Extragradient Method for (L_0, L_1)-Lipschitz Variational Inequalities

Anonymous Author(s)
Affiliation
Address
email

Abstract

Introduced by Korpelevich in 1976, the extragradient method (EG) has become a cornerstone technique for solving min-max optimization problems and variational inequalities problems (VIPs). Despite its longstanding presence and significant attention within the optimization community, most works focusing on understanding its convergence guarantees assume the strong L-Lipschitz condition. In this work, building on the proposed assumptions by Zhang et al. [2019] for minimization and Vankov et al. [2024] for VIPs, we focus on the more relaxed α-symmetric (L_0, L_1)-Lipschitz condition. This condition generalizes the standard Lipschitz assumption by allowing the Lipschitz constant to scale with the operator norm, providing a more refined characterization of problem structures in modern machine learning. Under the α-symmetric (L_0, L_1)-Lipschitz condition, we propose a novel step size strategy for EG and establish sublinear convergence rates for monotone operators and linear convergence rates for strongly monotone operators. Additionally, we prove local convergence guarantees for weak Minty variational inequality problems. We supplement our analysis with experiments validating our theory and demonstrating the effectiveness and robustness of the proposed step sizes for EG.

1 Introduction

Min-max optimization problems have recently attracted significant interest due to their widespread applications in machine learning, such as reinforcement learning [Brown et al., 2020], distributionally robust optimization [Namkoong and Duchi, 2016], and generative adversarial network training [Goodfellow et al., 2020]. These problems are often studied using a variational inequality problem perspective [Ryu and Yin, 2022, Gidel et al., 2018]. In the unconstrained case, the variational inequality problem (VIP) is defined as follows [Gorbunov et al., 2022b]:

$$\text{find } x_* \in \mathbb{R}^d \text{ such that } F(x_*) = 0, \tag{1}$$

where $F : \mathbb{R}^d \to \mathbb{R}^d$ is an operator. One of the well-known algorithms for solving variational inequality problems is the extragradient (EG) method [Korpelevich, 1977] due to its superior convergence guarantees [Gorbunov et al., 2022b]. The algorithm is defined as follows

$$\begin{aligned} \hat{x}_k &= x_k - \gamma_k F(x_k), \\ x_{k+1} &= x_k - \omega_k F(\hat{x}_k) \end{aligned} \tag{2}$$

where $\gamma_k > 0$ and $\omega_k > 0$ are the extrapolation step size and update step size, respectively. Since its original inception by Korpelevich, the EG method was revisited and extended in various ways, e.g., non-monotone VIPs [Diakonikolas et al., 2021, Fan et al., 2023] stochastic [Mishchenko et al., 2020, Gorbunov et al., 2022a, Choudhury et al., 2024, Li et al., 2022], distributed [Beznosikov et al., 2022]. Despite a rich literature for analysing EG and its variants, most of the existing convergence guarantees heavily rely on the L-Lipschitz assumption of the operator F [Korpelevich, 1977, Diakonikolas et al., 2021], i.e.

$$\|F(x) - F(y)\| \leq L\|x - y\| \tag{3}$$

for all $x, y \in \mathbb{R}^d$. However, this assumption can be restrictive; for instance, the operator $F(x) = x^2$ for $x \in \mathbb{R}$ does not satisfy (3) for any finite L [Zhang et al., 2019]. *The primary goal of this work is to relax this assumption and establish convergence guarantees under a more general framework.*

Relaxing the L-Lipschitz Assumption. Recently, Zhang et al. [2019] introduced the (L_0, L_1)-smoothness assumption for the minimization problems. Specifically, for $\min_{x \in \mathbb{R}^d} f(x)$, Zhang et al. [2019] assume $\|\nabla^2 f(x)\| \leq L_0 + L_1\|\nabla f(x)\|$ (when f is twice differentiable) and later [Chen et al., 2023] proved that this is equivalent to:

$$\|\nabla f(x) - \nabla f(y)\| \leq (L_0 + L_1\|\nabla f(x)\|)\|x - y\|. \tag{4}$$

Zhang et al. [2019] demonstrated that modern neural networks, such as LSTMs (Long Short-Term Memorys) [Hochreiter and Schmidhuber, 1997], align with (L_0, L_1)-smoothness assumption rather than the traditional L-smoothness assumption (i.e. $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$). Moreover, they used this assumption to justify why gradient clipping speeds up neural network training. Later, Ahn et al. [2023] showed that similar trends hold for the transformer [Vaswani et al., 2017] architecture.

In Figure 1, we present an example demonstrating the validity of the (L_0, L_1)-smoothness condition (4) for the iterates of the EG. Similar plots have been presented for gradient descent methods [Zhang et al., 2019, Gorbunov et al., 2024], but to the best of our knowledge, this linear connection between $\|\nabla^2 f(x_k)\|$ and $\|\nabla f(x_k)\|$ for EG iterates was never reported before. Following Gorbunov et al. [2024], we consider the minimization problem $\min_{x \in \mathbb{R}^d} f(x) := \log\left(1 + \exp(-a^\top x)\right)$, and plot the values of $\|\nabla^2 f(x_k)\|$ on the y-axis against $\|\nabla f(x_k)\|$ on the x-axis. Each point is colored according to the iteration index k, as indicated by the accompanying colorbar.

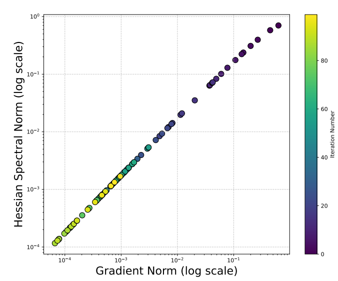


Figure 1: Scatter plot of $\|\nabla^2 f(x_k)\|$ on y-axis and $\|\nabla f(x_k)\|$ on x-axis.

The resulting plot reveals an approximately linear relationship between $\|\nabla^2 f(x_k)\|$ and $\|\nabla f(x_k)\|$, thereby supporting the modeling of this function within the $\|\nabla^2 f(x)\| \leq L_0 + L_1\|\nabla f(x)\|$ or (L_0, L_1)-smoothness framework.

In min-max optimization problem $\min_{w_1 \in \mathbb{R}^{d_1}} \max_{w_2 \in \mathbb{R}^{d_2}} \mathcal{L}(w_1, w_2)$, finding the equilibrium (i.e., find $x_* = (w_{1*}, w_{2*})^\top \in \mathbb{R}^d$ where $d = d_1 + d_2$ such that $\mathcal{L}(w_{1*}, w_2) \leq \mathcal{L}(w_{1*}, w_{2*}.) \leq \mathcal{L}(w_1, w_{2*})$, for every $w_1 \in \mathbb{R}^{d_1}$ and $w_2 \in \mathbb{R}^{d_2}$) is equivalent to solving the VIP (1) when $\mathcal{L}$ is convex-concave and F is defined as $F(x) := \left(\nabla_{w_1}\mathcal{L}(w_1, w_2)^\top, -\nabla_{w_2}\mathcal{L}(w_1, w_2)^\top\right)^\top$ [Gorbunov et al., 2022b]. If the operator F is L-Lipschitz, then its Jacobian matrix $\mathbf{J}(x)$, defined in (10), satisfies $\|\mathbf{J}(x)\| \leq L$ for all $x = (w_1^\top, w_2^\top)^\top$(follows from Theorem 2.1 with $L_1 = 0$). For example, consider the quadratic min-max objective $\min_{w_1} \max_{w_2} \mathcal{L}(w_1, w_2) = \frac{1}{2}w_1^2 + w_1 w_2 - \frac{1}{2}w_2^2$. In this case, implementing the EG method and plotting the Jacobian norm $\|\mathbf{J}(x_k)\|$ (on the y-axis) against the operator norm $\|F(x_k)\|$ (on the x-axis) yields a horizontal line parallel to x-axis (check Appendix A).

However, this behaviour does not persist for more complex problems. For instance, for the cubic objective

$$\min_{w_1} \max_{w_2} \mathcal{L}(w_1, w_2) = \tfrac{1}{3}w_1^3 + w_1 w_2 - \tfrac{1}{3}w_2^3, \tag{5}$$

$\|\mathbf{J}(x_k)\|$ increases with the $\|F(x_k)\|$. This observation suggests that the standard Lipschitz assumption may be overly restrictive for capturing the structure of such problems (check Figure 2).

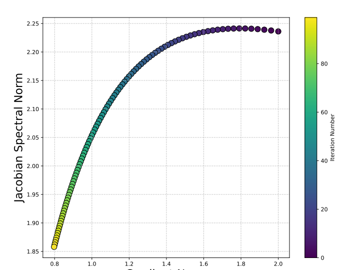


Figure 2: Scatter plot of $\|\mathbf{J}(x_k)\|$ on y-axis and $\|F(x_k)\|$ on x-axis.

To better model this relationship, we investigate a relaxed condition of the form $\|\mathbf{J}(x)\| \leq L_0 + L_1\|F(x)\|^\alpha$ with

$\alpha \in (0, 1]$ which generalizes the standard Lipschitz bound (for $L_1 = 0$, this boils down to $\|\mathbf{J}(x)\| \leq L_0$, which is the Lipschitz property). Note that, instead of $\alpha = 1$, our formulation permits α to lie in the broader interval $(0, 1]$. This condition is motivated by the plot in Figure 2, which suggests a sublinear relationship, resembling the form $h(r) = L_0 + L_1 r^\alpha$ for some $\alpha \in (0, 1)$, rather than a linear trend.

As we will prove later (in Theorem 2.1), for any doubly differentiable min-max optimization problem $\min_{w_1} \max_{w_2} \mathcal{L}(w_1, w_2)$, the condition $\|\mathbf{J}(x)\| \leq L_0 + L_1 \|F(x)\|^\alpha$ is equivalent to the operator F satisfying the α-symmetric (L_0, L_1)-Lipschitz condition (see Assumption 1.1). The equivalent α-symmetric (L_0, L_1)-Lipschitz condition does not rely on second-order information and applies to a broader class of problems (no need for double differentiability). Therefore, in the remainder of this work, we focus on analyzing the convergence of EG under the α-symmetric (L_0, L_1)-Lipschitz assumption [Vankov et al., 2024] on the operator F, defined below.

Assumption 1.1. F is called α-symmetric (L_0, L_1)-Lipschitz operator if for some $L_0, L_1 \geq 0$ and $\alpha \in (0, 1]$,

$$\|F(x) - F(y)\| \leq \left(L_0 + L_1 \max_{\theta \in [0,1]} \|F(\theta x + (1 - \theta) y)\|^\alpha\right) \|x - y\| \quad \forall x, y \in \mathbb{R}^d. \tag{6}$$

Instead of a fixed Lipschitz constant in (3), Assumption 1.1 allows the Lipschitz-like quantity to depend on the norm of the operator itself along the path from x to y. This assumption generalizes the standard L-Lipschitz condition (3), corresponding to the special case where $L_0 = L$ and $L_1 = 0$. Moreover, the α-symmetric (L_0, L_1)-Lipschitz condition provides a more refined characterisation of operators whose Lipschitz constant depends on their norm, offering a tighter bound by balancing $L_0 \ll L$ and $L_1 \ll L$. Additionally, (6) provides a more relaxed bound compared to (4) with $\alpha = 1$.

Classes of VIPs. Apart from the condition (6), we will also assume additional structure on the operator F to prove convergence. We say the operator F is monotone or strongly monotone if it satisfies the following assumption.

Assumption 1.2. F is called monotone if

$$\langle F(x) - F(y), x - y \rangle \geq 0 \quad \forall x, y \in \mathbb{R}^d \tag{7}$$

and strongly monotone if there is some $\mu > 0$ such that

$$\langle F(x) - F(y), x - y \rangle \geq \mu \|x - y\|^2 \quad \forall x, y \in \mathbb{R}^d. \tag{8}$$

This captures convex minimization and convex-concave min-max optimization problems as a special case. Apart from the monotone operators, we are also interested in some non-monotone operators, weak Minty operators [Diakonikolas et al., 2021], which satisfy the following assumption.

Assumption 1.3. Operator F is called weak Minty if for some $\rho \geq 0$,

$$\langle F(x), x - x_* \rangle \geq -\rho \|F(x)\|^2 \qquad \forall x \in \mathbb{R}^d. \tag{9}$$

1.1 Main Contributions

We summarize the main contributions of our work below.

- **Tighter analysis for strongly monotone:** We establish linear convergence guarantees for strongly monotone (8) α-symmetric (L_0, L_1)-Lipschitz problems (see Theorem 3.2, 3.4). In contrast to the results in Vankov et al. [2024] for $\alpha = 1$, our analysis shows that linear convergence can be achieved without incurring exponential dependence on the initial distance to the solution $\|x_0 - x_*\|$ (see Corollary 3.3).
- **First analysis for monotone and weak Minty:** We provide the first convergence analysis of EG for solving monotone (7) and weak Minty (9) variational inequality problems under α-symmetric (L_0, L_1)-Lipschitz assumption. We establish global sublinear convergence for monotone problems (see Theorem 3.5, 3.7) and local sublinear convergence for weak Minty problems (see Theorem 3.8, 3.9).
- **Novel step size for EG:** We propose a novel adaptive step-size strategy for the EG method designed to handle α-symmetric (L_0, L_1)-Lipschitz operators. Specifically, all our step-size schemes adopt

Table 1: Summary of step size selection for EG under the L-Lipschitz and α-symmetric (L_0, L_1)-Lipschitz assumptions. Our proposed step size strategy is of the general form $\gamma_k = \frac{1}{c_0 + c_1 \|F(x_k)\|^\alpha}$, tailored for solving problems involving α-symmetric (L_0, L_1)-Lipschitz operators.

Setup	Assumption	α	γ_k	ω_k
Strongly Monotone [(1)]	L-Lipschitz [(2)] [Mokhtari et al., 2020]	-	$\frac{0.25}{L}$	γ_k
	α-symmetric (L_0, L_1)-Lipschitz [Vankov et al., 2024]	1	$\min\left\{\frac{1}{4\mu}, \frac{1}{2\sqrt{2}eL_0}, \frac{1}{2\sqrt{2}eL_1\|F(x_k)\|}\right\}$	γ_k
	α-symmetric (L_0, L_1)-Lipschitz (Theorem 3.2)	1	$\frac{0.21}{L_0+L_1\|F(x_k)\|}$	γ_k
	α-symmetric (L_0, L_1)-Lipschitz (Theorem 3.4)	$(0,1)$	$\frac{0.61}{2K_0+\left(2K_1+2^{1-\alpha}K_2^{1-\alpha}\right)\|F(x_k)\|^\alpha}$ [(2)]	γ_k
Monotone [(1)]	L-Lipschitz [Gorbunov et al., 2022b]	-	$\frac{1}{L}$	$\frac{\gamma_k}{2}$
	α-symmetric (L_0, L_1)-Lipschitz (Theorem 3.5)	1	$\frac{0.45}{L_0+L_1\|F(x_k)\|}$	γ_k
	α-symmetric (L_0, L_1)-Lipschitz (Theorem 3.7)	$(0,1)$	$\frac{1}{2\sqrt{2}K_0+\left(2\sqrt{2}K_1+2^{3(1-\alpha)/2}K_2^{1-\alpha}\right)\|F(x_k)\|^\alpha}$	γ_k
Weak Minty [(1)]	L-Lipschitz [Diakonikolas et al., 2021]	-	$\frac{1}{L}$	$\frac{\gamma_k}{2}$
	L-Lipschitz [Pethick et al., 2023]	-	$\frac{1}{L}$	$\rho + \frac{\langle F(\hat{x}_k), x_k - \hat{x}_k\rangle}{\|F(\hat{x}_k)\|^2}$
	α-symmetric (L_0, L_1)-Lipschitz (Theorem 3.8)	1	$\frac{0.56}{L_0+L_1\|F(x_k)\|}$	$\frac{\gamma_k}{2}$
	α-symmetric (L_0, L_1)-Lipschitz (Theorem 3.9)	$(0,1)$	$\frac{1}{2\sqrt{2}K_0+\left(2\sqrt{2}K_1+2^{3(1-\alpha)/2}K_2^{1-\alpha}\right)\|F(x_k)\|^\alpha}$	$\frac{\gamma_k}{2}$

[(1)] Convergence measure • strongly monotone: $\|x_K - x_*\|^2$, • monotone: $\min_{0\le k\le K}\|F(x_k)\|^2$, • weak minty: $\min_{0\le k\le K}\|F(\hat{x}_k)\|^2$.
[(2)] For K_0, K_1, K_2, check Proposition 3.1. Note that, for $L_1 = 0$ we have $K_1 = K_2 = 0$.

the general form $\gamma_k = \frac{1}{c_0 + c_1\|F(x_k)\|^\alpha}$, where $c_0, c_1 > 0$ are constants determined by the problem-dependent parameters L_0, L_1, and α. In Table 1 we included a detailed summary of our proposed step size selection for different classes of VIPs and compared it with closely related works.

- **Numerical experiments:** Finally, in Section 4, we present experiments validating different aspects of our theoretical results. We compare our proposed step size selections with existing alternatives, demonstrating the effectiveness and robustness of our approach.

2 On the α-Symmetric (L_0, L_1)-Lipschitz Assumption

We divide this section into two parts. In the first subsection, we present an equivalent reformulation of the α-symmetric (L_0, L_1)-Lipschitz condition (6) in the context of min-max optimization. In the second subsection, we provide some examples of operators that satisfy (6) and highlight its significance.

2.1 Equivalent Formulation of α-Symmetric (L_0, L_1)-Lipschitz Assumption

In this subsection, we consider the min-max optimization problem $\min_{w_1}\max_{w_2}\mathcal{L}(w_1, w_2)$. The corresponding operator F and Jacobian $\mathbf{J}$ are defined as

$$F(x) = \begin{bmatrix} \nabla_{w_1}\mathcal{L}(w_1, w_2) \\ -\nabla_{w_2}\mathcal{L}(w_1, w_2) \end{bmatrix} \text{ and } \mathbf{J}(x) = \begin{bmatrix} \nabla^2_{w_1w_1}\mathcal{L}(w_1, w_2) & \nabla^2_{w_2w_1}\mathcal{L}(w_1, w_2) \\ -\nabla^2_{w_1w_2}\mathcal{L}(w_1, w_2) & -\nabla^2_{w_2w_2}\mathcal{L}(w_1, w_2) \end{bmatrix}, \tag{10}$$

where $x = (w_1^\top, w_2^\top)^\top$. Assuming that F is α-symmetric (L_0, L_1)-Lipschitz, we obtain the following theorem.

Theorem 2.1. Suppose F is the differentiable operator associated with the problem $\min_{w_1}\max_{w_2}\mathcal{L}(w_1, w_2)$. Then F satisfies the α-symmetric (L_0, L_1)-Lipschitz condition (6) if and only if

$$\|\mathbf{J}(x)\| \le L_0 + L_1\|F(x)\|^\alpha. \tag{11}$$

Here $\mathbf{J}(x)$ is the Jacobian defined in (10) and $\|\mathbf{J}(x)\| = \sigma_{\max}(\mathbf{J}(x))$ i.e. maximum singular value of $\mathbf{J}(x)$. In particular, we have $\|\mathbf{J}(x)\| \le L$ when operator F is L-Lipschitz.

This result provides an equivalent characterization of the α-symmetric (L_0, L_1)-Lipschitz condition (6) for min-max optimization problems. In practice, it is often easier to verify (11) than to directly check (6). In Appendix D, we provide an example where we use Theorem 2.1 to verify if an operator satisfies (6).

2.2 Examples of α-Symmetric (L_0, L_1)-Lipschitz Operators

To motivate the significance of this relaxed assumption (6), we present a few instances of α-symmetric (L_0, L_1)-Lipschitz operators that highlight its advantages over the conventional L-Lipschitz assumption.

Example 1[Gorbunov et al., 2024]: We start with an example from the minimization setting. Consider the logistic regression loss function $f(x) = \log\left(1 + \exp\left(-a^\top x\right)\right)$. Then the corresponding gradient operator $F = \nabla f$ satisfies the L-Lipschitz assumption with $L = \|a\|^2$ and α-symmetric (L_0, L_1)-Lipschitz assumption with $L_0 = 0, L_1 = \|a\|, \alpha = 1$. Therefore, when $\|a\| \gg 1$, the bound provided by 1-symmetric (L_0, L_1)-Lipschitz can be significantly tighter than the one imposed by the L-Lipschitz condition. This example emphasizes the benefit of the α-symmetric (L_0, L_1)-Lipschitz framework in scenarios where standard Lipschitz constants are overly pessimistic.

Example 2: Consider the operator $F(x) = (u_1^2, u_2^2)$ for $x = (u_1, u_2) \in \mathbb{R}^2$ with $x_* = (0, 0)$. Then we can show that

$$\|F(x) - F(y)\| \leq 2\left\|F\left(\tfrac{x+y}{2}\right)\right\|^{1/2}\|x - y\| \leq 2\left\|\max_{\theta\in[0,1]} F\left(\theta x + (1-\theta)y\right)\right\|^{1/2}\|x - y\|.$$

This establishes that F is $1/2$-symmetric $(0, 2)$-Lipschitz operator. However, this operator F does not satisfy the standard L-Lipschitz assumption for any finite choice of L. We add the related details to Appendix B. Therefore, this example highlights the need for relaxed assumptions on operators beyond standard L-Lipschitz (3).

We provide additional examples in Appendix B to illustrate cases where the operator associated with a bilinearly coupled min-max optimization problem or an N-player game satisfies the α-symmetric (L_0, L_1)-Lipschitz condition.

3 Convergence Analysis

In this section, we present the convergence guarantees of EG for solving monotone, strongly monotone, and weakly Minty operators. For strongly monotone operators, we have linear convergence, while for monotone and weak Minty operators, we provide sublinear convergence guarantees. To prove these results, we rely on the similar expression presented in Chen et al. [2023] for the (L_0, L_1)-smooth minimization problem. For completeness, we include the proof for α-symmetric (L_0, L_1)-Lipschitz operators in the Appendix.

Proposition 3.1. Suppose F is α-symmetric (L_0, L_1)-Lipschitz operator. Then, for $\alpha = 1$

$$\|F(x) - F(y)\| \leq (L_0 + L_1\|F(x)\|)\exp\left(L_1\|x - y\|\right)\|x - y\|, \tag{12}$$

and for $\alpha \in (0, 1)$ we have

$$\|F(x) - F(y)\| \leq \left(K_0 + K_1\|F(x)\|^\alpha + K_2\|x - y\|^{\alpha/1-\alpha}\right)\|x - y\| \tag{13}$$

where $K_0 = L_0(2^{\alpha^2/1-\alpha} + 1)$, $K_1 = L_1 \cdot 2^{\alpha^2/1-\alpha}$ and $K_2 = L_1^{1/1-\alpha} \cdot 2^{\alpha^2/1-\alpha} \cdot 3^\alpha(1-\alpha)^{\alpha/1-\alpha}$.

Proposition 3.1 eliminates the maximum over $\theta \in [0, 1]$ in (6) and provides a simpler upper bound on the $\|F(x) - F(y)\|$. We divide the rest of the section into three subsections based on the structure of operators. Moreover, each of these subsections is divided into two parts depending on the value of α, i.e. $\alpha = 1$ and $\alpha \in (0, 1)$.

3.1 Convergence Guarantees for Strongly Monotone Operators

In case of strongly monotone operators (8), we achieve linear convergence rates, analogous to those obtained under standard L-Lipschitz assumptions [Tseng, 1995, Mokhtari et al., 2020]. For $\alpha = 1$, operator F satisfies the condition (12). To guarantee convergence for this class of operators, we use the EG with step size $\gamma_k = \omega_k = \nu/(L_0+L_1\|F(x_k)\|)$ and $\nu > 0$. Gorbunov et al. [2024] used similar step sizes for the Gradient Descent algorithm to solve convex minimization problems.

Theorem 3.2. Suppose F is μ-strongly monotone and 1-symmetric (L_0, L_1)-Lipschitz operator. Then EG with step size $\gamma_k = \omega_k = \frac{\nu}{L_0+L_1\|F(x_k)\|}$ satisfy

$$\|x_{k+1} - x_*\|^2 \quad \leq \quad \left(1 - \frac{\nu\mu}{L_0\left(1 + L_1 \exp\left(L_1\|x_0 - x_*\|\right)\|x_0 - x_*\|\right)}\right)^{k+1} \|x_0 - x_*\|^2$$

where $\nu > 0$ satisfy $1 - 2\nu - \nu^2 \exp 2\nu = 0$.

The equation $1 - 2\nu - \nu^2 \exp(2\nu) = 0$ admits a positive solution, approximately $\nu \approx 0.363$. Specifically, to ensure $\|x_K - x_*\|^2 \leq \varepsilon$, we require $K = \mathcal{O}\left(\left(\frac{L_0}{\mu} + \frac{L_0 L_1\|x_0-x_*\| \exp\left(L_1\|x_0-x_*\|\right)}{\mu}\right)\log\frac{1}{\varepsilon}\right)$ iterations. When $L_1 = 0$, we recover the best-known results for the strongly monotone L-Lipschitz setting [Tseng, 1995, Mokhtari et al., 2020]. Vankov et al. [2024] also studied constrained strongly monotone problems and obtained similar guarantees with an alternative step size scheme.

However, using a refined proof technique, we can eliminate the $\exp(L_1\|x_0 - x_*\|)$ term from the convergence rate and establish a tighter bound. One of the intermediate steps of Theorem 3.2 is proving a lower bound on the step size γ_k, which can be very loose for large k. We now show that after a certain number of iterations K' (31), the operator norm satisfies $\|F(x_k)\| \leq L_0/L_1$ for all $k \geq K'$, which implies $\gamma_k = \omega_k \geq \nu/2L_0$ for all $k \geq K'$.

Corollary 3.3. Suppose F is a μ-strongly monotone and 1-symmetric (L_0, L_1)-Lipschitz operator. Then, EG with step sizes $\gamma_k = \omega_k = \frac{\nu}{L_0+L_1\|F(x_k)\|}$ guarantees $\|x_{K+1} - x_*\|^2 \leq \varepsilon$ after at most

$$K = \underbrace{\frac{2L_0}{\nu\mu}\log\left(\frac{\|x_0 - x_*\|^2}{\varepsilon}\right)}_{\text{Term I}} + \underbrace{\frac{1}{\zeta\mu}\log\left(\frac{2L_1\|x_0 - x_*\|^2}{\zeta^2 L_0}\right)}_{\text{Term II}} \tag{14}$$

iterations, where we have $\zeta \coloneqq \nu/L_0(1+L_1\exp(L_1\|x_0-x_*\|)\|x_0-x_*\|)$, and $\nu > 0$ satisfies $1 - 4\nu - 2\nu^2\exp(2\nu) = 0$.

This result shows that to reach an accuracy of $\varepsilon > 0$, we need (14) iterations. Importantly, Term II in (14) is independent of ε, and the Term I of (14) no longer depends on $\exp(L_1\|x_0 - x_*\|)$. Technically, Term II corresponds to the number of iterations required for the step sizes γ_k and ω_k to exceed $\nu/2L_0$, while Term I captures the iteration complexity of EG with a fixed step size $\nu/2L_0$.

Now, we investigate the behavior of α-symmetric (L_0, L_1)-Lipschitz operators for $0 < \alpha < 1$. In this regime, we adopt a step size of the order $\mathcal{O}\left(\|F(x_k)\|^{-\alpha}\right)$ and prove the following result.

Theorem 3.4. Suppose F is μ-strongly monotone and α-symmetric (L_0, L_1)-Lipschitz operator with $\alpha \in (0, 1)$. Then EG with $\gamma_k = \omega_k = \frac{\nu}{2K_0+\left(2K_1+2^{1-\alpha}K_2^{1-\alpha}\right)\|F(x_k)\|^\alpha}$ satisfy

$$\|x_{k+1} - x_*\|^2 \leq \left(1 - \frac{\nu\mu}{2K_0+(2K_1+2^{1-\alpha}K_2^{1-\alpha})(K_0+K_2\|x_0-x_*\|^{\alpha/1-\alpha})^\alpha\|x_0-x_*\|^\alpha}\right)^{k+1} \|x_0 - x_*\|^2$$

where $\nu \in (0, 1)$ is a constant such that $1 - \nu - \nu^2 = 0$.

This result establishes linear convergence. In particular, to ensure $\|x_K - x_*\|^2 \leq \varepsilon$, it suffices to run $K = \mathcal{O}\left(\left(\frac{K_0}{\mu} + \frac{(K_1K_2^\alpha+K_2)\|x_0-x_*\|^{\alpha/1-\alpha}}{\mu}\right)\log\frac{1}{\varepsilon}\right)$ iterations. Compared to the L-Lipschitz setting, the bound here includes an additional dependence on $\|x_0 - x_*\|^{\alpha/1-\alpha}$, which grows larger as $\alpha \to 1$.

3.2 Convergence Guarantees for Monotone Operators

In this subsection, we focus on the monotone operators (7). Here we provide the first analysis for the monotone 1-symmetric (L_0, L_1)-Lipschitz operators.

Theorem 3.5. Suppose F is monotone and 1-symmetric (L_0, L_1)-Lipschitz operator. Then EG with step size $\gamma_k = \omega_k = \frac{\nu}{L_0 + L_1\|F(x_k)\|}$ satisfy

$$\min_{0\le k\le K} \|F(x_k)\|^2 \le \frac{2L_0^2\left(1 + L_1 \exp\left(L_1\|x_0 - x_*\|\right)\|x_0 - x_*\|\right)^2 \|x_0 - x_*\|^2}{\nu^2(K+1)}. \tag{15}$$

where $\nu \exp \nu = 1/\sqrt{2}$.

Note that the solution of $\nu \exp \nu = 1/\sqrt{2}$ is approximately 0.45. Hence, this result proves sublinear convergence of EG when F is monotone. Moreover, (15) implies, EG will need $K = \mathcal{O}\left(\frac{L_0^2\|x_0 - x_*\|^2}{\varepsilon} + \frac{L_0^2 L_1^2 \exp\left(2L_1\|x_0 - x_*\|\right)\|x_0 - x_*\|^4}{\varepsilon}\right)$ iterations to get $\|F(x_k)\|^2 \le \varepsilon$ for some $k \le K$. Therefore the convergence rate exponentially depends on $\|x_0 - x_*\|$ when $L_1 > 0$. This shows that 1-symmetric (L_0, L_1)-Lipschitz operators potentially require more iterations of EG compared to L-Lipschitz operators when initialization x_0 is far from x_*. However, (15) recovers the best known dependence on $\|x_0 - x_*\|$ as a special case when $L_1 = 0$, i.e. F is a standard Lipschitz operator [Gorbunov et al., 2022b].

Theorem 3.5 shows that the EG's convergence rate has an extra term $\exp\left(L_1\|x_0 - x_*\|\right)$ compared to the results of the Lipschitz setting. One of the intermediate steps in this proof involves an upper bound on $\sum_{k=0}^K \gamma_k^2\|F(x_k)\|^2$ (see (35) in Appendix C). Then the simple approach is to get a lower bound on γ_k^2 for all k and derive (15). This lower bound on γ_k^2 involves the $\exp\left(L_1\|x_0 - x_*\|\right)$ term (see (34) in Appendix C) and can be potentially very small. However, it is possible to eliminate this exponential dependence using a refined proof technique.

Theorem 3.6. Suppose F is monotone and 1-symmetric (L_0, L_1)-Lipschitz operator. Then EG with step size $\gamma_k = \omega_k = \frac{\nu}{L_0 + L_1\|F(x_k)\|}$ satisfy

$$\min_{0\le k\le K} \|F(x_k)\| \le \frac{\sqrt{2}L_0\|x_0 - x_*\|}{\nu\sqrt{K+1} - \sqrt{2}L_1\|x_0 - x_*\|}$$

where $\nu \exp \nu = 1/\sqrt{2}$ and $K + 1 \ge \frac{2L_1^2\|x_0 - x_*\|^2}{\nu^2}$.

Note that to obtain this convergence guarantee, a sufficiently large number of iterations is required, specifically $K + 1 \ge \left(2L_1^2\|x_0 - x_*\|^2\right)/\nu^2$. Gorbunov et al. [2024] employed a similar proof technique to eliminate the exponential dependence on the initial distance $\exp(L_1\|x_0 - x_*\|)$ in the context of the Adaptive Gradient method.

Next we state our result for α-symmetric (L_0, L_1)-Lipschitz monotone operator with $\alpha \in (0, 1)$.

Theorem 3.7. Suppose F is monotone and α-symmetric (L_0, L_1)-Lipschitz operator with $\alpha \in (0, 1)$. Then EG with $\gamma_k = \omega_k = \frac{1}{2\sqrt{2}K_0 + \left(2\sqrt{2}K_1 + 2^{3(1-\alpha)/2}K_2^{1-\alpha}\right)\|F(x_k)\|^\alpha}$ satisfy

$$\min_{0\le k\le K} \|F(x_k)\|^2 \le \frac{16\left(K_0 + (K_1 + 2^{-3/2}K_2^{1-\alpha})(K_0 + K_2\|x_0 - x_*\|^{\alpha/1-\alpha})^\alpha \|x_0 - x_*\|^\alpha\right)^2 \|x_0 - x_*\|^2}{K+1}.$$

This theorem establishes a sublinear convergence rate for $\alpha \in (0, 1)$. In the special case where $L_1 = 0$ (i.e., the standard L-Lipschitz setting), we have $K_1 = K_2 = 0$ by Proposition 3.1. Thus, our result recovers the best-known rate $\mathcal{O}\left(L_0^2\|x_0 - x_*\|^2/K+1\right)$ from Gorbunov et al. [2022b]. On the other hand, when $L_1 > 0$, we obtain a convergence rate of $\mathcal{O}\left(\|x_0 - x_*\|^{\frac{2+4\alpha-2\alpha^2}{1-\alpha}}/K+1\right)$. Furthermore, as $\alpha \to 0$—which corresponds again to the L-Lipschitz setting—our step sizes γ_k and ω_k become constant, and we recover the standard convergence rate $\mathcal{O}\left(\|x_0 - x_*\|^2/K+1\right)$. This matches the classical result for monotone L-Lipschitz operators up to constants, emphasizing the tightness of our analysis.

3.3 Local Convergence Guarantees for Weak Minty Operators

Beyond the monotone operators, it is also possible to provide convergence for weak Minty operators (9) under some restrictions on $\rho > 0$. In contrast to the monotone problems where we used the same extrapolation and update step γ_k, ω_k, here we use smaller update step size ω_k. Specifically, we employ $\omega_k = \gamma_k/2$, similar to Diakonikolas et al. [2021] for handling weak Minty L-Lipschitz operators.

Theorem 3.8. Suppose F is weak Minty and 1-symmetric (L_0, L_1)-Lipschitz assumption. Moreover we assume

$$\Delta_1 := \frac{\nu}{L_0\left(1 + L_1\|x_0 - x_*\|e^{L_1\|x_0 - x_*\|}\right)} - 4\rho > 0. \tag{16}$$

Then EG with step size $\gamma_k = \frac{\nu}{L_0 + L_1\|F(x_k)\|}$ and $\omega_k = \gamma_k/2$ satisfies

$$\min_{0\le k\le K} \|F(\hat{x}_k)\|^2 \le \frac{4L_0\left(1 + L_1 \exp\left(L_1\|x_0 - x_*\|\right)\|x_0 - x_*\|\right)\|x_0 - x_*\|^2}{\nu\Delta_1(K+1)} \tag{17}$$

where $\nu \exp \nu = 1$.

To the best of our knowledge, this is the first result establishing convergence guarantees for weak Minty, α-symmetric (L_0, L_1)-Lipschitz operators. Similar to the monotone case, we obtain a sublinear convergence rate for weak Minty operators. However, the condition in (16) indicates that the initialization point x_0 must be sufficiently close to the solution x_* in order to ensure convergence. Consequently, Theorem 3.8 only provides a local convergence guarantee.

In the special case where $L_1 = 0$—i.e., the standard L-Lipschitz setting—condition (16) reduces to the simpler requirement $\rho < \nu/4L_0$. Similar assumptions on ρ have been made in prior works such as Diakonikolas et al. [2021] and Pethick et al. [2023] for the L-Lipschitz weak Minty setting. Finally, we extend our analysis to the case $\alpha \in (0, 1)$, and present a corresponding theorem establishing sublinear convergence under analogous restrictions on ρ.

Theorem 3.9. Suppose F is weak Minty and α-symmetric (L_0, L_1)-Lipschitz operator with $\alpha \in (0, 1)$. Moreover we assume

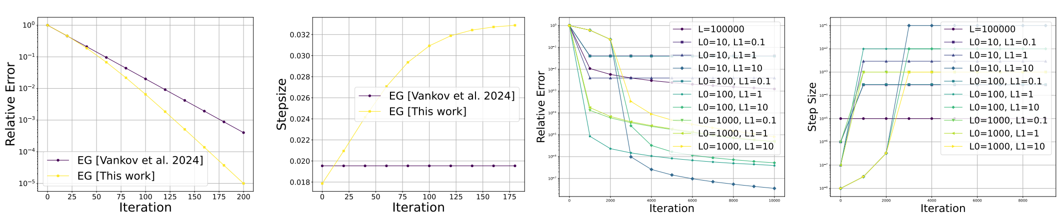
$$\Delta_\alpha := \frac{1}{2\sqrt{2}K_0 + 2\sqrt{2}(K_1 + 2^{-3/2}K_2^{1-\alpha})(K_0 + K_2\|x_0 - x_*\|^{\alpha/1-\alpha})^\alpha\|x_0 - x_*\|^\alpha} - 4\rho > 0. \tag{18}$$

Then EG with step size $\gamma_k = \frac{1}{2\sqrt{2}K_0 + \left(2\sqrt{2}K_1 + 2^{3(1-\alpha)/2}K_2^{1-\alpha}\right)\|F(x_k)\|^\alpha}$ and $\omega_k = \gamma_k/2$ satisfy

$$\min_{0\le k\le K} \|F(\hat{x}_k)\|^2 \le \frac{4\left(K_0 + \left(K_1 + 2^{-3/2}K_2^{1-\alpha}\right)(K_0 + K_2\|x_0 - x_*\|^{\alpha/1-\alpha})^\alpha\|x_0 - x_*\|^\alpha\right)\|x_0 - x_*\|^2}{\Delta_\alpha(K+1)}.$$

4 Numerical Experiments

In this section, we conduct experiments to validate the efficiency of our proposed step size strategy $\gamma_k = \frac{1}{c_0 + c_1\|F(x_k)\|^\alpha}$ with $\alpha = 1$. In the first experiment, we compare our step size choice with that of Vankov et al. [2024] on a strongly monotone problem, and in the second experiment, we make a comparison with the constant step size strategy for solving a monotone problem. Finally, we evaluate our scheme for solving the GlobalForsaken problem from Pethick et al. [2023]. All experiments in this work were conducted using a personal MacBook with an Apple M3 chip and 16GB of RAM.



(a) (b) (c) (d)

Figure 3: In Figures 3a and 3b, we compare our proposed adaptive step size strategy with that of Vankov et al. [2024]. In Figures 3c and 3d, we evaluate the performance of the EG method on the problem in (19), using both a constant step size and the (L_0, L_1)-adaptive step size. For both sets of experiments, we report the relative error and the magnitude of the step size over iterations.

Performance on a Strongly Monotone Problem. In this experiment, we compare our theoretical step sizes with those from Vankov et al. [2024]. Here, we implement EG for solving the operator $F(x) = (\text{sign}(u_1)|u_1| + u_2, \text{sign}(u_2)|u_2| - u_1)$. This problem has constants $L_0 = 1 + 2\sqrt{2}$ and $L_1 = 2\sqrt{2}$. For our method, we use $\gamma_k = \omega_k = \frac{\nu}{L_0 + L_1\|F(x_k)\|}$ while for EG [Vankov et al., 2024] we use stepsize $\gamma_k = \omega_k = \min\left\{\frac{1}{4\mu}, \frac{1}{2\sqrt{2}eL_0}, \frac{1}{2\sqrt{2}eL_1\|F(x_k)\|}\right\}$. In Figure 3a, we plot the relative error $\frac{\|x_k - x_*\|^2}{\|x_0 - x_*\|^2}$ on the y-axis while number of iterations on the x-axis. We find that our proposed step size outperforms that of Vankov et al. [2024]. Moreover, in Figure 3b, we compare the magnitude of the step size and how it evolves over the iterations. We find that the step size of Vankov et al. [2024] remains constant at approximately 0.02, whereas our proposed step size increases to a value larger than 0.032. These experiments highlight the efficiency of our proposed step size.

Performance on a Monotone Problem. Here we consider the following min-max optimization problem

$$\min_{w_1 \in \mathbb{R}^d} \max_{w_2 \in \mathbb{R}^d} \mathcal{L}(w_1, w_2) = \frac{1}{3}\left(w_1^\top \mathbf{A} w_1\right)^{3/2} + w_1^\top \mathbf{B} w_2 - \frac{1}{3}\left(w_2^\top \mathbf{C} w_2\right)^{3/2}. \tag{19}$$

where $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbb{R}^{d \times d}$ are positive definite matrices. Note that, when $d = 1$, and $\mathbf{A}, \mathbf{B}, \mathbf{C}$ are just scalars equal to 1, this problem reduces to (5). The corresponding operator of this problem is given by

$$F(x) = \begin{bmatrix} \nabla_{w_1} \mathcal{L}(w_1, w_2) \\ -\nabla_{w_2} \mathcal{L}(w_1, w_2) \end{bmatrix} = \begin{bmatrix} \left(w_1^\top \mathbf{A} w_1\right)^{1/2} \mathbf{A} w_1 + \mathbf{B} w_2 \\ \left(w_2^\top \mathbf{C} w_2\right)^{1/2} \mathbf{C} w_2 - \mathbf{B}^\top w_1 \end{bmatrix}.$$

Furthermore, we show that $\mathcal{L}$ is convex-concave and has an equilibrium only at $w_1, w_2 = 0 \in \mathbb{R}^d$ (check Appendix E). To solve (19), we implement the EG method using two types of step size strategies: (1) a constant step size $\gamma_k = \omega_k = {}^1/_c$, and (2) an adaptive step size $\gamma_k = \omega_k = {}^1/_{(c_0 + c_1\|F(x_k)\|)}$. For the constant step size EG, we perform a grid search over $c \in \{10^2, 10^3, 10^4, 10^5, 10^6, 10^7\}$. We find that $c = 10^5$ yields the best performance: larger values lead to slower convergence, while smaller values cause divergence. Figures 3c and 3d present the relative error and step size for the case $c = 10^5$. For our adaptive EG method, we perform a grid search over $c_0 \in \{10, 100, 1000\}$ and $c_1 \in \{0.1, 1, 10\}$, evaluating all 9 possible combinations. The performance of all adaptive variants is plotted in Figures 3c and 3d. We observe that most combinations outperform the constant step size EG, with $(c_0, c_1) = (10, 10)$ achieving the best results (see Figure 3c).

Interestingly, while the adaptive EG starts with smaller step sizes compared to the constant step size EG, its step sizes increase over time and eventually surpass those of the constant step size approach (see Figure 3d). This highlights the practical effectiveness of our proposed method in handling non-Lipschitz operators such as the one in (19).

Performance on a Weak Minty Problem. Here we consider the unconstrained GlobalForsaken problem from Pethick et al. [2023] given by

$$\min_{w_1 \in \mathbb{R}} \max_{w_2 \in \mathbb{R}} \mathcal{L}(w_1, w_2) := w_1 w_2 + \psi(w_1) - \psi(w_2), \tag{20}$$

where $\psi(w) = \frac{2w^6}{21} - \frac{w^4}{3} + \frac{w^2}{3}$. As shown in Pethick et al. [2023], the saddle-point problem in (20) admits a global Nash equilibrium at $(w_1, w_2) = (0, 0)$ and satisfies the weak Minty condition (9) with parameter $\rho \approx 0.119732$. We implement AdaptiveEG+, EG+ and EG with our step size strategy to solve this problem. For each algorithm, we perform step size tuning on a a grid of $\{10^{-5}, 10^{-4}, \cdots, 10^2\}$. We observe that both AdaptiveEG+ and EG+ perform best with a fixed step size of $\gamma_k = 0.1$. For our method, we set the step size parameters as $(c_0, c_1) = (1, 1)$. In Figure 4, we present the trajectory plots of these algorithms, all initialized at $(w_1, w_2) = (1, 1)$. Our findings indicate that all algorithms eventually converge to the equilibrium $(0, 0)$, but the convergence of our method is significantly faster. This demonstrates the advantage of our step-size strategy in solving challenging problems that satisfy only weak Minty conditions.

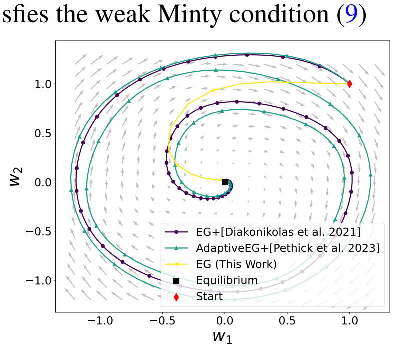


Figure 4: Trajectories of algorithms for solving problem (20).

References

K. Ahn, X. Cheng, M. Song, C. Yun, A. Jadbabaie, and S. Sra. Linear attention is (maybe) all you need (to understand transformer optimization). *arXiv preprint arXiv:2310.01082*, 2023.

D. Balduzzi, S. Racaniere, J. Martens, J. Foerster, K. Tuyls, and T. Graepel. The mechanics of n-player differentiable games. In *International Conference on Machine Learning*, pages 354–363. PMLR, 2018.

A. Beznosikov, P. Dvurechenskii, A. Koloskova, V. Samokhin, S. U. Stich, and A. Gasnikov. Decentralized local stochastic extra-gradient for variational inequalities. *Advances in Neural Information Processing Systems*, 35: 38116–38133, 2022.

N. Brown, A. Bakhtin, A. Lerer, and Q. Gong. Combining deep reinforcement learning and search for imperfect-information games. *Advances in neural information processing systems*, 33:17057–17069, 2020.

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

Z. Chen, Y. Zhou, Y. Liang, and Z. Lu. Generalized-smooth nonconvex optimization is as efficient as smooth nonconvex optimization. In *International Conference on Machine Learning*, pages 5396–5427. PMLR, 2023.

S. Choudhury, E. Gorbunov, and N. Loizou. Single-call stochastic extragradient methods for structured non-monotone variational inequalities: Improved analysis under weaker conditions. *Advances in Neural Information Processing Systems*, 36, 2024.

J. Diakonikolas, C. Daskalakis, and M. I. Jordan. Efficient methods for structured nonconvex-nonconcave min-max optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 2746–2754. PMLR, 2021.

Y. Fan, Y. Li, and B. Chen. Weaker mvi condition: Extragradient methods with multi-step exploration. In *The Twelfth International Conference on Learning Representations*, 2023.

G. Gidel, H. Berard, G. Vignoud, P. Vincent, and S. Lacoste-Julien. A variational inequality perspective on generative adversarial networks. *arXiv preprint arXiv:1802.10551*, 2018.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

E. Gorbunov, H. Berard, G. Gidel, and N. Loizou. Stochastic extragradient: General analysis and improved rates. In *International Conference on Artificial Intelligence and Statistics*, pages 7865–7901. PMLR, 2022a.

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

E. Gorbunov, N. Tupitsa, S. Choudhury, A. Aliev, P. Richtárik, S. Horváth, and M. Takáč. Methods for convex (l_0, l_1)-smooth optimization: Clipping, acceleration, and adaptivity. *arXiv preprint arXiv:2409.14989*, 2024.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

G. Korpelevich. Extragradient method for finding saddle points and other problems. *Matekon*, 13(4):35–49, 1977.

C. J. Li, Y. Yu, N. Loizou, G. Gidel, Y. Ma, N. Le Roux, and M. Jordan. On the convergence of stochastic extragradient for bilinear games using restarted iteration averaging. In *International Conference on Artificial Intelligence and Statistics*, pages 9793–9826. PMLR, 2022.

N. Loizou, S. Vaswani, I. H. Laradji, and S. Lacoste-Julien. Stochastic polyak step-size for sgd: An adaptive learning rate for fast convergence. In *International Conference on Artificial Intelligence and Statistics*, pages 1306–1314. PMLR, 2021.

K. Mishchenko, D. Kovalev, E. Shulgin, P. Richtárik, and Y. Malitsky. Revisiting stochastic extragradient. In *International Conference on Artificial Intelligence and Statistics*, pages 4573–4582. PMLR, 2020.

A. Mokhtari, A. Ozdaglar, and S. Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *International Conference on Artificial Intelligence and Statistics*, pages 1497–1507. PMLR, 2020.

H. Namkoong and J. C. Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. *Advances in neural information processing systems*, 29, 2016.

T. Pethick, P. Latafat, P. Patrinos, O. Fercoq, and V. Cevher. Escaping limit cycles: Global convergence for constrained nonconvex-nonconcave minimax problems. *arXiv preprint arXiv:2302.09831*, 2023.

E. K. Ryu and W. Yin. *Large-scale convex optimization: algorithms & analyses via monotone operators*. Cambridge University Press, 2022.

P. Tseng. On linear convergence of iterative methods for the variational inequality problem. *Journal of Computational and Applied Mathematics*, 60(1-2):237–252, 1995.

D. Vankov, A. Nedich, and L. Sankar. Generalized smooth variational inequalities: Methods with adaptive stepsizes. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 49137–49170. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/vankov24a.html.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

T. Yoon, S. Choudhury, and N. Loizou. Multiplayer federated learning: Reaching equilibrium with less communication. *arXiv preprint arXiv:2501.08263*, 2025.

J. Zhang, T. He, S. Sra, and A. Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. *arXiv preprint arXiv:1905.11881*, 2019.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: see Section 3 and 4.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: see Section 3.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: see Section 1 and 3.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.

- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: see Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: see Section 4.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: see Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: all experiments are deterministic.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: see Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: the paper follows NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: the paper is mostly theoretical and does not have a direct societal impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: we do not release data or models.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: see Section 4.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: the paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: not applicable.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: not applicable.

    Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLMs were not used.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

Supplementary Material

The supplementary material is organized as follows. Section A presents several technical lemmas used in our theoretical analysis. Section B provides illustrative examples of α-symmetric (L_0, L_1)-Lipschitz operators, while Section C contains the detailed proofs of the main convergence theorems presented in the paper. Next, in Section D, we discuss an equivalent formulation of the α-symmetric (L_0, L_1)-Lipschitz condition. Finally, Section E offers additional details on the experimental setup and results.

Contents

A Technical Lemmas

In this section, we present some technical lemmas, which will be used to prove the main results of the work in subsequent sections.

Lemma A.1. For $a, b \in \mathbb{R}^d$, we have

$$2\langle a, b\rangle = \|a\|^2 + \|b\|^2 - \|a - b\|^2. \tag{21}$$

Lemma A.2. For $a, b \in \mathbb{R}^d$, we have

$$-\|a\|^2 \le -\frac{1}{2}\|a + b\|^2 + \|b\|^2. \tag{22}$$

Lemma A.3. [Chen et al., 2023] Operator F is α-symmetric (L_0, L_1)-Lipschitz if and only if

$$\|F(x) - F(y)\| \le \left(L_0 + L_1 \int_0^1 \|F(\theta x + (1-\theta)y)\|^\alpha \, d\theta\right) \|x - y\| \qquad \forall x, y \in \mathbb{R}^d. \tag{23}$$

Lemma A.4. For a 2×2 symmetric matrix, the maximum eigenvalue is given by

$$\lambda_{\max}\left(\begin{bmatrix} a & b \\ b & d \end{bmatrix}\right) = \frac{(a+d) + \sqrt{(a-d)^2 + 4b^2}}{2} \tag{24}$$

where $a, b, d \in \mathbb{R}$.

Proof. Let A be a symmetric 2×2 matrix given by

$$A = \begin{bmatrix} a & b \\ b & d \end{bmatrix}$$

where $a, b, d \in \mathbb{R}$. Since A is symmetric, it has real eigenvalues. The eigenvalues of A are the roots of its characteristic polynomial:

$$\det(A - \lambda I) = \det\left(\begin{bmatrix} a - \lambda & b \\ b & d - \lambda \end{bmatrix}\right) = (a - \lambda)(d - \lambda) - b^2.$$

Expanding the determinant, we obtain the characteristic equation:

$$\lambda^2 - (a + d)\lambda + (ad - b^2) = 0.$$

This is a quadratic equation in λ, and its solutions are:

$$\lambda = \frac{(a+d) \pm \sqrt{(a-d)^2 + 4b^2}}{2}.$$

Thus, the maximum eigenvalue is the larger of the two roots:

$$\lambda_{\max}(A) = \frac{(a+d) + \sqrt{(a-d)^2 + 4b^2}}{2}.$$

This completes the proof. $\square$

Lemma A.5. For the quadratic problem $\min_{w_1} \max_{w_2} \mathcal{L}(w_1, w_2) = \frac{1}{2}w_1^2 + w_1 w_2 - \frac{1}{2}w_2^2$, the Jacobian $\mathbf{J}(x)$ is given by

$$\mathbf{J}(x) = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}.$$

In this case we get $\|\mathbf{J}(x)\| = \sigma_{\max}(\mathbf{J}(x)) = \sqrt{2}$.

Proof. Note that

$$\mathbf{J}(x)^\top \mathbf{J}(x) = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} = 2\mathbf{I}.$$

which has maximum eigenvalue $\sqrt{2}$. Hence, $\|\mathbf{J}(x)\| = \sigma_{\max}(\mathbf{J}(x)) = \lambda_{\max}(\mathbf{J}(x)^\top \mathbf{J}(x)) = \sqrt{2}$. $\square$

B Further Examples of α-Symmetric (L_0, L_1)-Lipschitz Operators

In this section, we provide further examples of operators that satisfy the α-symmetric (L_0, L_1)-Lipschitz assumption. We first start with the details of Example 2 from Section 2.2, then we provide two more examples for the min-max optimization problem and N-player game that satisfy the assumption.

Example 2: Here we consider the operator $F(x) = (u_1^2, u_2^2)$ for $x = (u_1, u_2)$. Then for $y = (v_1, v_2)$, we have

$$\begin{aligned}
\|F(x) - F(y)\| &= \left\|\left(u_1^2 - v_1^2, u_2^2 - v_2^2\right)\right\| \\
&= \left(\left(u_1^2 - v_1^2\right)^2 + \left(u_2^2 - v_2^2\right)^2\right)^{1/2} \\
&= \left((u_1 - v_1)^2 (u_1 + v_1)^2 + (u_2 - v_2)^2 (u_2 + v_2)^2\right)^{1/2} \\
&\leq \left((u_1 - v_1)^4 + (u_2 - v_2)^4\right)^{1/4} \left((u_1 + v_1)^4 + (u_2 + v_2)^4\right)^{1/4} \\
&\leq \left((u_1 - v_1)^2 + (u_2 - v_2)^2\right)^{1/2} \left((u_1 + v_1)^4 + (u_2 + v_2)^4\right)^{1/4} \\
&= \left((u_1 + v_1)^4 + (u_2 + v_2)^4\right)^{1/4} \|x - y\| \\
&= 2\left(\left(\frac{u_1 + v_1}{2}\right)^4 + \left(\frac{u_2 + v_2}{2}\right)^4\right)^{1/4} \|x - y\| \\
&= 2\left\|\left(\frac{u_1 + v_1}{2}\right)^2, \left(\frac{u_2 + v_2}{2}\right)^2\right\|^{1/2} \|x - y\| \\
&= 2\left\|F\left(\frac{x + y}{2}\right)\right\|^{1/2} \|x - y\| \\
&\leq 2 \max_{\theta \in [0,1]} \|F(\theta x + (1 - \theta) y)\|^{1/2} \|x - y\|.
\end{aligned}$$

Here, the first inequality follows from the Cauchy-Schwarz inequality. This completes the proof of $\frac{1}{2}$-symmetric-$(0, 2)$ Lipschitz property of F. Now, we consider the vectors $x = \alpha \mathbf{1}_2$ and $y = \mathbf{1}_2$ where $\mathbf{1}_2 = (1, 1)$. Then we have

$$\begin{aligned}
\frac{\|F(x) - F(y)\|}{\|x - y\|} &= \frac{\sqrt{(\alpha^2 - 1)^2 + (\alpha^2 - 1)^2}}{\sqrt{(\alpha - 1)^2 + (\alpha - 1)^2}} \\
&= \frac{\sqrt{2(\alpha^2 - 1)^2}}{\sqrt{2(\alpha - 1)^2}} \\
&= \sqrt{\frac{(\alpha - 1)^2 (\alpha + 1)^2}{(\alpha - 1)^2}} \\
&= |\alpha + 1|.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\lim_{\alpha \to \infty} \frac{\|F(x) - F(y)\|}{\|x - y\|} &= \lim_{\alpha \to \infty} |\alpha + 1| \\
&= \infty
\end{aligned}$$

Example 3: Min-max optimization problems can be studied using variational inequality formulations. Here we consider one such example, the bilinearly coupled min-max optimization [Chambolle and Pock, 2016] problem

$$\min_{\|w_1\| \leq R} \max_{\|w_2\| \leq R} \mathcal{L}(w_1, w_2) := f(w_1) + w_1^\top \mathbf{B} w_2 - g(w_2)$$

for matrix $\mathbf{B} \in \mathbb{R}^{d \times d}$ and functions $f, g : \mathbb{R}^d \to \mathbb{R}$. The associated operator for this problem is given by $F(x) = H(x) + \mathbf{M}x$ where

$$H(x) = \begin{bmatrix} \nabla f(w_1) \\ \nabla g(w_2) \end{bmatrix} \quad \text{and} \quad \mathbf{M} = \begin{bmatrix} 0 & \mathbf{B} \\ -\mathbf{B}^\top & 0 \end{bmatrix}.$$

If f, g are individually (L_0, L_1)-smooth [Zhang et al., 2019], we show that

$$\|F(x) - F(y)\| \leq \left(2L_0 + (1 + 2L_1 R)\|\mathbf{M}\| + \sqrt{2} L_1 \|F(x)\|\right) \|x - y\|.$$

Thus, F is 1-symmetric $\left(2L_0 + (1 + 2L_1 R)\|\mathbf{M}\|, \sqrt{2} L_1\right)$-Lipschitz. We add the details of this example to Appendix **??**.

Consider the min-max problem

$$\min_{\|w_1\|\le R}\max_{\|w_2\|\le R}\mathcal{L}(w_1,w_2):=f(w_1)+w_1^\top\mathbf{B}w_2-g(w_2).$$

where f,g are (L_0,L_1)-smooth. Then for $x=(w_1,w_2)$ we have

$$F(x)=\begin{pmatrix}\nabla f(w_1)\\ \nabla g(w_2)\end{pmatrix}+\begin{pmatrix}0 & \mathbf{B}\\ -\mathbf{B}^\top & 0\end{pmatrix}\begin{pmatrix}w_1\\ w_2\end{pmatrix}=H(x)+\mathbf{M}x$$

where $\mathbf{M}$ is a matrix and H is an operator. Now note that for $x=(w_1,w_2)$ and $y=(v_1,v_2)$ we get

$$\begin{aligned}
\|H(x)-H(y)\| &= \left\|\begin{bmatrix}\nabla f(w_1)-\nabla f(v_1)\\ \nabla g(w_2)-\nabla g(v_2)\end{bmatrix}\right\|\\
&= \left(\|\nabla f(w_1)-\nabla f(v_1)\|^2+\|g(w_2)-g(v_2)\|^2\right)^{1/2}\\
&\le \left((L_0+L_1\|\nabla f(w_1)\|)^2\|w_1-v_1\|^2+(L_0+L_1\|\nabla g(w_2)\|)^2\|w_2-v_2\|^2\right)^{1/2}\\
&\le \left((L_0+L_1\|\nabla f(w_1)\|)^4+(L_0+L_1\|\nabla g(w_2)\|)^4\right)^{1/4}\left(\|w_1-v_1\|^4+\|w_2-v_2\|^4\right)^{1/4}\\
&\le \left((L_0+L_1\|\nabla f(w_1)\|)^2+(L_0+L_1\|\nabla g(w_2)\|)^2\right)^{1/2}\left(\|w_1-v_1\|^2+\|w_2-v_2\|^2\right)^{1/2}\\
&\le \left(4L_0^2+2L_1^2\left(\|\nabla f(w_1)\|^2+\|\nabla g(w_2)\|^2\right)\right)^{1/2}\left(\|w_1-v_1\|^2+\|w_2-v_2\|^2\right)^{1/2}\\
&= \left(4L_0^2+2L_1^2\|H(x)\|^2\right)^{1/2}\|x-y\|\\
&\le \left(2L_0+\sqrt{2}L_1\|H(x)\|\right)\|x-y\|
\end{aligned}$$

Therefore, using the above inequality, we derive

$$\begin{aligned}
\|F(x)-F(y)\| &\le \|H(x)-H(y)\|+\|\mathbf{M}x-\mathbf{M}y\|\\
&\le \left(2L_0+\sqrt{2}L_1\|H(x)\|\right)\|x-y\|+\|\mathbf{M}\|\|x-y\|\\
&\le \left(2L_0+\|\mathbf{M}\|+\sqrt{2}L_1\|H(x)\|\right)\|x-y\|
\end{aligned}$$

Now using $\|H(x)\|=\|H(x)+\mathbf{M}x-\mathbf{M}x\|\le\|F(x)\|+\|\mathbf{M}\|\|x\|\le\|F(x)\|+\sqrt{2}R\|\mathbf{M}\|$ we get

$$\|F(x)-F(y)\| \le \left(2L_0+(1+2L_1R)\|\mathbf{M}\|+\sqrt{2}L_1\|F(x)\|\right)\|x-y\|.$$

Example 4: In this example, we consider an N-player game [Balduzzi et al., 2018, Loizou et al., 2021, Yoon et al., 2025], where each player $i\in[N]$ selects an action $w_i\in\mathbb{R}^{d_i}$, and the joint action vector of all players is denoted as $x=(w_1,w_2,\cdots,w_N)\in\mathbb{R}^{d_1+\cdots d_N}$. Each player i aims to minimize their loss function f_i for their action w_i. The objective is to find an equilibrium $x_*=(w_{1*},w_{2*},\cdots,w_{N*})$ such that

$$w_{i*}=\text{argmin}_{w_i\in\mathbb{R}^{d_i}}f_i(w_i,w_{-i*}).$$

Here we abuse the notation to denote $w_{-i}=(w_1,\cdots,w_{i-1},w_{i+1},\cdots,w_N)$ and $f_i(w_i,w_{-i})=f_i(w_1,\cdots,w_N)$. When the functions f_i are convex, this equilibrium corresponds to solving $F(x_*)=0$, where the operator F is defined as

$$F(x)=\left(\nabla_1 f_1(x),\nabla_2 f_2(x),\ldots,\nabla_N f_N(x)\right).$$

If each partial gradient $\nabla_i f_i$ satisfies the (L_0,L_1)-smoothness condition [Zhang et al., 2019], i.e., $\|\nabla_i f_i(x)-\nabla_i f_i(y)\|\le(L_0+L_1\|\nabla_i f_i(x)\|)\|x-y\|$, then we can show that the operator F satisfies the 1-symmetric $(\sqrt{2N}L_0,\sqrt{2}L_1)$-Lipschitz condition:

$$\|F(x)-F(y)\|\le\left(\sqrt{2N}L_0+\sqrt{2}L_1\|F(x)\|\right)\|x-y\|.$$

Here we consider the operator

$$F(x) = (\nabla_1 f_1(x),\nabla_2 f_2(x),\cdots\nabla_N f_N(x)).$$

In case each of these partial gradients $\nabla_i f_i$ are (L_0,L_1)-Lipschitz i.e.

$$\|\nabla_i f_i(x)-\nabla_i f_i(y)\| \le (L_0+L_1\|\nabla_i f_i(x)\|)\|x-y\|$$

then we obtain

$$
\begin{aligned}
\|F(x)-F(y)\|^2 &= \sum_{i=1}^{N} \|\nabla_i f_i(x) - \nabla_i f_i(y)\|^2 \\
&\leq \sum_{i=1}^{N} \left(L_0 + L_1 \|\nabla_i f_i(x)\|\right)^2 \|x-y\|^2 \\
&\leq \|x-y\|^2 \sum_{i=1}^{N} (2L_0^2 + 2L_1^2 \|\nabla_i f_i(x)\|^2) \\
&= \|x-y\|^2 (2NL_0^2 + 2L_1^2 \|F(x)\|^2) \\
&\leq \|x-y\|^2 (\sqrt{2N} L_0 + \sqrt{2} L_1 \|F(x)\|)^2.
\end{aligned}
$$

This completes the proof.

C Convergence Analysis

In this section, we present the missing proofs from Section 3. We start with the proof of Proposition 3.1 and then provide the results related to strongly monotone, monotone and weak Minty problems.

C.1 Proof of Proposition 3.1

Proposition C.1. Suppose F is α-symmetric (L_0, L_1)-Lipschitz operator. Then, for $\alpha = 1$

$$\|F(x) - F(y)\| \leq (L_0 + L_1\|F(x)\|) \exp\left(L_1\|x - y\|\right)\|x - y\|,$$

and for $\alpha \in (0, 1)$ we have

$$\|F(x) - F(y)\| \leq \left(K_0 + K_1\|F(x)\|^{\alpha} + K_2\|x - y\|^{\alpha/1-\alpha}\right)\|x - y\|$$

where $K_0 = L_0(2^{\alpha^2/1-\alpha} + 1)$, $K_1 = L_1 \cdot 2^{\alpha^2/1-\alpha}$ and $K_2 = L_1^{1/1-\alpha} \cdot 2^{\alpha^2/1-\alpha} \cdot 3^{\alpha}(1-\alpha)^{\alpha/1-\alpha}$.

Proof. For proving this theorem, we follow the proof technique similar to [Chen et al., 2023]. We start with $\alpha = 1$ case. Let $x, y \in \mathbb{R}^d$ and define $x_\theta := \theta x + (1-\theta)y$. Since F is 1-symmetric (L_0, L_1)-Lipschitz , we have for all $\theta \in [0, 1]$,

$$\|F(x_\theta) - F(y)\| \overset{(23)}{\leq} \left(L_0 + L_1\int_0^1 \|F(x_{\theta\tau})\|d\tau\right)\|x_\theta - y\|.$$

Note that

$$x_{\theta\tau} = \tau x_\theta + (1-\tau)y = \tau(\theta x + (1-\theta)y) + (1-\tau)y = \theta\tau x + (1-\theta\tau)y.$$

Let us define a function

$$H(\theta) := L_0\theta + L_1\int_0^\theta \|F(x_u)\|du.$$

Then, note that $H'(\theta) = L_0 + L_1\|F(x_\theta)\|$. Moreover, we have

$$\begin{aligned}
\|F(x_\theta) - F(y)\| &\leq \left(L_0 + L_1\int_0^1 \|F(x_{\theta\tau})\|d\tau\right)\|x_\theta - y\| \\
&= \left(L_0 + L_1\int_0^1 \|F(x_{\theta\tau})\|d\tau\right)\|\theta x + (1-\theta)y - y\| \\
&= \left(L_0 + L_1\int_0^1 \|F(x_{\theta\tau})\|d\tau\right)\|\theta x - \theta y\| \\
&= \left(L_0\theta + L_1\int_0^1 \|F(x_{\theta\tau})\|\theta d\tau\right)\|x - y\| \\
&= \left(L_0\theta + L_1\int_0^1 \|F(\theta\tau x + (1-\theta\tau)y)\|\theta d\tau\right)\|x - y\| \\
&= \left(L_0\theta + L_1\int_0^\theta \|F(ux + (1-u)y)\|du\right)\|x - y\| \\
&= \left(L_0\theta + L_1\int_0^\theta \|F(x_u)\|du\right)\|x - y\| \\
&= H(\theta)\|x - y\|.
\end{aligned}$$

Therefore we obtain

$$\begin{aligned}
H'(\theta) &= L_0 + L_1\|F(x_\theta)\| \\
&\leq L_0 + L_1\left(\|F(x_\theta) - F(y)\| + \|F(y)\|\right) \\
&\leq L_0 + L_1\left(H(\theta)\|x - y\| + \|F(y)\|\right) \\
&= aH(\theta) + b,
\end{aligned}$$

where $a = L_1\|x - y\|, b = L_0 + L_1\|F(y)\|$. Then we integrate both sides for $\theta \in [0, \theta']$ to get

$$H(\theta') \leq \frac{b}{a}(e^{a\theta'} - 1).$$

Here, we set $\theta' = 1$ to obtain

$$\begin{aligned}
H(1) &\leq \frac{b}{a}(e^a - 1) \\
&= \frac{L_0 + L_1\|F(y)\|}{L_1\|x - y\|}(e^{L_1\|x-y\|} - 1).
\end{aligned}$$

Now, put this back into the original inequality

$$\begin{aligned}\|F(x)-F(y)\| &\leq H(1)\|x-y\| \\ &\leq (L_0+L_1\|F(y)\|)\cdot\frac{e^{L_1\|x-y\|}-1}{L_1} \\ &= \left(\frac{L_0}{L_1}+\|F(y)\|\right)(e^{L_1\|x-y\|}-1).\end{aligned}$$

Finally, using the inequality $e^z-1\leq ze^z$ for $z\geq 0$, we get

$$\begin{aligned}\|F(x)-F(y)\| &\leq \left(\frac{L_0}{L_1}+\|F(y)\|\right)L_1\|x-y\|e^{L_1\|x-y\|} \\ &\leq (L_0+L_1\|F(y)\|)e^{L_1\|x-y\|}\|x-y\|\end{aligned}$$

This completes the proof for $\alpha=1$. The proof for $\alpha\in(0,1)$ follows similarly from [Chen et al., 2023]. $\square$

C.2 Convergence Guarantees for Strongly Monotone Operators

C.2.1 Proof of Theorem 3.2

Theorem C.2. Suppose F is μ-strongly monotone and 1-symmetric (L_0,L_1)-Lipschitz operator. Then Extragradient method with step size $\gamma_k=\frac{\nu}{L_0+L_1\|F(x_k)\|}$ satisfy

$$\|x_{k+1}-x_*\|^2 \leq \left(1-\frac{\nu\mu}{L_0\left(1+L_1\exp\left(L_1\|x_0-x_*\|\right)\|x_0-x_*\|\right)}\right)^{k+1}\|x_0-x_*\|^2$$

where $\nu>0$ root of $1-2\nu-\nu^2\exp 2\nu=0$.

Proof. From the update step of the Extragradient method, we obtain

$$\begin{aligned}\|x_{k+1}-x_*\|^2 &= \|x_k-\gamma_kF(\hat{x}_k)-x_*\|^2 \\ &= \|x_k-x_*\|^2-2\gamma_k\langle F(\hat{x}_k),x_k-x_*\rangle+\gamma_k^2\|F(\hat{x}_k)\|^2 \\ &= \|x_k-x_*\|^2-2\gamma_k\langle F(\hat{x}_k),\hat{x}_k-x_*\rangle-2\gamma_k\langle F(\hat{x}_k),x_k-\hat{x}_k\rangle+\gamma_k^2\|F(\hat{x}_k)\|^2 \\ &\overset{(7)}{\leq} \|x_k-x_*\|^2-2\gamma_k\mu\|\hat{x}_k-x_*\|^2-2\gamma_k\langle F(\hat{x}_k),x_k-\hat{x}_k\rangle+\gamma_k^2\|F(\hat{x}_k)\|^2 \\ &\overset{(22)}{\leq} \|x_k-x_*\|^2-\gamma_k\mu\|x_k-x_*\|^2+2\gamma_k\mu\|x_k-\hat{x}_k\|^2-2\gamma_k\langle F(\hat{x}_k),x_k-\hat{x}_k\rangle \\ &\quad +\gamma_k^2\|F(\hat{x}_k)\|^2 \\ &= (1-\gamma_k\mu)\|x_k-x_*\|^2+2\gamma_k\mu\|x_k-\hat{x}_k\|^2-2\gamma_k\langle F(\hat{x}_k),x_k-\hat{x}_k\rangle+\gamma_k^2\|F(\hat{x}_k)\|^2 \\ &= (1-\gamma_k\mu)\|x_k-x_*\|^2+2\gamma_k^3\mu\|F(x_k)\|^2-2\gamma_k^2\langle F(\hat{x}_k),F(x_k)\rangle+\gamma_k^2\|F(\hat{x}_k)\|^2 \\ &\overset{(21)}{=} (1-\gamma_k\mu)\|x_k-x_*\|^2+2\gamma_k^3\mu\|F(x_k)\|^2 \\ &\quad -\gamma_k^2\left(\|F(\hat{x}_k)\|^2+\|F(x_k)\|^2-\|F(x_k)-F(\hat{x}_k)\|^2\right)+\gamma_k^2\|F(\hat{x}_k)\|^2 \\ &= (1-\gamma_k\mu)\|x_k-x_*\|^2-\gamma_k^2(1-2\gamma_k\mu)\|F(x_k)\|^2+\gamma_k^2\|F(x_k)-F(\hat{x}_k)\|^2 \\ &\overset{(12)}{\leq} (1-\gamma_k\mu)\|x_k-x_*\|^2-\gamma_k^2(1-2\gamma_k\mu)\|F(x_k)\|^2 \\ &\quad +\gamma_k^2(L_0+L_1\|F(x_k)\|)^2\exp\left(2L_1\|x_k-\hat{x}_k\|\right)\|x_k-\hat{x}_k\|^2 \\ &= (1-\gamma_k\mu)\|x_k-x_*\|^2 \\ &\quad -\gamma_k^2\left(1-2\gamma_k\mu-\gamma_k^2(L_0+L_1\|F(x_k)\|)^2\exp\left(2\gamma_kL_1\|F(x_k)\|\right)\right)\|F(x_k)\|^2. \end{aligned}\tag{25}$$

Now we set $\gamma_k=\frac{\nu}{L_0+L_1\|F(x_k)\|}$. Then we want to choose ν such that

$$1-\frac{2\nu\mu}{L_0+L_1\|F(x_k)\|}-\nu^2\exp 2\nu\geq 0.$$

Note that $\mu\leq L_0+L_1\|F(x_k)\|$ for any x_k. Therefore, we get $1-\frac{2\nu\mu}{L_0+L_1\|F(x_k)\|}-\nu^2\exp 2\nu\geq 1-2\nu-\nu^2\exp 2\nu$ and it is enough to choose ν such that

$$1-2\nu-\nu^2\exp 2\nu\geq 0.$$

This inequality holds for any $\nu\leq 0.22$. Hence, for this choice of ν we get the following inequality from (25).

$$\|x_{k+1}-x_*\|^2 \leq (1-\gamma_k\mu)\|x_k-x_*\|^2. \tag{26}$$

This proves that the distance of the iterates x_k from x_* are bounded by $\|x_0 - x_*\|$. Now note that using (12) with $x = x_k, y = x_*$ with $\|x_k - x_*\| \leq \|x_0 - x_*\|$ we get

$$\begin{aligned} \|F(x_k)\| &\overset{(12)}{\leq} L_0 \exp\left(L_1\|x_k - x_*\|\right)\|x_k - x_*\| \\ &\overset{(26)}{\leq} L_0 \exp\left(L_1\|x_0 - x_*\|\right)\|x_0 - x_*\|. \end{aligned} \tag{27}$$

Therefore, we have the following lower bound on the step-size

$$\begin{aligned} \gamma_k &= \frac{\nu}{L_0 + L_1\|F(x_k)\|} \\ &\overset{(27)}{\geq} \frac{\nu}{L_0\left(1 + L_1\exp\left(L_1\|x_0 - x_*\|\right)\|x_0 - x_*\|\right)}. \end{aligned} \tag{28}$$

Hence, we get

$$\begin{aligned} \|x_{k+1} - x_*\|^2 &\overset{(26),(28)}{\leq} \left(1 - \frac{\nu\mu}{L_0\left(1 + L_1\exp\left(L_1\|x_0 - x_*\|\right)\|x_0 - x_*\|\right)}\right)\|x_k - x_*\|^2 \\ &\leq \left(1 - \frac{\nu\mu}{L_0\left(1 + L_1\exp\left(L_1\|x_0 - x_*\|\right)\|x_0 - x_*\|\right)}\right)^{k+1}\|x_0 - x_*\|^2. \end{aligned}$$

□

C.2.2 Proof of Corollary 3.3

Corollary C.3. Suppose F is μ-strongly monotone and 1-symmetric (L_0, L_1)-Lipschitz operator. Then Extragradient with step size $\gamma_k = \frac{\nu}{L_0 + L_1\|F(x_k)\|}$ satisfy $\|x_{K+1} - x_*\|^2 \leq \varepsilon$ after

$$K = \frac{2L_0}{\nu\mu}\log\left(\frac{\|x_0 - x_*\|^2}{\varepsilon}\right) + \frac{1}{\gamma\mu}\log\left(\frac{2L_1\|x_0 - x_*\|^2}{\gamma^2 L_0}\right)$$

many iterations, where $\nu > 0$ satisfy $1 - 4\nu - 2\nu^2\exp 2\nu = 0$ and

$$\gamma := \frac{\nu}{L_0\left(1 + L_1\exp\left(L_1\|x_0 - x_*\|\right)\|x_0 - x_*\|\right)}.$$

Proof. We set $\gamma_k = \frac{\nu}{L_0 + L_1\|F(x_k)\|}$ and we choose $\nu \in (0, 1)$ such that $1 - 4\nu - 2\nu^2\exp 2\nu = 0$. Then we have

$$\begin{aligned} 1 - 2\gamma_k\mu - \gamma_k^2(L_0 + L_1\|F(x_k)\|)^2\exp\left(2\gamma_k L_1\|F(x_k)\|\right) &\geq 1 - \frac{2\nu\mu}{L_0 + L_1\|F(x_k)\|} - \nu^2\exp 2\nu \\ &\geq 1 - 2\nu - \nu^2\exp 2\nu \\ &= \frac{1}{2}. \end{aligned} \tag{29}$$

Therefore, from (25) and (29) we get

$$\|x_{k+1} - x_*\|^2 \leq (1 - \gamma_k\mu)\|x_k - x_*\|^2 - \frac{\gamma_k^2}{2}\|F(x_k)\|^2. \tag{30}$$

In (28), we found that a lower bound of γ_k is

$$\gamma_k \geq \gamma := \frac{\nu}{L_0\left(1 + L_1\exp\left(L_1\|x_0 - x_*\|\right)\|x_0 - x_*\|\right)}.$$

Then from (30) we get

$$\frac{\gamma^2}{2}\|F(x_k)\|^2 \leq (1 - \gamma\mu)\|x_k - x_*\|^2 \leq (1 - \gamma\mu)^{k+1}\|x_0 - x_*\|^2 \tag{31}$$

This can be rearranged to write

$$\|F(x_k)\|^2 \leq \frac{2(1 - \gamma\mu)^{k+1}}{\gamma^2}\|x_0 - x_*\|^2.$$

This implies $\|F(x_k)\|^2 \leq \frac{L_0}{L_1}$ after

$$K' = \frac{1}{\gamma\mu}\log\left(\frac{2L_1\|x_0 - x_*\|^2}{\gamma^2 L_0}\right) \tag{32}$$

many iterations. Hence for $k \geq K'$ we have

$$\begin{aligned} \gamma_k &= \frac{\nu}{L_0 + L_1\|F(x_k)\|} \\ &\geq \frac{\nu}{2L_0}. \end{aligned}$$

In the last inequality we used $\|F(x_k)\|^2 \leq \frac{L_0}{L_1}$ for $k \geq K'$. Therefore for $k \geq K'$ we obtain

$$\begin{aligned} \|x_{k+1} - x_*\|^2 &\leq \left(1 - \frac{\nu\mu}{2L_0}\right)\|x_k - x_*\|^2 \\ &\leq \left(1 - \frac{\nu\mu}{2L_0}\right)^{k+1-K'} \|x_{K'} - x_*\|^2 \\ &\leq \left(1 - \frac{\nu\mu}{2L_0}\right)^{k+1-K'} \|x_0 - x_*\|^2 \end{aligned}$$

Thus we conclude that, $\|x_{K+1} - x_*\|^2 \leq \varepsilon$ after atmost

$$K = \frac{2L_0}{\nu\mu}\log\left(\frac{\|x_0 - x_*\|^2}{\varepsilon}\right) + K'$$

many iterations. □

C.2.3 Proof of Theorem 3.4

Theorem C.4. Suppose F is μ-strongly monotone and α-symmetric (L_0, L_1)-Lipschitz operator with $\alpha \in (0,1)$. Then Extragradient method with $\gamma_k = \frac{\nu}{2K_0 + \left(2K_1 + 2^{1-\alpha}K_2^{1-\alpha}\right)\|F(x_k)\|^\alpha}$ satisfy

$$\|x_{k+1} - x_*\|^2 \leq \left(1 - \frac{\nu\mu}{2K_0 + (2K_1 + 2^{1-\alpha}K_2^{1-\alpha})(K_0 + K_2\|x_0 - x_*\|^{\alpha/1-\alpha})^\alpha \|x_0 - x_*\|^\alpha}\right)^{k+1} \|x_0 - x_*\|^2.$$

where $\nu \in (0,1)$ is a constant such that $1 - \nu - \nu^2 \geq 0$.

Proof. Using the update steps of the Extragradient method, we have

$$\begin{aligned} \|x_{k+1} - x_*\|^2 &= \|x_k - \gamma_k F(\hat{x}_k) - x_*\|^2 \\ &= \|x_k - x_*\|^2 - 2\gamma_k \langle F(\hat{x}_k), x_k - x_*\rangle + \gamma_k^2\|F(\hat{x}_k)\|^2 \\ &= \|x_k - x_*\|^2 - 2\gamma_k \langle F(\hat{x}_k), \hat{x}_k - x_*\rangle - 2\gamma_k \langle F(\hat{x}_k), x_k - \hat{x}_k\rangle + \gamma_k^2\|F(\hat{x}_k)\|^2 \\ &\overset{(7)}{\leq} \|x_k - x_*\|^2 - 2\gamma_k\mu\|\hat{x}_k - x_*\|^2 - 2\gamma_k \langle F(\hat{x}_k), x_k - \hat{x}_k\rangle + \gamma_k^2\|F(\hat{x}_k)\|^2 \\ &\overset{(22)}{\leq} \|x_k - x_*\|^2 - \gamma_k\mu\|x_k - x_*\|^2 + 2\gamma_k\mu\|x_k - \hat{x}_k\|^2 - 2\gamma_k \langle F(\hat{x}_k), x_k - \hat{x}_k\rangle + \gamma_k^2\|F(\hat{x}_k)\|^2 \\ &= (1 - \gamma_k\mu)\|x_k - x_*\|^2 + 2\gamma_k\mu\|x_k - \hat{x}_k\|^2 - 2\gamma_k \langle F(\hat{x}_k), x_k - \hat{x}_k\rangle + \gamma_k^2\|F(\hat{x}_k)\|^2 \\ &= (1 - \gamma_k\mu)\|x_k - x_*\|^2 + 2\gamma_k^3\mu\|F(x_k)\|^2 - 2\gamma_k^2 \langle F(\hat{x}_k), F(x_k)\rangle + \gamma_k^2\|F(\hat{x}_k)\|^2 \\ &\overset{(21)}{=} (1 - \gamma_k\mu)\|x_k - x_*\|^2 + 2\gamma_k^3\mu\|F(x_k)\|^2 - \gamma_k^2\left(\|F(\hat{x}_k)\|^2 + \|F(x_k)\|^2 - \|F(x_k) - F(\hat{x}_k)\|^2\right) + \gamma_k^2\|F(\hat{x}_k)\|^2 \\ &= (1 - \gamma_k\mu)\|x_k - x_*\|^2 - \gamma_k^2(1 - 2\gamma_k\mu)\|F(x_k)\|^2 - \gamma_k^2\|F(x_k) - F(\hat{x}_k)\|^2 \\ &\overset{(13)}{\leq} (1 - \gamma_k\mu)\|x_k - x_*\|^2 - \gamma_k^2(1 - 2\gamma_k\mu)\|F(x_k)\|^2 \\ &\quad -\gamma_k^2\left(K_0 + K_1\|F(x_k)\|^\alpha + K_2\|x_k - \hat{x}_k\|^{\alpha/1-\alpha}\right)^2\|x_k - \hat{x}_k\|^2 \\ &= (1 - \gamma_k\mu)\|x_k - x_*\|^2 - \gamma_k^2\left(1 - 2\gamma_k\mu - \gamma_k^2\left(K_0 + K_1\|F(x_k)\|^\alpha + \gamma_k^{\alpha/1-\alpha}K_2\|F(x_k)\|^{\alpha/1-\alpha}\right)^2\right)\|F(x_k)\|^2. \end{aligned}$$

Here we will choose $\gamma_k > 0$ such that

$$1 - 2\gamma_k\mu - \gamma_k^2\left(K_0 + K_1\|F(x_k)\|^\alpha + \gamma_k^{\alpha/1-\alpha}K_2\|F(x_k)\|^{\alpha/1-\alpha}\right)^2 \geq 0$$

Let us choose $\gamma_k = \frac{\nu}{2K_0 + \left(2K_1 + 2^{1-\alpha}K_2^{1-\alpha}\right)\|F(x_k)\|^\alpha}$ for some $\nu \in (0,1)$. Then we observe that

$$\begin{aligned}
\gamma_k\left(K_0 + K_1\|F(x_k)\|^\alpha\right) + \gamma_k^{1/1-\alpha}K_2\|F(x_k)\|^{\alpha/1-\alpha} &\leq \frac{\nu\left(K_0 + K_1\|F(x_k)\|^\alpha\right)}{2K_0 + \left(2K_1 + 2^{1-\alpha}K_2^{1-\alpha}\right)\|F(x_k)\|^\alpha} \\
&\quad + \frac{\nu^{1/1-\alpha}K_2\|F(x_k)\|^{\alpha/1-\alpha}}{\left(2K_0 + \left(2K_1 + 2^{1-\alpha}K_2^{1-\alpha}\right)\|F(x_k)\|^\alpha\right)^{1/1-\alpha}} \\
&\leq \frac{\nu}{2} + \frac{\nu^{1/1-\alpha}}{2} \\
&\leq \nu.
\end{aligned}$$

The last inequality follows from $\nu \in (0,1)$. Therefore it is enough to choose $\nu \in (0,1)$ such that

$$1 - \frac{2\nu\mu}{2K_0 + \left(2K_1 + 2^{1-\alpha}K_2^{1-\alpha}\right)\|F(x_k)\|^\alpha} - \nu^2 \geq 0$$

However, note that, we always have $\mu \leq K_0$, thus it is enough to choose $\nu \in (0,1)$ such that

$$1 - \nu - \nu^2 \geq 0.$$

Hence, for this choice of ν we get

$$\|x_{k+1} - x_*\|^2 \quad \leq \quad (1 - \gamma_k\mu)\|x_0 - x_*\|^2.$$

Here we lower bound the step size γ_k with

$$\gamma_k \geq \frac{\nu}{2K_0 + (2K_1 + 2^{1-\alpha}K_2^{1-\alpha})(K_0 + K_2\|x_0 - x_*\|^{\alpha/1-\alpha})^\alpha\|x_0 - x_*\|^\alpha}.$$

Hence we obtain

$$\begin{aligned}
\|x_{k+1} - x_*\|^2 &\leq \left(1 - \frac{\nu\mu}{2K_0 + (2K_1 + 2^{1-\alpha}K_2^{1-\alpha})(K_0 + K_2\|x_0 - x_*\|^{\alpha/1-\alpha})^\alpha\|x_0 - x_*\|^\alpha}\right)\|x_k - x_*\|^2 \\
&\leq \left(1 - \frac{\nu\mu}{2K_0 + (2K_1 + 2^{1-\alpha}K_2^{1-\alpha})(K_0 + K_2\|x_0 - x_*\|^{\alpha/1-\alpha})^\alpha\|x_0 - x_*\|^\alpha}\right)^{k+1}\|x_0 - x_*\|^2.
\end{aligned}$$

This completes the proof of the theorem. $\square$

C.3 Convergence Guarantees for Monotone Operators

C.3.1 Proof of Theorem 3.5

Theorem C.5. Suppose F is monotone and 1-symmetric (L_0, L_1)-Lipschitz operator. Then EG with step size $\gamma_k = \omega_k = \frac{\nu}{L_0 + L_1\|F(x_k)\|}$ satisfy

$$\min_{0 \le k \le K} \|F(x_k)\|^2 \le \frac{2L_0^2\left(1 + L_1 \exp\left(L_1\|x_0 - x_*\|\right)\|x_0 - x_*\|\right)^2 \|x_0 - x_*\|^2}{\nu^2(K+1)}.$$

where $\nu \exp \nu = 1/\sqrt{2}$.

Proof. From the update rule of the Extragradient method, we have

$$\begin{aligned}
\|x_{k+1} - x_*\|^2 &= \|x_k - \gamma_k F(\hat{x}_k) - x_*\|^2 \\
&= \|x_k - x_*\|^2 - 2\gamma_k \langle F(\hat{x}_k), x_k - x_*\rangle + \gamma_k^2 \|F(\hat{x}_k)\|^2 \\
&= \|x_k - x_*\|^2 - 2\gamma_k \langle F(\hat{x}_k), \hat{x}_k - x_*\rangle - 2\gamma_k \langle F(\hat{x}_k), x_k - \hat{x}_k\rangle + \gamma_k^2 \|F(\hat{x}_k)\|^2 \\
&\overset{(7)}{\le} \|x_k - x_*\|^2 - 2\gamma_k \langle F(\hat{x}_k), x_k - \hat{x}_k\rangle + \gamma_k^2 \|F(\hat{x}_k)\|^2 \\
&= \|x_k - x_*\|^2 - 2\gamma_k^2 \langle F(\hat{x}_k), F(x_k)\rangle + \gamma_k^2 \|F(\hat{x}_k)\|^2 \\
&\overset{(21)}{=} \|x_k - x_*\|^2 - \gamma_k^2 \|F(\hat{x}_k)\|^2 - \gamma_k^2 \|F(x_k)\|^2 + \gamma_k^2 \|F(x_k) - F(\hat{x}_k)\|^2 + \gamma_k^2 \|F(\hat{x}_k)\|^2 \\
&= \|x_k - x_*\|^2 - \gamma_k^2 \|F(x_k)\|^2 + \gamma_k^2 \|F(x_k) - F(\hat{x}_k)\|^2 \\
&\overset{(12)}{\le} \|x_k - x_*\|^2 - \gamma_k^2 \|F(x_k)\|^2 + \gamma_k^2 (L_0 + L_1\|F(x_k)\|)^2 \exp\left(2L_1\|x_k - \hat{x}_k\|\right) \|x_k - \hat{x}_k\|^2 \\
&= \|x_k - x_*\|^2 - \gamma_k^2 \left(1 - \gamma_k^2 (L_0 + L_1\|F(x_k)\|)^2 \exp\left(2\gamma_k L_1 \|F(x_k)\|\right)\right) \|F(x_k)\|^2 \\
&\le \|x_k - x_*\|^2 - \gamma_k^2 \left(1 - \gamma_k^2 (L_0 + L_1\|F(x_k)\|)^2 \exp\left(2\gamma_k (L_0 + L_1\|F(x_k)\|)\right)\right) \|F(x_k)\|^2.
\end{aligned}$$

Here we use $\gamma_k = \frac{\nu}{L_0 + L_1\|F(x_k)\|}$ for some $\nu \in (0,1)$ to get

$$\|x_{k+1} - x_*\|^2 \le \|x_k - x_*\|^2 - \gamma_k^2 \left(1 - \nu^2 \exp(2\nu)\right) \|F(x_k)\|^2$$

Then we choose ν such that $\nu \exp \nu = 1/\sqrt{2}$ to obtain

$$\|x_{k+1} - x_*\|^2 \le \|x_k - x_*\|^2 - \frac{\gamma_k^2}{2} \|F(x_k)\|^2. \tag{33}$$

In particular the distance of the iterates x_k from x_* are bounded i.e. $\|x_{k+1} - x_*\| \le \|x_k - x_*\| \le \|x_0 - x_*\|$. Therefore, using (12) with $y = x_*$ and $x = x_k$, we get

$$\begin{aligned}
\|F(x_k)\| &\le L_0 \exp\left(L_1\|x_k - x_*\|\right) \|x_k - x_*\| \\
&\le L_0 \exp\left(L_1\|x_0 - x_*\|\right) \|x_0 - x_*\|.
\end{aligned}$$

Then we have the lower bound on step-size given as follows

$$\gamma_k = \frac{\nu}{L_0 + L_1\|F(x_k)\|} \ge \frac{\nu}{L_0\left(1 + L_1 \exp\left(L_1\|x_0 - x_*\|\right)\|x_0 - x_*\|\right)}. \tag{34}$$

Rearranging (33) we have

$$\frac{\gamma_k^2}{2} \|F(x_k)\|^2 \le \|x_k - x_*\|^2 - \|x_{k+1} - x_*\|^2.$$

Summing up the above inequality for $k = 0, 1, \cdots, K$ and dividing by $K+1$ we get

$$\frac{1}{K+1} \sum_{k=0}^{K} \frac{\gamma_k^2}{2} \|F(x_k)\|^2 \le \frac{\|x_0 - x_*\|^2 - \|x_{K+1} - x_*\|^2}{K+1} \le \frac{\|x_0 - x_*\|^2}{K+1}. \tag{35}$$

Here we will use the lower bound on step-size γ_k given in (34) to get

$$\frac{1}{K+1} \sum_{k=0}^{K} \|F(x_k)\|^2 \le \frac{2L_0^2\left(1 + L_1 \exp\left(L_1\|x_0 - x_*\|\right)\|x_0 - x_*\|\right)^2 \|x_0 - x_*\|^2}{\nu^2(K+1)}.$$

Finally, note that $\min_{0 \le k \le K} \|F(x_k)\|^2 \le \frac{1}{K+1} \sum_{k=0}^{K} \|F(x_k)\|^2$. Therefore we have

$$\min_{0 \le k \le K} \|F(x_k)\|^2 \le \frac{2L_0^2\left(1 + L_1 \exp\left(L_1\|x_0 - x_*\|\right)\|x_0 - x_*\|\right)^2 \|x_0 - x_*\|^2}{\nu^2(K+1)}.$$

This completes the proof of the Theorem. □

C.3.2 Proof of Theorem 3.6

Theorem C.6. Suppose F is monotone and 1-symmetric (L_0, L_1)-Lipschitz operator. Then Extragradient method with step size $\gamma_k = \frac{\nu}{L_0 + L_1\|F(x_k)\|}$ satisfy

$$\min_{0\le k\le K} \|F(x_k)\| \quad \le \quad \frac{\sqrt{2}L_0\|x_0 - x_*\|}{\nu\sqrt{K+1} - \sqrt{2}L_1\|x_0 - x_*\|}$$

where $\nu \exp \nu = {}^1\!/\!\sqrt{2}$ and $K + 1 \ge \frac{2L_1^2\|x_0 - x_*\|^2}{\nu^2}$.

Proof. From (35), we know steps of Extragradient method satisfy

$$\frac{1}{K+1}\sum_{k=0}^{K} \frac{\gamma_k^2}{2}\|F(x_k)\|^2 \quad \le \quad \frac{\|x_0 - x_*\|^2}{K+1}.$$

Taking the minimum on the left-hand side we have

$$\min_{0\le k\le K} \gamma_k^2\|F(x_k)\|^2 \quad \le \quad \frac{2\|x_0 - x_*\|^2}{K+1},$$

or equivalently,

$$\min_{0\le k\le K} \frac{\nu^2\|F(x_k)\|^2}{(L_0 + L_1\|F(x_k)\|)^2} \quad \le \quad \frac{2\|x_0 - x_*\|^2}{K+1}.$$

Taking the square root on both sides we have

$$\min_{0\le k\le K} \frac{\nu\|F(x_k)\|}{L_0 + L_1\|F(x_k)\|} \quad \le \quad \frac{\sqrt{2}\|x_0 - x_*\|}{\sqrt{K+1}}.$$

Therefore, for some $0 \le k_0 \le K$ we have

$$\frac{\|F(x_{k_0})\|}{L_0 + L_1\|F(x_{k_0})\|} \quad \le \quad \frac{\sqrt{2}\|x_0 - x_*\|}{\nu\sqrt{K+1}}.$$

Therefore, rearranging these terms, we get

$$\left(\nu\sqrt{K+1} - \sqrt{2}L_1\|x_0 - x_*\|\right)\|F(x_{k_0})\| \quad \le \quad \sqrt{2}L_0\|x_0 - x_*\|.$$

When we have $K + 1 \ge \frac{2L_1^2\|x_0 - x_*\|^2}{\nu^2}$ then we can rearrange the terms to obtain

$$\|F(x_{k_0})\| \quad \le \quad \frac{\sqrt{2}L_0\|x_0 - x_*\|}{\left(\nu\sqrt{K+1} - \sqrt{2}L_1\|x_0 - x_*\|\right)}.$$

for some $0 \le k_0 \le K$. Hence, we complete the proof of the theorem

$$\min_{0\le k\le K} \|F(x_k)\| \quad \le \quad \frac{\sqrt{2}L_0\|x_0 - x_*\|}{\nu\sqrt{K+1} - \sqrt{2}L_1\|x_0 - x_*\|}.$$

□

C.3.3 Proof of Theorem 3.7

Theorem C.7. Suppose F is monotone and α-symmetric (L_0, L_1)-Lipschitz operator with $\alpha \in (0, 1)$. Then Extragradient method with step size $\gamma_k = \frac{1}{2\sqrt{2}K_0 + \left(2\sqrt{2}K_1 + 2^{3(1-\alpha)/2}K_2^{1-\alpha}\right)\|F(x_k)\|^\alpha}$ satisfy

$$\min_{0\le k\le K} \|F(x_k)\|^2 \le \frac{16\left(K_0 + \left(K_1 + 2^{-3/2}K_2^{1-\alpha}\right)(K_0 + K_2\|x_0 - x_*\|^{\alpha/1-\alpha})^\alpha\|x_0 - x_*\|^\alpha\right)^2\|x_0 - x_*\|^2}{K+1}.$$

Proof. Here operator F is α-symmetric (L_0, L_1)-Lipschitz i.e. it satisfies (13). For the update steps of the Extragradient method, we have

$$\begin{aligned}
\|x_{k+1} - x_*\|^2 &= \|x_k - \gamma_k F(\hat{x}_k) - x_*\|^2 \\
&= \|x_k - x_*\|^2 - 2\gamma_k \langle F(\hat{x}_k), x_k - x_* \rangle + \gamma_k^2 \|F(\hat{x}_k)\|^2 \\
&= \|x_k - x_*\|^2 - 2\gamma_k \langle F(\hat{x}_k), \hat{x}_k - x_* \rangle - 2\gamma_k \langle F(\hat{x}_k), x_k - \hat{x}_k \rangle + \gamma_k^2 \|F(\hat{x}_k)\|^2 \\
&\overset{(7)}{\leq} \|x_k - x_*\|^2 - 2\gamma_k \langle F(\hat{x}_k), x_k - \hat{x}_k \rangle + \gamma_k^2 \|F(\hat{x}_k)\|^2 \\
&= \|x_k - x_*\|^2 - 2\gamma_k^2 \langle F(\hat{x}_k), F(x_k) \rangle + \gamma_k^2 \|F(\hat{x}_k)\|^2 \\
&\overset{(21)}{=} \|x_k - x_*\|^2 - \gamma_k^2 \|F(\hat{x}_k)\|^2 - \gamma_k^2 \|F(x_k)\|^2 + \gamma_k^2 \|F(x_k) - F(\hat{x}_k)\|^2 + \gamma_k^2 \|F(\hat{x}_k)\|^2 \\
&= \|x_k - x_*\|^2 - \gamma_k^2 \|F(x_k)\|^2 + \gamma_k^2 \|F(x_k) - F(\hat{x}_k)\|^2 \\
&\overset{(13)}{\leq} \|x_k - x_*\|^2 - \gamma_k^2 \|F(x_k)\|^2 + \gamma_k^2 \left(K_0 + K_1 \|F(x_k)\|^\alpha + K_2 \|x_k - \hat{x}_k\|^{\alpha/1-\alpha}\right)^2 \|x_k - \hat{x}_k\|^2 \\
&= \|x_k - x_*\|^2 - \gamma_k^2 \left(1 - \gamma_k^2 \left(K_0 + K_1 \|F(x_k)\|^\alpha + \gamma_k^{\alpha/1-\alpha} K_2 \|F(x_k)\|^{\alpha/1-\alpha}\right)^2\right) \|F(x_k)\|^2 \qquad (36)
\end{aligned}$$

Here we want to choose γ_k such that

$$\gamma_k \left(K_0 + K_1 \|F(x_k)\|^\alpha\right) + \gamma_k^{1/1-\alpha} K_2 \|F(x_k)\|^{\alpha/1-\alpha} \quad \leq \quad \frac{1}{\sqrt{2}}.$$

For this, it is enough to make sure

$$\gamma_k \left(K_0 + K_1 \|F(x_k)\|^\alpha\right) \leq \frac{1}{2\sqrt{2}} \quad \text{and} \quad \gamma_k^{1/1-\alpha} K_2 \|F(x_k)\|^{\alpha/1-\alpha} \leq \frac{1}{2\sqrt{2}}.$$

Therefore, we choose $\gamma_k = \frac{1}{2\sqrt{2}(K_0 + K_1\|F(x_k)\|^\alpha) + 2^{3(1-\alpha)/2} K_2^{1-\alpha} \|F(x_k)\|^\alpha}$ and we get the following from (36)

$$\|x_{k+1} - x_*\|^2 \quad \leq \quad \|x_k - x_*\|^2 - \frac{\gamma_k^2}{2} \|F(x_k)\|^2. \qquad (37)$$

Rearranging this inequality, we have

$$\frac{\gamma_k^2}{2} \|F(x_k)\|^2 \quad \leq \quad \|x_k - x_*\|^2 - \|x_{k+1} - x_*\|^2.$$

Then we sum up this inequality for $k = 0, 1, \cdots K$ to get

$$\frac{1}{K+1} \sum_{k=0}^{K} \gamma_k^2 \|F(x_k)\|^2 \quad \leq \quad \frac{2\|x_0 - x_*\|^2}{K+1}. \qquad (38)$$

For this step size, we also have $\|x_k - x_0\|^2 \leq \|x_0 - x_*\|^2$ from (41). Now note that from (13) we obtain the following bound with $x = x_k$ and $y = x_*$

$$\begin{aligned}
\|F(x_k)\|^\alpha &\leq (K_0 + K_2 \|x_k - x_*\|^{\alpha/1-\alpha})^\alpha \|x_k - x_*\|^\alpha \\
&\overset{(41)}{\leq} (K_0 + K_2 \|x_0 - x_*\|^{\alpha/1-\alpha})^\alpha \|x_0 - x_*\|^\alpha.
\end{aligned}$$

We use this to lower bound the step size γ_k as follows

$$\begin{aligned}
\gamma_k &= \frac{1}{2\sqrt{2}(K_0 + K_1 \|F(x_k)\|^\alpha) + 2^{3(1-\alpha)/2} K_2^{1-\alpha} \|F(x_k)\|^\alpha} \\
&\geq \frac{1}{2\sqrt{2} K_0 + 2\sqrt{2}(K_1 + 2^{-3/2} K_2^{1-\alpha})(K_0 + K_2 \|x_0 - x_*\|^{\alpha/1-\alpha})^\alpha \|x_0 - x_*\|^\alpha}.
\end{aligned}$$

Therefore from (42) we obtain

$$\min_{0 \leq k \leq K} \|F(x_k)\|^2 \leq \frac{16 \left(K_0 + (K_1 + 2^{-3/2} K_2^{1-\alpha})(K_0 + K_2 \|x_0 - x_*\|^{\alpha/1-\alpha})^\alpha \|x_0 - x_*\|^\alpha\right)^2 \|x_0 - x_*\|^2}{K+1}.$$

□

C.4 Local Convergence Guarantees for Weak Minty Operator

C.4.1 Proof of Theorem 3.8

Theorem C.8. Suppose F is weak Minty and 1-symmetric (L_0, L_1)-Lipschitz assumption. Moreover we assume

$$\Delta_1 := \frac{\nu}{L_0\left(1 + L_1\|x_0 - x_*\| e^{L_1\|x_0 - x_*\|}\right)} - 4\rho > 0. \tag{39}$$

Then EG with step size $\gamma_k = \frac{\nu}{L_0 + L_1\|F(x_k)\|}$ and $\omega_k = \gamma_k/2$ satisfies

$$\min_{0 \le k \le K} \|F(\hat{x}_k)\|^2 \le \frac{4L_0\left(1 + L_1 \exp\left(L_1\|x_0 - x_*\|\right)\|x_0 - x_*\|\right)\|x_0 - x_*\|^2}{\nu\Delta_1(K+1)} \tag{40}$$

where $\nu \exp \nu = 1$.

Proof. From the update rule of the Extragradient method, we have

$$\begin{aligned}
\|x_{k+1} - x_*\|^2 &= \left\|x_k - \frac{\gamma_k}{2}F(\hat{x}_k) - x_*\right\|^2 \\
&= \|x_k - x_*\|^2 - \gamma_k\left\langle F(\hat{x}_k), x_k - x_*\right\rangle + \frac{\gamma_k^2}{4}\|F(\hat{x}_k)\|^2 \\
&= \|x_k - x_*\|^2 - \gamma_k\left\langle F(\hat{x}_k), \hat{x}_k - x_*\right\rangle - \gamma_k\left\langle F(\hat{x}_k), x_k - \hat{x}_k\right\rangle + \frac{\gamma_k^2}{4}\|F(\hat{x}_k)\|^2 \\
&\overset{(9)}{\le} \|x_k - x_*\|^2 + \gamma_k\rho\|F(\hat{x}_k)\|^2 - \gamma_k\left\langle F(\hat{x}_k), x_k - \hat{x}_k\right\rangle + \frac{\gamma_k^2}{4}\|F(\hat{x}_k)\|^2 \\
&= \|x_k - x_*\|^2 + \gamma_k\rho\|F(\hat{x}_k)\|^2 - \gamma_k^2\left\langle F(\hat{x}_k), F(x_k)\right\rangle + \frac{\gamma_k^2}{4}\|F(\hat{x}_k)\|^2 \\
&\overset{(21)}{=} \|x_k - x_*\|^2 + \gamma_k\rho\|F(\hat{x}_k)\|^2 - \frac{\gamma_k^2}{2}\|F(\hat{x}_k)\|^2 - \frac{\gamma_k^2}{2}\|F(x_k)\|^2 + \frac{\gamma_k^2}{2}\|F(x_k) - F(\hat{x}_k)\|^2 \\
&\quad + \frac{\gamma_k^2}{4}\|F(\hat{x}_k)\|^2 \\
&= \|x_k - x_*\|^2 + \gamma_k\rho\|F(\hat{x}_k)\|^2 - \frac{\gamma_k^2}{4}\|F(\hat{x}_k)\|^2 - \frac{\gamma_k^2}{2}\|F(x_k)\|^2 + \frac{\gamma_k^2}{2}\|F(x_k) - F(\hat{x}_k)\|^2 \\
&\overset{(12)}{\le} \|x_k - x_*\|^2 + \gamma_k\rho\|F(\hat{x}_k)\|^2 - \frac{\gamma_k^2}{4}\|F(\hat{x}_k)\|^2 - \frac{\gamma_k^2}{2}\|F(x_k)\|^2 \\
&\quad + \frac{\gamma_k^2}{2}\left(L_0 + L_1\|F(x_k)\|\right)^2 \exp\left(2L_1\|x_k - \hat{x}_k\|\right)\|x_k - \hat{x}_k\|^2 \\
&= \|x_k - x_*\|^2 - \frac{\gamma_k}{4}\left(\gamma_k - 4\rho\right)\|F(\hat{x}_k)\|^2 \\
&\quad - \frac{\gamma_k^2}{2}\left(1 - \gamma_k^2(L_0 + L_1\|F(x_k)\|)^2 \exp\left(2\gamma_k L_1\|F(x_k)\|\right)\right)\|F(x_k)\|^2 \\
&\le \|x_k - x_*\|^2 - \frac{\gamma_k}{4}\left(\gamma_k - 4\rho\right)\|F(\hat{x}_k)\|^2 \\
&\quad - \frac{\gamma_k^2}{2}\left(1 - \gamma_k^2(L_0 + L_1\|F(x_k)\|)^2 \exp\left(2\gamma_k(L_0 + L_1\|F(x_k)\|)\right)\right)\|F(x_k)\|^2.
\end{aligned}$$

Similar to the proof of Theorem 3.5, we have

$$\|x_{k+1} - x_*\|^2 \quad \le \quad \|x_k - x_*\|^2 - \frac{\gamma_k}{4}(\gamma_k - 4\rho)\|F(\hat{x}_k)\|^2$$

for $\gamma_k = \frac{\nu}{L_0 + L_1\|F(x_k)\|}$ and $\nu \exp \nu = 1$. Again similar to Theorem 3.5, step size γ_k is lower bounded with

$$\gamma_k = \frac{\nu}{L_0 + L_1\|F(x_k)\|} \ge \frac{\nu}{L_0\left(1 + L_1 \exp\left(L_1\|x_0 - x_*\|\right)\|x_0 - x_*\|\right)}.$$

Hence from (16) we get

$$\frac{\gamma_k\Delta_1}{4}\|F(\hat{x}_k)\|^2 \quad \le \quad \|x_k - x_*\|^2 - \|x_{k+1} - x_*\|^2$$

Then we sum up this inequality for $k = 0, 1, \cdots, K$ to get

$$\frac{1}{K+1}\sum_{k=0}^{K}\frac{\gamma_k\Delta_1}{4}\|F(\hat{x}_k)\|^2 \quad \le \quad \frac{\|x_0 - x_*\|^2}{K+1}.$$

Therefore, we get

$$\min_{0 \leq k \leq K} \|F(\hat{x}_k)\|^2 \leq \frac{4\|x_0 - x_*\|^2}{\gamma \Delta_1 (K+1)}.$$

□

C.4.2 Proof of Theorem 3.9

Theorem C.9. Suppose F is weak Minty and α-symmetric (L_0, L_1)-Lipschitz with $\alpha \in (0,1)$. Moreover we assume

$$\Delta_\alpha := \frac{1}{2\sqrt{2}K_0 + 2\sqrt{2}(K_1 + 2^{-3/2}K_2^{1-\alpha})(K_0 + K_2\|x_0 - x_*\|^{\alpha/1-\alpha})^\alpha \|x_0 - x_*\|^\alpha} - 4\rho > 0.$$

Then EG with step size $\gamma_k = \frac{1}{2\sqrt{2}K_0 + \left(2\sqrt{2}K_1 + 2^{3(1-\alpha)/2}K_2^{1-\alpha}\right)\|F(x_k)\|^\alpha}$ and $\omega_k = \frac{\gamma_k}{2}$ satisfy

$$\min_{0 \leq k \leq K} \|F(\hat{x}_k)\|^2 \leq \frac{4\left(K_0 + \left(K_1 + 2^{-3/2}K_2^{1-\alpha}\right)(K_0 + K_2\|x_0 - x_*\|^{\alpha/1-\alpha})^\alpha \|x_0 - x_*\|^\alpha\right)\|x_0 - x_*\|^2}{\Delta_\alpha (K+1)}.$$

Proof. From the update rule of the Extragradient method, we have

$$\begin{aligned}
\|x_{k+1} - x_*\|^2 &= \left\|x_k - \frac{\gamma_k}{2}F(\hat{x}_k) - x_*\right\|^2 \\
&= \|x_k - x_*\|^2 - \gamma_k \langle F(\hat{x}_k), x_k - x_* \rangle + \frac{\gamma_k^2}{4}\|F(\hat{x}_k)\|^2 \\
&= \|x_k - x_*\|^2 - \gamma_k \langle F(\hat{x}_k), \hat{x}_k - x_* \rangle - \gamma_k \langle F(\hat{x}_k), x_k - \hat{x}_k \rangle + \frac{\gamma_k^2}{4}\|F(\hat{x}_k)\|^2 \\
&\overset{(9)}{\leq} \|x_k - x_*\|^2 + \gamma_k \rho \|F(\hat{x}_k)\|^2 - \gamma_k \langle F(\hat{x}_k), x_k - \hat{x}_k \rangle + \frac{\gamma_k^2}{4}\|F(\hat{x}_k)\|^2 \\
&= \|x_k - x_*\|^2 + \gamma_k \rho \|F(\hat{x}_k)\|^2 - \gamma_k^2 \langle F(\hat{x}_k), F(x_k) \rangle + \frac{\gamma_k^2}{4}\|F(\hat{x}_k)\|^2 \\
&\overset{(21)}{=} \|x_k - x_*\|^2 + \gamma_k \rho \|F(\hat{x}_k)\|^2 - \frac{\gamma_k^2}{2}\|F(\hat{x}_k)\|^2 - \frac{\gamma_k^2}{2}\|F(x_k)\|^2 + \frac{\gamma_k^2}{2}\|F(x_k) - F(\hat{x}_k)\|^2 \\
&\quad + \frac{\gamma_k^2}{4}\|F(\hat{x}_k)\|^2 \\
&= \|x_k - x_*\|^2 + \gamma_k \rho \|F(\hat{x}_k)\|^2 - \frac{\gamma_k^2}{4}\|F(\hat{x}_k)\|^2 - \frac{\gamma_k^2}{2}\|F(x_k)\|^2 + \frac{\gamma_k^2}{2}\|F(x_k) - F(\hat{x}_k)\|^2 \\
&\overset{(12)}{\leq} \|x_k - x_*\|^2 + \gamma_k \rho \|F(\hat{x}_k)\|^2 - \frac{\gamma_k^2}{4}\|F(\hat{x}_k)\|^2 - \frac{\gamma_k^2}{2}\|F(x_k)\|^2 \\
&\quad + \frac{\gamma_k^2}{2}\left(K_0 + K_1\|F(x_k)\|^\alpha + K_2\|x_k - \hat{x}_k\|^{\alpha/1-\alpha}\right)^2 \|x_k - \hat{x}_k\|^2 \\
&= \|x_k - x_*\|^2 - \frac{\gamma_k}{4}(\gamma_k - 4\rho)\|F(\hat{x}_k)\|^2 \\
&\quad - \frac{\gamma_k^2}{2}\left(1 - \gamma_k^2\left(K_0 + K_1\|F(x_k)\|^\alpha + \gamma_k^{\alpha/1-\alpha}K_2\|F(x_k)\|^{\alpha/1-\alpha}\right)^2\right)\|F(x_k)\|^2.
\end{aligned}$$

Here we choose $\gamma_k = \frac{1}{2(K_0 + K_1\|F(x_k)\|^\alpha) + 2^{3(1-\alpha)/2}K_2^{1-\alpha}\|F(x_k)\|^\alpha}$ and we get the following from (36)

$$\|x_{k+1} - x_*\|^2 \leq \|x_k - x_*\|^2 - \frac{\gamma_k}{4}(\gamma_k - 4\rho)\|F(\hat{x}_k)\|^2. \tag{41}$$

Rearranging this inequality, we have

$$\frac{\gamma_k}{4}(\gamma_k - 4\rho)\|F(\hat{x}_k)\|^2 \leq \|x_k - x_*\|^2 - \|x_{k+1} - x_*\|^2.$$

Then we sum up this inequality for $k = 0, 1, \cdots K$ to get

$$\frac{1}{K+1}\sum_{k=0}^{K} \frac{\gamma_k}{4}(\gamma_k - 4\rho)\|F(\hat{x}_k)\|^2 \leq \frac{2\|x_0 - x_*\|^2}{K+1}. \tag{42}$$

For this step size, we also have $\|x_k - x_0\|^2 \leq \|x_0 - x_*\|^2$ from (41). Now note that from (13) we obtain the following bound with $x = x_k$ and $y = x_*$

$$\begin{aligned}
\|F(x_k)\|^\alpha &\leq (K_0 + K_2\|x_k - x_*\|^{\alpha/1-\alpha})^\alpha \|x_k - x_*\|^\alpha \\
&\overset{(41)}{\leq} (K_0 + K_2\|x_0 - x_*\|^{\alpha/1-\alpha})^\alpha \|x_0 - x_*\|^\alpha.
\end{aligned}$$

We use this to lower bound the step size γ_k as follows

$$
\begin{aligned}
\gamma_k &= \frac{1}{2\sqrt{2}(K_0 + K_1\|F(x_k)\|^{\alpha}) + 2^{3(1-\alpha)/2}K_2^{1-\alpha}\|F(x_k)\|^{\alpha}} \\
&\geq \frac{1}{2\sqrt{2}K_0 + 2\sqrt{2}(K_1 + 2^{-3/2}K_2^{1-\alpha})(K_0 + K_2\|x_0 - x_*\|^{\alpha/1-\alpha})^{\alpha}\|x_0 - x_*\|^{\alpha}}.
\end{aligned}
$$

Therefore from (42) we obtain

$$
\min_{0\leq k\leq K}\|F(x_k)\|^2 \leq \frac{4\|x_0 - x_*\|^2}{\gamma\Delta(K+1)}.
$$

□

D Equivalent Formulation of α-Symmetric (L_0, L_1)-Lipschitz Assumption

In this section, we consider the min-max optimization problem given by $\min_{w_1} \max_{w_2} \mathcal{L}(w_1, w_2)$ and provide an equivalent formulation of α-symmetric (L_0, L_1)-Lipschitz operator. Next, we provide an example where we use this formulation to compute the constants α, L_0, L_1.

D.1 Proof of Theorem 2.1

Theorem D.1. Suppose F is the operator for the problem

$$\min_{w_1} \max_{w_2} \mathcal{L}(w_1, w_2).$$

Then F satisfies α-symmetric (L_0, L_1)-Lipschitz assumption if and only if

$$\|\mathbf{J}(x)\| = \sup_{\|u\|=1} \|\mathbf{J}(x)u\| \leq L_0 + L_1 \|F(x)\|^\alpha$$

where

$$\mathbf{J}(x) = \begin{bmatrix} \nabla^2_{w_1 w_1} \mathcal{L}(w_1, w_2) & \nabla^2_{w_2 w_1} \mathcal{L}(w_1, w_2) \\ -\nabla^2_{w_1 w_2} \mathcal{L}(w_1, w_2) & -\nabla^2_{w_2 w_2} \mathcal{L}(w_1, w_2) \end{bmatrix}.$$

Here $\|\mathbf{J}(x)\| = \sigma_{\max}(\mathbf{J}(x))$ i.e. maximum singular value of $\mathbf{J}(x)$.

Proof. Following (23), we have the equivalent characterization of F given by

$$\|F(y) - F(x)\| \leq \left(L_0 + L_1 \int_0^1 \|F(\theta y + (1-\theta)x)\|^\alpha \, d\theta\right) \|y - x\| \qquad \forall x, y \in \mathbb{R}^d.$$

As this inequality holds for any $x, y \in \mathbb{R}^d$, we choose $y = x + \theta' u$ where $\|u\| = 1$ and $\theta' \in (0, 1)$. Then we get

$$\|F(x + \theta' u) - F(x)\| \leq \left(L_0 + L_1 \int_0^1 \left\|F(x + \theta' \theta u)\right\|^\alpha d\theta\right) \|\theta' u\| \qquad \forall x \in \mathbb{R}^d.$$

The right-hand side of this inequality can be rewritten as

$$\begin{aligned} \left(L_0 + L_1 \int_0^1 \left\|F(x + \theta'\theta u)\right\|^\alpha d\theta\right) \|\theta' u\| &= \theta' \left(L_0 + L_1 \int_0^1 \left\|F(x + \theta'\theta u)\right\|^\alpha d\theta\right) \\ &= L_0 \theta' + L_1 \int_0^1 \left\|F(x + \theta'\theta u)\right\|^\alpha \theta' d\theta \\ &= L_0 \theta' + L_1 \int_0^{\theta'} \|F(x + \varphi u)\|^\alpha d\varphi. \end{aligned}$$

In the last line, we used the change of variable with $\varphi = \theta'\theta$. Therefore, we get

$$\frac{\|F(x + \theta' u) - F(x)\|}{\theta'} \leq L_0 + \frac{L_1}{\theta'} \int_0^{\theta'} \|F(x + \varphi u)\|^\alpha d\varphi.$$

Then we take $\theta' \to 0$ and use L'Hôpital's rule and Leibniz Integral rule to obtain

$$\lim_{\theta' \to 0} \frac{\|F(x + \theta' u) - F(x)\|}{\theta'} \leq L_0 + L_1 \|F(x)\|^\alpha.$$

Moreover, note that the left-hand side is given by $\|\mathbf{J}(x)u\|$ where

$$\mathbf{J}(x) = \begin{bmatrix} \nabla^2_{w_1 w_1} \mathcal{L}(w_1, w_2) & \nabla^2_{w_2 w_1} \mathcal{L}(w_1, w_2) \\ -\nabla^2_{w_1 w_2} \mathcal{L}(w_1, w_2) & -\nabla^2_{w_2 w_2} \mathcal{L}(w_1, w_2) \end{bmatrix}.$$

Therefore, for any $\|u\| = 1$ we have

$$\|\mathbf{J}(x)u\| \leq L_0 + L_1 \|F(x)\|^\alpha.$$

Hence we get

$$\|\mathbf{J}(x)\| = \sup_{\|u\|=1} \|\mathbf{J}(x)u\| \leq L_0 + L_1 \|F(x)\|^\alpha.$$

Now we want to show the other way, i.e. suppose we have $\|\mathbf{J}(x)\| \leq L_0 + L_1 \|F(x)\|^\alpha$. For this we define,

$$q(\theta) := F(\theta x + (1-\theta)y).$$

Then $q(1) = F(x)$ and $q(0) = F(y)$ and we have

$$\begin{aligned}
\|F(x) - F(y)\| &= \|q(1) - q(0)\| \\
&= \left\| \int_0^1 \frac{dq(\theta)}{d\theta} d\theta \right\| \\
&= \left\| \int_0^1 \frac{dF(\theta x + (1-\theta)y)}{d\theta} d\theta \right\| \\
&= \left\| \int_0^1 \mathbf{J}(\theta x + (1-\theta)y)(x-y) d\theta \right\| \\
&\leq \int_0^1 \|\mathbf{J}(\theta x + (1-\theta)y)\| \, \|x-y\| \, d\theta \\
&= \left(\int_0^1 \|\mathbf{J}(\theta x + (1-\theta)y)\| \, d\theta \right) \|x-y\| \\
&\leq \left(\int_0^1 L_0 + L_1 \|F(\theta x + (1-\theta)y)\|^\alpha d\theta \right) \|x-y\| \\
&= \left(L_0 + L_1 \int_0^1 \|F(\theta x + (1-\theta)y)\|^\alpha d\theta \right) \|x-y\| \, .
\end{aligned}$$

Then, using Lemma A.3, we have the result. □

D.2 Computation of α, L_0, L_1 for $\mathcal{L}(w_1, w_2)$.

We now revisit the min-max problem defined in (5). Note that, the operator corresponding to this problem is given by

$$F(x) = \begin{bmatrix} w_1^2 + w_2 \\ w_2^2 - w_1 \end{bmatrix}$$

Then the norm of operator is $\|F(x)\| = \sqrt{\left(w_1^2 + w_2\right)^2 + \left(w_2^2 - w_1\right)^2}$. Moreover, the Jacobian matrix is given by

$$\mathbf{J}(x) = \begin{bmatrix} 2w_1 & 1 \\ -1 & 2w_2 \end{bmatrix}.$$

Then the maximum singular value at any point x is given by

$$\begin{aligned}
\|\mathbf{J}(x)\| &= \lambda_{\max}\left(\mathbf{J}(x)^\top \mathbf{J}(x)\right) \\
&= \lambda_{\max}\left(\begin{bmatrix} 4w_1^2 + 1 & 2(w_1 - w_2) \\ 2(w_1 - w_2) & 4w_2^2 + 1 \end{bmatrix} \right) \\
&\overset{(24)}{=} \sqrt{2(w_1^2 + w_2^2) + 1 + 2\sqrt{(w_1 - w_2)^2 + (w_1^2 - w_2^2)^2}}
\end{aligned} \tag{43}$$

To validate whether the operator F satisfies the condition (6), we examine whether the following function is non-negative:

$$g(w_1, w_2) = L_0 + L_1 \|F(x)\| - \|\mathbf{J}(x)\|. \tag{44}$$

In Figure 5, we plot $g(w_1, w_2)$ using $L_0 = 10$ and $L_1 = 10$. We observe that $g(w_1, w_2)$ has no real solution and remains positive for all $w_1, w_2 \in \mathbb{R}$, confirming that the function (5) satisfies (11) with $(\alpha, L_0, L_1) = (1, 10, 10)$. Thus, the corresponding operator F is 1-symmetric $(10, 10)$-Lipschitz.

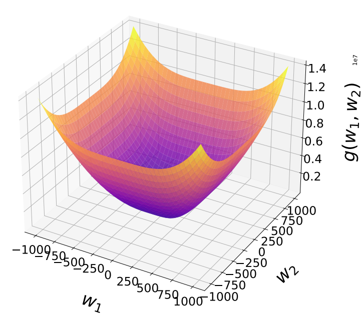


Figure 5: Plot of $g(w_1, w_2)$ (44). Here, the z-axis is in 10^7 scale.

E Additional Details on Numerical Experiments

In this section, we provide additional details on the second experiment related to the monotone problem. First, we show that

$$\mathcal{L}(w_1, w_2) = \frac{1}{3}\left(w_1^\top \mathbf{A} w_1\right)^{3/2} + w_1^\top \mathbf{B} w_2 - \frac{1}{3}\left(w_2^\top \mathbf{C} w_2\right)^{3/2}$$

is convex-concave, and then we find the equilibrium point of $\mathcal{L}$.

Convex-Concave $\mathcal{L}(w_1, w_2)$. Note that for $\mathcal{L}(w_1, w_2)$ in (19) we have $\nabla_{w_1}\mathcal{L}(w_1, w_2) = \left(w_1^\top \mathbf{A} w_1\right)^{1/2} \mathbf{A} w_1 + \mathbf{B} w_2$ and $\nabla_{w_2}\mathcal{L}(w_1, w_2) = \mathbf{B}^\top w_1 - \left(w_2^\top \mathbf{C} w_2\right)^{1/2} \mathbf{C} w_2$. Then the second-order derivatives are given by

$$\nabla^2_{w_1 w_1}\mathcal{L}(w_1, w_2) = \|\mathbf{A}^{1/2} w_1\|\mathbf{A} + \frac{\mathbf{A} w_1 w_1^\top \mathbf{A}^\top}{\|\mathbf{A}^{1/2} w_1\|}$$

Here, $\mathbf{A}$ is positive definite and $\mathbf{A} w_1 w_1^\top \mathbf{A}^\top$ is a positive semidefinite matrix. Hence, $\nabla^2_{w_1 w_1}\mathcal{L}(w_1, w_2)$ is a positive definite matrix as well and $\mathcal{L}(\cdot, w_2)$ is convex for any w_2. Similarly, we show that

$$-\nabla^2_{w_2 w_2}\mathcal{L}(w_1, w_2) = \|\mathbf{C}^{1/2} w_2\|\mathbf{C} + \frac{\mathbf{C} w_2 w_2^\top \mathbf{C}^\top}{\|\mathbf{C}^{1/2} w_2\|}$$

and $-\nabla^2_{w_2 w_2}\mathcal{L}(w_1, w_2)$ is positive definite. Therefore, $\mathcal{L}(w_1, \cdot)$ is concave for any w_1. This proves that $\mathcal{L}(w_1, w_2)$ is convex with respect to w_1 and concave with respect to w_2. Thus, we conclude that the corresponding operator F is monotone.

Equilibrium of $\mathcal{L}(w_1, w_2)$. To find the equilibrium points, we solve the set of equations given by $\nabla_{w_1}\mathcal{L}(w_1, w_2) = 0$ and $\nabla_{w_2}\mathcal{L}(w_1, w_2) = 0$, i.e. solve for

$$\begin{aligned} \left(w_1^\top \mathbf{A} w_1\right)^{1/2} \mathbf{A} w_1 + \mathbf{B} w_2 &= 0, \\ \mathbf{B}^\top w_1 - \left(w_2^\top \mathbf{C} w_2\right)^{1/2} \mathbf{C} w_2 &= 0. \end{aligned}$$

Now multiplying the first equation with $w_1^\top$ and second one with $w_2^\top$, we have

$$\begin{aligned} \left(w_1^\top \mathbf{A} w_1\right)^{1/2} w_1^\top \mathbf{A} w_1 + w_1^\top \mathbf{B} w_2 &= 0 \\ w_2^\top \mathbf{B}^\top w_1 - \left(w_2^\top \mathbf{C} w_2\right)^{1/2} w_2^\top \mathbf{C} w_2 &= 0 \end{aligned}$$

Combining these two equations, we get

$$\left(w_1^\top \mathbf{A} w_1\right)^{1/2} w_1^\top \mathbf{A} w_1 + \left(w_2^\top \mathbf{C} w_2\right)^{1/2} w_2^\top \mathbf{C} w_2 = 0$$

which can be equivalently written as

$$\left\|\mathbf{A}^{1/2} w_1\right\|^3 + \left\|\mathbf{C}^{1/2} w_2\right\|^3 = 0$$

which implies $w_1 = w_2 = 0$ as both $\mathbf{A}, \mathbf{C}$ are positive definite matrices (hence $\mathbf{A}^{1/2}, \mathbf{C}^{1/2}$ are invertible).