Convergence Rate of First-Order Methods for Saddle Point Problems

Ali Zindari, Parham Yazdkhasti

1 Introduction

Min-max optimization problems have recently gained a lot of attention. It has a wide range of usage in machine learning (ML) such as: adversarial attacks, generative adversarial networks (GANs) and also adversarial robustness. However, finding a solution for these problems is more challenging than a regular minimization problem which arises frequently in ML. There has been many works recently around proposing methods for finding the saddle point of a min-max game based on first-order information. In this project, we aim to provide convergence rates for different first-order methods used for min-max optimization problems. We considered the methods: Gradient Descent Ascent (GDA), Proximal Point (PP) and Extra Gradient Descent (EG). In the following, we first review some recent works on related to min-max optimization problems. Then we introduce some assumptions that will be used throughout this work and next we provide several convergence rates based on different assumptions and different optimizers.

2 Related Works

After the introduction of GANs [GPAM+14] which works based on solving a min-max game between two agents, saddle point problems have gained more attention. Different methods have been proposed for solving these problems. In [GLG22], the authors proposed extra gradient methods which could solve bi-linear problems which GDA couldn't solve before. Also [CPS+20] proposed the look ahead method which takes several steps of GD first to find an intermediate point and uses a weighted average of this point and the current point to find the next point. This leads to a faster convergence but comes at the cost of more gradient computations. Min-max problems also have been studied in the distributed setting such as in the work [ZCSL23]. Adibi et. al. [AMH23] investigated GDA under delay and showed that it can converge even if we don't have access to the most recent gradients. Another interesting direction is the relation between different optimizers that have been used so far. [MOP20] showed the rates for OGDA and EG with respect to the fact that they can be used as an approximation of PP.

3 Methods

In this section, we briefly introduce the methods that we used in this work and their update rules.

3.1 Gradient Descent Ascent (GDA)

This is the most basic method used for solving saddle point problems. The update rule for this method is as follows:

$$\begin{aligned}\mathbf{x}_{t+1} &= \mathbf{x}_t - \gamma \nabla_x f(\mathbf{x}_t, \mathbf{y}_t)\\ \mathbf{y}_{t+1} &= \mathbf{y}_t + \gamma \nabla_y f(\mathbf{x}_t, \mathbf{y}_t)\end{aligned}$$

3.2 Proximal Point (PP)

The update rule for this method is as follows:

$$\begin{aligned}\mathbf{x}_{t+1} &= \mathbf{x}_t - \gamma \nabla_x f(\mathbf{x}_{t+1}, \mathbf{y}_{t+1})\\ \mathbf{y}_{t+1} &= \mathbf{y}_t + \gamma \nabla_y f(\mathbf{x}_{t+1}, \mathbf{y}_{t+1})\end{aligned}$$

3.3 Extra Gradient (EG)

The update rule for this method is as follows:

$$\begin{aligned}\mathbf{x}_{t+\frac{1}{2}} &= \mathbf{x}_t - \gamma \nabla_x f(\mathbf{x}_t, \mathbf{y}_t), \quad \mathbf{y}_{t+\frac{1}{2}} = \mathbf{y}_t + \gamma \nabla_y f(\mathbf{x}_t, \mathbf{y}_t)\\ \mathbf{x}_{t+1} &= \mathbf{x}_t - \gamma \nabla_x f(\mathbf{x}_{t+\frac{1}{2}}, \mathbf{y}_{t+\frac{1}{2}}), \quad \mathbf{y}_{t+1} = \mathbf{y}_t + \gamma \nabla_y f(\mathbf{x}_{t+\frac{1}{2}}, \mathbf{y}_{t+\frac{1}{2}})\end{aligned}$$

4 Convergence Rates

In this section, we first introduce a set of assumptions that will be used in the theorems. Then we provide the convergence rates for different optimizers based on different assumptions. In each theorem, we will explicitly mention which assumptions are being used.

Assumption 1 (**L-Lipschitz**). *Let $F : S \to \mathbb{R}^n$ be an operator. F is an L-Lipschitz map on S iff there exists a positive L such that:*

$$\|F(\mathbf{z}) - F(\mathbf{z}')\| \leq L\|\mathbf{z} - \mathbf{z}'\|, \quad \mathbf{z}, \mathbf{z}' \in S \tag{1}$$

Assumption 2 (**Monotone**). *Let $F : S \to \mathbb{R}^n$ be an operator. F is a monotone map on S iff following inequality holds:*

$$\langle F(\mathbf{z}) - F(\mathbf{z}'), \mathbf{z} - \mathbf{z}' \rangle \geq 0, \quad \mathbf{z}, \mathbf{z}' \in S \tag{2}$$

Assumption 3 (**Strongly Monotone**). *Let $F : S \to \mathbb{R}^n$ be an operator. F is a monotone map on S iff following inequality holds:*

$$\langle F(\mathbf{z}) - F(\mathbf{z}'), \mathbf{z} - \mathbf{z}' \rangle \geq \mu\|\mathbf{z} - \mathbf{z}'\|^2, \quad \mathbf{z}, \mathbf{z}' \in S \tag{3}$$

In this part, we provide convergence rates for different first-order methods for solving min-max optimization problems.

Theorem 1. *Let f be a bi-linear function in the form of $f(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{B}\mathbf{y}$ where $\mathbf{B}$ is a symmetric positive definite matrix. After T iterations of PP method on this class of function, we have the following rate:*

$$\|\mathbf{x}_T\|^2 + \|\mathbf{y}_T\|^2 \leq \Big(\frac{1}{1 + \gamma^2 \lambda_{min}(\mathbf{B}\mathbf{B}^\top)}\Big)^T (\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2) \tag{4}$$

Please note that we know $(\mathbf{x}^\star, \mathbf{y}^\star) = (0, 0)$.

Theorem 2. *Let $F : S \to \mathbb{R}^n$ be a L-Lipschitz and μ-strongly monotone operator; and let*

$$\langle F(\mathbf{z}), \mathbf{z}' - \mathbf{z} \rangle \geq 0 \ \ \forall \mathbf{z}'$$

define the variational inequality that we want to solve. By running T iterations of GDA algorithm with step size $\gamma \leq \frac{\mu}{L^2}$, we have the following rate:

$$\|\mathbf{z}_T - \mathbf{z}^\star\|^2 \leq exp\left(\frac{-T\mu^2}{L^2}\right) \|\mathbf{z}_0 - \mathbf{z}^\star\|^2 \tag{5}$$

Theorem 3. *Let $F : S \to \mathbb{R}^n$ be a L-Lipschitz and μ-strongly monotone operator; and let*

$$\langle F(\mathbf{z}), \mathbf{z}' - \mathbf{z} \rangle \geq 0 \ \ \forall \mathbf{z}'$$

define the variational inequality that we want to solve. By running T iterations of EG algorithm with step size $\gamma \leq \frac{\mu}{2L^2}$, we have the following rate:

$$\|\mathbf{z}_T - \mathbf{z}^\star\|^2 \leq \left(1 + \frac{\mu^4}{16L^4} - \frac{\mu^2}{4L^2}\right)^T \|\mathbf{z}_0 - \mathbf{z}^\star\|^2 \tag{6}$$

Theorem 4. *Let $f(\mathbf{x}, \mathbf{y})$ be a general convex-concave function. After T iterations of PP on this class of function, we have the following rate for the average iterate:*

$$f(\bar{\mathbf{x}}_{t+1}, \bar{\mathbf{y}}_{t+1}) - f(\mathbf{x}^\star, \mathbf{y}^\star) \leq \frac{\|\mathbf{z}_0 - \mathbf{z}^\star\|^2}{2\gamma T}$$

Where $\bar{\mathbf{x}}_{t+1} = \frac{1}{T}\sum_{t=0}^{T-1} \mathbf{x}_{t+1}$ and $\bar{\mathbf{y}}_{t+1} = \frac{1}{T}\sum_{t=0}^{T-1} \mathbf{y}_{t+1}$.

5 Contributors

This project was done in a group of two by Parham Yazdkhasti and Ali Zindari. Theorems 1 and 4 are developed by Ali Zindari and Theorems 2 and 3 are developed by Parham Yazdkhasti. We used some slides and papers for gaining some intuitions and having some ideas about how the general way of proving in Min-Max optimization problems work. All of the Theorems are reviewed by both members.

References

[AMH23] Arman Adibi, Aritra Mitra, and Hamed Hassani. Min-max optimization under delays. *arXiv preprint arXiv:2307.06886*, 2023.

[CPS+20] Tatjana Chavdarova, Matteo Pagliardini, Sebastian U Stich, François Fleuret, and Martin Jaggi. Taming gans with lookahead-minmax. *arXiv preprint arXiv:2006.14567*, 2020.

[GLG22] Eduard Gorbunov, Nicolas Loizou, and Gauthier 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, 2022.

[GPAM+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[MOP20] Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *International Conference on Artificial Intelligence and Statistics*, pages 1497–1507. PMLR, 2020.

[ZCSL23] Siqi Zhang, Sayantan Choudhury, Sebastian U Stich, and Nicolas Loizou. Communication-efficient gradient descent-accent methods for distributed variational inequalities: Unified analysis and local updates. *arXiv preprint arXiv:2306.05100*, 2023.

6 Proofs

We first make a list of useful lemmas which are going to be used in the proofs later on.

Lemma 5. *Let* $\mathbf{B}$ *be a full rank square matrix* $\mathbf{B} \in \mathbb{R}^{d\times d}$ *and* $\mathbf{P}_x := (\mathbf{I} + \gamma^2 \mathbf{B}\mathbf{B}^\top)^{-1}$, $\mathbf{P}_y := (\mathbf{I} + \gamma^2 \mathbf{B}^\top\mathbf{B})^{-1}$. *Then we have the following property:*

$$\mathbf{P}_x^2 \mathbf{B} = \mathbf{B}\mathbf{P}_y^2$$
$$\mathbf{P}_y^2 \mathbf{B}^\top = \mathbf{B}^\top \mathbf{P}_x^2$$

Proof. We start by using the singular value decomposition for matrix $\mathbf{B} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^\top$:

$$\begin{aligned}
\mathbf{P}_x^2\mathbf{B} &= \left(\mathbf{I} + \gamma^2 \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^\top\mathbf{V}\boldsymbol{\Lambda}\mathbf{U}^\top\right)^{-2}\mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^\top \\
&= \left(\mathbf{U}\left(\gamma^2\boldsymbol{\Lambda}^2 + \mathbf{I}\right)\mathbf{U}^\top\right)^{-2}\mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^\top \\
&= (\mathbf{U}^\top)^{-2}\left(\gamma^2\boldsymbol{\Lambda}^2 + \mathbf{I}\right)^{-2}\mathbf{U}^{-2}\mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^\top \\
&= \mathbf{U}^2\left(\gamma^2\boldsymbol{\Lambda}^2 + \mathbf{I}\right)^{-2}\mathbf{U}^\top\boldsymbol{\Lambda}\mathbf{V}^\top \\
&= \left(\gamma^2\boldsymbol{\Lambda}^2 + \mathbf{I}\right)^{-2}\mathbf{U}\mathbf{U}\mathbf{U}^\top\boldsymbol{\Lambda}\mathbf{V}^\top \\
&= \left(\gamma^2\boldsymbol{\Lambda}^2 + \mathbf{I}\right)^{-2}\mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^\top \\
&= \left(\gamma^2\boldsymbol{\Lambda}^2 + \mathbf{I}\right)^{-2}\boldsymbol{\Lambda}\mathbf{U}\mathbf{V}^\top
\end{aligned}$$

Here we used the fact that $\mathbf{U}$ and $\mathbf{V}$ are orthogonal so we have that $\mathbf{U}^\top\mathbf{U} = \mathbf{V}^\top\mathbf{V} = \mathbf{I}$. By starting from the other side we have:

$$\begin{aligned}
\mathbf{B}\mathbf{P}_y^2 &= \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^\top\left(\mathbf{I} + \gamma^2\mathbf{V}\boldsymbol{\Lambda}\mathbf{U}^\top\mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^\top\right)^{-2} \\
&= \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^\top\left(\mathbf{V}\left(\gamma^2\boldsymbol{\Lambda}^2 + \mathbf{I}\right)\mathbf{V}^\top\right)^{-2} \\
&= \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^\top(\mathbf{V}^\top)^{-2}\left(\gamma^2\boldsymbol{\Lambda}^2 + \mathbf{I}\right)^{-2}\mathbf{V}^{-2} \\
&= \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^\top\mathbf{V}^2\left(\gamma^2\boldsymbol{\Lambda}^2 + \mathbf{I}\right)^{-2}\mathbf{V}^{-2} \\
&= \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^\top\mathbf{V}^2\mathbf{V}^{-2}\left(\gamma^2\boldsymbol{\Lambda}^2 + \mathbf{I}\right)^{-2} \\
&= \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^\top\left(\gamma^2\boldsymbol{\Lambda}^2 + \mathbf{I}\right)^{-2} \\
&= \left(\gamma^2\boldsymbol{\Lambda}^2 + \mathbf{I}\right)^{-2}\boldsymbol{\Lambda}\mathbf{U}\mathbf{V}^\top
\end{aligned}$$

And we know that $\mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^\top\left(\gamma^2\boldsymbol{\Lambda}^2 + \mathbf{I}\right)^{-2} = \left(\gamma^2\boldsymbol{\Lambda}^2 + \mathbf{I}\right)^{-2}\boldsymbol{\Lambda}\mathbf{U}\mathbf{V}^\top$ as $\boldsymbol{\Lambda}$ and $\left(\gamma^2\boldsymbol{\Lambda}^2 + \mathbf{I}\right)^{-2}$ are both diagonal so we can switch them. □

6.1 Proof of Theorem 1

Proof. In this theorem, we assume that our objective is bi-linear in the form of $f_{\mathbf{x},\mathbf{y}}(\mathbf{x},\mathbf{y}) = \mathbf{x}^\top\mathbf{B}\mathbf{y}$ where $\mathbf{B}$ is a full rank square matrix. The update rule of PP for this problem can be written as:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma\mathbf{B}\mathbf{y}_{t+1}$$
$$\mathbf{y}_{t+1} = \mathbf{y}_t + \gamma\mathbf{B}^\top\mathbf{x}_{t+1}$$

We can write the explicit form of the next iterate as follows:

$$\mathbf{x}_{t+1} = (\mathbf{I} + \gamma^2\mathbf{B}\mathbf{B}^\top)^{-1}(\mathbf{x}_t - \gamma\mathbf{B}\mathbf{y}_t) = (\mathbf{I} + \gamma^2\mathbf{B}\mathbf{B}^\top)^{-1}\mathbf{x}_t - \gamma(\mathbf{I} + \gamma^2\mathbf{B}\mathbf{B}^\top)^{-1}\mathbf{B}\mathbf{y}_t$$
$$\mathbf{y}_{t+1} = (\mathbf{I} + \gamma^2\mathbf{B}^\top\mathbf{B})^{-1}(\mathbf{y}_t + \gamma\mathbf{B}^\top\mathbf{x}_t) = (\mathbf{I} + \gamma^2\mathbf{B}^\top\mathbf{B})^{-1}\mathbf{y}_t + \gamma(\mathbf{I} + \gamma^2\mathbf{B}^\top\mathbf{B})^{-1}\mathbf{B}^\top\mathbf{x}_t$$

Where we know that both matrices $\mathbf{P}_x := (\mathbf{I} + \gamma^2\mathbf{B}\mathbf{B}^\top)^{-1}$ and $P_y := (\mathbf{I} + \gamma^2\mathbf{B}^\top\mathbf{B})^{-1}$ are symmetric. Next we take the norm of both sides in the above equations and we have:

$$\|\mathbf{x}_{t+1}\|^2 = \|\mathbf{P}_x\mathbf{x}_t\|^2 + \gamma^2\|\mathbf{P}_x\mathbf{B}\mathbf{y}_t\|^2 - 2\gamma\mathbf{x}_t^\top\mathbf{P}_x^2\mathbf{B}\mathbf{y}_t$$
$$\|\mathbf{y}_{t+1}\|^2 = \|\mathbf{P}_y\mathbf{y}_t\|^2 + \gamma^2\left\|\mathbf{P}_y\mathbf{B}^\top\mathbf{x}_t\right\|^2 + 2\gamma\mathbf{y}_t^\top\mathbf{P}_y^2\mathbf{B}^\top\mathbf{x}_t$$

Now we show that $2\gamma\mathbf{x}_t^\top\mathbf{P}_x^2\mathbf{B}\mathbf{y}_t = 2\gamma\mathbf{y}_t^\top\mathbf{P}_y^2\mathbf{B}\mathbf{x}_t$:

$$2\gamma\mathbf{x}_t^\top\mathbf{P}_x^2\mathbf{B}\mathbf{y}_t = 2\gamma\mathbf{x}_t^\top\mathbf{B}\mathbf{P}_y^2\mathbf{y}_t = 2\gamma(\mathbf{B}\mathbf{P}_y^2\mathbf{y}_t)^\top\mathbf{x}_t = 2\gamma\mathbf{y}_t^\top\mathbf{P}_y^2\mathbf{B}^\top\mathbf{x}_t$$

Then we have:

$$\|\mathbf{x}_{t+1}\|^2 + \|\mathbf{y}_{t+1}\|^2 = \|\mathbf{P}_x\mathbf{x}_t\|^2 + \gamma^2\|\mathbf{P}_x\mathbf{B}\mathbf{y}_t\|^2 + \|\mathbf{P}_y\mathbf{y}_t\|^2 + \gamma^2\left\|\mathbf{P}_y\mathbf{B}^\top\mathbf{x}_t\right\|^2$$

Now we simplify the above equation and start with the terms:

$$\begin{aligned}
&\|\mathbf{P}_x\mathbf{x}_t\|^2 + \gamma^2\left\|\mathbf{P}_y\mathbf{B}^\top\mathbf{x}_t\right\|^2 \\
&= (\mathbf{P}_x\mathbf{x}_t)^\top(\mathbf{P}_x\mathbf{x}_t) + \gamma^2(\mathbf{P}_y\mathbf{B}^\top\mathbf{x}_t)^\top(\mathbf{P}_y\mathbf{B}^\top\mathbf{x}_t) \\
&= \mathbf{x}_t^\top\mathbf{P}_x^\top\mathbf{P}_x\mathbf{x}_t + \gamma^2\mathbf{x}_t^\top\mathbf{B}\mathbf{P}_y^\top\mathbf{P}_y\mathbf{B}^\top\mathbf{x}_t \\
&= \mathbf{x}_t^\top(\mathbf{P}_x^2 + \gamma^2\mathbf{B}\mathbf{P}_y^2\mathbf{B}^\top)\mathbf{x}_t \\
&= \mathbf{x}_t^\top(\mathbf{P}_x^2 + \gamma^2\mathbf{B}\mathbf{B}^\top\mathbf{P}_x^2)\mathbf{x}_t \\
&= \mathbf{x}_t^\top(\mathbf{I} + \gamma^2\mathbf{B}\mathbf{B}^\top)\mathbf{P}_x^2\mathbf{x}_t \\
&= \mathbf{x}_t^\top(\mathbf{I} + \gamma^2\mathbf{B}\mathbf{B}^\top)(\mathbf{I} + \gamma^2\mathbf{B}\mathbf{B}^\top)^{-2}\mathbf{x}_t \\
&= \mathbf{x}_t^\top(\mathbf{I} + \gamma^2\mathbf{B}\mathbf{B}^\top)^{-1}\mathbf{x}_t
\end{aligned}$$

After following a similar approach for $\mathbf{y}_t$ we have that:

$$\|\mathbf{P}_y\mathbf{y}_t\|^2 + \gamma^2\|\mathbf{P}_x\mathbf{B}\mathbf{y}_t\|^2 = \mathbf{y}_t^\top(\mathbf{I} + \gamma^2\mathbf{B}^\top\mathbf{B})^{-1}\mathbf{y}_t$$

Now by replacing everything into the main equation we have:

$$\begin{aligned}
\|\mathbf{x}_{t+1}\|^2 + \|\mathbf{y}_{t+1}\|^2 &= \mathbf{x}_t^\top(\mathbf{I} + \gamma^2\mathbf{B}\mathbf{B}^\top)^{-1}\mathbf{x}_t + \mathbf{y}_t^\top(\mathbf{I} + \gamma^2\mathbf{B}^\top\mathbf{B})^{-1}\mathbf{y}_t \\
&\leq \frac{1}{1 + \gamma^2\lambda_{min}(\mathbf{B}\mathbf{B}^\top)}(\|\mathbf{x}_t\|^2 + \|\mathbf{y}_t\|^2)
\end{aligned}$$

Where the LHS shows the distance of each parameter to the saddle point. Please note that in this specific problem the saddle point $(\mathbf{x}^\star, \mathbf{y}^\star) = (0, 0)$. □

6.2 Proof of Theorem 2

Proof. Since GDA can be equivalently written as $\mathbf{z}_{t+1} = \mathbf{z}_t - \gamma F(\mathbf{z}_t)$.

$$\begin{aligned}
\|\mathbf{z}_{t+1} - \mathbf{z}^\star\|^2 &= \|\mathbf{z}_t - \mathbf{z}^\star - \gamma F(\mathbf{z}_t)\|^2 \\
&= \|\mathbf{z}_t - \mathbf{z}^\star\|^2 + \gamma^2\|F(\mathbf{z}_t)\|^2 - 2\gamma\langle\mathbf{z}_t - \mathbf{z}^\star, F(\mathbf{z}_t)\rangle \\
&= \|\mathbf{z}_t - \mathbf{z}^\star\|^2 + \gamma^2\|F(\mathbf{z}_t)\|^2 - 2\gamma\langle\mathbf{z}_t - \mathbf{z}^\star, F(\mathbf{z}_t) - F(\mathbf{z}^\star)\rangle \\
&\overset{(3)}{\leq} \|\mathbf{z}_t - \mathbf{z}^\star\|^2 + \gamma^2\|F(\mathbf{z}_t)\|^2 - 2\gamma\mu\|\mathbf{z}_t - \mathbf{z}^\star\|^2 \\
&= \|\mathbf{z}_t - \mathbf{z}^\star\|^2 + \gamma^2\|F(\mathbf{z}_t) - F(\mathbf{z}^\star)\|^2 - 2\gamma\mu\|\mathbf{z}_t - \mathbf{z}^\star\|^2 \\
&\overset{(1)}{\leq} \|\mathbf{z}_t - \mathbf{z}^\star\|^2 + \gamma^2L^2\|\mathbf{z}_t - \mathbf{z}^\star\|^2 - 2\gamma\mu\|\mathbf{z}_t - \mathbf{z}^\star\|^2 \\
&= (1 + \gamma^2L^2 - 2\gamma\mu)\|\mathbf{z}_t - \mathbf{z}^\star\|^2
\end{aligned}$$

We need to set γ such that $(1 + \gamma^2L^2 - 2\gamma\mu) < 1$. Therefore we can conclude; $\gamma < \frac{2\mu}{L^2}$. Setting $\gamma \leq \frac{\mu}{L^2}$,

$$\|\mathbf{z}_{t+1} - \mathbf{z}^\star\|^2 \leq \left(1 - \frac{\mu^2}{L^2}\right)\|\mathbf{z}_t - \mathbf{z}^\star\|^2$$

By recursively replacing from $t = 0$ *to* $T - 1$,

$$\begin{aligned}
\|\mathbf{z}_{t+1} - \mathbf{z}^\star\|^2 &\leq \left(1 - \frac{\mu^2}{L^2}\right)^T\|\mathbf{z}_0 - \mathbf{z}^\star\|^2 \\
&\leq exp\left(\frac{-T\mu^2}{L^2}\right)\|\mathbf{z}_0 - \mathbf{z}^\star\|^2
\end{aligned}$$

□

6.3 Proof of Theorem 3

Proof. Since EG can be equivalently written as

$$\mathbf{z}_{t+\frac{1}{2}} = \mathbf{z}_t - \gamma F(\mathbf{z}_t) \tag{7}$$

$$\mathbf{z}_{t+1} = \mathbf{z}_t - \gamma F(\mathbf{z}_{t+\frac{1}{2}}) \tag{8}$$

Here we start by expanding the $t+1$ iteration's distance to the $z^\star$,

$$\begin{aligned}\|\mathbf{z}_{t+1} - \mathbf{z}^\star\|^2 &= \|\mathbf{z}_t - \mathbf{z}^\star - \gamma F(\mathbf{z}_{t+\frac{1}{2}})\|^2 \\ &= \|\mathbf{z}_t - \mathbf{z}^\star\|^2 + \gamma^2 \underbrace{\|F(\mathbf{z}_{t+\frac{1}{2}})\|^2}_{:=A} - 2\gamma \underbrace{\langle \mathbf{z}_t - \mathbf{z}^\star, F(\mathbf{z}_{t+\frac{1}{2}})\rangle}_{:=B}\end{aligned}$$

Let us first simplify term A.

$$\begin{aligned}A :&= \|F(\mathbf{z}_{t+\frac{1}{2}}) - F(\mathbf{z}^\star)\|^2 \\ &\overset{(1)}{\leq} L^2\|\mathbf{z}_{t+\frac{1}{2}} - \mathbf{z}^\star\|^2\end{aligned}$$

Now we simplify term B.

$$\begin{aligned}B :&= -\langle \mathbf{z}_t - \mathbf{z}^\star, F(\mathbf{z}_{t+\frac{1}{2}})\rangle \\ &= -\langle \mathbf{z}_t - \gamma F(\mathbf{z}_t) - \mathbf{z}^\star, F(\mathbf{z}_{t+\frac{1}{2}})\rangle - \gamma\langle F(\mathbf{z}_t), F(\mathbf{z}_{t+\frac{1}{2}})\rangle \\ &\overset{(7)}{=} -\langle \mathbf{z}_{t+\frac{1}{2}} - \mathbf{z}^\star, F(\mathbf{z}_{t+\frac{1}{2}})\rangle - \gamma\langle F(\mathbf{z}_t), F(\mathbf{z}_{t+\frac{1}{2}})\rangle \\ &= -\langle \mathbf{z}_{t+\frac{1}{2}} - \mathbf{z}^\star, F(\mathbf{z}_{t+\frac{1}{2}})\rangle - \frac{\gamma}{2}\|F(\mathbf{z}_t)\|^2 - \frac{\gamma}{2}\|F(\mathbf{z}_{t+\frac{1}{2}})\|^2 + \frac{\gamma}{2}\|F(\mathbf{z}_t) - F(\mathbf{z}_{t+\frac{1}{2}})\|^2 \\ &\overset{(3)}{\leq} -\mu\|\mathbf{z}_{t+\frac{1}{2}} - \mathbf{z}^\star\|^2 - \frac{\gamma}{2}\|F(\mathbf{z}_t)\|^2 - \frac{\gamma}{2}\|F(\mathbf{z}_{t+\frac{1}{2}})\|^2 + \frac{\gamma}{2}\|F(\mathbf{z}_t) - F(\mathbf{z}_{t+\frac{1}{2}})\|^2 \\ &\overset{(1)}{\leq} -\mu\|\mathbf{z}_{t+\frac{1}{2}} - \mathbf{z}^\star\|^2 - \frac{\gamma}{2}\|F(\mathbf{z}_t)\|^2 + \frac{\gamma L^2}{2}\|\mathbf{z}_t - \mathbf{z}_{t+\frac{1}{2}}\|^2 \\ &= -\mu\|\mathbf{z}_{t+\frac{1}{2}} - \mathbf{z}^\star\|^2 - \frac{\gamma}{2}\|F(\mathbf{z}_t)\|^2 + \frac{\gamma^3 L^2}{2}\|F(\mathbf{z}_t)\|^2 \\ &\overset{(7)}{=} -\mu\|\mathbf{z}_{t+\frac{1}{2}} - \mathbf{z}^\star\|^2 + \left(\frac{\gamma^3 L^2}{2} - \frac{\gamma}{2}\right)\|F(\mathbf{z}_t)\|^2 \\ &\overset{(1)}{\leq} -\mu\|\mathbf{z}_{t+\frac{1}{2}} - \mathbf{z}^\star\|^2 + L^2\left(\frac{\gamma^3 L^2}{2} - \frac{\gamma}{2}\right)\|\mathbf{z}_t - \mathbf{z}^\star\|^2\end{aligned}$$

$$\|\mathbf{z}_{t+1} - \mathbf{z}^\star\|^2 \leq \|\mathbf{z}_t - \mathbf{z}^\star\|^2 + L^2\gamma^2\|\mathbf{z}_{t+\frac{1}{2}} - \mathbf{z}^\star\|^2 - 2\gamma\mu\|\mathbf{z}_{t+\frac{1}{2}} - \mathbf{z}^\star\|^2 + 2\gamma L^2\left(\frac{\gamma^3 L^2}{2} - \frac{\gamma}{2}\right)\|\mathbf{z}_t - \mathbf{z}^\star\|^2 \tag{9}$$

Let us assume $\gamma \leq \frac{2\mu}{L^2}$,now we can conclude,

$$-2\gamma\mu\|\mathbf{z}_{t+\frac{1}{2}} - \mathbf{z}^\star\|^2 + L^2\gamma^2\|\mathbf{z}_{t+\frac{1}{2}} - \mathbf{z}^\star\|^2 \leq 0$$

Using the above inequality we can rewrite (9),

$$\|\mathbf{z}_{t+1} - \mathbf{z}^\star\|^2 \leq \left(1 + \gamma^4 L^4 - L^2\gamma^2\right)\|\mathbf{z}_t - \mathbf{z}^\star\|^2$$

For driving a convergence proof we need to set γ such that $1 + \gamma^4 L^4 - L^2\gamma^2 < 1$, therefore by setting $\gamma = \frac{\mu}{2L^2}$ we have,

$$\|\mathbf{z}_{t+1} - \mathbf{z}^\star\|^2 \leq \left(1 + \frac{\mu^4}{16L^4} - \frac{\mu^2}{4L^2}\right)\|\mathbf{z}_t - \mathbf{z}^\star\|^2$$

By recursively replacing from $t = 0$ *to* $T-1$,

$$\|\mathbf{z}_T - \mathbf{z}^\star\|^2 \leq \left(1 + \frac{\mu^4}{16L^4} - \frac{\mu^2}{4L^2}\right)^T\|\mathbf{z}_0 - \mathbf{z}^\star\|^2$$

□

6.4 Proof of Theorem 4

Proof. We start with the update rule of PP:

$$\mathbf{z}_{t+1} = \mathbf{z}_t - \gamma F(\mathbf{z}_{t+1})$$

Then for the distance from the saddle point $\mathbf{z}^\star$ we have:

$$\begin{aligned}
\|\mathbf{z}_{t+1} - \mathbf{z}^\star\|^2 &= \|\mathbf{z}_t - \gamma F(\mathbf{z}_{t+1}) - \mathbf{z}^\star\|^2 \\
&= \|\mathbf{z}_t - \mathbf{z}^\star\|^2 + \gamma^2 \|F(\mathbf{z}_{t+1})\|^2 - 2\gamma(\mathbf{z}_t - \mathbf{z}^\star)^\top F(\mathbf{z}_{t+1}) \\
&= \|\mathbf{z}_t - \mathbf{z}^\star\|^2 + \gamma^2 \|F(\mathbf{z}_{t+1})\|^2 - 2\gamma(\mathbf{z}_t - \mathbf{z}^\star)^\top F(\mathbf{z}_{t+1}) - 2\gamma \mathbf{z}_{t+1}^\top F(\mathbf{z}_{t+1}) + 2\gamma \mathbf{z}_{t+1}^\top F(\mathbf{z}_{t+1}) \\
&= \|\mathbf{z}_t - \mathbf{z}^\star\|^2 - 2\gamma(\mathbf{z}_{t+1} - \mathbf{z}^\star)^\top F(\mathbf{z}_{t+1}) - 2\gamma(\mathbf{z}_t - \mathbf{z}_{t+1})^\top F(\mathbf{z}_{t+1}) + \gamma^2 \|F(\mathbf{z}_{t+1})\|^2
\end{aligned}$$

Using the update rule for PP we have:

$$\|\mathbf{z}_{t+1} - \mathbf{z}^\star\|^2 = \|\mathbf{z}_t - \mathbf{z}^\star\|^2 - 2\gamma(\mathbf{z}_{t+1} - \mathbf{z}^\star)^\top F(\mathbf{z}_{t+1}) - \|\mathbf{z}_{t+1} - \mathbf{z}_t\|^2$$

By rearranging the terms we have:

$$(\mathbf{z}_{t+1} - \mathbf{z}^\star)^\top F(\mathbf{z}_{t+1}) \leq \frac{1}{2\gamma}\|\mathbf{z}_t - \mathbf{z}^\star\|^2 - \frac{1}{2\gamma}\|\mathbf{z}_{t+1} - \mathbf{z}^\star\|^2$$

We lower bound the LHS using the fact that f is convex-concave:

$$\begin{aligned}
(\mathbf{z}_{t+1} - \mathbf{z}^\star)^\top F(\mathbf{z}_{t+1}) &\geq f(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}) - f(\mathbf{x}^\star, \mathbf{y}_{t+1}) + f(\mathbf{x}_{t+1}, \mathbf{y}^\star) - f(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}) \\
&= f(\mathbf{x}_{t+1}, \mathbf{y}^\star) - f(\mathbf{x}^\star, \mathbf{y}_{t+1})
\end{aligned}$$

After replacing in the main inequality we have:

$$f(\mathbf{x}_{t+1}, \mathbf{y}^\star) - f(\mathbf{x}^\star, \mathbf{y}_{t+1}) \leq \frac{1}{2\gamma}\|\mathbf{z}_t - \mathbf{z}^\star\|^2 - \frac{1}{2\gamma}\|\mathbf{z}_{t+1} - \mathbf{z}^\star\|^2$$

Summing from $t = 0, ..., T-1$ we get:

$$\frac{1}{T}\sum_{t=0}^{T-1} f(\mathbf{x}_{t+1}, \mathbf{y}^\star) - \frac{1}{T}\sum_{t=0}^{T-1} f(\mathbf{x}^\star, \mathbf{y}_{t+1}) \leq \frac{\|\mathbf{z}_0 - \mathbf{z}^\star\|^2}{2\gamma T}$$

Then we lower bound the LHS and we have:

$$f(\bar{\mathbf{x}}_{t+1}, \bar{\mathbf{y}}_{t+1}) - f(\mathbf{x}^\star, \mathbf{y}^\star) \leq \frac{\|\mathbf{z}_0 - \mathbf{z}^\star\|^2}{2\gamma T}$$

Where $\bar{\mathbf{x}}_{t+1} = \frac{1}{T}\sum_{t=0}^{T-1} \mathbf{x}_{t+1}$ and $\bar{\mathbf{y}}_{t+1} = \frac{1}{T}\sum_{t=0}^{T-1} \mathbf{y}_{t+1}$. □