Acceleration of Generalized Optimistic Method with Anchoring

Anonymous Author
Anonymous Institution

Abstract

We study first-order methods for solving monotone variational inequalities arising in min-max optimization. Classical approaches such as the extragradient method rely on two gradient queries per iteration, which limits their analysis and applicability in the online and stochastic settings. We propose a family of Generalized Optimistic Methods with Anchoring (GOMA), which combine two time-scale optimistic updates with an anchoring term inspired by Halpern iteration. In particular, we show that for monotone Lipschitz operators, GOMA achieves an accelerated last-iterate convergence rate of $\mathcal{O}(1/k^2)$ in the squared gradient norm which is optimal. We also show that in stochastic games where classical methods, such as the extragradient and optimistic method, fail, GOMA can converge. Theoretically, we show that it has a last-iterate convergence rate of $\mathcal{O}(1/\sqrt{k})$ for monotone Lipschitz operators in stochastic regimes with linearly increasing minibatches.

1 Introduction

Minimax optimization and more generally, Variational Inequality (VI) problems, naturally arise in adversarial training, constrained optimization and multi-agent reinforcement learning, where the goal is to find equilibrium solutions under structured interaction of agents or competing objectives. When solving VIs classical gradient descent fails to converge even in simple bilinear games (Mertikopoulos et al., 2020). A breakthrough came with the extragradient method (EG) of Korpelevich (1976), which introduces a correction step and guarantees convergence under monotonicity. However, (i) EG requires two gradient evaluations per iteration, which is computationally expensive and makes it impractical in online or stochastic environments (Golowich et al., 2020). Moreover, subsequent work revealed that (ii) EG may fail in adversarial or stochastic regimes, motivating two-time-scale methods, such as EG+ algorithm (Hsieh et al., 2020). Above all (iii) EG has a last-iterate convergence guarantee of $O(1/k)$ for Monotone and Lipschitz operator in terms of operator norm, which is not optimal.

The optimistic method (Popov, 1980) reduces the per-iteration cost to a single gradient call by leveraging past gradients, addressing (i). Generalized optimistic method is a variant of optimistic method with two time-scales, addressing (ii). Also, anchoring, drawing inspiration from the Halpern fixed-point iteration (Halpern, 1967; Lieder, 2020), has recently emerged as a powerful mechanism for accelerating VI algorithms, addressing (iii). In this paper we introduce Generalized Optimistic Method with Anchoring (GOMA) which addresses these three issues simultaneously. The contributions of our paper are as follows:

1.1 Contributions

Our contributions are:

- We introduce GOMA, an accelerated two-time-scale algorithm requiring only one gradient evaluation per step.
- In the deterministic setting with monotone Lipschitz operators, GOMA attains an accelerated last-iterate convergence rate of $\mathcal{O}(1/k^2)$ in the squared operator norm, matching the complexity lower bounds.
- In the stochastic setting, where existing methods such as EG and optimistic gradient descent fail to converge, a simplified variant of GOMA (with zero exploration step size) achieves a last-iterate convergence rate of $\mathcal{O}(1/k)$ to a ball around the solution.

Preliminary work. Under review by AISTATS 2026. Do not distribute.

2 Related Work

2.1 Extragradient-type algorithms

For monotone Lipschitz operators, the extra-gradient algorithm (Korpelevich, 1976) achieves an optimal rate of $O\big(\frac{1}{k}\big)$ *ergodic* duality gap (Nesterov, 2007). This matches the lower bound of $O\big(\frac{1}{k}\big)$ from Nemirovski (2004). However, *the last-iterate* convergence of extra-gradient in terms of squared gradient norm for the same class of operators is $O\big(\frac{1}{k}\big)$ (Gorbunov et al., 2022a). However, this rate is not optimal, as the lower bound for this class is $O\big(\frac{1}{k^2}\big)$ (Yoon & Ryu, 2021) Similarly, optimistic gradient method also has a $O\big(\frac{1}{k}\big)$ last-iterate convergence rate for monotone Lipschitz operators (Gorbunov et al., 2022b).

2.2 Two time-scale Methods

Two time-scale strategies were introduced to stabilize extragradient dynamics in regimes where single-step methods fail. Hsieh et al. (2020) showed that running extrapolation with a larger stepsize than the correction step prevents divergence of stochastic EG, yielding almost sure convergence with last-iterate rates up to $O\big(\frac{1}{k}\big)$ in affine problems. In parallel, Lee & Kim (2021) proposed EG+, which extends this idea to smooth problems under the assumption of negative comonotonicity, ensuring convergence with an $O\big(\frac{1}{k}\big)$ squared-gradient rate.

2.3 Halpern-Type Acceleration

The *Halpern method* (Halpern, 1967) was presented for solving the fixed-point problem of Eq. (1) for some nonexpansive operator $T : \mathbb{R}^d \to \mathbb{R}^d$. If we would like to find the zeros of a gradient operator F we will define the operator $T = (\mathbb{I} + \alpha F)^{-1}$.

$$z_{k+1} = z_k - T(z_k) + \beta_k(z_0 - z_k) \tag{1}$$

When $\beta_k = \frac{1}{k+2}$ in Eq. (1), the method is called *optimized Halpern method (OHM)* (Lieder, 2020) and achieves $O\big(\frac{1}{k^2}\big)$ convergence rate.

To give a bit of intuition on why anchoring can accelerate the convergence of gradient descent on a monotone Lipschitz operator F, note that it turns the operator into an "almost" strongly-monotone one. Let

$$\widetilde{T}_k(z) \;=\; T(z) \;+\; \beta_k\,(z - z_0),$$

so that

$$z_{k+1} = z_k \;-\; \widetilde{T}_k(z_k) = z_k - \big[T(z_k) + \beta_k(z_k - z_0)\big].$$

Moreover, since T is monotone, we have

$$\begin{aligned}\big\langle \widetilde{T}_k(x) - \widetilde{T}_k(y),\, x - y \big\rangle &= \langle T(x) - T(y),\, x - y\rangle \\ &\quad + \beta_k\|x - y\|^2 \\ &\geq \beta_k\|x - y\|^2.\end{aligned}$$

i.e., $\widetilde{T}_k$ is β_k-strongly monotone. As $\beta_k \downarrow 0$, this "almost" strong convexity vanishes, recovering the original convex method.

Anchoring mechanism has emerged as the key mechanism for last-iterate acceleration of variational inequalities. Methods like *Extra Anchored Gradient (EAG)* algorithm (Yoon & Ryu, 2021), *Fast Extragradient (FEG)* algorithm (Lee & Kim, 2021), and *Anchored Popov*: (Tran-Dinh & Luo, 2021) all use anchoring mechanism to achieve acceleration.

Anchoring has also been applied in reinforcement learning, where it provides stability and accelerates convergence. Lee et al. (2021) shows that anchoring connects RL, quantal response equilibria, and zero-sum games by damping oscillations and guiding updates toward equilibria. More recently, anchoring has been used to accelerate value iteration Shi et al. (2022), yielding faster convergence without sacrificing optimality. These results highlight anchoring as a general mechanism for stabilizing and accelerating learning in sequential decision-making.

2.4 Stochastic Variational Inequalites

Under stochastic assumption, common variational inequality methods like extra-gradient and optimistic do not converge Hsieh et al. (2020). Pethick et al. (2023), Crespi (2004) show the convergence of Mirror-Prox-type methods under weak Minty (non-monotone-like) operators. Under monotone and Lipschitz assumptions, however, Lee & Kim (2021) shows convergence only with a growing batch sizes.

3 Preliminaries

In this section, we briefly review some basics for the class of problems under consideration, saddle-point problems and the associated vector field formulation. Let $G : \mathbb{R}^d \to \mathbb{R}^d$ be monotone and L-Lipschitz. Our performance metric is the last-iterate residual $\|G(x_k)\|^2$

We consider the saddle-point problem

$$\min_{x\in\mathcal{X}}\;\max_{y\in\mathcal{Y}}\; f(x, y),$$

where $\mathcal{X} \subseteq \mathbb{R}^{d_x}$ and $\mathcal{Y} \subseteq \mathbb{R}^{d_y}$ are closed convex sets and f is continuously differentiable. We define the

Choice of parameters	Resulting method
$\beta_k = 0$	Generalized Optimistic Method (Mokhtari et al., 2020)
$\gamma_k = \eta_k$	Anchored Popov (Tran-Dinh & Luo, 2021)
$\beta_k = 0,\ \gamma_k = \eta_k$	Optimistic Method / Popov / PEG (Popov, 1980)

Table 1: Special cases of Eq. (GOMA).

saddle-gradient operator

$$G(z) \ := \ \begin{pmatrix} \nabla_x f(x,y) \\ -\nabla_y f(x,y) \end{pmatrix}, \qquad z = (x,y).$$

A point $z^\star = (x^\star, y^\star)$ is a saddle point if and only if $G(z^\star) = 0$.

Saddle-point problems. Formally, let $\mathcal{L} : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}$ be a value function that assigns a cost $\mathcal{L}(\theta, \phi)$ to a player choosing $\theta \in \mathbb{R}^{d_1}$ and a payoff to a player choosing $\phi \in \mathbb{R}^{d_2}$. The saddle-point problem associated with $\mathcal{L}$ is to find a profile $(\theta^\star, \phi^\star) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ such that, for all $\theta \in \mathbb{R}^{d_1}$ and $\phi \in \mathbb{R}^{d_2}$,

$$\mathcal{L}(\theta^\star, \phi) \ \le \ \mathcal{L}(\theta^\star, \phi^\star) \ \le \ \mathcal{L}(\theta, \phi^\star). \tag{SP}$$

In this setting, the pair $(\theta^\star, \phi^\star)$ is a (global) saddle point of $\mathcal{L}$—or, in game-theoretic terms, a *Nash equilibrium.* For concision we often write $x = (\theta, \phi) \in \mathbb{R}^d$ with $d = d_1 + d_2$.

Definitions In this paper, we will make the following standard assumption for (stochastic) monotone Lipschitz variational inequalities.

- *Monotonicity.* $\langle G(x) - G(y),\, x - y\rangle \ge 0$, $\forall x, y$.
- *L-Lipschitz.* $\|G(x) - G(y)\| \le L\,\|x - y\|$, $\forall x, y$.
- *Oracle.* One query at x returns $G(x)$, or an unbiased estimate $\widehat{G}(x, \xi)$ with

$$\mathbb{E}[\widehat{G}(x,\xi) \mid x] = G(x), \tag{2}$$
$$\mathbb{E}\big[\|\widehat{G}(x,\xi) - G(x)\|^2 \mid x\big] \le \sigma^2. \tag{3}$$

In particular in Section 4 we will focus on monotone Lipschitz and in Section 5 on stochastic monotone Lipschitz operators.

4 Generalized Optimistic Method with Anchoring (GOMA)

In this section, we introduce the *Generalized Optimistic Method with Anchoring (GOMA)*, a family of algorithms that extend classical optimistic methods by incorporating an anchoring mechanism.

Generalized Optimistic Method with Anchoring (GOMA)

$$\begin{aligned} y_k &= \beta_k x_0 + (1-\beta_k)x_k - \gamma_k G(y_{k-1}) \\ x_{k+1} &= \beta_k x_0 + (1-\beta_k)x_k - \eta_k G(y_k), \end{aligned} \tag{GOMA}$$

Here, γ_k and η_k denote the step-sizes for the exploration and update steps, respectively. The coefficient $\beta_k \in [0,1)$ is the anchoring parameter, which gradually decays to zero as $k \to \infty$. Throughout this paper, we assume

$$\beta_k = \frac{a}{k+b} \quad \text{for} \quad a, b > 0.$$

Several well-known algorithms arise as special cases of GOMA as shown in the Table 1:

- Setting $\beta_k = 0$ recovers the *generalized optimistic method* (Mokhtari et al., 2020).
- Setting $\gamma_k = \eta_k$ yields the *anchored Popov algorithm* (Tran-Dinh & Luo, 2021).
- Setting both $\beta_k = 0$ and $\gamma_k = \eta_k$ reduces the scheme to the *classical Popov method* (Popov, 1980), also known as the *optimistic method* or *past extragradient (PEG).*

4.1 Proof Outline

We have two parameter choice: In one of them the exploration step-size is larger by a factor of $(1-\beta_k)$: $\eta_k = (1-\beta_k)\gamma_*$. and in the other one the update step-size is larger by the same factor: $\gamma_k = (1-\beta_k)\eta_*$. Each of these choices might be preferable depending on if an algorithm with a higher exploration step-size is needed or a higher update step.

To analyze the dynamics, we follow the standard Lyapunov approach and construct a potential function, which will serve as the basis for the descent argument.

Define the Lyapunov function

$$\begin{aligned} V_k = a_k\|G(x_k)\|^2 &+ b_k\langle G(x_k),\, x_k - x_0\rangle \\ &+ c_k L^2\|x_k - y_{k-1}\|^2. \end{aligned} \tag{4}$$

Where $a_k = c_k = \frac{b_k \eta_k}{2\beta_k}$, and $b_{k+1} = \frac{b_k}{1-\beta_k}$.

Case I: larger update step. We first study the schedule with a larger update step-size, $\eta_k = \eta_*$ and exploration scaled by $(1-\beta_k)$. The next lemma states that the one-step Lyapunov decrease holds whenever (η_k, β_k) satisfy three elementary conditions that arise from bounding Lipschitz and monotonicity cross-terms.

Lemma 1 (One-step Lyapunov decrease)**.** *Let G be monotone and L-Lipschitz, and consider the iterates*

$$\begin{aligned} y_k &= \beta_k x_0 + (1-\beta_k)x_k - \eta_*(1-\beta_k)G(y_{k-1}), \\ x_{k+1} &= \beta_k x_0 + (1-\beta_k)x_k - \eta_* G(y_k). \end{aligned} \tag{$\triangle$}$$

We prove that the Lyapunov function (4) is decreasing if the step-size η_k satisfies the following conditions:

$$\eta_{k+1} \;\le\; \frac{\beta_{k+1}}{2M\,\eta_k\,\beta_k(1-\beta_k)}, \tag{5}$$

$$1 - 2M\eta_k^2(1-\beta_k)^2 - M\eta_k^2\beta_k^2 \ge 0, \tag{6}$$

$$\eta_{k+1} \le \frac{\beta_{k+1}}{2\,\beta_k(1-\beta_k)}\Big[\frac{2(1-\beta_k^2) - 4M\eta_k^2(1-\beta_k)^2 - \beta_k^2}{1 - 2M\eta_k^2(1-\beta_k)^2 - M\eta_k^2\beta_k^2}\eta_k\Big]. \tag{7}$$

Lemma 2. *Conditions of Eq. (5), Eq. (6), Eq. (7) are satisfied, with constant step-size $\eta_k = \eta_* \in (0, \frac{1}{2\sqrt{3}L})$, and the choice of $\beta_k = \frac{2}{k+6}$.*

Theorem 1. *Suppose G is monotone and L-Lipschitz continuous. Consider the updates of Eq. ($\triangle$) With the parameter choices*

$$\beta_k = \frac{2}{k+6}, \quad \eta_* \in \Big(0, \tfrac{1}{2\sqrt{3}L}\Big),$$

$$a_k = c_k = \frac{b_0\,\eta_*}{80}\,(k+4)(k+5)(k+6),$$

$$b_k = \frac{b_0}{20}\,(k+4)(k+5).$$

the Lyapunov decrease from Lemma 1 implies the bound

$$\|G(x_k)\|^2 \;\le\; \frac{16/\eta_*^2 + 40L^2}{(k+6)^2}\,\|x_0 - x^\star\|^2. \tag{8}$$

Also if we pick the largest admissible constant stepsize $\eta_* = \frac{1}{2\sqrt{3}L}$ we get the bound:

$$\|G(x_k)\|^2 \le \frac{232L^2}{(k+6)^2}\,\|x_0 - x^\star\|^2.$$

The bound (8) gives an $O(1/k^2)$ decay of the residual $\|G(x_k)\|^2$ with explicit constants and a single scalar hyperparameter η_*.

Case II: larger exploration step. We next analyze the complementary schedule with a larger exploration step-size, keeping the update scaled by $(1-\beta_k)$. This variant can be preferable when exploration requires a larger look-ahead while updates must remain conservative.

Lemma 3. *Let G be monotone and L-Lipschitz, and consider the iterates*

$$\begin{aligned} y_k &= \beta_k x_0 + (1-\beta_k)x_k - \gamma_*\,G(y_{k-1}), \\ x_{k+1} &= \beta_k x_0 + (1-\beta_k)x_k - (1-\beta_k)\gamma_*\,G(y_k). \end{aligned} \tag{$\square$}$$

The potential in (4) is non-increasing for the algorithm if the following conditions are satisfied.

$$\gamma_{k+1} \;\le\; \frac{\beta_{k+1}}{2M\,\beta_k\,\gamma_k}\frac{(1-\beta_k)^2}{(1-\beta_{k+1})}, \tag{9}$$

$$1 - 2M\gamma_k^2 - M\beta_k^2\gamma_k^2 \;\ge\; 0, \tag{10}$$

$$\gamma_{k+1} \le \frac{\beta_{k+1}}{2\,\beta_k(1-\beta_{k+1})}\left[\frac{(1-\beta_k)^2\,(2-4M\gamma_k^2) - \beta_k^2}{1 - 2M\gamma_k^2 - M\beta_k^2\gamma_k^2}\right]\gamma_k. \tag{11}$$

Lemma 4. *Conditions of Eq. (9), Eq. (10), Eq. (11) are satisfied, with constant step-size $\gamma_k = \gamma_* \in \Big(0, \frac{1}{\sqrt{3}L}\Big)$, and the choice of $\beta_k = \frac{2}{k+6}$.*

Theorem 2. *Suppose G is monotone and L-Lipschitz, and let $x^\star$ satisfy $G(x^\star) = 0$. Consider the update of ($\square$). If the step-size γ and anchoring coefficient β_k satisfy the conditions of Lemma 4, then the Lyapunov function of Eq. (4) is decreasing. This implies the bound*

$$\|G(x_k)\|^2 \;\le\; \frac{16/\gamma_*^2 + 48L^2}{(k+6)^2}\,\|x_0 - x^\star\|^2. \tag{12}$$

In particular, with the largest admissible constant step-size $\gamma_* = \frac{1}{\sqrt{3}L}$ we obtain

$$\|G(x_k)\|^2 \;\le\; \frac{96L^2}{(k+6)^2}\,\|x_0 - x^\star\|^2.$$

Summary We outlined the proof for the two cases with fixed η_* and γ_* both of them resulted in a last-iterate acceleration. Note that this scheme can be viewed as a pseudo fixed step-size algorithm: in practice, hyperparameter tuning reduces to adjusting only η_* or γ_*. Meanwhile, the two-time-scale structure is maintained via the $(1-\beta_k)$ factor. This highlights the advantage of our bound over the anchored Popov method Tran-Dinh & Luo (2021), which requires a more intricate changing step-size analysis for a similar method.

4.2 Bilinear Example

Let $f : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ be $f(x,y) = Lxy$. Its saddle gradient operator and solution are $\boldsymbol{F}(x,y) = (Ly, -Lx)$ and $\boldsymbol{z}_* = (0,0)$, respectively. For the learning rate

$\frac{1}{L\theta}$, a random start point (x_0, y_0), we initialize at point $\boldsymbol{z}_0 = (-\frac{\theta}{b}y_0, \frac{\theta}{b}x_0)$, the sequence $\{\boldsymbol{z}_k\}_{k\geq 0}$ generated by GOMA Eq. ($\triangle$) satisfies $\boldsymbol{z}_k = (-\frac{\theta}{k+b}y_0, \frac{\theta}{k+b}x_0)$ for all $k \geq 0$. Hence, we have

$$\|G(z_k)\|^2 = \frac{(2\theta)^2 L^2}{(k+b-1)^2}\,\|z_0 - z^\star\|^2$$

UB vs. BB (d=6, same iterate z_k). Let $\eta_\star = \frac{1}{\theta L}$ with $\theta > 0$ and use $\beta_k = \frac{2}{k+6}$.

Bilinear Bound (BB). from the result above, we have

$$\|G(z_k)\|^2 \;=\; \frac{4\theta^2 L^2}{(k+4)^2}\,\|z_0 - z^\star\|^2. \tag{BB}$$

Universal Bound (UB).

$$\|G(z_k)\|^2 \;\leq\; \frac{L^2\,(16\theta^2+40)}{(k+6)^2}\,\|z_0 - z^\star\|^2. \tag{UB}$$

For the same initialization norm,

$$\frac{\mathrm{UB}}{\mathrm{BB}} \;\leq\; \left(\frac{k+4}{k+6}\right)^2 \frac{16\theta^2+40}{4\theta^2} \;\leq\; 6.5 \quad (\theta = 2).$$

Both bounds have the same $O(1/k^2)$ decay; UB is within a small constant factor and differs only by the index shift $(k+6)$ vs. $(k+4)$.

5 Variant of GOMA for Stochastic Settings

In this section we will discuss a variant of GOMA in which $\gamma_k = 0$. This results in the following conveniently simple algorithm:

$$\begin{aligned} y_k &= \beta_k x_0 + (1-\beta_k)x_k \\ x_{k+1} &= y_k - \eta_k G(y_k). \end{aligned} \tag{$\diamond$}$$

This method computes each update by evaluating the operator at an interpolated point between the initial iterate and the current iterate, followed by a gradient step anchored at this interpolation.

In what follows, we first establish a convergence rate of $\mathcal{O}(1/k)$ in the full-batch setting. Although this rate is not optimal in the full-batch setting, we show that the algorithm achieves favorable performance in the stochastic setting. In particular, it not only relaxes the noise assumptions of Lee & Kim (2021), but also demonstrates improved convergence behavior in bilinear experiments (see Figure Fig. 1a, Fig. 1b).

5.1 Non-Stochastic Rate

Theorem 3. *Assume G is monotone and L-Lipschitz. Consider*

$$y_k = \beta_k x_0 + (1-\beta_k)x_k, \qquad x_{k+1} = y_k - \eta_k\, G(y_k),$$

with $(y_{-1} = x_0)$. Potential of (4) with the coefficients $a_{k+1} = \frac{b_k\eta_k}{2\beta_k(1-\beta_k)}$ and $b_{k+1} = \frac{b_k}{1-\beta_k}$, $c_k = a_k$, and the choice of parameters:

$$\eta_k = \frac{c}{L}\sqrt{\frac{\beta_k}{2}}, \quad c \in (0, \frac{1}{\sqrt{2}}], \quad \beta_k = \frac{1}{k+2}.$$

admits a one-step Lyapunov decrease:

$$V_{k+1} \leq V_k$$

Thus we show that for all $k \geq 0$, with constants $C_{\text{init}} = \frac{8a_0L}{b_0c}$ and $C_\star = \frac{32}{c^2}$:

$$\|G(x_k)\|^2 \;\leq\; \frac{C_{\text{init}}\,L^2}{(k+1)^{3/2}}\,\|x_0 - x^\star\|^2 + \frac{C_\star\,L^2}{k+1}\|x_0 - x^\star\|^2.$$

In particular, $\|G(x_k)\|^2 = \mathcal{O}((k+1)^{-1})$.

5.2 Stochastic Rate

Assumption 4 (unbiased stochastic oracle with bounded variance)**.** *We assume that $\hat{G}(x,\xi)$ is an unbiased* stochastic oracle *of $G(x)$, i.e.,*

$$\mathbb{E}_\xi[\widehat{G}(x,\xi) \mid x] = G(x), \tag{13}$$

and that hat $\hat{G}(x,\xi)$ has bounded variance, i.e.,

$$\mathbb{E}_\xi\Big[\|\widehat{G}(x,\xi) - G(x)\|^2 \mid x\Big] \leq \sigma^2. \tag{14}$$

Note that on can leverage Assumption 4 to get vanishing variance by considering minibatches of independently sampled stochastic gradient,

$$\mathbb{E}_\xi\left[\left\|\frac{1}{B}\sum_{i=1}^{B}\widehat{G}(x,\xi_i) - G(x)\right\|^2 \mid x\right] \leq \frac{\sigma^2}{B}. \tag{15}$$

We will leverage that idea to get a convergence result in the stochastic case using linearly growing minibatches. More precisely, let us consider the updates

$$\begin{aligned} y_k &= \beta_k x_0 + (1-\beta_k)x_k, \\ x_{k+1} &= y_k - \frac{\eta_k}{k+1}\sum_{i=0}^{k}\widehat{G}(y_k,\xi_{k,i}), \end{aligned} \tag{16}$$

where $(\xi_{k,i})$ are i.i.d. By (15), the stochastic update has linearly vanishing variance $\sigma_k \leq \frac{\sigma^2}{k+1}$. Note that our analysis can be extended to a general sequence of stochastics oracles with such linearly vanishing variance.[1] This assumption is similar to the one in FEG where they assume that $\sigma_k \le \frac{\epsilon}{6(k+1)}$. Note that our assumption is slightly weaker since they need to consider an arbitrarily small epsilon to get convergence, while for us, $\epsilon = 1$ suffices for convergence.

Theorem 5 (Last-iterate bound for stochastic GOMA, bounded-variance oracle)**.** *Let $G : \mathbb{R}^d \to \mathbb{R}^d$ be monotone and L-Lipschitz and $\hat{G}(x,\xi)$ be a stochastic oracle following Assumption 4. Then for the updates described in (16) with $\beta_k = \frac{1}{k+2}$ and $\eta_k = \frac{1}{2L\sqrt{k+2}}$, we have that for all $N \ge 0$,*

$$\mathbb{E}\|G(x_N)\|^2 \;\le\; \frac{16L^2\|x_0 - x^*\|^2}{N+2} + \frac{6\sigma^2}{\sqrt{N+2}}$$

In comparison, FEG make a stronger assumption on σ_k and proves a convergence into a ball of size ϵ. To get a similar rate as ours, they would need to assume that $\sigma_k \le \frac{\sigma^2}{(k+1)^{3/2}}$, which is more restrictive that our assumption $\sigma_k \le \frac{\sigma^2}{k+1}$.

5.3 Bilinear Game

We consider the finite-sum saddle problem

$$\min_{\theta\in\mathbb{R}^d} \max_{\varphi\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n}\Big(\theta^\top b_i + \theta^\top A_i\varphi + c_i^\top \varphi\Big),$$

with saddle operator $F(\theta,\varphi) = [\bar{b} + \bar{A}\varphi;\ -(\bar{A}^\top\theta + \bar{c})]$, where $\bar{A} = \frac{1}{n}\sum_i A_i$ and $\bar{b}, \bar{c}$ are the sample means, $n = d$; $A_i = \mathrm{diag}(0,\dots,\lambda_i,\dots,0)$ with $\lambda_i \in [\tau, 1]$. The SFO returns $F_i(\theta,\varphi) = [b_i + A_i\varphi;\ -(A_i^\top\theta + c_i)]$ by sampling i uniformly.

See Fig. 1a. EG+ with decaying stepsizes quickly stalls at a high plateau and does not improve further, while FEG becomes unstable and grows over time. Overall, among the tested methods, only GOMA attains consistent decay of $\|F(z_k)\|^2$ to small values within the allotted gradient calls.

EG+ convergence on this instance, Recall from EG+ :

$$\begin{aligned}\mathbb{E}\big[\|X_{t+1} - x^\star\|^2 \,\big|\, \mathcal{F}_t\big] \le{}& (1 + C_t\kappa^2)\,\|X_t - x^\star\|^2 \\ &- \gamma_t\eta_t\big(1 - \gamma_t^2L^2 - 8\gamma_t\eta_t\kappa^2\big) \\ &\quad\times\|V(X_t)\|^2 \\ &+ C_t\sigma^2.\end{aligned} \tag{17}$$

where $C_t = 4\gamma_t^2\eta_t L + 2\gamma_t^3\eta_t L^2 + 4\eta_t^2 + 16\gamma_t^2\eta_t^2\kappa^2$. Using their error bound assumption:

$$\|V(x)\| \ge \tau\,\mathrm{dist}(x, X^\star)$$

a sufficient contraction condition is

$$\begin{aligned} 1 - \gamma_t^2L^2 - 8\gamma_t\eta_t\kappa^2 > 0,& \\ \text{and}\quad \gamma_t\eta_t\big(1 - \gamma_t^2L^2 - 8\gamma_t\eta_t\kappa^2\big)\tau^2 > C_t\kappa^2.&\end{aligned} \tag{18}$$

Since $1 - \gamma_t^2L^2 - 8\gamma_t\eta_t\kappa^2 \le 1$ and $C_t \ge 4\eta_t^2$

a necessary condition for the inequality above is

$$\gamma_t\eta_t\,\tau^2 \;>\; 4\,\eta_t^2\,\kappa^2 \qquad\Longrightarrow\qquad \frac{\gamma_t}{\eta_t} \;>\; \frac{4\kappa^2}{\tau^2}\,.$$

we set $\gamma_t = \alpha/\beta_{\text{alg}}$, $\eta_t = \alpha$(where $\beta_{\text{alg}} \ge 1$ is the fixed step–size ratio, based on their setting) yields the ratio

$$\frac{\gamma_t}{\eta_t} > \frac{4\kappa^2}{\tau^2} > 1, \quad \text{whereas} \quad \frac{\gamma_t}{\eta_t} = \frac{1}{\beta_{\text{alg}}} \le 1.$$

For high-dimensional games, the quantity $\tau \ll 1$ is tiny and $\kappa = 1 - \frac{1}{n} \approx 1$, which leads to contraction only for vanishingly small step sizes and leads to an arbitrarily slow convergence rate for EG+ (See for instance Fig 1).

5.4 Special Case: $\kappa = 0$

The preceding example satisfies the moment-control assumption

$$\mathbb{E}[\|Z_t\|^q \mid \mathcal{F}_t]\,\mathbf{1}_{\{X_t\in U\}} \le \big(\sigma + \kappa\|X_t - x^\star\|\big)^q,$$

with a strictly positive state-dependent term $\kappa \neq 0$. We now turn to the case $\kappa = 0$ (bounded variance without state dependence) and discuss the resulting behavior; See Fig. 1b.

On the stochastic bilinear game $f(x,y) = Lxy$ with $L = 1$, GOMA continues to drive the last-iterate residual $\|F(z_k)\|^2$ down by several orders of magnitude under $\kappa \neq 0$, whereas EG+ is essentially flat and FEG decreases only mildly. With gradient-call accounting on the x-axis, GOMA attains the best residual for a fixed oracle-query budget in this setting.

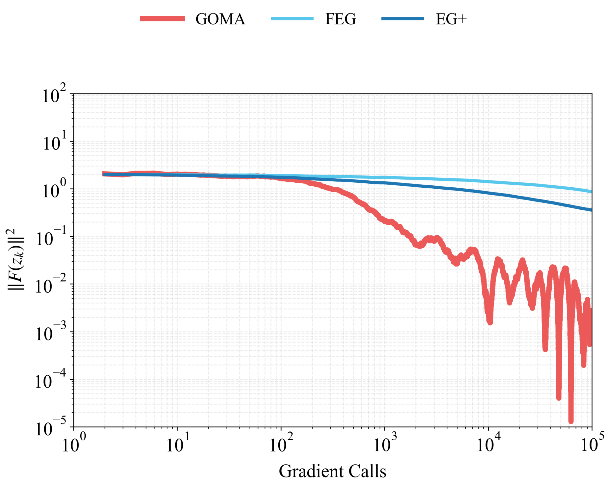
6 Experiments

Setup. We performed a toy experiment on a simple quadratic function also used in Lee & Kim (2021),

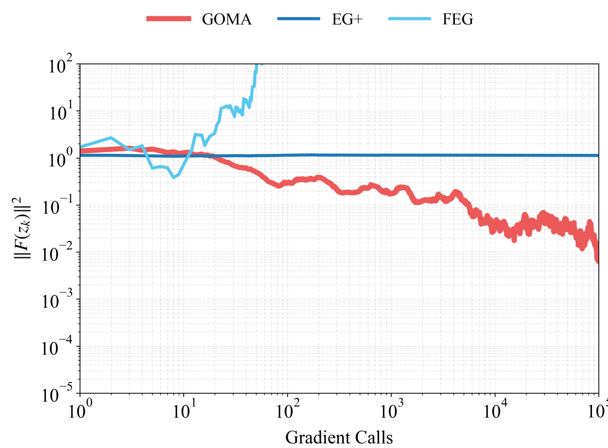
$$f(x,y) = -\tfrac{1}{6}x^2 + \tfrac{2\sqrt{2}}{3}xy + \tfrac{1}{6}y^2$$

We run all methods from a fixed nonzero initialization z_0 and evaluate convergence by the residual $\|F(z_k)\|^2$ versus total gradient calls (on a log–log scale), up to 10^5 calls.

[1] Since growing minibatches is the most practical way to achieve such a bound on σ_k, we decided to focus on that case for simplicity of the presentation and motivations.



(a) $\kappa = 0$: stochastic bilinear $f(x, y) = xy$ $(L = 1)$



(b) $\kappa \neq 0$: finite-sum stochastic saddle

Figure 1: Last-iterate residual vs. gradient calls (shared setup). Metric: $\|F(z_k)\|^2$ on log–log axes; x-axis counts oracle calls (one-query$= 1$/iter, two-query$= 2$/iter). Compared: GOMA (red), EG+ (blue), FEG (light blue).

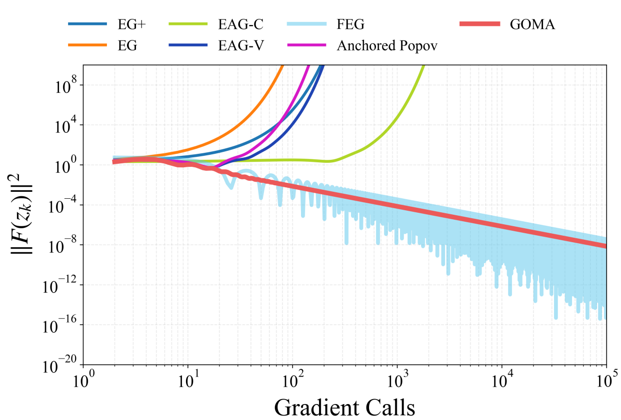


Figure 2: Numerical result with $f(x, y) = -\frac{1}{6}x^2 + \frac{2\sqrt{2}}{3}xy + \frac{1}{6}y^2$. Only GOMA and FEG converge. GOMA converges without oscillations, unlike FEG's dynamics.

Method. For (GOMA) with a $\eta = 0.2$ and $\gamma_k = 0.8(1 - \beta_k)$, where $\beta_k = \frac{2}{k+6}$. Figure 2 shows the squared norm of the operator $\|F(z_k)\|^2$ against the number of iterations for several algorithms, including EG+, EAG-C, EAG-V, Halpern, FEG, anchored popov and our proposed GOMA.

Results. See 2. GOMA and FEG converge with an accelerated rate, whereas EG+, EAG-C, EAG-V, and the (explicit) Halpern iteration diverge. Moreover, on this instance (under our tuning), GOMA yields uniformly smaller residuals than FEG by an approximately constant factor, with near-parallel log–log curves indicating the same asymptotic rate but a better constant

7 Discussion

Our results highlight that anchoring, when combined with generalized optimistic dynamics, provides a principled path toward overcoming three fundamental barriers in variational inequality algorithms: per-iteration cost, robustness to stochasticity, and last-iterate acceleration.

In particular, GOMA achieves the optimal $O(1/k^2)$ last-iterate rate in deterministic monotone Lipschitz settings with only a single gradient query per iteration, thus matching lower bounds while being more computationally efficient than extragradient-based methods. Moreover, the simplified stochastic variant of GOMA demonstrates convergence in regimes where both extragradient and optimistic methods diverge, underscoring anchoring as a stabilizing mechanism in noisy environments.

These findings suggest several promising directions for future work: extending the analysis to broader operator classes such as negative comonotone or weak Minty operators, refining the stochastic guarantees to obtain exact last-iterate convergence without growing minibatches, and exploring applications in reinforcement learning and large-scale adversarial training, where the tradeoff between gradient efficiency and stability is most critical.

References

Crespi, G. P. Minty variational inequalities and monotone trajectories of differential inclusions. *Journal of Inequalities in Pure and Applied Mathematics*, 5(2):Article 48, 2004. URL https://www.kurims.kyoto-u.ac.jp/EMIS/journals/JIPAM/images/137_03_JIPAM/137_03.pdf. (Cit. on p. 2)

Golowich, N., Pattathil, S., and Daskalakis, C. 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. Curran Associates, Inc., 2020. (Cit. on p. 1)

Gorbunov, E., Loizou, N., and Gidel, G. Extragradient method: O (1/k) last-iterate convergence for monotone variational inequalities and connections with cocoercivity. In *International Conference on Artificial Intelligence and Statistics*, pp. 366–402. PMLR, 2022a. (Cit. on p. 2)

Gorbunov, E., Loizou, N., and Gidel, G. Extragradient method: O(1/k) last-iterate convergence for monotone variational inequalities and connections with cocoercivity. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I. (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 366–402. PMLR, 28–30 Mar 2022b. URL https://proceedings.mlr.press/v151/gorbunov22a.html. (Cit. on p. 2)

Halpern, B. Fixed points of nonexpanding maps. *Bulletin of the American Mathematical Society*, 73:957–961, 1967. (Cit. on p. 1, 2)

Hsieh, Y.-G., Iutzeler, F., Malick, J., and Mertikopoulos, P. Explore aggressively, update conservatively: Stochastic extragradient methods with variable stepsize scaling. In *Advances in Neural Information Processing Systems*, volume 33, pp. 16223–16234. Curran Associates, Inc., 2020. (Cit. on p. 1, 2)

Korpelevich, G. M. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976. (Cit. on p. 1, 2)

Lee, S. and Kim, D. Fast extra gradient methods for smooth structured nonconvex-nonconcave minimax problems. In *Advances in Neural Information Processing Systems*, 2021. (Cit. on p. 2, 5, 6)

Lee, S., Kim, D., and Ryu, E. K. A unified approach to reinforcement learning, quantal response equilibria, and two-player zero-sum games. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 13450–13462, 2021. (Cit. on p. 2)

Lieder, F. On the convergence rate of the halpern-iteration. *Optimization Letters*, 15(2):405–418, 2020. ISSN 1862-4480. (Cit. on p. 1, 2)

Mertikopoulos, P., Papadimitriou, C., and Piliouras, G. Optimistic mirror descent in saddle-point problems: Going the extra(-gradient) mile. *Theoretical Computer Science*, 807:442–470, 2020. (Cit. on p. 1)

Mokhtari, A., Ozdaglar, A., and Pattathil, S. 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. PMLR, 26–28 Aug 2020. (Cit. on p. 3)

Nemirovski, A. 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. (Cit. on p. 2)

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

Pethick, T., Fercoq, O., Latafat, P., Patrinos, P., and Cevher, V. Solving stochastic weak minty variational inequalities without increasing batch size. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2302.09029. (Cit. on p. 2)

Popov, L. D. 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. (Cit. on p. 1, 3)

Shi, B., Lee, S., and Ryu, E. K. Accelerating value iteration with anchoring. In *International Conference on Machine Learning (ICML)*, pp. 19975–19994, 2022. (Cit. on p. 2)

Tran-Dinh, Q. and Luo, Y. Halpern-type accelerated and splitting algorithms for monotone inclusions. 2021. (Cit. on p. 2, 3, 4)

Yoon, T. and Ryu, E. K. Accelerated algorithms for smooth convex-concave minimax problems with $O(1/k^2)$ rate on squared gradient norm. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12098–12109. PMLR, 18–24 Jul 2021. (Cit. on p. 2)

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]
   (d) Information about consent from data providers/curators. [Yes]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]