EXACT PARAMETER REGIONS FOR GENERALIZED OPTIMISTIC GRADIENT METHODS

Anonymous authors
Paper under double-blind review

ABSTRACT

The generalized optimistic gradient method (GOGD) is a one-call method for operator equations that includes the optimistic gradient method as the special case of equal step and correction parameters, yet its full admissible parameter region has remained unknown. For L-Lipschitz operators satisfying a weak-Minty condition with parameter ρ and $-1 \leq \rho L < 1$, we characterize exactly the parameter pairs that guarantee bounded iterates with square-summable increments and residuals, and show by a two-dimensional linear counterexample that no other pair does. In particular, equal step and correction parameters are admissible only for small ρL. For monotone operators, the region is exactly the set of parameters yielding weak convergence to a zero from every initialization; for weak-Minty operators, weak convergence follows under standard demiclosedness and solution-set assumptions, with norm convergence in finite dimensions. The proof combines a Lyapunov identity with a matching spectral obstruction and yields best-iterate $O(1/N)$ bounds for squared residuals and increments. Under the stronger pairwise condition of co-hypomonotonicity, we further obtain last-iterate $o(1/N)$ rates on a subregion, and on the full region for affine operators. The new region substantially enlarges previously known sufficient regions, especially as $\rho L \to 1$: on a quadratic minimax example with $\rho L = 0.9$, a parameter rule from the new region uses about 17 times fewer operator evaluations than tuned candidates from earlier regions.

1 INTRODUCTION

Optimistic-gradient methods solve operator equations arising in smooth saddle-point problems by correcting a gradient step with the difference between successive operator values. They require one new evaluation per iteration, rather than the two used by extragradient methods. Their convergence depends on the balance between the step and the correction, especially beyond monotone problems, where the operator can point away from a solution.

We determine which parameter choices guarantee convergence for the generalized optimistic gradient method (GOGD) (Zhang & Yu, 2020; Anagnostides & Panageas, 2022). Given a mapping $F\colon \mathcal{H} \to \mathcal{H}$ on a real Hilbert space, the method seeks a zero of F through the recurrence

$$x_{k+1} = x_k - (\alpha+\beta)F(x_k) + \beta F(x_{k-1}), \qquad k \geq 1, \tag{1}$$

from initial points $x_0, x_1 \in \mathcal{H}$. The positive parameters α and β control the gradient step and the correction, respectively. Setting $\alpha = \beta$ gives the usual optimistic-gradient (OG) update (Popov, 1980); allowing them to differ gives the two-parameter variant studied as OGDA+ (Böhm, 2023). The root-finding formulation also covers smooth optimization and unconstrained variational inequalities.

We assume that F is L-Lipschitz, with $L > 0$, and satisfies the weak-Minty condition, meaning there is a zero $x_\star$ and a constant $\rho \in \mathbb{R}$ such that

$$\langle F(x), x - x_\star \rangle \geq -\rho \|F(x)\|^2 \qquad (x \in \mathcal{H}).$$

Unlike monotonicity, this condition compares each point with one distinguished zero, rather than comparing every pair of points. For $\rho > 0$, it allows $\langle F(x), x - x_\star \rangle$ to be negative, but by at most $\rho\|F(x)\|^2$. For $\rho > 0$, this is also called ρ-star-negative comonotonicity (Gorbunov et al., 2023). Conventions for ρ differ: Diakonikolas et al. (2021) write $-(\rho/2)\|F(x)\|^2$, and Pethick et al. (2022) use the opposite sign. We convert all cited results to our convention. Existing analyses guarantee

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.

convergence of GOGD on several parameter regions, and establish that suitable pairs exist whenever $0 \le \rho L < 1$ (Böhm, 2023). For $-1 \le \rho L < 1$, we characterize all pairs for which every trajectory, for every such operator and every initialization, is bounded with square-summable increments and residuals. Outside this region, a two-dimensional linear example produces residuals that do not vanish. If, in addition, F is demiclosed at zero and either its zero is unique or the same weak-Minty inequality holds at every zero, the iterates converge weakly to a zero, and in norm in finite dimensions. Monotone Lipschitz operators satisfy these conditions, as do co-hypomonotone ones with $\rho L < 1$; for monotone operators, the region therefore characterizes iterate convergence.

The proof combines two arguments that give the same boundary: a Lyapunov identity proves descent over the full weak-Minty class, and a spectral calculation identifies when the iteration becomes unstable on a planar linear operator. The descent argument also gives best-iterate $O(1/N)$ bounds for squared residuals and increments. Under co-hypomonotonicity, pairing the same potential with a nonincreasing residual function gives last-iterate $o(1/N)$ rates on a subregion, and on the full region for affine operators.

Beyond the existence of admissible pairs, the exact region has two consequences. First, at the same correction parameter β, it permits steps αL of order $1 - \rho L$ as $\rho L \to 1$, compared with order $(1 - \rho L)^2$ in Böhm (2023, Theorem A.1). Second, equal parameters are admissible if and only if $\rho L < \sqrt{6}/16$; beyond this threshold, the correction parameter must differ from the step. Section 7 illustrates the effect of this larger region on operator-query counts.

2 Related work

Extragradient (Korpelevich, 1976) and mirror-prox (Nemirovski, 2004) evaluate the operator twice per iteration. Popov's method (Popov, 1980), reflected gradient (Malitsky, 2015), and forward-reflected-backward splitting (Malitsky & Tam, 2020) instead reuse earlier information. In the unconstrained constant-step setting, forward-reflected-backward, past-extragradient, and optimistic-gradient updates reduce to the same recursion (Alacaoglu et al., 2023, Section 2.2). Analyses of standard OG ($\alpha = \beta$) use proximal-point approximations and computer-assisted potentials (Mokhtari et al., 2020a; Gorbunov et al., 2022); for monotone problems, constant-step OG has a tight $O(1/\sqrt{k})$ last-iterate rate (Golowich et al., 2020a; Cai et al., 2022). For the argument-reflected Popov update, Nguyen et al. (2026) prove convergence for $\alpha L < 1/\sqrt{3}$ under boundedness and give a rotational counterexample at equality.

Unequal update and optimism parameters have been studied in smooth convex-concave games (Ryu et al., 2019, Theorem 2), bilinear saddle problems (Mokhtari et al., 2020b), and restricted parameter families under stronger assumptions (Anagnostides & Panageas, 2022). Under the weak-Minty condition, the OGDA+ analysis of Böhm (2023), the bounds of Gorbunov et al. (2023), and the one-block specialization of Tran-Dinh & Luo (2026) provide sufficient regions for the same recurrence. We compare these regions in detail after stating our characterization.

For weak-Minty problems in finite-dimensional Euclidean spaces, Diakonikolas et al. (2021, Assumption 1 and Theorem 3.2) analyze EG+, which uses two new operator evaluations per iteration and gives an $O(1/N)$ average squared-residual bound at its extrapolated points. Adaptive GRAAL and constrained extragradient methods also address locally Lipschitz weak-Minty problems (Alacaoglu et al., 2023; Pethick et al., 2022).

Stronger convergence rates are available under pairwise assumptions. The comonotone framework of Bauschke et al. (2021) includes ρ-co-hypomonotonicity, the pairwise condition (6) in Section 3. It implies the weak-Minty condition, but also constrains the operator away from its zeros. Tran-Dinh & Nguyen-Trung (2026) distinguish best-iterate rates under the weak-Minty condition from last-iterate rates under co-hypomonotonicity. In finite dimensions, under Lipschitz continuity, pairwise co-hypomonotonicity, and suitable parameter restrictions, Tran-Dinh & Nguyen-Trung (2025, Theorems 4.1 and 6.1) establish $O(1/k)$ last-iterate residual-norm bounds, or $O(1/k^2)$ squared-residual bounds, for anchored and Nesterov-type extragradient schemes, including past-gradient variants, with time-varying coefficients. For monotone Lipschitz equations in Hilbert spaces, Upadhyaya et al. (2026b, Corollary 2.5) prove a last-iterate $o(1/k)$ squared-residual rate for extragradient with step size in $(0, 1/L)$. For equal-parameter OG under co-hypomonotonicity, last-iterate $O(1/k)$ squared-residual bounds hold for small ρL (Gorbunov et al., 2023, Theorem 4.4, second part). Our main result

assumes only the weak-Minty condition and gives best-iterate rates; under co-hypomonotonicity, we also obtain last-iterate rates on a subregion that is nonempty for every $\rho L < 1$ (Theorem 3.5). We consider only the unconstrained deterministic recurrence; nonlinear projection and momentum corrections (Giselsson, 2021; Morin et al., 2023), as well as stochastic, higher-order, and anchored variants (Hsieh et al., 2020; Jiang & Mokhtari, 2025; Sohrabi et al., 2026), are outside our scope.

3 Main results

Throughout, $\mathcal{H}$ is a real Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\|$. We make the following standing assumption.

Assumption 3.1. The mapping $F\colon \mathcal{H} \to \mathcal{H}$ is L-Lipschitz for some $L > 0$, its zero set $\operatorname{zer} F := \{x \in \mathcal{H} \mid F(x) = 0\}$ is nonempty, and there exist $x_\star \in \operatorname{zer} F$ and $\rho \in \mathbb{R}$ such that

$$\langle F(x), x - x_\star \rangle \geq -\rho \|F(x)\|^2 \qquad (x \in \mathcal{H}). \tag{2}$$

We use (1) with arbitrary $x_0, x_1 \in \mathcal{H}$ and fixed positive parameters α, β. Negative ρ imposes a one-point cocoercivity condition; positive ρ allows nonmonotone operators. For $\rho L < 1$ and $\max\{0, \rho\} < \beta < 1/L$, define

$$\Psi_{\rho,L}(\beta) := \frac{2(\beta - \rho)(1 - \beta^2 L^2)}{1 + \beta^2 L^2 - 2\beta\rho L^2}. \tag{3}$$

The associated parameter region is

$$\mathcal{R}_{\rho,L} := \left\{ (\alpha, \beta) \;\middle|\; \max\{0, \rho\} < \beta < \frac{1}{L}, \quad 0 < \alpha < \Psi_{\rho,L}(\beta) \right\}. \tag{4}$$

Theorem 3.2. *Fix $L > 0$, $\rho \in \mathbb{R}$, and $\alpha, \beta > 0$. If $-1 \leq \rho L < 1$, then the following statements are equivalent:*

(i) $(\alpha, \beta) \in \mathcal{R}_{\rho,L}$;

(ii) *for every real Hilbert space, every mapping satisfying Assumption 3.1 with constants L and ρ, and every initial pair, the trajectory generated by* (1) *is bounded, and its increments and operator residuals are square summable.*

In particular, $x_{k+1} - x_k \to 0$ and $F(x_k) \to 0$. If $\rho L < -1$, membership in $\mathcal{R}_{\rho,L}$ still implies statement (ii)*. If $\rho L \geq 1$, no positive parameter pair satisfies statement* (ii)*.*

We prove sufficiency in Section 5 and necessity in Section 6; for $\rho L < -1$, whether the converse holds is open, and Remark A.1 explains why the necessity argument does not extend to that range. The proof also shows that, for $-1 \leq \rho L < 1$, statement (i) is equivalent to $F(x_k) \to 0$ for every such mapping and every initial pair. For monotone operators, the region also characterizes weak iterate convergence.

Corollary 3.3. *Fix $L > 0$ and $\alpha, \beta > 0$. For every real Hilbert space, every monotone L-Lipschitz mapping with a nonempty zero set, and every initial pair, the trajectory generated by* (1) *converges weakly to a zero if and only if*

$$0 < \beta < \frac{1}{L}, \qquad 0 < \alpha < \frac{2\beta(1 - \beta^2 L^2)}{1 + \beta^2 L^2}. \tag{5}$$

Within this region, increments and residuals are also square summable. In finite dimensions, the convergence is in norm.

For bilinear games, (5) is the exact linear-convergence region of Zhang & Yu (2020, Theorem 3.3); Corollary 3.3 extends it to all monotone Lipschitz operators.

For weak-Minty operators, iterate convergence requires additional assumptions. A mapping F is *demiclosed at zero* if $x_n \rightharpoonup x$ and $F(x_n) \to 0$ imply $F(x) = 0$ (Bauschke & Combettes, 2017, Definition 4.26). In finite dimensions, continuity ensures this property. In Hilbert spaces, monotone

Lipschitz operators are maximally monotone and hence demiclosed at zero (Bauschke & Combettes, 2017, Corollary 20.28 and Proposition 20.38(ii)).

Proposition 3.4. *Suppose Assumption 3.1 holds and* $(\alpha, \beta) \in \mathcal{R}_{\rho,L}$.

(a) *If* $\mathcal{H}$ *is finite dimensional and* $\operatorname{zer} F = \{x_\star\}$, *then* $x_k \to x_\star$ *in norm.*

(b) *If* F *is demiclosed at zero and* $\operatorname{zer} F = \{x_\star\}$, *then* $x_k \rightharpoonup x_\star$.

(c) *If* F *is demiclosed at zero and every* $z \in \operatorname{zer} F$ *satisfies* (2) *with* $x_\star$ *replaced by* z *and the same* ρ, *then* (x_k) *converges weakly to a zero. The convergence is in norm in finite dimensions.*

For $-1 \le \rho L < 1$, membership in $\mathcal{R}_{\rho,L}$ is also necessary for the conclusions of parts (b) and (c) to hold over the respective classes: the planar counterexamples in Section 6 have unique zeros and satisfy both sets of assumptions. Proofs of Corollary 3.3 and Proposition 3.4 are given in Appendix A.1.

The assumptions of Proposition 3.4(c) hold under the stronger pairwise condition

$$\langle F(x) - F(y), x - y \rangle \ge -\rho \|F(x) - F(y)\|^2 \qquad (x, y \in \mathcal{H}). \tag{6}$$

For $\rho \ge 0$, this is ρ-co-hypomonotonicity (also called ρ-negative comonotonicity); for $\rho < 0$, it is $(-\rho)$-cocoercivity. Taking $y = x_\star$ shows that (6) implies (2) at every zero with the same ρ, and the two conditions coincide for affine operators with a zero. For brevity, we refer to (6) as co-hypomonotonicity for either sign of ρ. If F is L-Lipschitz and $\rho L < 1$, then $(I + \beta F)^{-1}$ has full domain for $\max\{0, \rho\} < \beta < 1/L$. By Minty's theorem, $F^{-1} + \rho \operatorname{Id}$ is then maximally monotone, that is, F is maximally co-hypomonotone (Bauschke & Combettes, 2017, Proposition 23.14). As in the monotone case, it follows that F is demiclosed at zero.

Co-hypomonotonicity also yields last-iterate rates. Our nonlinear guarantee, however, does not cover all of $\mathcal{R}_{\rho,L}$: it requires $\alpha \le \Phi_{\rho,L}(\beta) \coloneqq 2(\beta - \rho)(1 - \beta L)/(1 + \beta L)$, which defines a strict subregion when $\rho L > -1$.

Theorem 3.5. *Suppose* F *is* L*-Lipschitz,* $\operatorname{zer} F \neq \emptyset$, *and* (6) *holds with* $\rho L < 1$. *Let* $(\alpha, \beta) \in \mathcal{R}_{\rho,L}$.

(a) *The sequence generated by* (1) *converges weakly to a zero of* F, *and in norm in finite dimensions.*

(b) *If* $\alpha \le \Phi_{\rho,L}(\beta)$, *then* $\|F(x_N)\|^2 = o(1/N)$ *and* $\|x_{N+1} - x_N\|^2 = o(1/N)$.

(c) *If* F *is affine, the conclusion of* (b) *holds without the restriction* $\alpha \le \Phi_{\rho,L}(\beta)$.

Appendix A.5 gives the proofs and explicit $O(1/N)$ bounds, and shows for $\rho \ge 0$ that the order $1/N$ cannot be improved uniformly. The subregion in part (b) is nonempty for every $\rho L < 1$; Section 4 compares it with earlier regions. Since the rotations in Section 6 are linear, part (c) shows that, over affine operators, $\mathcal{R}_{\rho,L}$ is necessary and sufficient for last-iterate residual convergence when $-1 \le \rho L < 1$. Whether this extends to nonlinear operators remains open.

4 Comparison and parameter choices

The region $\mathcal{R}_{\rho,L}$ contains the earlier sufficient regions for GOGD. For finite-dimensional weak-Minty problems, Böhm (2023, Theorem 3.1) analyzes (1) under the name OGDA+. In our notation, the sufficient region is

$$2\rho < \beta < \frac{1}{L}, \qquad 0 < \alpha \le \frac{\beta(1 - \beta L)}{1 + \beta L}. \tag{7}$$

Its upper boundary is the constant-step GRAAL family (Malitsky, 2020), for which Alacaoglu et al. (2023, Theorem 7) give a best-iterate residual bound. Böhm's Theorem A.1 provides a second region: eliminating its auxiliary parameter and converting the weak-Minty convention gives, for $\rho \ge 0$,

$$\rho < \beta < \frac{1}{L}, \qquad 0 < \alpha \le \beta, \qquad \alpha < \frac{2(\beta - \rho)(1 - \beta L)}{1 + \beta L}. \tag{8}$$

This region is nonempty for every $0 \le \rho L < 1$, but its upper bound on α is smaller than $\Psi_{\rho,L}(\beta)$.

Gorbunov et al. (2023, Theorem 4.4, first part) give another sufficient region for the same recurrence. In our notation, their residual bound applies when

$$0 \le \rho L < \frac{1}{2}, \qquad 2\rho < \beta < \frac{1}{L}, \qquad 0 < \alpha \le \beta - 2\rho, \qquad \alpha < \frac{1}{L} - \beta. \tag{9}$$

Their result is finite dimensional, uses an initialization generated by an auxiliary recursion, and bounds the average squared residual. Regions (7) and (9) are not nested, but both lie strictly inside $\mathcal{R}_{\rho,L}$.

Tran-Dinh & Luo (2026) study randomized block-coordinate optimistic updates under the weak-Minty condition. For $\rho \ge 0$, the one-block specialization of their Section 3.2 descent lemma gives

$$\rho < \beta < \frac{1}{\sqrt{2}L}, \qquad 0 < \alpha < \frac{(\beta-\rho)(1-2\beta^2L^2)}{1-\beta^2L^2}. \tag{10}$$

This is larger than the region obtained from their stated step-size restriction, but retains the threshold $\rho L < 1/\sqrt{2}$ and remains strictly inside our region. The same lemma gives an $O(1/N)$ best-iterate squared-residual rate on (10). Appendix B gives all parameter conversions and strict-containment arguments; Figure 1 compares the boundaries.

Membership in $\mathcal{R}_{\rho,L}$ can be expressed as a single inequality. Define the descent coefficient

$$c := \frac{2(\beta-\rho)(1-\beta^2L^2) - \alpha(1+\beta^2L^2-2\beta\rho L^2)}{\alpha}. \tag{11}$$

Its positivity controls both Lyapunov descent and spectral stability.

Proposition 4.1. *Let $L > 0$, $\rho \in \mathbb{R}$, and $\alpha, \beta > 0$, with $\rho L < 1$. Then $(\alpha, \beta) \in \mathcal{R}_{\rho,L}$ if and only if $c > 0$. For equal parameters $\alpha = \beta$, this condition becomes*

$$0 < \alpha L < 1, \qquad \alpha L(1-3\alpha^2L^2) - 2\rho L(1-2\alpha^2L^2) > 0. \tag{12}$$

The proof is given in Appendix A.2.

For $\rho \ge 0$, the equal-parameter condition requires $\alpha L < 1/\sqrt{3}$, which recovers the threshold of Nguyen et al. (2026) when $\rho = 0$, and is feasible if and only if $\rho L < \sqrt{6}/16$. As $\rho L \uparrow \sqrt{6}/16$, the admissible interval shrinks to the point $\alpha L = 1/\sqrt{6}$. By contrast, unequal parameters remain admissible up to $\rho L < 1$, as shown in the left panel of Figure 1. A useful special case is $\alpha = \beta = 1/(2L)$, which is admissible if and only if $\rho L < 1/8$.

The difference from the earlier bounds is particularly pronounced as ρL approaches one. To compare with (8), choose $\beta L = 1 - q(1-\rho L)$, with fixed $0 < q < 1$. As $\rho L \uparrow 1$, the two upper bounds satisfy

$$\begin{aligned} L\Psi_{\rho,L}(\beta) &= 2q(1-q)(1-\rho L) + O((1-\rho L)^2), \\ \frac{2(\beta L - \rho L)(1-\beta L)}{1+\beta L} &= q(1-q)(1-\rho L)^2 + O((1-\rho L)^3). \end{aligned} \tag{13}$$

The ratio of the upper bounds is therefore asymptotic to $2/(1-\rho L)$. Taking $q = 1/4$ gives a simple parameter rule with this scaling:

$$\alpha L = \frac{3(1-\rho L)}{8}, \qquad \beta L = 1 - \frac{1-\rho L}{4}. \tag{14}$$

It lies in $\mathcal{R}_{\rho,L}$ for $0 < \rho L < 1$ (Appendix A.2) and matches the leading term of $L\Psi_{\rho,L}(\beta)$. For this choice, $c = 3(1-\rho L)^2/16$.

The enlarged region also improves the certified constant, although not at this rule. When $x_0 = x_1$ and $\rho \ge 0$, the right-hand side of (24) is at most an absolute constant times $\|x_1 - x_\star\|^2/(\alpha^2 c)$. For fixed β, $\alpha^2 c = \alpha\big[2(\beta-\rho)(1-\beta^2L^2) - \alpha(1+\beta^2L^2-2\beta\rho L^2)\big]$ is maximized at $\alpha = \Psi_{\rho,L}(\beta)/2$. As $\rho L \uparrow 1$, its maximum over $\mathcal{R}_{\rho,L}$ is asymptotic to $(1-\rho L)^3/(8L^2)$, compared with $(1-\rho L)^4/(4L^2)$ over (8), a gain of order $1/(1-\rho L)$. The rule (14) instead lies near the boundary, where $L^2\alpha^2 c = 27(1-\rho L)^4/1024$: it trades the certified constant for a larger step. We use it as an untuned choice in the experiments.

The last-iterate subregion of Theorem 3.5(b) behaves differently. For $\rho \ge 0$, it contains Böhm's region (8) without the restriction $\alpha \le \beta$, so by (13) it permits only steps of order $(1-\rho L)^2$ as $\rho L \to 1$. The equal-parameter last-iterate bound of Gorbunov et al. (2023, Theorem 4.4, second part) requires $\rho L \le 5/62$ and $4\rho \le \alpha = \beta \le 10/(31L)$; this range is not contained in the subregion, which is, however, nonempty for every $\rho L < 1$.

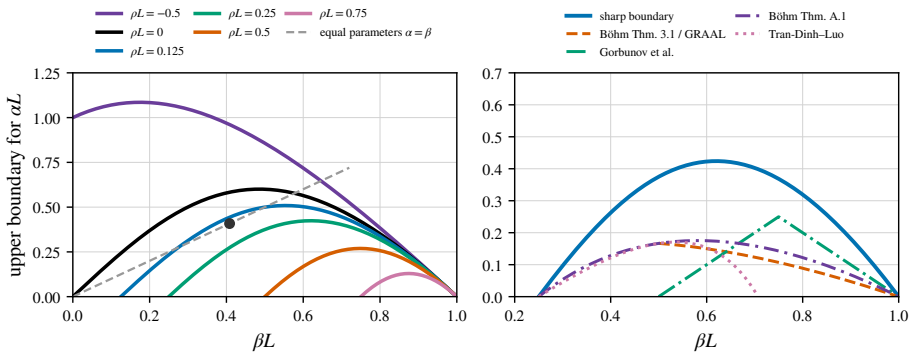


Figure 1: Left: sharp upper boundaries for αL as functions of βL, for representative values of ρL; the admissible set lies below each curve for $\max\{0, \rho L\} < \beta L < 1$. The dashed line denotes equal parameters. The dot marks $\alpha L = \beta L = 1/\sqrt{6}$, where the equal-parameter region collapses as $\rho L \uparrow \sqrt{6}/16$. Right: at $\rho L = 0.25$, the sharp boundary is compared with the sufficient-region boundaries from Böhm's Theorem 3.1 (7) (the constant-step GRAAL curve), Böhm's Theorem A.1 (8), Gorbunov et al. (9), and the one-block specialization of Tran-Dinh–Luo's descent lemma (10). At this ρL, the Theorem A.1 curve coincides with the boundary $\alpha = \Phi_{\rho,L}(\beta)$ of the last-iterate subregion in Theorem 3.5(b).

5 Lyapunov analysis and stationarity

We prove sufficiency by combining two properties of a single potential: its descent controls the accumulated increments, and its coercivity bounds the iterates. The recursion then transfers the increment bound to the operator residuals. Fix the zero $x_\star$ in Assumption 3.1, and write $F_k := F(x_k)$ for $k \geq 0$ and $d_k := x_k - x_{k-1}$ for $k \geq 1$. The recursion gives

$$d_{k+1} = -(\alpha + \beta)F_k + \beta F_{k-1} = -\alpha F_k - \beta(F_k - F_{k-1}). \tag{15}$$

Set

$$B := \frac{\beta}{\alpha}\big[\alpha(\beta - 2\rho) + 2\beta(\beta - \rho)\big], \tag{16}$$

so that $B = \beta^2(\alpha + 2\beta)/\alpha$ when $\rho = 0$. For $(\alpha, \beta) \in \mathcal{R}_{\rho,L}$, both c and B are positive, as shown in Proposition 4.1 and the proof of Lemma 5.2. Consider the potential

$$\begin{aligned} V_k := \|x_k - x_\star\|^2 + BL^2\|d_k\|^2 - 2\alpha\langle x_k - x_\star, F_k\rangle + 2\beta\langle x_k - x_\star, F_{k-1}\rangle \\ - 2\alpha\rho\|F_k\|^2 + \beta(2\rho - \beta)\|F_{k-1}\|^2. \end{aligned} \tag{17}$$

The certificate below uses only Lipschitz continuity between consecutive iterates and relative to $x_\star$, together with (2). The potential was found with the help of AutoLyap (Upadhyaya et al., 2025; 2026a).

5.1 One-step identity

Lemma 5.1. *For every $k \geq 1$ and all positive α, β, the potential in* (17) *satisfies*

$$\begin{aligned} V_k - V_{k+1} = c\|d_{k+1}\|^2 + B\big(L^2\|d_k\|^2 - \|F_k - F_{k-1}\|^2\big) \\ + 2\alpha\big(\langle F_{k+1}, x_{k+1} - x_\star\rangle + \rho\|F_{k+1}\|^2\big). \end{aligned} \tag{18}$$

Consequently, under Assumption 3.1 and $(\alpha, \beta) \in \mathcal{R}_{\rho,L}$,

$$V_{k+1} \leq V_k - c\|d_{k+1}\|^2. \tag{19}$$

Proof. Expanding the potential and substituting (15) gives (18); Appendix A.3 provides the algebra. The second and third terms of (18) are nonnegative by Lipschitz continuity and (2). Together with $B > 0$ by the proof of Lemma 5.2 and $c > 0$ by Proposition 4.1, this yields (19). □

5.2 Positivity and coercivity

The potential is not manifestly nonnegative because the coefficients of $\|F_k\|^2$ and $\|F_{k-1}\|^2$ can have either sign. The following decomposition shows that it is coercive.

Lemma 5.2. *For every $k \geq 1$ and $\beta \neq \rho$, the potential in* (17) *admits the representation*

$$\begin{aligned} V_k = {} & B\big(L^2\|d_k\|^2 - \|F_k - F_{k-1}\|^2\big) \\ & + \frac{\beta(2\beta - 2\rho - \alpha)}{\beta - \rho}\big(\langle x_k - x_\star, F_k\rangle + \rho\|F_k\|^2\big) \\ & + \frac{\alpha B}{2(\beta - \rho)}\big(L^2\|x_k - x_\star\|^2 - \|F_k\|^2\big) \\ & + \frac{\beta^2}{2\alpha(\beta - \rho)}\left\|\frac{\alpha B}{\beta^2}F_k - 2(\beta - \rho)F_{k-1} - \frac{\alpha}{\beta}(x_k - x_\star)\right\|^2 \\ & + \frac{\alpha c}{2(\beta - \rho)}\|x_k - x_\star\|^2. \end{aligned} \tag{20}$$

If $(\alpha, \beta) \in \mathcal{R}_{\rho,L}$, every term on the right is nonnegative and

$$V_k \geq \frac{\alpha c}{2(\beta - \rho)}\|x_k - x_\star\|^2 \geq 0. \tag{21}$$

Proof. See Appendix A.3. □

5.3 Summability and best-iterate bounds

Lemmas 5.1 and 5.2 yield the following bounds, which prove statement (ii) of Theorem 3.2.

Theorem 5.3. *Suppose Assumption 3.1 holds and $(\alpha, \beta) \in \mathcal{R}_{\rho,L}$. For every initial pair and every $N \geq 1$, the sequence generated by* (1) *satisfies*

$$\sum_{k=1}^{N}\|x_{k+1} - x_k\|^2 \leq \frac{V_1}{c}, \tag{22}$$

$$\min_{1\leq k\leq N}\|x_{k+1} - x_k\|^2 \leq \frac{V_1}{cN}, \tag{23}$$

$$\sum_{k=1}^{N}\|F(x_k)\|^2 \leq \frac{\beta}{\alpha}\|F(x_0)\|^2 + \frac{V_1}{\alpha^2 c}, \tag{24}$$

$$\min_{1\leq k\leq N}\|F(x_k)\|^2 \leq \frac{1}{N}\left[\frac{\beta}{\alpha}\|F(x_0)\|^2 + \frac{V_1}{\alpha^2 c}\right]. \tag{25}$$

Both minima are $o(1/N)$. Moreover, the trajectory is bounded, and

$$x_{k+1} - x_k \to 0, \qquad F(x_k) \to 0. \tag{26}$$

Proof. Telescoping (19) bounds the increments. Expanding $\|d_{k+1}\|^2$ with (15) turns this into a bound on the residuals, and coercivity (21) gives boundedness. The details are given in Appendix A.3. □

6 Sharpness via a planar rotation

We now show that the sufficient region cannot be enlarged when $-1 \leq \rho L < 1$. Following the spectral approach of Zhang et al. (2022, Section 5.1), we use a planar operator that satisfies the weak-Minty condition with equality. For this operator, Schur stability reduces to $c > 0$. By rescaling, we may assume $L = 1$. For $-1 \leq \rho < 1$, set

$$F_\rho(x) \coloneqq \left(-\rho I + \sqrt{1 - \rho^2}J\right)x, \qquad J \coloneqq \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}. \tag{27}$$

The matrix defining F_ρ is orthogonal, so F_ρ is 1-Lipschitz, and its unique zero is the origin. Moreover,

$$\langle F_\rho(x), x\rangle = -\rho\|x\|^2 = -\rho\|F_\rho(x)\|^2. \tag{28}$$

Identify $\mathbb{R}^2$ with $\mathbb{C}$, let w_k denote the complex representative of x_k, and set $\mu_\rho = -\rho + i\sqrt{1-\rho^2}$. The scalar recurrence for w_k has characteristic polynomial

$$p(\lambda) = \lambda^2 - \big(1 - (\alpha+\beta)\mu_\rho\big)\lambda - \beta\mu_\rho. \tag{29}$$

Theorem 6.1. *Let $-1 \le \rho < 1$ and $\alpha, \beta > 0$. Both roots of* (29) *lie strictly inside the unit disk if and only if $(\alpha, \beta) \in \mathcal{R}_{\rho,1}$. On the curved boundary $\max\{0, \rho\} < \beta < 1$ and $\alpha = \Psi_{\rho,1}(\beta)$, the polynomial has the unit-modulus root*

$$\lambda_\star = \frac{1 - \alpha\beta - \beta^2 - \alpha\mu_\rho}{1 - \beta^2}. \tag{30}$$

Consequently, the trajectory with $w_0 = 1$ and $w_1 = \lambda_\star$ satisfies $\|F_\rho(x_k)\| = 1$ for all k.

Proof. Applied to (29), the degree-two Schur–Cohn criterion (Zhang et al., 2022, Appendix D, Lemma D.2) states that both roots lie in the open unit disk if and only if

$$\beta < 1, \qquad |\alpha\beta + \beta^2 - 1 + \alpha\mu_\rho| < 1 - \beta^2. \tag{31}$$

Using $\operatorname{Re}\mu_\rho = -\rho$ and $|\mu_\rho| = 1$, a direct computation gives

$$(1-\beta^2)^2 - |\alpha\beta + \beta^2 - 1 + \alpha\mu_\rho|^2 = \alpha\big[2(\beta-\rho)(1-\beta^2) - \alpha(1 + \beta^2 - 2\beta\rho)\big] = \alpha^2 c. \tag{32}$$

The two inequalities in (31) are therefore equivalent to $(\alpha, \beta) \in \mathcal{R}_{\rho,1}$. On the curved boundary, substitution gives $p(\lambda_\star) = 0$ and (32) gives $|\lambda_\star| = 1$; details appear in Appendix A.4. Whenever Schur stability fails, p has a root λ with $|\lambda| \ge 1$. The initial values $w_0 = 1$ and $w_1 = \lambda$ select the mode $w_k = \lambda^k$. Since the matrix defining F_ρ is orthogonal, the residual norm equals $|\lambda|^k$; it does not vanish when $|\lambda| = 1$, and the trajectory is unbounded when $|\lambda| > 1$.

□

For $\rho = 1/2$, F_ρ is the rotation that Gorbunov et al. (2023, Theorem 4.5) use to argue that no choice of step sizes guarantees convergence of optimistic gradient once $\rho L \ge 1/2$. Their argument bounds the spectral norm of the iteration matrix, whereas Theorem 6.1 shows that its spectral radius is below one for every $(\alpha, \beta) \in \mathcal{R}_{1/2,1}$, in agreement with Böhm (2023, Theorem A.1).

The case $\rho \ge 1$ in Theorem 3.2 is simpler. The map $F(x) = -x$ is 1-Lipschitz and satisfies (2) for every $\rho \ge 1$, while its characteristic polynomial $\lambda^2 - (1 + \alpha + \beta)\lambda + \beta$ takes the value $-\alpha < 0$ at $\lambda = 1$. It therefore has a real root larger than one, and the corresponding trajectory diverges for every $\alpha, \beta > 0$.

7 Numerical experiments

We compare operator-query costs on a quadratic minimax problem, a high-dimensional extension, and a nonlinear Polar game. The quadratic benchmark comes from Pethick et al. (2022, Example 5) and was also used to evaluate HPE (Pethick et al., 2025, Section 9). The underlying objective is $axy + b(x^2 - y^2)/2$, with $a > 0$ and $b < 0$. Setting $b = -r$ and $a = \sqrt{1 - r^2}$ gives

$$F_r(x, y) = (-rx + \sqrt{1-r^2}\, y, -\sqrt{1-r^2}\, x - ry), \qquad L = 1, \quad \rho = r.$$

We compare GOGD tuned over $\mathcal{R}_{r,1}$ and over the union of the four earlier regions in Section 4, together with AdaptiveEG+ (Pethick et al., 2022) and the error-corrected inexact proximal point method of Pethick et al. (2025), denoted HPE. We also test the untuned rule (14), referred to as the explicit rule. For the quadratic problem and its extension, each tuned method receives 32 candidates. We count operator queries to relative squared residual $\|F(z)\|^2/\|F(z_1)\|^2 \le 10^{-8}$, using the accounting convention in Appendix C.2.

On the two-dimensional problem at $r = 1/3$, with every method tuned on the instance itself, HPE is fastest (Figure 2, left; query counts in Appendix C.1).

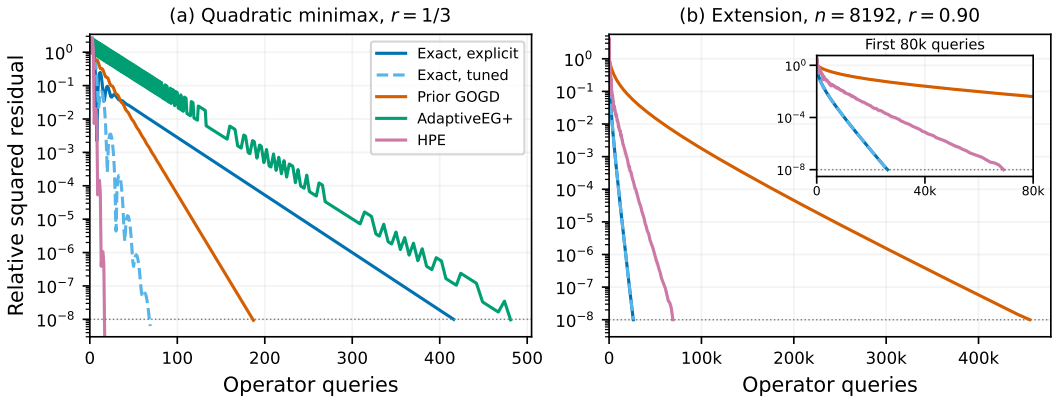


Figure 2: Relative squared residual versus operator queries (including inner evaluations). Left: the planar problem at $r = 1/3$, tuned on the instance. Right: the extension at $r = 0.90$, $n = 8192$, $\kappa = 3$; median and interquartile range over ten held-out seeds. The inset shows the first $80{,}000$ queries.

We then extend the planar operator to block-diagonal instances with singular values in $[1/\kappa, 1]$ that retain $\langle F(z), z\rangle = -r\|F(z)\|^2$ and $L = 1$. For each r, we tune on two instances with $n = 256$ and evaluate on ten held-out instances with $n = 8192$ and $\kappa = 3$ (Appendix C). Because every instance satisfies the weak-Minty inequality with equality and contains the extremal operator of Section 6, these experiments illustrate what the enlarged region gains on the operators that determine it, rather than typical performance. After an orthogonal change of variables, GOGD decouples into planar recursions, so its query count is governed by the slowest block. Table 1 reports median query counts. At all four tested values of r, tuning over $\mathcal{R}_{r,1}$ selects the explicit rule, and the improvement over prior-region tuning is largest at $r = 0.90$.

Table 1: Median queries over ten held-out seeds ($n = 8192$, $\kappa = 3$); 32 candidates per tuned method. "Prior" searches the union of earlier GOGD regions. A dash marks values of r for which the evaluated AdaptiveEG+ specialization has no admissible configuration.

r	Exact-explicit	Exact-tuned	Prior GOGD	AdaptiveEG+	HPE
0.25	**494**	**494**	852	583	537.5
0.45	905	905	3,116.5	4,357	**884.5**
0.75	**4,254**	**4,254**	29,952.5	—	6,470.5
0.90	**26,109.5**	**26,109.5**	454,147.5	—	68,718.5

At $r = 0.90$, the explicit rule reaches the target in 2.6×10^4 median operator queries, compared with 4.5×10^5 for the best prior-region configuration and 6.9×10^4 for HPE, that is, about 17 and 2.6 times fewer. Appendix C.3 repeats this comparison with a larger tuning budget. The right panel of Figure 2 shows the corresponding residual trajectories. Appendix C.4 complements the query counts with wall-clock timings on three held-out instances.

As a nonlinear example, we use a globally Lipschitz continuation of the Polar game of Pethick et al. (2022, Example 3(iii)), for which $\rho L \approx 0.10$ (Appendix D). Exact-region GOGD uses 961 median queries, versus 1,224 for the best prior-region configuration. Since $\rho L < \sqrt{6}/16$, equal parameters remain admissible, and tuned equal-parameter GOGD is nearly as fast (973); AdaptiveEG+ is fastest (664.5).

8 Conclusion

We have characterized the GOGD parameter region for weak-Minty operators with $-1 \le \rho L < 1$; for monotone operators, it is exactly the region of iterate convergence. Under co-hypomonotonicity, the same potential also yields last-iterate $o(1/N)$ rates on a subregion, and on the full region for affine operators. Open questions include full-region last-iterate rates for nonlinear operators, adaptation to unknown constants, and constrained or stochastic extensions.

AI USE STATEMENT

Generative AI tools assisted throughout the research process, including with literature retrieval and discovery, identifying the Lyapunov function, developing and presenting parts of the mathematical proofs, writing experimental code, and drafting and polishing parts of the manuscript. The authors take full responsibility for the correctness, originality, and final content.

REPRODUCIBILITY STATEMENT

The main text and appendices state the assumptions and give complete proofs of the theoretical results. Appendices C and D describe the experimental instances, tuning and evaluation protocols, operator-query accounting, and wall-clock measurements. Code is available from the authors upon request.

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.

Ioannis Anagnostides and Ioannis Panageas. Frequency-domain representation of first-order methods: A simple and robust framework of analysis. In *Symposium on Simplicity in Algorithms (SOSA)*, pp. 131–160. SIAM, 2022. doi: 10.1137/1.9781611977066.10.

Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics. Springer, Cham, 2nd edition, 2017. doi: 10.1007/978-3-319-48311-5.

Heinz H. Bauschke, Walaa M. Moursi, and Xianfu Wang. Generalized monotone operators and their averaged resolvents. *Mathematical Programming*, 189(1):55–74, 2021. doi: 10.1007/s10107-020-01500-6.

Axel Böhm. Solving nonconvex–nonconcave min–max problems exhibiting weak Minty solutions. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=Gp0pHyUyrb.

Yang Cai, Argyris Oikonomou, and Weiqiang Zheng. Finite-time last-iterate convergence for learning in multi-player games. In *Advances in Neural Information Processing Systems*, volume 35, pp. 33904–33919, 2022.

Sayantan Choudhury, Eduard Gorbunov, and Nicolas Loizou. Single-call stochastic extragradient methods for structured non-monotone variational inequalities: Improved analysis under weaker conditions. In *Advances in Neural Information Processing Systems*, volume 36, pp. 64918–64956, 2023. doi: 10.52202/075280-2833.

Jelena Diakonikolas, Constantinos Daskalakis, and Michael I. Jordan. Efficient methods for structured nonconvex–nonconcave min–max optimization. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 2746–2754. PMLR, 2021. URL https://proceedings.mlr.press/v130/diakonikolas21a.html.

Pontus Giselsson. Nonlinear forward-backward splitting with projection correction. *SIAM Journal on Optimization*, 31:2199–2226, 2021. doi: 10.1137/20M1345062.

Noah Golowich, Sarath Pattathil, and Constantinos Daskalakis. Tight last-iterate convergence rates for no-regret learning in multi-player games. In *Advances in Neural Information Processing Systems*, volume 33, pp. 20766–20778, 2020a.

Noah Golowich, Sarath Pattathil, Constantinos Daskalakis, and Asuman Ozdaglar. Last iterate is slower than averaged iterate in smooth convex–concave saddle point problems. In *Proceedings of the Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 1758–1784. PMLR, 2020b.

Eduard Gorbunov, Adrien B. 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, 2022.

Eduard Gorbunov, Adrien B. Taylor, Samuel Horváth, and Gauthier Gidel. Convergence of proximal point and extragradient-based methods beyond monotonicity: The case of negative comonotonicity. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 11614–11641. PMLR, 2023. URL https://proceedings.mlr.press/v202/gorbunov23a.html.

Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. Explore aggressively, update conservatively: Stochastic extragradient methods with variable stepsize scaling. In *Advances in Neural Information Processing Systems*, volume 33, pp. 16223–16234, 2020.

Ruichen Jiang and Aryan Mokhtari. Generalized optimistic methods for convex–concave saddle point problems. *SIAM Journal on Optimization*, 35(3):2066–2097, 2025. doi: 10.1137/24M1630475.

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

Yura Malitsky. Projected reflected gradient methods for monotone variational inequalities. *SIAM Journal on Optimization*, 25(1):502–520, 2015. doi: 10.1137/14097238X.

Yura Malitsky. Golden ratio algorithms for variational inequalities. *Mathematical Programming*, 184 (1–2):383–410, 2020. doi: 10.1007/s10107-019-01416-w.

Yura Malitsky and Matthew K. Tam. A forward-backward splitting method for monotone inclusions without cocoercivity. *SIAM Journal on Optimization*, 30(2):1451–1472, 2020. doi: 10.1137/18M1207260.

Aryan Mokhtari, Asuman E. 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, 2020a. doi: 10.1137/19M127375X.

Aryan Mokhtari, Asuman E. Ozdaglar, and Sarath Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1497–1507, 2020b. URL https://proceedings.mlr.press/v108/mokhtari20a.html.

Martin Morin, Sebastian Banert, and Pontus Giselsson. Nonlinear forward-backward splitting with momentum correction. *Set-Valued and Variational Analysis*, 31(4):37, 2023. doi: 10.1007/s11228-023-00700-4.

Arkadi Nemirovski. Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex–concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004. doi: 10.1137/S1052623403425629.

Nhung Hong Nguyen, Thanh Quoc Trinh, and Phan Tu Vuong. The Popov's algorithm with optimal bounded stepsize for generalized monotone variational inequalities, 2026. URL https://arxiv.org/abs/2603.06442. arXiv:2603.06442.

Thomas Pethick, Puya Latafat, Panagiotis Patrinos, Olivier Fercoq, and Volkan Cevher. Escaping limit cycles: Global convergence for constrained nonconvex–nonconcave minimax problems. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=2_vhkAMARk.

Thomas Pethick, Ioannis Mavrothalassitis, and Volkan Cevher. Efficient interpolation between extragradient and proximal methods for weak MVIs. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Y7slJZPGCy.

Leonid 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. doi: 10.1007/BF01141092.

Ernest K. Ryu, Kun Yuan, and Wotao Yin. ODE analysis of stochastic gradient methods with optimism and anchoring for minimax problems, 2019. URL https://arxiv.org/abs/1905.10899. arXiv:1905.10899.

Motahareh Sohrabi, Jianxin You, Simon Lacoste-Julien, Eduard Gorbunov, and Gauthier Gidel. Accelerated and stable convergence with anchored generalized optimistic method. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026. URL https://arxiv.org/abs/2606.21528.

Quoc Tran-Dinh and Yang Luo. Randomized block-coordinate optimistic gradient algorithms for root-finding problems. *Mathematics of Operations Research*, 51(1):746–782, 2026. doi: 10.1287/moor.2024.0414. URL https://arxiv.org/abs/2301.03113.

Quoc Tran-Dinh and Nghia Nguyen-Trung. Accelerated extragradient-type methods: Part 2—generalization and sublinear convergence rates under co-hypomonotonicity, 2025. URL https://arxiv.org/abs/2501.04585. arXiv:2501.04585.

Quoc Tran-Dinh and Nghia Nguyen-Trung. Revisiting extragradient-type methods: Part 1—generalizations and sublinear convergence rates. *Computational Optimization and Applications*, 94(1):141–210, 2026. doi: 10.1007/s10589-025-00727-3. URL https://arxiv.org/abs/2409.16859.

Manu Upadhyaya, Sebastian Banert, Adrien B. Taylor, and Pontus Giselsson. Automated tight Lyapunov analysis for first-order methods. *Mathematical Programming*, 209(1–2):133–170, 2025. doi: 10.1007/s10107-024-02061-8.

Manu Upadhyaya, Shuvomoy Das Gupta, Adrien B. Taylor, Sebastian Banert, and Pontus Giselsson. The AutoLyap software suite for computer-assisted Lyapunov analyses of first-order methods, 2026a. URL https://arxiv.org/abs/2506.24076. arXiv:2506.24076.

Manu Upadhyaya, Puya Latafat, and Pontus Giselsson. A Lyapunov analysis of Korpelevich's extragradient method with fast and flexible extensions. *Mathematical Programming*, 2026b. doi: 10.1007/s10107-025-02322-0.

Guojun Zhang and Yaoliang Yu. Convergence of gradient methods on bilinear zero-sum games. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SJlVY04FwH.

Guojun Zhang, Pascal Poupart, and Yaoliang Yu. Optimality and stability in non-convex smooth games. *Journal of Machine Learning Research*, 23(35):1–71, 2022.

A Additional proofs

A.1 Proofs for Section 3

Proof of Proposition 3.4. The sequence is bounded and $F(x_k) \to 0$. In finite dimensions, continuity makes every cluster point a zero; uniqueness then gives part (a). If F is demiclosed at zero, every weak cluster point is a zero, and uniqueness gives part (b) by Bauschke & Combettes (2017, Lemma 2.46). For part (c), repeat the Lyapunov construction with any $z \in \operatorname{zer} F$. Equation (36) then shows that $\|x_k - z\|$ has a limit for every zero. Opial's lemma, in the form of Bauschke & Combettes (2017, Lemma 2.47), yields weak convergence; weak and norm convergence coincide in finite dimensions. □

Proof of Corollary 3.3. Set $\rho = 0$. If $z \in \operatorname{zer} F$, monotonicity implies (2) relative to z. Moreover, F is demiclosed at zero by Bauschke & Combettes (2017, Corollary 20.28 and Proposition 20.38(ii)). Thus Theorem 5.3 and Proposition 3.4 give weak convergence and square summability within the region. For necessity, the scaled rotation $F = LJ$ in Theorem 6.1 is monotone and has the unique zero 0. Outside the stated region, an initial pair selects a characteristic mode of modulus at least one, so the iterates do not converge to 0. Since this example is finite dimensional, it also rules out weak convergence to a zero. □

A.2 Proofs for Section 4

Proof of Proposition 4.1. The denominator $1+\beta^2L^2-2\beta\rho L^2$ is positive when $\rho L<1$: for $\rho\le 0$ this is immediate, while for $0<\rho L<1$ its minimum over $\beta>0$ is $1-\rho^2L^2>0$. Hence the condition $c>0$ first requires $2(\beta-\rho)(1-\beta^2L^2)>0$. For $\beta>0$ and $\rho L<1$, this is equivalent to $\max\{0,\rho\}<\beta<1/L$, after which $c>0$ is equivalent to $\alpha<\Psi_{\rho,L}(\beta)$. Substituting $\alpha=\beta$ and collecting terms gives

$$\alpha L(1-3\alpha^2L^2)-2\rho L(1-2\alpha^2L^2)>0,$$

which is (12). For $\rho\ge 0$, this requires $0<\alpha L<1/\sqrt{3}$ and is equivalent to

$$0\le\rho L<\frac{\alpha L(1-3\alpha^2L^2)}{2(1-2\alpha^2L^2)},\qquad 0<\alpha L<\frac{1}{\sqrt{3}}.$$

Differentiating gives

$$\frac{\mathrm{d}}{\mathrm{d}(\alpha L)}\frac{\alpha L(1-3\alpha^2L^2)}{2(1-2\alpha^2L^2)}=\frac{(\alpha L-1)(\alpha L+1)(6\alpha^2L^2-1)}{2(2\alpha^2L^2-1)^2}.$$

The unique maximizer on $0<\alpha L<1/\sqrt{3}$ is therefore $\alpha L=1/\sqrt{6}$, where the ratio equals $\sqrt{6}/16$. Thus

$$\max_{0<\alpha L<1/\sqrt{3}}\frac{\alpha L(1-3\alpha^2L^2)}{2(1-2\alpha^2L^2)}=\frac{\sqrt{6}}{16},\qquad\text{at }\alpha L=\frac{1}{\sqrt{6}}.\tag{33}$$

□

We next verify that the rule (14) is admissible. Set $t=1-\rho L\in(0,1)$. The rule (14) satisfies $\rho L<\beta L<1$, $\alpha L>0$, and

$$2(\beta L-\rho L)(1-\beta^2L^2)-\alpha L(1+\beta^2L^2-2\beta\rho L^2)=\frac{9t^3}{128}>0.$$

Hence it lies strictly inside $\mathcal{R}_{\rho,L}$ for every $0<\rho L<1$. It takes $q=1/4$ in (13), with $\alpha L=3t/8$ matching the leading term of the exact upper boundary and $L\Psi_{\rho,L}(\beta)-\alpha L=9t^2/[8(32-7t)]>0$.

A.3 Proofs for Section 5

Proof of Lemma 5.1. Expanding V_k-V_{k+1} and using (15) gives

$$\begin{aligned}V_k-V_{k+1}=BL^2\big(\|d_k\|^2-\|d_{k+1}\|^2\big)-\|d_{k+1}\|^2-2\beta\langle d_{k+1},F_k\rangle\\+\big[\beta^2-2\rho(\alpha+\beta)\big]\|F_k\|^2+\beta(2\rho-\beta)\|F_{k-1}\|^2\\+2\alpha\big(\langle F_{k+1},x_{k+1}-x_\star\rangle+\rho\|F_{k+1}\|^2\big).\end{aligned}$$

Substituting $d_{k+1}=-\alpha F_k-\beta(F_k-F_{k-1})$ into the middle terms yields

$$\begin{aligned}&-\|d_{k+1}\|^2-2\beta\langle d_{k+1},F_k\rangle+\big[\beta^2-2\rho(\alpha+\beta)\big]\|F_k\|^2+\beta(2\rho-\beta)\|F_{k-1}\|^2\\&\quad=\frac{2\beta-2\rho-\alpha}{\alpha}\|d_{k+1}\|^2-B\|F_k-F_{k-1}\|^2.\end{aligned}$$

Since $c=(2\beta-2\rho-\alpha)/\alpha-BL^2$, this proves (18).

□

Proof of Lemma 5.2. Expanding the squared norm in (20) produces $-2B\langle F_k,F_{k-1}\rangle$, which cancels the $2B\langle F_k,F_{k-1}\rangle$ obtained from $-B\|F_k-F_{k-1}\|^2$. Using (16) and (11), the remaining nontrivial coefficients reduce as follows:

$$\begin{aligned}\|x_k-x_\star\|^2:&\quad\frac{\alpha(BL^2+1+c)}{2(\beta-\rho)}=1,\\\langle x_k-x_\star,F_k\rangle:&\quad\frac{\beta(2\beta-2\rho-\alpha)}{\beta-\rho}-\frac{\alpha B}{\beta(\beta-\rho)}=-2\alpha,\end{aligned}$$

$$\|F_{k-1}\|^2: \quad -B + \frac{2\beta^2(\beta-\rho)}{\alpha} = \beta(2\rho-\beta),$$
$$\|F_k\|^2: \quad -B + \frac{\rho\beta(2\beta-2\rho-\alpha)}{\beta-\rho} - \frac{\alpha B}{2(\beta-\rho)} + \frac{\alpha B^2}{2\beta^2(\beta-\rho)} = -2\alpha\rho.$$

The remaining coefficients, of $\|d_k\|^2$ and $\langle x_k - x_\star, F_{k-1}\rangle$, are BL^2 and 2β, respectively. Thus the expansion recovers (17).

On $\mathcal{R}_{\rho,L}$, we have $\beta-\rho>0$, $c>0$, and $\alpha<\Psi_{\rho,L}(\beta)<2(\beta-\rho)$, so the weak-Minty coefficient is positive. From (16), $B>0$ is immediate when $2\rho\le\beta$. If $\beta<2\rho<2\beta$, then

$$\Psi_{\rho,L}(\beta) < \frac{2\beta(\beta-\rho)}{2\rho-\beta},$$

because, after cancelling $2(\beta-\rho)$ and clearing the positive denominators, this inequality reduces to $\rho<\beta$. Hence $\alpha<\Psi_{\rho,L}(\beta)$ again gives $B>0$. Thus all coefficients in (20) are positive. The terms $L^2\|d_k\|^2-\|F_k-F_{k-1}\|^2$ and $L^2\|x_k-x_\star\|^2-\|F_k\|^2$ are nonnegative by Lipschitz continuity and $F(x_\star)=0$, and $\langle x_k-x_\star,F_k\rangle+\rho\|F_k\|^2\ge 0$ by (2). Dropping these terms and the squared norm leaves the last term of (20), which gives (21).

□

Proof of Theorem 5.3. Telescoping (19) and using $V_{N+1}\ge 0$ gives (22); averaging gives (23). Moreover, since $d_{k+1}=-\alpha F_k-\beta(F_k-F_{k-1})$ by (15), expanding the squares gives

$$\|d_{k+1}\|^2+\alpha\beta\big(\|F_{k-1}\|^2-\|F_k\|^2\big)=\alpha^2\|F_k\|^2+\beta(\alpha+\beta)\|F_k-F_{k-1}\|^2.$$

Together with (19), this yields

$$\big(V_k+\alpha\beta c\|F_{k-1}\|^2\big)-\big(V_{k+1}+\alpha\beta c\|F_k\|^2\big)\ge\alpha^2 c\|F_k\|^2+\beta(\alpha+\beta)c\|F_k-F_{k-1}\|^2. \tag{34}$$

Telescoping and $V_{N+1}\ge 0$ give

$$\sum_{k=1}^{N}\Big(\|F_k\|^2+\frac{\beta(\alpha+\beta)}{\alpha^2}\|F_k-F_{k-1}\|^2\Big)\le\frac{\beta}{\alpha}\|F_0\|^2+\frac{V_1}{\alpha^2 c}. \tag{35}$$

Averaging proves (25). Letting $N\to\infty$ proves square summability and (26), and applying the same bounds to the tail sums $\sum_{k=\lceil N/2\rceil}^{N}$ gives the $o(1/N)$ refinements.

Finally, (21) and $V_k\le V_1$ imply boundedness. Since $d_k\to 0$ and $F_k\to 0$, every correction term in (17) vanishes. Thus

$$V_k-\|x_k-x_\star\|^2\to 0. \tag{36}$$

The decreasing nonnegative sequence (V_k) converges, so $\|x_k-x_\star\|$ has a limit. □

With $d_1=x_1-x_0$, $F_0=F(x_0)$, and $F_1=F(x_1)$,

$$\begin{aligned}V_1=\|x_1-x_\star\|^2+BL^2\|d_1\|^2-2\alpha\langle x_1-x_\star,F_1\rangle+2\beta\langle x_1-x_\star,F_0\rangle\\ -2\alpha\rho\|F_1\|^2+\beta(2\rho-\beta)\|F_0\|^2.\end{aligned} \tag{37}$$

Although some coefficients in (37) may be negative, Lemma 5.2 proves that it is nonnegative and controls $\|x_1-x_\star\|^2$.

A.4 Details for Section 6

To verify the boundary-root formula in Theorem 6.1, note that $\mu_\rho^2+2\rho\mu_\rho+1=0$. Substitution of (30) into (29) gives

$$(1-\beta^2)^2p(\lambda_\star)=-\alpha\beta\mu_\rho\big[2(\beta-\rho)(1-\beta^2)-\alpha(1+\beta^2-2\beta\rho)\big].$$

Thus $p(\lambda_\star)=0$ on the curved boundary. In addition,

$$(1-\beta^2)^2-|1-\alpha\beta-\beta^2-\alpha\mu_\rho|^2=\alpha\big[2(\beta-\rho)(1-\beta^2)-\alpha(1+\beta^2-2\beta\rho)\big],$$

which proves $|\lambda_\star|=1$ on the boundary.

Remark A.1. For $\rho<0$, the weak-Minty condition and Cauchy–Schwarz imply $|\rho|\|F(x)\|^2\le\langle F(x),x-x_\star\rangle\le\|F(x)\|\|x-x_\star\|$. Hence a linear example $F(x)=Mx$ has $\|M\|\le 1/|\rho|$, so for $\rho L<-1$ the construction in Section 6, which requires $\|M\|=L$, is unavailable.

A.5 Proofs for Theorem 3.5

Throughout this subsection, F is L-Lipschitz with $\operatorname{zer} F \neq \emptyset$ and satisfies (6), $\rho L < 1$, and $(\alpha, \beta) \in \mathcal{R}_{\rho,L}$; in particular, $\rho < \beta < 1/L$. Then Assumption 3.1 holds for every $x_\star \in \operatorname{zer} F$, so all results of Section 5 apply.

Proof of Theorem 3.5(a). Every zero satisfies (2) with the same ρ, so by Proposition 3.4(c) it suffices to show that F is demiclosed at zero. The operator $A \coloneqq F^{-1} + \rho \operatorname{Id}$ has graph $\{(F(x), x + \rho F(x)) \mid x \in \mathcal{H}\}$, and (6) is equivalent to its monotonicity. For $w \in \mathcal{H}$, the solution x of $x + \beta F(x) = (\beta - \rho)w$ satisfies $w = F(x) + (x + \rho F(x))/(\beta - \rho)$, so $w \in (\operatorname{Id} + A/(\beta - \rho))(F(x))$. Hence $\operatorname{ran}(\operatorname{Id} + A/(\beta - \rho)) = \mathcal{H}$, and Minty's theorem (Bauschke & Combettes, 2017, Theorem 21.1) shows that $A/(\beta - \rho)$, and therefore A, is maximally monotone. Thus F is maximally ρ-co-hypomonotone (Bauschke & Combettes, 2017, Proposition 23.14). Now let $z_n \rightharpoonup z$ with $F(z_n) \to 0$. The pairs $(F(z_n), z_n + \rho F(z_n)) \in \operatorname{gra} A$ converge strongly in the first and weakly in the second component to $(0, z)$, so Bauschke & Combettes (2017, Proposition 20.38(i)) gives $(0, z) \in \operatorname{gra} A$, that is, $F(z) = 0$. □

For part (b), we pair the shifted potential in (34) with a residual function that is nonincreasing along the iteration. Since $\beta L < 1$, for each $y \in \mathcal{H}$ the map $p \mapsto y - \beta F(p)$ is a contraction, so the equation $p + \beta F(p) = y$ has a unique solution. If p and p' solve it for y and y', then (6) gives

$$\langle F(p) - F(p'), y - y' \rangle = \langle F(p) - F(p'), p - p' \rangle + \beta \|F(p) - F(p')\|^2 \geq (\beta - \rho)\|F(p) - F(p')\|^2, \tag{38}$$

so $y \mapsto F(p)$ is $(\beta - \rho)$-cocoercive.

For $k \geq 1$, let p_k solve $p_k + \beta F(p_k) = x_k + \beta F_{k-1}$. By (1), $(x_{k+1} + \beta F_k) - (x_k + \beta F_{k-1}) = -\alpha F_k$. Thus GOGD takes a forward step with step size α on the cocoercive map $y \mapsto F(p)$ from the point $x_k + \beta F_{k-1}$, but evaluates F at x_k instead of at p_k. With exact evaluations, the residuals $\|F(p_k)\|$ would be nonincreasing whenever $\alpha \leq 2(\beta - \rho)$. The function

$$R_k \coloneqq \|F(p_k)\|^2 + \frac{1}{\beta L}\|F_k - F(p_k)\|^2$$

adds the evaluation error to this residual, and controlling the error costs the factor $(1 - \beta L)/(1 + \beta L)$ in $\Phi_{\rho,L}$. Finally, set

$$\omega \coloneqq 1 + \frac{\alpha^2 L(1 + \beta L)}{(\alpha + \beta)(1 - \beta L)^2} \quad \text{and} \quad S \coloneqq \frac{\beta}{\alpha}\|F(x_0)\|^2 + \frac{V_1}{\alpha^2 c}.$$

With this choice of ω, Lemma A.2(ii) bounds R_k by ω times $\|F_k\|^2 + \beta(\alpha + \beta)\|F_k - F_{k-1}\|^2/\alpha^2$, the quantity whose sum is bounded by S in (35).

Lemma A.2. *For every $k \geq 1$,*

(i) $\|F_k\|^2 \leq (1 + \beta L)R_k$;

(ii) $R_k \leq \omega\big(\|F_k\|^2 + \beta(\alpha + \beta)\|F_k - F_{k-1}\|^2/\alpha^2\big)$;

(iii) *if $\alpha \leq \Phi_{\rho,L}(\beta)$, then $R_{k+1} \leq R_k$.*

Proof. Since $x_k - p_k = \beta(F(p_k) - F_{k-1})$, Lipschitz continuity gives

$$\|F_k - F(p_k)\| \leq \beta L \|F(p_k) - F_{k-1}\|. \tag{39}$$

(i) Young's inequality gives $\|F_k\|^2 \leq (1 + \beta L)\|F(p_k)\|^2 + (1 + 1/(\beta L))\|F_k - F(p_k)\|^2 = (1 + \beta L)R_k$.

(ii) Writing $F(p_k) - F_{k-1} = (F_k - F_{k-1}) - (F_k - F(p_k))$ in (39) gives $\|F_k - F(p_k)\| \leq \beta L\|F_k - F_{k-1}\|/(1 - \beta L)$. Young's inequality with weight $\omega - 1$ then gives

$$R_k \leq \omega\|F_k\|^2 + \Big(1 + \frac{1}{\omega - 1} + \frac{1}{\beta L}\Big)\frac{\beta^2 L^2}{(1 - \beta L)^2}\|F_k - F_{k-1}\|^2.$$

By the definition of ω, the last coefficient equals $\big(\omega - 1 + \beta L/(1+\beta L)\big)\beta(\alpha+\beta)/\alpha^2 \le \omega\beta(\alpha+\beta)/\alpha^2$.

(iii) Applying (38) to $y = x_{k+1} + \beta F_k$ and $y' = x_k + \beta F_{k-1}$, whose difference is $-\alpha F_k$, gives

$$-\alpha\langle F(p_{k+1}) - F(p_k), F_k\rangle \ge (\beta - \rho)\|F(p_{k+1}) - F(p_k)\|^2. \tag{40}$$

Moreover, $x_{k+1} - p_{k+1} = \beta(F(p_{k+1}) - F_k)$, so Lipschitz continuity gives $\|F_{k+1} - F(p_{k+1})\| \le \beta L\|F(p_{k+1}) - F_k\|$, and hence $R_{k+1} \le \|F(p_{k+1})\|^2 + \beta L\|F(p_{k+1}) - F_k\|^2$. Expanding both squares around $F(p_k)$ gives

$$\|F(p_{k+1})\|^2 = \|F(p_k)\|^2 + 2\langle F(p_{k+1}) - F(p_k), F(p_k)\rangle + \|F(p_{k+1}) - F(p_k)\|^2,$$
$$\|F(p_{k+1}) - F_k\|^2 = \|F(p_{k+1}) - F(p_k)\|^2 - 2\langle F(p_{k+1}) - F(p_k), F_k - F(p_k)\rangle + \|F_k - F(p_k)\|^2.$$

Writing $F(p_k) = F_k - (F_k - F(p_k))$ in the first inner product and using (40), we obtain

$$\begin{aligned} R_{k+1} - R_k \le \Big(1 + \beta L - \frac{2(\beta-\rho)}{\alpha}\Big)\|F(p_{k+1}) - F(p_k)\|^2 - \frac{1-\beta^2L^2}{\beta L}\|F_k - F(p_k)\|^2 \\ - 2(1+\beta L)\langle F(p_{k+1}) - F(p_k), F_k - F(p_k)\rangle. \end{aligned}$$

Young's inequality bounds the inner product term by $\frac{1-\beta^2L^2}{\beta L}\|F_k - F(p_k)\|^2 + \frac{\beta L(1+\beta L)}{1-\beta L}\|F(p_{k+1}) - F(p_k)\|^2$. Therefore

$$R_{k+1} - R_k \le \Big(\frac{1+\beta L}{1-\beta L} - \frac{2(\beta-\rho)}{\alpha}\Big)\|F(p_{k+1}) - F(p_k)\|^2,$$

whose coefficient is nonpositive if and only if $\alpha \le 2(\beta-\rho)(1-\beta L)/(1+\beta L) = \Phi_{\rho,L}(\beta)$. □

The following theorem gives the explicit bounds behind Theorem 3.5(b).

Theorem A.3. *Under the assumptions of Theorem 3.5, suppose $\alpha \le \Phi_{\rho,L}(\beta)$. Then*

$$\|F(x_N)\|^2 \le \frac{(1+\beta L)\,\omega S}{N} \qquad (N \ge 1),$$
$$\|x_{N+1} - x_N\|^2 \le \frac{(\alpha+2\beta)^2(1+\beta L)\,\omega S}{N-1} \qquad (N \ge 2).$$

Proof of Theorem 3.5(b) *and Theorem A.3.* By (34), the shifted potential $V_k + \alpha\beta c\|F_{k-1}\|^2$ decreases by at least $\alpha^2 c\big(\|F_k\|^2 + \beta(\alpha+\beta)\|F_k - F_{k-1}\|^2/\alpha^2\big)$, so Lemma A.2(ii) gives

$$\big(V_k + \alpha\beta c\|F_{k-1}\|^2\big) - \big(V_{k+1} + \alpha\beta c\|F_k\|^2\big) \ge \frac{\alpha^2 c}{\omega}R_k,$$

and $R_{k+1} \le R_k$ by Lemma A.2(iii). Telescoping, $V_{N+1} \ge 0$, and $R_N \le R_k$ for $k \le N$ give

$$NR_N \le \sum_{k=1}^{N} R_k \le \frac{\omega}{\alpha^2 c}\big(V_1 + \alpha\beta c\|F_0\|^2\big) = \omega S.$$

Lemma A.2(i) gives the residual bound. For $N \ge 2$, $\|x_{N+1} - x_N\| \le (\alpha+\beta)\|F_N\| + \beta\|F_{N-1}\| \le (\alpha+2\beta)\sqrt{(1+\beta L)R_{N-1}}$, which gives the increment bound. Finally, since (R_k) is nonincreasing and summable, $(N/2)R_N \le \sum_{k=\lceil N/2\rceil}^{N} R_k \to 0$, so both bounds improve to $o(1/N)$. □

Proof of Theorem 3.5(c). Write $F(x) = M(x - x_\star)$ with M bounded linear and $\|M\| \le L$. Applying M to (1) gives $F_{k+1} = F_k - (\alpha+\beta)MF_k + \beta MF_{k-1}$. Thus (F_k) is a GOGD trajectory for the linear operator M, which is L-Lipschitz and, by (6), satisfies (2) at the zero 0. Its potential (17), with $x_k - x_\star$ replaced by F_k and F by M, is

$$\begin{aligned} \tilde{V}_k = \|F_k\|^2 + BL^2\|F_k - F_{k-1}\|^2 - 2\alpha\big(\langle F_k, MF_k\rangle + \rho\|MF_k\|^2\big) \\ + 2\beta\langle F_k, MF_{k-1}\rangle + \beta(2\rho - \beta)\|MF_{k-1}\|^2. \end{aligned}$$

Lemmas 5.1 and 5.2 give $\tilde{V}_{k+1} \le \tilde{V}_k$ and $\tilde{V}_k \ge \alpha c\|F_k\|^2/(2(\beta-\rho))$. The third term is nonpositive, and $\|M\| \le L$ bounds the others, so $\tilde{V}_k \le C(\|F_k\|^2 + \|F_{k-1}\|^2)$ with $C \coloneqq 1 + 2BL^2 + \beta L + \beta\max\{0, 2\rho-\beta\}L^2$. By (35), $N\tilde{V}_N \le \sum_{k=1}^{N}\tilde{V}_k \le C(\|F_0\|^2 + 2S)$, and therefore

$$\|F(x_N)\|^2 \le \frac{2(\beta-\rho)C}{\alpha c N}\big(\|F(x_0)\|^2 + 2S\big).$$

As in part (b), since $(\tilde{V}_k)$ is nonincreasing and summable, $N\tilde{V}_N \to 0$. Finally, $\|x_{N+1} - x_N\|^2 \le 2(\alpha+\beta)^2\|F_N\|^2 + 2\beta^2\|F_{N-1}\|^2$, which gives the increment rate. □

Remark A.4. The two upper boundaries satisfy, for $\max\{0,\rho\} < \beta < 1/L$,

$$\Psi_{\rho,L}(\beta) - \Phi_{\rho,L}(\beta) = \frac{4(\beta-\rho)\beta L(1-\beta L)(1+\rho L)}{(1+\beta L)(1+\beta^2L^2 - 2\beta\rho L^2)},$$

which is positive when $\rho L > -1$. For every $\rho L < 1$, the subregion in Theorem 3.5(b) is nonempty, since $\Phi_{\rho,L}(\beta) > 0$ for $\max\{0,\rho\} < \beta < 1/L$. For $0 \le \rho L < 1$, the choice $\beta L = (1+\rho L)/2$ and $\alpha L = (1-\rho L)^2/(3+\rho L)$ attains $\alpha = \Phi_{\rho,L}(\beta)$, and substitution gives $c = (1+\rho L)^2$.

Remark A.5. The order $1/N$ cannot be improved uniformly over the class. Let $\rho \ge 0$, $(\alpha,\beta) \in \mathcal{R}_{\rho,L}$, and $0 < \varepsilon \le L$. The operator εJ is monotone, so it satisfies (6), and its zero is 0. Replacing μ_ρ by $i\varepsilon$ in (29), the simple root 1 at $\varepsilon = 0$ continues to an analytic root

$$\lambda_\varepsilon = 1 - i\alpha\varepsilon - \alpha\beta\varepsilon^2 + O(\varepsilon^3), \qquad |\lambda_\varepsilon|^2 = 1 - \alpha(2\beta-\alpha)\varepsilon^2 + O(\varepsilon^3).$$

Since $\alpha < \Psi_{\rho,L}(\beta) < 2(\beta-\rho) \le 2\beta$, we have $|\lambda_\varepsilon| < 1$ for small ε. The initial values $w_0 = 1$ and $w_1 = \lambda_\varepsilon$ select $w_N = \lambda_\varepsilon^N$, so $\|F(x_N)\|^2 = \varepsilon^2|\lambda_\varepsilon|^{2N}$. With $\varepsilon = N^{-1/2}$, $N\|F(x_N)\|^2 = |\lambda_\varepsilon|^{2N} \to e^{-\alpha(2\beta-\alpha)}$. Thus no bound $CN^{-1-\eta}$ with $\eta > 0$ and C depending only on α, β, ρ, L and the initial distances to the solution holds over the class, consistent with the $\Omega(1/\sqrt{N})$ last-iterate lower bounds for constant-step extragradient and OG of Golowich et al. (2020a;b). Faster rates are obtained by changing the recurrence, as in the time-varying anchored and Nesterov-type schemes of Tran-Dinh & Nguyen-Trung (2025).

B Comparison with prior weak-Minty bounds

For Böhm (2023, Theorem 3.1), the identification $\beta = s$ and $\alpha = s\gamma$, together with the conversion of the weak-Minty convention, gives $s > 2\rho$ and $sL \le (1-\gamma)/(1+\gamma)$. These conditions yield (7). The constant-step GRAAL parameterization is $\beta = \eta/\varphi$ and $\alpha = \eta(\varphi-1)/\varphi$ (Alacaoglu et al., 2023, Theorem 7).

For $\rho \ge 0$, the sufficient region (7) from Böhm's Theorem 3.1 lies strictly inside $\mathcal{R}_{\rho,L}$. Indeed, for $2\rho < \beta < 1/L$,

$$\Psi_{\rho,L}(\beta) - \frac{\beta(1-\beta L)}{1+\beta L} = \frac{(1-\beta L)\big[\beta L(1+4\beta L+\beta^2L^2) - 2\rho L(1+2\beta L)\big]}{L(1+\beta L)(1+\beta^2L^2-2\beta\rho L^2)} > 0. \tag{41}$$

For Gorbunov et al. (2023, Theorem 4.4, first part), eliminating the auxiliary iterate identifies $\beta = \gamma_1$ and $\alpha = \gamma_2$; their weak-Minty convention agrees with (2). We next verify the strict inclusion of (9). When $2\rho < \beta \le (1/L + 2\rho)/2$, its active upper bound is $\beta - 2\rho$, and

$$\Psi_{\rho,L}(\beta) - (\beta - 2\rho) = \frac{\beta\big[1 - 4\rho^2L^2 - 3\beta L(\beta L - 2\rho L)\big]}{1+\beta^2L^2-2\beta\rho L^2} > 0. \tag{42}$$

Indeed, the bracket decreases with βL on this interval and equals $(1-4\rho^2L^2)/4$ at $\beta = (1/L+2\rho)/2$. When $(1/L+2\rho)/2 \le \beta < 1/L$, the active upper bound is $1/L - \beta$, and

$$\Psi_{\rho,L}(\beta) - \left(\frac{1}{L} - \beta\right) = \frac{(1/L-\beta)(\beta^2L^2 + 2\beta L - 2\rho L - 1)}{1+\beta^2L^2-2\beta\rho L^2} > 0, \tag{43}$$

because the second factor is increasing in βL and equals $(1+2\rho L)^2/4$ at $\beta = (1/L+2\rho)/2$. Thus (9) lies strictly inside $\mathcal{R}_{\rho,L}$. The two prior sufficient regions are not nested: at $(\rho L, \beta L) = (0, 7/10)$, their upper bounds on αL are $21/170$ and $3/10$, whereas at $(\rho L, \beta L) = (69/200, 7/10)$ they are

$21/170$ and $1/100$, respectively. The deterministic specialization of the stochastic past-extragradient method in Choudhury et al. (2023, Proposition B.6 and Corollary 4.6) also reduces to (1). It requires $\max\{2\rho, 1/(2L)\} < \beta < 1/L$ and $0 < \alpha < \min\{\beta - 2\rho, (1/L - \beta)/4\}$, a strict subregion of (9).

For Böhm (2023, Theorem A.1), the parameter identification $s = \beta$ and $\gamma = \alpha/\beta$ converts the conditions to

$$\lambda\alpha > 2\rho, \qquad 0 \le \lambda \le \frac{2\beta}{\alpha}, \qquad \beta L \le \frac{2\beta - \lambda\alpha - \alpha}{2\beta - \lambda\alpha + \alpha}, \qquad 0 < \alpha \le \beta.$$

The step-size inequality is equivalent to $\lambda\alpha \le 2\beta - \alpha(1+\beta L)/(1-\beta L)$. Hence an admissible λ exists if and only if (8) holds. For $0 \le \rho L < \beta L < 1$, the difference between the upper bounds is

$$L\Psi_{\rho,L}(\beta) - \frac{2(\beta L - \rho L)(1-\beta L)}{1+\beta L} = \frac{4\beta L(\beta L - \rho L)(1-\beta L)(1+\rho L)}{(1+\beta L)(1+\beta^2 L^2 - 2\beta\rho L^2)} > 0.$$

Hence (8) lies strictly inside $\mathcal{R}_{\rho,L}$.

For the one-block Tran-Dinh–Luo specialization, set $\eta = \alpha+\beta$ and $\gamma = \beta/(\alpha+\beta)$. Their Section 3.2 descent lemma gives $\beta^2 L^2(1+2\sigma) \le \sigma$ and $(1+\sigma)\alpha < \beta - \rho$ for an auxiliary parameter $\sigma > 0$. These conditions are equivalent to

$$\frac{\beta^2 L^2}{1 - 2\beta^2 L^2} \le \sigma < \frac{\beta - \rho}{\alpha} - 1, \qquad \beta L < \frac{1}{\sqrt{2}}.$$

Eliminating σ gives (10). For $0 \le \rho L < \beta L < 1/\sqrt{2}$, its upper bound is strictly below $\Psi_{\rho,L}(\beta)$ because

$$\begin{aligned} &2(1-\beta^2L^2)^2 - (1-2\beta^2L^2)(1+\beta^2L^2-2\beta\rho L^2) \\ &\quad = (1-\beta^2L^2)(1-2\beta^2L^2) + 2\beta^4L^4 + 2\beta\rho L^2(1-2\beta^2L^2) > 0. \end{aligned}$$

C Quadratic minimax experiments

We give the tuning protocol, baseline updates, and additional results for the quadratic experiments in Section 7.

C.1 Experimental setup

For the planar experiment, we use the operator in Section 7 at $r = 1/3$, with $z_0 = z_1$ a unit vector (seed 2001). Each tuned method evaluates 32 candidates on this instance and selects the one with the smallest query count. Rotational equivariance and linear scaling make the relative residual histories independent of the nonzero starting point. On the planar instance at $r = 1/3$, tuning over $\mathcal{R}_{r,1}$ gives GOGD with 69 queries, compared with 187 for prior-region GOGD, 481 for AdaptiveEG+, and 17 for HPE; the explicit rule uses 416 queries.

We extend the same planar family to heterogeneous blocks, preserving its weak-Minty constant r. For $n = 2m$, choose $s_i \in [1/\kappa, 1]$, including both endpoints, and set

$$M_{r,s} = s\begin{bmatrix} -rs & \sqrt{1-r^2s^2} \\ -\sqrt{1-r^2s^2} & -rs \end{bmatrix}, \qquad D = \operatorname{blkdiag}(M_{r,s_1}, \dots, M_{r,s_m}). \tag{44}$$

For an orthogonal matrix Q, let $F(z) = Q^\top DQz$. Since $M_{r,s}^\top M_{r,s} = s^2 I$ and $(M_{r,s} + M_{r,s}^\top)/2 = -rs^2 I$, we obtain

$$\|F\| = 1, \qquad \langle F(z), z\rangle = -r\|F(z)\|^2, \qquad \sigma_{\min}(F) = 1/\kappa.$$

Consequently, the origin is the unique zero, the smallest admissible weak-Minty constant is r, and $\|z\| \le \kappa\|F(z)\|$.

For the high-dimensional study, we use $r \in \{0.25, 0.45, 0.75, 0.90\}$ and $\kappa = 3$, with approximately logarithmically spaced block scales and uniform unit-sphere initializations. For GOGD, $z_0 = z_1$. We tune each method over 32 candidates on two $n = 256$ instances (seeds $101, 102$). We select the smallest mean query count among candidates succeeding on both instances and freeze the configuration for ten $n = 8192$ test instances (seeds $1001, \dots, 1010$). The target is relative squared

residual 10^{-8}, with a cap of 2×10^6 queries per instance. All methods receive $L = 1$ and $\rho = r$. For each GOGD region, the candidates form a deterministic low-discrepancy (Halton) design: β covers the admissible interval, and α is a fraction of the region's upper boundary at that β. The prior-region search splits its candidates evenly across the earlier regions that are nonempty at r. The exact-region search uses the explicit rule (14), also evaluated without tuning, and 31 design points. AdaptiveEG+ and HPE use fixed grids over their step size and relaxation parameter, and, for HPE, the inner tolerance. Appendix A.2 verifies that the rule lies in $\mathcal{R}_{r,1}$.

C.2 Baseline methods

We specialize Algorithm 1 of Pethick et al. (2022) (AdaptiveEG+) and Algorithm 1 of Pethick et al. (2025) (HPE) to the unconstrained case. Both papers use the signed weak-Minty parameter $-r$, and we set their auxiliary parameter $\delta = -r$. The constant extrapolation step below is a specialization of AdaptiveEG+ covered by its convergence guarantee; its projection correction remains adaptive.

In all experiments, every new operator evaluation counts, including initialization, terminal checks, and rejected inner candidates; cached values are reused without charge. The stopping test is applied at every queried point. The primary comparisons measure operator queries; Appendix C.4 additionally reports wall-clock times.

For an outer iterate z, initialize $w = z$ and repeatedly compute

$$w^+ = z - \gamma F(w), \qquad v = \gamma F(w^+).$$

HPE accepts the first candidate satisfying

$$\langle z - w^+, v\rangle \geq (1-\sigma)\|v\|^2, \qquad r < \gamma < 1, \quad 0 < \sigma < 1 - r/\gamma. \tag{45}$$

If the test fails, set $w = w^+$ and continue, reusing $F(w^+)$. On acceptance, take

$$z^+ = z - \lambda \left(\frac{\langle z - w^+, v\rangle}{\|v\|^2} - \frac{r}{\gamma} \right) v, \qquad 0 < \lambda < 2. \tag{46}$$

AdaptiveEG+ uses one candidate $w^+ = z - \gamma F(z)$ and the same correction (46), with $2r < \gamma < 1$; this specialization is therefore available only for $r < 1/2$.

The residual test precedes acceptance and division by $\|v\|^2$, so an HPE inner candidate may terminate before passing (45). There is no additional projection-oracle cost in this unconstrained setting.

C.3 Additional results

Table 2 reports the query-count distributions for the high-dimensional experiment. All test runs reach the target, and the exact-region search selects the explicit rule for every r. The advantage does not rest on the rule alone: for $r \in \{0.25, 0.45, 0.75\}$, the best other exact-region candidate also needs fewer queries on the tuning instances than every prior-region candidate, and for $r = 0.90$ every other exact-region candidate was stopped early once it exceeded the rule's query count.

The additional GOGD candidates continue the Halton design, and the additional baseline candidates are drawn, with a fixed seed, uniformly from the interiors of the same parameter ranges. At $r = 0.90$, increasing the budget to 128 candidates reduces the prior-region and HPE medians to 385,612 and 59,845 queries, while exact-region GOGD remains at 26,109.5. The improvement factors therefore decrease from about 17.4 and 2.6 to about 14.8 and 2.3; the gains persist but depend on the search budget.

C.4 Wall-clock times

To complement the oracle counts, we measure wall-clock time on three of the held-out instances (seeds 1001–1003), with $n = 8192$, $\kappa = 3$, and the frozen configurations at $r \in \{0.45, 0.90\}$. All methods execute their direct recurrences with cached operator values and the same $O(n)$ block oracle, represented in complex coordinates. Each configuration receives a 128-query warm-up followed by three timed repetitions in randomized order. Tuning, instance construction, and output validation are excluded from timing. Table 3 summarizes the query counts and runtimes.

Table 2: Operator queries in the high-dimensional experiment: first quartile, median, and third quartile over ten held-out seeds, with linearly interpolated quartiles. AdaptiveEG+ is omitted for $r \geq 1/2$ (Appendix C.2).

r	Method	Q1	Median	Q3
0.25	Exact, explicit	493	494	495
0.25	Exact, tuned	493	494	495
0.25	Prior GOGD	850.25	852	853
0.25	AdaptiveEG+	581	583	583
0.25	HPE	536	537.5	539
0.45	Exact, explicit	903.25	905	906
0.45	Exact, tuned	903.25	905	906
0.45	Prior GOGD	3,110.5	3,116.5	3,120
0.45	AdaptiveEG+	4,346	4,357	4,377
0.45	HPE	883	884.5	886
0.75	Exact, explicit	4,246.5	4,254	4,259.75
0.75	Exact, tuned	4,246.5	4,254	4,259.75
0.75	Prior GOGD	29,897.5	29,952.5	29,990.25
0.75	HPE	6,460	6,470.5	6,483.75
0.90	Exact, explicit	26,060	26,109.5	26,143
0.90	Exact, tuned	26,060	26,109.5	26,143
0.90	Prior GOGD	453,295.25	454,147.5	454,728.5
0.90	HPE	68,600	68,718.5	68,825.25

Table 3: Wall-clock time on three held-out instances. Query counts are medians across instances. For time, we first take the median of three repetitions per instance, then report the median [minimum, maximum] across instances. Measurements use an Intel Core i7-1165G7 CPU and NumPy 2.5.3, single-threaded.

r	Method	Median queries	Seconds: median [min, max]
0.45	Exact GOGD	903	0.033 [0.032, 0.039]
0.45	Prior GOGD	3,110	0.094 [0.089, 0.095]
0.45	AdaptiveEG+	4,349	0.122 [0.117, 0.151]
0.45	HPE	883	0.028 [0.026, 0.035]
0.90	Exact GOGD	26,056	0.815 [0.785, 0.850]
0.90	Prior GOGD	453,228	15.094 [14.052, 15.514]
0.90	HPE	68,587	2.447 [2.438, 2.541]

The median runtime is lower for HPE at $r = 0.45$ and for exact-region GOGD at $r = 0.90$, consistent with the query-count ranking on this subset. The measurements are specific to this implementation on a shared machine: the synthetic block oracle does not model the cost of a dense operator, projection, or data-driven gradient.

D Polar game

D.1 Experimental setup

The Polar game of Pethick et al. (2022, Example 3(iii)) provides a nonlinear complement to the linear experiments. We use their parameter-$1/3$ instance and extend it to a globally Lipschitz operator. For $z = (x, y) \in \mathbb{R}^2$ and the rotation J in (27), set

$$F_{\mathrm{p}}(z) = \bar{h}(\|z\|^2)z + Jz, \qquad \bar{h}(t) = \frac{(\bar{t} - 1)(16\bar{t} - 9)}{48}, \qquad \bar{t} = \min\{t, 2(1.1)^2\}. \tag{47}$$

This agrees with the operator in that example on the disk of radius $\sqrt{2(1.1)^2}$, which contains their domain $[-1.1, 1.1]^2$. The origin is the unique zero, while $\dot{z} = -F_{\mathrm{p}}(z)$ has an attracting periodic orbit on the unit circle. The continuation preserves the constants from their Appendix C.2, with the weak-Minty sign converted to our convention (see Appendix D.3):

$$\rho \approx 0.01595, \qquad L \approx 6.306, \qquad \rho L \approx 0.1006.$$

We compare seven GOGD parameter families and the constant-extrapolation AdaptiveEG+ specialization described in Appendix C.2, applied to F_{p}/L, whose weak-Minty constant is ρL. Each search uses $1{,}600$ candidates on eight logarithmically spaced tuning radii in $[0.1, 1.5]$. For the equal-parameter family, these candidates are equally spaced interior points of the admissible interval in (12), whose endpoints are approximately 0.2115 and 0.5256 in $\alpha L = \beta L$. Selection minimizes tuning failures, then mean capped query count. We freeze the selected configurations and test them on 64 held-out radii drawn log-uniformly from $[0.1, 1.5]$.

We initialize GOGD at $z_0 = z_1 = (\varrho, 0)$, where ϱ is the radius; rotational equivariance makes the starting angle immaterial. The target is relative residual norm 10^{-8}, with a budget of $20{,}001$ operator evaluations per run. We use the query-accounting convention in Appendix C.2.

D.2 Numerical results

Table 4 reports the selected parameters and query-count quartiles. All selected configurations reach the target on all 64 test radii. Searching the exact region reduces the median GOGD cost from $1{,}224$ to 961 operator evaluations relative to the best prior-region search, about 21% fewer. Tuned equal-parameter GOGD uses 973 evaluations, about 1% more than exact-region GOGD, and AdaptiveEG+ uses 664.5.

Table 4: Operator evaluations on 64 held-out Polar radii, to relative residual norm 10^{-8} within $20{,}001$ queries, with linearly interpolated quartiles; parameters are rounded.

GOGD family	αL	βL	Queries: median [Q1, Q3]
Exact region	0.5014	0.6051	961 [952, 1,008.25]
Böhm, Theorem 3.1	0.1633	0.4934	2,888 [2,862, 3,013]
Böhm, Theorem A.1	0.2526	0.5612	1,836 [1,820, 1,912.75]
Gorbunov et al.	0.3802	0.6103	1,224 [1,213, 1,276.25]
Tran-Dinh–Luo	0.2584	0.5148	1,847 [1,830.75, 1,932]
GRAAL curve	0.1666	0.5005	2,821 [2,796, 2,941.75]
Equal parameters	0.5254	0.5254	973 [963, 1,032.75]
Adaptive method	γL	λ	Queries: median [Q1, Q3]
AdaptiveEG+	0.9805	1.8049	664.5 [657, 721]

All seven selected GOGD pairs and the selected adaptive configuration stay within radius 1.5 on both tuning and test runs. The continuation is therefore inactive on these runs.

D.3 GLOBAL CONTINUATION

We verify that the continuation in (47) preserves the stated weak-Minty and Lipschitz constants. The polynomial Polar operator is equivariant under rotations, so its Jacobian norm and weak-Minty ratio depend only on $\|z\|$. Every radius up to $\sqrt{2(1.1)^2}$ occurs in the original domain $[-1.1, 1.1]^2$; hence its bounds hold throughout that disk. Outside the disk, the continued radial coefficient is fixed and positive, so $\langle F_{\mathrm{p}}(z), z\rangle \geq 0$ implies the weak-Minty inequality. The outer linear map has norm $\sqrt{1 + \bar{h}(2(1.1)^2)^2}$. This equals the norm of the inner Jacobian applied to a unit tangent vector at the boundary, and is therefore at most L. Continuity across the circle and integration along line segments give the same global Lipschitz bound.