Convergence of First-Order Methods for Saddle-Point Problems

Parham Yazdkhasti
Ali Zindari

Saarland University

INTRODUCTION

In this project, we provide the convergence rate for some first-order methods that were discussed in the course. More specifically, we will discuss the rates for:

- Proximal Point method for bi-linear objectives.
- Proximal Point method for convex-concave objectives.
- GDA for strongly monotone objectives.
- EG for strongly monotone objectives.

ASSUMPTIONS

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:

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

Theorem

Theorem 1 (PP Convergence for Bi-linear Games)

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 full-rank matrix. After T iterations of PP method on this class of function, we have the following rate:

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

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

- We have *exponential* (linear) convergence.
- We observe that $\frac{1}{1+\gamma^2\lambda_{\text{min}}(\mathbf{B}\mathbf{B}^\top)} \leq 1$. This inequality arises from the fact that $\mathbf{B}^\top\mathbf{B}$ is positive definite, ensuring that $\lambda_{\text{min}} > 0$.
- As suggested by the theorem, employing larger stepsizes enables faster convergence.

Proof

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

$$\begin{aligned}\mathbf{x}_{t+1} &= (\mathbf{I} + \gamma^2 \mathbf{B}\mathbf{B}^\top)^{-1}(\mathbf{x}_t - \gamma \mathbf{B}\mathbf{y}_t) = (\mathbf{I} + \gamma^2 \mathbf{B}\mathbf{B}^\top)^{-1}\mathbf{x}_t - \gamma(\mathbf{I} + \gamma^2 \mathbf{B}\mathbf{B}^\top)^{-1}\mathbf{B}\mathbf{y}_t \\ \mathbf{y}_{t+1} &= (\mathbf{I} + \gamma^2 \mathbf{B}^\top\mathbf{B})^{-1}(\mathbf{y}_t + \gamma \mathbf{B}^\top\mathbf{x}_t) = (\mathbf{I} + \gamma^2 \mathbf{B}^\top\mathbf{B})^{-1}\mathbf{y}_t + \gamma(\mathbf{I} + \gamma^2 \mathbf{B}^\top\mathbf{B})^{-1}\mathbf{B}^\top\mathbf{x}_t\end{aligned} \tag{5}$$

Where we define $\mathbf{P}_x := (\mathbf{I} + \gamma^2 \mathbf{B}\mathbf{B}^\top)^{-1}$ and $P_y := (\mathbf{I} + \gamma^2 \mathbf{B}^\top\mathbf{B})^{-1}$. The following equations are also crucial for the proof:

$$\begin{aligned}\mathbf{P}_x^2\mathbf{B} &= \mathbf{B}\mathbf{P}_y^2 \\ \mathbf{P}_y^2\mathbf{B}^\top &= \mathbf{B}^\top\mathbf{P}_x^2\end{aligned} \tag{6}$$

Then we compute the squared norm of both sides in (5):

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

After summing up the equations and with the use of (6) the blue terms cancel out.

$$\|\mathbf{x}_{t+1}\|^2 + \|\mathbf{y}_{t+1}\|^2 = \|\mathbf{P}_x\mathbf{x}_t\|^2 + \gamma^2\|\mathbf{P}_x\mathbf{B}\mathbf{y}_t\|^2 + \|\mathbf{P}_y\mathbf{y}_t\|^2 + \gamma^2\left\|\mathbf{P}_y\mathbf{B}^\top\mathbf{x}_t\right\|^2 \tag{7}$$

Proof

We can simplify the RHS using the following equations:

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

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

Now by replacing everything into (7) and unrolling the recursion we have:

$$\begin{aligned}
\|\mathbf{x}_{t+1}\|^2 + \|\mathbf{y}_{t+1}\|^2 &= \mathbf{x}_t^\top(\mathbf{I}+\gamma^2\mathbf{B}\mathbf{B}^\top)^{-1}\mathbf{x}_t + \mathbf{y}_t^\top(\mathbf{I}+\gamma^2\mathbf{B}^\top\mathbf{B})^{-1}\mathbf{y}_t \\
&\leq \frac{1}{1+\gamma^2\lambda_{min}(\mathbf{B}\mathbf{B}^\top)}(\|\mathbf{x}_t\|^2 + \|\mathbf{y}_t\|^2) \\
&\leq \left(\frac{1}{1+\gamma^2\lambda_{min}(\mathbf{B}\mathbf{B}^\top)}\right)^T(\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2)
\end{aligned}$$

Theorem 2 (PP Convergence for Monotone)

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:

$$\Big[f(\bar{\mathbf{x}}_t, \mathbf{y}^\star) - f(\mathbf{x}^\star, \mathbf{y}^\star)\Big] + \Big[f(\mathbf{x}^\star, \mathbf{y}^\star) - f(\mathbf{x}^\star, \bar{\mathbf{y}}_t)\Big] \leq \frac{\|\mathbf{z}_0 - \mathbf{z}^\star\|^2}{2\gamma T}$$

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

- The rate holds for average iterate not last iterate.

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:

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

Proof

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 $t = 0, ..., T-1$ and adding and subtracting $f(\mathbf{x}^\star, \mathbf{y}^\star)$ we get:

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

Then we lower bound the LHS and we have:

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

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

Theorem 3 (GDA Convergence for Strongly-Monotone)

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:

$$\|\mathbf{z}_T - \mathbf{z}^\star\|^2 \leq exp\left(\frac{-T\mu^2}{L^2}\right)\|\mathbf{z}_0 - \mathbf{z}^\star\|^2 \tag{8}$$

- The distance between the Tth iteration and the equilibrium is *exponentially* decreasing.
- μ and L are problem-dependent parameters and they effect the convergence speed.
- The theorem suggests that the best rate is achieved when $L = \mu$.

Proof

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

For bounding A Lipschitz Continuity is used,

$$\begin{aligned}\|F(\mathbf{z}_t)\|^2 &= \|F(\mathbf{z}_t)-F(\mathbf{z}^\star)\|^2\\ &\le \gamma^2 L^2\|\mathbf{z}_t-\mathbf{z}^\star\|^2\end{aligned}$$

and for bounding B we use strong monotone property,

$$2\gamma\langle \mathbf{z}_t-\mathbf{z}^\star, F(\mathbf{z}_t)-F(\mathbf{z}^\star)\rangle \ge 2\gamma\mu\|\mathbf{z}_t-\mathbf{z}^\star\|^2$$

Then we put everything together and choosing $\gamma \le \frac{\mu}{L^2}$,

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

Theorem

Theorem 4 (EG Convergence for Strongly-Monotone)

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:

$$\|\mathbf{z}_T - \mathbf{z}^\star\|^2 \leq \left(1 + \frac{\mu^4}{16L^4} - \frac{\mu^2}{4L^2}\right)^T \|\mathbf{z}_0 - \mathbf{z}^\star\|^2 \tag{9}$$

Proof

EG update rule,

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

We start by expanding the distance term,

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

For bounding A Lipschitz Continuity is used,

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

PROOF

Now we simplify term B.

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

where $-a^T b = -\frac{1}{2}\|a\|^2 - \frac{1}{2}\|b\|^2 + \frac{1}{2}\|a-b\|^2$ is used for the third equality.

Proof

Then we put everything together and choosing $\gamma \leq \frac{\mu}{2L^2}$,

$$
\begin{aligned}
&\|\mathbf{z}_{t+1} - \mathbf{z}^\star\|^2 \\
&\leq \|\mathbf{z}_t - \mathbf{z}^\star\|^2 + L^2\gamma^2\|\mathbf{z}_{t+\frac{1}{2}} - \mathbf{z}^\star\|^2 - 2\gamma\mu\|\mathbf{z}_{t+\frac{1}{2}} - \mathbf{z}^\star\|^2 + 2\gamma L^2\left(\frac{\gamma^3 L^2}{2} - \frac{\gamma}{2}\right)\|\mathbf{z}_t - \mathbf{z}^\star\|^2 \\
&\leq \left(1 + \gamma^4 L^4 - L^2\gamma^2\right)\|\mathbf{z}_t - \mathbf{z}^\star\|^2 \\
&\leq \underbrace{\left(1 + \frac{\mu^4}{16L^4} - \frac{\mu^2}{4L^2}\right)}_{\leq 1}\|\mathbf{z}_t - \mathbf{z}^\star\|^2
\end{aligned}
$$