave $L_*$. In this case, $\partial L_p (x, y) = (\partial f)(x, \partial g(y))$, and $J_{\alpha g}(x, y) = (\text{Prox}_{\alpha f}(x), \text{Prox}_{\alpha g}(y))$, where

$$\text{Prox}_{\alpha f}(x) = \text{argmin}_{x' \in \mathbb{R}^n} \left\{ \alpha f(x') + \frac{1}{2} \|x' - x\|^2 \right\}, \quad \text{Prox}_{\alpha g}(y) = \text{argmin}_{y' \in \mathbb{R}^m} \left\{ \alpha g(y') + \frac{1}{2} \|y' - y\|^2 \right\}$$

are the proximal operators. When $f(x) = \delta_X(x) = \begin{cases} 0 & x \in \mathcal{X} \\ +\infty & x \notin \mathcal{X} \end{cases}$, the proximal operator becomes the projection operator: $\text{Prox}_{\alpha f}(x) = \Pi_X(x)$ for any $\alpha > 0$. Therefore, the constrained problem (1) with $\mathcal{X} \subseteq \mathbb{R}^n$, $\mathcal{Y} \subseteq \mathbb{R}^m$ is recovered with $L_p(x, y) = \delta_X(x) - \delta_Y(y)$.

Note that $z_*$ is a solution of (6) if and only if $z_* = J_{\alpha A}(z_\alpha - \alpha B z_\alpha)$. Therefore, the norm of the forward-backward residual $G_\alpha(z) = \frac{1}{\alpha} (z - J_{\alpha A}(z - \alpha B z))$ serves as a measure for the convergence of algorithms solving (6), and it generalizes the operator norm and the fixed-point residual norm.

### 5.2 Preliminaries: Accelerated proximal splitting algorithm

Consider

$$w_k = J_{\alpha B}(u_k)$$

$$u_{k+1} = \beta_k u_0 + (1 - \beta_k) (J_{\alpha A}(w_k - \alpha B w_k) + \alpha B w_k)$$

(OMH-DRS)

with $\beta_k = \frac{1}{k+2}$, which is the application of OHM to the Douglas–Rachford splitting (DRS) operator [27, 53] defined by $T_{\text{DRS}} = I - J_{\alpha B} + J_{\alpha A} \circ (2J_{\alpha B} - 1)$. Since $T_{\text{DRS}}$ is nonexpansive [53], OHM-DRS exhibits the accelerated rate $\|\alpha G_\alpha(w_k)\|^2 = \|w_k - T_{\text{DRS}}(u_k)\|^2 = O(1/k^2)$. (Here we use the alternative representation of OHM $u_{k+1} = \beta_k u_0 + (1 - \beta_k) T u_k$ with $\beta_k = \frac{1}{k+2}$, which is equivalent to (4) under the identification $u_k = w_{k+1/2}$. We clarify some details in Appendix A.6).

OMH-DRS will serve as a conceptual algorithm (not implemented) that our APG* (implementable) is designed to $O(1/k^2)$-MP approximate, so that APG* inherits the accelerated rate of OHM-DRS. (OMH-DRS is not prox-grad-type, while APG* is.)

### 5.3 MP property and convergence of APG*

We now describe the design of APG*. The forms of APG* and OHM-DRS bear a clear resemblance. The second lines are identical. In the first lines, the resolvent computation $J_{\alpha B}$ of OHM-DRS is replaced in APG* by an inner loop with SM-EAG+. Note, $z = J_{\alpha B}(\xi_k)$ if and only if it solves

$$\text{find } z \in \mathbb{R}^d \quad 0 = z + \alpha B z - \xi_k = (\mathbf{I} + \alpha B - \xi_k)(z),$$

(7)

and $I + \alpha B - \xi_k$ is a 1-strongly monotone, $(1 + \alpha L)$-Lipschitz operator. Therefore, (7) is an instance of the problem class of Section 4, and SM-EAG+ can efficiently solve it to accuracy $\|I + \alpha B\|(z) - \xi_k \leq \varepsilon_k$ in $O(\log(1/\varepsilon_k))$ iterations. Therefore the $z_k$-iterate of APG* approximates the $w_k = J_{\alpha B}(u_k)$-iterate of OHM-DRS.

**Theorem 6** (APG* ≈ OHM-DRS). With $\alpha \in (0, \frac{1}{L})$, $\beta_k = \frac{1}{k+2}$, and $\varepsilon_k = \frac{1 + L^{-1} \|\xi_0\|}{(k+1)(k+2)}$, APG* is $O(1/k^2)$-MP to OHM-DRS; specifically, if $\xi_k, z_k$ are the APG* iterates and $u_k, w_k$ are the iterates of OHM-DRS, and if $\xi_0 = u_0$, then

$$\max \left\{ \|\xi_k - u_k\|^2, \|z_k - w_k\|^2 \right\} \leq \frac{C(\xi_0)^2}{L^2(k + 1)^2}, \quad \text{for } k = 0, 1, \ldots,$$

where $C(\xi_0) = L(\|\xi_0 - \xi_*\| + 1) + \|B \xi_*\|$ and $\xi_* = \Pi_{\text{Fix}(T_{\text{DRS}})}(\xi_0)$.

8
Corollary 7 (Fast rate of APG'). With the conditions and notations of Theorem 6, $\xi_k \to \xi^*$ and $z_k \to J_{\alpha \beta}(\xi^*) \in \text{Zer}(A + B)$ as $k \to \infty$. Additionally,

$$\|G_\alpha(z_k)\|^2 \leq \frac{(3 + \alpha L)^2 C(\xi_0)^2}{\alpha^2 L^2(k + 1)^2}, \quad \text{for } k = 0, 1, \ldots.$$ 

Corollary 8 (Oracle complexity of APG'). With the conditions and notations of Theorem 6, the $k$th inner loop of SM-EAG+ in APG' requires $\mathcal{O}(\log(1/\epsilon))$ oracle calls of $B$ for $k = 0, 1, \ldots$. This implies that the total computational cost needed to achieve $\|G_\alpha(z_k)\| \leq \epsilon$ is

$$\mathcal{O}\left(\frac{C(\xi_0)}{\epsilon} \left(C_A + C_B \log \frac{C(\xi_0)}{\epsilon}\right)\right),$$

where $C_A, C_B$ are upper bounds on the respective costs of evaluating $J_{\alpha \beta}$ and $B$.

Discussion. We clarify some key differences between APG' and the prior work of Diakonikolas [24, Algorithm 3]. First, APG' is a more general proximal-gradient-type algorithm, while the prior algorithm of Diakonikolas is a projected-gradient-type algorithm. Second, APG' is an anytime algorithm, as it requires no a priori specification of the terminal accuracy or the total iteration count, while the prior algorithm is not an anytime algorithm, as it requires the terminal accuracy to be pre-specified. Third, APG' has $\mathcal{O}\left(\frac{C_A}{\epsilon} + C_B \log \frac{C(\xi_0)}{\epsilon}\right)$ total cost (hiding dependency on $\xi_0, \xi^*$ and $L$), while the prior algorithm has $\mathcal{O}\left(\frac{C_A + C_B \log \frac{C(\xi_0)}{\epsilon}}{\epsilon}\right)$ total cost, as its inner loops require projections (corresponding to $J_{\alpha \beta}$). Therefore, APG' offers a speedup for constrained minimax problems when the projection represents the dominant cost, i.e., when $C_A \gg C_B$.

The initial point $\xi_k$ for the inner loop SM-EAG+ ($I + \alpha B - \xi_k, \xi_k, \epsilon_k$) of APG' is a crude choice. However, while a more careful warm starting may improve the iteration complexity by a constant factor, we do not believe it alone can eliminate the $\mathcal{O}(\log(1/\epsilon))$-factor. In our analysis of APG' presented in Appendix D, we induce the MP property by solving the $k$th inner loop up to accuracy $\epsilon_k$ satisfying $\sum_{k=0}^{\infty} k \epsilon_k < \infty$. Therefore, unless a mechanism for finding a warm-starting point $z$ satisfying $\|(I + \alpha B)(z) - \xi^*\| = O(\epsilon_k) = o(1/k)$ is proposed (which seems nontrivial), the $\mathcal{O}(\log(1/\epsilon))$-factor will persist. Eliminating the logarithmic factor probably requires a new additional insight, and we leave it as a future work.

6 Conclusion

In this paper, we identified the novel merging path (MP) property among accelerated minimax algorithms, EAG [100], FEG [49], and APS [92], and the accelerated proximal algorithm OHM [35, 51, 43]. We then used this MP property to design new state-of-the-art algorithms.

We quickly mention that our results straightforwardly extend to the setup where the underlying space $\mathbb{K}^d$ is replaced with an infinite-dimensional Hilbert space, since our proofs do not rely on the finite-dimensionality of $\mathbb{K}^d$. In particular, the point convergence results of Corollaries 3 and 7 hold as strong convergence.

The MP property provides the insight that the four different accelerated algorithms, EAG, FEG, APS, and OHM, actually reduce to one single acceleration mechanism. However, elucidating the fundamental principle behind this single anchoring-based acceleration mechanism still remains an open problem, and further investigating the phenomenon from multiple different viewpoints, as was done for Nesterov’s acceleration, would be an interesting direction of future work.

The presence or applicability of the MP property in other setups is another interesting direction of future work. One question is whether other prior optimization algorithms exhibit the MP property. In our analysis, we found that anchoring induces the MP property, so algorithms using similar regularizations, such as the recursive regularization [4] or the Katyusha momentum [3], may be approximating some other conceptual algorithm. Another question is whether the MP property can be used to design new algorithms in other setups. The strategy of designing efficient algorithms by approximating a conceptual algorithm, in the MP sense, may be widely applicable in setups beyond non-stochastic minimax optimization.
References

[1] K. Ahn and S. Sra. Understanding Nesterov’s acceleration via proximal point method. Symposium on Simplicity in Algorithms, 2022.

[2] M. S. Alkousa, A. V. Gasnikov, D. M. Dvinskikh, D. A. Kovalev, and F. S. Stonyakin. Accelerated methods for saddle-point problem. Computational Mathematics and Mathematical Physics, 60(11):1787–1809, 2020.

[3] Z. Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. Journal of Machine Learning Research, 18(1):8194–8244, 2017.

[4] Z. Allen-Zhu. How to make the gradients small stochastically: Even faster convex and nonconvex SGD. Neural Information Processing Systems, 2018.

[5] W. Azizian, I. Mitliagkas, S. Lacoste-Julien, and G. Gidel. A tight and unified analysis of gradient-based methods for a whole spectrum of differentiable games. International Conference on Artificial Intelligence and Statistics, 2020.

[6] B. Van Scoy, R. A. Freeman, and K. M. Lynch. The fastest known globally convergent first-order method for minimizing strongly convex functions. IEEE Control Systems Letters, 2(1):49–54, 2018.

[7] H. H. Bauschke and P. L. Combettes. Convex Analysis and Monotone Operator Theory in Hilbert Spaces. Springer International Publishing, second edition, 2017.

[8] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM Journal on Imaging Sciences, 2(1):183–202, 2009.

[9] Y. Censor, A. Gibali, and S. Reich. Strong convergence of subgradient extragradient methods for the variational inequality problem in Hilbert space. Optimization Methods and Software, 26(4-5):827–845, 2011.

[10] Y. Censor, A. Gibali, and S. Reich. The subgradient extragradient method for solving variational inequalities in Hilbert space. Journal of Optimization Theory and Applications, 148(2):318–335, 2011.

[11] A. Chambolle and C. Dossal. On the convergence of the iterates of the “fast iterative shrinkage/thresholding algorithm”. Journal of Optimization Theory and Applications, 166(3):968–982, 2015.

[12] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. Journal of Mathematical Imaging and Vision, 40(1):120–145, 2011.

[13] A. Chambolle and T. Pock. On the ergodic convergence rates of a first-order primal-dual algorithm. Mathematical Programming, 159(1):253–287, 2016.

[14] T. Chavdarova, G. Gidel, F. Fleuret, and S. Lacoste-Julien. Reducing noise in GAN training with variance reduced extragradient. Neural Information Processing Systems, 2019.

[15] Y. Chen, G. Lan, and Y. Ouyang. Optimal primal-dual methods for a class of saddle point problems. SIAM Journal on Optimization, 24(4):1779–1814, 2014.

[16] Y. Chen, G. Lan, and Y. Ouyang. Accelerated schemes for a class of variational inequalities. Mathematical Programming, 165(1):113–149, 2017.

[17] P. L. Combettes and J.-C. Pesquet. Proximal splitting methods in signal processing. In H. H. Bauschke, R. S. Burachik, P. L. Combettes, V. Elser, D. R. Luke, and H. Wolkowicz, editors, Fixed-Point Algorithms for Inverse Problems in Science and Engineering, pages 185–212. Springer, 2011.

[18] L. Condat. A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms. Journal of Optimization Theory and Applications, 158(2):460–479, 2013.

[19] E. R. Csetnek, Y. Malitsky, and M. K. Tam. Shadow Douglas–Rachford splitting for monotone inclusions. Applied Mathematics & Optimization, 80(3):665–678, 2019.

[20] C. Daskalakis, A. Deckelbaum, and A. Kim. Near-optimal no-regret algorithms for zero-sum games. Games and Economic Behavior, 92:327–348, 2015.

[21] C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng. Training GANs with optimism. International Conference on Learning Representations, 2018.
[22] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57(11):1413–1457, 2004.

[23] D. Davis and B. Grimmer. Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems. *SIAM Journal on Optimization*, 29(3):1908–1930, 2019.

[24] J. Diakonikolas. Halpern iteration for near-optimal and parameter-free monotone inclusion and strong solutions to variational inequalities. *Conference on Learning Theory*, 2020.

[25] J. Diakonikolas, C. Daskalakis, and M. Jordan. Efficient methods for structured nonconvex-nonconcave min-max optimization. *International Conference on Artificial Intelligence and Statistics*, 2021.

[26] J. Diakonikolas and P. Wang. Potential function-based framework for making the gradients small in convex and min-max optimization. *SIAM Journal on Optimization, To appear: arXiv:2101.12101*, 2022.

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

[28] D. Drusvyatskiy. The proximal point method revisited. *arXiv preprint arXiv:1712.06038*, 2017.

[29] H. Edwards and A. J. Storkey. Censoring representations with an adversary. *International Conference on Learning Representations*, 2016.

[30] G. Gidel, H. Berard, G. Vignoud, P. Vincent, and S. Lacoste-Julien. A variational inequality perspective on generative adversarial networks. *International Conference on Learning Representations*, 2019.

[31] G. Gidel, R. A. Hemmat, M. Pezeshki, R. L. Priol, G. Huang, S. Lacoste-Julien, and I. Mitliagkas. Negative momentum for improved game dynamics. *International Conference on Artificial Intelligence and Statistics*, 2019.

[32] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Neural Information Processing Systems*, 2014.

[33] E. Gorbunov, N. Loizou, and G. Gidel. Extragradient method: $O(1/K)$ last-iterate convergence for monotone variational inequalities and connections with cocoercivity. *International Conference on Artificial Intelligence and Statistics*, 2022.

[34] O. Güler. New proximal point algorithms for convex minimization. *SIAM Journal on Optimization*, 2(4):649–664, 1992.

[35] B. Halpern. Fixed points of nonexpanding maps. *Bulletin of the American Mathematical Society*, 73(6):957–961, 1967.

[36] E. Y. Hamedani and N. S. Aybat. A primal-dual algorithm with line search for general convex-concave saddle point problems. *SIAM Journal on Optimization*, 31(2):1299–1329, 2021.

[37] B. He and X. Yuan. Convergence analysis of primal-dual algorithms for a saddle-point problem: From contraction perspective. *SIAM Journal on Imaging Sciences*, 5(1):119–149, 2012.

[38] Y. He and R. D. C. Monteiro. An accelerated HPE-Type algorithm for a class of composite convex-concave saddle-point problems. *SIAM Journal on Optimization*, 26(1):29–56, 2016.

[39] Y.-G. Hsieh, K. Antonakopoulos, and P. Mertikopoulos. Adaptive learning in continuous games: Optimal regret bounds and convergence to Nash equilibrium. *Conference on Learning Theory*, 2021.

[40] Y.-G. Hsieh, F. Iutzeler, J. Malick, and P. Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. *Neural Information Processing Systems*, 2019.

[41] Y. Jin, A. Sidford, and K. Tian. Sharper rates for separable minimax and finite sum optimization via primal-dual extragradient methods. *arXiv preprint arXiv:2202.04640*, 2022.

[42] A. Juditsky, A. Nemirovski, and C. Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.
[43] D. Kim. Accelerated proximal point method for maximally monotone operators. *Mathematical Programming*, 190(1–2):57–87, 2021.

[44] D. Kim and J. A. Fessler. Optimized first-order methods for smooth convex minimization. *Mathematical Programming*, 159(1–2):81–107, 2016.

[45] D. Kim and J. A. Fessler. Optimizing the efficiency of first-order methods for decreasing the gradient of smooth convex functions. *Journal of Optimization Theory and Applications*, 188(1):192–219, 2021.

[46] O. Kolosvoski and R. Monteiro. An accelerated non-Euclidean hybrid proximal extragradient-type algorithm for convex–concave saddle-point problems. *Optimization Methods and Software*, 32(6):1244–1272, 2017.

[47] G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Ekonomika i Matematicheskie Metody*, 12(4):747–756, 1976.

[48] J. Lee, C. Park, and E. K. Ryu. A geometric structure of acceleration and its role in making gradients small fast. *Neural Information Processing Systems*, 2021.

[49] S. Lee and D. Kim. Fast extra gradient methods for smooth structured nonconvex-nonconcave minimax problems. *Neural Information Processing Systems*, 2021.

[50] T. Liang and J. Stokes. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. *International Conference on Artificial Intelligence and Statistics*, 2019.

[51] F. Lieder. On the convergence rate of the Halpern-iteration. *Optimization Letters*, 15(2):405–418, 2021.

[52] T. Lin, C. Jin, and M. I. Jordan. Near-optimal algorithms for minimax optimization. *Conference on Learning Theory*, 125, 2020.

[53] P. L. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16(6):964–979, 1979.

[54] S. I. Lyashko, V. V. Semenov, and T. A. Voitova. Low-cost modification of Korpelevich’s methods for monotone equilibrium problems. *Cybernetics and Systems Analysis*, 47(4):631, 2011.

[55] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations*, 2018.

[56] Y. Malitsky. Projected reflected gradient methods for monotone variational inequalities. *SIAM Journal on Optimization*, 25(1):502–520, 2015.

[57] Y. Malitsky. Golden ratio algorithms for variational inequalities. *Mathematical Programming*, 184(1):383–410, 2020.

[58] Y. Malitsky and M. K. Tam. A forward-backward splitting method for monotone inclusions without cocoercivity. *SIAM Journal on Optimization*, 30(2):1451–1472, 2020.

[59] Y. V. Malitsky and V. V. Semenov. An extragradient algorithm for monotone variational inequalities. *Cybernetics and Systems Analysis*, 50(2):271–277, 2014.

[60] B. Martinet. Régularisation d’inéquations variationnelles par approximations successives. *Revue Française d’Informatique et de Recherche Opérationnelle, Série Rouge*, 4(3):154–158, 1970.

[61] P. Mertikopoulos, B. Lecouat, H. Zenati, C.-S. Foo, V. Chandrasekhar, and G. Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. *International Conference on Learning Representations*, 2019.

[62] G. J. Minty. Monotone (nonlinear) operators in Hilbert space. *Duke Mathematical Journal*, 29(3):341–346, 1962.

[63] A. Mokhtari, A. Ozdaglar, and S. Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. *International Conference on Artificial Intelligence and Statistics*, 2020.

[64] A. Mokhtari, A. E. Ozdaglar, and S. Pattathil. Convergence rate of $O(1/k)$ for optimistic gradient and extragradient methods in smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 30(4):3230–3251, 2020.
[65] R. D. C. Monteiro and B. F. Svaiter. On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean. *SIAM Journal on Optimization*, 20(6):2755–2787, 2010.

[66] R. 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. *SIAM Journal on Optimization*, 21(4):1688–1720, 2011.

[67] J. Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société Mathématique de France*, 93:273–299, 1965.

[68] A. Nemirovski. Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

[69] Y. Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Doklady Akademii Nauk SSSR*, 269(3):543–547, 1983.

[70] Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005.

[71] Y. Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2):319–344, 2007.

[72] Y. Nesterov, A. Gasnikov, S. Guminov, and P. Dvurechensky. Primal–dual accelerated gradient methods with small-dimensional relaxation oracle. *Optimization Methods and Software*, pages 1–38, 2020.

[73] Y. Nesterov and L. Scrimali. Solving strongly monotone variational and quasi-variational inequalities. *Discrete & Continuous Dynamical Systems*, 31(4):1383–1396, 2011.

[74] M. A. Noor. New extragradient-type methods for general variational inequalities. *Journal of Mathematical Analysis and Applications*, 277(2):379–394, 2003.

[75] C. Park, J. Park, and E. K. Ryu. Factor-$\sqrt{2}$ acceleration of accelerated gradient methods. *arXiv:2102.07366*, 2021.

[76] J. Park and E. K. Ryu. Exact optimal accelerated complexity for fixed-point iterations. *International Conference on Machine Learning*, 2022.

[77] W. Peng, Y.-H. Dai, H. Zhang, and L. Cheng. Training GANs with centripetal acceleration. *Optimization Methods and Software*, 35(5):955–973, 2020.

[78] R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.

[79] E. K. Ryu and W. Yin. *Large-Scale Convex Optimization via Monotone Operators*. Cambridge University Press, 2022.

[80] E. K. Ryu, K. Yuan, and J. C. Duchi. Certifying some distributional robustness with principled adversarial training. *International Conference on Learning Representations*, 2018.

[81] A. Sinha, H. Namkoong, and J. C. Duchi. Certifying some distributional robustness with principled adversarial training. *International Conference on Learning Representations*, 2018.
[88] V. Syrgkanis, A. Agarwal, H. Luo, and R. E. Schapire. Fast convergence of regularized learning in games. Neural Information Processing Systems, 2015.

[89] A. Taylor and Y. Drori. An optimal gradient method for smooth strongly convex minimization. arXiv:2101.09741, 2021.

[90] K. K. Thekumparampil, N. He, and S. Oh. Lifted primal-dual method for bilinearly coupled smooth minimax optimization. International Conference on Artificial Intelligence and Statistics, 2022.

[91] K. K. Thekumparampil, P. Jain, P. Netrapalli, and S. Oh. Efficient algorithms for smooth minimax optimization. Neural Information Processing Systems, 2019.

[92] Q. Tran-Dinh and Y. Luo. Halpern-type accelerated and splitting algorithms for monotone inclusions. arXiv:2110.08150, 2021.

[93] P. Tseng. A modified forward-backward splitting method for maximal monotone mappings. SIAM Journal on Control and Optimization, 38(2):431–446, 2000.

[94] B. C. Vū. A splitting algorithm for dual monotone inclusions involving cocoercive operators. Advances in Computational Mathematics, 38(3):667–681, 2013.

[95] Y. Wang and J. Li. Improved algorithms for convex-concave minimax optimization. Neural Information Processing Systems, 2020.

[96] G. Xie, Y. Han, and Z. Zhang. DIPPA: An improved method for bilinear saddle point problems. arXiv preprint arXiv:2103.08270, 2021.

[97] A. K. Yadav, S. Shah, Z. Xu, D. W. Jacobs, and T. Goldstein. Stabilizing adversarial nets with prediction methods. International Conference on Learning Representations, 2018.

[98] M. Yan. A new primal-dual algorithm for minimizing the sum of three functions with a linear operator. Journal of Scientific Computing, 76(3):1698–1717, 2018.

[99] J. Yang, S. Zhang, N. Kiyavash, and N. He. A catalyst framework for minimax optimization. Neural Information Processing Systems, 2020.

[100] T. Yoon and E. K. Ryu. Accelerated algorithms for smooth convex-concave minimax problems with $O(1/k^2)$ rate on squared gradient norm. International Conference on Machine Learning, 2021.

[101] K. Zhou, L. Tian, A. M.-C. So, and J. Cheng. Practical schemes for finding near-stationary points of convex finite-sums. International Conference on Artificial Intelligence and Statistics, 2022.
A Algorithm specifications

A.1 Classical minimax optimization algorithms

Simultaneous gradient descent-ascent for smooth minimax optimization is defined by

\[
\begin{align*}
x_{k+1} &= x_k - \alpha \nabla_x L(x_k, y_k) \\
y_{k+1} &= y_k + \alpha \nabla_y L(x_k, y_k)
\end{align*}
\]

where \(\alpha > 0\) is the step-size. Writing \(z_k = (x_k, y_k)\) and using the \(\nabla_\pm\) notation, we can more concisely rewrite this as

\[
z_{k+1} = z_k - \alpha \nabla_\pm L(z_k).
\]

The extragradient (EG) algorithm is defined by

\[
z_{k+1/2} = z_k - \alpha \nabla_\pm L(z_k),
\]

\[
z_{k+1} = z_k - \alpha \nabla_\pm L(z_{k+1/2}).
\]

The optimistic gradient descent (OG) algorithm is defined by

\[
z_{k+1} = z_k - \alpha \nabla_\pm L(z_k) - \alpha (\nabla_\pm L(z_k) - \nabla_\pm L(z_{k-1})),
\]

where \(z_0 = z_{-1}\) are the starting points.

A.2 Nesterov’s AGM and its momentum parameters

In his seminal work [69], Nesterov developed the accelerated gradient method (AGM) for the smooth convex minimization problem

\[
\text{minimize } f(x)
\]

where \(f: \mathbb{R}^n \to \mathbb{R}\) is convex and \(L\)-smooth (i.e., differentiable with \(L\)-Lipschitz continuous gradient). Nesterov’s original AGM has the form

\[
\begin{align*}
x_{k+1} &= y_k - \alpha \nabla f(y_k) \\
y_{k+1} &= x_{k+1} + \frac{t_k - 1}{t_k} (x_{k+1} - x_k)
\end{align*}
\]

with parameters \(\alpha = \frac{1}{1+\sqrt{1+4t_k^2}}\) for \(k = 0, 1, \ldots\) and has the convergence rate

\[
f(x_k) - f_* \leq \frac{\|x_0 - x_*\|^2}{2\alpha t_k^2} = \mathcal{O}\left(\frac{L\|x_0 - x_*\|^2}{k^2}\right)
\]

for \(k = 1, 2, \ldots\), where \(f_* = \min_{x \in \mathbb{R}^d} f(x)\).

In the study [11] on the point convergence of the fast iterative shrinkage-thresholding algorithm (FISTA), which generalizes AGM to the prox-grad setup, a relaxed parameter selection rule \(\alpha \leq \frac{1}{L}\) and \(t_{k+1}^2 - t_{k+1} - t_k^2 \leq 0\) was used. In particular, with the choice \(t_k = (k + a - 1)/a\) with \(a > 2\), the AGM (and FISTA) exhibits point convergence to a minimizer \((x_k \to x_*)\) for some minimizer \(x_*\) while retaining \(\mathcal{O}(\|x_0 - x_*\|^2/\alpha k^2)\) convergence rate [11]. (While \(a = 2\) leads to the fastest rate in terms of the function value suboptimality, whether the iterates convergence when \(a = 2\) remains an open question.) In Figure 1, we use AGM with parameters \(t_k = k+a-1\).

A.3 EAG and APS with iteration-dependent step-sizes

In Section 3, we introduced EAG and APS with constant step-size \(\alpha\). On the other hand, in the prior works introducing EAG [100] and APS [92], these algorithms were used with iteration-dependent step-sizes. Specifically, the EAG-V algorithm [100] takes the form

\[
\begin{align*}
z_{k+1/2} &= \beta_k z_0 + (1 - \beta_k) z_k - \alpha_k B z_k \\
z_{k+1} &= \beta_k z_0 + (1 - \beta_k) z_k - \alpha_k B z_{k+1/2}
\end{align*}
\]
where $\alpha_{k+1} = \alpha_k \left(1 - \frac{1}{(k+1)(k+3)} \frac{\alpha_k^2 L^2}{1-\alpha_k^2 L^2}\right)$ and $\beta_k = \frac{1}{k+2}$. The original anchored Popov’s scheme [92] takes the form

\begin{align*}
    w_{k+1} &= \beta_k z_0 + (1 - \beta_k) z_k - \alpha_k B u_k \\
    z_{k+1} &= \beta_k z_0 + (1 - \beta_k) z_k - \alpha_k B u_{k+1}
\end{align*}

and the convergence analysis was done for $\alpha_{k+1} = \frac{\alpha_k \beta_{k+1} (1 - \beta_k^2 - M \alpha_k^2)}{(1-M \alpha_k^2) \beta_k (1-\beta_k)}$, $\beta_k = \frac{1}{k+2}$, and $M = 2L^2(1 + \theta)$ for some $\theta > 0$. We provide a brief discussion on the MP results for algorithms with iteration-dependent step-sizes in Section F.

### A.4 Equivalent forms of OHM and their convergence

Recall that we used the two distinct forms of OHM (in Sections 3 and 5):

\begin{align*}
    u_{k+1} &= \frac{1}{k+1} u_0 + \frac{k}{k+1} w_k \\
    w_{k+1} &= T(w_{k+1/2})
\end{align*}

and

\begin{align*}
    u_{k+1} &= \frac{1}{k+2} u_0 + \frac{k+1}{k+2} T u_k. \\
    w_{k+1/2} &= u_k
\end{align*}

We show that the two forms are equivalent under the identification $w_{k+1/2} = u_k$ in the sense that if $u_0 = w_0 = w_{1/2}$, then $w_{k+1/2} = u_k$ for all $k = 0, 1, \ldots$. Suppose that $w_{k+1/2} = u_k$ holds for some $k \geq 0$. Then $w_{k+1} = T w_{k+1/2} = T u_k$, so

\begin{align*}
    w_{k+3/2} &= \frac{1}{k+2} w_0 + \frac{k+1}{k+2} w_{k+1} = \frac{1}{k+2} u_0 + \frac{k+1}{k+2} T u_k = u_{k+1}.
\end{align*}

By induction, this shows that if $u_0 = w_0$, then $w_{k+1/2} = u_k$ for all $k = 0, 1, \ldots$.

It is well-known that if $T$ is nonexpansive and $\text{Fix } T \neq \emptyset$, the Halpern iteration of the form

\begin{align*}
    u_{k+1} &= \beta_k u_0 + (1 - \beta_k) T u_k
\end{align*}

satisfying $\beta_k \in (0, 1)$, $\sum_{k=0}^\infty \beta_k = \infty$, and $\sum_{k=0}^\infty |\beta_k - \beta_{k+1}| < \infty$ strongly converges to $\Pi_{\text{Fix } T}(u_0)$ in Hilbert spaces [7, Theorem 30.1]. The particular choice $\beta_k = \frac{1}{k+2}$ satisfies these assumptions, so we see that the OHM iterates $u_k$ converge to $\Pi_{\text{Fix } T}(u_0)$. Because $T$ is continuous, this implies $w_{k+1} = T w_{k+1/2} = T u_k \rightarrow T \left( \Pi_{\text{Fix } T}(u_0) \right) = \Pi_{\text{Fix } T}(u_0)$ as $k \rightarrow \infty$. The tight convergence rate of OHM with respect to the fixed-point residual is [51, Theorem 2.1]

\begin{align*}
    \|w_{k+1/2} - T w_{k+1/2}\|^2 &= \|u_k - T u_k\|^2 \leq \frac{4\|u_0 - u_*\|^2}{(k+1)^2}
\end{align*}

for any $u_* \in \text{Fix } T$ and $k = 0, 1, \ldots$.

### A.5 Equivalent forms of OC-Halpern

The original form of OC-Halpern [76] was given by

\begin{align*}
    u_{k+1} &= \frac{u_0}{\varphi_k} + \left(1 - \frac{1}{\varphi_k}\right) T u_k
\end{align*}

where $\varphi_k = \sum_{j=0}^{k+1} \gamma_j$. Through similar analysis as in Section A.4, one can check that under the identification $w_{k+1/2} = u_k$, the above form is equivalent to

\begin{align*}
    w_{k+1/2} &= \frac{1}{\sum_{j=0}^{k+1} \gamma_j} w_0 + \left(1 - \frac{1}{\sum_{j=0}^{k+1} \gamma_j}\right) w_k \\
    w_{k+1} &= T(w_{k+1/2})
\end{align*}

which takes the form of (4) with $\beta_k = \left(\sum_{j=0}^{k+1} \gamma_j\right)^{-1}$, as we described in Section 4.
A.6 Halpern-accelerated Douglas–Rachford splitting

The DRS operator, defined by $T_{\text{DRS}} = I - J_{\alpha B} + J_{\alpha A} \circ (2J_{\alpha B} - I)$, is connected to the monotone inclusion problem

$$\begin{align*}
\text{find } z &\in \mathbb{R}^d \quad 0 \in (A + B)(z)
\end{align*}$$

in the following way: $u_* \in \text{Fix } T_{\text{DRS}}$ if and only if $w_* = J_{\alpha B}(u_*) \in \text{Zer } (A + B)$. This holds because

$$\begin{align*}
u_* \in \text{Fix } T_{\text{DRS}} &\iff u_* = T_{\text{DRS}}(u_*) = u_* - J_{\alpha B}(u_*) + J_{\alpha A} \circ (2J_{\alpha B} - I)(u_*) \\
&\iff w_* = J_{\alpha A}(2w_* - u_*) = J_{\alpha A}(w_* - \alpha Bw_*) \\
&\iff w_* + \alpha Aw_* = (I + \alpha A)(w_*) \ni w_* - \alpha Bw_* \\
&\iff 0 \in (A + B)(w_*),
\end{align*}$$

where in the second line, we use $u_* = (I + \alpha B)(w_*) \implies 2w_* - u_* = w_* - \alpha Bw_*$. Given $u_k$, letting $w_k = J_{\alpha B}(u_k)$ and $v_k = J_{\alpha A} \circ (2J_{\alpha B} - I)(u_k) = J_{\alpha A}(2w_k - u_k)$, we can write $T_{\text{DRS}}(u_k) = u_k - w_k + v_k$. Therefore, the OHM (11) applied with $T = T_{\text{DRS}}$ can be written as

$$\begin{align*}
w_k &= J_{\alpha B}(u_k) \\
v_k &= J_{\alpha A}(2w_k - u_k) \\
u_{k+1} &= \beta_k u_0 + (1 - \beta_k)(u_k + v_k - w_k)
\end{align*}$$

(13)

where $\beta_k = \frac{1}{k+2}$. But $2w_k - u_k = 2w_k - (w_k + \alpha Bw_k) = w_k - \alpha Bw_k$, so $v_k = J_{\alpha A}(w_k - \alpha Bw_k)$. Therefore, we can rewrite (13) as

$$\begin{align*}
w_k &= J_{\alpha B}(u_k) \\
u_{k+1} &= \beta_k u_0 + (1 - \beta_k)(J_{\alpha A}(w_k - \alpha Bw_k) + \alpha Bw_k)
\end{align*}$$

by eliminating the second line and replacing $u_k$ by $w_k + \alpha Bw_k$ and $v_k$ by $J_{\alpha A}(w_k - \alpha Bw_k)$ in the third line of (13). The rewritten version precisely agrees with OHM-DRS we presented in Section 5.2. Additionally, observe that

$$u_k - T_{\text{DRS}}(u_k) = w_k - v_k = w_k - J_{\alpha A}(w_k - \alpha Bw_k) = \alpha G_{\alpha}(w_k).$$

As $T_{\text{DRS}}$ is nonexpansive [53], by the discussion of Section A.4, (13) and OHM-DRS satisfy

$$\|\alpha G_{\alpha}(w_k)\|^2 = \|u_k - T_{\text{DRS}}(u_k)\|^2 \leq \frac{4\|u_0 - u_*\|^2}{(k+1)^2}$$

(14)

where $u_* \in \text{Fix } T_{\text{DRS}}$.

B Omitted proofs for Section 3

B.1 Proof of Lemma 2

It remains to prove that

$$V_k = p_k \|Bz_k\|^2 + q_k \langle Bz_k, z_k - z_0 \rangle + \frac{1}{2\alpha} \|z_0 - z_*\|^2$$

with $p_k = \frac{\alpha k^2}{2}$ and $q_k = k$ satisfies $V_k \geq 0$ and

$$V_k - V_{k+1} \geq \frac{\alpha (1 - \alpha^2 L^2)}{2} \|kBz_k - (k + 1)Bz_{k+1/2}\|^2$$

(15)

for $k = 0, 1, \ldots$. The first part (nonnegativity) is easy to show:

$$V_k \geq \frac{\alpha k^2}{2} \|Bz_k\|^2 + k \langle Bz_k, z_* - z_0 \rangle + \frac{1}{2\alpha} \|z_0 - z_*\|^2 \geq 0,$$

where the first inequality follows from monotonicity $\langle Bz_k, z_k - z_* \rangle \geq 0$ (as $Bz_* = 0$) and the second one is Young’s inequality. Next, we show (15).
Case $k = 0$. We have $z_{1/2} = z_0$ and $z_1 = z_0 - \alpha B z_{1/2} = z_0 - \alpha B z_0$, $V_0 = \frac{1}{2\alpha} \|z_0 - z_*\|^2$, and
\[
V_1 = \frac{\alpha}{2} \|B z_1\|^2 + \langle B z_1, z_1 - z_0 \rangle + \frac{1}{2\alpha} \|z_0 - z_*\|^2
\]
\[
= \frac{\alpha}{2} \|B z_1\|^2 - \alpha \langle B z_1, B z_0 \rangle + \frac{1}{2\alpha} \|z_0 - z_*\|^2.
\]
Because $B$ is $L$-Lipschitz,
\[
\|B z_1 - B z_0\|^2 \leq L^2 \|z_1 - z_0\|^2 = \alpha^2 L^2 \|B z_0\|^2
\]
and by multiplying $\frac{\alpha}{2}$ on both sides and rearranging terms, we get
\[
\frac{\alpha}{2} \|B z_1\|^2 - \alpha \langle B z_1, B z_0 \rangle \leq -\frac{\alpha}{2} (1 - \alpha^2 L^2) \|B z_0\|^2 = -\frac{\alpha}{2} (1 - \alpha^2 L^2) \|B z_{1/2}\|^2.
\]
Adding $\frac{1}{2\alpha} \|z_0 - z_*\|^2$, the above inequality becomes
\[
V_1 \leq V_0 - \frac{\alpha}{2} (1 - \alpha^2 L^2) \|B z_{1/2}\|^2,
\]
proving (15) for $k = 0$.

Case $k \geq 1$. From the definition of the FEG iterates, we have the identities
\[
z_{k+1} - z_0 = (1 - \beta_k)(z_k - z_0) - \alpha B z_{k+1/2}
\]
\[
z_k - z_{k+1} = \beta_k(z_k - z_0) + \alpha B z_{k+1/2}
\]
\[
z_{k+1/2} - z_k = \alpha B z_{k+1/2} - \alpha (1 - \beta_k) B z_k.
\]
Therefore, because $q_{k+1} = \frac{q_k}{\beta_k}$,
\[
V_k - V_{k+1} \geq V_k - V_{k+1} - \frac{q_k}{\beta_k} \langle z_k - z_{k+1}, B z_k - B z_{k+1} \rangle
\]
\[
= p_k \|B z_k\|^2 + q_k \langle B z_k, z_k - z_0 \rangle
\]
\[
- p_{k+1} \|B z_{k+1}\|^2 - q_k (1 - \beta_k) (z_k - z_0) - \alpha B z_{k+1/2}
\]
\[
- \frac{q_k}{\beta_k} \beta_k (z_k - z_0) + \alpha B z_{k+1/2}, B z_k - B z_{k+1})
\]
\[
= p_k \|B z_k\|^2 - \frac{q_k}{\beta_k} \|B z_k, z_{k+1/2}\|^2 + \frac{q_k}{\beta_k} \frac{1}{1 - \beta_k} \|B z_{k+1/2}, B z_{k+1}\| - p_{k+1} \|B z_{k+1}\|^2.
\]
Plugging in $\beta_k = \frac{1}{k+1}$, $p_k = \frac{k^2}{2}$, $p_{k+1} = \frac{q(k+1)^2}{2}$ and $q_k = k$, the inequality becomes
\[
V_k - V_{k+1} \geq \frac{\alpha k^2}{2} \|B z_k\|^2 - k \|B z_k, B z_{k+1/2}\| - \frac{\alpha (k+1)^2}{2} \|B z_{k+1}\|^2.
\]
(16)
Recall that $B$ is $L$-Lipschitz, so
\[
0 \geq L^2 \|z_{k+1/2} - z_k\|^2 - \|B z_{k+1/2} - B z_k\|^2
\]
\[
= \alpha^2 L^2 \left( B z_{k+1/2} - \frac{k}{k+1} B z_k \right)^2 - \|B z_{k+1/2} - B z_k\|^2.
\]
Multiplying $\frac{\alpha (k+1)^2}{2}$ to the above inequality and combining with (16), we lower-bound $V_k - V_{k+1}$ as
\[
\geq \frac{\alpha k^2}{2} \|B z_k\|^2 - \alpha k (k+1) \langle B z_k, B z_{k+1/2} \rangle + \alpha (k+1) \langle B z_{k+1/2}, B z_k \rangle - \frac{\alpha (k+1)^2}{2} \|B z_{k+1}\|^2
\]
\[
- \frac{\alpha (k+1)^2}{2} \left( \frac{k^2}{2} \left| B z_{k+1/2} - \frac{k}{k+1} B z_k \right| - \|B z_{k+1/2} - B z_k\|^2 \right)
\]
\[
= \frac{\alpha k^2}{2} - \frac{\alpha^3 L^2}{2} \|B z_k\|^2 + \alpha^3 L^2 k \|B z_{k+1/2}\|^2 - \alpha (k+1)(B z_k, B z_{k+1/2})
\]
\[
- \frac{\alpha (k+1)^2}{2} (1 - \alpha^2 L^2) \|B z_{k+1/2}\|^2
\]
\[
= \frac{\alpha (1 - \alpha^2 L^2)}{2} \|B z_k - (k+1) B z_{k+1/2}\|^2.
\]
We first consider the case of EAG. As before, denote the Halpern iterates by \( \{z_k\}_{k=0,1,...} \) and the EAG iterates by \( \{w_k\}_{k=0,1,...} \). Then in the exact same way as before, we have
\[
z_{k+1} - w_{k+1} = (\beta_k z_k + (1 - \beta_k) w_k - \alpha B z_{k+1/2}) - (\beta_k w_0 + (1 - \beta_k) w_k - \alpha B w_{k+1})
\]
and
\[
\|z_{k+1} - w_{k+1}\|^2 = (1 - \beta_k)^2 \|z_k - w_k\|^2 + 2 \langle \alpha(1 - \beta_k)(z_k - w_k), B w_{k+1} - B z_{k+1/2} \rangle
+ \alpha^2 \|B w_{k+1} - B z_{k+1/2}\|^2.
\]
(17)
The different part is
\[
z_{k+1/2} - w_{k+1} = (1 - \beta_k)(z_k - w_k) - \alpha B z_k + \alpha B w_{k+1}
\]
so that inner product term in (17) is rewritten as
\[
2 \langle \alpha(1 - \beta_k)(z_k - w_k), B w_{k+1} - B z_{k+1/2} \rangle
= 2 \langle \alpha(z_{k+1/2} - w_{k+1}) + \alpha^2 (B z_k - B w_{k+1}), B w_{k+1} - B z_{k+1/2} \rangle
\]
\[
\leq 2\alpha^2 \|B z_k - B w_{k+1}, B w_{k+1} - B z_{k+1/2}\|.
\]
Combining this with (17), we obtain
\[
\|z_{k+1} - w_{k+1}\|^2
\leq (1 - \beta_k)^2 \|z_k - w_k\|^2 + 2\alpha^2 \langle B z_k - B w_{k+1}, B w_{k+1} - B z_{k+1/2} \rangle + \alpha^2 \|B w_{k+1} - B z_{k+1/2}\|^2
\]
\[
= (1 - \beta_k)^2 \|z_k - w_k\|^2 - \alpha^2 \|B w_{k+1}\|^2 + 2\alpha^2 \langle B z_k, B w_{k+1} \rangle - 2\alpha^2 \|B z_k, B z_{k+1/2} \rangle + \alpha^2 \|B z_{k+1/2}\|^2
\]
Plugging in \( \beta_k = \frac{1}{k+1} \) and multiplying both sides by \((k + 1)^2\) gives
\[
(k + 1)^2 \|z_{k+1} - w_{k+1}\|^2 \leq k^2 \|z_k - w_k\|^2 + \alpha^2(k + 1)^2 \|B z_k - B z_{k+1/2}\|^2.
\]
Now, with the following Lemma 9, the proof for EAG is done.

**Lemma 9.** Let \( B : \mathbb{R}^d \rightarrow \mathbb{R}^d \) be monotone and \( L\)-Lipschitz and assume \( z* \in \text{Zer} B \) exists. Let \( \{z_k\}_{k=0,1,...} \) be the iterates of EAG with \( \beta_k = \frac{1}{k+1} \). Then there exists \( \eta \in (0,1) \) and \( C > 0 \) such that for \( \alpha \in (0, \frac{1}{L}) \),
\[
\sum_{k=0}^{\infty} (k + 1)^2 \|B z_k - B z_{k+1/2}\|^2 \leq \frac{C}{\alpha^2} \|z_0 - z*\|^2 < \infty.
\]
For the case of APS, proceeding with \( B z_k, B z_{k+1/2} \) replaced by \( B v_k, B v_{k+1} \) respectively, we get
\[
(k + 1)^2 \|z_{k+1} - w_{k+1}\|^2 \leq k^2 \|z_k - w_k\|^2 + \alpha^2(k + 1)^2 \|B v_k - B v_{k+1}\|^2
\]
and the proof is complete once we show:

**Lemma 10.** Let \( B : \mathbb{R}^d \rightarrow \mathbb{R}^d \) be monotone and \( L\)-Lipschitz and assume \( z* \in \text{Zer} B \) exists. Let \( \{z_k\}_{k=0,1,...} \) be the iterates of APS with \( \beta_k = \frac{1}{k+1} \). Then there exists \( \eta \in (0,1) \) and \( C > 0 \) such that for \( \alpha \in (0, \frac{1}{L}) \),
\[
\sum_{k=0}^{\infty} (k + 1)^2 \|B v_k - B v_{k+1}\|^2 \leq \frac{C}{\alpha^2} \|z_0 - z*\|^2 < \infty.
\]
The proofs of Lemmas 9 and 10 follow from arguments similar to that of Lemma 2. However, they are much more lengthy, so defer their proofs to Appendix E.
C Omitted proofs for Section 4

C.1 Proof of Lemma 4

Denote the iterates of OC-Halpern (12) by \( \{w_k\}_{k=0,1,...} \) and SM-EAG+ iterates by \( \{z_k\}_{k=0,1,...} \). Then as in Section 3, we have

\[
\|z_{k+1} - w_{k+1}\|^2 = (1 - \beta_k)^2 \|z_k - w_k\|^2 + 2 \langle \alpha(1 - \beta_k)(z_k - w_k), Bw_{k+1} - Bz_{k+1/2} \rangle \\
+ \alpha^2 \|Bw_{k+1} - Bz_{k+1/2}\|^2.
\]

Using

\[
z_{k+1/2} - w_{k+1} = (1 - \beta_k)(z_k - w_k) - \eta_k \alpha B z_k + \alpha B w_{k+1},
\]
we get

\[
2 \langle \alpha(1 - \beta_k)(z_k - w_k), Bw_{k+1} - Bz_{k+1/2} \rangle \\
= 2 \langle \alpha(z_{k+1/2} - w_{k+1}) + \alpha^2(\eta_k B z_k - B w_{k+1}), B w_{k+1} - B z_{k+1/2} \rangle \\
\leq -2 \mu \|z_{k+1/2} - w_{k+1}\|^2 + 2 \alpha^2 \langle \eta_k B z_k - B w_{k+1}, B w_{k+1} - B z_{k+1/2} \rangle
\]

where the last line uses \( \mu \)-strong monotonicity of \( B \). For any \( \epsilon \in (0, 1) \) and \( x, y \in \mathbb{R}^d \), we have

\[
\epsilon \|x\|^2 + \frac{1}{\epsilon} \|y\|^2 \geq 2 \langle x, y \rangle \implies \|x - y\|^2 \geq (1 - \epsilon) \|x\|^2 - \left( \frac{1}{\epsilon} - 1 \right) \|y\|^2.
\]

Using the last inequality with \( x = z_{k+1} - w_{k+1} \) and \( y = z_{k+1} - z_{k+1/2} = \alpha(\eta_k B z_k - B z_{k+1/2}) \), we can upper-bound (18) by

\[
-2 \mu \left( (1 - \epsilon) \|z_k - w_k\|^2 - \frac{1}{\epsilon} \|w_{k+1}\|^2 \right) + 2 \alpha^2 \|\eta_k B z_k - B w_{k+1}, B w_{k+1} - B z_{k+1/2}\|^2.
\]

Thus,

\[
\|z_{k+1} - w_{k+1}\|^2 \\
\leq (1 - \beta_k)^2 \|z_k - w_k\|^2 - 2 \alpha \mu \left( (1 - \epsilon) \|z_k - w_k\|^2 - \frac{1}{\epsilon} \|w_{k+1}\|^2 \right) + 2 \alpha^2 \|\eta_k B z_k - B w_{k+1}, B w_{k+1} - B z_{k+1/2}\|^2
\]

and rearranging, we obtain

\[
(1 + 2 \alpha \mu(1 - \epsilon))\|z_{k+1} - w_{k+1}\|^2 \\
\leq (1 - \beta_k)^2 \|z_k - w_k\|^2 + 2 \alpha \mu \left( \frac{1}{\epsilon} - 1 \right) \|\eta_k B z_k - B z_{k+1/2}\|^2 \\
- \alpha^2 \|B w_{k+1}\|^2 + 2 \epsilon \alpha \|B w_{k+1}, B z_k\|^2 - 2 \eta \alpha^2 \|z_k, B z_{k+1/2}\| + \alpha^2 \|B z_{k+1/2}\|^2
\]

\[
\leq (1 - \beta_k)^2 \|z_k - w_k\|^2 + \left( 1 + 2 \alpha \mu \left( \frac{1}{\epsilon} - 1 \right) \right) \|\eta_k B z_k - B z_{k+1/2}\|^2.
\]

Thus, provided that the following Lemma 11 (in particular (19)) holds, we have

\[
(1 + 2 \alpha \mu(1 - \epsilon))^{k+1}\|z_{k+1} - w_{k+1}\|^2 \\
\leq (1 + 2 \alpha \mu(1 - \epsilon))^k \|z_k - w_k\|^2 + (1 + 2 \alpha \mu)^k \left( 1 + 2 \alpha \mu \left( \frac{1}{\epsilon} - 1 \right) \right) \|\eta_k B z_k - B z_{k+1/2}\|^2
\]

\[
\leq \cdots \leq \left( 1 + 2 \alpha \mu \left( \frac{1}{\epsilon} - 1 \right) \right) \alpha^2 \sum_{j=0}^{k} (1 + 2 \alpha \mu)^j \|\eta_j B z_j - B z_{j+1/2}\|^2
\]

\[
\leq \left( 1 + 2 \alpha \mu \left( \frac{1}{\epsilon} - 1 \right) \right) \frac{1 + 2 \alpha \mu}{1 + 2 \alpha \mu - \alpha^2 \epsilon^2} \|z_0 - z_*\|^2,
\]

i.e., the two algorithms have \( \mathcal{O} \left( \frac{\|z_0 - z_*\|^2}{(1 + 2 \alpha \mu(1 - \epsilon))^k} \right) = \mathcal{O} \left( e^{-2k \alpha \mu(1 - \epsilon)} \|z_0 - z_*\|^2 \right) \)-MP.
Lemma 11. Let \( B : \mathbb{R}^d \to \mathbb{R}^d \) be a \( \mu \)-strongly monotone, \( L \)-Lipschitz operator \((0 < \mu \leq L)\), and let \( \{z_k\}_{k=0}^{\infty} \) be the iterates of SM-EAG+ with \( \beta_k = \frac{1}{\sum_{j=0}^{\infty} (1 + 2\alpha\mu)^j} \). Then, the quantities

\[
V_k = p_k \|Bz_k\|^2 + q_k \langle Bz_k - \mu(z_k - z_0), z_k - z_0 \rangle + \left( \frac{1}{2\alpha} + \mu \right) \|z_0 - z_*\|^2
\]

with \( p_0 = q_0 = 0 \) and \( q_k = \frac{1}{(1 + 2\alpha\mu)^k - 1} \), \( p_k = \frac{m^\alpha k}{2\beta_k} \) for \( k = 1, 2, \ldots \) satisfy \( V_k \geq 0 \). Then, the quantities

\[
V_k - V_{k+1} \geq \frac{\alpha(1 + 2\alpha\mu - \alpha^2 L^2)q_k}{2\beta_k(1 - \beta_k)} \|\eta_k Bz_k - Bz_{k+1/2}\|^2
\]

for \( k = 1, 2, \ldots \). In particular, if \( \alpha \in \left(0, \frac{\sqrt{L^2 + \mu^2} + \mu}{L^2}\right)\), \( V_k \) is a Lyapunov function for the algorithm, and if \( \alpha \in \left(0, \frac{\sqrt{L^2 + \mu^2} + \mu}{L^2}\right)\), we additionally have the summability result

\[
\sum_{k=0}^{\infty} (1 + 2\alpha\mu)^k \|\eta_k Bz_k - Bz_{k+1/2}\|^2 \leq \frac{1 + 2\alpha\mu}{\alpha^2(1 + 2\alpha\mu - \alpha^2 L^2)} \|z_0 - z_*\| < \infty. \tag{19}
\]

C.2 Proof of Lemma 11

Due to the subtlety caused by setting \( \beta_0 = 1 \), we divide the cases \( k = 0 \) and \( k \geq 1 \).

Case \( k = 0 \). We have \( z_{1/2} = z_0, z_1 = z_0 - \alpha Bz_0, V_0 = \left( \frac{1}{2\alpha} + \mu \right) \|z_0 - z_*\|^2 \), and

\[
V_1 = \frac{\alpha}{2} \|Bz_1\|^2 + \langle Bz_1 - \mu(z_1 - z_0), z_1 - z_0 \rangle + \left( \frac{1}{2\alpha} + \mu \right) \|z_0 - z_*\|^2
\]

\[
= \frac{\alpha}{2} \|Bz_1\|^2 - \alpha \langle Bz_1, Bz_0 \rangle - \alpha^2 \mu \|Bz_0\|^2 + \left( \frac{1}{2\alpha} + \mu \right) \|z_0 - z_*\|^2.
\]

Because \( B \) is \( L \)-Lipschitz,

\[
0 \leq L^2 \|z_1 - z_0\|^2 - \|Bz_1 - Bz_0\|^2
\]

\[
= (\alpha^2 L^2 - 1) \|Bz_0\|^2 + 2 \langle Bz_0, Bz_1 \rangle - \|Bz_1\|^2.
\]

Therefore,

\[
V_0 - V_1 = \alpha^2 \mu \|Bz_0\|^2 + \alpha \langle Bz_0, Bz_1 \rangle - \frac{\alpha}{2} \|Bz_1\|^2
\]

\[
= \frac{\alpha}{2} \left(1 + 2\alpha\mu - \alpha^2 L^2\right) \|Bz_0\|^2 + \frac{\alpha}{2} \left((\alpha^2 L^2 - 1) \|Bz_0\|^2 + 2 \langle Bz_0, Bz_1 \rangle - \|Bz_1\|^2\right)
\]

\[
\geq \frac{\alpha}{2} \left(1 + 2\alpha\mu - \alpha^2 L^2\right) \|Bz_0\|^2
\]

\[
= \frac{\alpha}{2} \left(1 + 2\alpha\mu - \alpha^2 L^2\right) \|\eta_0 Bz_0 - Bz_{1/2}\|^2
\]

where the last equality holds because \( \eta_0 = 0 \) and \( z_{1/2} = z_0 \).

Case \( k \geq 1 \). We have the identities

\[
z_{k+1} - z_0 = (1 - \beta_k)(z_k - z_0) - \alpha Bz_{k+1/2}
\]

\[
z_k - z_{k+1} = \beta_k (z_k - z_0) + \alpha Bz_{k+1/2}
\]

\[
z_{k+1/2} - z_{k+1} = \alpha Bz_{k+1/2} - \eta_k \alpha Bz_k.
\]

Note that \( \beta_k = \frac{1}{\sum_{j=0}^{\infty} (1 + 2\alpha\mu)^j} \) satisfies

\[
\frac{1}{\beta_{k+1}} = \sum_{j=0}^{k+1} (1 + 2\alpha\mu)^j = 1 + (1 + 2\alpha\mu) \sum_{j=0}^{k} (1 + 2\alpha\mu)^j = 1 + \frac{1 + 2\alpha\mu}{\beta_k}
\]
for all $k = 0, 1, \ldots$, so $q_k = \frac{1}{(1 + 2\alpha \mu)^k}$ satisfies the recurrence relation

$$q_{k+1} = \frac{q_k}{\beta_k} \frac{\beta_{k-1}}{(1 + 2\alpha \mu)} = \frac{q_k}{\beta_k} \frac{1}{1 - 1/\beta_k} = \frac{q_k}{2} \quad (23)$$

for $k = 1, 2, \ldots$. Additionally, because $\eta_k = \frac{1 - \beta_k}{1 + 2\alpha \mu}$ and $p_k = \frac{\eta_k q_k}{\beta_k}$,

$$p_k = \frac{\alpha (1 - \beta_k) q_k}{2 \beta_k (1 + 2\alpha \mu)} = \frac{\alpha}{2} \left( \frac{1}{\beta_k} - 1 \right) \frac{q_k}{1 + 2\alpha \mu} = \frac{\alpha}{2} \frac{1}{\beta_{k-1}} q_k = \frac{\alpha}{2} (1 + 2\alpha \mu)^{k-1} q_k^2 \quad (24)$$

and

$$p_{k+1} = \frac{\alpha \eta_{k+1} q_{k+1}}{2 \beta_{k+1}} = \frac{\alpha q_{k+1} 1 - \beta_{k+1}}{2 \beta_{k+1} (1 + 2\alpha \mu)} = \frac{\alpha q_k}{2 (1 - \beta_k) (1 + 2\alpha \mu) \beta_{k+1}} (1 - 1/\beta_k) = \frac{\alpha q_k}{2 \beta_k (1 - \beta_k)} \quad (25)$$

The identities (23) through (25) will be useful for the subsequent arguments.

Using $\mu$-strong monotonicity of $B$ and the identities (20), (21) and (23), we have

$$V_k - V_{k+1} - \frac{q_k}{\beta_k} (z_k - z_{k+1}, B z_k - B z_{k+1} - \mu(z_k - z_{k+1}))$$

$$= p_k \|B z_k\|^2 + q_k (B z_k, z_k - z_0) - q_k \|z_k - z_0\|^2$$

$$- p_{k+1} \|B z_{k+1}\|^2 - q_{k+1} (B z_{k+1}, (1 - \beta_k)(z_k - z_0) - \alpha B z_{k+1})$$

$$+ q_{k+1} \|z_k - (1 - \beta_k)(z_k - z_0) - \alpha B z_{k+1}\|^2$$

$$- \frac{q_k}{\beta_k} (\beta_k(z_k - z_0) + \alpha B z_{k+1})$$

$$+ q_k \mu \|\beta_k(z_k - z_0) + \alpha B z_{k+1}\|^2$$

$$= p_k \|B z_k\|^2 - \frac{\alpha q_k}{\beta_k} (B z_k, B z_{k+1}/\beta_k) + \mu \alpha^2 \left( \frac{q_k}{\beta_k} + q_{k+1} \right) \|B z_{k+1}/\beta_k\|^2$$

$$+ \alpha \left( \frac{q_k}{\beta_k} + q_{k+1} \right) \|B z_{k+1}/\beta_k\|$$

$$+ 2 \mu \alpha \left( q_k - (1 - \beta_k) q_{k+1} \right) (B z_{k+1}/\beta_k, z_k - z_0)$$

$$+ (q_k - (1 - \beta_k) q_{k+1}) (B z_{k+1}/\beta_k - B z_k - \beta_k)$$

$$+ \mu (1 - \beta_k)(1 - \beta_k) \|z_k - z_0\|^2$$

$$= p_k \|B z_k\|^2 - \frac{\alpha q_k}{\beta_k} (B z_k, B z_{k+1}/\beta_k) + \frac{\mu \alpha^2 q_k}{\beta_k (1 - \beta_k)} \|B z_{k+1}/\beta_k\|^2$$

$$+ \frac{\alpha q_k}{\beta_k (1 - \beta_k)} (B z_{k+1}/\beta_k - B z_k - \beta_k) \|B z_{k+1}/\beta_k\|^2 \quad (26)$$

Now because $B$ is $L$-Lipschitz,

$$\|B z_{k+1}/\beta_k - B z_k\|^2 \leq L^2 \|z_{k+1}/\beta_k - z_k\|^2 = L^2 \|\eta_k \alpha B z_k\|^2$$

where we applied the identity (22) for the last equality. Hence, using the above inequality, we obtain

$$(26) - p_{k+1} \left( L^2 \|\eta_k \alpha B z_k\|^2 - \|B z_{k+1}/\beta_k - B z_k\|^2 \right)$$

$$= (p_k - q_{k+1} \eta_k \alpha^2 L^2) \|B z_k\|^2 - \left( \frac{\alpha q_k}{\beta_k} - 2 p_{k+1} \eta_k \alpha^2 L^2 \right) (B z_k, B z_{k+1}/\beta_k)$$

$$+ \left( \frac{\mu \alpha^2 q_k}{\beta_k (1 - \beta_k)} + p_{k+1} (1 - \alpha^2 L^2) \right) \|B z_{k+1}/\beta_k\|^2 + \left( \frac{\alpha q_k}{\beta_k (1 - \beta_k)} - 2 p_{k+1} \right) (B z_{k+1}/\beta_k, B z_k)$$

$$= \frac{\eta_k \alpha q_k}{\beta_k (1 - \beta_k)} \left( 1 - \eta_k \alpha^2 L^2 \right) \|B z_k\|^2 - \frac{\alpha q_k}{\beta_k} \left( 1 - \eta_k \alpha^2 L^2 \right) \|B z_{k+1}/\beta_k\|^2$$

$$+ \frac{\alpha q_k}{\beta_k (1 - \beta_k)} (1 + 2 \mu - \alpha^2 L^2) \|B z_{k+1}/\beta_k\|^2 \quad (27)$$
where the coefficient of the \( \|Bz_{k+1}\|^2 \) vanishes due to (25). Because \( \eta_k = \frac{1 - \beta_k}{1 + 2\alpha \mu} \), we have
\[
1 - \eta_k \alpha^2 L^2 \frac{1}{1 - \beta_k} = \eta_k \left( \frac{1}{\eta_k} - \alpha^2 L^2 \frac{1}{1 - \beta_k} \right) = \frac{\eta_k (1 + 2\alpha \mu - \alpha^2 L^2)}{1 - \beta_k}
\]
and thus, (27) becomes
\[
\frac{\alpha (1 + 2\alpha \mu - \alpha^2 L^2) q_k}{2\beta_k (1 - \beta_k)} \| \eta_k Bz_k - Bz_{k+1/2} \|^2.
\]
As \( 1 + 2\alpha \mu - \alpha^2 L^2 \geq 0 \) if \( 0 < \alpha \leq \frac{\sqrt{L^2 + \mu^2 + \mu}}{L^2} \), for this range of \( \alpha \),
\[
V_k - V_{k+1} \geq \frac{\alpha (1 + 2\alpha \mu - \alpha^2 L^2) q_k}{2\beta_k (1 - \beta_k)} \| \eta_k Bz_k - Bz_{k+1/2} \|^2
\]
\[
\geq \frac{\alpha (1 + 2\alpha \mu - \alpha^2 L^2)}{2} (1 + 2\alpha \mu)^k \| \eta Bz_k - Bz_{k+1} \|^2 \geq 0,
\]
where we used \( \frac{q_k}{1 - \beta_k} = q_{k+1} > 1 \) and \( \frac{1}{\beta_k} = \sum_{j=0}^{k} (1 + 2\alpha \mu)^j > (1 + 2\alpha \mu)^k \). Therefore, once we show that \( V_N \geq 0 \) holds for any \( N \in \mathbb{N} \), we obtain
\[
\sum_{k=0}^{N} (1 + 2\alpha \mu)^k \| \eta Bz_k - Bz_{k+1} \|^2 \leq \frac{2}{\alpha (1 + 2\alpha \mu - \alpha^2 L^2)} (V_0 - V_{N+1})
\]
\[
\leq \frac{2}{\alpha (1 + 2\alpha \mu - \alpha^2 L^2)} \left( \frac{1}{2\alpha} + \mu \right) \| z_0 - z_* \|^2 = \frac{1 + 2\alpha \mu}{\alpha^2 (1 + 2\alpha \mu - \alpha^2 L^2)} \| z_0 - z_* \|^2.
\]
Taking \( N \to \infty \), we conclude that (19) holds.

**Nonnegativity of \( V_k \).** Define
\[
W_k = V_k - \left( \frac{1}{2\alpha} + \mu \right) \| z_0 - z_* \|^2 = p_k \| Bz_k \|^2 + q_k (Bz_k - \mu(z_k - z_0), z_k - z_0),
\]
which is a shifted version of \( V_k \). Because \( B \) is \( \mu \)-strongly monotone, we have
\[
\langle Bz_k - \mu(z_k - z_*), z_k - z_* \rangle \geq 0.
\]
Subtracting \( \lambda \geq 0 \) times the above inequality from \( W_k \) gives
\[
W_k \geq p_k \| Bz_k \|^2 + q_k (Bz_k - \mu(z_k - z_0), z_k - z_0) - \lambda (Bz_k - \mu(z_k - z_*), z_k - z_*)
\]
\[
= p_k \| Bz_k \|^2 + (Bz_k, q_k(z_k - z_0) - \lambda(z_k - z_*)) - q_k \mu \| z_k - z_0 \|^2 + \lambda \mu \| z_k - z_* \|^2
\]
\[
= p_k \| Bz_k \|^2 + (Bz_k, (q_k - \lambda)(z_k - z_0)) + \mu (\lambda - q_k) \| z_k \|^2 + 2\mu (z_k, q_k z_0 - \lambda z_* - q_k \mu z_0)^2 + \lambda \mu \| z_* \|^2 \tag{28}
\]
Observe that if \( \lambda > q_k \), we can rewrite the terms not involving \( Bz_k \) in (28) as
\[
\mu (\lambda - q_k) \| z_k \|^2 + 2\mu (z_k, q_k z_0 - \lambda z_* - q_k \mu z_0)^2 + \lambda \mu \| z_* \|^2
\]
\[
= \frac{\mu \lambda - q_k}{\lambda - q_k} \| z_k + q_k z_0 - \lambda z_* \|^2 - \frac{\mu}{\lambda - q_k} \| q_k z_0 - \lambda z_* \|^2 - q_k \mu \| z_0 \|^2 + \lambda \mu \| z_* \|^2
\]
\[
= \frac{\mu \lambda - q_k}{\lambda - q_k} \| z_k + q_k z_0 - \lambda z_* \|^2 - \frac{\mu}{\lambda - q_k} \| q_k z_0 - \lambda z_* \|^2 + \frac{2q_k \lambda \mu}{\lambda - q_k} (z_0, z_*)
\]
\[
= \frac{\mu}{\lambda - q_k} \| (\lambda - q_k) z_k + q_k z_0 - \lambda z_* \|^2 - \frac{q_k \lambda \mu}{\lambda - q_k} \| z_0 - z_* \|^2.
\]
Therefore, by adding \( \frac{q_k \lambda \mu}{\lambda - q_k} \| z_0 - z_* \|^2 \) to both sides of the inequality (28), we obtain

\[
W_k + \frac{q_k \lambda \mu}{\lambda - q_k} \| z_0 - z_* \|^2 \geq p_k \| B z_k \|^2 + (B z_k, (q_k - \lambda) z_k - q_k z_0 + \lambda z_*)
\]

\[
+ \frac{\mu}{\lambda - q_k} \left( \frac{\lambda - q_k}{4 \mu} \right) \| B z_k \|^2
\]

\[
= \left( p_k - \frac{\lambda - q_k}{4 \mu} \right) \| B z_k \|^2 + \frac{\mu}{\lambda - q_k} \left( \frac{\lambda - q_k}{2 \mu} \right) \| B z_k - (\lambda - q_k) z_k - q_k z_0 + \lambda z_* \|^2
\]

(29)

provided that \( \lambda > q_k \). In particular, taking \( \lambda = q_k + 4 \mu p_k \) in (29), we conclude that

\[
0 \leq \frac{1}{4 p_k} \| 2 p_k B z_k - 4 \mu p_k z_k - q_k z_0 + (q_k + 4 \mu p_k) z_* \|^2
\]

\[
\leq W_k + \left( \frac{q_k^2}{4 p_k} + \mu q_k \right) \| z_0 - z_* \|^2
\]

\[
= W_k + \left( \frac{1}{2 \alpha (1 + 2 \alpha \mu)^{k-1}} + \frac{\mu (1 + 2 \alpha \mu)^k - 1}{2 \alpha \mu (1 + 2 \alpha \mu)^{k-1}} \right) \| z_0 - z_* \|^2
\]

\[
= W_k + \frac{1 + 2 \alpha \mu}{2 \alpha} \| z_0 - z_* \|^2 = V_k,
\]

where in the third line we use (24) and \( q_k = \frac{1}{(1 + 2 \alpha \mu)^{k-1} \beta_{k-1}} = \frac{1}{(1 + 2 \alpha \mu)^{k-1}} \).\( \frac{1 + 2 \alpha \mu}{2 \alpha} \). \( \frac{1}{\beta_{k-1}} \)

### C.3 Proof of Theorem 5

Recall the shifted version \( W_k \) of \( V_k \), defined within the proof of Lemma 11. Because the subtracted term \( \left( \frac{1}{2 \alpha \mu} + \mu \right) \| z_0 - z_* \|^2 \) does not depend on \( k \), we immediately see that \( W_k \) inherits nonincreasingness from \( V_k \), i.e., \( W_k \leq \cdots \leq W_0 = 0 \). Therefore, from (29), we have

\[
\frac{q_k \lambda \mu}{\lambda - q_k} \| z_0 - z_* \|^2 \geq W_k + \frac{q_k \lambda \mu}{\lambda - q_k} \| z_0 - z_* \|^2 \geq \left( p_k - \frac{\lambda - q_k}{4 \mu} \right) \| B z_k \|^2
\]

\[
\iff \| B z_k \|^2 \leq \left( p_k - \frac{\lambda - q_k}{4 \mu} \right)^{-1} \frac{q_k \lambda \mu}{\lambda - q_k} \| z_0 - z_* \|^2
\]

(30)

provided that \( q_k < \lambda < q_k + 4 \mu p_k \). One can straightforwardly verify that

\[
\lambda_* = (q_k (q_k + 4 \mu p_k))^{1/2}
\]

(31)

lies between \( q_k \) and \( q_k + 4 \mu p_k \), and minimizes the factor \( \left( p_k - \frac{\lambda - q_k}{4 \mu} \right)^{-1} \frac{q_k \lambda \mu}{\lambda - q_k} \) in the convergence rate (30). Plugging in the explicit expressions

\[
q_k = \frac{(1 + 2 \alpha \mu)^k - 1}{2 \alpha \mu (1 + 2 \alpha \mu)^{k-1}}
\]

\[
p_k = \frac{\alpha}{2} \left( (1 + 2 \alpha \mu)^{k-1} q_k^2 = \frac{(1 + 2 \alpha \mu)^k - 1)^2}{8 \alpha \mu^2 (1 + 2 \alpha \mu)^{k-1}}
\]

into (31) (where the second line uses (24)), we obtain

\[
4 \mu p_k + q_k = \frac{(1 + 2 \alpha \mu)(1 + 2 \alpha \mu)^k - 1)}{2 \alpha \mu} = (1 + 2 \alpha \mu)^k q_k
\]

\[
\lambda_* = (1 + 2 \alpha \mu)^{1/2} q_k
\]
and therefore,
\[
\left(\frac{p_k - \lambda_* - q_k}{4\mu}\right)^{-1} q_k \frac{\lambda_*}{\lambda_* - q_k} = \frac{4\mu^2}{(1 + 2\alpha\mu)^2 - 1} = \frac{4\mu^2}{(\sqrt{1 + 2\alpha\mu} - 1)^2 \left(\sum_{j=0}^{k-1}(1 + 2\alpha\mu)^{\frac{j}{2}}\right)^2} = \frac{\alpha^2}{\alpha^2 \left(\sum_{j=0}^{k-1}(1 + 2\alpha\mu)^{\frac{j}{2}}\right)^2},
\]
which concludes the proof.

D Omitted proofs for Section 5

D.1 Proof of Theorem 6

We first state and prove some preliminary results. The following Lemma 12, which characterizes the forward-backward residual, has been used in [92] for developing accelerated splitting algorithms.

Lemma 12 ([92, Lemma 3]). Suppose that \(A, B\) are maximal monotone operators on \(\mathbb{R}^d\), and \(B\) is single-valued. For \(\alpha > 0\), let \(G_\alpha(v) = \frac{1}{\alpha} (v - J_{\alpha A}(v - \alpha Bv))\). Then for any \(v, w \in \mathbb{R}^d\),
\[
\langle G_\alpha(v) - G_\alpha(w), (I + \alpha B)(v) - (I + \alpha B)(w) \rangle \geq \alpha \|G_\alpha(v) - G_\alpha(w)\|^2.
\]

Proof. Observe that
\[
G_\alpha(v) = M_{\alpha A}(v - \alpha Bv) + Bv,
\]
where \(M_{\alpha A}(u) := \frac{1}{\alpha}(u - J_{\alpha A}(u))\) of \(A\) is the Moreau–Yosida approximation of \(A\). It is known that \(M_{\alpha A}\) is \(\alpha\)-cocoercive [7, Corollary 23.11], and in particular, for \(\bar{v} = v - \alpha Bv\) and \(\bar{w} = w - \alpha Bw\), we have
\[
\alpha \|M_{\alpha A}(\bar{v}) - M_{\alpha A}(\bar{w})\|^2 \
= \langle M_{\alpha A}(\bar{v}) - M_{\alpha A}(\bar{w}), \bar{v} - \bar{w} \rangle \
= \langle (G_\alpha(v) - G_\alpha(w)) - (Bv - Bw), (v - w) - \alpha(Bv - Bw) \rangle.
\]

Expanding and rearranging the terms, and using monotonicity of \(B\), we get
\[
\alpha \|G_\alpha(v) - G_\alpha(w)\|^2 \leq \langle G_\alpha(v) - G_\alpha(w), (v - w) + \alpha(Bv - Bw) \rangle - \langle Bv - Bw, v - w \rangle
\leq \langle G_\alpha(v) - G_\alpha(w), (v - w) + \alpha(Bv - Bw) \rangle
\]
as desired. \(\square\)

Lemma 13. Let \(\{(u_k, w_k)\}_{k=0,1,...}\) be the iterates of OHM-DRS and \(\{(\xi_k, z_k)\}_{k=0,1,...}\) be the iterates of APG*. If \(u_0 = \xi_0\), then for \(k = 1, 2, \ldots\), we have
\[
(k + 1)\|\xi_k - u_k\| \leq k\|\xi_{k-1} + \alpha Bz_{k-1} - u_{k-1}\| \tag{32}
\]
and
\[
(k + 1)\|z_k + \alpha Bz_k - u_k\| \leq k\|\xi_{k-1} + \alpha Bz_{k-1} - u_{k-1}\| + (k + 1)\epsilon_k. \tag{33}
\]
where each \(\epsilon_k\) is an upper bound on \(\|z_k + \alpha Bz_k - \xi_k\|\).

Proof. Observe that \(J_{\alpha A}(z - \alpha Bz) + \alpha Bz = z + \alpha Bz - \alpha G_\alpha(z)\) for any \(z \in \mathbb{R}^d\), so
\[
\xi_k = \frac{1}{k + 1}\xi_0 + \frac{k}{k + 1}(z_{k-1} + \alpha Bz_{k-1} - \alpha G_\alpha(z_{k-1})).
\]
Similarly, we have
\[ u_k = \frac{1}{k+1} u_0 + \frac{k}{k+1} (w_{k-1} + \alpha B w_{k-1} - \alpha G_{\alpha}(w_{k-1})). \]
Therefore, provided that \( u_0 = \xi_0 \),
\[ \xi_k - u_k = \frac{k}{k+1} \left( (I + \alpha B)(z_{k-1}) - (I + \alpha B)(w_{k-1}) - \alpha \left( G_{\alpha}(z_{k-1}) - G_{\alpha}(w_{k-1}) \right) \right), \]
which gives
\[
\|\xi_k - u_k\|^2 = \left( \frac{k}{k+1} \right)^2 \left\| (I + \alpha B)(z_{k-1}) - (I + \alpha B)(w_{k-1}) - \alpha \left( G_{\alpha}(z_{k-1}) - G_{\alpha}(w_{k-1}) \right) \right\|^2 \\
= \left( \frac{k}{k+1} \right)^2 \left\| (I + \alpha B)(z_{k-1}) - (I + \alpha B)(w_{k-1}) \right\|^2 + \alpha^2 \left\| G_{\alpha}(z_{k-1}) - G_{\alpha}(w_{k-1}) \right\|^2 \\
- 2\alpha \,( (I + \alpha B)(z_{k-1}) - (I + \alpha B)(w_{k-1}), G_{\alpha}(z_{k-1}) - G_{\alpha}(w_{k-1})) \\
\leq \left( \frac{k}{k+1} \right)^2 \left\| (I + \alpha B)(z_{k-1}) - (I + \alpha B)(w_{k-1}) \right\|^2 + \alpha^2 \left\| G_{\alpha}(z_{k-1}) - G_{\alpha}(w_{k-1}) \right\|^2 \\
= \left( \frac{k}{k+1} \right)^2 \left\| (I + \alpha B)(z_{k-1}) - u_{k-1} \right\|^2 \\
\end{align*}
where the first inequality follows from Lemma 12 and the last equality from \( w_{k-1} = J_{\alpha B}(u_{k-1}). \)
This proves (32). Then (33) follows from
\[
\|z_k + \alpha B z_k - u_k\| \leq \|z_k + \alpha B z_k - \xi_k\| + \|\xi_k - u_k\| \\
\leq \epsilon_k + \frac{k}{k+1} \|(z_{k-1} + \alpha B z_{k-1}) - u_{k-1}\|.
\]

We now prove Theorem 6.

**Proof of Theorem 6.** Let \( M = 1 + \frac{1}{2} \|B\| \|\xi_0\| \), so that \( \epsilon_j = \frac{M}{(j+1)(j+2)} \) and
\[ M = \sum_{j=0}^{\infty} (j+1)\epsilon_j, \]

Then by Lemma 13, for each \( k = 1, 2, \ldots, \)
\[
(k+1)\|\xi_k - u_k\| \leq k\|z_{k-1} + \alpha B z_{k-1} - u_{k-1}\| \\
\leq (k-1)\|z_{k-2} + \alpha B z_{k-2} - u_{k-2}\| + k\epsilon_k \\
\leq \cdots \leq \|z_0 + \alpha B z_0 - u_0\| + \sum_{j=1}^{k} (j+1)\epsilon_j \\
\leq \sum_{j=0}^{k} (j+1)\epsilon_j \leq M.
\]
This shows that \( \|\xi_k - u_k\|^2 \leq \left( \frac{M}{k+1} \right)^2 \) for \( k = 0, 1, \ldots \) (note that \( \|\xi_0 - u_0\| = 0 \)). Additionally, for each \( k = 0, 1, \ldots \) we have \( (k+1)\|z_k + \alpha B z_k - u_k\| \leq M \), and because \( J_{\alpha B} = (I + \alpha B)^{-1} \) is nonexpansive, we have
\[
(k+1)\|z_k - u_k\| = (k+1)\|J_{\alpha B} \circ (I + \alpha B)(z_k) - J_{\alpha B}(u_k)\| \\
\leq (k+1)\|z_k + \alpha B z_k - u_k\| \leq M,
\]
26
We first state and prove some lemmas, which will be useful for the main proof.

D.2 Proof of Corollary 7

As $T_{\text{DRS}}$ is nonexpansive [53], by the discussion of Section A.4, we have $u_k \rightarrow \Pi_{\text{Fix}(T_{\text{DRS}})}(u_0) = \Pi_{\text{Fix}(T_{\text{DRS}})}(\xi_0) = \xi_*$. Because $J_{\alpha \beta}$ is nonexpansive, this implies $z_k = J_{\alpha \beta}(u_k) \rightarrow J_{\alpha \beta}(\xi_*) \in \text{Zer } (A + B)$. Due to the MP property proved in Theorem 6, we immediately obtain $\xi_k \rightarrow \xi_*$ and $z_k \rightarrow J_{\alpha \beta}(\xi_*)$.

Next, note that Lemma 12 implies that $G_\alpha$ is $(\frac{1 + \alpha L}{\alpha})$-Lipschitz, because

\[ \alpha \| G_\alpha(v) - G_\alpha(w) \|^2 \leq \langle G_\alpha(v) - G_\alpha(w), (I + \alpha B)(v) - (I + \alpha B)(w) \rangle \]

\[ \leq \| G_\alpha(v) - G_\alpha(w) \| \| (I + \alpha B)(v) - (I + \alpha B)(w) \| \]

\[ \leq \| G_\alpha(v) - G_\alpha(w) \| (1 + \alpha L) \| v - w \| \]

for any $v, w \in \mathbb{R}^d$. Therefore, we obtain

\[ \| G_\alpha(z_k) \|^2 = \| G_\alpha(w_k) + (G_\alpha(z_k) - G_\alpha(w_k)) \|^2 \]

\[ \leq (1 + \nu) \| G_\alpha(w_k) \|^2 + \left( 1 + \frac{1}{\nu} \right) \| G_\alpha(z_k) - G_\alpha(w_k) \|^2 \]  \hspace{1cm} (34)

\[ \leq (1 + \nu) \frac{4\| \xi_0 - \xi_* \|^2}{\alpha^2 (k + 1)^2} + \left( 1 + \frac{1}{\nu} \right) \frac{1 + \alpha L}{\alpha} \| z_k - w_k \|^2 \]  \hspace{1cm} (35)

\[ \leq (1 + \nu) \frac{4C(\xi_0)^2}{\alpha^2 L^2 (k + 1)^2} + \left( 1 + \frac{1}{\nu} \right) \frac{1 + \alpha L}{\alpha} \frac{C(\xi_0)^2}{L^2 (k + 1)^2} \]  \hspace{1cm} (36)

\[ = \frac{C(\xi_0)^2}{\alpha^2 L^2 (k + 1)^2} \left( 4(1 + \nu) + \left( 1 + \frac{1}{\nu} \right) (1 + \alpha L) \right) \]  \hspace{1cm} (37)

where $\nu > 0$ is arbitrary, (34) uses Young’s inequality, (35) uses the convergence rate of OHM (from Section A.4) and Lipschitzness of $G_\alpha$, and (36) uses $L\| \xi_0 - \xi_* \| \leq C(\xi_0)$ and Theorem 6. Now choosing $\nu = \frac{1 + \alpha L}{2}$, which minimizes the quantity (37), we obtain

\[ \| G_\alpha(z_k) \|^2 \leq \frac{C(\xi_0)^2}{\alpha^2 L^2 (k + 1)^2} \left( 4 + (1 + \alpha L)^2 + 4(1 + \alpha L) \right) = \frac{(3 + \alpha L)^2 C(\xi_0)^2}{\alpha^2 L^2 (k + 1)^2}, \]

which completes the proof.

D.3 Proof of Corollary 8

We first state and prove some lemmas, which will be useful for the main proof.

Lemma 14. Let $B : \mathbb{R}^d \rightarrow \mathbb{R}^d$ be a monotone, $L$-Lipschitz operator, and let $\alpha \in (0, \frac{1}{2L})$. Then for any $u \in \mathbb{R}^d$,

\[ \| u - J_{\alpha \beta}(u) \| \leq \frac{\alpha}{1 - \alpha L} \| Bu \|. \]

Proof. Let $v = J_{\alpha \beta}(u)$. Then $u = v + \alpha Bv$, so

\[ \| Bu \| - \| Bu \| \leq \| Bv - Bu \| \leq L \| v - u \| = \alpha L \| Bv \| \implies \| Bv \| \leq \frac{1}{1 - \alpha L} \| Bu \|. \]

Therefore,

\[ \| u - J_{\alpha \beta}(u) \| = \| u - v \| = \alpha \| Bv \| \leq \frac{\alpha}{1 - \alpha L} \| Bu \|. \]
Lemma 15. Let $\xi_* \in \text{Fix } T_{\text{DRS}}$ and $z_* = J_{\alpha B}(\xi_*)$. $\epsilon_k > 0 \quad (k = 0, 1, \ldots)$ satisfy

$$\sum_{k=0}^{\infty} (k+1)\epsilon_k = M < \infty.$$ 

Then APG* iterates $\xi_k, z_k$ satisfy $\max \{\|\xi_k - \xi_*\|, \|z_k - z_*\|\} \leq \|\xi_0 - \xi_*\| + M$ for $k = 0, 1, \ldots$.

Proof. Let $u_k, w_k$ be the iterates of OHM-DRS with $u_0 = \xi_0$. Because $\xi_* \in \text{Fix}(T_{\text{DRS}})$ and $T_{\text{DRS}}$ is nonexpansive, we have

$$\|u_{k+1} - \xi_*\| = \|\beta_k u_k + (1 - \beta_k) T_{\text{DRS}}(u_k) - \xi_*\|
\leq \beta_k \|u_k - \xi_*\| + (1 - \beta_k) \|T_{\text{DRS}}(u_k) - \xi_*\|
\leq \beta_k \|\xi_0 - \xi_*\| + (1 - \beta_k) \|u_k - \xi_*\|$$

for $k = 0, 1, \ldots$. Therefore, by induction on $k$, $\|u_k - \xi_*\| \leq \|\xi_0 - \xi_*\|$ for all $k = 0, 1, \ldots$. Then by Theorem 6,

$$\|\xi_k - \xi_*\| \leq \|\xi_k - u_k\| + \|u_k - \xi_*\| \leq \frac{M}{k+1} + \|\xi_0 - \xi_*\| \leq \|\xi_0 - \xi_*\| + M$$

and

$$\|z_k - z_*\| \leq \|z_k - w_k\| + \|w_k - z_*\| \leq \frac{M}{k+1} + ||J_{\alpha B}(u_k) - J_{\alpha B}(\xi_*)||
\leq \frac{M}{k+1} + \|u_k - \xi_*\| \leq \|\xi_0 - \xi_*\| + M$$

for $k = 0, 1, \ldots$. \hfill \Box

We now prove Corollary 8.

Proof of Corollary 8. Recall that each outer loop of APG* performs a single evaluation of $J_{\alpha A}$, a forward evaluation $Bz_k$, and an inner loop, which starts with $z_k^{(0)} = \xi_k$ and solves

$$\text{find } z \in \mathbb{R}^d \quad 0 = z + \alpha Bz - \xi_k$$

using SM-EAG+. The distance from the initial point $z_k^{(0)}$ to the true solution $J_{\alpha B}(\xi_k)$ of (38) is

$$d_k^{(0)} := \|z_k^{(0)} - J_{\alpha B}(\xi_k)\| = \|\xi_k - J_{\alpha B}(\xi_k)\| \leq \frac{\alpha}{1 - \alpha L} \|B\xi_k\|$$

where the inequality holds due to Lemma 14. Observe that by $L$-Lipschitzness of $B$ and Lemma 15, we have

$$\|B\xi_k\| \leq \|B\xi_*\| + \|B(z_k - B\xi_*)\| \leq \|B\xi_*\| + L\|z_k - \xi_*\| \leq \|B\xi_*\| + L\|\xi_0 - \xi_*\| + LM,$$

so $d_k^{(0)} \leq \frac{\alpha}{1 - \alpha L} (\|B\xi_*\| + L\|\xi_0 - \xi_*\| + LM)$. Because the inner loop objective (38) is 1-strongly monotone and $(\alpha L + 1)$-Lipschitz, by Theorem 5, the total number of inner loop iterations needed to achieve the error $\|z + \alpha Bz - \xi_k\| \leq \epsilon_k$ is

$$\mathcal{O} \left( (\alpha L + 1) \log \frac{d_k^{(0)}}{\epsilon_k} \right) = \mathcal{O} \left( \frac{d_k^{(0)}}{\epsilon_k} \right) = \mathcal{O} \left( \frac{\log k \left( \|B\xi_0\| + 2L\|\xi_0 - \xi_*\| + LM \right)}{LM} \right)$$

where the last $\mathcal{O}$ expression holds because $\alpha L \in (0, 1)$ is a constant. Now recalling that $M = 1 + \frac{1}{L} \|B\xi_0\|$, we obtain

$$\log \frac{k \left( \|B\xi_0\| + 2L\|\xi_0 - \xi_*\| + LM \right)}{LM} \leq \log \left( k \left( \frac{\|B\xi_0\| + 2\|\xi_0 - \xi_*\|}{LM} + 1 \right) \right)
\leq \log (k (1 + 2\|\xi_0 - \xi_*\| + 1)) = \mathcal{O}(\log(k(\|\xi_0 - \xi_*\| + 1))).$$
We first state a simple lemma, which will be useful within the proofs of Lemmas 9 and 10 for arguing

\[ H(0) = H(N)(r)k^N + \cdots + H(1)(r)k + H(0). \]

If \( H(N)(0) > 0 \) and \( H(1)(0) \geq 0 \) for all \( \ell = 0, \ldots, N - 1 \), then there exists \( \eta > 0 \) such that \( H(r, k) > 0 \) for all \( r \in (0, \eta) \) and \( k = 1, 2, \ldots \).

**Proof.** Let \( J = \{ \ell \in \{0, \ldots, N-1\} \mid H^{(\ell)}(0) = 0 \} \) (so \( H^{(\ell)}(r) > 0 \) for \( \ell \in \{0, \ldots, N-1\} \setminus J \)). As \( H^{(\ell)}(r) \) are continuous, one can choose \( \eta > 0 \) small enough so that \( |r| < \eta \) implies

\[
\sum_{\ell \in J} |H^{(\ell)}(r)| < \frac{H^{(N)}(0)}{2}
\]

and \( H^{(\ell)}(r) > 0 \) for all \( \ell \in \{0, \ldots, N-1\} \setminus J \). Then for \( k = 1, 2, \ldots \), we have

\[
H(r, k) = k^N \left( H^{(N)}(0) + \sum_{\ell=0}^{N-1} H^{(\ell)}(r) k^{\ell-N} \right)
> k^N \left( H^{(N)}(0) + \sum_{\ell \in J} H^{(\ell)}(r) k^{\ell-N} \right)
\geq k^N \left( H^{(N)}(0) - \sum_{\ell \in J} |H^{(\ell)}(r)| k^{\ell-N} \right)
\geq k^N \left( H^{(N)}(0) - \sum_{\ell \in J} |H^{(\ell)}(r)| \right)
> \frac{H^{(N)}(0)k^N}{2} > 0
\]

provided that \( r \in (0, \eta) \). \qed

**Proof of Lemma 9.** We show the given statement by establishing the following result: by defining the sequences \( p_k, q_k \) by \( q_k = k \) \( (k = 0, 1, \ldots) \) and

\[
p_0 = \frac{\alpha(1 - 3\alpha L + 6\alpha^2 L^2 - 2\alpha^4 L^4)}{2(1 - \alpha L)^2(1 + \alpha L)(2 + \alpha L)}
\]

\[
p_k = \frac{\alpha(k+1)(k + \alpha L(k-1))}{2(1 + \alpha L)}, \quad k = 1, 2, \ldots.
\]
we have $p_0 > 0$,
\[
V_k = p_k \|Bz_k\|^2 + q_k (Bz_k, z_k - z_0) + \frac{1}{2\alpha} \|z_0 - z_*\|^2 \geq 0,
\]
and
\[
V_k - V_{k+1} \geq \epsilon (k + 1)^2 \|Bz_k - Bz_{k+1/2}\|^2
\]
for $\alpha \in (0, \frac{\eta}{2})$, $\epsilon = \frac{\alpha^2 L(1 - \alpha^2 L^2)}{2}$ with $\eta \in (0, 1)$ sufficiently small. Once this is shown, we can conclude that for any positive integer $N$,
\[
\frac{\alpha^2 L(1 - \alpha^2 L^2)}{2} \sum_{k=0}^{N} (k + 1)^2 \|Bz_k - Bz_{k+1/2}\|^2 \leq V_0 - V_{N+1}
\]
holds, which gives
\[
\sum_{k=0}^{N} (k + 1)^2 \|Bz_k - Bz_{k+1/2}\|^2 \leq \frac{2}{\alpha^2 L(1 - \alpha^2 L^2)} \left( p_0 L^2 + \frac{1}{2\alpha} \|z_0 - z_*\|^2 \right)
\]
where
\[
C = \frac{2 - r - 2r^2 - 2r^3 + 7r^4 - 2r^6}{r (1 - r)^3 (1 + r)^2 (2 + r)} \quad (r = \alpha L)
\]
is a positive constant provided that $r < \eta$ for $\eta$ small enough.

**Case $k = 0$.** We have $z_{1/2} = z_0 - \alpha Bz_0$ and $z_1 = z_0 - \alpha Bz_{1/2}$. By monotonicity and Lipschitzness of $B$, we have
\[
V_1 + \frac{\alpha^2 L(1 - \alpha^2 L^2)}{2} \|Bz_0 - Bz_{1/2}\|^2
\]
\[
= \alpha \|Bz_1\|^2 + \langle Bz_1, z_1 - z_0 \rangle + \frac{1}{2\alpha} \|z_0 - z_*\|^2 + \frac{\alpha^2 L(1 - \alpha^2 L^2)}{2} \|Bz_0 - Bz_{1/2}\|^2 \quad (39)
\]
\[
\leq \alpha \|Bz_1\|^2 + \langle Bz_1, z_1 - z_0 \rangle + \frac{1}{2\alpha} \|z_0 - z_*\|^2 + \frac{\alpha^2 L(1 - \alpha^2 L^2)}{2} \|Bz_0 - Bz_{1/2}\|^2
\]
\[
+ \frac{\alpha}{1 + \alpha L} \left[ L^2 (z_{1/2} - z_1)^2 - \|Bz_{1/2} - Bz_1\|^2 \right] + \frac{1 - \alpha L}{1 + \alpha L} \langle z_0 - z_1, Bz_0 - Bz_1 \rangle
\]
\[
= \frac{\alpha}{1 + \alpha L} \|Bz_1\|^2 + \langle Bz_1, -\alpha Bz_{1/2} \rangle + \frac{1}{2\alpha} \|z_0 - z_*\|^2 + \frac{\alpha^2 L(1 - \alpha^2 L^2)}{2} \|Bz_0 - Bz_{1/2}\|^2
\]
\[
+ \frac{\alpha}{1 + \alpha L} \left[ \alpha^2 L^2 \|Bz_0 - Bz_{1/2}\|^2 - \|Bz_{1/2} - Bz_1\|^2 \right] + \frac{1 - \alpha L}{1 + \alpha L} \langle \alpha Bz_{1/2}, Bz_0 - Bz_1 \rangle
\]
\[
= \frac{\alpha}{2(1 + r)} \left( r(1 + 3r - r^2 - r^3) \right) \|Bz_0\|^2 - (1 - r)^2 (1 + r)(2 + r)\|Bz_{1/2}\|^2
\]
\[
+ 2(1 + r)(1 - 3r + r^3) \langle Bz_0, Bz_{1/2} \rangle + \frac{1}{2\alpha} \|z_0 - z_*\|^2 \quad (40)
\]
where $r = \alpha L$. Next, using Young’s inequality, we can bound the inner product term in (40) as
\[
2(1 + r)(1 - 3r + r^3) \langle Bz_0, Bz_{1/2} \rangle
\]
\[
\leq (1 - r)^2 (1 + r)(2 + r) \|Bz_{1/2}\|^2 + \frac{(1 + r)(1 - 3r + r^3)^2}{(1 - r)^2 (1 + r)(2 + r)} \|Bz_0\|^2.
\]
Plugging this back into (40) and rearranging, we obtain
\[
V_1 + \frac{\alpha^2 L(1 - \alpha^2 L^2)}{2} \|Bz_0 - Bz_{1/2}\|^2 \leq \frac{\alpha(1 - 3r + 6r^2 - 2r^4)}{2(1 + r)(1 - r)(2 + r)} \|Bz_0\|^2 + \frac{1}{2\alpha} \|z_0 - z_*\|^2 = V_0.
\]
Clearly, if $r$ is small enough, $p_0 = \frac{\alpha(1 - 3r + 6r^2 - 2r^4)}{2(1 - r)^2 (1 + r)(2 + r)} > 0$.  

30
Case $k \geq 1$. From the definition of the EAG iterates, we have
\[
\begin{align*}
z_{k+1} - z_0 &= (1 - \beta_k)(z_k - z_0) - \alpha B z_{k+1/2} \\
z_k - z_{k+1} &= \beta_k (z_k - z_0) + \alpha B z_{k+1/2} \\
z_{k+1/2} - z_k &= \alpha B z_{k+1/2} - \beta_k B z_k
\end{align*}
\]
Therefore, as in previous sections, we can lower-bound $V_k - V_{k+1}$ as
\[
V_k - V_{k+1} \geq \frac{q_k}{\beta_k} (z_k - z_{k+1}) - \frac{\alpha q_k}{\beta_k^2} (z_k, B z_{k+1/2}) + \alpha q_k (B z_{k+1/2}, B z_k) - p_{k+1} \|B z_{k+1}\|^2
\]
As $B$ is $L$-Lipschitz,
\[
0 \leq L^2 \|z_{k+1/2} - z_k\|^2 - \|B z_{k+1/2} - B z_k\|^2
\]
Combining this with (41), we obtain
\[
V_k - V_{k+1} - \epsilon (k+1)^2 \|B z_k - B z_{k+1/2}\|^2 \\
\geq \frac{p_k}{\beta_k} \|B z_k\|^2 - \alpha k (k + 1) (B z_k, B z_{k+1/2}) + \alpha (k + 1)^2 (B z_{k+1/2}, B z_k) - p_{k+1} \|B z_{k+1}\|^2
\]
\[
- \epsilon (k+1)^2 \|B z_k - B z_{k+1/2}\|^2 - \tau (\alpha^2 L^2 \|B z_{k+1/2} - B z_k\|^2 - \|B z_{k+1/2} - B z_k\|^2)
\]
\[
= (p_k - \alpha^2 L^2 \tau - \epsilon (k+1)^2) \|B z_k\|^2 + (\tau (1 - \alpha^2 L^2) - \epsilon (k+1)^2) \|B z_{k+1/2}\|^2 + (\tau - p_{k+1}) \|B z_{k+1}\|^2
\]
\[
+ (2 \alpha^2 L^2 \tau + 2 \epsilon (k+1)^2 - \alpha (k+1)) \langle B z_k, B z_{k+1/2} \rangle + (\alpha (k+1)^2 - 2 \tau) \langle B z_{k+1/2}, B z_k \rangle
\]
\[
= \tau \text{Tr} (M_k S_k M_k^T),
\]
where $\tau > 0$, $M_k := [B z_k, B z_{k+1/2}, B z_{k+1}]$ and
\[
S_k = \begin{bmatrix}
p_k - \alpha^2 L^2 \tau - \epsilon (k+1)^2 & \alpha^2 L^2 \tau + \epsilon (k+1)^2 - \frac{1}{2} k (k+1) & 0 \\
\alpha^2 L^2 \tau + \epsilon (k+1)^2 - \frac{1}{2} k (k+1) & \tau (1 - \alpha^2 L^2) - \epsilon (k+1)^2 & \frac{1}{2} \alpha (k+1)^2 - \tau \\
0 & \frac{1}{2} \alpha (k+1)^2 - \tau & \tau - p_{k+1}
\end{bmatrix}
\]
Now, we choose $\tau$ as
\[
\tau = \frac{(k+1) (k(1 - \alpha L + \alpha^2 L^2 + \alpha^3 L^3) - \alpha L (2 - \alpha L - \alpha^2 L^2))}{4 \alpha L^2}
\]
For notational simplicity, let $\alpha L = r$. Then by direct computation we obtain
\[
s_{11} = \frac{\alpha}{4 (1 + r)} \left( (1 - r^2)^2 k^2 + (1 - 2r - 3r^2 + 2r^4) k - r^2 (2 + r - r^3) \right)
\]
\[
s_{12} = - \frac{\alpha (1 - r)(k+1) ((1 - r^2) k - r^2)}{4}
\]
By Lemma 16, $s_{11} > 0$ for sufficiently small $r$, and in that case, Young’s inequality gives
\[
s_{11} \|B z_k\|^2 + 2 s_{12} \langle B z_k, B z_{k+1/2} \rangle + \frac{s_{12}^2}{s_{11}} \|B z_{k+1/2}\|^2 \geq 0.
\]
Continuing on, we compute
\[
t_{22} = s_{22} - \frac{s_{12}^2}{s_{11}}
\]
\[
= \frac{(1 - r^2) T_{22}^{(3)}(r) k^3 + T_{22}^{(4)}(r) k^2 + T_{22}^{(1)}(r) (k + T_{22}^{(0)}(r))}{4 L r (1 - r^2)^2 k - r (2 + r - r^3)}
\]
where $T_{22}^{(\ell)}(r)$, $\ell = 0, 1, 2, 3,$ are polynomials in $r$ defined by

\[
T_{22}^{(0)}(r) = r^2(4 + r^2 - r^3)
\]

\[
T_{22}^{(1)}(r) = -r(4 - 6r - 2r^2 - 3r^3 + 3r^4)
\]

\[
T_{22}^{(2)}(r) = 1 - 5r + 4r^3 + 3r^4 - 3r^5
\]

\[
T_{22}^{(3)}(r) = (1 + r)^2(1 - r)^3,
\]

and

\[
s_{23} = \frac{(1 + r)(k + 1)((1 - r)^2 k - r(2 - r))}{4Lr}
\]

By Lemma 16 again, $t_{22} > 0$ for sufficiently small $r$ and then Young’s inequality implies that

\[
t_{22} ||Bz_{k+1/2}||^2 + 2s_{23}(Bz_{k+1/2}, Bz_{k+1}) + \frac{s_{23}^2}{t_{22}} ||Bz_{k+1}||^2 \geq 0.
\]

Finally, we have

\[
\text{Tr}(M_k S_k M_k^T)
\]

\[
= \left(s_{11} ||Bz_k||^2 + 2s_{12}(Bz_k, Bz_{k+1/2}) + \frac{s_{12}^2}{s_{11}} ||Bz_{k+1/2}||^2\right)
\]

\[
+ \left(t_{22} ||Bz_{k+1/2}||^2 + 2s_{23}(Bz_{k+1/2}, Bz_{k+1}) + \frac{s_{23}^2}{t_{22}} ||Bz_{k+1}||^2\right) + \left(s_{33} - \frac{s_{33}^2}{t_{22}} \right) ||Bz_{k+1}||^2
\]

\[
\geq 0,
\]

as the following computation and Lemma 16 shows that $s_{33} - \frac{s_{33}^2}{t_{22}} > 0$ for sufficiently small $r$:

\[
s_{33} - \frac{s_{33}^2}{t_{22}} = \frac{\alpha r ((1 - 4r - 2r^2 + r^4) k - r(6 - r - r^3))}{2(1 - r^2)((1 + r)^2(1 - r)k - r(2 + r + r^2))}.
\]

Therefore, we have established $V_k - V_{k+1} \geq 0$ for small $r$.

**Nonnegativity.** Observe that

\[
p_k - \frac{\alpha k^2}{2} = \frac{\alpha(k + 1)(k + \alpha L(k - 1))}{2(1 + \alpha L)} - \frac{\alpha(1 + \alpha L)k^2}{2(1 + \alpha L)} = \frac{\alpha(k - \alpha L)}{2(1 + \alpha L)} > 0
\]

provided that $k \geq 1$ and $\alpha < \frac{1}{2}$. Thus,

\[
V_k \geq \frac{\alpha k^2}{2} ||Bz_k||^2 + k||Bz_k, z_k - z_0|| + \frac{1}{2\alpha} ||z_0 - z_*||^2
\]

\[
\geq \frac{\alpha k^2}{2} ||Bz_k||^2 + k||Bz_k, z_k - z_0|| + \frac{1}{2\alpha} ||z_0 - z_*||^2 \geq 0
\]

by monotonicity and Young’s inequality.

**Proof of Lemma 10.** We establish the following Lyapunov analysis: with same $p_k, q_k$ as in Lemma 9,

\[
V_k = p_k ||Bz_k||^2 + q_k(Bz_k, z_k - z_0) + p_k L^2 ||z_k - v_k||^2 + \frac{1}{2\alpha} ||z_0 - z_*||^2
\]

satisfies $V_k \geq 0$ and $V_k - V_{k+1} \geq \epsilon(k + 1)^2 ||Bv_k - Bv_{k+1}||^2$ for $k = 0, 1, \ldots, \alpha \in (0, \frac{1}{2})$, and

\[
\epsilon = \frac{\alpha^2 L^2 (1 - \alpha L - \alpha^2 L - \alpha^2 L^2)}{(1 + \alpha L)}
\]

with $\eta \in (0, 1)$ sufficiently small. This will imply

\[
\epsilon \sum_{k=0}^{N} (k + 1)^2 ||Bz_k - Bz_{k+1/2}||^2 \leq V_0 \leq \left(p_0 L^2 + \frac{1}{2\alpha}\right) ||z_0 - z_*||^2,
\]

32
so that
\[
\sum_{k=0}^{\infty} (k+1)^2 \| B v_k - B v_{k+1} \|^2 \leq \frac{1}{\epsilon} \left( p_0 L^2 + \frac{1}{2\alpha} \right) \| z_0 - z_* \|^2 = \frac{C}{\alpha^2} \| z_0 - z_* \|^2
\]
with
\[
C = \frac{2 - r - 2r^2 - 2r^3 + 7r^4 - 2r^6}{r(1 - r)^2(2 + r)(1 - r - r^2 - r^3)} > 0
\]
provided that \( r = \alpha L \in (0, \eta) \) for \( \eta \) small enough.

**Nonnegativity.** Proved in the exact same way as in Lemma 9, except that we have an additional nonnegative term \( p_k L^2 \| z_k - v_k \|^2 \).

**Case** \( k = 0 \). We have \( v_1 = z_0 - \alpha B v_0 = z_0 - \alpha B z_0 \) and \( z_1 = z_0 - \alpha B v_1 \). Observe that
\[
V_1 + \epsilon \| B v_0 - B v_1 \|^2 = \frac{\alpha}{1 + \alpha L} \| B z_1 \|^2 + \langle B z_1, z_1 - z_0 \rangle + p_1 L^2 \| z_1 - v_1 \|^2 + \frac{1}{2\alpha} \| z_0 - z_* \|^2 + \epsilon \| B z_0 - B v_1 \|^2
\]
\[
= \frac{\alpha}{1 + \alpha L} \| B z_1 \|^2 + \langle B z_1, -\alpha B v_1 \rangle + p_1 L^2 \| \alpha (B z_0 - B v_1) \|^2 + \frac{1}{2\alpha} \| z_0 - z_* \|^2 + \epsilon \| B z_0 - B v_1 \|^2
\]
\[
= \frac{\alpha}{1 + \alpha L} \| B z_1 \|^2 + \langle B z_1, -\alpha B v_1 \rangle + \frac{1}{2\alpha} \| z_0 - z_* \|^2 + (\alpha^2 L^2 p_1 + \epsilon) \| B z_0 - B v_1 \|^2. \quad (42)
\]
Because
\[
\alpha^2 L^2 p_1 + \epsilon = \alpha^2 L^2 \frac{\alpha}{1 + \alpha L} + \frac{\alpha^2 L (1 - \alpha L - \alpha^2 L^2 - \alpha^3 L^3)}{2(1 + \alpha L)} = \alpha^2 L (1 - \alpha^2 L^2) / 2,
\]
we see that the expression (42) coincides with (39), except that \( B z_{1/2} \) is replaced with \( B v_1 \). Therefore, the rest of the proof is exactly the same as in Lemma 9 (add the same set of inequalities to (42) with \( v_1 \) and \( B v_1 \) in places of \( z_{1/2} \) and \( B z_{1/2} \), respectively).

**Case** \( k \geq 1 \). Using the identities
\[
z_{k+1} - z_0 = (1 - \beta_k) (z_k - z_0) - \alpha B v_{k+1}
\]
\[
z_k - z_{k+1} = \beta_k (z_k - z_0) + \alpha B v_{k+1}
\]
\[
v_k - z_k = \alpha B v_k - \alpha B v_{k+1}
\]
\[
v_{k+1} - z_{k+1} = \alpha B v_{k+1} - \alpha B v_k,
\]
we begin with
\[
V_k - V_{k+1} \geq V_k - V_{k+1} - \frac{q_k}{\beta_k} (z_k - z_{k+1}, B z_k - B z_{k+1})
\]
\[
= p_k \| B z_k \|^2 - \frac{\alpha q_k}{\beta_k} (B z_k, B v_{k+1}) + \frac{\alpha q_k}{\beta_k (1 - \beta_k)} (B v_{k+1}, B z_{k+1}) - p_{k+1} \| B z_{k+1} \|^2
\]
\[
+ \alpha^2 p_k L^2 \| B v_{k-1} - B v_k \|^2 - \alpha^2 p_{k+1} L^2 \| B v_k - B v_{k+1} \|^2
\]
\[
= p_k \| B z_k \|^2 - \alpha (k+1) (B z_k, B v_{k+1}) + \alpha (k+1)^2 (B v_{k+1}, B z_{k+1}) - p_{k+1} \| B z_{k+1} \|^2
\]
\[
+ \alpha^2 p_k L^2 \| B v_{k-1} - B v_k \|^2 - \alpha^2 p_{k+1} L^2 \| B v_k - B v_{k+1} \|^2. \quad (43)
\]
Following the idea in the prior work [92] which introduced the anchored Popov’s scheme, we use the following inequality:
\[
\| B v_k - B v_{k+1} \|^2 = \| (B v_k - B z_k) + (B z_k - B v_{k+1}) \|^2
\]
\[
\leq 2 \| B v_k - B z_k \|^2 + 2 \| B z_k - B v_{k+1} \|^2
\]
\[
\leq 2 L^2 \| v_k - z_k \|^2 + 2 \| z_k - B v_{k+1} \|^2
\]
\[
= 2 \alpha^2 L^2 \| B v_k - B v_{k-1} \|^2 + 2 \| z_k - B v_{k+1} \|^2,
\]
\[
= 2 \alpha^2 L^2 \| B v_{k+1} - B v_k \|^2 + 2 \| z_{k+1} - B v_{k+1} \|^2.
\]
which implies
\[ 2\alpha^2 L^2 \|B v_k - B v_{k-1}\|^2 + 2 \|B z_k - B v_{k+1}\|^2 - \|B v_k - B v_{k+1}\|^2 \geq 0. \]  
(44)

Additionally, note the simple Lipschitzness inequality
\[ 0 \leq \alpha^2 L^2 \|B v_k - B v_{k+1}\|^2 + \|B v_k - B v_{k+1}\|^2 - \|B ^{z_k+1} - B v_{k+1}\|^2 \geq 0. \]  
(45)

Then, given positive constants \( \tau_1, \tau_2 \), we can lower-bound \( V_k - V_{k+1} - \epsilon (k + 1)^2 \|B v_k - B v_{k+1}\|^2 \) by using (43) and subtracting \( \frac{\tau_1}{2} \times (44) \) and \( \tau_2 \times (45) \) as
\[
V_k - V_{k+1} - \epsilon (k + 1)^2 \|B v_k - B v_{k+1}\|^2 \\
\geq p_k \|B z_k\|^2 - \alpha (k + 1) (B z_k, B v_{k+1}) + \alpha (k + 1)^2 (B v_{k+1}, B z_k) - p_{k+1} \|B z_{k+1}\|^2 \\
+ \alpha^2 p_k L^2 \|B v_{k-1} - B v_k\|^2 - \alpha^2 p_{k+1} L^2 \|B v_k - B v_{k+1}\|^2 - \epsilon (k + 1)^2 \|B v_k - B v_{k+1}\|^2 \\
- \frac{\tau_1}{2} (2\alpha^2 L^2 \|B v_k - B v_{k-1}\|^2 + 2 \|B z_k - B v_{k+1}\|^2 - \|B v_k - B v_{k+1}\|^2) \\
- \tau_2 (\alpha^2 L^2 \|B v_k - B v_{k+1}\|^2 - \|B z_{k+1} - B v_{k+1}\|^2) \\
= \alpha^2 L^2 (p_k - \tau_1) \|B v_{k-1} - B v_k\|^2 \\
+ \left( \frac{\tau_1}{2} - \alpha^2 L^2 \tau_2 - \alpha^2 L^2 p_{k+1} - \epsilon (k + 1)^2 \right) \|B v_k - B v_{k+1}\|^2 \\
+ \left( p_k - \tau_1 \right) \|B z_k\|^2 + (2\tau_1 - \alpha (k + 1)) (B z_k, B v_{k+1}) + (\tau_2 - \tau_1) \|B v_{k+1}\|^2 \\
+ \left( \alpha (k + 1)^2 - 2\tau_2 \right) (B v_{k+1}, B z_{k+1}) + (\tau_2 - p_{k+1}) \|B z_{k+1}\|^2. \\
\]  
(46)

We take
\[
\tau_1 = \frac{\alpha (k + 1) \left((1 + 2\alpha^2 L^2)k - \alpha L(1 - 2\alpha L)\right)}{3}, \quad \tau_2 = \frac{\tau_1}{4\alpha^2 L^2},
\]
which are positive for \( \alpha L \) sufficiently small (by Lemma 16), and plug in \( \epsilon = \frac{\alpha^2 L (1 - \alpha L - \alpha^2 L^2 - \alpha^3 L^3)}{2(1 + \alpha L)} \) and
\[
p_k = \frac{\alpha (k + 1) (k + \alpha L (k - 1))}{2(1 + \alpha L)}, \quad p_{k+1} = \frac{\alpha (k + 2) (k + 1 + \alpha L)}{2(1 + \alpha L)}
\]




to obtain
\[
p_k - \tau_1 = \frac{\alpha (k + 1) \left((1 + r - 4r^2 - 4r^3)k - r(1 + 2r + 4r^2)\right)}{6(1 + r)}
\]
(47)
and
\[
\frac{\tau_1}{2} - \alpha^2 L^2 \tau_2 - \alpha^2 L^2 p_{k+1} - \epsilon (k + 1)^2 = \frac{\alpha \left(T(2)(r)k^2 + T(1)(r)k + T(0)(r)\right)}{12(1 + r)}
\]
where \( r = \alpha L \) and \( T(\ell)(r) \), \( \ell = 0, 1, 2 \), are polynomials in \( r \) defined by
\[
T(0)(r) = -r(7 + 5r - 8r^2 - 6r^3) \\
T(1)(r) = 1 - 12r - 3r^2 + 4r^3 + 12r^4 \\
T(2)(r) = 1 - 5r + 2r^2 + 2r^3 + 6r^4.
\]
Therefore, by Lemma 16, \( \tau_1, \tau_2 \) and the coefficients of the first two norm square terms in (46) are positive provided that \( r \) is sufficiently small, and thus we obtain
\[
V_k - V_{k+1} - \epsilon (k + 1)^2 \|B v_k - B v_{k+1}\|^2 \\
\geq (p_k - \tau_1) \|B z_k\|^2 + (2\tau_1 - \alpha (k + 1)) (B z_k, B v_{k+1}) + (\tau_2 - \tau_1) \|B v_{k+1}\|^2 \\
+ \left( \alpha (k + 1)^2 - 2\tau_2 \right) (B v_{k+1}, B z_{k+1}) + (\tau_2 - p_{k+1}) \|B z_{k+1}\|^2 \\
= \text{Tr}(M_k S_k M_k^T),
\]
We have seen in (47) that $s_{11} = p_k - \tau_1 > 0$ if $r$ is small. Next, we have

$$t_{22} = s_{22} - \frac{s_{22}^2}{s_{11}} = \frac{(1 - 2r)(k + 1) \left((1 + 3r + 2r^2)k - r(1 - 2r)\right) \left((1 - 4r^2)k - r(1 + 4r)\right)}{12lr \left((1 + r - 4r^2 - 4r^3)k - r(1 + 2r + 4r^2)\right)}$$

and

$$t_{33} = s_{33} - \frac{s_{23}^2}{t_{22}} = \alpha r \frac{((1 - 8r - 4r^2 - 4r^3)k - 2r(5 - 2r + 2r^2))}{2(1 + r)(1 - 2r)\left((1 + 3r + 2r^2)k - r(1 - 2r)\right)}.$$

and as Lemma 16 shows that these quantities are positive for small $r$, we conclude that

$$V_k - V_{k+1} - \epsilon(k + 1)^2 \|BV_k - Bv_{k+1}\|^2 \geq \text{Tr}(M_k S_k M_k^\top)$$

$$= (s_{11} \|Bz_k\|^2 + 2s_{12}(Bz_k, Bv_{k+1}) + \frac{s_{22}^2}{s_{11}} \|Bv_{k+1}\|^2)$$

$$+ \left(t_{22} \|Bv_{k+1}\|^2 + 2s_{23}(Bv_{k+1}, Bz_{k+1}) + \frac{s_{23}^2}{t_{22}} \|Bz_{k+1}\|^2\right) + t_{33} \|Bz_{k+1}\|^2$$

$$\geq 0.$$

\[\square\]

### F MP analysis for algorithms with iteration-dependent step-sizes

In order to perform the MP analysis of Section 3 for the algorithms with step-sizes varying over iterations such as (9) and (10), we must consider the Halpern iteration with varying resolvent parameters:

$$w_{k+1/2} = \beta_k w_0 + (1 - \beta_k) w_k$$

$$w_{k+1} = J_{\alpha_k B}(w_{k+1/2}).$$

(48)

This is an instance of the generalized Halpern iteration using operators with common fixed points, i.e.,

$$u_{k+1} = \beta_k u_0 + (1 - \beta_k) T_k u_k,$$

where each $T_k : \mathbb{R}^d \to \mathbb{R}^d$ is nonexpansive and $\bigcap_{k=0}^\infty \text{Fix } T_k \neq \emptyset$.

Proceeding in the same way as in Section B.2, denoting by $z_k$ the iterates of (9) and $w_k$ the generalized Halpern iterates (48), we have

$$(k + 1)^2 \|z_{k+1} - w_{k+1}\|^2 \leq k^2 \|z_k - w_k\|^2 + \alpha_k^2 (k + 1)^2 \|Bz_k - Bz_{k+1/2}\|^2.$$

Therefore, once $\{\alpha_k\}_k$ are chosen appropriately so that a summability result

$$\sum_{k=0}^\infty \alpha_k^2 (k + 1)^2 \|Bz_k - Bz_{k+1/2}\|^2 \leq C \|z_0 - z_*\|^2 < \infty$$

holds (which analogous to Lemma 9), then an MP result between the two algorithms can be established. The same analysis can be carried out for the APS algorithm as well. Assuming that the step-size sequence $\{\alpha_k\}_k$ converges to some positive limit $\alpha_\infty$ (which is the case for EAG-V [100] and the original APS [92]), the operators $T_k = J_{z_0 B}$ uniformly converges to $T_\infty = J_{z_\infty B}$ on every compact subset of $\mathbb{R}^d$ [7, Proposition 23.31], and we anticipate that (48) will behave asymptotically similarly as the ordinary Halpern iteration for $T_\infty$ and that the formal convergence analysis could be carried out without significant obstacles. However, we do not further pursue this direction in the interest of keeping the length of this paper manageable.
### G Regularity of convex and convex-concave functions

Consider an extended real-valued convex function \( f : \mathbb{R}^n \to \mathbb{R} \cup \{\pm \infty\} \). We say \( f \) is proper if \( f(x) > -\infty \) for all \( x \in \mathbb{R}^n \) and \( f(x) < +\infty \) for at least one \( x \in \mathbb{R}^n \). The closure \( f \), written as \( \text{cl} f \), is defined as the constant function \( \text{cl} f \equiv -\infty \) if \( f(x) = -\infty \) for some \( x \in \mathbb{R}^n \) and otherwise
\[
\text{cl} f(x) := \lim_{x' \to x} f(x').
\]

We say \( f \) is closed if \( \text{cl} f = f \) (when \( f \) is proper, this is equivalent to lower semicontinuity). If \( f \) is closed, convex, and proper (CCP), its convex subdifferential operator \( \partial f : \mathbb{R}^n \rightharpoonup \mathbb{R}^n \) defined by
\[
\partial f(x) = \{ v \in \mathbb{R}^n \mid f(x') \geq f(x) + \langle v, x' - x \rangle, \quad \forall x' \in \mathbb{R}^n \}
\]
is maximal monotone [81, Corollary 31.5.2]. For an extended real-valued concave function \( g \), we define \( \text{cl} g = -\text{cl} (-g) \) and say it is proper (resp. closed) if \(-g\) is proper (resp. closed) as a convex function.

Let \( K : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R} \cup \{\pm \infty\} \) be an extended real-valued convex-concave function. Define
\[
\begin{align*}
dom_x K &= \{ x \in \mathbb{R}^n \mid K(x, y) < +\infty, \quad \forall y \in \mathbb{R}^m \} \\
\dom_y K &= \{ y \in \mathbb{R}^m \mid K(x, y) > -\infty, \quad \forall x \in \mathbb{R}^n \} \\
\dom K &= \dom_x K \times \dom_y K.
\end{align*}
\]

We say \( K \) is proper if \( \dom K \neq \emptyset \). Define \( \text{cl}_x K(x, y) \) as the function obtained by taking the convex closure of \( K(x, y) \) with respect to \( x \) for each fixed \( y \), and define \( \text{cl}_y K(x, y) \) analogously. Then both \( \text{cl}_x K \) and \( \text{cl}_y K \) are convex-concave [81, Corollary 33.1.1]. Respectively define the lower and upper closures of \( K \) by
\[
\begin{align*}
\underline{K} &= \text{cl}_x \text{cl}_y K, \quad \overline{K} = \text{cl}_y \text{cl}_x K,
\end{align*}
\]

which satisfy \( \underline{K} \leq \overline{K} \). We say \( K \) is closed if \( \text{cl}_x K = \text{cl}_x \text{cl}_y K \) and \( \text{cl}_y K = \text{cl}_y \text{cl}_x K \).

The saddle subdifferential operator \( \partial_{\pm} K : \mathbb{R}^n \times \mathbb{R}^m \rightharpoonup \mathbb{R}^n \times \mathbb{R}^m \) is defined by
\[
\partial_{\pm} K(x, y) = \{ (v, w) \in \mathbb{R}^n \times \mathbb{R}^m \mid v \in \partial_x K(x, y), \quad w \in \partial_y (\pm K)(x, y) \},
\]
where \( \partial_x K(x, y) \) denotes the convex subdifferential of \( K(x, y) \) as a convex function of \( x \) (for each fixed \( y \)), and \( \partial_y (\pm K)(x, y) \) denotes the convex subdifferential of \( \pm K(x, y) \) as a convex function of \( y \) (for each fixed \( x \)). If \( K \) is closed, convex-concave, and proper (CCCP), then \( \partial_{\pm} K \) is maximal monotone [81, Corollary 35.7.2] and \((x, y)\) is a minimax solution to
\[
\begin{align*}
\text{minimize} & \quad K(x, y) \\
\text{maximize} & \quad K(x, y)
\end{align*}
\]
if and only if \( 0 \in \partial_{\pm} K(x, y) \). Given \( z \in \mathbb{R}^n \times \mathbb{R}^m \), \( K \) is differentiable at \( z \) if and only if \( \partial_{\pm} K \) is singleton, and in that case, \( \partial_{\pm} K(z) = \nabla_{\pm} K(z) \) [81, Theorem 35.8].

We say two convex-concave functions \( K, L \) are equivalent if \( \text{cl}_x K = \text{cl}_x L \) and \( \text{cl}_y K = \text{cl}_y L \). If \( K, L \) are equivalent, then they share the set of minimax solutions, the function values at the solutions (if any) are equal, and \( \partial_{\pm} K = \partial_{\pm} L \) [81, Theorem 36.4, Corollary 37.4.1]. If \( K \) is a closed convex-concave function, then the equivalence class of closed convex-concave functions containing \( K \) is precisely characterized [81, Theorem 34.2] by
\[
\{ L : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R} \cup \{\pm \infty\} \mid L \text{ is closed, convex-concave, and } \underline{K} \leq L \leq \overline{K} \}.
\]

When \( f \) and \( g \) are respectively CCP functions on \( \mathbb{R}^n \) and \( \mathbb{R}^m \), we define
\[
f(x) - g(y) = \begin{cases} f(x) - g(y) & x \in \text{dom} f, \ y \in \text{dom} g \\ +\infty & x \notin \text{dom} f, \ y \in \text{dom} g \\ -\infty & x \in \text{dom} f, \ y \notin \text{dom} g \end{cases}
\]
but when \( x \notin \text{dom} f, \ y \notin \text{dom} g \), it seems unclear how to define \( f(x) - g(y) = +\infty - \infty \). However, it turns out that one can extend (49) to a CCCP function on \( \mathbb{R}^n \times \mathbb{R}^m \), and any such extensions are equivalent to one another [81, Theorem 34.4]. In particular, using the convention \( +\infty - \infty = -\infty \), one gets the lower closure (minimal element) of the equivalence class containing the equivalent CCCP extensions of (49). Alternatively, using the convention \( +\infty - \infty = +\infty \), one gets the upper closure (maximal element) of the equivalence class. In this regard, we can safely deal with the case \( \text{L}_{p}(x, y) = f(x) - g(y) \) (where the equality in fact means \( \text{L}_{p} \) belongs to the equivalence class of CCCP extensions of (49)) without pathology or ambiguities.