MA^{4}EG: MOVING-ANCHOR ADAPTIVE-ADAPTIVE ACCELERATED EXTRA GRADIENT METHOD

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose the moving-anchor adaptive-adaptive accelerated extragradient method (MA^{4}EG) for smooth convex-concave minimax problems, which combines extragradient updates with a moving anchor to achieve accelerated convergence while adapting to both the smoothness constant and the strong convexity-concavity parameter. MA^{4}EG uses no line search and no restarts, and its step size and anchor coefficient are adaptively updated. For a smooth convex-concave objective function, MA^{4}EG achieves an accelerated $\mathcal{O}(1/k^2)$ last-iterate convergence rate, similar to the extra anchored gradient (EAG) method. When the objective function is additionally μ-strongly convex-concave, MA^{4}EG has a linear last-iterate convergence rate. Numerical experiments in several diverse settings show that MA^{4}EG matches or beats prior methods, and significantly outperforms the baselines when the local smoothness constant and strong convexity-concavity parameter are more favorable than their global counterparts.

1 INTRODUCTION

We consider minimax optimization problems of the form

$$\min_{\boldsymbol{x}\in\mathbb{R}^{d_x}} \max_{\boldsymbol{y}\in\mathbb{R}^{d_y}} \mathcal{L}(\boldsymbol{x},\boldsymbol{y}). \tag{1}$$

Such problems arise in a wide range of machine learning applications, including adversarial training (Goodfellow et al., 2015; Madry et al., 2018), generative adversarial networks (Goodfellow et al., 2014), and reinforcement learning (Du et al., 2017). For smooth convex-concave objectives, accelerated minimax methods such as extra anchored gradient (EAG) (Yoon & Ryu, 2021) and fast extragradient (FEG) (Lee & Kim, 2021) attain the optimal $\mathcal{O}(1/k^2)$ last-iterate convergence rate in squared gradient norm.

However, existing accelerated minimax algorithms lack adaptivity to the smoothness and strong convexity-concavity parameters of the objective function in equation 1. Classical accelerated minimax methods choose their step sizes using the smoothness constant, which may be unknown or substantially larger than the local smoothness constant encountered along the iterates. Furthermore, these algorithms do not automatically yield linear convergence when the objective is additionally strongly convex-concave; i.e., the "accelerated" method is *slower* than the nonaccelerated method on problem instances with the additional structure of strong convexity-concavity. Prior works address these limitations through parameter-dependent tuning (Yoon & Ryu, 2025), backtracking line search (Lee & Kim, 2021), restarting (Diakonikolas, 2020), or implicit updates (Suh et al., 2023). In Section 1.2, we provide an overview of prior work and existing results on adaptivity.

Contributions. We propose the moving-anchor adaptive-adaptive accelerated extragradient method (MA^{4}EG) for smooth convex-concave minimax problems. MA^{4}EG uses explicit first-order updates, requires no prior knowledge of the smoothness constant or the strong convexity-concavity parameter, and employs neither line search nor restarts. We establish an optimal $\mathcal{O}(1/k^2)$ last-iterate convergence rate in squared gradient norm, along with linear convergence when the objective is additionally strongly convex-concave. MA^{4}EG matches the convergence rate and empirical performance of the strongest baselines in each regime: EAG and FEG in the convex-concave setting, and EG and Popov's method in the strongly convex-concave setting. Experiments further show that

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

MA^{4}EG substantially outperforms these baselines when the local curvature is more favorable than the global smoothness and strong convexity-concavity constants.

Algorithm 1 MA^{4}EG

Choose $\nu \in (0,1)$, $\gamma > 1$, $\rho > 1$, and $0 < \widehat{R}_0 < R_0$.
Initialize $\boldsymbol{z}_0 \in \mathbb{R}^d$, $\boldsymbol{a}_0 = \boldsymbol{z}_0$, and $h_0 = \nu/(\gamma R_0)$.
for $k = 0, 1, 2, \ldots$ **do**
  If $F(\boldsymbol{z}_k) = 0$, stop.
  $\tau_k = \dfrac{2[\langle F(\boldsymbol{z}_k), \boldsymbol{a}_k - \boldsymbol{z}_k\rangle]_+}{\|F(\boldsymbol{z}_k)\|^2}, \quad \beta_k = \dfrac{h_k}{\tau_k + h_k}.$
  $\boldsymbol{y}_k = (1-\beta_k)\boldsymbol{z}_k + \beta_k \boldsymbol{a}_k - (1-\beta_k) h_k F(\boldsymbol{z}_k).$
  $\boldsymbol{z}_{k+1} = (1-\beta_k)\boldsymbol{z}_k + \beta_k \boldsymbol{a}_k - h_k F(\boldsymbol{y}_k).$
  $\ell_k = \|F(\boldsymbol{z}_{k+1}) - F(\boldsymbol{y}_k)\| / \|\boldsymbol{z}_{k+1} - \boldsymbol{y}_k\|$, where a ratio with zero denominator is set to 0.
  $\widehat{R}_{k+1} = \max\{\widehat{R}_k, \ell_k\}, \quad R_{k+1} = \max\{\widehat{R}_{k+1},\, R_k/\rho\}.$
  $h_{k+1} = \min\{\rho h_k,\ \nu/(\gamma R_{k+1})\}.$
  $w_k = \left[\dfrac{\tau_k + h_k}{2}\|F(\boldsymbol{z}_{k+1})\|^2 + \langle F(\boldsymbol{z}_{k+1}), \boldsymbol{z}_{k+1} - \boldsymbol{a}_k\rangle\right]_+.$
  $\boldsymbol{a}_{k+1} = \boldsymbol{a}_k + \dfrac{w_k}{\|F(\boldsymbol{z}_{k+1})\|^2} F(\boldsymbol{z}_{k+1})$, where a ratio with zero denominator is set to 0.
end for

1.1 Preliminaries and Notation

We call a saddle function $\mathcal{L} : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R}$ *convex-concave* if $\mathcal{L}(\cdot, \boldsymbol{y})$ is convex for every fixed $\boldsymbol{y}$ and $\mathcal{L}(\boldsymbol{x}, \cdot)$ is concave for every fixed $\boldsymbol{x}$. For $\mu > 0$, $\mathcal{L}$ is *μ-strongly convex-concave* if $\mathcal{L}(\cdot, \boldsymbol{y}) - \frac{\mu}{2}\|\cdot\|^2$ is convex for every $\boldsymbol{y}$ and $\mathcal{L}(\boldsymbol{x}, \cdot) + \frac{\mu}{2}\|\cdot\|^2$ is concave for every $\boldsymbol{x}$. Here and throughout, $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ denote the Euclidean inner product and norm. A *saddle point* $(\boldsymbol{x}^\star, \boldsymbol{y}^\star)$ of $\mathcal{L}$ satisfies

$$\mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{y}) \le \mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{y}^\star) \le \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}^\star) \qquad \forall\, \boldsymbol{x} \in \mathbb{R}^{d_x},\ \boldsymbol{y} \in \mathbb{R}^{d_y}.$$

We seek a saddle point of $\mathcal{L}$ for equation 1.

For notational convenience, let $d = d_x + d_y$ and write $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^d$. For a differentiable objective $\mathcal{L}$, define its *saddle operator* by

$$F(\boldsymbol{z}) := \begin{pmatrix} \nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) \\ -\nabla_{\boldsymbol{y}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) \end{pmatrix}. \tag{2}$$

An operator $F : \mathbb{R}^d \to \mathbb{R}^d$ is *monotone* if

$$\langle F(u) - F(v), u - v\rangle \ge 0 \qquad \forall\, u, v \in \mathbb{R}^d.$$

It is *μ-strongly monotone*, for $\mu > 0$, if

$$\langle F(u) - F(v), u - v\rangle \ge \mu\|u - v\|^2 \qquad \forall\, u, v \in \mathbb{R}^d.$$

Convex-concavity of $\mathcal{L}$ implies monotonicity of its saddle operator, and μ-strong convexity-concavity implies μ-strong monotonicity. We use $\mu = 0$ for the merely convex-concave or monotone case. The operator is *L-Lipschitz continuous*, for $L > 0$, if

$$\|F(u) - F(v)\| \le L\|u - v\| \qquad \forall\, u, v \in \mathbb{R}^d.$$

We call the saddle function $\mathcal{L}$ *L-smooth* when its full gradient $\nabla\mathcal{L}$ is L-Lipschitz, which is equivalent to L-Lipschitz continuity of F. For a differentiable convex-concave objective, $\boldsymbol{z}^\star = (\boldsymbol{x}^\star, \boldsymbol{y}^\star)$ is a saddle point of $\mathcal{L}$ if and only if $F(\boldsymbol{z}^\star) = 0$.

Throughout the paper, for a scalar t, $[t]_+ := \max\{t, 0\}$ and $\lceil t \rceil$ is its ceiling. We write I for the identity map or matrix of the appropriate dimension. We follow the convention that empty sums are zero. For nonnegative sequences, $g(k) = \mathcal{O}(f(k))$ means that $g(k) \le c f(k)$ for all sufficiently large k, with c independent of k. Logarithms are natural unless a base is indicated.

1.2 Prior Work

Extragradient-type methods for minimax problems. The extragradient (EG) method (Korpelevich, 1976) and Popov's method (Popov, 1980) are classical first-order methods for smooth convex-concave minimax problems. EG uses two gradient evaluations per iteration, whereas Popov's method reuses a previous evaluation and requires only one new evaluation. Mokhtari et al. (2020a;b) analyzed these methods as approximate proximal-point iterations, establishing convergence guarantees in structured settings and an $\mathcal{O}(1/k)$ ergodic primal–dual gap bound in the general convex-concave setting. Subsequent work established an $\mathcal{O}(1/k)$ last-iterate rate in squared gradient norm for smooth convex-concave minimax problems with suitable fixed step sizes: Gorbunov et al. (2022a) for EG and Gorbunov et al. (2022b) for Popov's method. These two methods serve as the nonaccelerated baselines in our experiments.

Anchoring-based acceleration for minimax problems. A separate line of work accelerates extragradient-type methods by anchoring the iterates to a reference point with a vanishing weight, a mechanism that traces back to Halpern's fixed-point iteration (Halpern, 1967) and was subsequently used in minimax optimization (Ryu et al., 2019). Yoon & Ryu (2021) combined anchoring with extragradient steps to obtain the extra anchored gradient (EAG) methods, the first to achieve an accelerated $\mathcal{O}(1/k^2)$ last-iterate rate in squared gradient norm for general smooth convex-concave minimax problems. Subsequent work achieved the same rate through the fast extragradient (FEG) method (Lee & Kim, 2021), Halpern-type anchoring of Popov's method (Tran-Dinh & Luo, 2021), and Tikhonov regularization of the past-extragradient method (Reg-PEG) (Chen, 2026). Our proposed MA^{4}EG attains this optimal $\mathcal{O}(1/k^2)$ last-iterate rate in squared gradient norm for smooth convex-concave minimax problems.

Handling strong convexity-concavity within accelerated methods. Existing approaches exploit strong convexity-concavity through parameter-dependent tuning, restarting, or implicit updates. Yoon & Ryu (2025) developed SM-EAG+, which achieves linear convergence using coefficients that depend on the strong convexity-concavity parameter μ and reduces to FEG when $\mu = 0$. Diakonikolas (2020) instead uses restarting to exploit this additional curvature without knowing its parameter. Suh et al. (2023) obtain both an $\mathcal{O}(1/k^2)$ last-iterate rate in squared gradient norm for smooth convex-concave problems and linear convergence for strongly convex-concave problems through adaptive anchoring, but each iteration requires solving a proximal saddle-point subproblem. Our proposed MA^{4}EG achieves both guarantees using explicit first-order updates, without restarts or prior knowledge of either the smoothness constant or the strong convexity-concavity parameter.

Adaptive methods. A parallel line of work removes the need for prior knowledge of smoothness constants. In convex minimization, Malitsky & Mishchenko (2020) developed an adaptive gradient method without line search, and Suh & Ma (2025) extended this approach to Nesterov acceleration. For convex-concave minimax problems, universal Mirror-Prox (Bach & Levy, 2019) and AdaProx (Antonakopoulos et al., 2021) adapt to unknown smoothness with convergence rates for averaged iterates. Adaptive Golden Ratio methods (Malitsky, 2020; Alacaoglu et al., 2023) compute step sizes from recent iterates without line search. These adaptive minimax methods do not establish an accelerated $\mathcal{O}(1/k^2)$ last-iterate rate in squared gradient norm for general smooth convex-concave problems. FEG-A (Lee & Kim, 2021) achieves this rate without knowing the smoothness constant but requires backtracking. More recently, Sun et al. (2026) developed moving-anchored extragradient methods with the same accelerated last-iterate rate, whose adaptive variants need backtracking or restarts. Our proposed MA^{4}EG attains the accelerated last-iterate rate with adaptive step sizes, without line searches or restarts.

2 MA^{4}EG

The proposed algorithm MA^{4}EG is presented in Algorithm 1, which combines fast extragradient updates with an adaptive step size h_k, an adaptive anchor coefficient β_k, and a moving anchor $\boldsymbol{a}_k$. The initialization parameter $\widehat{R}_0$ is meant to be an estimate of the Lipschitz constant of the operator F. The step size h_0 is initialized through $R_0 > \widehat{R}_0$ and scaling factors $\nu \in (0, 1)$ and $\gamma > 1$. The parameter $\rho > 1$ controls how aggressively the step size can grow.

At iteration k, we first compute the anchor coefficient β_k according to the adaptive rule $\beta_k = h_k/(\tau_k + h_k)$, where τ_k is determined by the current residual and anchor position. Although β_k

need not decrease monotonically, we expect it to decrease as the adaptive clock τ_k advances. The updates of $\boldsymbol{y}_k$ and $\boldsymbol{z}_{k+1}$ in Lines 6 and 7 then follow the fast extragradient method (Lee & Kim, 2021), using the current step size h_k, anchor coefficient β_k, and anchor $\boldsymbol{a}_k$. Next, ℓ_k estimates the local smoothness from the ratio $\|F(\boldsymbol{z}_{k+1}) - F(\boldsymbol{y}_k)\|/\|\boldsymbol{z}_{k+1} - \boldsymbol{y}_k\|$, and $\widehat{R}_{k+1}$ tracks the maximum local curvature seen so far. The safeguarded estimate R_{k+1} incorporates this information while allowing the initial scale R_0 to decrease gradually if R_0 is too large, and is used to determine the next step size h_{k+1}. Finally, we update the moving anchor $\boldsymbol{a}_k$ through a nonnegative correction w_k.

We now state the convergence results for MA^{4}EG. The quantity J in Theorem 1 upper-bounds the number of unsafe iterations, which is small when $\widehat{R}_0$ is a good estimate of the smoothness constant of the objective. In our experiments in Section 3, we set $\nu = 0.95$, $\gamma = 1.10$, and $\rho = 1.30$, and MA^{4}EG matches or beats all baselines with tuned step sizes even when $\widehat{R}_0$ is initialized using only simple local finite-difference probes at $\boldsymbol{z}_0$ (see equation 10).

Theorem 1 (Convex-concave)**.** *Assume $\mathcal{L} : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R}$ is convex-concave and L-smooth for some $L > 0$, and admits a saddle point $\boldsymbol{z}^\star = (\boldsymbol{x}^\star, \boldsymbol{y}^\star)$. Let $(\boldsymbol{z}_k)_{k\geq 0}$ be generated by Algorithm 1, with the iterates kept constant after termination. Let*

$$\overline{h} := \frac{\nu}{\gamma \widehat{R}_0}, \qquad \underline{h} := \frac{\nu}{\gamma \max\{R_0, L\}}, \qquad G := 2 + L\overline{h}\left(1 + \frac{4(1+\nu)}{1-\nu}\right),$$

$$C_{\mathrm{anc}} := \frac{4(1+\nu)}{1-\nu} + L\overline{h} + \frac{5 + L\overline{h}}{2} G, \qquad J := \max\left\{0,\ \left\lceil \log_\gamma\left(\frac{L}{\widehat{R}_0}\right)\right\rceil - 1\right\}.$$

Then for every $k \geq 1$,

$$\|\nabla \mathcal{L}(\boldsymbol{z}_k)\| \leq \frac{2C_{\mathrm{anc}}^J}{k\underline{h}} \|\boldsymbol{z}_0 - \boldsymbol{z}^\star\|.$$

Consequently, MA^{4}EG achieves the last-iterate convergence rate $\|\nabla \mathcal{L}(\boldsymbol{z}_k)\|^2 = \mathcal{O}(k^{-2})$.

Theorem 2 (Strongly convex-concave)**.** *In addition to the hypotheses of Theorem 1, suppose $\mathcal{L}$ is μ-strongly convex-concave for some $\mu > 0$. Let $\underline{h}$, C_{anc}, and J be as defined in Theorem 1. Let*

$$\Theta := \frac{4}{\mu\underline{h}} + \frac{2(3+\nu)^2}{1-\nu^2}, \qquad q := 1 + \frac{2}{\Theta}.$$

For $k \geq 1$, let $m_k := \max\{k - J, 0\}$. Then

$$\|\nabla \mathcal{L}(\boldsymbol{z}_k)\| \leq \frac{2C_{\mathrm{anc}}^J \|\boldsymbol{z}_0 - \boldsymbol{z}^\star\|}{\underline{h}\left[k - m_k + \frac{\Theta}{2}\left(q^{m_k} - 1\right)\right]} \leq \frac{2C_{\mathrm{anc}}^J \|\boldsymbol{z}_0 - \boldsymbol{z}^\star\|}{\underline{h}\, q^{\max\{m_k, 1\}-1}}.$$

Consequently, MA^{4}EG achieves linear convergence $\|\nabla \mathcal{L}(\boldsymbol{z}_k)\| = \mathcal{O}(q^{-k})$.

2.1 Discussion

The influential work of Malitsky & Mishchenko (2020) establishes a line of research on adaptive methods, where step sizes are chosen based on local smoothness information without line search. This motivates our step-size design in MA^{4}EG, where we use ℓ_k to monitor local curvature and maintain a safeguarded estimate of the largest curvature observed so far. The resulting update adjusts the step size using information collected during the iteration, without requiring prior knowledge of L or a line search.

Suh et al. (2023) study the continuous-time dynamics of anchor acceleration and show that an adaptive choice of the anchor coefficient achieves accelerated convergence under monotonicity and, without any prior knowledge of μ, an even faster rate under strong monotonicity. Our choice of the anchor coefficient β_k in MA^{4}EG follows this insight. For a fixed anchor, the resulting anchoring term recovers their adaptive coefficient in the continuous-time limit.

Finally, Alcala et al. (2023) show that moving-anchor variants of EAG and FEG retain accelerated convergence. MA^{4}EG's anchor update (Line 12 of Algorithm 1) builds on this idea to cope with an unknown Lipschitz constant. Intuitively, because MA^{4}EG must choose its step size h_k without knowing L, an individual step can occasionally overshoot the progress it should have made toward $\boldsymbol{z}^\star$. MA^{4}EG moves the anchor only when needed to ensure the adaptive clock τ_k advances by at least the step size at every iteration ($\tau_{k+1} \geq \tau_k + h_k$), the key property behind the accelerated convergence rate, as discussed in the proof in Section 2.2.

2.2 PROOF OUTLINE

We outline the proof of Theorem 1. The proof of Theorem 2 follows a similar analysis that additionally uses strong monotonicity.

Proof of Theorem 1 (outline). The analysis is based on the adaptive clock τ_k defined in Algorithm 1 and, for any anchor $\boldsymbol{a} \in \mathbb{R}^d$, the quadratic potential

$$\Phi_k^{\boldsymbol{a}}(t) := \frac{t^2}{2}\|F(\boldsymbol{z}_k)\|^2 + t\,\langle F(\boldsymbol{z}_k), \boldsymbol{z}_k - \boldsymbol{a}\rangle\,. \tag{3}$$

By construction, τ_k is the largest nonnegative root of $\Phi_k^{\boldsymbol{a}_k}$, i.e.,

$$\Phi_k^{\boldsymbol{a}_k}(\tau_k) = 0. \tag{4}$$

The MA^{4}EG updates imply

$$\begin{aligned}
\boldsymbol{z}_{k+1} - \boldsymbol{a}_k &= (1-\beta_k)(\boldsymbol{z}_k - \boldsymbol{a}_k) - h_k F(\boldsymbol{y}_k),\\
\boldsymbol{z}_k - \boldsymbol{z}_{k+1} &= \beta_k(\boldsymbol{z}_k - \boldsymbol{a}_k) + h_k F(\boldsymbol{y}_k),\\
\boldsymbol{z}_{k+1} - \boldsymbol{y}_k &= -h_k F(\boldsymbol{y}_k) + (1-\beta_k) h_k F(\boldsymbol{z}_k).
\end{aligned}$$

Using these identities and $\tau_k\beta_k = (1-\beta_k)h_k$, one obtains the key one-step relation

$$\begin{aligned}
\Phi_k^{\boldsymbol{a}_k}(\tau_k) - \Phi_{k+1}^{\boldsymbol{a}_k}(\tau_k + h_k) = \frac{\tau_k(\tau_k+h_k)}{h_k}\,\langle F(\boldsymbol{z}_k) - F(\boldsymbol{z}_{k+1}), \boldsymbol{z}_k - \boldsymbol{z}_{k+1}\rangle\\
+ \frac{(\tau_k+h_k)^2}{2}\left(\frac{\|\boldsymbol{z}_{k+1} - \boldsymbol{y}_k\|^2}{h_k^2} - \|F(\boldsymbol{z}_{k+1}) - F(\boldsymbol{y}_k)\|^2\right).
\end{aligned} \tag{5}$$

The first term on the right-hand side is nonnegative by monotonicity. If iteration k is safe, namely if $h_k\ell_k \le \nu < 1$, then the definition of ℓ_k makes the second term nonnegative as well. Hence $\Phi_{k+1}^{\boldsymbol{a}_k}(\tau_k + h_k) \le 0$, so the positive-part correction gives $w_k = 0$ and $\boldsymbol{a}_{k+1} = \boldsymbol{a}_k$. On an unsafe iteration, the anchor correction is chosen so that

$$\Phi_{k+1}^{\boldsymbol{a}_{k+1}}(\tau_k + h_k) \le 0.$$

Since τ_{k+1} is the largest nonnegative root of $\Phi_{k+1}^{\boldsymbol{a}_{k+1}}$, both cases yield

$$\tau_{k+1} \ge \tau_k + h_k. \tag{6}$$

Next, Lipschitz continuity gives $\ell_k \le L$. Consequently, the adaptive estimates satisfy $R_k \le \max\{R_0, L\}$ and

$$h_k \ge \underline{h} := \frac{\nu}{\gamma\max\{R_0, L\}}. \tag{7}$$

Moreover, if $h_k\ell_k > \nu$, then $\ell_k > \gamma R_k \ge \gamma\widehat{R}_k$. Thus every unsafe iteration increases $\widehat{R}_k$ by at least a factor of γ. Since $\ell_k \le L$, the number of unsafe iterations is at most

$$J := \max\left\{0, \left\lceil \log_\gamma\left(\frac{L}{\widehat{R}_0}\right)\right\rceil - 1\right\}.$$

Combining equation 7 with equation 6 gives

$$\tau_k \ge \sum_{i=0}^{k-1} h_i \ge k\underline{h}.$$

It remains to control the anchor. Let b_k be the number of unsafe iterations before iteration k. The proof establishes, by induction, that

$$\|\boldsymbol{a}_k - \boldsymbol{z}^\star\| \le C_{\mathrm{anc}}^{b_k}\|\boldsymbol{z}_0 - \boldsymbol{z}^\star\|, \tag{8}$$

$$\|\boldsymbol{z}_k - \boldsymbol{a}_k\| \le \frac{2(1+\nu)}{1-\nu} C_{\mathrm{anc}}^{b_k}\|\boldsymbol{z}_0 - \boldsymbol{z}^\star\|. \tag{9}$$

Finally, dividing equation 4 by $\tau_k \geq k\underline{h} > 0$ at iteration k gives

$$\begin{aligned}\frac{\tau_k}{2}\|F(\boldsymbol{z}_k)\|^2 &= \langle F(\boldsymbol{z}_k), \boldsymbol{a}_k - \boldsymbol{z}_k\rangle \\ &= \langle F(\boldsymbol{z}_k), \boldsymbol{a}_k - \boldsymbol{z}^\star\rangle + \langle F(\boldsymbol{z}_k), \boldsymbol{z}^\star - \boldsymbol{z}_k\rangle \\ &\leq \langle F(\boldsymbol{z}_k), \boldsymbol{a}_k - \boldsymbol{z}^\star\rangle \\ &\leq \|F(\boldsymbol{z}_k)\|\,\|\boldsymbol{a}_k - \boldsymbol{z}^\star\|,\end{aligned}$$

where the inequality uses monotonicity and $F(\boldsymbol{z}^\star) = 0$, since $\langle F(\boldsymbol{z}_k), \boldsymbol{z}^\star - \boldsymbol{z}_k\rangle = -\langle F(\boldsymbol{z}_k) - F(\boldsymbol{z}^\star), \boldsymbol{z}_k - \boldsymbol{z}^\star\rangle \leq 0$. Therefore, using equation 9, $b_k \leq J$, and $\tau_k \geq k\underline{h}$, we obtain

$$\|F(\boldsymbol{z}_k)\| \leq \frac{2C_{\mathrm{anc}}^J}{k\underline{h}}\|\boldsymbol{z}_0 - \boldsymbol{z}^\star\|.$$

Thus MA^{4}EG has an $\mathcal{O}(1/k)$ last-iterate residual rate, or equivalently an $\mathcal{O}(1/k^2)$ rate for the squared residual. □

3 Experiments

We now present experiments illustrating the practical performance of MA^{4}EG on smooth convex-concave and strongly convex-concave minimax problems of the form equation 1. On every problem we compare MA^{4}EG against seven existing methods: EG (Korpelevich, 1976), Popov's method (Popov, 1980), FEG (Lee & Kim, 2021), EAG-C and EAG-V (Yoon & Ryu, 2021), the Halpern-type anchored Popov method (Tran-Dinh & Luo, 2021), and Reg-PEG (Chen, 2026), whose precise iterate formulas are restated in Appendix A. All methods are initialized at $\boldsymbol{z}_0 = 0$, except for the Experiment 3 benchmark, which follows the original setting of Yoon & Ryu (2021) and uses $\boldsymbol{z}_0 = (1, 0)$. We plot the full gradient norm $\|\nabla\mathcal{L}(\boldsymbol{z}_k)\| = \|F(\boldsymbol{z}_k)\|$ against gradient evaluations. Each gradient evaluation returns both partial gradients $(\nabla_{\boldsymbol{x}}\mathcal{L}, \nabla_{\boldsymbol{y}}\mathcal{L})$, corresponding to one call to F in equation 2. EG, FEG, EAG-C, EAG-V, and MA^{4}EG use two evaluations per iteration. Popov, anchored Popov, and Reg-PEG use one, at the auxiliary point $\boldsymbol{y}_k$ of Appendix A.

For all methods except MA^{4}EG, the step size is tuned by a grid search against the certified global smoothness constant of each minimax objective. MA^{4}EG uses no prior knowledge of the strong convexity-concavity parameter μ or the smoothness constant L. The initialization of MA^{4}EG fixes $\nu = 0.95$, $\gamma = 1.10$, $\rho = 1.30$ throughout. We obtain the base scale R_0 from 100 finite-difference probes around $\boldsymbol{z}_0$,

$$R_0 = \underset{j=1,\dots,100}{\mathrm{median}} \frac{\|\nabla\mathcal{L}(\boldsymbol{z}_0 + \delta_j\boldsymbol{v}_j) - \nabla\mathcal{L}(\boldsymbol{z}_0)\|}{\delta_j}, \quad \delta_j \sim \mathrm{LogUnif}(10^{-3}, 0.5), \quad \boldsymbol{v}_j \sim \mathrm{Unif}(\mathbb{S}^{d-1}), \tag{10}$$

where $\mathbb{S}^{d-1} := \{v \in \mathbb{R}^d : \|v\| = 1\}$, Unif denotes the uniform distribution on the indicated set, and $\delta \sim \mathrm{LogUnif}(a, b)$ means $\log\delta \sim \mathrm{Unif}(\log a, \log b)$. We then set the safeguarded initial estimate $\widehat{R}_0$ required by Algorithm 1 by shrinking R_0 with an independent random multiplier,

$$\widehat{R}_0 = u\,R_0, \qquad u \sim \mathrm{Unif}(1/\rho, 1),$$

so that $0 < \widehat{R}_0 < R_0$ almost surely. Because this initialization is randomized, we plot 20 independent MA^{4}EG trials in every figure. This initialization of MA^{4}EG requires substantially less effort and far fewer gradient evaluations than tuning the step sizes of the other baselines.

Experiment 1: bilinearly coupled minimax problem. Figure 1 presents experiments on the unconstrained minimax problem

$$\begin{aligned}&\min_{\boldsymbol{x}\in\mathbb{R}^{d_x}}\max_{\boldsymbol{y}\in\mathbb{R}^{d_y}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) := \boldsymbol{x}^\top K\boldsymbol{y} - \tfrac{1}{2}\boldsymbol{y}^\top D\boldsymbol{y} + \boldsymbol{b}_x^\top\boldsymbol{x} - \boldsymbol{b}_y^\top\boldsymbol{y}, \\ &F(\boldsymbol{x}, \boldsymbol{y}) = (K\boldsymbol{y} + \boldsymbol{b}_x,\ -K^\top\boldsymbol{x} + D\boldsymbol{y} + \boldsymbol{b}_y).\end{aligned} \tag{11}$$

Here $d_x = d_y = 50$ (so $d = 100$), $K = \mathrm{diag}(\sigma_1, \dots, \sigma_{50})$ is diagonal with $\sigma_i \in [0.7, 1]$, $D = 0.01I$, and $\boldsymbol{b} = (\boldsymbol{b}_x, \boldsymbol{b}_y) \in \mathbb{R}^{100}$ is a small affine bias so that $\boldsymbol{z}_0 = 0$ is not already a saddle point. The objective $\mathcal{L}$ is convex-concave with exact global smoothness constant $L \approx 1.005$, but is not strongly convex-concave ($\mu = 0$), since it is affine in $\boldsymbol{x}$. Figure 1 shows that MA^{4}EG performs comparably to the best of the tuned baselines on this problem.

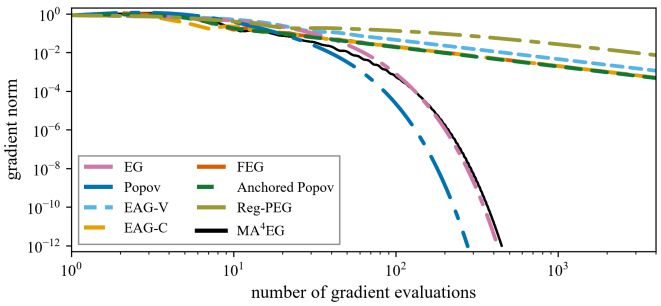
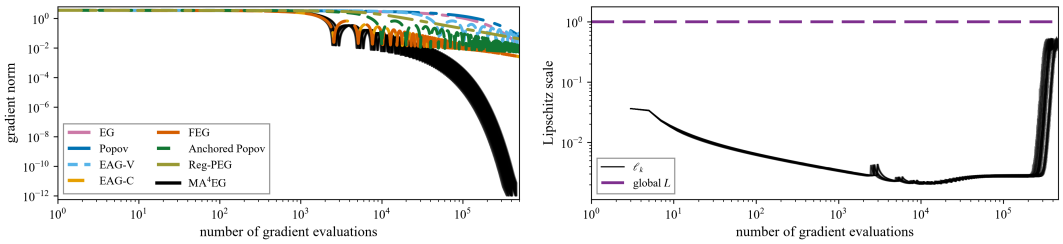


Figure 1: Gradient norm $\|\nabla\mathcal{L}(\boldsymbol{z}_k)\|$ versus gradient evaluations for solving equation 11. MA^{4}EG performs comparably to the best of the tuned baselines.



(a) Gradient norm $\|\nabla\mathcal{L}(\boldsymbol{z}_k)\|$ vs. gradient evaluations for solving equation 12.

(b) MA^{4}EG's adaptive local smoothness estimate ℓ_k (Line 8 of Algorithm 1) vs. the certified global smoothness bound $L = 1$.

Figure 2: Results for Experiment 2 over 5×10^5 gradient evaluations. MA^{4}EG's advantage in (a) is explained by (b): the certified global smoothness constant L substantially overestimates the local curvature of the minimax objective encountered by the iterates. Since MA^{4}EG adapts its step size to this smaller local scale ℓ_k rather than to the fixed global bound L used by every other method, it converges faster than even the accelerated baselines.

Experiment 2: quadratic Lagrangian minimax problem. We next reproduce the second example of Yoon & Ryu (2021), used there to demonstrate the accelerated $\mathcal{O}(1/k^2)$ rate in squared gradient norm of EAG-C and EAG-V over the nonaccelerated EG and Popov baselines. The minimax objective is the Lagrangian of a linearly constrained convex quadratic program,

$$\min_{\boldsymbol{x}\in\mathbb{R}^n}\max_{\boldsymbol{y}\in\mathbb{R}^n}\mathcal{L}(\boldsymbol{x},\boldsymbol{y}) := \tfrac{1}{2}\boldsymbol{x}^\top H\boldsymbol{x} - \boldsymbol{h}^\top\boldsymbol{x} - \langle A\boldsymbol{x}-\boldsymbol{b},\boldsymbol{y}\rangle, \qquad F(\boldsymbol{x},\boldsymbol{y}) = (H\boldsymbol{x}-\boldsymbol{h}-A^\top\boldsymbol{y},\ A\boldsymbol{x}-\boldsymbol{b}). \tag{12}$$

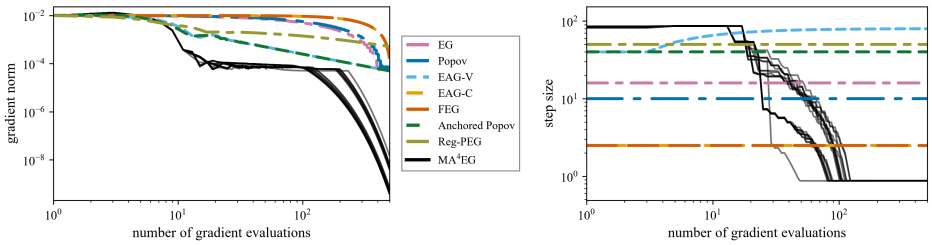
Here $H, A \in \mathbb{R}^{n\times n}$, $\boldsymbol{h}, \boldsymbol{b} \in \mathbb{R}^n$, and $H \succeq 0$ means that H is symmetric positive semidefinite. We adopt the same $H, \boldsymbol{h}, A, \boldsymbol{b}$ construction as Yoon & Ryu (2021), in turn taken from the adversarial lower-bound instance of Ouyang & Xu (2021). We use $d_x = d_y = n = 200$ and run the methods for 5×10^5 gradient evaluations. The objective $\mathcal{L}$ is convex-concave but not strongly convex-concave ($\mu = 0$), since it is affine in $\boldsymbol{y}$, and has exact global smoothness constant $\|\nabla^2\mathcal{L}\|_2 \approx 0.809$. Following Yoon & Ryu (2021), every method that requires a global smoothness constant uses the certified bound $L = 1$ instead. Figure 2 plots gradient norms and MA^{4}EG's local smoothness estimates relative to the global bound over 5×10^5 gradient evaluations, showing that MA^{4}EG outperforms all baselines on this problem.

Experiment 3: smooth absolute-value minimax problem. Our third benchmark is the two-dimensional example of Yoon & Ryu (2021): for $\epsilon > 0$, let $f_\epsilon(u) = \epsilon|u| - \frac{1}{2}\epsilon^2$ if $|u| \geq \epsilon$ and $f_\epsilon(u) = \frac{1}{2}u^2$ otherwise, and consider the Huber-smoothed absolute-value minimax problem

$$\min_{x\in\mathbb{R}}\max_{y\in\mathbb{R}}\ \mathcal{L}_{\delta,\epsilon}(x,y) = \delta xy + (1-\delta)f_\epsilon(x) - (1-\delta)f_\epsilon(y), \tag{13}$$

with saddle operator

$$F(x,y) = \big((1-\delta)f_\epsilon'(x) + \delta y,\ (1-\delta)f_\epsilon'(y) - \delta x\big),$$



(a) Gradient norm $\|\nabla \mathcal{L}_{\delta,\epsilon}(\boldsymbol{z}_k)\|$ vs. gradient evaluations for solving equation 13.

(b) MA^{4}EG's adaptive step size h_k (Line 10 of Algorithm 1) vs. the step size of each nonadaptive baseline.

Figure 3: Results for Experiment 3 over 500 gradient evaluations. MA^{4}EG's advantage in (a) is explained by (b): its step size starts large to cross the nearly flat outer region quickly, then contracts as the local smoothness estimate ℓ_k increases near the strongly convex-concave core. The baselines' step sizes are tuned using global smoothness information and must trade off between these two regimes; only MA^{4}EG converges rapidly to the saddle point while the rest remain confined to the outer region.

where $f'_\epsilon(u) = u$ for $|u| \le \epsilon$ and $f'_\epsilon(u) = \epsilon\,\mathrm{sign}(u)$ otherwise. We use $\delta = 10^{-2}$, $\epsilon = 5 \times 10^{-5}$, and $\boldsymbol{z}_0 = (1, 0)$. The objective is globally convex-concave ($\mu = 0$), but is $(1-\delta)$-strongly convex-concave inside the core $|x|, |y| \le \epsilon$. The certified global smoothness bound $L = 1$ reflects the curvature in this core. In regions where both $|x| > \epsilon$ and $|y| > \epsilon$, the local smoothness constant is only δ. Figure 3 compares the gradient norms and step sizes of MA^{4}EG and the nonadaptive baselines over 500 gradient evaluations.

Experiment 4: regularized logistic minimax problem. Our fourth benchmark is a regularized logistic-regression minimax problem adapted from Liu et al. (2022), who study minimax optimization with unbalanced primal–dual dimensionality and objectives of the form

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{m_x}\sum_{i=1}^{m_x} \log\left(1 + e^{\boldsymbol{c}_i^\top \boldsymbol{x} - s_i}\right) + \frac{\lambda_x}{2}\|\boldsymbol{x}\|^2 + \boldsymbol{x}^\top K \boldsymbol{y} - \frac{1}{m_y}\sum_{j=1}^{m_y} \log\left(1 + e^{\boldsymbol{d}_j^\top \boldsymbol{y} - t_j}\right) - \frac{\lambda_y}{2}\|\boldsymbol{y}\|^2. \tag{14}$$

Here $\boldsymbol{c}_i \in \mathbb{R}^{d_x}$ and $\boldsymbol{d}_j \in \mathbb{R}^{d_y}$ are sample vectors, $s_i, t_j \in \mathbb{R}$ are scalar offsets, m_x, m_y are the sample sizes, $\lambda_x, \lambda_y > 0$ are regularization parameters, and $K \in \mathbb{R}^{d_x \times d_y}$ is the coupling matrix. We set the primal and dual dimensions $d_x = d_y = 25$, sample sizes $m_x = m_y = 300$, regularization $\lambda_x = \lambda_y = 0.01$, and dense bilinear coupling K, with one rare, high-leverage sample per player to separate the global and local curvature. The objective $\mathcal{L}$ is μ-strongly convex-concave and L-smooth with $\mu = 0.01$ and certified global smoothness constant $L \approx 18.85$. Figure 4 plots gradient norms for all methods and MA^{4}EG's adaptive local smoothness estimate relative to the certified global bound over 6000 gradient evaluations.

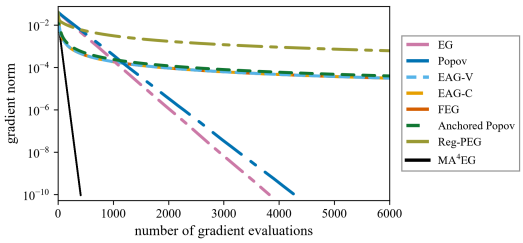
Experiment 5: minimax problem with non-bilinear coupling. Our final example is adapted from Lee et al. (2026), who consider the minimax objective of the form

$$\min_{x\in\mathbb{R}} \max_{y\in\mathbb{R}} \; \mathcal{L}(x, y) = g(x) + \sin x \sin y - h(y), \tag{15}$$

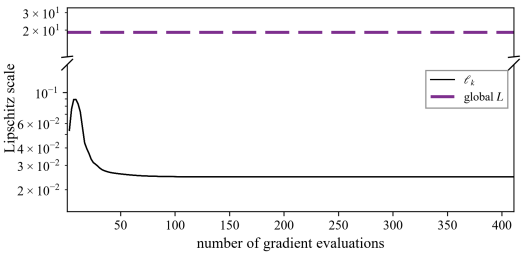
with $g(x) = x^2$ and $h(y) = y^2$, the first known example whose Asplund decomposition admits a non-skew-linear coupling. In our experiment, we choose

$$g(x) = \tfrac{a}{2}x^2 - b_x x + \alpha \log(1 + e^{x-s}), \qquad h(y) = \tfrac{a}{2}y^2 - b_y y + \alpha \log(1 + e^{y-s}),$$

with $a = 1.05$, $s = 20$, $\alpha = 391.8$, and independent shifts $b_x, b_y \sim \mathcal{N}(0, 0.5^2)$, which denotes a normal distribution with mean 0 and variance 0.5^2. The objective $\mathcal{L}$ is μ-strongly convex-concave and L-smooth with $\mu = 0.05$ and certified global smoothness constant $L = 100$. Figure 5 compares gradient norms across methods and MA^{4}EG's adaptive local smoothness estimate against the global bound $L = 100$ over 6000 gradient evaluations.

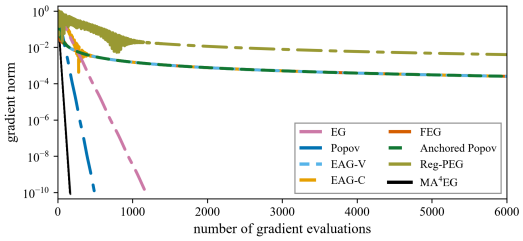


(a) Gradient norm $\|\nabla \mathcal{L}(z_k)\|$ vs. gradient evaluations for the minimax objective in equation 14.

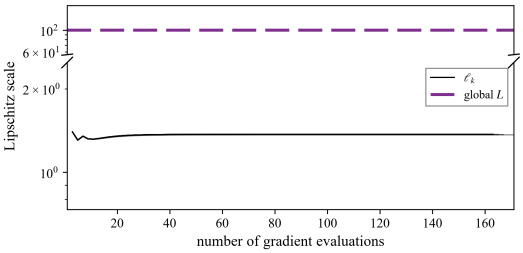


(b) MA^{4}EG's adaptive local smoothness estimate ℓ_k (Line 8 of Algorithm 1) vs. the certified global smoothness bound $L \approx 18.85$.

Figure 4: Results for Experiment 4 over 6000 gradient evaluations. MA^{4}EG's advantage in (a) is explained by (b): the global smoothness constant $L \approx 18.85$ is roughly 200 times larger than the local smoothness estimate ℓ_k along the iterates. Since MA^{4}EG adapts its step size to this much smaller local scale rather than to the fixed global bound L used by every other method, it converges substantially faster.



(a) Gradient norm $\|\nabla \mathcal{L}(z_k)\|$ vs. gradient evaluations for solving equation 15.



(b) MA^{4}EG's adaptive local smoothness estimate ℓ_k (Line 8 of Algorithm 1) vs. the certified global smoothness bound $L = 100$.

Figure 5: Results for Experiment 5 over 6000 gradient evaluations. MA^{4}EG's performance in (a) is explained by (b): the certified global smoothness constant $L = 100$ is substantially larger than the local smoothness estimate ℓ_k along the iterates. Since MA^{4}EG adapts its step size to this smaller local scale without ever needing to know the global smoothness constant L in advance, it performs even better than EG and Popov baselines with tuned step sizes that rely on the certified global smoothness constant.

4 Conclusion

This work presents MA^{4}EG, an adaptive extragradient method for solving smooth convex-concave minimax problems. MA^{4}EG adaptively chooses step sizes and anchor coefficients, requires neither line search nor restarts, and uses no prior knowledge of the smoothness constant or the strong convexity-concavity parameter. We establish an accelerated $\mathcal{O}(1/k^2)$ last-iterate convergence rate for the squared gradient norm and linear convergence under strong convexity-concavity. Numerical experiments demonstrate competitive performance against baselines with tuned step sizes and substantial improvements on problems where the local curvature is more favorable than the global constants suggest.

Beyond these results, MA^{4}EG has novel components with broader potential. First, its moving-anchor correction supports accelerated convergence with adaptive step sizes without line search or restarts. This suggests that moving anchors may be useful in designing adaptive algorithms in other settings, including stochastic problems. Second, to the best of our knowledge, MA^{4}EG is the first method to show that adaptivity to smoothness constants, adaptivity to strong convexity-concavity, and acceleration can be achieved simultaneously without line search or restarts. Their compatibility encourages the development of optimization algorithms that combine desirable properties previously established separately into a single method.

AI USE STATEMENT

In this work, we used generative AI tools for implementing methods. We have not used generative AI tools for generating synthetic data sets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, assisting with translation, cleaning and reformatting datasets, supporting qualitative and thematic data analysis and interpreting results. Additionally, we used generative AI tools for drafting parts of a research paper and editing a research paper to improve readability. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Ahmet Alacaoglu, Axel Böhm, and Yura Malitsky. Beyond the golden ratio for variational inequality algorithms. *Journal of Machine Learning Research*, 24(172):1–33, 2023.

James K. Alcala, Yat Tin Chow, and Mahesh Sunkula. Moving anchor extragradient methods for smooth structured minimax problems. *arXiv preprint arXiv:2308.12359*, 2023.

Kimon Antonakopoulos, E. Veronica Belmega, and Panayotis Mertikopoulos. Adaptive extragradient methods for min-max optimization and games. In *International Conference on Learning Representations*, 2021.

Francis Bach and Kfir Y. Levy. A universal algorithm for variational inequalities adaptive to smoothness and noise. In *Proceedings of the 32nd Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 164–194. PMLR, 2019.

Zihao Chen. A unifying view of anchoring via operator-side Tikhonov regularization. *arXiv preprint arXiv:2605.30905*, 2026.

Jelena Diakonikolas. Halpern iteration for near-optimal and parameter-free monotone inclusion and strong solutions to variational inequalities. In *Proceedings of the 33rd Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 1428–1451. PMLR, 2020.

Simon S. Du, Jianshu Chen, Lihong Li, Lin Xiao, and Dengyong Zhou. Stochastic variance reduction methods for policy evaluation. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1049–1058. PMLR, 2017.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, pp. 2672–2680, 2014.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Eduard Gorbunov, Nicolas Loizou, and Gauthier Gidel. Extragradient method: $\mathcal{O}(1/K)$ last-iterate convergence for monotone variational inequalities and connections with cocoercivity. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 366–402. PMLR, 2022a.

Eduard Gorbunov, Adrien Taylor, and Gauthier Gidel. Last-iterate convergence of optimistic gradient method for monotone variational inequalities. In *Advances in Neural Information Processing Systems*, volume 35, pp. 21858–21870, 2022b.

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

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

Jaewook Lee, Ernest K. Ryu, and Chulhee Yun. Nesterov acceleration with operator decomposition. *arXiv preprint arXiv:2604.11105*, 2026.

Sucheol Lee and Donghwan Kim. Fast extra gradient methods for smooth structured nonconvex-nonconcave minimax problems. In *Advances in Neural Information Processing Systems*, volume 34, pp. 22588–22600, 2021.

Chengchang Liu, Shuxian Bi, Luo Luo, and John C. S. Lui. Partial-quasi-Newton methods: Efficient algorithms for minimax optimization problems with unbalanced dimensionality. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1031–1041, 2022.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Yura Malitsky. Golden ratio algorithms for variational inequalities. *Mathematical Programming*, 184(1–2):383–410, 2020.

Yura Malitsky and Konstantin Mishchenko. Adaptive gradient descent without descent. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6702–6712. PMLR, 2020.

Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1497–1507. PMLR, 2020a.

Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. Convergence rate of $\mathcal{O}(1/k)$ for optimistic gradient and extragradient methods in smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 30(4):3230–3251, 2020b.

Yuyuan Ouyang and Yangyang Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, 185(1–2):1–35, 2021.

L. D. Popov. A modification of the Arrow–Hurwicz method for search of saddle points. *Mathematical Notes of the Academy of Sciences of the USSR*, 28(5):845–848, 1980.

Ernest K. Ryu, Kun Yuan, and Wotao Yin. ODE analysis of stochastic gradient methods with optimism and anchoring for minimax problems. *arXiv preprint arXiv:1905.10899*, 2019.

Jaewook J. Suh and Shiqian Ma. An adaptive and parameter-free Nesterov's accelerated gradient method for convex optimization. *arXiv preprint arXiv:2505.11670*, 2025.

Jaewook J. Suh, Jisun Park, and Ernest K. Ryu. Continuous-time analysis of anchor acceleration. In *Advances in Neural Information Processing Systems*, volume 36, pp. 32782–32866, 2023.

Defeng Sun, Liping Zhang, and Wei Zhao. Convergence analysis of the restarted moving-anchored extra-gradient method in the absence of local Lipschitz continuity. *arXiv preprint arXiv:2607.07585*, 2026.

Quoc Tran-Dinh and Yang Luo. Halpern-type accelerated and splitting algorithms for monotone inclusions. *arXiv preprint arXiv:2110.08150*, 2021.

Taeho Yoon and Ernest K. Ryu. Accelerated algorithms for smooth convex-concave minimax problems with $\mathcal{O}(1/k^2)$ rate on squared gradient norm. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12098–12109. PMLR, 2021.

Taeho Yoon and Ernest K. Ryu. Accelerated minimax algorithms flock together. *SIAM Journal on Optimization*, 35(1):180–209, 2025.

A ALGORITHM SPECIFICATIONS

For clarity, we precisely specify the update rule of every baseline algorithm compared against MA^{4}EG in our experiments. Throughout, F denotes the operator whose zero is sought, $\boldsymbol{z}_k \in \mathbb{R}^d$ is the k-th iterate, $\boldsymbol{y}_k \in \mathbb{R}^d$ is an auxiliary extragradient point, and $h > 0$ is a fixed step size unless stated otherwise.

Extragradient (EG) (Korpelevich, 1976) is defined as

$$\boldsymbol{y}_k = \boldsymbol{z}_k - hF(\boldsymbol{z}_k), \qquad \boldsymbol{z}_{k+1} = \boldsymbol{z}_k - hF(\boldsymbol{y}_k).$$

Popov's method (Popov, 1980), with $\boldsymbol{y}_0 = \boldsymbol{z}_0$, is defined as

$$\boldsymbol{z}_{k+1} = \boldsymbol{z}_k - hF(\boldsymbol{y}_k), \qquad \boldsymbol{y}_{k+1} = \boldsymbol{z}_{k+1} - hF(\boldsymbol{y}_k).$$

Fast extragradient (FEG) (Lee & Kim, 2021), with anchor $\boldsymbol{a} = \boldsymbol{z}_0$ and $\beta_k = 1/(k+1)$, is defined as

$$\boldsymbol{y}_k = \boldsymbol{z}_k + \beta_k(\boldsymbol{a} - \boldsymbol{z}_k) - (1-\beta_k)hF(\boldsymbol{z}_k), \qquad \boldsymbol{z}_{k+1} = \boldsymbol{z}_k + \beta_k(\boldsymbol{a} - \boldsymbol{z}_k) - hF(\boldsymbol{y}_k).$$

Extra anchored gradient with constant step (EAG-C) (Yoon & Ryu, 2021), with anchor $\boldsymbol{a} = \boldsymbol{z}_0$ and $\beta_k = 1/(k+2)$, is defined as

$$\boldsymbol{y}_k = \boldsymbol{z}_k + \beta_k(\boldsymbol{a} - \boldsymbol{z}_k) - hF(\boldsymbol{z}_k), \qquad \boldsymbol{z}_{k+1} = \boldsymbol{z}_k + \beta_k(\boldsymbol{a} - \boldsymbol{z}_k) - hF(\boldsymbol{y}_k).$$

Extra anchored gradient with varying step (EAG-V) (Yoon & Ryu, 2021) uses the same update as EAG-C, but replaces the constant step h by a sequence α_k:

$$\boldsymbol{y}_k = \boldsymbol{z}_k + \beta_k(\boldsymbol{a} - \boldsymbol{z}_k) - \alpha_k F(\boldsymbol{z}_k), \qquad \boldsymbol{z}_{k+1} = \boldsymbol{z}_k + \beta_k(\boldsymbol{a} - \boldsymbol{z}_k) - \alpha_k F(\boldsymbol{y}_k),$$

where $\boldsymbol{a} = \boldsymbol{z}_0$, $\beta_k = 1/(k+2)$, and, for a certified Lipschitz constant L and initial step $\alpha_0 > 0$ (chosen by the grid search, which is not restricted to the range $\alpha_0 < 3/(4L)$ covered by the theory of Yoon & Ryu (2021)), the step size is updated as

$$\alpha_{k+1} = \alpha_k \left(1 - \frac{\alpha_k^2 L^2}{(k+1)(k+3)(1-\alpha_k^2 L^2)}\right), \qquad k \geq 0.$$

Anchored Popov (Tran-Dinh & Luo, 2021), with anchor $\boldsymbol{a} = \boldsymbol{z}_0$, $\boldsymbol{y}_0 = \boldsymbol{z}_0$, $\boldsymbol{g}_0 = F(\boldsymbol{y}_0)$, and $\boldsymbol{z}_1 = \boldsymbol{z}_0 - h\boldsymbol{g}_0$, is defined for $k \geq 1$ as

$$\boldsymbol{p}_k = \frac{\boldsymbol{a} + k\boldsymbol{z}_k}{k+1}, \qquad \boldsymbol{y}_k = \boldsymbol{p}_k - \frac{k}{k+1}hF(\boldsymbol{y}_{k-1}), \qquad \boldsymbol{z}_{k+1} = \boldsymbol{p}_k - hF(\boldsymbol{y}_k).$$

Regularized past-extragradient (Reg-PEG) (Chen, 2026), with anchor $\boldsymbol{a} = \boldsymbol{z}_0$, $\tilde{\boldsymbol{z}}_{-1} = \boldsymbol{z}_0$, step size $h > 0$, and vanishing regularization $\delta_k = 8/(k+64)$, is defined as

$$\tilde{\boldsymbol{z}}_k = \boldsymbol{z}_k - hF(\tilde{\boldsymbol{z}}_{k-1}) + \delta_k(\boldsymbol{a} - \tilde{\boldsymbol{z}}_{k-1}), \qquad \boldsymbol{z}_{k+1} = \boldsymbol{z}_k - hF(\tilde{\boldsymbol{z}}_k) + \delta_k(\boldsymbol{a} - \tilde{\boldsymbol{z}}_k).$$

For every method above, the step size (h, or the initial EAG-V step α_0) is tuned by grid search against the certified global Lipschitz constant of each experiment, as described in Section 3. MA^{4}EG itself requires no such tuning; its update rule is given in Algorithm 1.

B OMITTED PROOFS OF SECTION 2

We give the full proofs of Theorem 1 and Theorem 2, whose statements and proof outline already appeared in Section 2. All identities involving the adaptive clock are used before termination; if a zero is reached, the residual bounds hold immediately with the constant continuation specified in Section 2.

For a fixed anchor $\boldsymbol{a} \in \mathbb{R}^d$, recall the Lyapunov root potential

$$\Phi_k^{\boldsymbol{a}}(t) := \frac{t^2}{2}\|F(\boldsymbol{z}_k)\|^2 + t\,\langle F(\boldsymbol{z}_k), \boldsymbol{z}_k - \boldsymbol{a}\rangle\,. \tag{16}$$

By construction, $\Phi_k^{\boldsymbol{a}_k}(\tau_k) = 0$, as in equation 4. The MA^{4}EG updates in Algorithm 1 give

$$\boldsymbol{z}_{k+1} - \boldsymbol{a}_k = (1-\beta_k)(\boldsymbol{z}_k - \boldsymbol{a}_k) - h_k F(\boldsymbol{y}_k), \tag{17}$$

$$\boldsymbol{z}_k - \boldsymbol{z}_{k+1} = \beta_k(\boldsymbol{z}_k - \boldsymbol{a}_k) + h_k F(\boldsymbol{y}_k), \tag{18}$$

$$\boldsymbol{z}_{k+1} - \boldsymbol{y}_k = -h_k F(\boldsymbol{y}_k) + (1-\beta_k) h_k F(\boldsymbol{z}_k), \tag{19}$$

together with $\tau_k\beta_k = (1-\beta_k)h_k$. From equation 18, $\boldsymbol{z}_k - \boldsymbol{a}_k = \frac{\tau_k+h_k}{h_k}(\boldsymbol{z}_k - \boldsymbol{z}_{k+1}) - (\tau_k+h_k)F(\boldsymbol{y}_k)$, and then equation 17 gives $\boldsymbol{z}_{k+1} - \boldsymbol{a}_k = \frac{\tau_k}{h_k}(\boldsymbol{z}_k - \boldsymbol{z}_{k+1}) - (\tau_k+h_k)F(\boldsymbol{y}_k)$. Substituting these two relations into $\Phi_k^{\boldsymbol{a}_k}(\tau_k) - \Phi_{k+1}^{\boldsymbol{a}_k}(\tau_k+h_k)$ and collecting the terms with $\boldsymbol{z}_k - \boldsymbol{z}_{k+1}$ gives

$$\begin{aligned}\Phi_k^{\boldsymbol{a}_k}(\tau_k) - \Phi_{k+1}^{\boldsymbol{a}_k}(\tau_k+h_k) = {}& \frac{\tau_k(\tau_k+h_k)}{h_k}\langle F(\boldsymbol{z}_k) - F(\boldsymbol{z}_{k+1}), \boldsymbol{z}_k - \boldsymbol{z}_{k+1}\rangle \\ & + \frac{1}{2}\|\tau_k F(\boldsymbol{z}_k) - (\tau_k+h_k)F(\boldsymbol{y}_k)\|^2 \\ & - \frac{(\tau_k+h_k)^2}{2}\|F(\boldsymbol{z}_{k+1}) - F(\boldsymbol{y}_k)\|^2. \end{aligned}\tag{20}$$

The last two operator terms arise from completing the square after the linear terms in $F(\boldsymbol{y}_k)$ are grouped with $\frac{\tau_k^2}{2}\|F(\boldsymbol{z}_k)\|^2 - \frac{(\tau_k+h_k)^2}{2}\|F(\boldsymbol{z}_{k+1})\|^2$. Using equation 19 and $(1-\beta_k)(\tau_k+h_k) = \tau_k$, we have $\tau_k F(\boldsymbol{z}_k) - (\tau_k+h_k)F(\boldsymbol{y}_k) = \frac{\tau_k+h_k}{h_k}(\boldsymbol{z}_{k+1} - \boldsymbol{y}_k)$, so equation 20 becomes

$$\begin{aligned}\Phi_k^{\boldsymbol{a}_k}(\tau_k) - \Phi_{k+1}^{\boldsymbol{a}_k}(\tau_k+h_k) = {}& \frac{\tau_k(\tau_k+h_k)}{h_k}\langle F(\boldsymbol{z}_k) - F(\boldsymbol{z}_{k+1}), \boldsymbol{z}_k - \boldsymbol{z}_{k+1}\rangle \\ & + \frac{(\tau_k+h_k)^2}{2}\left(\frac{\|\boldsymbol{z}_{k+1} - \boldsymbol{y}_k\|^2}{h_k^2} - \|F(\boldsymbol{z}_{k+1}) - F(\boldsymbol{y}_k)\|^2\right). \end{aligned}\tag{21}$$

For a nonterminal iterate with $F(\boldsymbol{z}_{k+1}) \neq 0$, we have $\Phi_{k+1}^{\boldsymbol{a}_k}(\tau_k+h_k) = (\tau_k+h_k)\left[\frac{\tau_k+h_k}{2}\|F(\boldsymbol{z}_{k+1})\|^2 + \langle F(\boldsymbol{z}_{k+1}), \boldsymbol{z}_{k+1} - \boldsymbol{a}_k\rangle\right]$. Using the definition of w_k and the anchor update in Algorithm 1,

$$\Phi_{k+1}^{\boldsymbol{a}_{k+1}}(\tau_k+h_k) = \Phi_{k+1}^{\boldsymbol{a}_k}(\tau_k+h_k) - (\tau_k+h_k)w_k \le 0,$$

because w_k is exactly the positive part subtracted from the bracketed quantity above. If $F(\boldsymbol{z}_{k+1}) = 0$, the method has already reached a zero and no further clock estimate is needed. Otherwise, τ_{k+1} is the largest nonnegative root of $\Phi_{k+1}^{\boldsymbol{a}_{k+1}}$ and $\Phi_{k+1}^{\boldsymbol{a}_{k+1}}(0) = 0$, so

$$\tau_{k+1} \ge \tau_k + h_k. \tag{22}$$

Turning to the step sizes, since F is L-Lipschitz, $\ell_k \le L$. Thus $\widehat{R}_k \le \max\{\widehat{R}_0, L\} \le \max\{R_0, L\}$ for all k, and the update rule for R_{k+1} gives $R_k \le \max\{R_0, L\}$ by induction. Therefore

$$h_k \ge \underline{h} = \frac{\nu}{\gamma \max\{R_0, L\}} \qquad \forall\, k, \tag{23}$$

and, since $R_k \ge \widehat{R}_k \ge \widehat{R}_0 > 0$,

$$h_k \le \overline{h} = \frac{\nu}{\gamma \widehat{R}_0} \qquad \forall\, k. \tag{24}$$

Both bounds follow by induction: they hold at $k = 0$ by the choice of h_0, and the update rule for h_{k+1} preserves them at every step.

Call iteration k *unsafe* if $h_k\ell_k > \nu$. If k is unsafe, then $\ell_k > \nu/h_k \ge \gamma R_k \ge \gamma\widehat{R}_k$, so $\widehat{R}_{k+1} \ge \ell_k > \gamma\widehat{R}_k$: every unsafe iteration increases $\widehat{R}_k$ by at least a factor of γ. If $i_1 < i_2 < \cdots < i_m$ are the unsafe iterations among the first k, then $\widehat{R}_{i_1} \ge \widehat{R}_0$, so $\ell_{i_1} > \gamma\widehat{R}_0$ and $\widehat{R}_{i_1+1} > \gamma\widehat{R}_0$; since $\widehat{R}_k$ is nondecreasing, $\widehat{R}_{i_2} \ge \widehat{R}_{i_1+1} > \gamma\widehat{R}_0$, and hence $\ell_{i_2} > \gamma^2\widehat{R}_0$. Continuing in this way gives $\ell_{i_s} > \gamma^s\widehat{R}_0$ for $s = 1, \ldots, m$, and since $\ell_{i_s} \le L$, every unsafe index satisfies $\gamma^s\widehat{R}_0 < L$. Therefore the number of unsafe iterations among any prefix is at most

$$J := \max\left\{0, \left\lceil \log_\gamma\left(\frac{L}{\widehat{R}_0}\right)\right\rceil - 1\right\}. \tag{25}$$

On a safe step, $h_k \ell_k \le \nu < 1$, so $\|F(\boldsymbol{z}_{k+1}) - F(\boldsymbol{y}_k)\| \le \frac{\nu}{h_k}\|\boldsymbol{z}_{k+1} - \boldsymbol{y}_k\|$. By equation 21 and monotonicity of F, this gives $\Phi_{k+1}^{\boldsymbol{a}_k}(\tau_k + h_k) \le 0$, so $w_k = 0$ and

$$\boldsymbol{a}_{k+1} = \boldsymbol{a}_k \qquad \text{whenever } h_k \ell_k \le \nu. \tag{26}$$

We next bound the anchor. Let b_k denote the number of unsafe iterations before iteration k.

Lemma 1 (Anchor claim)**.** *For every $k \ge 0$,*

$$\|\boldsymbol{a}_k - \boldsymbol{z}^\star\| \le C_{\mathrm{anc}}^{b_k}\|\boldsymbol{z}_0 - \boldsymbol{z}^\star\|, \qquad \|\boldsymbol{z}_k - \boldsymbol{a}_k\| \le \frac{2(1+\nu)}{1-\nu} C_{\mathrm{anc}}^{b_k}\|\boldsymbol{z}_0 - \boldsymbol{z}^\star\|,$$

where C_{anc} is as defined in Theorem 1.

Proof. We proceed by induction on k. The case $k = 0$ follows from $\boldsymbol{a}_0 = \boldsymbol{z}_0$.

Assume the claim holds at iteration k. The root identity $\Phi_k^{\boldsymbol{a}_k}(\tau_k) = 0$ and monotonicity against $\boldsymbol{z}^\star$ give

$$\tau_k\|F(\boldsymbol{z}_k)\| \le 2C_{\mathrm{anc}}^{b_k}\|\boldsymbol{z}_0 - \boldsymbol{z}^\star\| : \tag{27}$$

if $\tau_k = 0$ this is immediate, and if $\tau_k > 0$, then $\frac{\tau_k}{2}\|F(\boldsymbol{z}_k)\|^2 = \langle F(\boldsymbol{z}_k), \boldsymbol{a}_k - \boldsymbol{z}_k\rangle = \langle F(\boldsymbol{z}_k), \boldsymbol{a}_k - \boldsymbol{z}^\star\rangle + \langle F(\boldsymbol{z}_k), \boldsymbol{z}^\star - \boldsymbol{z}_k\rangle \le \|F(\boldsymbol{z}_k)\|\,\|\boldsymbol{a}_k - \boldsymbol{z}^\star\|$, and the induction hypothesis gives equation 27. We also use the elementary estimate $\|\boldsymbol{z}_k - \boldsymbol{a}_k + \tau_k F(\boldsymbol{z}_k)\| \le \|\boldsymbol{z}_k - \boldsymbol{a}_k\|$, which is an equality if $\tau_k = 0$ and otherwise follows from $\langle F(\boldsymbol{z}_k), \boldsymbol{z}_k - \boldsymbol{a}_k\rangle = -\frac{\tau_k}{2}\|F(\boldsymbol{z}_k)\|^2$.

From equation 24, Lipschitz continuity, and the induction hypothesis,

$$\begin{aligned} h_k\|F(\boldsymbol{z}_k)\| &\le h_k L\|\boldsymbol{z}_k - \boldsymbol{z}^\star\| \\ &\le \overline{h}L\left(\|\boldsymbol{z}_k - \boldsymbol{a}_k\| + \|\boldsymbol{a}_k - \boldsymbol{z}^\star\|\right) \\ &\le \overline{h}L\left(1 + \frac{2(1+\nu)}{1-\nu}\right) C_{\mathrm{anc}}^{b_k}\|\boldsymbol{z}_0 - \boldsymbol{z}^\star\|. \end{aligned} \tag{28}$$

Furthermore, using the update rule for $\boldsymbol{y}_k$ and the estimate above,

$$\begin{aligned} (\tau_k + h_k)\|F(\boldsymbol{y}_k) - F(\boldsymbol{z}_k)\| &\le (\tau_k + h_k)L\|\boldsymbol{y}_k - \boldsymbol{z}_k\| \\ &= h_k L\|-\boldsymbol{z}_k + \boldsymbol{a}_k - \tau_k F(\boldsymbol{z}_k)\| \\ &\le h_k L\|\boldsymbol{z}_k - \boldsymbol{a}_k\| \\ &\le \overline{h}L\frac{2(1+\nu)}{1-\nu} C_{\mathrm{anc}}^{b_k}\|\boldsymbol{z}_0 - \boldsymbol{z}^\star\|. \end{aligned} \tag{29}$$

Combining equation 27, equation 28, and equation 29,

$$\begin{aligned} (\tau_k + h_k)\|F(\boldsymbol{y}_k)\| &\le \left[2 + \overline{h}L\left(1 + \frac{4(1+\nu)}{1-\nu}\right)\right] C_{\mathrm{anc}}^{b_k}\|\boldsymbol{z}_0 - \boldsymbol{z}^\star\| \\ &= G C_{\mathrm{anc}}^{b_k}\|\boldsymbol{z}_0 - \boldsymbol{z}^\star\|, \end{aligned} \tag{30}$$

where G is as defined in Theorem 1.

If k is safe, then $h_k\ell_k \le \nu$, and by equation 26, $\boldsymbol{a}_{k+1} = \boldsymbol{a}_k$ and $b_{k+1} = b_k$. The inequality $\Phi_{k+1}^{\boldsymbol{a}_k}(\tau_k + h_k) \le 0$ gives $(\tau_k + h_k)\|F(\boldsymbol{z}_{k+1})\| \le 2\|\boldsymbol{a}_k - \boldsymbol{z}^\star\| \le 2C_{\mathrm{anc}}^{b_k}\|\boldsymbol{z}_0 - \boldsymbol{z}^\star\|$. Using the relation $(1-\beta_k)h_k = \beta_k\tau_k$ and the definition of ℓ_k,

$$\begin{aligned} \|\boldsymbol{z}_{k+1} - \boldsymbol{y}_k\| &= \|\beta_k\tau_k F(\boldsymbol{z}_k) - h_k F(\boldsymbol{z}_{k+1}) + h_k(F(\boldsymbol{z}_{k+1}) - F(\boldsymbol{y}_k))\| \\ &\le 2\beta_k C_{\mathrm{anc}}^{b_k}\|\boldsymbol{z}_0 - \boldsymbol{z}^\star\| + 2\beta_k C_{\mathrm{anc}}^{b_k}\|\boldsymbol{z}_0 - \boldsymbol{z}^\star\| + h_k\ell_k\|\boldsymbol{z}_{k+1} - \boldsymbol{y}_k\|, \end{aligned}$$

so $\|\boldsymbol{z}_{k+1} - \boldsymbol{y}_k\| \le \frac{4\beta_k}{1-\nu} C_{\mathrm{anc}}^{b_k}\|\boldsymbol{z}_0 - \boldsymbol{z}^\star\|$. Then

$$\begin{aligned} \|\boldsymbol{z}_{k+1} - \boldsymbol{a}_k\| &\le (1-\beta_k)\|\boldsymbol{z}_k - \boldsymbol{a}_k\| + 2\beta_k C_{\mathrm{anc}}^{b_k}\|\boldsymbol{z}_0 - \boldsymbol{z}^\star\| + h_k\ell_k\|\boldsymbol{z}_{k+1} - \boldsymbol{y}_k\| \\ &\le \frac{2(1+\nu)}{1-\nu} C_{\mathrm{anc}}^{b_k}\|\boldsymbol{z}_0 - \boldsymbol{z}^\star\|, \end{aligned}$$

and since $\boldsymbol{a}_{k+1} = \boldsymbol{a}_k$, the induction inequalities hold at $k+1$.

If instead k is unsafe, then $h_k \ell_k > \nu$, and by equation 30, $h_k \|F(\boldsymbol{y}_k)\| \le G C_{\mathrm{anc}}^{b_k} \|\boldsymbol{z}_0 - \boldsymbol{z}^\star\|$. Hence $\|\boldsymbol{z}_{k+1} - \boldsymbol{a}_k\| \le \left(\frac{2(1+\nu)}{1-\nu} + G\right) C_{\mathrm{anc}}^{b_k} \|\boldsymbol{z}_0 - \boldsymbol{z}^\star\|$, and, similarly to equation 30,

$$(\tau_k + h_k)\|F(\boldsymbol{z}_{k+1})\| \le \left(G + \overline{h}LG + 2\overline{h}L\right) C_{\mathrm{anc}}^{b_k} \|\boldsymbol{z}_0 - \boldsymbol{z}^\star\|.$$

The anchor correction length satisfies

$$\begin{aligned}
\|\boldsymbol{a}_{k+1} - \boldsymbol{a}_k\| &= \frac{w_k}{\|F(\boldsymbol{z}_{k+1})\|} \\
&\le \frac{\tau_k + h_k}{2}\|F(\boldsymbol{z}_{k+1})\| + \|\boldsymbol{z}_{k+1} - \boldsymbol{a}_k\| \\
&\le \left[\frac{1}{2}\left(G + \overline{h}LG + 2\overline{h}L\right) + \frac{2(1+\nu)}{1-\nu} + G\right] C_{\mathrm{anc}}^{b_k} \|\boldsymbol{z}_0 - \boldsymbol{z}^\star\|,
\end{aligned}$$

and, if $F(\boldsymbol{z}_{k+1}) = 0$, the algorithm stops and the same bound is harmless. Combining this with the induction hypothesis gives

$$\|\boldsymbol{a}_{k+1} - \boldsymbol{z}^\star\| \le \left[1 + \frac{2(1+\nu)}{1-\nu} + \frac{3}{2}G + \frac{1}{2}\overline{h}LG + \overline{h}L\right] C_{\mathrm{anc}}^{b_k} \|\boldsymbol{z}_0 - \boldsymbol{z}^\star\|.$$

By the definition of C_{anc} in Theorem 1,

$$C_{\mathrm{anc}} - \left[1 + \frac{2(1+\nu)}{1-\nu} + \frac{3}{2}G + \frac{1}{2}\overline{h}LG + \overline{h}L\right] = \frac{2(1+\nu)}{1-\nu} + G - 1 > 0,$$

and since $b_{k+1} = b_k + 1$ on an unsafe step, $\|\boldsymbol{a}_{k+1} - \boldsymbol{z}^\star\| \le C_{\mathrm{anc}}^{b_{k+1}} \|\boldsymbol{z}_0 - \boldsymbol{z}^\star\|$. Finally, $\|\boldsymbol{z}_{k+1} - \boldsymbol{a}_{k+1}\| \le \|\boldsymbol{z}_{k+1} - \boldsymbol{a}_k\| + \|\boldsymbol{a}_k - \boldsymbol{a}_{k+1}\| \le \frac{2(1+\nu)}{1-\nu} C_{\mathrm{anc}}^{b_{k+1}} \|\boldsymbol{z}_0 - \boldsymbol{z}^\star\|$, because the sum of the two bounds above is exactly C_{anc}, and $2(1+\nu)/(1-\nu) > 1$. This closes the induction, in both the safe and unsafe cases. □

B.1 Proof of Theorem 1

With Lemma 1 in hand, we can prove Theorem 1. If $F(\boldsymbol{z}_k) = 0$, the result is immediate. Otherwise, by $\Phi_k^{\boldsymbol{a}_k}(\tau_k) = 0$,

$$\frac{\tau_k}{2}\|F(\boldsymbol{z}_k)\|^2 = \langle F(\boldsymbol{z}_k), \boldsymbol{a}_k - \boldsymbol{z}_k\rangle = \langle F(\boldsymbol{z}_k), \boldsymbol{a}_k - \boldsymbol{z}^\star\rangle + \langle F(\boldsymbol{z}_k), \boldsymbol{z}^\star - \boldsymbol{z}_k\rangle.$$

Monotonicity gives $\langle F(\boldsymbol{z}_k), \boldsymbol{z}_k - \boldsymbol{z}^\star\rangle \ge 0$, hence

$$\|F(\boldsymbol{z}_k)\| \le \frac{2}{\tau_k}\|\boldsymbol{a}_k - \boldsymbol{z}^\star\|. \tag{31}$$

By equation 22 and equation 23, $\tau_k \ge \sum_{i=0}^{k-1} h_i \ge k\underline{h}$. Using Lemma 1 and $b_k \le J$ from equation 25 gives

$$\|F(\boldsymbol{z}_k)\| \le \frac{2C_{\mathrm{anc}}^J}{k\underline{h}}\|\boldsymbol{z}_0 - \boldsymbol{z}^\star\|,$$

which is the claim of Theorem 1.

B.2 Proof of Theorem 2

We now turn to Theorem 2, additionally assuming F is μ-strongly monotone. On a safe iteration, $h_k \ell_k \le \nu$ and $\boldsymbol{a}_{k+1} = \boldsymbol{a}_k$. From equation 21, strong monotonicity, and the definition of ℓ_k,

$$-\Phi_{k+1}^{\boldsymbol{a}_k}(\tau_k + h_k) \ge \tau_k(\tau_k + h_k)\left[\frac{\mu}{h_k}\|\boldsymbol{z}_k - \boldsymbol{z}_{k+1}\|^2 + \frac{1-\nu^2}{2h_k^2}\|\boldsymbol{z}_{k+1} - \boldsymbol{y}_k\|^2\right]. \tag{32}$$

We relate the right side to $\|F(\boldsymbol{z}_{k+1})\|^2$. On a safe iteration,

$$\begin{aligned}
\|F(\boldsymbol{z}_{k+1})\| &\le \|F(\boldsymbol{z}_{k+1}) - F(\boldsymbol{y}_k)\| + \|F(\boldsymbol{y}_k)\| \\
&\le \frac{1+\nu}{h_k}\|\boldsymbol{z}_{k+1} - \boldsymbol{y}_k\| + (1 - \beta_k)\|F(\boldsymbol{z}_k)\|.
\end{aligned} \tag{33}$$

If $\tau_k = 0$, then $\beta_k = 1$, so $(1-\beta_k)\|F(\boldsymbol{z}_k)\| = 0$. If $\tau_k > 0$, then $\Phi_k^{\boldsymbol{a}_k}(\tau_k) = 0$ gives $\langle F(\boldsymbol{z}_k), \boldsymbol{a}_k - \boldsymbol{z}_k\rangle = \frac{\tau_k}{2}\|F(\boldsymbol{z}_k)\|^2$, and, using $\boldsymbol{y}_k = (1-\beta_k)\boldsymbol{z}_k + \beta_k\boldsymbol{a}_k - (1-\beta_k)h_kF(\boldsymbol{z}_k)$ and $\tau_k\beta_k = (1-\beta_k)h_k$, we obtain $\langle F(\boldsymbol{z}_k), \boldsymbol{y}_k - \boldsymbol{z}_k\rangle = -\frac{(1-\beta_k)h_k}{2}\|F(\boldsymbol{z}_k)\|^2$. In both cases, $(1-\beta_k)\|F(\boldsymbol{z}_k)\| \le \frac{2}{h_k}\|\boldsymbol{y}_k - \boldsymbol{z}_k\|$, so equation 33 gives

$$\|F(\boldsymbol{z}_{k+1})\| \le \frac{3+\nu}{h_k}\|\boldsymbol{z}_{k+1} - \boldsymbol{y}_k\| + \frac{2}{h_k}\|\boldsymbol{z}_{k+1} - \boldsymbol{z}_k\|. \tag{34}$$

By the weighted Cauchy inequality, with weights $\omega_1 = \mu h_k/4$ and $\omega_2 = (1-\nu^2)/(2(3+\nu)^2)$,

$$\begin{aligned}\|F(\boldsymbol{z}_{k+1})\|^2 &\le \left(\frac{4}{\mu h_k} + \frac{2(3+\nu)^2}{1-\nu^2}\right)\left(\frac{\mu}{h_k}\|\boldsymbol{z}_k - \boldsymbol{z}_{k+1}\|^2 + \frac{1-\nu^2}{2h_k^2}\|\boldsymbol{z}_{k+1} - \boldsymbol{y}_k\|^2\right)\\ &\le \Theta\left(\frac{\mu}{h_k}\|\boldsymbol{z}_k - \boldsymbol{z}_{k+1}\|^2 + \frac{1-\nu^2}{2h_k^2}\|\boldsymbol{z}_{k+1} - \boldsymbol{y}_k\|^2\right),\end{aligned} \tag{35}$$

where Θ is as defined in Theorem 2, and the last inequality uses $h_k \ge \underline{h}$. Combining equation 32 and equation 35,

$$-\Phi_{k+1}^{\boldsymbol{a}_k}(\tau_k + h_k) \ge \frac{\tau_k(\tau_k + h_k)}{\Theta}\|F(\boldsymbol{z}_{k+1})\|^2. \tag{36}$$

On the other hand, since $\boldsymbol{a}_{k+1} = \boldsymbol{a}_k$ on a safe iteration and $\Phi_{k+1}^{\boldsymbol{a}_{k+1}}(\tau_{k+1}) = 0$,

$$-\Phi_{k+1}^{\boldsymbol{a}_{k+1}}(\tau_k + h_k) = \frac{(\tau_k + h_k)(\tau_{k+1} - \tau_k - h_k)}{2}\|F(\boldsymbol{z}_{k+1})\|^2. \tag{37}$$

Combining equation 36 and equation 37, and recalling $q = 1 + 2/\Theta$,

$$\tau_{k+1} \ge q\tau_k + h_k \qquad \text{on safe iterations.} \tag{38}$$

For the first k iterations, let b_k be the number of unsafe iterations and $s_k = k - b_k$ the number of safe iterations. By equation 25, $b_k \le J$, hence $s_k \ge \max\{0, k-J\} = m_k$. Every iteration satisfies $\tau_{i+1} \ge \tau_i + h_i$, while safe iterations satisfy equation 38. For a fixed number s_k of safe iterations, the smallest lower bound on τ_k is obtained by placing all safe iterations before all unsafe ones: with $T_s(t) := qt + \underline{h}$ and $T_u(t) := t + \underline{h}$, $T_s(T_u(t)) - T_u(T_s(t)) = (q-1)\underline{h} \ge 0$, so moving a safe update earlier cannot increase the lower bound. Thus

$$\tau_k \ge \left[\sum_{i=0}^{s_k-1} q^i + k - s_k\right]\underline{h}.$$

The function $\psi(s) := \left[\sum_{i=0}^{s-1} q^i + k - s\right]\underline{h}$ is nondecreasing in s, because $\psi(s+1) - \psi(s) = (q^s - 1)\underline{h} \ge 0$. Hence

$$\tau_k \ge \left[\frac{\Theta}{2}\left(q^{m_k} - 1\right) + k - m_k\right]\underline{h}. \tag{39}$$

Finally, equation 31, Lemma 1, and equation 39 give

$$\|F(\boldsymbol{z}_k)\| \le \frac{2C_{\mathrm{anc}}^J\|\boldsymbol{z}_0 - \boldsymbol{z}^\star\|}{\underline{h}\left[k - m_k + \frac{\Theta}{2}\left(q^{m_k} - 1\right)\right]},$$

which proves Theorem 2.

Comparison of the convergence bounds. For a strongly convex-concave objective, the first bound in Theorem 2 is no worse than the convex-concave bound in Theorem 1 at every iteration. Indeed, since $q = 1 + 2/\Theta > 1$, for every $k \ge 1$,

$$k - m_k + \frac{\Theta}{2}\left(q^{m_k} - 1\right) = k + \sum_{i=0}^{m_k-1}\left(q^i - 1\right) \ge k.$$

Equality holds when $m_k \le 1$, and the inequality is strict when $m_k \ge 2$. Thus the two bounds coincide for $k \le J+1$, while the strongly convex-concave bound is strictly tighter for $k \ge J+2$ whenever $\boldsymbol{z}_0 \ne \boldsymbol{z}^\star$. Since J is independent of k, the denominator in the strongly convex-concave bound grows geometrically as $k \to \infty$, whereas the denominator in the convex-concave bound grows only linearly.

C AN AFFINE-POPOV GENERALIZATION OF MA^{4}EG

The moving-anchor construction of MA^{4}EG also accommodates a predictor that reuses a past operator evaluation. In this appendix we develop an *affine-Popov extension* of MA^{4}EG, denoted MA^{4}EG-AP. It retains the adaptive step size, the adaptive anchor coefficient, and the anchor correction of Algorithm 1, and adds a memory parameter $\theta \in [0, 1]$. At $\theta = 0$, the method reduces exactly to MA^{4}EG. For $\theta > 0$, the predictor blends the current operator value with the operator value at the previous predictor. The resulting method preserves the accelerated last-iterate rate for convex-concave objectives and automatically achieves linear convergence under strong convexity-concavity, without knowledge of L or μ.

Popov's method and motivation. Popov's method (Popov, 1980), also called past extragradient, reuses the previous predictor evaluation in the first stage. In the unconstrained operator formulation, with a fixed step size $h > 0$, its updates can be written as

$$\boldsymbol{y}_k = \boldsymbol{z}_k - hF(\boldsymbol{y}_{k-1}), \qquad \boldsymbol{z}_{k+1} = \boldsymbol{z}_k - hF(\boldsymbol{y}_k). \tag{40}$$

Caching $F(\boldsymbol{y}_{k-1})$ leaves one new operator evaluation per iteration. Anchoring can accelerate this basic construction: Tran-Dinh & Luo (2021) give an anchored Popov method with an $\mathcal{O}(1/k^2)$ last-iterate rate for the squared operator residual. These results raise the question of whether the moving anchor and the simultaneous adaptation to smoothness and strong convexity-concavity in MA^{4}EG can also support a predictor with memory.

Simply replacing $F(\boldsymbol{z}_k)$ by $F(\boldsymbol{y}_{k-1})$ in MA^{4}EG introduces an additional error in its one-step Lyapunov relation equation 21. We absorb this error by augmenting the root potential equation 16 with a memory term that is nonnegative at the selected root, and by choosing the memory weight from the progress of the adaptive clock. The anchor correction continues to guarantee clock growth even when an iteration encounters unexpectedly large local curvature. Thus the extension uses the same proof mechanism as MA^{4}EG: a root potential, safe and unsafe iterations, finitely many anchor corrections, and faster clock growth under strong monotonicity.

We use the notation of Section 1.1. In particular, $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^d$ with $d = d_x + d_y$, and F is the saddle operator equation 2 of $\mathcal{L}$, so that $\|F(\boldsymbol{z})\| = \|\nabla\mathcal{L}(\boldsymbol{z})\|$ and a saddle point $\boldsymbol{z}^\star = (\boldsymbol{x}^\star, \boldsymbol{y}^\star)$ satisfies $F(\boldsymbol{z}^\star) = 0$. As in Algorithm 1, $\boldsymbol{y}_k \in \mathbb{R}^d$ denotes the predictor; it is distinct from the dual variable $\boldsymbol{y} \in \mathbb{R}^{d_y}$ of equation 1. The proofs apply more generally to any monotone Lipschitz operator with a zero.

C.1 ALGORITHM AND RELATION TO MA^{4}EG

In addition to the MA^{4}EG state $(\boldsymbol{z}_k, \boldsymbol{a}_k, h_k, \widehat{R}_k, R_k)$, MA^{4}EG-AP maintains the cached value $F(\boldsymbol{y}_{k-1})$ and a *certified clock* $S_k \geq 0$. The adaptive clock τ_k determines $\beta_k = h_k/(\tau_k + h_k)$, as before. The additional memory weight is

$$\lambda_k = \begin{cases} (\tau_k - S_k)/\tau_k, & \tau_k > 0, \\ 0, & \tau_k = 0, \end{cases} \qquad \theta_k = \theta\sqrt{\lambda_k}.$$

We show in equation 46 that $\tau_k \geq S_k$, so that $0 \leq \theta_k \leq \theta \leq 1$. Only the predictor direction and the root rule change; the smoothness estimates and the anchor correction are those of MA^{4}EG. The method is summarized in Algorithm 2.

If $\theta = 0$, then $M_k = F(\boldsymbol{z}_k)$ and

$$\tau_k = \frac{2[\langle F(\boldsymbol{z}_k), \boldsymbol{a}_k - \boldsymbol{z}_k\rangle]_+}{\|F(\boldsymbol{z}_k)\|^2}.$$

Algorithm 2 therefore reproduces every update of Algorithm 1, with S_k an unused auxiliary variable. The analysis below uses $\nu < 1/\sqrt{2}$ to control the memory errors. At $\theta = 0$ these errors vanish, and the full range $\nu \in (0, 1)$ and the sharper constants of Theorems 1 and 2 apply. For $\theta > 0$, the choice $\theta = 1$ permits the largest memory weight in this family, but θ_k is still adaptive, so the method does not coincide with the classical Popov iteration equation 40.

With cached values, each iteration uses at most two new operator evaluations, $F(\boldsymbol{y}_k)$ and $F(\boldsymbol{z}_{k+1})$, just as MA^{4}EG. The latter is needed for the root rule, the smoothness estimate, and the anchor

Algorithm 2 MA^{4}EG-AP: affine-Popov extension of MA^{4}EG

Choose $\theta \in [0,1]$, $\nu \in (0, 1/\sqrt{2})$, $\gamma > 1$, $\rho > 1$, and $0 < \widehat{R}_0 < R_0$.
Initialize $\boldsymbol{z}_0 \in \mathbb{R}^d$, $\boldsymbol{a}_0 = \boldsymbol{z}_0$, $\boldsymbol{y}_{-1} = \boldsymbol{z}_0$, $S_0 = 0$, and $h_0 = \nu/(\gamma R_0)$.
for $k = 0, 1, 2, \ldots$ **do**
  If $F(\boldsymbol{z}_k) = 0$, stop.
  $\tau_k = \dfrac{2\big[\theta^2 S_k \|F(\boldsymbol{z}_k) - F(\boldsymbol{y}_{k-1})\|^2 + \langle F(\boldsymbol{z}_k), \boldsymbol{a}_k - \boldsymbol{z}_k\rangle\big]_+}{\|F(\boldsymbol{z}_k)\|^2 + 2\theta^2 \|F(\boldsymbol{z}_k) - F(\boldsymbol{y}_{k-1})\|^2}, \quad \beta_k = \dfrac{h_k}{\tau_k + h_k}.$
  $\lambda_k = (\tau_k - S_k)/\tau_k$ if $\tau_k > 0$, and $\lambda_k = 0$ otherwise; set $\theta_k = \theta\sqrt{\lambda_k}$.
  $M_k = (1 - \theta_k) F(\boldsymbol{z}_k) + \theta_k F(\boldsymbol{y}_{k-1})$.
  $\boldsymbol{y}_k = (1 - \beta_k)\boldsymbol{z}_k + \beta_k \boldsymbol{a}_k - (1 - \beta_k) h_k M_k$.
  $\boldsymbol{z}_{k+1} = (1 - \beta_k)\boldsymbol{z}_k + \beta_k \boldsymbol{a}_k - h_k F(\boldsymbol{y}_k)$.
  $\ell_k = \|F(\boldsymbol{z}_{k+1}) - F(\boldsymbol{y}_k)\| / \|\boldsymbol{z}_{k+1} - \boldsymbol{y}_k\|$, where a ratio with zero denominator is set to 0.
  $\widehat{R}_{k+1} = \max\{\widehat{R}_k, \ell_k\}, \quad R_{k+1} = \max\{\widehat{R}_{k+1},\, R_k/\rho\}$.
  $h_{k+1} = \min\{\rho h_k,\, \nu/(\gamma R_{k+1})\}$; set $S_{k+1} = \tau_k + h_k$.
  $w_k = \left[\dfrac{\tau_k + h_k}{2}\|F(\boldsymbol{z}_{k+1})\|^2 + \langle F(\boldsymbol{z}_{k+1}), \boldsymbol{z}_{k+1} - \boldsymbol{a}_k\rangle\right]_+.$
  $\boldsymbol{a}_{k+1} = \boldsymbol{a}_k + \dfrac{w_k}{\|F(\boldsymbol{z}_{k+1})\|^2} F(\boldsymbol{z}_{k+1})$, where a ratio with zero denominator is set to 0.
end for

correction. The extension therefore carries Popov's memory mechanism into the adaptive framework while retaining the evaluation budget of MA^{4}EG. Every computed iterate is accepted; there is no line search or restart. As in Section 2, the iterates are kept constant after a zero is reached.

C.2 Convergence guarantees

The step-size bounds and the bound on the number of unsafe iterations are the same as for MA^{4}EG. The constants $C_{s,\theta}$, G_θ, $C_{\mathrm{anc},\theta}$, Θ_θ, and q_θ below belong to the extension and do not replace the constants in Theorems 1 and 2. The anchor constants $C_{s,\theta}$, G_θ, and $C_{\mathrm{anc},\theta}$ are in fact uniform over $\theta \in [0,1]$; only Θ_θ and q_θ depend on θ.

Theorem 3 (Convex-concave: affine-Popov extension)**.** *Assume $\mathcal{L} : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R}$ is convex-concave and L-smooth for some $L > 0$, and admits a saddle point $\boldsymbol{z}^\star = (\boldsymbol{x}^\star, \boldsymbol{y}^\star)$. Let $(\boldsymbol{z}_k)_{k\geq 0}$ be generated by Algorithm 2, with the iterates kept constant after termination. Let*

$$\overline{h} := \frac{\nu}{\gamma \widehat{R}_0}, \qquad \underline{h} := \frac{\nu}{\gamma \max\{R_0, L\}}, \qquad J := \max\left\{0,\ \left\lceil \log_\gamma\left(\frac{L}{\widehat{R}_0}\right)\right\rceil - 1\right\},$$

$$C_{s,\theta} := 2 + \frac{\nu(4 + \sqrt{2})}{1 - \nu}, \qquad G_\theta := 2 + L\overline{h}\left(2C_{s,\theta} + 3 + \sqrt{2}\right),$$

$$C_{\mathrm{anc},\theta} := 1 + 2C_{s,\theta} + L\overline{h}\left(2 + \sqrt{2}\right) + \left(3 + L\overline{h}\right) G_\theta.$$

Then for every $k \geq 1$,

$$\|\nabla \mathcal{L}(\boldsymbol{z}_k)\| \leq \frac{2C_{\mathrm{anc},\theta}^J}{k\underline{h}} \|\boldsymbol{z}_0 - \boldsymbol{z}^\star\|. \tag{41}$$

Consequently, $\|\nabla \mathcal{L}(\boldsymbol{z}_k)\|^2 = \mathcal{O}(k^{-2})$.

Theorem 4 (Strongly convex-concave: affine-Popov extension)**.** *In addition to the hypotheses of Theorem 3, suppose $\mathcal{L}$ is μ-strongly convex-concave for some $\mu > 0$. Let $\underline{h}$, $C_{\mathrm{anc},\theta}$, and J be as defined in Theorem 3, and let*

$$\delta_\nu := \frac{1}{2} - \nu^2 > 0, \qquad \Theta_\theta := \frac{18}{\mu\underline{h}} + \frac{2\left[3\left(\sqrt{6} + 1 + \nu\right)^2 + 2\theta^2\nu^2\right]}{\delta_\nu} + 48, \qquad q_\theta := 1 + \frac{2}{\Theta_\theta}.$$

For $k \geq 1$, let $m_k := \max\{k - J, 0\}$. Then

$$\|\nabla \mathcal{L}(\boldsymbol{z}_k)\| \leq \frac{2C_{\mathrm{anc},\theta}^J \|\boldsymbol{z}_0 - \boldsymbol{z}^\star\|}{\underline{h}\left[k - m_k + \frac{\Theta_\theta}{2}\left(q_\theta^{m_k} - 1\right)\right]} \leq \frac{2C_{\mathrm{anc},\theta}^J \|\boldsymbol{z}_0 - \boldsymbol{z}^\star\|}{\underline{h}\, q_\theta^{\max\{m_k,1\}-1}}. \tag{42}$$

Consequently, $\|\nabla\mathcal{L}(\boldsymbol{z}_k)\| = \mathcal{O}(q_\theta^{-k})$. The distance to the solution has the same geometric rate, since $\|\boldsymbol{z}_k - \boldsymbol{z}^\star\| \le \mu^{-1}\|F(\boldsymbol{z}_k)\|$ by strong monotonicity.

Relation to the main results. These are the same two last-iterate regimes as in Theorems 1 and 2. The memory terms change the constants and the admissible range of ν, while the algorithm still uses neither L nor μ. Moreover, since $q_\theta > 1$,

$$k - m_k + \frac{\Theta_\theta}{2}\left(q_\theta^{m_k} - 1\right) = k + \sum_{i=0}^{m_k-1}\left(q_\theta^i - 1\right) \ge k.$$

Thus the strongly convex-concave bound is no worse than the convex-concave bound at any iteration, with a strictly larger denominator when $m_k \ge 2$. As in Appendix B, the integer J bounds the total number of unsafe iterations; it need not bound the index of the last unsafe iteration.

C.3 Lyapunov potential and certified clock

Write $D_0 := \|\boldsymbol{z}_0 - \boldsymbol{z}^\star\|$ throughout the proofs. All clock identities below concern nonterminal iterations; if a zero is reached, the residual bounds follow immediately from the constant continuation. Recall the MA^{4}EG potential $\Phi_k^{\boldsymbol{a}}$ from equation 16. For the extension, we add a memory term that vanishes at the certified clock: for $\boldsymbol{a} \in \mathbb{R}^d$ and $S \ge 0$, let

$$\widetilde{\Phi}_k^{\boldsymbol{a}}(t; S) := \Phi_k^{\boldsymbol{a}}(t) + \theta^2 t(t - S)\|F(\boldsymbol{z}_k) - F(\boldsymbol{y}_{k-1})\|^2. \tag{43}$$

This is a quadratic in t, and at $\theta = 0$ it is exactly $\Phi_k^{\boldsymbol{a}}$. Along the algorithm we evaluate it at the certified clock S_k and write $\widetilde{\Phi}_k^{\boldsymbol{a}}(t) := \widetilde{\Phi}_k^{\boldsymbol{a}}(t; S_k)$ for any anchor $\boldsymbol{a} \in \mathbb{R}^d$. Before termination, the quadratic $\widetilde{\Phi}_k^{\boldsymbol{a}_k}$ has the roots 0 and $2\big(\theta^2 S_k\|F(\boldsymbol{z}_k) - F(\boldsymbol{y}_{k-1})\|^2 + \langle F(\boldsymbol{z}_k), \boldsymbol{a}_k - \boldsymbol{z}_k\rangle\big)/\big(\|F(\boldsymbol{z}_k)\|^2 + 2\theta^2\|F(\boldsymbol{z}_k) - F(\boldsymbol{y}_{k-1})\|^2\big)$. Hence, by Line 5 of Algorithm 2, τ_k is its largest nonnegative root, and

$$\widetilde{\Phi}_k^{\boldsymbol{a}_k}(\tau_k) = 0. \tag{44}$$

Initially $S_0 = 0$, so the following invariant holds at $k = 0$; we show in Appendix C.5 that the anchor correction preserves it at every subsequent iteration:

$$\widetilde{\Phi}_k^{\boldsymbol{a}_k}(S_k) \le 0. \tag{45}$$

Before termination, the quadratic $\widetilde{\Phi}_k^{\boldsymbol{a}_k}$ has a positive leading coefficient, so its nonpositive sublevel set on $[0, \infty)$ is $[0, \tau_k]$. Therefore equation 45 implies

$$\tau_k \ge S_k. \tag{46}$$

Consequently, $\lambda_k \in [0, 1]$ and $\theta_k \in [0, \theta] \subseteq [0, 1]$, and

$$\tau_k^2\theta_k^2 = \theta^2\tau_k(\tau_k - S_k). \tag{47}$$

C.4 One-step relation

Recall the affine-Popov direction $M_k = (1 - \theta_k)F(\boldsymbol{z}_k) + \theta_k F(\boldsymbol{y}_{k-1})$ from Line 7 of Algorithm 2. Since $1 - \beta_k = \tau_k/(\tau_k + h_k)$, the updates of Algorithm 2 give

$$\boldsymbol{z}_{k+1} - \boldsymbol{a}_k = \frac{\tau_k}{\tau_k + h_k}(\boldsymbol{z}_k - \boldsymbol{a}_k) - h_k F(\boldsymbol{y}_k), \tag{48}$$

$$\boldsymbol{z}_k - \boldsymbol{z}_{k+1} = \frac{h_k}{\tau_k + h_k}(\boldsymbol{z}_k - \boldsymbol{a}_k) + h_k F(\boldsymbol{y}_k), \tag{49}$$

$$\boldsymbol{z}_{k+1} - \boldsymbol{y}_k = h_k\left(\frac{\tau_k}{\tau_k + h_k}M_k - F(\boldsymbol{y}_k)\right). \tag{50}$$

Thus

$$\frac{\tau_k + h_k}{h_k}(\boldsymbol{z}_{k+1} - \boldsymbol{y}_k) = \tau_k F(\boldsymbol{z}_k) - \tau_k\theta_k\big(F(\boldsymbol{z}_k) - F(\boldsymbol{y}_{k-1})\big) - (\tau_k + h_k)F(\boldsymbol{y}_k), \tag{51}$$

so

$$\tau_k F(\boldsymbol{z}_k) - (\tau_k + h_k)F(\boldsymbol{y}_k) = \frac{\tau_k + h_k}{h_k}(\boldsymbol{z}_{k+1} - \boldsymbol{y}_k) + \tau_k\theta_k\big(F(\boldsymbol{z}_k) - F(\boldsymbol{y}_{k-1})\big). \tag{52}$$

As in Appendix B, equation 49 and equation 48 imply

$$\begin{aligned}\boldsymbol{z}_k - \boldsymbol{a}_k &= \frac{\tau_k + h_k}{h_k}(\boldsymbol{z}_k - \boldsymbol{z}_{k+1}) - (\tau_k + h_k)F(\boldsymbol{y}_k),\\ \boldsymbol{z}_{k+1} - \boldsymbol{a}_k &= \frac{\tau_k}{h_k}(\boldsymbol{z}_k - \boldsymbol{z}_{k+1}) - (\tau_k + h_k)F(\boldsymbol{y}_k).\end{aligned}$$

These are the only relations used to derive equation 20, so the non-memory part of $\widetilde{\Phi}_k^{\boldsymbol{a}_k}(\tau_k) - \widetilde{\Phi}_{k+1}^{\boldsymbol{a}_k}(\tau_k + h_k)$ equals the right-hand side of equation 20. The memory terms contribute

$$\theta^2\tau_k(\tau_k - S_k)\|F(\boldsymbol{z}_k) - F(\boldsymbol{y}_{k-1})\|^2,$$

because the new certified clock is $S_{k+1} = \tau_k + h_k$, so the memory term in $\widetilde{\Phi}_{k+1}^{\boldsymbol{a}_k}(\tau_k + h_k) = \widetilde{\Phi}_{k+1}^{\boldsymbol{a}_k}(\tau_k + h_k; S_{k+1})$ vanishes. Using equation 52, we therefore obtain

$$\begin{aligned}\widetilde{\Phi}_k^{\boldsymbol{a}_k}(\tau_k) - \widetilde{\Phi}_{k+1}^{\boldsymbol{a}_k}(\tau_k + h_k) = {}& \frac{\tau_k(\tau_k + h_k)}{h_k}\,\langle F(\boldsymbol{z}_k) - F(\boldsymbol{z}_{k+1}), \boldsymbol{z}_k - \boldsymbol{z}_{k+1}\rangle\\ &+ \frac{1}{2}\left\|\frac{\tau_k + h_k}{h_k}(\boldsymbol{z}_{k+1} - \boldsymbol{y}_k) + \tau_k\theta_k\big(F(\boldsymbol{z}_k) - F(\boldsymbol{y}_{k-1})\big)\right\|^2\\ &- \frac{(\tau_k + h_k)^2}{2}\|F(\boldsymbol{z}_{k+1}) - F(\boldsymbol{y}_k)\|^2\\ &+ \theta^2\tau_k(\tau_k - S_k)\|F(\boldsymbol{z}_k) - F(\boldsymbol{y}_{k-1})\|^2.\end{aligned} \tag{53}$$

Using the elementary inequality $\|P + E\|^2 \geq \frac{1}{2}\|P\|^2 - \|E\|^2$ for $P, E \in \mathbb{R}^d$, together with equation 47, we get

$$\begin{aligned}\widetilde{\Phi}_k^{\boldsymbol{a}_k}(\tau_k) - \widetilde{\Phi}_{k+1}^{\boldsymbol{a}_k}(\tau_k + h_k) \geq {}& \frac{\tau_k(\tau_k + h_k)}{h_k}\,\langle F(\boldsymbol{z}_k) - F(\boldsymbol{z}_{k+1}), \boldsymbol{z}_k - \boldsymbol{z}_{k+1}\rangle\\ &+ \frac{(\tau_k + h_k)^2}{4h_k^2}\|\boldsymbol{z}_{k+1} - \boldsymbol{y}_k\|^2 - \frac{(\tau_k + h_k)^2}{2}\|F(\boldsymbol{z}_{k+1}) - F(\boldsymbol{y}_k)\|^2\\ &+ \frac{\theta^2}{2}\tau_k(\tau_k - S_k)\|F(\boldsymbol{z}_k) - F(\boldsymbol{y}_{k-1})\|^2.\end{aligned} \tag{54}$$

As in Appendix B, call iteration k *safe* if $h_k\ell_k \leq \nu$, and *unsafe* otherwise. On a safe iteration,

$$\|F(\boldsymbol{z}_{k+1}) - F(\boldsymbol{y}_k)\| \leq \ell_k\|\boldsymbol{z}_{k+1} - \boldsymbol{y}_k\| \leq \frac{\nu}{h_k}\|\boldsymbol{z}_{k+1} - \boldsymbol{y}_k\|.$$

Since $\frac{1}{4} - \frac{\nu^2}{2} = \frac{\delta_\nu}{2}$ with $\delta_\nu = \frac{1}{2} - \nu^2 > 0$, equation 54 yields, on a safe iteration,

$$\begin{aligned}\widetilde{\Phi}_k^{\boldsymbol{a}_k}(\tau_k) - \widetilde{\Phi}_{k+1}^{\boldsymbol{a}_k}(\tau_k + h_k) \geq {}& \frac{\tau_k(\tau_k + h_k)}{h_k}\,\langle F(\boldsymbol{z}_k) - F(\boldsymbol{z}_{k+1}), \boldsymbol{z}_k - \boldsymbol{z}_{k+1}\rangle\\ &+ \frac{\delta_\nu}{2}\frac{(\tau_k + h_k)^2}{h_k^2}\|\boldsymbol{z}_{k+1} - \boldsymbol{y}_k\|^2\\ &+ \frac{\theta^2}{2}\tau_k(\tau_k - S_k)\|F(\boldsymbol{z}_k) - F(\boldsymbol{y}_{k-1})\|^2.\end{aligned} \tag{55}$$

If F is monotone, the first term is nonnegative. Since $\widetilde{\Phi}_k^{\boldsymbol{a}_k}(\tau_k) = 0$ by equation 44, every safe iteration satisfies

$$\widetilde{\Phi}_{k+1}^{\boldsymbol{a}_k}(\tau_k + h_k) \leq 0. \tag{56}$$

C.5 Anchor correction, clock growth, and step sizes

At the new certified clock $S_{k+1} = \tau_k + h_k$, the memory term in $\widetilde{\Phi}^{\boldsymbol{a}}_{k+1}$ vanishes for every anchor $\boldsymbol{a}$. Therefore

$$\widetilde{\Phi}^{\boldsymbol{a}_k}_{k+1}(\tau_k + h_k) = (\tau_k + h_k)\left[\frac{\tau_k + h_k}{2}\|F(\boldsymbol{z}_{k+1})\|^2 + \langle F(\boldsymbol{z}_{k+1}), \boldsymbol{z}_{k+1} - \boldsymbol{a}_k\rangle\right].$$

Let

$$B_k := \frac{\tau_k + h_k}{2}\|F(\boldsymbol{z}_{k+1})\|^2 + \langle F(\boldsymbol{z}_{k+1}), \boldsymbol{z}_{k+1} - \boldsymbol{a}_k\rangle,$$

so that $w_k = [B_k]_+$. If $F(\boldsymbol{z}_{k+1}) \neq 0$, the anchor update in Line 14 of Algorithm 2 gives $\langle F(\boldsymbol{z}_{k+1}), \boldsymbol{z}_{k+1} - \boldsymbol{a}_{k+1}\rangle = \langle F(\boldsymbol{z}_{k+1}), \boldsymbol{z}_{k+1} - \boldsymbol{a}_k\rangle - w_k$, and hence

$$\begin{aligned}\widetilde{\Phi}^{\boldsymbol{a}_{k+1}}_{k+1}(\tau_k + h_k) &= \widetilde{\Phi}^{\boldsymbol{a}_k}_{k+1}(\tau_k + h_k) - (\tau_k + h_k)[B_k]_+ \\ &= (\tau_k + h_k)\,(B_k - [B_k]_+) \\ &= -(\tau_k + h_k)[-B_k]_+ \\ &\leq 0.\end{aligned}$$

If $F(\boldsymbol{z}_{k+1}) = 0$, the method has reached a zero. Hence the invariant equation 45 holds at iteration $k+1$, and equation 46 gives

$$\tau_{k+1} \geq S_{k+1} = \tau_k + h_k. \tag{57}$$

Moreover, on a safe iteration, equation 56 implies $B_k \leq 0$, so $w_k = 0$ and

$$\boldsymbol{a}_{k+1} = \boldsymbol{a}_k \qquad \text{whenever } h_k\ell_k \leq \nu. \tag{58}$$

Step-size bounds and unsafe iterations. The smoothness estimates and the step-size rule in Lines 10–12 of Algorithm 2 are those of Algorithm 1, and the corresponding arguments in Appendix B use only these lines and the L-Lipschitz continuity of F. We recall them for completeness. Since F is L-Lipschitz, $\ell_k \leq L$, so the two smoothness estimates satisfy

$$\widehat{R}_0 \leq \widehat{R}_k \leq \max\{\widehat{R}_0, L\}, \qquad \widehat{R}_k \leq R_k \leq \max\{R_0, L\},$$

where the bound for R_k follows by induction from $R_{k+1} = \max\{\widehat{R}_{k+1}, R_k/\rho\}$. Starting from $h_0 = \nu/(\gamma R_0)$, the step-size update therefore gives

$$\underline{h} \leq h_k \leq \overline{h}, \qquad h_k \leq \frac{\nu}{\gamma R_k} \qquad (k \geq 0). \tag{59}$$

Indeed, both entries in the minimum defining h_{k+1} are at least $\underline{h}$, and its second entry is at most $\overline{h}$ because $R_{k+1} \geq \widehat{R}_0$. If iteration k is unsafe, then

$$\ell_k > \frac{\nu}{h_k} \geq \gamma R_k \geq \gamma\widehat{R}_k, \qquad \widehat{R}_{k+1} \geq \ell_k > \gamma\widehat{R}_k.$$

Thus every unsafe iteration increases the nondecreasing estimate $\widehat{R}_k$ by a factor strictly greater than γ. If $i_1 < \cdots < i_m$ are unsafe indices, induction gives $\ell_{i_s} > \gamma^s\widehat{R}_0$ for $s = 1, \ldots, m$. Since $\ell_{i_s} \leq L$, necessarily $\gamma^s\widehat{R}_0 < L$. Consequently, there are at most J unsafe iterations, with J as in Theorem 3.

C.6 Boundedness of the moving anchor

Let

$$b_k := |\{0 \leq i < k : \; h_i\ell_i > \nu\}|$$

be the number of unsafe iterations before iteration k.

Lemma 2 (Anchor bound for the affine-Popov extension)**.** *For every iteration k before termination,*

$$\|\boldsymbol{a}_k - \boldsymbol{z}^\star\| \leq C^{b_k}_{\mathrm{anc},\theta}D_0, \qquad \|\boldsymbol{z}_k - \boldsymbol{a}_k\| \leq C_{s,\theta}C^{b_k}_{\mathrm{anc},\theta}D_0,$$

where $C_{s,\theta}$ and $C_{\mathrm{anc},\theta}$ are as defined in Theorem 3. Consequently, since $b_k \leq J$, $\|\boldsymbol{a}_k - \boldsymbol{z}^\star\| \leq C^J_{\mathrm{anc},\theta}D_0$.

Proof. We proceed by induction on k. The case $k = 0$ follows from $\boldsymbol{a}_0 = \boldsymbol{z}_0$. Assume the claim holds at iteration k.

We first collect three estimates at iteration k. If $\tau_k = 0$, they hold trivially. Otherwise, by the root identity equation 44 and $\tau_k \geq S_k$, the memory term is nonnegative, so

$$\frac{\tau_k^2}{2}\|F(\boldsymbol{z}_k)\|^2 \leq \tau_k \langle F(\boldsymbol{z}_k), \boldsymbol{a}_k - \boldsymbol{z}_k \rangle.$$

By monotonicity against $\boldsymbol{z}^\star$, $\langle F(\boldsymbol{z}_k), \boldsymbol{z}_k - \boldsymbol{z}^\star \rangle \geq 0$, hence

$$\langle F(\boldsymbol{z}_k), \boldsymbol{a}_k - \boldsymbol{z}_k \rangle \leq \langle F(\boldsymbol{z}_k), \boldsymbol{a}_k - \boldsymbol{z}^\star \rangle \leq \|F(\boldsymbol{z}_k)\| \, \|\boldsymbol{a}_k - \boldsymbol{z}^\star\|. \tag{60}$$

Therefore

$$\tau_k \|F(\boldsymbol{z}_k)\| \leq 2C_{\mathrm{anc},\theta}^{b_k} D_0. \tag{61}$$

Furthermore, equation 44, equation 60, and equation 61 give

$$\begin{aligned}
\theta^2 \tau_k(\tau_k - S_k)\|F(\boldsymbol{z}_k) - F(\boldsymbol{y}_{k-1})\|^2 &= \tau_k \langle F(\boldsymbol{z}_k), \boldsymbol{a}_k - \boldsymbol{z}_k \rangle - \frac{\tau_k^2}{2}\|F(\boldsymbol{z}_k)\|^2 \\
&\leq \tau_k \|F(\boldsymbol{z}_k)\| \, \|\boldsymbol{a}_k - \boldsymbol{z}^\star\| \\
&\leq 2C_{\mathrm{anc},\theta}^{2b_k} D_0^2,
\end{aligned}$$

and so, by equation 47,

$$\tau_k \theta_k \|F(\boldsymbol{z}_k) - F(\boldsymbol{y}_{k-1})\| \leq \sqrt{2}\, C_{\mathrm{anc},\theta}^{b_k} D_0. \tag{62}$$

Since $M_k = F(\boldsymbol{z}_k) - \theta_k\big(F(\boldsymbol{z}_k) - F(\boldsymbol{y}_{k-1})\big)$, combining equation 61 and equation 62 gives

$$\tau_k \|M_k\| \leq \left(2 + \sqrt{2}\right) C_{\mathrm{anc},\theta}^{b_k} D_0. \tag{63}$$

Safe iteration. Suppose k is safe. Then $\boldsymbol{a}_{k+1} = \boldsymbol{a}_k$ by equation 58, and $b_{k+1} = b_k$. Also, $\widetilde{\Phi}_{k+1}^{\boldsymbol{a}_k}(\tau_k + h_k) \leq 0$ by equation 56, and the memory term vanishes at $\tau_k + h_k$. Therefore

$$\frac{(\tau_k + h_k)^2}{2}\|F(\boldsymbol{z}_{k+1})\|^2 + (\tau_k + h_k)\langle F(\boldsymbol{z}_{k+1}), \boldsymbol{z}_{k+1} - \boldsymbol{a}_k \rangle \leq 0.$$

Using monotonicity against $\boldsymbol{z}^\star$ as in equation 60 gives

$$(\tau_k + h_k)\|F(\boldsymbol{z}_{k+1})\| \leq 2C_{\mathrm{anc},\theta}^{b_k} D_0. \tag{64}$$

By equation 50 and $\beta_k = h_k/(\tau_k + h_k)$,

$$\boldsymbol{z}_{k+1} - \boldsymbol{y}_k = \beta_k \tau_k M_k - h_k F(\boldsymbol{y}_k).$$

Since $F(\boldsymbol{y}_k) = F(\boldsymbol{z}_{k+1}) - \big(F(\boldsymbol{z}_{k+1}) - F(\boldsymbol{y}_k)\big)$ and $h_k = \beta_k(\tau_k + h_k)$,

$$\|\boldsymbol{z}_{k+1} - \boldsymbol{y}_k\| \leq \beta_k \tau_k \|M_k\| + \beta_k(\tau_k + h_k)\|F(\boldsymbol{z}_{k+1})\| + h_k \ell_k \|\boldsymbol{z}_{k+1} - \boldsymbol{y}_k\|.$$

Using $h_k \ell_k \leq \nu$, equation 63, and equation 64,

$$\|\boldsymbol{z}_{k+1} - \boldsymbol{y}_k\| \leq \frac{\beta_k(4 + \sqrt{2})}{1 - \nu} C_{\mathrm{anc},\theta}^{b_k} D_0. \tag{65}$$

Next, $\boldsymbol{z}_{k+1} - \boldsymbol{a}_k = (1 - \beta_k)(\boldsymbol{z}_k - \boldsymbol{a}_k) - h_k F(\boldsymbol{y}_k)$ by Line 9 of Algorithm 2. Using $h_k F(\boldsymbol{y}_k) = h_k F(\boldsymbol{z}_{k+1}) - h_k\big(F(\boldsymbol{z}_{k+1}) - F(\boldsymbol{y}_k)\big)$, the induction hypothesis, equation 64, and equation 65,

$$\begin{aligned}
\|\boldsymbol{z}_{k+1} - \boldsymbol{a}_k\| &\leq (1 - \beta_k)\|\boldsymbol{z}_k - \boldsymbol{a}_k\| + \beta_k(\tau_k + h_k)\|F(\boldsymbol{z}_{k+1})\| + h_k \ell_k \|\boldsymbol{z}_{k+1} - \boldsymbol{y}_k\| \\
&\leq (1 - \beta_k) C_{s,\theta} C_{\mathrm{anc},\theta}^{b_k} D_0 + \beta_k \left(2 + \frac{\nu(4 + \sqrt{2})}{1 - \nu}\right) C_{\mathrm{anc},\theta}^{b_k} D_0 \\
&= C_{s,\theta} C_{\mathrm{anc},\theta}^{b_k} D_0.
\end{aligned}$$

Since $\boldsymbol{a}_{k+1} = \boldsymbol{a}_k$ and $b_{k+1} = b_k$, the induction inequalities hold at $k + 1$.

Unsafe iteration. Suppose k is unsafe. Then $b_{k+1} = b_k + 1$. First, by equation 61, $h_k \le \overline{h}$, Lipschitz continuity, and the induction hypothesis,

$$\begin{aligned}(\tau_k + h_k)\|F(\boldsymbol{z}_k)\| &\le \tau_k \|F(\boldsymbol{z}_k)\| + h_k L\left(\|\boldsymbol{z}_k - \boldsymbol{a}_k\| + \|\boldsymbol{a}_k - \boldsymbol{z}^\star\|\right) \\ &\le 2C_{\mathrm{anc},\theta}^{b_k} D_0 + L\overline{h}\left(C_{s,\theta} + 1\right) C_{\mathrm{anc},\theta}^{b_k} D_0. \end{aligned} \tag{66}$$

Also, by the predictor update in Line 8 of Algorithm 2,

$$\boldsymbol{y}_k - \boldsymbol{z}_k = \beta_k(\boldsymbol{a}_k - \boldsymbol{z}_k) - (1-\beta_k)h_k M_k.$$

Multiplying by $\tau_k + h_k$ and using $(\tau_k + h_k)\beta_k = h_k$, $(\tau_k + h_k)(1-\beta_k) = \tau_k$, and equation 63,

$$\begin{aligned}(\tau_k + h_k)\|\boldsymbol{y}_k - \boldsymbol{z}_k\| &\le h_k \|\boldsymbol{z}_k - \boldsymbol{a}_k\| + h_k \tau_k \|M_k\| \\ &\le \overline{h}\left(C_{s,\theta} + 2 + \sqrt{2}\right) C_{\mathrm{anc},\theta}^{b_k} D_0.\end{aligned}$$

Therefore

$$(\tau_k + h_k)\|F(\boldsymbol{y}_k) - F(\boldsymbol{z}_k)\| \le L\overline{h}\left(C_{s,\theta} + 2 + \sqrt{2}\right) C_{\mathrm{anc},\theta}^{b_k} D_0.$$

Combining this with equation 66 gives

$$(\tau_k + h_k)\|F(\boldsymbol{y}_k)\| \le G_\theta C_{\mathrm{anc},\theta}^{b_k} D_0, \tag{67}$$

with $G_\theta = 2 + L\overline{h}(2C_{s,\theta} + 3 + \sqrt{2})$ as defined in Theorem 3. Next, by equation 50, $\boldsymbol{z}_{k+1} - \boldsymbol{y}_k = \beta_k \tau_k M_k - h_k F(\boldsymbol{y}_k)$, and hence, by equation 63 and equation 67,

$$\begin{aligned}(\tau_k + h_k)\|\boldsymbol{z}_{k+1} - \boldsymbol{y}_k\| &\le h_k \tau_k \|M_k\| + h_k(\tau_k + h_k)\|F(\boldsymbol{y}_k)\| \\ &\le \overline{h}\left(2 + \sqrt{2} + G_\theta\right) C_{\mathrm{anc},\theta}^{b_k} D_0.\end{aligned}$$

By Lipschitz continuity,

$$(\tau_k + h_k)\|F(\boldsymbol{z}_{k+1}) - F(\boldsymbol{y}_k)\| \le L\overline{h}\left(2 + \sqrt{2} + G_\theta\right) C_{\mathrm{anc},\theta}^{b_k} D_0.$$

Combining this with equation 67,

$$(\tau_k + h_k)\|F(\boldsymbol{z}_{k+1})\| \le \left[G_\theta + L\overline{h}\left(2 + \sqrt{2} + G_\theta\right)\right] C_{\mathrm{anc},\theta}^{b_k} D_0. \tag{68}$$

Moreover, from Line 9 of Algorithm 2, $h_k \le \tau_k + h_k$, the induction hypothesis, and equation 67,

$$\|\boldsymbol{z}_{k+1} - \boldsymbol{a}_k\| \le (1-\beta_k)\|\boldsymbol{z}_k - \boldsymbol{a}_k\| + h_k\|F(\boldsymbol{y}_k)\| \le \left(C_{s,\theta} + G_\theta\right) C_{\mathrm{anc},\theta}^{b_k} D_0. \tag{69}$$

If $F(\boldsymbol{z}_{k+1}) = 0$, the algorithm terminates and no further anchor bound is needed. Otherwise, since $w_k = [B_k]_+ \le \frac{\tau_k + h_k}{2}\|F(\boldsymbol{z}_{k+1})\|^2 + \|F(\boldsymbol{z}_{k+1})\|\,\|\boldsymbol{z}_{k+1} - \boldsymbol{a}_k\|$, the anchor correction length satisfies, by equation 68 and equation 69,

$$\begin{aligned}\|\boldsymbol{a}_{k+1} - \boldsymbol{a}_k\| &= \frac{w_k}{\|F(\boldsymbol{z}_{k+1})\|} \\ &\le \frac{\tau_k + h_k}{2}\|F(\boldsymbol{z}_{k+1})\| + \|\boldsymbol{z}_{k+1} - \boldsymbol{a}_k\| \\ &\le \left[\frac{G_\theta + L\overline{h}(2 + \sqrt{2} + G_\theta)}{2} + C_{s,\theta} + G_\theta\right] C_{\mathrm{anc},\theta}^{b_k} D_0.\end{aligned}$$

Consequently,

$$\begin{aligned}\|\boldsymbol{a}_{k+1} - \boldsymbol{z}^\star\| &\le \|\boldsymbol{a}_k - \boldsymbol{z}^\star\| + \|\boldsymbol{a}_{k+1} - \boldsymbol{a}_k\| \\ &\le \left[1 + C_{s,\theta} + \frac{3}{2}G_\theta + \frac{L\overline{h}}{2}\left(2 + \sqrt{2} + G_\theta\right)\right] C_{\mathrm{anc},\theta}^{b_k} D_0 \\ &\le C_{\mathrm{anc},\theta}^{b_k+1} D_0,\end{aligned}$$

where the last inequality follows from the definition $C_{\mathrm{anc},\theta} = 1 + 2C_{s,\theta} + L\overline{h}(2+\sqrt{2}) + (3 + L\overline{h})G_\theta$, which dominates the bracket term by term. Finally,

$$\begin{aligned}\|\boldsymbol{z}_{k+1} - \boldsymbol{a}_{k+1}\| &\le \|\boldsymbol{z}_{k+1} - \boldsymbol{a}_k\| + \|\boldsymbol{a}_{k+1} - \boldsymbol{a}_k\| \\ &\le \left[2\left(C_{s,\theta} + G_\theta\right) + \frac{G_\theta + L\overline{h}(2 + \sqrt{2} + G_\theta)}{2}\right] C_{\mathrm{anc},\theta}^{b_k} D_0 \\ &\le C_{s,\theta} C_{\mathrm{anc},\theta}^{b_k+1} D_0,\end{aligned}$$

because the bracket is at most $C_{\mathrm{anc},\theta}$, again term by term, and $C_{s,\theta} \ge 2$. Since $b_{k+1} = b_k + 1$, this closes the induction. □

C.7 Proof of Theorem 3

If $F(\boldsymbol{z}_k) = 0$, the result is immediate. Otherwise, by the root identity equation 44 and $\tau_k \geq S_k$, the memory term is nonnegative, so

$$\frac{\tau_k^2}{2}\|F(\boldsymbol{z}_k)\|^2 + \tau_k\langle F(\boldsymbol{z}_k), \boldsymbol{z}_k - \boldsymbol{a}_k\rangle \leq 0.$$

By equation 57 and equation 59, $\tau_k \geq \sum_{i=0}^{k-1} h_i \geq k\underline{h} > 0$ for $k \geq 1$. Dividing by $\tau_k^2/2$ and using equation 60, we obtain

$$\|F(\boldsymbol{z}_k)\| \leq \frac{2}{\tau_k}\|\boldsymbol{a}_k - \boldsymbol{z}^\star\|. \tag{70}$$

Combining this with $\tau_k \geq k\underline{h}$ and Lemma 2 gives

$$\|F(\boldsymbol{z}_k)\| \leq \frac{2C_{\mathrm{anc},\theta}^J}{k\underline{h}} D_0.$$

Since $\|\nabla\mathcal{L}(\boldsymbol{z}_k)\| = \|F(\boldsymbol{z}_k)\|$, this proves equation 41, and squaring gives the stated $\mathcal{O}(k^{-2})$ rate.

C.8 Proof of Theorem 4

Assume now that the hypotheses of Theorem 4 hold, and write $\Delta_k := F(\boldsymbol{z}_k) - F(\boldsymbol{y}_{k-1})$, so that $F(\boldsymbol{z}_k) = M_k + \theta_k\Delta_k$. On a safe iteration, equation 55, equation 44, and strong monotonicity imply

$$-\widetilde{\Phi}_{k+1}^{\boldsymbol{a}_k}(\tau_k + h_k) \geq \tau_k(\tau_k + h_k)\left[\frac{\mu}{h_k}\|\boldsymbol{z}_k - \boldsymbol{z}_{k+1}\|^2 + \frac{\delta_\nu}{2h_k^2}\|\boldsymbol{z}_{k+1} - \boldsymbol{y}_k\|^2 + \frac{\theta^2}{2}\frac{\tau_k - S_k}{\tau_k + h_k}\|\Delta_k\|^2\right], \tag{71}$$

where we also used $(\tau_k + h_k)^2 \geq \tau_k(\tau_k + h_k)$ in the second term. Define the next quadratic coefficient

$$\mathcal{A}_{k+1} := \|F(\boldsymbol{z}_{k+1})\|^2 + 2\theta^2\|F(\boldsymbol{z}_{k+1}) - F(\boldsymbol{y}_k)\|^2,$$

so that $\widetilde{\Phi}_{k+1}^{\boldsymbol{a}}(t)$ has leading coefficient $\mathcal{A}_{k+1}/2$ for every anchor $\boldsymbol{a}$. We show that $\mathcal{A}_{k+1}$ is controlled by the bracket in equation 71. From equation 49 and equation 51,

$$\boldsymbol{z}_k - \boldsymbol{a}_k + \tau_k M_k = \frac{\tau_k + h_k}{h_k}\left(\boldsymbol{z}_k - \boldsymbol{z}_{k+1} + \boldsymbol{z}_{k+1} - \boldsymbol{y}_k\right). \tag{72}$$

If $\tau_k = 0$, the basic clock growth equation 57 already gives $\tau_{k+1} \geq h_k = q_\theta\tau_k + h_k$. We may therefore assume $\tau_k > 0$ for the following residual estimates. Dividing equation 44 by τ_k^2 and using $\theta_k^2 = \theta^2(\tau_k - S_k)/\tau_k$ gives

$$0 = \frac{1}{2}\|F(\boldsymbol{z}_k)\|^2 + \left\langle F(\boldsymbol{z}_k), \frac{\boldsymbol{z}_k - \boldsymbol{a}_k}{\tau_k}\right\rangle + \theta_k^2\|\Delta_k\|^2.$$

Substituting

$$\frac{\boldsymbol{z}_k - \boldsymbol{a}_k}{\tau_k} = \frac{\tau_k + h_k}{\tau_k h_k}\left(\boldsymbol{z}_k - \boldsymbol{z}_{k+1} + \boldsymbol{z}_{k+1} - \boldsymbol{y}_k\right) - M_k$$

from equation 72, and using $\frac{1}{2}\|F(\boldsymbol{z}_k)\|^2 - \langle F(\boldsymbol{z}_k), M_k\rangle = -\frac{1}{2}\|M_k\|^2 + \frac{1}{2}\theta_k^2\|\Delta_k\|^2$, gives

$$\frac{1}{2}\|M_k\|^2 = \frac{\tau_k + h_k}{\tau_k h_k}\langle F(\boldsymbol{z}_k), \boldsymbol{z}_k - \boldsymbol{z}_{k+1} + \boldsymbol{z}_{k+1} - \boldsymbol{y}_k\rangle + \frac{3}{2}\theta_k^2\|\Delta_k\|^2.$$

Since $\|F(\boldsymbol{z}_k)\| \leq \|M_k\| + \theta_k\|\Delta_k\|$ and $\theta_k = \theta\sqrt{\lambda_k}$, this yields

$$\frac{1}{2}\|M_k\|^2 \leq \left(\|M_k\| + \theta\sqrt{\lambda_k}\|\Delta_k\|\right)\frac{\tau_k + h_k}{\tau_k h_k}\|\boldsymbol{z}_k - \boldsymbol{z}_{k+1} + \boldsymbol{z}_{k+1} - \boldsymbol{y}_k\| + \frac{3\theta^2}{2}\lambda_k\|\Delta_k\|^2. \tag{73}$$

Write $c := \frac{\tau_k + h_k}{\tau_k h_k}\|\boldsymbol{z}_k - \boldsymbol{z}_{k+1} + \boldsymbol{z}_{k+1} - \boldsymbol{y}_k\|$ and $e := \theta\sqrt{\lambda_k}\|\Delta_k\|$. Multiplying equation 73 by 2 and applying Young's inequality in the forms $2\|M_k\|c \leq \frac{1}{2}\|M_k\|^2 + 2c^2$ and $2ec \leq e^2 + c^2$ gives $\frac{1}{2}\|M_k\|^2 \leq 3c^2 + 4e^2$, that is,

$$\|M_k\|^2 \leq 6\left(\frac{\tau_k + h_k}{\tau_k h_k}\right)^2\|\boldsymbol{z}_k - \boldsymbol{z}_{k+1} + \boldsymbol{z}_{k+1} - \boldsymbol{y}_k\|^2 + 8\theta^2\lambda_k\|\Delta_k\|^2. \tag{74}$$

By equation 50,

$$F(\boldsymbol{z}_{k+1}) = \frac{\tau_k}{\tau_k + h_k} M_k - \frac{1}{h_k}(\boldsymbol{z}_{k+1} - \boldsymbol{y}_k) + \big(F(\boldsymbol{z}_{k+1}) - F(\boldsymbol{y}_k)\big),$$

and on a safe iteration

$$\|F(\boldsymbol{z}_{k+1}) - F(\boldsymbol{y}_k)\| \le \ell_k \|\boldsymbol{z}_{k+1} - \boldsymbol{y}_k\| \le \frac{\nu}{h_k} \|\boldsymbol{z}_{k+1} - \boldsymbol{y}_k\|.$$

Taking square roots in equation 74, multiplying by $\tau_k/(\tau_k+h_k)$, and applying the triangle inequality and $(a+b+c)^2 \le 3(a^2+b^2+c^2)$, we obtain

$$\begin{aligned}\mathcal{A}_{k+1} \le \frac{18}{h_k^2}\|\boldsymbol{z}_k - \boldsymbol{z}_{k+1}\|^2 + \frac{3(\sqrt{6}+1+\nu)^2 + 2\theta^2\nu^2}{h_k^2}\|\boldsymbol{z}_{k+1} - \boldsymbol{y}_k\|^2 \\ + 24\theta^2 \left(\frac{\tau_k}{\tau_k + h_k}\right)^2 \lambda_k \|\Delta_k\|^2. \end{aligned} \tag{75}$$

Using $h_k \ge \underline{h}$, $\tau_k/(\tau_k + h_k) \le 1$, and $\lambda_k = (\tau_k - S_k)/\tau_k$, the definition of Θ_θ implies

$$\mathcal{A}_{k+1} \le \Theta_\theta \left[\frac{\mu}{h_k}\|\boldsymbol{z}_k - \boldsymbol{z}_{k+1}\|^2 + \frac{\delta_\nu}{2h_k^2}\|\boldsymbol{z}_{k+1} - \boldsymbol{y}_k\|^2 + \frac{\theta^2}{2}\frac{\tau_k - S_k}{\tau_k + h_k}\|\Delta_k\|^2\right]. \tag{76}$$

Indeed, the three coefficients in equation 75 are at most

$$\frac{18}{\mu\underline{h}} \cdot \frac{\mu}{h_k}, \qquad \frac{2\left[3(\sqrt{6}+1+\nu)^2 + 2\theta^2\nu^2\right]}{\delta_\nu} \cdot \frac{\delta_\nu}{2h_k^2}, \qquad 48 \cdot \frac{\theta^2}{2}\frac{\tau_k - S_k}{\tau_k + h_k},$$

respectively, where the last one uses $24\theta^2\big(\frac{\tau_k}{\tau_k+h_k}\big)^2 \frac{\tau_k - S_k}{\tau_k} \le 24\theta^2 \frac{\tau_k - S_k}{\tau_k + h_k}$; each factor in front is at most Θ_θ. Combining equation 71 and equation 76, every safe iteration satisfies

$$-\widetilde{\Phi}_{k+1}^{\boldsymbol{a}_k}(\tau_k + h_k) \ge \frac{\tau_k(\tau_k + h_k)}{\Theta_\theta}\mathcal{A}_{k+1}. \tag{77}$$

On a safe iteration, $\boldsymbol{a}_{k+1} = \boldsymbol{a}_k$ and $S_{k+1} = \tau_k + h_k$. Since $\tau_{k+1} \ge S_{k+1} > 0$ is the positive root of $\widetilde{\Phi}_{k+1}^{\boldsymbol{a}_{k+1}}$, we have

$$\widetilde{\Phi}_{k+1}^{\boldsymbol{a}_{k+1}}(t) = \frac{\mathcal{A}_{k+1}}{2} t\,(t - \tau_{k+1}).$$

Evaluating at $t = \tau_k + h_k$ and using equation 77,

$$\frac{\mathcal{A}_{k+1}}{2}(\tau_k + h_k)(\tau_{k+1} - \tau_k - h_k) \ge \frac{\tau_k(\tau_k + h_k)}{\Theta_\theta}\mathcal{A}_{k+1}.$$

If $\mathcal{A}_{k+1} = 0$, then $F(\boldsymbol{z}_{k+1}) = 0$ and the method has solved the problem. Otherwise, cancelling $\mathcal{A}_{k+1}(\tau_k + h_k)/2$ gives the safe-iteration geometric clock growth

$$\tau_{k+1} \ge \left(1 + \frac{2}{\Theta_\theta}\right)\tau_k + h_k = q_\theta \tau_k + h_k. \tag{78}$$

Unsafe iterations still satisfy the weaker clock growth equation 57.

From safe-iteration growth to the rate. Among the first k iterations, let s_k be the number of safe iterations, so that $b_k = k - s_k$ is the number of unsafe iterations. Since $b_k \le J$,

$$s_k \ge m_k = \max\{k - J, 0\}.$$

By equation 59, safe iterations satisfy $\tau_{i+1} \ge T_s(\tau_i)$ and unsafe iterations satisfy $\tau_{i+1} \ge T_u(\tau_i)$, where $T_s(t) := q_\theta t + \underline{h}$ and $T_u(t) := t + \underline{h}$ are increasing maps. For fixed s_k and b_k, the smallest lower bound is obtained when all safe iterations come first, since $T_s(T_u(t)) - T_u(T_s(t)) = (q_\theta - 1)\underline{h} \ge 0$. Thus, since $\tau_0 \ge 0$,

$$\tau_k \ge \underline{h}\left[\sum_{i=0}^{s_k - 1} q_\theta^i + k - s_k\right] = \underline{h}\left[\frac{q_\theta^{s_k} - 1}{q_\theta - 1} + k - s_k\right].$$

The function

$$\psi(s) := \frac{q_\theta^s - 1}{q_\theta - 1} + k - s$$

is nondecreasing on the nonnegative integers, because $\psi(s+1) - \psi(s) = q_\theta^s - 1 \geq 0$. Since $s_k \geq m_k$ and $q_\theta - 1 = 2/\Theta_\theta$,

$$\tau_k \geq \underline{h} \left[k - m_k + \frac{\Theta_\theta}{2} \left(q_\theta^{m_k} - 1 \right) \right]. \tag{79}$$

Finally, equation 70 and Lemma 2 give

$$\|F(\boldsymbol{z}_k)\| \leq \frac{2 C_{\mathrm{anc},\theta}^J D_0}{\tau_k}.$$

Combining this with equation 79 and $\|\nabla \mathcal{L}(\boldsymbol{z}_k)\| = \|F(\boldsymbol{z}_k)\|$ proves the first bound in equation 42.

For the second bound, if $m_k \geq 1$, then

$$q_\theta^{m_k} - 1 = (q_\theta - 1)\left(1 + q_\theta + \cdots + q_\theta^{m_k - 1}\right) \geq (q_\theta - 1) q_\theta^{m_k - 1},$$

so $\frac{\Theta_\theta}{2}(q_\theta^{m_k} - 1) \geq q_\theta^{m_k - 1}$, and $k - m_k \geq 0$. If $m_k = 0$, then the denominator in equation 42 is $k\underline{h} \geq \underline{h} = \underline{h}\, q_\theta^0$. This proves the second bound. For $k > J$, $m_k = k - J$, which gives the geometric rate $\mathcal{O}(q_\theta^{-k})$.