MASTERARBEIT | MASTER'S THESIS

Titel | Title

Convergence rates of extragradient type methods for solving smooth convex-concave problem

verfasst von | submitted by

Tuan Minh Le BSc

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2024

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 821
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Mathematik
Betreut von | Supervisor:	Univ.-Prof. Dr. Radu Ioan Bot Privatdoz.

Contents

1 Introduction

Given a function $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$, we consider methods for finding the saddle points of the following problem:

$$\min_{u \in \mathbb{R}^n} \max_{v \in \mathbb{R}^m} f(u, v). \tag{1.1}$$

We assume that f is convex-concave i.e for any $v \in \mathbb{R}^m$, $f(\cdot, v)$ is a convex function and for any $u \in \mathbb{R}^n$, $f(u, \cdot)$ is a concave function. Furthermore f is everywhere differentiable. This formulation appears in several areas such as robust optimization [BTEGN09] or in machine learning in generative adversarial networks (GANs) [ACB17] [GPAM$^+$14].

The solutions we are looking for for this class of problem are saddle points. The extragradient (EG) method introduced by G.M. Koperlevich in 1976 [Kor76] is a well known algorithm designed to find a saddle point. In 2020 Ozdaglar has shown [MOP20] that the optimistic gradient descent-ascent and extragradient method have the convergence rate of $\mathcal{O}(1/k)$ in terms of the gap function of the averaged iterate and saddle point. Our goal in this thesis is to use similar technique like in [MOP20] to show that the algorithms introduced in the paper by Quoc Tran-Dinh [TD23] (page 12-13) can archive similar rates.
The content of this thesis is follow: we'll go briefly over the necessary definitions and properties of the problem in Section 2. In Section 3 we'll look at the key techniques introduced by Ozdaglar. Section 4 will be about the algorithms shown in [TD23] which are variants of classical extragradient method and the analysis of said algorithms will be done in the Sections 5 and 6. Section 7 contains small numerical experiments with these algorithms.

2 Preliminaries

Definition 2.1. An operator $F : \mathbb{R}^d \to \mathbb{R}^d$is called L-Lipschitz continuous if for all $x, y \in \mathbb{R}^d$ we have:

$$\|F(x) - F(y)\| \leq L\|x - y\|.$$

Definition 2.2. The pair (u_*, v_*) is called saddle point of problem (1.1) if for any $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^m$ we have:

$$f(u_*, v) \leq f(u_*, v_*) \leq f(u, v_*).$$

Definition 2.3. A function $\phi : \mathbb{R}^d \to \mathbb{R}$ is convex if for all $x, y \in \mathbb{R}^d$ and $\lambda \in [0, 1]$ we have:

$$\phi(\lambda x + (1 - \lambda)y) \leq \lambda\phi(x) + (1 - \lambda)\phi(y).$$

The function ϕ is concave if $-\phi$ is convex

The following proposition is one of the most used properties of convex functions.

Proposition 2.4. *A differentiable function $\phi : \mathbb{R}^d \to \mathbb{R}$ is convex if and only if for all $x, y \in \mathbb{R}^d$ we have:*

$$\phi(y) - \phi(x) \geq \nabla\phi(x)^\top(y - x).$$

In this thesis we denote the k-th iterate by u_k, v_k and the average iterate by $\hat{u}_k, \hat{v}_k$, which are defined as:

$$\hat{u}_k := \frac{1}{k}\sum_{i=1}^{k} u_i \quad \hat{v}_k := \frac{1}{k}\sum_{i=1}^{k} v_i \quad \forall k \geq 1.$$

We also define the vector $z = (u, v) \in \mathbb{R}^n \times \mathbb{R}^m$ and the operator $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n \times \mathbb{R}^m$ as:

$$F(z) = (\nabla_u f(u, v); -\nabla_v f(u, v)). \tag{2.1}$$

The following Lemma show give us the properties of F:

Lemma 2.5. *Assume that the gradient $\nabla_u f$ and $\nabla_v f$ satisfy:*

$$\begin{aligned}
\|\nabla_u f(u_1, v) - \nabla_u f(u_2, v)\| &\leq L_{uu}\|u_1 - u_2\| \quad \forall v, u_1, u_2, \\
\|\nabla_u f(u, v_1) - \nabla_u f(u, v_2)\| &\leq L_{uv}\|v_1 - v_2\| \quad \forall u, v_1, v_2, \\
\|\nabla_v f(u_1, v) - \nabla_v f(u_2, v)\| &\leq L_{vu}\|u_1 - u_2\| \quad \forall v, u_1, u_2, \\
\|\nabla_v f(u, v_1) - \nabla_v f(u, v_2)\| &\leq L_{vv}\|v_1 - v_2\| \quad \forall u, v_1, v_2.
\end{aligned}$$

For some positive numbers $L_{uu}, L_{uv}, L_{vu}, L_{vv}$. Then the operator $F(z) = (\nabla_u f(u, v); -\nabla_v f(u, v))$ is L-Lipschitz continuous with $L = 2 \times max\{L_{uu}, L_{uv}, L_{vu}, L_{vv}\}$.

Lemma 2.6. *Let F defined in* (2.1) *and assume that $\mathcal{Z}^*$ the set of solutions of* (1.1) *is not empty. Then:*

- *F is a monotone operator, i.e for all $z_1, z_2 \in \mathbb{R}^n \times \mathbb{R}^m$ we have:*

$$(F(z_1) - F(z_2))^\top (z_1 - z_2) \geq 0.$$

- *For any $z^* = (u^*, v^*) \in \mathcal{Z}^*$ we have $F(z^*) = 0$.*

Proof. Using the definition of F and z, for all $z_1 = (u_1, v_1), z_2 = (u_2.v_2) \in \mathbb{R}^n \times \mathbb{R}^m$ we have:

$$\begin{aligned}(F(z_1) - F(z_2))^\top (z_1 - z_2) = &(\nabla_u f(u_1, v_1) - \nabla_u f(u_2, v_2))^\top (u_1 - u_2) \\ &- (\nabla_v f(u_1, v_1) - \nabla_v f(u_2, v_2))^\top (u_1 - u_2).\end{aligned}$$

Since $f(\cdot, v)$ is convex function and $f(u, \cdot)$ is concave we obtain:

$$\begin{aligned}\nabla_u f(u_1, v_1)^\top (u_1 - u_2) &\geq f(u_1, v_1) - f(u_2, v_1), \\ -\nabla_u f(u_2, v_2)^\top (u_1 - u_2) &\geq f(u_2, v_2) - f(u_1, v_2), \\ -\nabla_v f(u_1, v_1)^\top (v_1 - v_2) &\geq -f(u_1, v_1) + f(u_1, v_2), \\ \nabla_v f(u_2, v_2)^\top (v_1 - v_2) &\geq f(u_2, v_1) - f(u_2, v_2).\end{aligned}$$

Adding all 4 equation give us $(F(z_1) - F(z_2))^\top (z_1 - z_2) \geq 0$.
For the proof of the second part, we use the fact that if x is the minimum of a convex function ϕ then $\nabla\phi(x) = 0$ and similarly $\nabla\psi(y) = 0$ if y is maximum of concave function ψ. □

The above Lemma is well known as shown in [MOP20] and [Nem04], however I couldn't find the proof so I prove it for completeness. The following result are also well known, as mentioned in [MOP20]:

Proposition 2.7. *In the definition of F and z, suppose that f is continuously differentiable and a convex-concave function. Furthermore assume that a solution for* (1.1) *exists. Then for any $z = (u, v) \in \mathbb{R}^n \times \mathbb{R}^m$ and a sequence $\{(u_k, v_k)\}_{k\geq 1}$ with $z_k = (u_k, v_k)$, the following holds for all $N \geq 1$:*

$$f(\hat{u}_N, v) - f(u, \hat{v}_N) \leq \frac{1}{N} \sum_{k=1}^{N} F(z_k)^\top (z_k - z).$$

Proof. Using the definition of F we can write for all $N \geq 1$:

$$\begin{aligned}\frac{1}{N}\sum_{k=1}^{N} F(z_k)^\top (z_k - z) &= \frac{1}{N}\sum_{k=1}^{N}(\nabla_x f(u_k, v_k)^\top (u_k - u) + \nabla_y f(u_k, v_k)^\top (v - v_k)) \\ &\geq \frac{1}{N}\sum_{k=1}^{N} (f(u_k, v_k) - f(u, v_k) + f(u_k, v) - f(u_k, v_k)) \qquad (2.2) \\ &= \frac{1}{N}\sum_{k=1}^{N}(f(u_k, v) - f(u, v_k)).\end{aligned}$$

Here the inequality follow from property of f being convex-concave. Using again this property of f we get:

$$\frac{1}{N}\sum_{k=1}^{N} f(u_k, v) \geq f(\hat{u}_N, v) \quad \textit{and} \quad \frac{1}{N}\sum_{k=1}^{N} f(u, v_k) \leq f(u, \hat{v}_N) \quad \forall N \geq 1.$$

This implies the statement of the proposition. $\square$

3 Proximal Point method with Error

One of the classical approaches to solve the saddle point problem (1.1) is the proximal point method. This method is introduced in [Mar70] and studied in [Roc76]. The proximal point method generates the iterate (u_{k+1}, v_{k+1}) which is defined as the unique solution to the saddle point problem:

$$\min_{u\in\mathbb{R}^n} \max_{v\in\mathbb{R}^m} \left\{ f(u,v) + \frac{1}{2\eta}\|u-u_k\|^2 + \frac{1}{2\eta}\|v-v_k\|^2 \right\}. \tag{3.1}$$

It can be shown that if for every $k \geq 1$, (u_{k+1}, v_{k+1}) is the solution of (3.1), then u_{k+1} and v_{k+1} satisfy:

$$\begin{aligned} u_{k+1} &= \arg\min_{u\in\mathbb{R}^n}\{f(u, v_{k+1}) + \frac{1}{2\eta}\|u-u_k\|^2\}, \\ v_{k+1} &= \arg\max_{v\in\mathbb{R}^m}\{f(u_{k+1}, v) + \frac{1}{2\eta}\|v-v_k\|^2\}. \end{aligned}$$

Since the first function is strongly convex and the second one is strongly concave, the update of proximal point method for saddle point problem can be written as

$$\begin{aligned} u_{k+1} &= u_k - \eta\nabla_u f(u_{k+1}, v_{k+1}), \\ v_{k+1} &= v_k + \eta\nabla_v f(u_{k+1}, v_{k+1}). \end{aligned}$$

Using our definition of $z = (u, v)$ and operator F the update for proximal point method can be rewritten as:

$$z_{k+1} = z_k - \eta F(z_{k+1}) \quad \forall k \geq 1.$$

Recalling Lemma 2.6, finding a saddle point $z_* = (u_*, v_*)$ would be equivalent to finding $z \in \mathbb{R}^d$ such that $F(z) = 0$, where $d = n + m$.
As shown in [Nem04] the proximal point method archives a sublinear rate of $\mathcal{O}(1/k)$ for convex minimization and solving monotone variational inequalities. Ozdaglar [MOP20] then shows that by interpreting ODGA and EG as approximate of proximal point method, the gap between the function value at average iterate $(\hat{u}_k, \hat{v}_k)$ and function value at the saddle points approaches 0 at the rate of $\mathcal{O}(1/k)$. The following proposition establishes a relation for the iterates of a proximal point method with error.

Proposition 3.1. *Consider the sequence of iterate $\{z_k\}_{k\geq 1} \in \mathbb{R}^d$, generated by the following update:*

$$z_{k+1} = z_k - \eta F(z_{k+1}) + \epsilon_k, \tag{3.2}$$

here $F : \mathbb{R}^d \to \mathbb{R}^d$ is a monotone and L-Lipschitz continuous operator, $\epsilon_k \in \mathbb{R}^d$ arbitrary *vector, η a positive constant. Then for $z \in \mathbb{R}^d$ and for each $k \geq 1$ we have:*

$$F(z_{k+1})^\top (z_{k+1}-z) = \frac{1}{2\eta}\|z_k-z\|^2 - \frac{1}{2\eta}\|z_{k+1}-z\|^2 - \frac{1}{2\eta}\|z_{k+1}-z_k\|^2 + \frac{1}{\eta}\epsilon_k^\top (z_{k+1}-z). \quad (3.3)$$

Proof. Using the update scheme in (3.2) we have for $k \geq 1$:

$$\begin{aligned} \|z_{k+1} - z\|^2 = \|z_k - z\|^2 + \eta^2 \|F(z_{k+1})\|^2 + \|\epsilon_k\|^2 \\ - 2\eta(z_k - z)^\top F(z_{k+1}) + 2\epsilon_k^\top (z_k - z - \eta F(z_{k+1})) \\ = \|z_k - z\|^2 + \eta^2 \|F(z_{k+1})\|^2 + \|\epsilon_k\|^2 + 2\epsilon_k^\top (z_k - z - \eta F(z_{k+1})) \\ - 2\eta(z_{k+1} - z)^\top F(z_{k+1}) - 2\eta(z_k - z_{k+1})^\top F(z_{k+1}). \end{aligned} \quad (3.4)$$

Replacing $F(z_{k+1})$ with $\frac{1}{\eta}(-z_{k+1} + z_k + \epsilon_k)$ it yields:

$$\begin{aligned} \|z_{k+1} - z\|^2 = \|z_k - z\|^2 + \|-z_{k+1} + z_k + \epsilon_k\|^2 + \|\epsilon_k\|^2 - 2\eta(z_{k+1} - z)^\top F(z_{k+1}) \\ + 2(z_k - z_{k+1})^\top (z_{k+1} - z_k - \epsilon_k) + 2\epsilon_k^\top (z_{k+1} - z - \epsilon_k). \end{aligned} \quad (3.5)$$

Rearranging the last 2 terms, we obtain:

$$\begin{aligned} &2(z_k - z_{k+1})^\top (z_{k+1} - z_k - \epsilon_k) + 2\epsilon_k^\top (z_{k+1} - z - \epsilon_k) \\ &= -\|-z_{k+1} + z_k + \epsilon_k\|^2 - \|z_k - z_{k+1}\|^2 + \epsilon_k^\top ((z_{k+1} - z_k) \\ &\quad + (z_k - z) + (z_{k+1} - z - \epsilon_k)) \\ &= -\|-z_{k+1} + z_k + \epsilon_k\|^2 - \|z_k - z_{k+1}\|^2 - \|\epsilon_k\|^2 + 2\epsilon_k^\top (z_{k+1} - z). \end{aligned} \quad (3.6)$$

Hence for all $k \geq 1$:

$$\|z_{k+1} - z\|^2 = \|z_k - z\|^2 - 2\eta(z_{k+1} - z)^\top F(z_{k+1}) - \|z_k - z_{k+1}\|^2 + 2\epsilon_k^\top (z_{k+1} - z). \quad (3.7)$$

Rearranging the term we get

$$\begin{aligned} F(z_{k+1})^\top (z_{k+1} - z) = &\frac{1}{2\eta}\|z_k - z\|^2 - \frac{1}{2\eta}\|z_{k+1} - z\|^2 \\ &- \frac{1}{2\eta}\|z_{k+1} - z_k\|^2 + \frac{1}{\eta}\epsilon_k^\top (z_{k+1} - z). \end{aligned}$$

□

4 Generalized Extra gradient methods

In this section we will look at a more generalized version of EG and ODGA as seen in [TD23] page 12 and 13. Using a similar proof like in [MOP20], we will later show that these classes of algorithm also have the rate of $\mathcal{O}(1/k)$ for the gap between the function value at average iterate $(\hat{u}_k, \hat{v}_k)$ and function value at the saddle points.

Definition 4.1. Extra gradient The class of extragradient methods for solving:

$$x \in zer(F) = \{x \in \mathbb{R}^d | F(x) = 0\}, \tag{NE}$$

starting from $x_0 \in \mathbb{R}^d$ we update:

$$\begin{cases} y_k := x_k - \frac{\eta}{\beta} d_k, \\ x_{k+1} := x_k - \eta F(y_k), \end{cases} \tag{EG}$$

where $\eta > 0$ is the step size and $\beta \in (0, 1]$ is scaling factor and d_k is given as $d_k := F(x_k)$ or $d_k := F(y_{k-1})$

Lemma 4.2. *Let $\{(x_k, y_k)\}_{k\geq 1}$ be generated by* (EG)*. Then for any $\gamma > 0$ and $\hat{x} \in \mathbb{R}^d$ we have:*

$$\begin{aligned} \|x_{k+1} - \hat{x}\|^2 \leq \|x_k - \hat{x}\|^2 - \beta\|y_k - x_k\|^2 + \frac{\eta^2}{\gamma}\|F(y_k) - d_k\|^2 - 2\eta F(y_k)^\top (y_k - \hat{x}) \\ - (\beta - \gamma)\|y_k - x_{k+1}\|^2 - (1 - \beta)\|x_k - x_{k+1}\|^2. \end{aligned} \tag{4.1}$$

Proof. Using $x_{k+1} - x_k = -\eta F(y_k)$ we have for all $\hat{x} \in \mathbb{R}^n \times \mathbb{R}^m$:

$$\begin{aligned} \|x_{k+1} - \hat{x}\|^2 &= \|x_k - \hat{x}\|^2 + 2(x_{k+1} - x_k)^\top (x_k - \hat{x}) + \|x_{k+1} - x_k\|^2 \\ &= \|x_k - \hat{x}\|^2 + 2(x_{k+1} - x_k)^\top (x_{k+1} - \hat{x}) - \|x_{k+1} - x_k\|^2 \\ &= \|x_k - \hat{x}\|^2 - 2\eta F(y_k)^\top (x_{k+1} - \hat{x}) - \|x_{k+1} - x_k\|^2. \end{aligned} \tag{4.2}$$

Next using $\eta d_k = \beta(x_k - y_k)$, Cauchy-Schwarz inequality and the identity

$$2(x_{k+1} - y_k)^\top (x_k - y_k) = \|x_k - y_k\|^2 + \|x_{k+1} - y_k\|^2 - \|x_{k+1} - x_k\|^2,$$

and the fact that for any $\gamma > 0$: $2|x^\top y| \leq \gamma\|x\|^2 + \frac{1}{\gamma}\|y\|^2$, we can conclude for all $k \geq 1$

and $\gamma > 0$ that:

$$\begin{aligned} 2\eta F(y_k)^\top(x_{k+1} - \hat{x}) &= 2\eta F(y_k)^\top(y_k - \hat{x}) + 2\eta(F(y_k) - d_k)^\top(x_{k+1} - y_k) \\ &\quad + 2\eta d_k^\top(x_{k+1} - y_k) \\ &\geq 2\eta F(y_k)^\top(y_k - \hat{x}) - \frac{\eta^2}{\gamma}\|F(y_k) - d_k\|^2 - \gamma\|x_{k+1} - y_k\|^2 \\ &\quad + \beta[\|x_k - y_k\|^2 + \|x_{k+1} - y_k\|^2 - \|x_{k+1} - x_k\|^2] \\ &= 2\eta F(y_k)^\top(y_k - \hat{x}) + \beta\|y_k - x_k\|^2 - \frac{\eta^2}{\gamma}\|F(y_k) - d_k\|^2 \\ &\quad + (\beta - \gamma)\|x_{k+1} - y_k\|^2 - \beta\|x_{k+1} - x_k\|^2. \end{aligned} \tag{4.3}$$

□

Lemma 4.3. *Let F be a monotone and L-Lipschitz continuous operator. Let $\{(x_k, y_k)\}_{k\geq 1}$ be the sequence generated by* (EG)*. Then for any $\omega > 0$ and $k \geq 1$ we have:*

$$\begin{aligned} \|F(x_{k+1})\|^2 \leq \|F(x_k)\|^2 - \|F(y_k) - F(x_k)\|^2 + \frac{\eta^2\omega L^2}{\beta^2}\|\beta F(y_k) - d_k\|^2 \\ - (\omega - 1)\|F(x_{k+1}) - F(y_k)\|^2. \end{aligned} \tag{4.4}$$

Proof. Since F is monotone, using the relation in second line of EG:

$$x_{k+1} - x_k = -\eta F(y_k),$$

we get:

‘

$$\begin{aligned} 0 &\leq (F(x_{k+1}) - F(x_k))^\top(x_{k+1} - x_k) \\ &= (F(x_{k+1}) - F(x_k))^\top(-\eta F(y_k)) \\ &= (F(x_k) - F(x_{k+1}))^\top(\eta F(y_k)). \end{aligned} \tag{4.5}$$

Hence we have

$$\begin{aligned} 0 &\leq 2(F(x_k) - F(x_{k+1}))^\top F(y_k) \\ &= \|F(x_k)\|^2 - \|F(x_{k+1})\|^2 + \|F(y_k) - F(x_{k+1})\|^2 - \|F(x_k) - F(y_k)\|^2. \end{aligned} \tag{4.6}$$

Moving $\|F(x_{k+1})\|^2$ to left hand side, we get:

$$\begin{aligned} \|F(x_{k+1})\|^2 \leq \|F(x_k)\|^2 + \omega\|F(y_k) - F(x_{k+1})\|^2 - (\omega - 1)\|F(y_k) - F(x_{k+1})\|^2 \\ - \|F(x_k) - F(y_k)\|^2. \end{aligned} \tag{4.7}$$

Using the fact that F is L-Lipschitz continuous and $y_k - x_{k+1} = \eta F(y_k) - \frac{\eta}{\beta} d_k$, we obtain:

$$\begin{aligned}
\|F(x_{k+1})\|^2 &\leq \|F(x_k)\|^2 + \omega L^2 \|y_k - x_{k+1}\|^2 - (\omega - 1)\|F(y_k) - F(x_{k+1})\|^2 \\
&\quad - \|F(x_k) - F(y_k)\|^2 \\
&= \|F(x_k)\|^2 + \omega L^2 \|\eta F(y_k) - \frac{\eta}{\beta} d_k\|^2 - (\omega - 1)\|F(y_k) - F(x_{k+1})\|^2 \\
&\quad - \|F(x_k) - F(y_k)\|^2 \\
&= \|F(x_k)\|^2 - \|F(y_k) - F(x_k)\|^2 + \frac{\eta^2 \omega L^2}{\beta^2} \|\beta F(y_k) - d_k\|^2 \\
&\quad - (\omega - 1)\|F(x_{k+1}) - F(y_k)\|^2.
\end{aligned} \tag{4.8}$$

□

5 Forward Reflected Backward Splitting

Letting $d_k = F(y_{k-1})$ in (EG), we have $x_k = y_k + \frac{\eta}{\beta} y_{k-1}$. Substituting to the second line of (EG) we get $x_{k+1} = y_k + \frac{\eta}{\beta} y_{k-1} - \eta F(y_k)$. We can then eliminate x_k to obtain the update of y_k for any $k \geq 0$ and initial points y_{-1} and y_0:

$$y_{k+1} := y_k - \frac{\eta}{\beta}((1+\beta)F(y_k) - F(y_{k-1})). \tag{FRBS}$$

This method is also called Popov's method to solve (NE). By setting $\beta = 1$ we get the Optimistic gradient descent ascent (OGDA) in [MOP20]

Theorem 5.1. *Let the sequence $\{y_k\}_{k\geq 1}$ be generated by* (FRBS) *with $y_{-1} = y_0$ and F a monotone and L-Lipschitz continuous operator. Let $\beta \in (0,1]$ and $\eta \leq \frac{\beta}{2L}$. Then:*
a) For any y and $k \geq 0$ we get:

$$\begin{aligned} F(y_{k+1})^\top (y_{k+1} - y) \leq & \frac{1}{2\eta}\|y_k - y\|^2 - \frac{1}{2\eta}\|y_{k+1} - y\|^2 + \frac{L}{2\beta}\|y_k - y_{k-1}\|^2 \\ & + (F(y_{k+1}) - F(y_k))^\top (y_{k+1} - y) - \frac{L}{2\beta}\|y_{k+1} - y_k\|^2 \\ & - \frac{1}{\beta}(F(y_k) - F(y_{k-1}))^\top (y_k - y). \end{aligned} \tag{5.1}$$

b) The iterates $\{y_k\}_{k\geq 1}$ stay within the compact set $\mathcal{D}$ defined as:

$$\mathcal{D} := \left\{ y \in \mathbb{R}^d : \|y - y_*\|^2 \leq \frac{2}{\beta(2-\beta)}\|y_0 - y*\|^2 \right\},$$

where $y_* \in zer(F)$

Proof. a) We rewrite the update in (FRBS) in the form $y_{k+1} = y_k - \eta F(y_{k+1}) + \epsilon_k$, where $\epsilon_k = \eta(F(y_{k+1}) - F(y_k)) - \frac{\eta}{\beta}(F(y_k) - F(y_{k-1}))$. Substituting this into (3.3) we obtain for all $k \geq 0$:

$$\begin{aligned}
&F(y_{k+1})^\top (y_{k+1} - y) \\
&= \frac{1}{2\eta}\|y_k - y\|^2 - \frac{1}{2\eta}\|y_{k+1} - y\|^2 - \frac{1}{2\eta}\|y_{k+1} - y_k\|^2 + \frac{1}{\eta}\epsilon_k^\top (y_{k+1} - y) \\
&= \frac{1}{2\eta}\|y_k - y\|^2 - \frac{1}{2\eta}\|y_{k+1} - y\|^2 - \frac{1}{2\eta}\|y_{k+1} - y_k\|^2 \\
&\quad + (F(y_{k+1}) - F(y_k))^\top (y_{k+1} - y) - \frac{1}{\beta}(F(y_k) - F(y_{k-1}))^\top (y_{k+1} - y) \\
&= \frac{1}{2\eta}\|y_k - y\|^2 - \frac{1}{2\eta}\|y_{k+1} - y\|^2 - \frac{1}{2\eta}\|y_{k+1} - y_k\|^2 \\
&\quad + (F(y_{k+1}) - F(y_k))^\top (y_{k+1} - y) - \frac{1}{\beta}(F(y_k) - F(y_{k-1}))^\top (y_k - y) \\
&\quad + \frac{1}{\beta}(F(y_k) - F(y_{k-1}))^\top (y_k - y_{k+1}).
\end{aligned} \tag{5.2}$$

Using Cauchy Schwarz inequality, the fact that F is L-Lipschitz continuous and $ab \leq \frac{1}{2}(a^2 + b^2)$ we then get:

$$\begin{aligned}
\frac{1}{\beta}(F(y_k) - F(y_{k-1}))^\top (y_k - y_{k+1}) &\leq \frac{1}{\beta}\|F(y_k) - F(y_{k-1})\|\|y_k - y_{k+1}\| \\
&\leq \frac{L}{\beta}\|y_k - y_{k-1}\|\|y_k - y_{k+1}\| \\
&\leq \frac{L}{2\beta}\|y_k - y_{k-1}\|^2 + \frac{L}{2\beta}\|y_k - y_{k+1}\|^2.
\end{aligned} \tag{5.3}$$

Substituting back in (5.2) we get:

$$\begin{aligned}
&F(y_{k+1})^\top (y_{k+1} - y) \\
&\leq \frac{1}{2\eta}\|y_k - y\|^2 - \frac{1}{2\eta}\|y_{k+1} - y\|^2 - \frac{1}{2\eta}\|y_{k+1} - y_k\|^2 \\
&\quad + (F(y_{k+1}) - F(y_k))^\top (y_{k+1} - y) - \frac{1}{\beta}(F(y_k) - F(y_{k-1}))^\top (y_k - y) \\
&\quad + \frac{L}{2\beta}\|y_k - y_{k-1}\|^2 + \frac{L}{2\beta}\|y_k - y_{k+1}\|^2 \\
&\leq \frac{1}{2\eta}\|y_k - y\|^2 - \frac{1}{2\eta}\|y_{k+1} - y\|^2 + (F(y_{k+1}) - F(y_k))^\top (y_{k+1} - y) \\
&\quad - \frac{1}{\beta}(F(y_k) - F(y_{k-1}))^\top (y_k - y) + \frac{L}{2\beta}\|y_k - y_{k-1}\|^2 - \frac{L}{2\beta}\|y_{k+1} - y_k\|^2.
\end{aligned} \tag{5.4}$$

Here the last inequality holds because $\eta \leq \frac{\beta}{2L}$

b) Taking the sum over $k = 0$ to $N - 1$ we get for all $N \geq 1$:

$$\begin{aligned}\sum_{k=0}^{N-1} F(y_{k+1})^\top (y_{k+1} - y) \leq & \frac{1}{2\eta}\|y_0 - y\|^2 - \frac{1}{2\eta}\|y_N - y\|^2 - \frac{L}{2\beta}\|y_N - y_{N-1}\|^2 \\ & + \frac{L}{2\beta}\|y_0 - y_{-1}\|^2 + (F(y_N) - F(y_{N-1}))^\top (y_N - y) \\ & - \frac{1}{\beta}(F(y_0) - F(y_{-1}))^\top (y_0 - y) \\ & + \sum_{k=1}^{N-1} (1 - \frac{1}{\beta})(F(y_k) - F(y_{k-1}))^\top (y_k - y).\end{aligned} \tag{5.5}$$

Since $y_{-1} = y_0$ we get:

$$\begin{aligned}\sum_{k=0}^{N-1} F(y_{k+1})^\top (y_{k+1} - y) \leq & \frac{1}{2\eta}\|y_0 - y\|^2 - \frac{1}{2\eta}\|y_N - y\|^2 - \frac{L}{2\beta}\|y_N - y_{N-1}\|^2 \\ & + (F(y_N) - F(y_{N-1}))^\top (y_N - y) \\ & + \sum_{k=1}^{N-1} (1 - \frac{1}{\beta})(F(y_k) - F(y_{k-1}))^\top (y_k - y).\end{aligned} \tag{5.6}$$

We can rewrite $y_{k+1} := y_k - \frac{\eta}{\beta}((1+\beta)F(y_k) - F(y_{k-1}))$ in the form:

$$F(y_k) - F(y_{k-1}) = \frac{\beta}{\eta}(y_k - y_{k+1}) - \beta F(y_k).$$

Hence for any $k \geq 1$:

$$(F(y_k) - F(y_{k-1}))^\top (y_k - y) = \left(\frac{\beta}{\eta}(y_k - y_{k+1}) - \beta F(y_k)\right)^\top (y_k - y).$$

Applying this to (5.6) to get:

$$\begin{aligned}
\sum_{k=0}^{N-1} F(y_{k+1})^\top (y_{k+1}-y) \leq{}& \frac{1}{2\eta}\|y_0-y\|^2 - \frac{1}{2\eta}\|y_N-y\|^2 - \frac{L}{2\beta}\|y_N-y_{N-1}\|^2 \\
&+ (F(y_N)-F(y_{N-1}))^\top (y_N-y) \\
&+ \sum_{k=1}^{N-1}(1-\frac{1}{\beta})\left(\frac{\beta}{\eta}(y_k-y_{k-1}) - \beta F(y_k)\right)^\top (y_k-y) \\
={}& \frac{1}{2\eta}\|y_0-y\|^2 - \frac{1}{2\eta}\|y_N-y\|^2 - \frac{L}{2\beta}\|y_N-y_{N-1}\|^2 \\
&+ (F(y_N)-F(y_{N-1}))^\top (y_N-y) \\
&+ \sum_{k=0}^{N-2}(1-\beta)(F(y_{k+1})^\top (y_{k+1}-y) \\
&+ \frac{1}{\eta}\sum_{k=1}^{N-1}(\beta-1)(y_k-y_{k+1})^\top (y_k-y).
\end{aligned} \tag{5.7}$$

Bring $\sum_{k=0}^{N-2}(1-\beta)(F(y_{k+1})^\top (y_{k+1}-y)$ to the other side to obtain:

$$\begin{aligned}
F(y_N)^\top (y_N-y) + \beta\sum_{k=0}^{N-2} F(y_{k+1})^\top (y_{k+1}-y) \leq{}& \frac{1}{2\eta}\|y_0-y\|^2 - \frac{1}{2\eta}\|y_N-y\|^2 \\
&- \frac{L}{2\beta}\|y_N-y_{N-1}\|^2 \\
&+ (F(y_N)-F(y_{N-1}))^\top (y_N-y) \\
&+ \frac{\beta-1}{\eta}\sum_{k=1}^{N-1}(y_k-y_{k+1})^\top (y_k-y).
\end{aligned} \tag{5.8}$$

Using the relation $2(y_k-y_{k+1})^\top (y_k-y) = \|y_k-y_{k+1}\|^2 + \|y_k-y\|^2 - \|y_{k+1}-y\|^2$ we get:

$$\frac{1}{\eta}\sum_{k=1}^{N-1}(\beta-1)(y_k-y_{k+1})^\top (y_k-y) = \frac{\beta-1}{2\eta}(\|y_1-y\|^2 - \|y_N-y\|^2 + \sum_{k=1}^{N-1}\|y_k-y_{k+1}\|^2).$$

Substitute it to (5.8) we obtain:

$$\begin{aligned}F(y_N)^\top(y_N - y) + \beta\sum_{k=0}^{N-2} F(y_{k+1})^\top(y_{k+1} - y) \leq & \frac{1}{2\eta}\|y_0 - y\|^2 - \frac{1}{2\eta}\|y_N - y\|^2 \\ & - \frac{L}{2\beta}\|y_N - y_{N-1}\|^2 \\ & + (F(y_N) - F(y_{N-1}))^\top(y_N - y) \\ & + \frac{\beta - 1}{2\eta}(\|y_1 - y\|^2 - \|y_N - y\|^2 \\ & + \sum_{k=1}^{N-1}\|y_k - y_{k+1}\|^2).\end{aligned} \tag{5.9}$$

Since $\beta \in (0, 1]$, which mean $\beta - 1 \leq 0$, we get the following inequality:

$$\begin{aligned}\beta\sum_{k=0}^{N-1} F(y_{k+1})^\top(y_{k+1} - y) \leq & \frac{1}{2\eta}\|y_0 - y\|^2 - \frac{1}{2\eta}\|y_N - y\|^2 - \frac{L}{2\beta}\|y_N - y_{N-1}\|^2 \\ & + (F(y_N) - F(y_{N-1}))^\top(y_N - y) \\ & - \frac{\beta - 1}{2\eta}\|y_N - y\|^2 \\ = & \frac{1}{2\eta}\|y_0 - y\|^2 - \frac{\beta}{2\eta}\|y_N - y\|^2 - \frac{L}{2\beta}\|y_N - y_{N-1}\|^2 \\ & + (F(y_N) - F(y_{N-1}))^\top(y_N - y).\end{aligned} \tag{5.10}$$

Now let $y = y_* \in zer(F)$ we get:

$$F(y_{k+1})^\top(y_{k+1} - y_*) = (F(y_{k+1}) - F(y_*))^\top(y_{k+1} - y_*) \geq 0$$

for all $k \geq 1$. Hence we get:

$$\begin{aligned}0 \leq & \frac{1}{2\eta}\|y_0 - y_*\|^2 - \frac{\beta}{2\eta}\|y_N - y_*\|^2 - \frac{L}{2\beta}\|y_N - y_{N-1}\|^2 \\ & + (F(y_N) - F(y_{N-1}))^\top(y_N - y_*).\end{aligned} \tag{5.11}$$

Using Cauchy-Schwarz inequality, the fact that F is L-Lipschitz continuous and $ab \leq \frac{1}{2}(\frac{1}{\beta}a^2 + \beta b^2)$ for $\beta > 0$, we obtain:

$$\begin{aligned}F(y_{N-1}))^\top(y_N - y_*) & \leq L\|y_N - y_{N-1}\|\|y_N - y_*\| \\ & \leq \frac{L}{2\beta}\|y_N - y_{N-1}\|^2 + \frac{\beta L}{2}\|y_N - y_*\|^2.\end{aligned}$$

Applying this relation to (5.11), we get:

$$
\begin{aligned}
0 \leq & \frac{1}{2\eta}\|y_0 - y_*\|^2 - \frac{\beta}{2\eta}\|y_N - y_*\|^2 - \frac{L}{2\beta}\|y_N - y_{N-1}\|^2 \\
& + \frac{L}{2\beta}\|y_N - y_{N-1}\|^2 + \frac{\beta L}{2}\|y_N - y_*\|^2.
\end{aligned}
$$

Hence:

$$
0 \leq \frac{1}{2\eta}\|y_0 - y_*\|^2 + \frac{\beta(\eta L - 1)}{2\eta}\|y_N - y_*\|^2.
$$

Therefore we obtain:

$$
\|y_N - y_*\|^2 \leq \frac{1}{\beta(1-\eta L)}\|y_0 - y_*\|^2 \leq \frac{2}{\beta(2-\beta)}\|y_0 - y_*\|^2.
$$

In the last inequality we use the fact that $\eta \leq \frac{\beta}{2L}$. $\square$

The above theorem shows that the sequence of iterates $\{(u_k, v_k)\}_{k\geq 1}$ stays inside a closed and bounded convex set. We will use this result in the following theorem to show the convergence rate of $\mathcal{O}(1/k)$ for the gap between function value at average iterates and at saddle point.

Theorem 5.2. *Suppose that f is a differentiable convex-concave function. Furthermore assume that the gradient of f satisfy the assumption in Lemma 2.5. Let $y_k = (u_k, v_k)$ for all $k \geq 1$ is generated by* (FRBS) *with $y_{-1} = y_0$ and F defined as in* (2.1)*. Let $\beta \in (0, 1]$ and $\eta \leq \frac{\beta}{2L}$. Then we get:*

$$
\left[\max_{v:(\hat{u}_N, v)\in\mathcal{D}} f(\hat{u}_N, v) - f^*\right] + \left[f^* - \min_{u:(u,\hat{v}_N)\in\mathcal{D}} f(u, \hat{v}_N)\right] \leq \frac{D\left(\frac{8L}{\beta(2-\beta)} + \frac{1}{2\eta}\left[2 + \frac{4}{\beta(2-\beta)}\right]\right)}{N\beta}.
$$

Here $f^* = f(u_*, v_*)$, $D = \|u_0 - u_*\|^2 + \|v_0 - v_*\|^2$ and

$$
\mathcal{D} = \left\{(u,v) \| \|u - u_*\|^2 + \|v - v_*^2\| \leq \frac{2}{\beta(2-\beta)}(\|u_0 - u_*\|^2 + \|v_0 - v_*^2\|)\right\}.
$$

Proof. Recall (5.10), we get for all $N \geq 1$:

$$
\begin{aligned}
\beta\sum_{k=1}^{N} F(y_k)^\top(y_k - y) \leq \frac{1}{2\eta}\|y_0 - y\|^2 - \frac{\beta}{2\eta}\|y_N - y\|^2 - \frac{L}{2\beta}\|y_N - y_{N-1}\|^2 \\
+ (F(y_N) - F(y_{N-1}))^\top(y_N - y).
\end{aligned} \tag{5.12}
$$

Hence:

$$
\frac{1}{N}\sum_{k=1}^{N} F(y_k)^\top(y_k - y) \leq \frac{\frac{1}{2\eta}\|y_0 - y\|^2 + (F(y_N) - F(y_{N-1}))^\top(y_N - y)}{N\beta}.
$$

Note that:

$$\|y_0 - y\|^2 \leq 2(\|y_0 - y_*\|^2 + \|y_* - y\|^2) \leq 2\left(D + \frac{2}{\beta(2-\beta)}\right),$$

here the last inequality follow from part b) of the previous theorem. Similarly, using Cauchy-Schwarz inequality and the fact F is a Lipschitz operator:

$$(F(y_N) - F(y_{N-1}))^\top (y_N - y) \leq L\|y_N - y_{N-1}\|\|y_N - y\|.$$

On the other hand for any $x, y \in \mathcal{D}$:

$$\|x - y\| \leq \|x - y_*\| + \|y - y_*\| \leq 2\sqrt{\frac{2D}{\beta(2-\beta)}}.$$

Using Proposition 2.7 we finally obtain for all $y = (u, v) \in \mathcal{D}$ and $N \geq 1$ that:

$$f(\hat{u}_N, v) - f(u, \hat{v}_N) \leq \frac{1}{N}\sum_{k=1}^{N} F(y_k)^\top (y_k - y) \leq \frac{D\left(\frac{8L}{\beta(2-\beta)} + \frac{1}{2\eta}[2 + \frac{4}{\beta(2-\beta)}]\right)}{N\beta}.$$

□

The result of Theorem 5.2 show the $\mathcal{O}(\frac{1}{k})$ convergence rate of the primal-dual gap $\max_{v:(\hat{u}_N, v)\in\mathcal{D}} f(\hat{u}_N, v) - \min_{u:(u,\hat{v}_N)\in\mathcal{D}} f(u, \hat{v}_N)$, which is often used to measure the closeness to the true solution in convex-concave setting [Nem04]. The duality gap is zero if and only if $(\hat{u}_n, \hat{v}_n)$ is a saddle point of the problem.

Corollar 5.3. *If the assumption of previous theorem is satisfied then:*

$$|f(\hat{u}_N, \hat{v}_N) - f^*| \leq \frac{D\left(\frac{8L}{\beta(2-\beta)} + \frac{1}{2\eta}[2 + \frac{4}{\beta(2-\beta)}]\right)}{N\beta} \quad \forall N \geq 1.$$

Proof. Since $y_* = (u_*, v_*) \in \mathcal{D}$ and it's the saddle point we have both $\left[\max_{v:(\hat{u}_N, y)\in\mathcal{D}} f(\hat{u}_N, y) - f^*\right]$ and $\left[f^* - \min_{u:(u,\hat{v}_N)\in\mathcal{D}} f(u, \hat{v}_N)\right]$ are non negative. Furthermore:

$$f(\hat{u}_N, \hat{v}_N) - f^* \leq \max_{v:(\hat{u}_N, y)\in\mathcal{D}} f(\hat{u}_N, y) - f^*,$$

and

$$f^* - f(\hat{u}_N, \hat{v}_N) \leq f^* - \min_{u:(u,\hat{v}_N)\in\mathcal{D}} f(u, \hat{v}_N).$$

Therefore $|f(\hat{u}_N, \hat{v}_N) - f^*| \leq \frac{D\left(\frac{8L}{\beta(2-\beta)} + \frac{1}{2\eta}[2+\frac{4}{\beta(2-\beta)}]\right)}{N\beta} \quad \forall N \geq 1.$ □

6 Forward Backward Forward Splitting

Recall the formulation of extra gradient method:

$$\begin{cases} y_k := x_k - \frac{\eta}{\beta} d_k, \\ x_{k+1} := x_k - \eta F(y_k) \quad \forall k \geq 0. \end{cases}$$

We now set $d_k = F(x_k)$ and derive a formula of y_{k+1} for all $k \geq 0$:

$$y_{k+1} = x_{k+1} - \frac{\eta}{\beta} F(x_{k+1}) = y_k - \eta(F(y_k) - \frac{1}{\beta} F(x_k)) - \frac{\eta}{\beta} F(x_{k+1}).$$

Here the second equality is obtained by setting $x_k = y_k + \frac{\eta}{\beta} F(x_k)$ in the second line of (EG). In order to do similar analysis like in FRBS we reorder the term to get for $k \geq 0$:

$$y_{k+1} = y_k - \eta F(y_{k+1}) + \eta(F(y_{k+1}) - F(y_k)) - \frac{\eta}{\beta}(F(x_{k+1}) - F(x_k))).$$

This method is called Forward Backward Forward Splitting (FBFS).

Theorem 6.1. *Let $\{(x_k, y_k)\}_{k \geq 0}$ be generated by FBFS with F a monotone and L-Lipschitz operator. Furthermore let $\eta \leq \frac{\beta}{L}$. Let $x_* \in zer(F)$. Then we have for all $k \geq 1$*

$$\|x_0 - x_*\|^2 \leq \|x_k - x_*\|^2 \quad \text{and} \quad \|y_k - x_*\|^2 \leq (2 + \frac{2}{\beta - L\eta}) \|x_0 - x_*\|^2.$$

Proof. Using Lemma 4.2 we obtain for any $\gamma > 0$ and $k \geq 0$:

$$\begin{aligned} \|x_{k+1} - \hat{x}\|^2 \leq \|x_k - \hat{x}\|^2 - \beta \|y_k - x_k\|^2 + \frac{\eta^2}{\gamma} \|F(y_k) - d_k\|^2 - 2\eta F(y_k)^\top (y_k - \hat{x}) \\ - (\beta - \gamma)\|y_k - x_{k+1}\|^2 - (1 - \beta)\|x_k - x_{k+1}\|^2. \end{aligned} \tag{6.1}$$

Replacing $\hat{x}$ with $x_* \in zer(F)$ we get $F(y_k)^\top (y_k - x_*) = (F(y_k) - F(x_*))^\top (y_k - x_*) \geq 0$. Furthermore, let $d_k = F(x_k)$ and $\|F(y_k) - F(x_k)\|^2 \leq L^2 \|y_k - x_k\|^2$, we obtain:

$$\begin{aligned} \|x_{k+1} - x_*\|^2 \leq \|x_k - x_*\|^2 - (\beta - \frac{L^2 \eta^2}{\gamma}) \|y_k - x_k\|^2 \\ - (\beta - \gamma)\|y_k - x_{k+1}\|^2 - (1 - \beta)\|x_k - x_{k+1}\|^2. \end{aligned} \tag{6.2}$$

Choose $\gamma = L\eta$ and since $\beta \in (0,1]$ the inequality implies:

$$\|x_{k+1} - x_*\|^2 + (\beta - L\eta)(\|y_k - x_k\|^2 + \|y_k - x_{k+1}\|^2) \leq \|x_k - x_*\|^2. \tag{6.3}$$

Hence we obtain

$$\|x_{k+1} - x_*\|^2 \leq \|x_k - x_*\|^2 \leq \cdots \leq \|x_0 - x_*\|^2,$$

due to the fact that $(\beta - L\eta) \geq 0$ from our choice of η. Furthermore we have $\|y_k - x_k\|^2 \leq \|x_k - x_*\|^2$. Therefore for all $k \geq 0$:

$$\begin{aligned}
\|y_k - x_*\|^2 \leq 2(\|x_{k+1} - x_*\|^2 + \|y_k - x_{k+1}\|^2) &\leq \left(2 + \frac{2}{\beta - L\eta}\right) \|x_k - x_*\|^2 \\
&\leq \left(2 + \frac{2}{\beta - L\eta}\right) \|x_0 - x_*\|^2.
\end{aligned}$$

□

This shows that the iteration $\{(u_k, v_k)\}_{k\geq 1}$ generated by the algorithm also stays within a bounded set $\mathcal{D}$ like in FRBS.

Theorem 6.2. *Suppose that f is a differentiable convex-concave function. Furthermore assume that the gradient of f satisfy the assumption in Lemma 2.5. Let $\{(x_k, y_k)\}_{k\geq 1}$, generated by FBFS and F defined as in* (2.1).
a) For all $k \geq 0$ and $y \in \mathbb{R}^n \times \mathbb{R}^m$:

$$\begin{aligned}
F(y_{k+1})^\top (y_{k+1} - y) \leq &\frac{1}{2\eta}\|y_k - y\|^2 - \frac{1}{2\eta}\|y_{k+1} - y\|^2 + \frac{1}{2\eta}\|x_{k+1} - y_k\|^2 \\
&+ (F(y_{k+1}) - \frac{1}{\beta}F(x_{k+1}))^\top (y_{k+1} - y) \\
&- (F(y_k) - \frac{1}{\beta}F(x_k))^\top (y_k - y).
\end{aligned} \tag{6.4}$$

b) Let $y_k = (u_k, v_k)$, then we get for all $N \in \mathbb{N}$:

$$\left[\max_{v:(\hat{u}_N, y)\in\mathcal{D}} f(\hat{u}_N, y) - f^*\right] + \left[f^* - \min_{u:(u,\hat{v}_N)\in\mathcal{D}} f(u, \hat{v}_N)\right] \leq \frac{D}{N}\left[\frac{11}{\eta} + \frac{21}{2\eta(\beta - L\eta)}\right],$$

here $f^* = f(u_*, v_*)$, $D = \|u_0 - u_*\|^2 + \|v_0 - v_*\|^2$ and

$$\mathcal{D} = \{(u,v) | \|u - u_*\|^2 + \|v - v_*^2\| \leq (2 + \frac{2}{\beta - L\eta})(\|u_0 - u_*\|^2 + \|v_0 - v_*^2\|)\}.$$

Proof. Applying (3.3) to y_{k+1} we get for all $k \geq 0$ and $y \in \mathbb{R}^n \times \mathbb{R}^m$:

$$\begin{aligned} F(y_{k+1})^\top (y_{k+1} - y) =& \frac{1}{2\eta}\|y_k - y\|^2 - \frac{1}{2\eta}\|y_{k+1} - y\|^2 \\ & - \frac{1}{2\eta}\|y_{k+1} - y_k\|^2 + \frac{1}{\eta}\epsilon_k^\top (y_{k+1} - y). \end{aligned} \tag{6.5}$$

Where $\epsilon_k = \eta(F(y_{k+1} - F(y_k)) - \frac{\eta}{\beta}(F(x_{k+1} - F(x_k)))$. First, we work on the term $\frac{1}{\eta}\epsilon_k^\top (y_{k+1} - y)$:

$$\begin{aligned} \frac{1}{\eta}\epsilon_k^\top (y_{k+1} - y) &= (F(y_{k+1}) - \frac{1}{\beta}F(x_{k+1}))^\top (y_{k+1} - y) \\ &\quad - (F(y_k) - \frac{1}{\beta}F(x_k))^\top (y_{k+1} - y) \\ &= (F(y_{k+1}) - \frac{1}{\beta}F(x_{k+1}))^\top (y_{k+1} - y) \\ &\quad - (F(y_k) - \frac{1}{\beta}F(x_k))^\top (y_k - y) \\ &\quad - (F(y_k) - \frac{1}{\beta}F(x_k))^\top (y_{k+1} - y_k). \end{aligned} \tag{6.6}$$

We get from our formula of FBFS that $F(y_k) - \frac{1}{\beta}F(x_k) = \frac{1}{\eta}(y_k - x_{k+1})$. Putting it into the last term and using Cauchy-Schwarz inequality as well as Young inequality, for all $k \geq 0$ we obtain:

$$\begin{aligned} -(F(y_k) - \frac{1}{\beta}F(x_k))^\top (y_{k+1} - y_k) &\leq \frac{1}{\eta}\|y_k - x_{k+1}\|\|y_{k+1} - y_k\| \\ &\leq \frac{1}{2\eta}(\|y_k - x_{k+1}\|^2 + \|y_{k+1} - y_k\|^2). \end{aligned}$$

Together with (6.5) and (6.6) we get:

$$\begin{aligned} F(y_{k+1})^\top (y_{k+1} - y) \leq& \frac{1}{2\eta}\|y_k - y\|^2 - \frac{1}{2\eta}\|y_{k+1} - y\|^2 + \frac{1}{2\eta}\|x_{k+1} - y_k\|^2 \\ & + (F(y_{k+1}) - \frac{1}{\beta}F(x_{k+1}))^\top (y_{k+1} - y) \\ & - (F(y_k) - \frac{1}{\beta}F(x_k))^\top (y_k - y). \end{aligned} \tag{6.7}$$

Which conclude part a).
For part b) of the proof we once again taking the sum of the above inequality over

$k = 0 \dots N-1$:

$$\begin{aligned}\sum_{k=1}^{N} F(y_k)^\top (y_k - y) \leq & \frac{1}{2\eta}\|y_0 - y\|^2 - \frac{1}{2\eta}\|y_N - y\|^2 + \frac{1}{2\eta}\sum_{k=0}^{N-1}\|x_{k+1} - y_k\|^2 \\ & + (F(y_N) - \frac{1}{\beta}F(x_N))^\top (y_N - y) \\ & - (F(y_0) - \frac{1}{\beta}F(x_0))^\top (y_k - y).\end{aligned} \tag{6.8}$$

For all $N \geq 1$. Once again using $F(y_k) - \frac{1}{\beta}F(x_k) = \frac{1}{\eta}(y_k - x_{k+1})$, Cauchy-Schwarz inequality and Young inequality, we get:

$$\begin{aligned}\sum_{k=1}^{N} F(y_k)^\top (y_k - y) \leq & \frac{1}{2\eta}\|y_0 - y\|^2 + \frac{1}{2\eta}\sum_{k=0}^{N}\|x_{k+1} - y_k\|^2 \\ & + \frac{1}{2\eta}(\|x_1 - y_0\|^2 + \|y_0 - y\|^2) \\ = & \frac{1}{\eta}\|y_0 - y\|^2 + \frac{1}{2\eta}\sum_{k=0}^{N}\|x_{k+1} - y_k\|^2 + \frac{1}{2\eta}\|x_1 - y_0\|^2.\end{aligned} \tag{6.9}$$

Now using the proof of Theorem 6.1 for all

$$y \in \mathcal{D} = \{(u,v) | \|u - u_*\|^2 + \|v - v_*^2\| \leq (2 + \frac{2}{\beta - L\eta})(\|u_0 - u_*\|^2 + \|v_0 - v_*^2\|)\}$$

and set $x_* = (u_*, v_*)$, we obtain:

$$\|y_0 - y\|^2 \leq 2(\|y_0 - x_*\|^2 + \|x_* - y\|^2) \leq 4\left(2 + \frac{2}{\beta - L\eta}\right) D.$$

$$\|x_1 - y_0\|^2 \leq 2(\|x_1 - x_*\|^2 + \|x_* - y_0\|^2) \leq 2\left(D + \left(2 + \frac{2}{\beta - L\eta}\right) D\right).$$

By (6.3), we get for all $k \geq 0$:

$$\|x_{k+1} - x_*\|^2 + (\beta - L\eta)(\|y_k - x_k\|^2 + \|y_k - x_{k+1}\|^2) \leq \|x_k - x_*\|^2.$$

Summing over $k = 0 \dots N$ to obtain:

$$(\beta - L\eta)\sum_{k=0}^{N}\|x_{k+1} - y_k\|^2 \leq \|x_0 - x_*\|^2.$$

With all these term we have the following inequality:

$$\begin{aligned}\frac{1}{N}\sum_{k=1}^{N}F(y_k)^\top(y_k-y) &\le \frac{D}{N}\left[\frac{1}{\eta}\left(8+\frac{8}{\beta-L\eta}\right)+\frac{1}{\eta}\left(3+\frac{2}{\beta-L\eta}\right)\right.\\ &\qquad\left.+\frac{1}{2\eta(\beta-L\eta)}\right]\\ &=\frac{D}{N}\left[\frac{11}{\eta}+\frac{21}{2\eta(\beta-L\eta)}\right].\end{aligned}\tag{6.10}$$

Similar to the proof of FRBS in the last section we use Proposition 2.7 and get:

$$f(\hat{u}_N,v)-f(u,\hat{v}_N)\le\frac{1}{N}\sum_{k=1}^{N}F(y_k)^\top(y_k-y)\le\frac{D}{N}\left[\frac{11}{\eta}+\frac{21}{2\eta(\beta-L\eta)}\right].$$

□

Similarly, we get the bound for the gap between the function value at average iterate and the true solution:

$$|f(\hat{u}_N,\hat{v}_N)-f^*|\le\frac{D}{N}\left[\frac{11}{\eta}+\frac{21}{2\eta(\beta-L\eta)}\right].$$

Remark 6.3. The FBFS algorithm requires us to make an additional calculation of gradient compared to FRBS. On the other hand FBFS allow us to use better step size $\frac{\beta}{L}$ compared to FRBS $\frac{\beta}{2L}$.

7 Numerical Experiment

In this section we perform numerical experiments to observe the convergence rates of FBRS and FBFS with different step size and choice of β. We consider the following minmax problem:

$$\min_{u\in\mathbb{R}^n}\max_{v\in\mathbb{R}^n}\mathcal{L}(u,v) := \frac{1}{2}u^\top Hu - u^\top h - v^\top(Au-b),$$

here $H := 2A^\top A$. In the first experiment we set $n = 1000$ with the following choice of A, b and h:

$$A := \frac{1}{4}\begin{pmatrix} & & & -1 & 1 \\ & & \iddots & \iddots & \\ & -1 & 1 & & \\ -1 & 1 & & & \\ 1 & & & & \end{pmatrix}, \quad b := \frac{1}{4}\begin{pmatrix}1\\1\\\vdots\\1\\1\end{pmatrix}, \quad h := \frac{1}{4}\begin{pmatrix}0\\0\\\vdots\\0\\1\end{pmatrix}.$$

The problem is introduced in [OX21] and it was shown that $\|A\| \leq \frac{1}{2}$ hence $\|H\| \leq \frac{1}{2}$. That why the operator $(u,v) \to [\nabla_u f(u,v); -\nabla_v f(u,v)]$ can take the Lipschitz constant $L = 1$. Since the solution can be calculated rather easily, we can calculate the difference $|f(\hat{u}_N, \hat{v}_N) - f^*|$ and use it for our experiment.
To see the effect of different pairs of (β,η) we choose 7 different values of β and for each β we choose $\eta = \frac{\sigma\beta}{2L}$ for FRBS and $\eta = \frac{\sigma\beta}{L}$ for FBFS. The Figure 7.1 is the result of running both FRBS and FBFS for a random starting point $z_0 = (u_0, v_0)$ such that $\|z_0 - z^*\| \leq 10$, where $z^* = (u^*, v^*)$ is the solution. We see that the function values oscillate near the solution. We can also see that the gap do reach the rate $\mathcal{O}\left(\frac{1}{k}\right)$. However, due to the oscillation near the solution, it is difficult to determine which pair of (β,η) is the optimal.

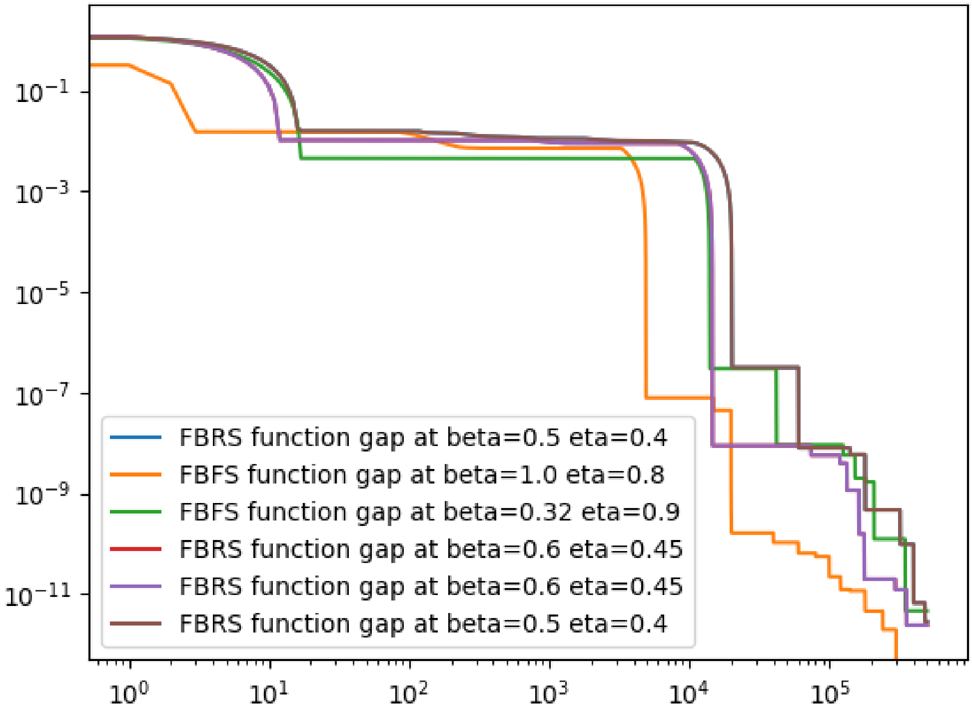

Figure 7.1: Sample from different β and η

Therefore, to see the impact of the choice of (β, η) I choose the following method: For each pair of (β, η), I generate 10 starting points $z_0 = (u_0, v_0)$ such that $\|z_0 - z^*\| \leq 10$, where $z^* = (u^*, v^*)$ is the solution. Since the function values oscillate near the solution, I run the algorithm for each starting point until the values $|f(\hat{u}_N, \hat{v}_N) - f^*|$ is less than 10^{-6} for 1000 consecutive iterations. The tables below show the average number of iterations needed for each pair (β, η):

	0.2	0.25	0.325	0.4	0.45
0.25	317936.3	254560.1	196102.4	159451.9	141860.3
0.32	248714.1	199075.7	153445.8	124835.3	111075.2
0.5	159458.2	127788.0	98506.4	80246.7	71419.3
0.6	133110.4	106677.7	82241.1	67035.0	59696.7
0.75	106677.1	85521.4	66023.8	53827.9	85515.0
0.85	94214.6	75567.5	58367.4	94226.4	83856.4
1.0	80218.1	64396.5	88777.0	80222.1	71428.8

Table 7.1: Average number of iterations for FRBS

	0.4	0.5	0.65	0.8	0.9
0.25	159469.9	127801.5	98530.5	80201.4	71422.6
0.32	124812.0	100062.9	77174.9	62892.8	56038.3
0.5	80264.5	64400.9	83895.4	80224.6	71436.3
0.6	67043.6	53823.5	82244.6	67020.6	59684.0
0.75	59122.2	85513.0	65999.9	53823.0	64372.8
0.85	94209.8	75558.4	58358.2	70900.0	63129.8
1.0	80215.1	64382.7	57067.6	60414.6	53821.6

Table 7.2: Average number of iterations for FBFS

Here the first column indicates the choice of β and the first row is the choice of σ that we use to calculate η. We can observe that $\beta = 1.0$ isn't always the most optimal choice for this example. In fact $\beta = 0.75$ and $\sigma = 0.8$ is the optimal among the 30 pairs we observe for this example. Also as noticed in Remark 6.3 the usage of larger step size in FBFS allows us to converge with less iterations.
For the second experiment, we reduce the dimension to $n = 200$ but now A is a random sparse $\mathbb{R}^n \times \mathbb{R}^n$ matrix and b and h are random $\mathbb{R}^n$ vectors. After running the experiments 6 times for all pairs of (β, η) like above, I take the average as shown in the tables below

	0.4	0.5	0.65	0.8	0.9
0.25	191341.55	157855.8667	127874.3333	101718.3667	90590.06667
0.32	151226.1333	119529.8667	97749.63333	84995.25	81168.68333
0.5	105338.0333	90830.86667	66929.21667	53880.63333	49911.3
0.6	88357.61667	69847.03333	57444.38333	51131.76667	44629.03333
0.75	72919.05	63491.46667	47800.06667	39502.51667	38356.81667
0.85	64546.4	53691.18333	41184.56667	37699.15	33158.86667
1	58105.06667	48189.43333	38448.56667	31149.85	29845.18333

	0.4	0.5	0.65	0.8	0.9
0.25	104815.85	86915.88333	65837.18333	57846.08333	54862.78333
0.32	85009.5	69201.21667	53955.18333	47028.3	45949.2
0.5	58216.1	48130.98333	37343.18333	32883.51667	29226.08333
0.6	50066.15	42359.01667	33298.61667	27792.51667	25663.6
0.75	42499.01667	34423.13333	28320.2	23328.15	20303.93333
0.85	36969.41667	30465.68333	23869.31667	22035.08333	19418.13333
1	32727.55	27467.88333	21738.8	19186.71667	17578.5

Table 7.3: Average number of iterations in random matrix case. First table: FBRS, second table: FBFS.

In this experiment however, we can see that the largest pair of (β, η) is the optimal value. But on some individual case, a different choice of (β, η), i.e $\beta < 1$, could be the better choice.
In conclusion the methods introduce in [TD23] retain the $\mathcal{O}\left(\frac{1}{k}\right)$ convergence rate for solving smooth convex-concave saddle point problem from ODGA and EG, while allow the algorithm to be more flexible by the choosing of β.

Bibliography

[ACB17] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[BTEGN09] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*, volume 28. Princeton university press, 2009.

[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.

[Kor76] Galina M Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

[Mar70] Bernard Martinet. Brève communication. régularisation d'inéquations variationnelles par approximations successives. *Revue française d'informatique et de recherche opérationnelle. Série rouge*, 4(R3):154–158, 1970.

[MOP20] Aryan Mokhtari, Asuman E Ozdaglar, and Sarath Pattathil. Convergence rate of o(1/k) for optimistic gradient and extragradient methods in smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 30(4):3230–3251, 2020.

[Nem04] Arkadi Nemirovski. Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

[OX21] Yuyuan Ouyang and Yangyang Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, 185(1):1–35, 2021.

[Roc76] R Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976.

[TD23] Quoc Tran-Dinh. Sublinear convergence rates of extragradient-type methods: A survey on classical and recent developments. *arXiv preprint arXiv:2303.17192*, 2023.

Abstract: Das Min-Max-Problem spielt in verschiedenen Bereichen eine wichtige Rolle, z. B. bei der Roburst-Optimierung oder bei Nullsummenspielen. Die Lösungen für solche Probleme werden Sattelpunkte genannt. In meiner Arbeit werden wir uns mit der Extragradienten-Methode beschäftigen - eine bekannte Methode, um einen Sattelpunkt einer glatten konvex-konkaven Funktion zu finden. Wir werden die von Ozdaglar eingeführte Technik verwenden, um zu zeigen, dass diese Methoden die Konvergenzrate von O(1/k) erreichen.