Lecture III: Variational Inequalities

Alina Ene

ADFOCS ’21: Convex Optimization and Graph Algorithms
Problem Definition

\[ F: \mathbb{R}^d \rightarrow \mathbb{R}^d \] vector-valued function, called an operator

\[ K \subseteq \mathbb{R}^d \] convex and bounded constraint set
Problem Definition

$F: \mathbb{R}^d \rightarrow \mathbb{R}^d$ vector-valued function, called an operator

$K \subseteq \mathbb{R}^d$ convex and bounded constraint set

The variational inequality problem associated with $F$ asks for a strong solution, i.e., a point $x^* \in K$ satisfying:

$$\langle F(x^*), x^* - y \rangle \leq 0 \quad \forall y \in K$$
**Problem Definition**

\( F : \mathbb{R}^d \rightarrow \mathbb{R}^d \) vector-valued function, called an operator

\( K \subseteq \mathbb{R}^d \) convex and bounded constraint set

The variational inequality problem associated with \( F \) asks for a **strong solution**, i.e., a point \( x^* \in K \) satisfying:

\[
\langle F(x^*), x^* - y \rangle \leq 0 \quad \forall y \in K
\]

A **weak solution** is a point \( x^* \in K \) satisfying

\[
\langle F(y), x^* - y \rangle \leq 0 \quad \forall y \in K
\]
**Problem Definition**

\[ F: \mathbb{R}^d \rightarrow \mathbb{R}^d \] vector-valued function, called an operator

\[ K \subseteq \mathbb{R}^d \] convex and bounded constraint set

The variational inequality problem associated with \( F \) asks for a **strong solution**, i.e., a point \( x^* \in K \) satisfying:

\[ \langle F(x^*), x^* - y \rangle \leq 0 \quad \forall y \in K \]

A **weak solution** is a point \( x^* \in K \) satisfying

\[ \langle F(y), x^* - y \rangle \leq 0 \quad \forall y \in K \]

For the operators we will consider in this lecture, a weak solution is a strong solution and vice versa.
Problem Definition

\[ F: \mathbb{R}^d \rightarrow \mathbb{R}^d \] vector-valued function, called an operator

\[ K \subseteq \mathbb{R}^d \] convex and bounded constraint set

Computational model: operator access via first-order oracle

\[ x \in K \quad \text{Blackbox} \quad F(x) \]

Goal: minimize number of queries \( x_1, x_2, \ldots, x_T \) to obtain

\[ \text{Err}(x_{out}) := \sup_{y \in K} \langle F(y), x_{out} - y \rangle \leq \epsilon \]
Example: Convex Minimization

Let $f: \mathbb{R}^d \rightarrow \mathbb{R}$ be a differentiable loss function.
Example: Convex Minimization

Let $f: \mathbb{R}^d \to \mathbb{R}$ be a differentiable loss function.

Consider the operator $F(x) = \nabla f(x)$. 
Example: Convex Minimization

Let \( f: \mathbb{R}^d \to \mathbb{R} \) be a differentiable loss function.

Consider the operator \( F(x) = \nabla f(x) \).

The strong solutions are the minimizers of \( f \) over \( K \):

\[
\langle \nabla f(x^*), x^* - y \rangle \leq 0 \quad \forall y \in K \iff x^* \in \arg\min_{x \in K} f(x)
\]

If \( f \) is convex, we have the following \( \forall x, y \):

\[
f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle
\]

\[
f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle
\]

\[
\Rightarrow \quad \langle \nabla f(x) - \nabla f(y), x - y \rangle \geq 0 \quad \forall x, y
\]

This gives us a way to extend convexity to operators.
Example: Convex Minimization

Let $f: \mathbb{R}^d \rightarrow \mathbb{R}$ be a differentiable loss function.

Consider the operator $F(x) = \nabla f(x)$.

The strong solutions are the minimizers of $f$ over $K$:

$$\langle \nabla f(x^*), x^* - y \rangle \leq 0 \quad \forall y \in K \iff x^* \in \arg \min_{x \in K} f(x)$$

The operator analogue of convexity is monotonicity:

$$\langle F(x) - F(y), x - y \rangle \geq 0 \quad \forall x, y \in K$$
Example: Convex Minimization

Let $f: \mathbb{R}^d \to \mathbb{R}$ be a differentiable loss function.

Consider the operator $F(x) = \nabla f(x)$.

The strong solutions are the minimizers of $f$ over $K$:

$$\langle \nabla f(x^*), x^* - y \rangle \leq 0 \quad \forall y \in K \iff x^* \in \arg\min_{x \in K} f(x)$$

The operator analogue of convexity is monotonicity:

$$\langle F(x) - F(y), x - y \rangle \geq 0 \quad \forall x, y \in K$$

Throughout, we will assume that $F$ is monotone and continuous.

For such operators, weak solutions are strong solutions, and we can measure convergence via the error function:

$$\text{Err}(x) := \sup_{y \in K} \langle F(y), x - y \rangle$$
Example: Nash Equilibria in Games

Consider a 2-player zero-sum game, such as:

Payoff matrix:

\[
\begin{pmatrix}
0 & -1 & 1 \\
1 & 0 & -1 \\
-1 & 1 & 0
\end{pmatrix}
\]
Example: Nash Equilibria in Games

Consider a 2-player zero-sum game, such as:

**Payoff matrix:**

\[
\mathbb{A} = \begin{pmatrix}
0 & -1 & 1 \\
1 & 0 & -1 \\
-1 & 1 & 0
\end{pmatrix}
\]

Alice chooses a distribution \( p \in \Delta_3 \) over the strategies
Bob chooses a distribution \( q \in \Delta_3 \) over the strategies

Alice’s expected payoff is \( f(p, q) := p^\top \mathbb{A} q \)
Example: Nash Equilibria in Games

Consider a 2-player zero-sum game, such as:

\[
A = \begin{pmatrix}
0 & -1 & 1 \\
1 & 0 & -1 \\
-1 & 1 & 0
\end{pmatrix}
\]

Alice chooses a distribution \( p \in \Delta_3 \) over the strategies
Bob chooses a distribution \( q \in \Delta_3 \) over the strategies

Alice’s expected payoff is \( f(p, q) := p^\top A q \)

A pair of strategies \((p^*, q^*)\) is a mixed Nash equilibrium if

\[
f(p, q^*) \leq f(p^*, q^*) \leq f(p^*, q) \quad \forall p, q \in \Delta_3
\]

No player is better off switching if the other player’s strategy remains fixed.
Consider a 2-player zero-sum game with payoff matrix \( A \in \mathbb{R}^{m \times n} \).

Alice’s expected payoff is \( f(p, q) := p^T A q \).

A pair of strategies \((p^*, q^*)\) is a mixed Nash equilibrium if

\[
\max_{p \in \Delta_m} f(p, q^*) \leq f(p^*, q^*) \leq \min_{q \in \Delta_n} f(p^*, q)
\]

Consider the monotone operator:

\[
F((p, q)) = \left( -\nabla_p f(p, q), \nabla_q f(p, q) \right) = (-Aq, Ap)
\]

Suppose \((p^*, q^*)\) is a strong solution.

\[
\left( (-Aq^*, A^T p^*), (p^*, q^*) - (p, q^*) \right) \leq 0 \quad \forall \ p
\]

\[
\Rightarrow p^T A q^* \geq p^T A q^* \quad \forall \ p
\]

Similarly, \( p^T A q^* \leq p^T A q^* \quad \forall \ q \)
Consider a 2-player zero-sum game with payoff matrix $A \in \mathbb{R}^{m \times n}$.

Alice's expected payoff is $f(p, q) := p^T A q$.

A pair of strategies $(p^*, q^*)$ is a mixed Nash equilibrium if

$$\max_{p \in \Delta_m} f(p, q^*) \leq f(p^*, q^*) \leq \min_{q \in \Delta_n} f(p^*, q)$$

Consider the monotone operator:

$$F((p, q)) = \left(-\nabla_p f(p, q), \nabla_q f(p, q)\right) = (-A q, A p)$$

A strong solution for the VI is a Nash equilibrium for the game.
Example: Min-Max Optimization

2-player games are a special case of min-max optimization:

\[
\min_{q \in \Delta_n} \max_{p \in \Delta_m} p^T A q = \max_{p \in \Delta_m} \min_{q \in \Delta_n} p^T A q
\]

(Minimax Thm: Von Neumann)

Suppose players take turns and both play optimally.

If Alice goes first:

Alice's payoff = \( \min_{q} \max_{p} p^T A q \)

If Alice goes second:

Alice's payoff = \( \max_{p} \min_{q} p^T A q \)

Minimax Thm: there is no advantage to going second.
Example: Min-Max Optimization

2-player games are a special case of min-max optimization:

\[
\min_{q \in \Delta_n} \max_{p \in \Delta_m} p^T A q = \max_{p \in \Delta_m} \min_{q \in \Delta_n} p^T A q
\]

More generally, we can consider the min-max optimization

\[
\min_{u \in U} \max_{v \in V} f(u, v)
\]

where \( f(u, v) \) is convex in \( u \) and concave in \( v \)
Example: Min-Max Optimization

2-player games are a special case of min-max optimization:

$$\min_{q \in \Delta_n} \max_{p \in \Delta_m} p^T A q = \max_{p \in \Delta_m} \min_{q \in \Delta_n} p^T A q$$

More generally, we can consider the min-max optimization

$$\min_{u \in U} \max_{v \in V} f(u, v)$$

where $f(u, v)$ is convex in $u$ and concave in $v$

As before, we can consider the monotone operator:

$$F((u, v)) = \left( \nabla_u f(u, v), -\nabla_v f(u, v) \right)$$
Example: Min-Max Optimization

2-player games are a special case of min-max optimization:

\[
\min_{q \in \Delta_n} \max_{p \in \Delta_m} p^T A q = \max_{p \in \Delta_m} \min_{q \in \Delta_n} p^T A q
\]

More generally, we can consider the min-max optimization

\[
\min_{u \in U} \max_{v \in V} f(u, v)
\]

where \(f(u, v)\) is convex in \(u\) and concave in \(v\)

As before, we can consider the monotone operator:

\[
F((u, v)) = \left( \nabla_u f(u, v), -\nabla_v f(u, v) \right)
\]

A strong solution \((u^*, v^*)\) for the VI is an equilibrium:

\[
\max_{v \in V} f(u^*, v) \leq f(u^*, v^*) \leq \min_{u \in U} f(u, v^*)
\]

or saddle point
Variational Inequalities

\[ F : \mathbb{R}^d \rightarrow \mathbb{R}^d \quad \text{monotone operator} \]

\[ K \subseteq \mathbb{R}^d \quad \text{convex and bounded constraint set} \]

Computational model: operator access via first-order oracle

\[ x \in K \quad \text{Blackbox} \quad F(x) \]

Goal: minimize number of queries \( x_1, x_2, \ldots, x_T \) to obtain

\[ \text{Err}(x_{\text{out}}) := \sup_{y \in K} \langle F(y), x_{\text{out}} - y \rangle \leq \epsilon \]
“Gradient” descent:

\[ x_t = \arg\min_{x \in K} \left\{ \langle F(x_{t-1}), x \rangle + \frac{1}{2\eta} \|x - x_{t-1}\|^2 \right\} \]
"Gradient" descent:

\[ x_t = \arg\min_{x \in K} \left\{ \langle F(x_{t-1}), x \rangle + \frac{1}{2\eta} \|x - x_{t-1}\|^2 \right\} \]

Consider \( \min \max uv \):

Equilibrium is \((u^*, v^*) = (0, 0)\)

Start = \((u_0, v_0) = (1, 1)\)
Extra-"Gradient" Algorithm:

\[
x_t = \arg \min_{u \in K} \left\{ \left\langle F(z_{t-1}), u \right\rangle + \frac{1}{2\eta} \| u - z_{t-1} \|^2 \right\}
\]

\[
z_t = \arg \min_{u \in K} \left\{ \left\langle F(x_t), u \right\rangle + \frac{1}{2\eta} \| u - z_{t-1} \|^2 \right\}
\]
In Extra-Gradient Descent We Trust

Extra-"Gradient" Algorithm:

\[ x_t = \arg \min_{u \in K} \left\{ \langle F(z_{t-1}), u \rangle + \frac{1}{2\eta} \| u - z_{t-1} \|^2 \right\} \]

\[ z_t = \arg \min_{u \in K} \left\{ \langle F(x_t), u \rangle + \frac{1}{2\eta} \| u - z_{t-1} \|^2 \right\} \]

Unconstrained (\( K = \mathbb{R}^d \)):

\[ x_t = z_{t-1} - \eta F(z_{t-1}) \]

\[ z_t = z_{t-1} - \eta F(x_t) = z_{t-1} - \eta F(z_{t-1}) \]
Extra-Gradient Algorithm

Let $z_0 \in K$

For $t = 1, \ldots, T$:

$$x_t = \arg \min_{u \in K} \left\{ \langle F(z_{t-1}), u \rangle + \frac{1}{2\eta} \| u - z_{t-1} \|^2 \right\}$$

$$z_t = \arg \min_{u \in K} \left\{ \langle F(x_t), u \rangle + \frac{1}{2\eta} \| u - z_{t-1} \|^2 \right\}$$

Return $\bar{x}_T = \frac{1}{T} \sum_{t=1}^{T} x_t$
Extra-Gradient Analysis

We will analyze convergence via the error function:

\[ \text{Err}(x) := \sup_{y \in K} \langle F(y), x - y \rangle \]

Thus we want to upper bound \( \text{Err}(\bar{x}_T) \)

Analogously to GD, we consider two settings:

"non-smooth" setting: \( \|F(x)\| \leq G \quad \forall x \in K \)

"smooth" setting: \( \|F(x) - F(y)\| \leq \beta \|x - y\| \quad \forall x, y \in K \)
**Extra-Gradient Analysis**

We will analyze convergence via the error function:

\[
\text{Err}(x) := \sup_{y \in K} \langle F(y), x - y \rangle
\]

Thus we want to upper bound \( \text{Err}(\bar{x}_T) \)

Using that \( F \) is monotone (analogue of convexity):

\[
\text{Err}(\bar{x}_T) = \sup_{y \in K} \langle F(y), \bar{x}_T - y \rangle \quad \text{definition}
\]

\[
= \sup_{y \in K} \left( \frac{1}{T} \sum_{t=1}^{T} \langle F(y), x_t - y \rangle \right) \quad \text{monotonicity}
\]

\[
\leq \sup_{y \in K} \left( \frac{1}{T} \sum_{t=1}^{T} \langle F(x_t), x_t - y \rangle \right) \quad \langle F(x_t) - F(y), x_t - y \rangle \geq 0
\]

\[
\bar{x}_T = \frac{1}{T} \sum_{t=1}^{T} x_t
\]
Extra-Gradient Analysis

We have: \( \text{Err}(\bar{x}_T) \leq \frac{1}{T} \sup_{y \in K} \left( \sum_{t=1}^{T} \langle F(x_t), x_t - y \rangle \right) \) \( \text{Fix} \quad y \in K \)

We split each term so that we can use the optimality condition:

\[
x_t = \arg\min_{u \in K} \left\{ \langle F(z_{t-1}), u \rangle + \frac{1}{2\eta} \| u - z_{t-1} \|^2 \right\}
\]

\[
z_t = \arg\min_{u \in K} \left\{ \langle F(x_t), u \rangle + \frac{1}{2\eta} \| u - z_{t-1} \|^2 \right\}
\]
Extra-Gradient Analysis

We have: \( \text{Err}(\bar{x}_T) \leq \frac{1}{T} \sup_{y \in K} \left( \sum_{t=1}^{T} \langle F(x_t), x_t - y \rangle \right) \)

We split each term so that we can use the optimality condition:

\[
\langle F(x_t), x_t - y \rangle = \langle F(x_t), z_t - y \rangle + \langle F(z_{t-1}), x_t - z_t \rangle + \langle F(x_t) - F(z_{t-1}), x_t - z_t \rangle
\]
Extra-Gradient Analysis

We have:  
\[
\text{Err}(\bar{x}_T) \leq \frac{1}{T} \sup_{y \in K} \left( \sum_{t=1}^{T} \langle F(x_t), x_t - y \rangle \right)
\]

We split each term so that we can use the optimality condition:
\[
\langle F(x_t), x_t - y \rangle = \langle F(x_t), z_t - y \rangle + \langle F(z_{t-1}), x_t - z_t \rangle + \langle F(x_t) - F(z_{t-1}), x_t - z_t \rangle
\]

We bound the first term using the optimality condition for \( z_t \):
\[
z_t = \arg \min_{u \in K} \left\{ \langle F(x_t), u \rangle + \frac{1}{2\eta} \| u - z_{t-1} \|^2 \right\}
\]
We have: \[ \text{Err}(\bar{x}_T) \leq \frac{1}{T} \sup_{y \in K} \left( \sum_{t=1}^{T} \langle F(x_t), x_t - y \rangle \right) \]

We split each term so that we can use the optimality condition:

\[ \langle F(x_t), x_t - y \rangle = \langle F(x_t), z_t - y \rangle + \langle F(z_{t-1}), x_t - z_t \rangle + \langle F(x_t) - F(z_{t-1}), x_t - z_t \rangle \]

We bound the first term using the optimality condition for \( z_t \):

\[ z_t = \arg\min_{u \in K} \left\{ \langle F(x_t), u \rangle + \frac{1}{2\eta} \| u - z_{t-1} \| ^2 \right\} \]

\[ \langle F(x_t), z_t - y \rangle \leq \frac{1}{\eta} \langle z_{t-1} - z_t, z_t - y \rangle \]

\[ a^2 b \leq \frac{1}{2} \left( a + b \right)^2 - \frac{1}{2} a^2 - \frac{1}{2} b^2 \]

\[ \implies \frac{1}{2\eta} \left( \| z_{t-1} - y \| ^2 - \| z_t - y \| ^2 - \| z_{t-1} - z_t \| ^2 \right) \]
**Extra-Gradient Analysis**

We have:  
\[ \text{Err}(\bar{x}_T) \leq \frac{1}{T} \sup_{y \in K} \left( \sum_{t=1}^{T} \langle F(x_t), x_t - y \rangle \right) \]

We split each term so that we can use the optimality condition:

\[ \langle F(x_t), x_t - y \rangle = \langle F(x_t), z_t - y \rangle + \langle F(z_{t-1}), x_t - z_t \rangle + \langle F(x_t) - F(z_{t-1}), x_t - z_t \rangle \]

We bound the second term using the optimality condition for \( x_t \):

\[ x_t = \arg \min_{u \in K} \left\{ \langle F(z_{t-1}), u \rangle + \frac{1}{2\eta} \| u - z_{t-1} \| ^2 \right\} \]
Extra-Gradient Analysis

We have: \[ \text{Err}(\bar{x}_T) \leq \frac{1}{T} \sup_{y \in K} \left( \sum_{t=1}^{T} \langle F(x_t), x_t - y \rangle \right) \]

We split each term so that we can use the optimality condition:

\[ \langle F(x_t), x_t - y \rangle = \langle F(x_t), z_t - y \rangle + \langle F(z_{t-1}), x_t - z_t \rangle + \langle F(x_t) - F(z_{t-1}), x_t - z_t \rangle \]

We bound the second term using the optimality condition for \( x_t \):

\[
x_t = \arg \min_{u \in K} \left\{ \langle F(z_{t-1}), u \rangle + \frac{1}{2\eta} \| u - z_{t-1} \|^2 \right\}
\]

\[
\langle F(z_{t-1}), x_t - z_t \rangle \leq \frac{1}{\eta} \langle z_{t-1} - x_t, x_t - z_t \rangle
\]

\[
= \frac{1}{2\eta} \left( \| z_{t-1} - z_t \|^2 - \| x_t - z_{t-1} \|^2 - \| x_t - z_t \|^2 \right)
\]
We have:

\[
\langle F(x_t), x_t - y \rangle = \langle F(x_t), z_t - y \rangle + \langle F(z_{t-1}), x_t - z_t \rangle + \langle F(x_t) - F(z_{t-1}), x_t - z_t \rangle
\]

\[
\langle F(x_t), z_t - y \rangle \leq \frac{1}{2\eta} \left( \| z_{t-1} - y \|^2 - \| z_t - y \|^2 - \| z_{t-1} - z_t \|^2 \right)
\]

\[
+ \langle F(z_{t-1}), x_t - z_t \rangle \leq \frac{1}{2\eta} \left( \| z_{t-1} - z_t \|^2 - \| x_t - z_{t-1} \|^2 - \| x_t - z_t \|^2 \right)
\]
We have:

\[
\langle F(x_t), x_t - y \rangle = \langle F(x_t), z_t - y \rangle + \langle F(z_{t-1}), x_t - z_t \rangle + \langle F(x_t) - F(z_{t-1}), x_t - z_t \rangle
\]

\[
\langle F(x_t), z_t - y \rangle \leq \frac{1}{2\eta} \left( \| z_{t-1} - y \|^2 - \| z_t - y \|^2 - \| z_{t-1} - z_t \|^2 \right)
\]

\[
\langle F(z_{t-1}), x_t - z_t \rangle \leq \frac{1}{2\eta} \left( \| z_{t-1} - z_t \|^2 - \| x_t - z_{t-1} \|^2 - \| x_t - z_t \|^2 \right)
\]

Therefore

\[
\langle F(x_t), x_t - y \rangle \leq \frac{1}{2\eta} \| z_{t-1} - y \|^2 - \frac{1}{2\eta} \| z_t - y \|^2
\]

\[
+ \langle F(x_t) - F(z_{t-1}), x_t - z_t \rangle - \frac{1}{2\eta} \left( \| x_t - z_{t-1} \|^2 + \| x_t - z_t \|^2 \right)
\]
We have:

\[
\langle F(x_t), x_t - y \rangle = \langle F(x_t), z_t - y \rangle + \langle F(z_{t-1}), x_t - z_t \rangle + \langle F(x_t) - F(z_{t-1}), x_t - z_t \rangle
\]

\[
\langle F(x_t), z_t - y \rangle \leq \frac{1}{2\eta} \left( \| z_{t-1} - y \|^2 - \| z_t - y \|^2 - \| z_{t-1} - z_t \|^2 \right)
\]

\[
\langle F(z_{t-1}), x_t - z_t \rangle \leq \frac{1}{2\eta} \left( \| z_{t-1} - z_t \|^2 - \| x_t - z_{t-1} \|^2 - \| x_t - z_t \|^2 \right)
\]

Therefore

\[
\langle F(x_t), x_t - y \rangle \leq \frac{1}{2\eta} \| z_{t-1} - y \|^2 - \frac{1}{2\eta} \| z_t - y \|^2 + \langle F(x_t) - F(z_{t-1}), x_t - z_t \rangle - \frac{1}{2\eta} \left( \| x_t - z_{t-1} \|^2 + \| x_t - z_t \|^2 \right)
\]

\( \text{telescopes} \)

\( \text{loss} \)

\( \text{gain} \)
Next, we analyze the net loss:

\[
\langle F(x_t) - F(z_{t-1}), x_t - z_t \rangle - \frac{1}{2\eta} \left( \| x_t - z_{t-1} \|^2 + \| x_t - z_t \|^2 \right)
\]
Next, we analyze the net loss:

\[
\langle F(x_t) - F(z_{t-1}), x_t - z_t \rangle - \frac{1}{2\eta} \left( \| x_t - z_{t-1} \|^2 + \| x_t - z_t \|^2 \right)
\]

\[\text{loss}\]

\[\\text{gain}\]

In the “non-smooth” setting, we assume \( \| F(x) \| \leq G \)

We proceed similarly to the GD analysis, and obtain:

\[
\langle F(x_t) - F(z_{t-1}), x_t - z_t \rangle \leq \| F(x_t) - F(z_{t-1}) \| \| x_t - z_t \| \text{ Cauchy-Schwartz}
\]
Next, we analyze the net loss:

\[
\langle F(x_t) - F(z_{t-1}), x_t - z_t \rangle - \frac{1}{2\eta} \left( \| x_t - z_{t-1} \|^2 + \| x_t - z_t \|^2 \right)
\]

In the "non-smooth" setting, we assume \( \| F(x) \| \leq G \)

We proceed similarly to the GD analysis, and obtain:

\[
\langle F(x_t) - F(z_{t-1}), x_t - z_t \rangle \leq \| F(x_t) - F(z_{t-1}) \| \| x_t - z_t \| \text{ Cauchy-Schwartz}
\]
\[
\leq \left( \| F(x_t) \| + \| F(z_{t-1}) \| \right) \| x_t - z_t \| \text{ D-ineq.}
\]
Next, we analyze the net loss:

$$\langle F(x_t) - F(z_{t-1}), x_t - z_t \rangle - \frac{1}{2\eta} \left( \| x_t - z_{t-1} \|^2 + \| x_t - z_t \|^2 \right)$$

In the “non-smooth” setting, we assume $\|F(x)\| \leq G$

We proceed similarly to the GD analysis, and obtain:

$$\langle F(x_t) - F(z_{t-1}), x_t - z_t \rangle \leq \| F(x_t) - F(z_{t-1}) \| \| x_t - z_t \| \leq \left( \| F(x_t) \| + \| F(z_{t-1}) \| \right) \| x_t - z_t \| \leq 2G \| x_t - z_t \|$$
Next, we analyze the net loss:

\[
\langle F(x_t) - F(z_{t-1}), x_t - z_t \rangle - \frac{1}{2\eta} \left( \| x_t - z_{t-1} \|^2 + \| x_t - z_t \|^2 \right)
\]

In the “non-smooth” setting, we assume \( \| F(x) \| \leq G \)

We proceed similarly to the GD analysis, and obtain:

\[
\langle F(x_t) - F(z_{t-1}), x_t - z_t \rangle \leq \| F(x_t) - F(z_{t-1}) \| \| x_t - z_t \| \leq \left( \| F(x_t) \| + \| F(z_{t-1}) \| \right) \| x_t - z_t \| \leq 2G \| x_t - z_t \|
\]

\[
\leq 2\eta G^2 + \frac{1}{2\eta} \| x_t - z_t \|^2 \quad \text{ab} \leq \frac{\lambda}{2} \bar{\alpha}^2 + \frac{1}{2\lambda} \bar{\beta}^2 \quad \text{for any } \lambda > 0
\]
Extra-Gradient Analysis

Putting everything together:

$$\langle F(x_t), x_t - y \rangle \leq \frac{1}{2\eta} \| z_{t-1} - y \|^2 - \frac{1}{2\eta} \| z_t - y \|^2 + 2\eta G^2$$
Putting everything together:

$$\langle F(x_t), x_t - y \rangle \leq \frac{1}{2\eta} \| z_{t-1} - y \|^2 - \frac{1}{2\eta} \| z_t - y \|^2 + 2\eta G^2$$

Summing up and telescoping:

$$\sum_{t=1}^{T} \langle F(x_t), x_t - y \rangle \leq \frac{1}{2\eta} \| z_0 - y \|^2 + 2\eta G^2 T$$
Extra-Gradient Analysis

Putting everything together:

\[ \langle F(x_t), x_t - y \rangle \leq \frac{1}{2\eta} \| z_{t-1} - y \|^2 - \frac{1}{2\eta} \| z_t - y \|^2 + 2\eta G^2 \]

Summing up and telescoping:

\[ \sum_{t=1}^{T} \langle F(x_t), x_t - y \rangle \leq \frac{1}{2\eta} \| z_0 - y \|^2 + 2\eta G^2 T \]

We set \( \eta \) to balance the two terms:

\[ \eta = \frac{\| z_0 - y \|}{2G\sqrt{T}} \]
Extra-Gradient Analysis

Putting everything together:

$$\langle F(x_t), x_t - y \rangle \leq \frac{1}{2\eta} \| z_{t-1} - y \|^2 - \frac{1}{2\eta} \| z_t - y \|^2 + 2\eta G^2$$

Summing up and telescoping:

$$\sum_{t=1}^{T} \langle F(x_t), x_t - y \rangle \leq \frac{1}{2\eta} \| z_0 - y \|^2 + 2\eta G^2 T$$

We set $\eta$ to balance the two terms:

$$\eta = \frac{\| z_0 - y \|}{2G\sqrt{T}}$$

Thus

$$\sum_{t=1}^{T} \langle F(x_t), x_t - y \rangle \leq 2G \underbrace{\| z_0 - y \| \sqrt{T}}_{\leq R} \leq 2GR\sqrt{T}$$
Extra-Gradient Analysis

We have:

\[
\text{Err}(\bar{x}_T) \leq \frac{1}{T} \sup_{y \in K} \left( \sum_{t=1}^{T} \langle F(x_t), x_t - y \rangle \right)
\]

\[
\sum_{t=1}^{T} \langle F(x_t), x_t - y \rangle \leq 2GR\sqrt{T} \quad \forall \ y \in K
\]

Therefore we have our final convergence guarantee:

\[
\text{Err}(\bar{x}_T) \leq O \left( \frac{GR}{\sqrt{T}} \right)
\]
Extra-Gradient Analysis

Next, we consider the “smooth” (i.e., Lipschitz) setting:

$$\|F(x) - F(y)\| \leq \beta \|x - y\| \quad \forall x, y$$
Next, we consider the “smooth” (i.e., Lipschitz) setting:

\[ \|F(x) - F(y)\| \leq \beta \|x - y\| \quad \forall x, y \]

As before, we need to analyze the net loss:

\[
\begin{align*}
\langle F(x_t) - F(z_{t-1}), x_t - z_t \rangle - \frac{1}{2\eta} \left( \| x_t - z_{t-1} \|^2 + \| x_t - z_t \|^2 \right)
\end{align*}
\]

\begin{itemize}
\item \text{loss}
\item \text{gain}
\end{itemize}
Extra-Gradient Analysis

Next, we consider the “smooth” (i.e., Lipschitz) setting:

$$\|F(x) - F(y)\| \leq \beta \|x - y\| \quad \forall x, y$$

As before, we need to analyze the net loss:

$$\langle F(x_t) - F(z_{t-1}), x_t - z_t \rangle - \frac{1}{2\eta} \left( \| x_t - z_{t-1} \|^2 + \| x_t - z_t \|^2 \right)$$

Using the Lipschitz property, we obtain:

$$\langle F(x_t) - F(z_{t-1}), x_t - z_t \rangle \leq \| F(x_t) - F(z_{t-1}) \| \| x_t - z_t \| \quad \text{(Cauchy-Schwarz)}$$
Next, we consider the “smooth” (i.e., Lipschitz) setting:

\[ \|F(x) - F(y)\| \leq \beta \|x - y\| \quad \forall x, y \]

As before, we need to analyze the net loss:

\[
\begin{aligned}
&\langle F(x_t) - F(z_{t-1}), x_t - z_t \rangle - \frac{1}{2\eta} \left( \| x_t - z_{t-1} \|^2 + \| x_t - z_t \|^2 \right) \\
= &\quad \text{loss} \quad \text{gain}
\end{aligned}
\]

Using the Lipschitz property, we obtain:

\[
\begin{aligned}
&\langle F(x_t) - F(z_{t-1}), x_t - z_t \rangle \\
\leq &\quad \| F(x_t) - F(z_{t-1}) \| \| x_t - z_t \| \quad \text{CS} \\
\leq &\quad \beta \| x_t - z_{t-1} \| \| x_t - z_t \| \quad \text{Smoothness}
\end{aligned}
\]
Next, we consider the “smooth” (i.e., Lipschitz) setting:

$$\|F(x) - F(y)\| \leq \beta \|x - y\| \quad \forall x, y$$

As before, we need to analyze the net loss:

$$\langle F(x_t) - F(z_{t-1}), x_t - z_t \rangle - \frac{1}{2\eta} \left( \| x_t - z_{t-1} \|^2 + \| x_t - z_t \|^2 \right)$$

Using the Lipschitz property, we obtain:

$$\langle F(x_t) - F(z_{t-1}), x_t - z_t \rangle \leq \frac{\beta}{2} \| x_t - z_{t-1} \| \| x_t - z_t \|$$

$$\leq \frac{\beta}{2} \| x_t - z_{t-1} \|^2 + \frac{\beta}{2} \| x_t - z_t \|^2$$
As before, we need to analyze the net loss:

\[
\langle F(x_t) - F(z_{t-1}), x_t - z_t \rangle - \frac{1}{2\eta} \left( \| x_t - z_{t-1} \|^2 + \| x_t - z_t \|^2 \right)
\]

Using the Lipschitz property, we obtained:

\[
\langle F(x_t) - F(z_{t-1}), x_t - z_t \rangle \leq \frac{\beta}{2} \| x_t - z_{t-1} \|^2 + \frac{\beta}{2} \| x_t - z_t \|^2
\]
Extra-Gradient Analysis

As before, we need to analyze the net loss:

\[
\langle F(x_t) - F(z_{t-1}), x_t - z_t \rangle - \frac{1}{2\eta} \left( \| x_t - z_{t-1} \|^2 + \| x_t - z_t \|^2 \right)
\]

Using the Lipschitz property, we obtained:

\[
\langle F(x_t) - F(z_{t-1}), x_t - z_t \rangle \leq \frac{\beta}{2} \| x_t - z_{t-1} \|^2 + \frac{\beta}{2} \| x_t - z_t \|^2
\]

Thus, if we set \( \eta = \frac{1}{\beta} \), we obtain:

\[
\langle F(x_t) - F(z_{t-1}), x_t - z_t \rangle - \frac{1}{2\eta} \left( \| x_t - z_{t-1} \|^2 + \| x_t - z_t \|^2 \right) \leq 0
\]
Extra-Gradient Analysis

Putting everything together:

\[
\langle F(x_t), x_t - y \rangle \leq \frac{\beta}{2} \| z_{t-1} - y \|^2 - \frac{\beta}{2} \| z_t - y \|^2
\]
Putting everything together:

\[ \langle F(x_t), x_t - y \rangle \leq \frac{\beta}{2} \| z_{t-1} - y \|^2 - \frac{\beta}{2} \| z_t - y \|^2 \]

Summing up and telescoping:

\[ \sum_{t=1}^{T} \langle F(x_t), x_t - y \rangle \leq \frac{\beta}{2} \| z_0 - y \|^2 \leq \frac{\beta}{2} R^2 \]
Extra-Gradient Analysis

Putting everything together:

$$\langle F(x_t), x_t - y \rangle \leq \frac{\beta}{2} \| z_{t-1} - y \|^2 - \frac{\beta}{2} \| z_t - y \|^2$$

Summing up and telescoping:

$$\sum_{t=1}^{T} \langle F(x_t), x_t - y \rangle \leq \frac{\beta}{2} \| z_0 - y \|^2 \leq \frac{\beta}{2} R^2$$

Which gives us our final convergence guarantee:

$$\text{Err}(\bar{x}_T) \leq O \left( \frac{\beta R^2}{T} \right)$$
Extra-Gradient Analysis

Putting everything together:

\[ \langle F(x_t), x_t - y \rangle \leq \frac{\beta}{2} \| z_{t-1} - y \|^2 - \frac{\beta}{2} \| z_t - y \|^2 \]

Summing up and telescoping:

\[ \sum_{t=1}^{T} \langle F(x_t), x_t - y \rangle \leq \frac{\beta}{2} \| z_0 - y \|^2 \leq \frac{\beta}{2} R^2 \]

Which gives us our final convergence guarantee:

\[ \text{Err}(\bar{x}_T) \leq O \left( \frac{\beta R^2}{T} \right) \]

Both the “non-smooth” and “smooth” rates are optimal
Extension to Bregman Divergences

**Mirror-Prox Algorithm** (Nemirovski)

Let \( z_0 \in K \)

For \( t = 1, \ldots, T \):

\[
x_t = \arg \min_{u \in K} \left\{ \langle F(z_{t-1}), u \rangle + \frac{1}{2\eta} D_\psi(u, z_{t-1}) \right\}
\]

\[
z_t = \arg \min_{u \in K} \left\{ \langle F(x_t), u \rangle + \frac{1}{2\eta} D_\psi(u, z_{t-1}) \right\}
\]

Return \( \bar{x}_T = \frac{1}{T} \sum_{t=1}^{T} x_t \)

\( \psi \): strongly convex function

\[
D_\psi(x, y) = \psi(x) - \psi(y) - \langle \nabla \psi(y), x - y \rangle
\]

\( \psi \): strongly convex function

\[
D_\psi(x, y) = \psi(x) - \psi(y) - \langle \nabla \psi(y), x - y \rangle
\]
Adaptive Algorithms

Adaptive Algorithm: Iterate Movement

Let $z_0 \in K, \eta_0 > 0,R \geq \max_{x,y \in K} ||x - y||$

For $t = 1, \ldots, T$:

$$x_t = \arg \min_{u \in K} \left\{ \langle F(z_{t-1}), u \rangle + \frac{1}{2\eta_{t-1}} \| u - z_{t-1} \|^2 \right\}$$

$$z_t = \arg \min_{u \in K} \left\{ \langle F(x_t), u \rangle + \frac{1}{2\eta_{t-1}} \| u - z_{t-1} \|^2 \right\}$$

$$\frac{1}{\eta_t^2} = \frac{1}{\eta_{t-1}^2} \left( 1 + \frac{\| x_t - z_{t-1} \|^2 + \| x_t - z_t \|^2}{2R^2} \right)$$

Return $\bar{x}_T = \frac{1}{T} \sum_{t=1}^{T} x_t$
Adaptive Algorithms

Adaptive Algorithm: Operator Differences

Let $z_0 \in K, \eta_0 > 0, R \geq \max_{x,y \in K} \|x - y\|

For $t = 1, \ldots, T$:

$$x_t = \arg \min_{u \in K} \left\{ \langle F(z_{t-1}), u \rangle + \frac{1}{2\eta_{t-1}} \| u - z_{t-1} \|^2 \right\}$$

$$\eta_t = \frac{R}{\sqrt{\sum_{s=1}^{t} \| F(x_s) - F(z_{s-1}) \|^2}}$$

$$z_t = \arg \min_{u \in K} \left\{ \langle F(x_t), u \rangle + \frac{1}{2\eta_{t-1}} \| u - z_{t-1} \|^2 \right\}$$

Return $\bar{x}_T = \frac{1}{T} \sum_{t=1}^{T} x_t$