Online Learning in Games

DRAFT

Prof. Volkan Cevher
volkan.cevher@epfl.ch

Lecture 11: Equilibrium computation for nonmonontone problems

Laboratory for Information and Inference Systems (LIONS)
École Polytechnique Fédérale de Lausanne (EPFL)

EE-735 (Spring 2024)

License Information for Online Learning in Games Slides

Acknowledgements

These slides were originally prepared by Dongyang Fan and Wenjie Xu.

Outline

Introduction

Min-Max Optimization has wide applications in machine learning:

- Nonsmooth optimization
- Generative adversarial networks
- Distributionally robust optimization

Example (Minimax optimization)

$$\min_{x \in \mathbb{R}^{n_x}} \max_{y \in \mathbb{R}^{n_y}} \mathcal{L}(x, y) := \phi(x, y) + g(x) - h(y) \tag{1}$$

where ϕ is not necessarily convex in x and concave in y. $g(\cdot)$ and $h(\cdot)$ are lower semi-continuous and convex, for example, they are usually chosen as l_1, l_2 regularization terms or indicator functions (constrained case).

Remarks:

- $Fz = (\nabla_x \phi(x, y), -\nabla_y \phi(x, y))$, $Az = (\partial g(x), \partial h(y))$
- From an operator point of view: the problem can be translated as **finding zeros of $\mathbf{F} + \mathbf{A}$**
- We will start our analysis from the unconstrained case, i.e. $A \equiv 0$

Recap – Extragradient & Monotone Case

Definition (Operators)

- An operator $F : \mathbb{R}^d \to \mathbb{R}^d$ is said to be monotone if $\langle Fz - Fz', z - z' \rangle \geq 0, \forall z, z' \in \mathbb{R}^d$
- F is *maximal* monotone if there is no monotone operator that properly contains it.
- An operator $F : \mathbb{R}^d \to \mathbb{R}^d$ is said to be L-Lipschitz for $L > 0$ if $\|Fz - Fz'\| \leq L\|z - z'\|, \forall z, z' \in \mathbb{R}^d$

◦ Recall the extragradient (EG) algorithm from **?**], which is a de facto algorithm for **monotone** VI problems:

$$\begin{aligned} \bar{z}^t &= z^t - \gamma F z^t, \\ z^{t+1} &= z^t - \gamma F \bar{z}^t. \end{aligned} \tag{EG}$$

◦ `EG` can be seen as an approximation to the proximal point method $z^{t+1} = z^t - \gamma F(z^{t+1})$

◦ For F monotone and L-Lipschitz (ϕ convex-concave, L-Lipschitz smooth), `EG` has tight last iterate convergence rate $\mathcal{O}(\frac{1}{\sqrt{T}})$ in terms of gap function [1]. [**?**]. While average iterate enjoys $\mathcal{O}(\frac{1}{T})$ convergence rate. [**?**]

[1] Gap function: $\mathrm{GAP}_{Z,F,D}(z) = \max_{z' \in Z \cap \mathcal{B}(z,D)} \langle Fz, z - z' \rangle$

Convergence plot for ExtraGradient – Bilinear Game

$$\min_{x\in\mathbb{R}} \max_{y\in\mathbb{R}} xy$$

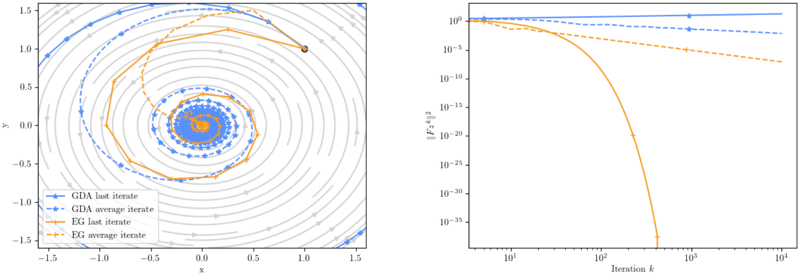


Figure: Left: Convergence plot of EG for bilinear game; Right: decrease of $\|Fz^t\|^2$ over number of iterations

- For strongly monotone and affine operators (e.g. bilinear games), last iterate converges exponentially fast [?] (compared with $\mathcal{O}(1/T)$ for average iterate)

Convergence plot of EG – Forsaken game (non-manontone)

- For non-monotone case, we do *not* have guarantees for average iterates (Jensen's ineq not applicable), and even no guarantees for best iterate

$$\min_{|x|\leq 3/2} \max_{|y|\leq 3/2} \phi(x,y) := x(y-0.45) + \psi(x) - \psi(y) \quad \text{where } \psi(z) = 1/4 z^2 - 1/2 z^4 + 1/6 z^6 \tag{Forsaken}$$

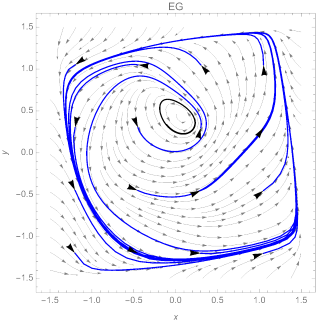


Figure: Last iterate convergence plot (limit cycle marked with dark cicle)

Go beyond monotone cases?

- Why Non-monotone? Nonmonotone *captures* nonconvex-nonconcave minimax by generalizing it.
- Weak Minty is the (*structured*) type of nonmonotonicity that we will study.
  - it captures the forsaken example we considered in the last slide
  - with weak Minty VI assumption, we could establish nice descent inequality (will appear in the next slide)

Definition (Variational Inequalities)

- Minty VI is a sufficient condition for optimality :$\langle F(z), z - z^\star \rangle \geq 0$
- **Weak Minty VI**: $\langle F(z), z - z^\star \rangle \geq \rho \|F(z)\|^2$, for some $\rho < 0$

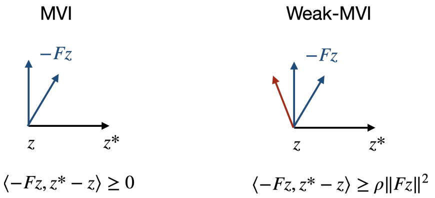


Figure: MVI versus weak MVI

Why can we have positive result for weak Minty?

- Consider a class of non-monotone problems characterized by Weak MVI
- Descent inequality for proximal point method ($z^{t+1} = z^t - \gamma F(z^{t+1})$):

Descent Inequality

$$\begin{aligned} \|z^{t+1} - z^\star\|^2 =& \|z^t - z^\star\|^2 + \gamma^2 \|F(z^{t+1})\|^2 - 2\langle z^t - z^\star, \gamma F(z^{t+1})\rangle \\ =& \|z^t - z^\star\|^2 + \gamma^2 \|F(z^{t+1})\|^2 - 2\langle z^{t+1} - z^\star, \gamma F(z^{t+1})\rangle - \underbrace{2\langle z^t - z^{t+1}, \gamma F(z^{t+1})\rangle}_{z^t - z^{t+1} = \gamma F z^{t+1}} \\ =& \|z^t - z^\star\|^2 - \gamma^2 \|z^{t+1} - z^t\|^2 - \underbrace{2\gamma\langle F(z^{t+1}), z^{t+1} - z^\star\rangle}_{\geq \rho \|F(z^{t+1})\|^2} \end{aligned} \tag{2}$$

To ensure strict descent between neighboring steps, we need

$$\gamma^2 + 2\gamma\rho > 0 \Rightarrow \rho > -\frac{\gamma}{2} \tag{3}$$

Weak Minty is the largest class for which PPM still has a decrease of distance to optimum every iteration.

An explicit scheme in unconstrained case: ExtraGradient+

EG+ is EG with a smaller 2nd step size

?] introduced EG+, which provably converges to a stationary point for a class of non-monotone problems provided Weak Minty Variational Inequalities (MVI).

$$\begin{aligned} \bar{z}^t &= (\mathrm{id} - \gamma F) z^t, \\ z^{t+1} &= z^t - \alpha \gamma F \bar{z}^t. \end{aligned} \tag{EG+}$$

We now show the choices of α, following the tighter analysis in **?**].

$$\|z^{t+1} - z^\star\|^2 = \|z^t - z^\star\|^2 + \alpha^2 \|\gamma F \bar{z}^t\|^2 - 2\alpha \langle \gamma F \bar{z}^t, z^t - z^\star \rangle \tag{4}$$

Cocoercivity assumption can "convert" $\langle F\bar{z}^t, z^t - z^\star\rangle$ into $\|F\bar{z}^t\|^2$!

Cocoercivity of $\mathrm{id} - \gamma F$

Definition (Cocoercivity)

An operator $F : \mathbb{R}^d \to \mathbb{R}^d$ is said to be β-cocoercive for $\beta > 0$ if $\langle Fz - Fz', z - z' \rangle \geq \beta \|Fz - Fz'\|^2$, for all $z, z' \in \mathbb{R}^d$.

Claim

Suppose F is L-Lipschitz and $\gamma \leq 1/L$. Then, the mapping $H = \mathrm{id} - \gamma F$ is 1/2-cocoercive

Proof.

$$\begin{aligned}
\langle Hz - H\bar{z}, z - \bar{z} \rangle &= \langle Hz - H\bar{z}, Hz - H\bar{z} + \gamma Fz - \gamma F\bar{z} \rangle \\
&= \|Hz - H\bar{z}\|^2 + \gamma \langle Hz - H\bar{z}, Fz - F\bar{z} \rangle \\
&\overset{?}{=} \frac{1}{2}\|Hz - H\bar{z}\|^2 - \frac{\gamma^2}{2}\|F\bar{z} - Fz\|^2 + \frac{1}{2}\|\bar{z} - z\|^2 \\
&= \frac{1}{2}\|Hz - H\bar{z}\|^2 + \Big(\frac{1}{2} - \frac{\gamma^2 L^2}{2}\Big)\|\bar{z} - z\|^2 \geq \frac{1}{2}\|Hz - H\bar{z}\|^2
\end{aligned} \tag{5}$$

?: $\frac{1}{2}\|\bar{z} - z\|^2 = \frac{1}{2}\|Hz - H\bar{z}\|^2 + \frac{\gamma^2}{2}\|F\bar{z} - Fz\|^2 + \gamma\langle H\bar{z} - Hz, F\bar{z} - Fz \rangle$ □

`EG+` requires a small constant step size

We use cocoercivity assumption and weak-Minty VI to bound the $\langle \gamma F\bar{z}^t, z^t - z^\star\rangle$ term:

$$\begin{aligned}\langle \gamma F\bar{z}^t, z^t - z^\star\rangle &= \underbrace{\langle \gamma F\bar{z}^t, z^t - \bar{z}^t\rangle}_{\text{id } -\gamma F \text{ is } 1/2 \text{ cocoercive}} + \underbrace{\langle \gamma F\bar{z}^t, \bar{z}^t - z^\star\rangle}_{\text{weak MVI}} \\ &\geq {}^1\!/\!{}_2\|\gamma F\bar{z}^t\|^2 + \gamma\rho\|F\bar{z}^t\|^2\end{aligned} \tag{6}$$

Thus, the descent inequality becomes

$$\|z^{t+1} - z^\star\|^2 \leq \|z^t - z^\star\|^2 + \underbrace{\alpha\gamma^2\left(\alpha - 1 - 2\rho/\gamma\right)}_{\text{make sure it is smaller than 0}} \|F\bar{z}^t\|^2 \tag{7}$$

Fixing $\gamma = 1/L$, to ensure a strict descent, we must have $\alpha < 1 + 2\rho L$. For a more negative ρ, i.e. a wider range of class, α needs to be very small.

Extend to a more general setting?

- Adding constraints characterized by the A operator: find $0 \in Tz \coloneqq Az + Fz$ for Lipschitz continuous operator F and maximally monotone operator A
- Allow for larger stepsizes (bigger α) and a wider range of problem

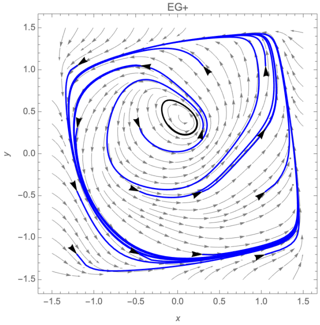


Figure: When the weak MVI constant ρ does not satisfy algorithmic requirements of (EG+), i.e. $\rho \in (-1/8L, 0]$, and (EG+) does not converge to a stationary point but rather the attracting limit cycle (marked in dark circle)

Algorithm

Compile error

Assumptions

Assumptions

For $T = A + F$. Given Assumptions

- Operator $A : \mathbb{R}^n \to \mathbb{R}^n$ is maximally monotone
- Operator $F : \mathbb{R}^n \to \mathbb{R}^n$ is L-lipschitz continuous
- Weak Minty condition holds: i.e. $\exists$ nonempty $S^\star \in \operatorname{zer} T$, s.t. $\forall z^\star \in S^\star$ and $\rho \in (-1/2L, +\infty)$

$$\langle v, z - z^\star \rangle \geq \rho \|v\|^2, \quad \forall (z, v) \in \mathbf{gph}(T)$$

Generalization of EG+

Theorem (Convergence to stationarity under weak MVI)

Let $\lambda_t \in (0,2)$, $\gamma_t \in \left(\lfloor -2\rho \rfloor_+, {}^1\!/_L\right]$ *where* $\lfloor x \rfloor_+ := \max\{0, x\}$, $\delta_t \in (-{}^{\gamma_t}\!/_2, \rho]$, $\liminf_{k\to\infty} \lambda_k(2-\lambda_k) > 0$, *and* $\liminf_{k\to\infty}(\delta_t + {}^{\gamma_t}\!/_2) > 0$. *Consider the sequences* $(z^t)_{k\in\mathbb{N}}$, $(\bar{z}^t)_{k\in\mathbb{N}}$ *generated by* $\left(\mathrm{id} + \gamma_t A\right)^{-1}\left(z^t - \gamma_t F z^t\right)$. *Then for all* $z^\star \in \mathcal{S}^\star$, *One can establish the following convergence result:*

$$\min_{k=0,..,m} \frac{1}{\gamma_t^2} \|H\bar{z}^t - Hz^t\|^2 \leq \frac{1}{\kappa(m+1)} \|z^0 - z^\star\|^2 \tag{8}$$

where $\kappa = \liminf_{k\to\infty} \lambda_k(2-\lambda_k)(\delta_k + {}^{\gamma_k}\!/_2)^2$.

Remark:

- $\mathbf{gph}(A) := \left\{(x,y) \in \mathbb{R}^n \times \mathbb{R}^d \mid y \in Ax\right\}$
- LHS denotes the gap between two updates: $z^{t+1} = z^t + \lambda_t \alpha_t (H\bar{z}^t - Hz^t)$

Proof ideas

- Construct a half space that contains the solution set:

$$\mathcal{D}(z) = \left\{ w \mid \langle \bar{v}, \bar{z} - w \rangle \geq \rho \|\bar{v}\|^2 \right\} \tag{9}$$

  onto which we could iteratively project: $\mathbf{\Pi}_{\mathcal{D}}(z^t)$
- To evaluate $\bar{v}$, use $\bar{v} \in T(\bar{z}^t)$. As $H = \mathrm{id} - \gamma F$, the update of $\bar{z}$ provides a way to evaluate T (show in the next slide)
- The algorithm update can be casted as KM iteration:

$$\begin{aligned} z^{t+1} &= (1 - \lambda_t)\, z^t + \lambda_t \mathbf{\Pi}_{\mathcal{D}^t}(z^t) \\ \text{with } \mathbf{\Pi}_{\mathcal{D}^t}(z^t) &= z^t + \alpha_t (H\bar{z}^t - Hz^t) \end{aligned} \tag{10}$$

- Use best iterate convergence result of KM iterate

Proof I – Preparation

- To use weak-MVI, we need to verify $\left(\bar{z}^t, {}^1\!/\gamma_t(Hz^t - H\bar{z}^t)\right) \in \mathbf{gph}(T)$
  As $H = \mathrm{id} - \gamma_t F$ and $\bar{z}^t = \left(\mathrm{id} + \gamma_t A\right)^{-1}\left(z^t - \gamma_t F z^t\right)$, we have

$$\bar{z}^t + \gamma_t A\bar{z}^t = Hz^t \tag{11}$$

$$\bar{z}^t - \gamma_t F\bar{z}^t + \gamma_t A\bar{z}^t = Hz^t - \gamma_t F\bar{z}^t \tag{12}$$

$$\frac{1}{\gamma_t}(Hz^t - H\bar{z}^t) \in A\bar{z}^t + F\bar{z}^t = T\bar{z}^t \tag{13}$$

- Verify H is 1/2-cocoercive: refer to slide 11

Proof II – Construct a half-space containing solution set

Following Lectue 3: each step of EG is a projection onto a particular hyperplane. We are aiming at the following construction:

$$\bar{z}^t = \left(\mathrm{id} + \gamma_t A\right)^{-1}\left(z^t - \gamma_t F z^t\right) \tag{14}$$

$$z^{t+1} = (1 - \lambda_t) z^t + \lambda_t \mathbf{\Pi}_{\mathcal{D}_t}(z^t) \tag{15}$$

where $\mathcal{D}_t$ is constructed to *contain the solution set*. As $\frac{1}{\gamma_t}(Hz^t - H\bar{z}^t) \in \mathrm{gph}(T)$, we have

$$\left\langle \frac{1}{\gamma_t}(Hz^t - H\bar{z}^t), \bar{z}^t - z^\star \right\rangle \overset{\text{Weak-MVI}}{\geq} \frac{\rho}{\gamma_t^2}\|Hz^t - H\bar{z}^t\|^2 \overset{\rho \geq \delta_t}{\geq} \frac{\delta_t}{\gamma_t^2}\|Hz^t - H\bar{z}^t\|^2 \tag{16}$$

Thus we can construct $\mathcal{D}_t$ as

$$\mathcal{D}_t \coloneqq \left\{ w \mid \langle Hz^t - H\bar{z}^t, \bar{z}^t - w \rangle \geq \frac{\delta_t}{\gamma_t}\|Hz^t - H\bar{z}^t\|^2 \right\} \tag{17}$$

Note due to the $1/2$-cocoercivity of H, our chosen step size α_t is positive and bounded away from 0:

$$\alpha_t = \frac{\delta_t}{\gamma_t} + \frac{\langle \bar{z}^t - z^t, H\bar{z}^t - Hz^t \rangle}{\|H\bar{z}^t - Hz^t\|^2} \geq \frac{1}{2} + \frac{\delta_t}{\gamma_t} \tag{18}$$

Proof III - Projection onto the constructed half space

Lemma

The projection $\mathbf{\Pi}_{\mathcal{D}}(x) := \arg\min_{z\in\mathcal{D}} \frac{1}{2}\|z-x\|^2$ *onto the set* $\mathcal{D} = \{z \mid \langle a, z\rangle \geq b\}$ *is given for* $x \notin \mathcal{D}$ *as,*

$$\mathbf{\Pi}_{\mathcal{D}}(x) = x - \frac{\langle a, x\rangle - b}{\|a\|^2} a. \tag{19}$$

The projection onto $\mathcal{D}_t$ for any $v \notin \mathcal{D}_t$ is given by

$$\mathbf{\Pi}_{\mathcal{D}_t}(v) = v + \frac{\langle \bar{z}^t - v, Hz^t - H\bar{z}^t\rangle - \frac{\delta_t}{\gamma_t}\|Hz^t - H\bar{z}^t\|^2}{\|Hz^t - H\bar{z}^t\|^2}(Hz^t - H\bar{z}^t)$$

For $v = z^t$, we have

$$\mathbf{\Pi}_{\mathcal{D}_t}(z^t) = z^t + \alpha_t(H\bar{z}^t - Hz^t) \tag{20}$$

We can thus rewrite an update step as

$$z^{t+1} = (1-\lambda_t)z^t + \lambda_t\left(z^t - \alpha(Hz^t - H\bar{z}^t)\right) \tag{21}$$

Proof IV - Formulate as KM iteration

Krasnosel'skii-Mann (KM) iteration

Let $S : \mathbb{R}^d \to \mathbb{R}^d$ be an operator and $\lambda > 0$. The KM iteration is given by

$$z^{t+1} = (1 - \lambda) z^t + \lambda S z^t \tag{KM}$$

Remarks

- An operator $S : \mathbb{R}^d \to \mathbb{R}^d$ is said to be firmly nonexpansive if

$$\|Sz - Sz'\|^2 + \|(\mathrm{id} - S)z - (\mathrm{id} - S)z'\|^2 \leq \|z - z'\|^2 \quad \forall z, z' \in \mathbb{R}^d. \tag{22}$$

- Let C be a nonempty closed convex subset. Then the projector $\mathbf{\Pi}_C$ is firmly nonexpansive.

As $\mathcal{D}_t$ is closed convex set, $\mathbf{\Pi}_{\mathcal{D}_t}$ is firmly non-expansive

Proof V - Best iterate result of KM

Theorem (Best iterate of KM - Recall Lecture 3)

Suppose $S : \mathbb{R}^d \to \mathbb{R}^d$ *is firmly nonexpansive. Consider the sequence* $(z^t)_{t\in\mathbb{N}}$ *generated by* **??** *with* $\lambda \in (0, 2)$*.*
Then, for all $z^\star \in \operatorname{fix} S$*, it holds that*

$$\min_{t\in\{0,\ldots,T-1\}} \|Sz^t - z^t\|^2 \leq \frac{\|z^0 - z^\star\|^2}{\lambda(2-\lambda)T}. \tag{23}$$

Directly apply the theorem, we have

$$\min_{t=0,\ldots,m} \|\mathbf{\Pi}_{\mathcal{D}_t} z^t - z^t\|^2 \leq \frac{\|z^0 - z^\star\|^2}{\lambda_t(2-\lambda_t)(m+1)} \tag{24}$$

$$\|\mathbf{\Pi}_{\mathcal{D}_t} z^t - z^t\|^2 = \alpha_t^2 \|H\bar{z}^t - Hz^t\|^2 \geq \Big(\frac{1}{2} + \frac{\delta_t}{\gamma_t}\Big)^2 \|H\bar{z}^t - Hz^t\|^2 \tag{25}$$

Choose $\kappa = \liminf_{t\to\infty} \lambda_t(2-\lambda_t)(\frac{\gamma_t}{2} + \delta_t)^2$, we directly have

$$\min_{t=0,\ldots,m} \frac{1}{\gamma_t^2} \|Hz^t - H\bar{z}^t\|^2 \leq \frac{1}{\kappa(1+m)} \|z^0 - z^\star\|^2 \tag{26}$$

Additional results on the last iterate – $\bar{z}^t$

Limit points of $\bar{z}^t$ belong to $\operatorname{zer} T$.

Claim 1

Following the same assumptions as the main theorem, $(\bar{z}^t)_{k\in\mathbb{N}}$ is bounded and its limit points belong to $\operatorname{zer} T$;

Proof Ideas.

- Since $\liminf_{t\to\infty} \varepsilon_t \coloneqq \lambda_t(2-\lambda_t)(\frac{\gamma_t}{2}+\delta_t)^2 > 0$, as

$$\|z^{t+1}-z^\star\|^2 \leq \|z^t - z^\star\|^2 - \frac{\varepsilon_t}{\gamma_t^2}\|Hz^t - H\bar{z}^t\|^2$$

  $(\frac{1}{\gamma_t^2}\|H\bar{z}^t - Hz^t\|^2)_{k\in\mathbb{N}}$ converges to zero.
- $\|z^t - z^\star\|^2$ converges, we have z^t bounded. γ_t is bounded, F and the resolvents $(\mathrm{id}+\gamma_t A)^{-1}$ are Lipschitz continuous, so is their composition. Thus, $\bar{z}^t$ is also bounded.
- $(\bar{z}^t, 1/\gamma_t(Hz^t - H\bar{z}^t)) \in \mathrm{gph}(T)$, that is

$$1/\gamma_t\left(Hz^t - H\bar{z}^t\right) \in T\bar{z}^t$$

  As $\lim_{t\to\infty} 1/\gamma_t\left(Hz^t - H\bar{z}^t\right) = 0$, we have $0 \in T\bar{z}^t$

Additional results on the last iterate – z^t

Limit points of z^t stay close to $\bar{z}^t$.

Claim 2

Following the same assumptions as the main theorem, if in addition $\limsup_{k\to\infty} \gamma_t < 1/L$ and $\mathcal{S}^\star = \operatorname{zer} T$, then $(z^t)_{k\in\mathbb{N}}$, $(\bar{z}^t)_{k\in\mathbb{N}}$ both converge to some $z^\star \in \operatorname{zer} T$.

Lemma

If A is L-Lipschitz, then $T = \operatorname{id} - \eta A$, $\eta \in (0, 1/L)$, is $(1-\eta L)$-monotone, and in particular $\|Tu - Tv\| \geq (1-\eta L)\|u - v\|$ for all $u, v \in \mathbb{R}^n$.

- if $\gamma = \limsup_{k\to\infty} \gamma_t < 1/L$, following Lemma **??**, we have

$$(1-\gamma L)\|\bar{z}^t - z^t\| \leq \|H\bar{z}^t - Hz^t\|$$

Therefore, $(\|\bar{z}^t - z^t\|)_{k\in\mathbb{N}}$ converges to zero. Hence, $(z^t)_{k\in\mathbb{N}}[k \in K]$ also converges to $\bar{z} \in \operatorname{zer} T$.

Connection to other existent algorithms

A constant step size variant

$$\bar{z}^t = (\mathrm{id} + \gamma_k A)^{-1}(z^t - \gamma F z^t), \quad z^{t+1} = z^t - \bar{\alpha}_t (H\bar{z}^t - Hz^t). \tag{CEG+}$$

According to the main theorem, convergence is guaranteed when $\lambda_t \in (0, 2)$, that is when $\bar{\alpha}_t < 2\alpha_t$. We could choose constant stepsize $\bar{\alpha} \in \left(0, 1 + \frac{2\delta}{\gamma}\right)$

$$\bar{\alpha}_t < 1 + \frac{2\delta_t}{\gamma_t} \leq \frac{2\delta_t}{\gamma_t} + \frac{2\langle \bar{z}^t - z^t, H\bar{z}^t - Hz^t \rangle}{\|H\bar{z}^t - Hz^t\|^2} = 2\alpha_t \tag{27}$$

- **Larger second stepsize than `EG+`.** As the setting of `EG+`, in the unconstrained case, i.e. $A \equiv 0$, for the smallest ρ that `EG+` allows ($\rho = -1/8L$), `EG+` chooses $\gamma_t = 1/L$ and $\alpha_t = 1/2$. While we can select $\gamma_t = 1/L$ and $\alpha_t \in (0, 3/4)$ here.
- **Connection to `FBF`.** In the monotone case. i.e. $\rho = 0$, choose $\gamma_t \in (0, 1/L)$ and $\alpha = 1$, we have $z^{t+1} = \bar{z}^t + \gamma F z^t - \gamma F \bar{z}^t$, which is the forward-backward-forward algorithm proposed by **?**]

Adaptive Step Size according to Local Curvature

- A larger stepsize γ_t would guarantee global convergence to a wider class of problem, as $\rho > -\frac{\gamma_t}{2}$
- Global Lipschitz constant is inherently pessimistic – use local curvature instead

○ Adaptive in γ_t

Algorithm Lipschitz constant backtracking

Initialize $z^t \in \mathbb{R}^n$, $\tau \in (0,1)$, $\nu \in (0,1)$

Set initial guess $\gamma = \nu \|JF(z^t)\|^{-1}$, and let $G_\gamma(z^t) := \left(\mathrm{id} + \gamma A\right)^{-1}\left(z^t - \gamma F z^t\right)$

while $\gamma\|F(G_\gamma(z^t)) - Fz^t\| > \nu\|G_\gamma(z^t) - z^t\|$ **do** $\gamma \leftarrow \tau\gamma$

Return $\gamma_t = \gamma$ and $\bar{z}^t = G_\gamma(z^t)$

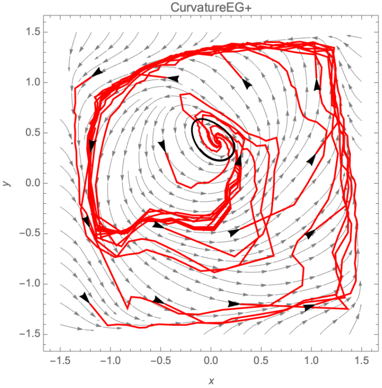


Figure: `CurvatureEG+` is able to escape the attractive limit cycle in the Forsaken game

Overview of the stochastic part

- In the last lecture,
  - Extra-gradient+ algorithms with adaptive stepsize;
  - Convergence result for deterministic case.
- This lecture,
  - Extra-gradient+ algorithm in the **stochastic** setting: *Bias-corrected* algorithm;
  - Prove the convergence of the bias-corrected algorithm in **unconstrained** case;
  - Extend the algorithm to **constrained** case;
  - Extend the algorithm to the primal-dual setting with asynchronous update;
- Thomas Pethick, Olivier Fercoq, Puya Latafat, Panagiotis Patrinos, and Volkan Cevher. *Solving Stochastic Weak Minty Variational Inequalities Without Increasing Batch Size*. In The Eleventh International Conference on Learning Representations, 2023. URL https://openreview.net/forum?id=ejR4E1jaH9k.

Recall: formulation

Find $z \in \mathbb{R}^n$, such that

$$0 \in Tz := Az + Fz, \tag{28}$$

where A maps to a set (captures the constraints) and F maps to a single vector (captures the objective).

Recall: formulation

Find $z \in \mathbb{R}^n$, such that

$$0 \in Tz := Az + Fz, \tag{28}$$

where A maps to a set (captures the constraints) and F maps to a single vector (captures the objective).

- **Limitation of the last lecture**: Need exact **deterministic** evaluation of Fz.

Recall: formulation

Find $z \in \mathbb{R}^n$, such that

$$0 \in Tz := Az + Fz, \tag{28}$$

where A maps to a set (captures the constraints) and F maps to a single vector (captures the objective).

- **Limitation of the last lecture**: Need exact **deterministic** evaluation of Fz.
- Such **deterministic** evaluation can be expensive or even unavailable.
  - E.g., the gradient of training a deep neural network.
- EG+: needs large enough exploration stepsize. Otherwise, may get stuck in a limit cycle.

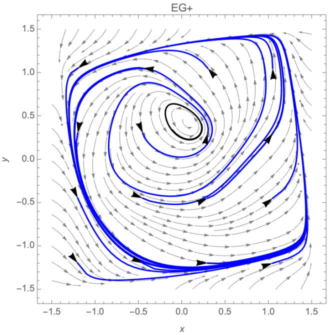
Recall: formulation

Find $z \in \mathbb{R}^n$, such that

$$0 \in Tz := Az + Fz, \tag{28}$$

where A maps to a set (captures the constraints) and F maps to a single vector (captures the objective).

- **Limitation of the last lecture**: Need exact **deterministic** evaluation of Fz.
- Such **deterministic** evaluation can be expensive or even unavailable.
  - E.g., the gradient of training a deep neural network.
- EG+: needs large enough exploration stepsize. Otherwise, may get stuck in a limit cycle.



Stochastic oracle

◦ Instead, the evaluation of Fz can be **stochastic**.

Stochastic oracle

- Instead, the evaluation of Fz can be **stochastic**.
- We can not directly evaluate the deterministic Fz, but get access to a stochastic oracle $\hat{F}(z, \xi)$.

Stochastic oracle

- Instead, the evaluation of Fz can be **stochastic**.
- We can not directly evaluate the deterministic Fz, but get access to a stochastic oracle $\hat{F}(z, \xi)$.
- E.g., Stochastic Gradient Descent (SGD) for training deep neural network.
  - $\min_x f(x) \coloneqq \frac{1}{N} \sum_{i=1}^{N} f_i(x)$
  - At step t, $x^{t+1} = x^t - \eta \nabla f_{i_t}(x^t)$, where i_t is randomly sampled from $[N]$.
  - In this example, $z \coloneqq x$, $\xi \coloneqq i_t$, $\hat{F}(z, \xi) \coloneqq \nabla f_{i_t}(x^t)$.

Stochastic oracle

- Instead, the evaluation of Fz can be **stochastic**.
- We can not directly evaluate the deterministic Fz, but get access to a stochastic oracle $\hat{F}(z,\xi)$.
- E.g., Stochastic Gradient Descent (SGD) for training deep neural network.
  - $\min_x f(x) := \frac{1}{N}\sum_{i=1}^{N} f_i(x)$
  - At step t, $x^{t+1} = x^t - \eta\nabla f_{i_t}(x^t)$, where i_t is randomly sampled from $[N]$.
  - In this example, $z := x$, $\xi := i_t$, $\hat{F}(z,\xi) := \nabla f_{i_t}(x^t)$.

Assumption (Assumptions on $\hat{F}(z,\xi)$)

For the operator $\hat{F}(\cdot,\xi) : \mathbb{R}^n \to \mathbb{R}^n$, the following holds.

1. ***Unbiased:*** $\mathbb{E}_\xi[\hat{F}(z,\xi)] = Fz \quad \forall z \in \mathbb{R}^n$.
2. ***Bounded variance:*** $\mathbb{E}_\xi\left[\|\hat{F}(z,\xi) - F(z)\|^2\right] \le \sigma_F^2 \quad \forall z \in \mathbb{R}^n$.

Stochastic oracle

- Instead, the evaluation of Fz can be **stochastic**.
- We can not directly evaluate the deterministic Fz, but get access to a stochastic oracle $\hat{F}(z, \xi)$.
- E.g., Stochastic Gradient Descent (SGD) for training deep neural network.
  - $\min_x f(x) := \frac{1}{N} \sum_{i=1}^{N} f_i(x)$
  - At step t, $x^{t+1} = x^t - \eta \nabla f_{i_t}(x^t)$, where i_t is randomly sampled from $[N]$.
  - In this example, $z := x$, $\xi := i_t$, $\hat{F}(z, \xi) := \nabla f_{i_t}(x^t)$.

Assumption (Assumptions on $\hat{F}(z, \xi)$)

For the operator $\hat{F}(\cdot, \xi) : \mathbb{R}^n \to \mathbb{R}^n$, the following holds.

1. ***Unbiased***: $\mathbb{E}_\xi[\hat{F}(z, \xi)] = Fz \quad \forall z \in \mathbb{R}^n$.
2. ***Bounded variance***: $\mathbb{E}_\xi\left[\|\hat{F}(z, \xi) - F(z)\|^2\right] \leq \sigma_F^2 \quad \forall z \in \mathbb{R}^n$.

Remarks:

- Assumption 1 holds for SGD when i_t is uniformly sampled.
- Assumption 2 is common. Easily satisfied when z restricted to a compact set.

Why 'unbiased' and 'bounded variance'?

- Stochastic Gradient Descent (SGD).
  - $\min f(x) := \frac{1}{N}\sum_{i=1}^{N} f_i(x)$
  - At step t, $x^{t+1} = x^t - \eta\nabla f_{i_t}(x^t)$, where i_t is randomly sampled from $[N]$.
- **Unbiased**: $\mathbb{E}[\nabla f_{i_t}(x^t)] = \frac{1}{N}\sum_{i=1}^{N} \nabla f_i(x) = \nabla f(x^t)$.
- **Bounded variance**: Assume $\mathbb{E}[\|\nabla f_{i_t}(x^t) - \nabla f(x^t)\|^2] \le \sigma_f^2$.

Why ‘unbiased’ and ‘bounded variance’?

◦ $x^\star \in \arg\min_x f(x)$.

◦ $g_t \coloneqq \nabla f_{i_t}(x^t)$.

Why 'unbiased' and 'bounded variance'?

- $x^\star \in \arg\min_x f(x)$.
- $g_t \coloneqq \nabla f_{i_t}(x^t)$.
- One-step iterate of SGD.

$$
\begin{aligned}
&\|x^{t+1} - x^\star\|^2 \\
=&\|x^t - x^\star - \eta g_t\|^2
\end{aligned}
$$

Why 'unbiased' and 'bounded variance'?

- $x^\star \in \arg\min_x f(x)$.
- $g_t := \nabla f_{i_t}(x^t)$.
- One-step iterate of SGD.

$$
\begin{aligned}
&\|x^{t+1} - x^\star\|^2 \\
=&\|x^t - x^\star - \eta g_t\|^2 \\
=&\|x^t - \eta g_t\|^2 - 2\eta\langle x^t - x^\star, g_t\rangle + \eta^2\|g_t\|^2 \\
=&\|x^t - x^\star\|^2 - 2\eta\langle x^t - x^\star, \nabla f(x^t)\rangle - 2\eta\langle x^t - x^\star, g_t - \nabla f(x^t)\rangle + \eta^2\|\nabla f(x^t) + g_t - \nabla f(x^t)\|^2
\end{aligned}
$$

Why 'unbiased' and 'bounded variance'?

- $x^\star \in \arg\min_x f(x)$.
- $g_t := \nabla f_{i_t}(x^t)$.
- One-step iterate of SGD.

$$
\begin{aligned}
&\|x^{t+1} - x^\star\|^2 \\
=&\|x^t - x^\star - \eta g_t\|^2 \\
=&\|x^t - \eta g_t\|^2 - 2\eta\langle x^t - x^\star, g_t\rangle + \eta^2\|g_t\|^2 \\
=&\|x^t - x^\star\|^2 - 2\eta\langle x^t - x^\star, \nabla f(x^t)\rangle - 2\eta\langle x^t - x^\star, g_t - \nabla f(x^t)\rangle + \eta^2\|\nabla f(x^t) + g_t - \nabla f(x^t)\|^2 \\
=&\underbrace{\|x^t - x^\star\|^2 - 2\eta\langle x^t - x^\star, \nabla f(x^t)\rangle + \eta^2\|\nabla f(x^t)\|^2}_{\text{deterministic terms}} \\
&\underbrace{-2\eta\langle x^t - x^\star, g_t - \nabla f(x^t)\rangle + 2\eta^2\langle \nabla f(x^t), g_t - \nabla f(x^t)\rangle}_{\text{has expectation 0 by the 'unbiased' property}} \\
&+ \qquad \underbrace{\eta^2\|g_t - \nabla f(x^t)\|^2}_{\leq \eta^2\sigma_f^2 \text{ in expectation by the 'bounded variance' property}}
\end{aligned}
$$

Recall: assumptions

Assumption (Assumptions on F and A)

1. *The operator $F : \mathbb{R}^n \to \mathbb{R}^n$ is L_F-Lipschitz with $L_F \in [0, \infty)$, i.e.,*

$$\left\| Fz - Fz' \right\| \leq L_F \left\| z - z' \right\| \quad \forall z, z' \in \mathbb{R}^n.$$

2. *The operator $A : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is a maximally monotone operator.*
3. *Weak Minty variational inequality (MVI) holds, i.e., there exists a nonempty set $\mathcal{S}^\star \subseteq$ zer T such that for all $z^\star \in \mathcal{S}^\star$ and some $\rho \in \left(-\frac{1}{2L_F}, \infty\right)$*

$$\langle v, z - z^\star \rangle \geq \rho \|v\|^2, \quad \text{for all } (z, v) \in \mathbf{gph} T.$$

Further assumptions on $\hat{F}(z, \xi)$

Assumption

- ***Two-point oracle***: *The stochastic oracle can be queried for any two points* $z, z' \in \mathbb{R}^n$, $\hat{F}(z, \xi), \hat{F}(z', \xi)$ *where* $\xi \sim \mathcal{P}$.
- ***Lipschitz continuity***: *The operator* $\hat{F}(\cdot, \xi) : \mathbb{R}^n \to \mathbb{R}^n$ *is Lipschitz continuous in mean with* $L_{\hat{F}} \in [0, \infty)$ *:*

$$\mathbb{E}_{\xi}\left[\left\|\hat{F}(z, \xi) - \hat{F}\left(z', \xi\right)\right\|^2\right] \leqslant L_{\hat{F}}^2 \left\|z - z'\right\|^2$$

for all $z, z' \in \mathbb{R}^n$.

Further assumptions on $\hat{F}(z, \xi)$

Assumption

- ***Two-point oracle***: *The stochastic oracle can be queried for any two points* $z, z' \in \mathbb{R}^n$, $\hat{F}(z, \xi), \hat{F}(z', \xi)$ *where* $\xi \sim \mathcal{P}$.
- ***Lipschitz continuity***: *The operator* $\hat{F}(\cdot, \xi) : \mathbb{R}^n \to \mathbb{R}^n$ *is Lipschitz continuous in mean with* $L_{\hat{F}} \in [0, \infty)$ *:*

$$\mathbb{E}_{\xi}\left[\left\|\hat{F}(z, \xi) - \hat{F}\left(z', \xi\right)\right\|^2\right] \leqslant L_{\hat{F}}^2 \left\|z - z'\right\|^2$$

  for all $z, z' \in \mathbb{R}^n$.

Remarks:

- **Two-point oracle** assumption is reasonable for SGD by choosing the same i_t.
- Will see: **Lipschitz continuity** critical for algorithm design and variance reduction.

Oblivious extension of EG+?

- Extension of EG+ with $A \equiv 0$ to the stochastic setting:

$$\bar{z}^t = z^t - \gamma \frac{1}{B} \sum_{i=1}^{B} \hat{F}\left(z^t, \xi_{t,i}\right), \quad z^{t+1} = z^t - \alpha_t \gamma \frac{1}{B} \sum_{i=1}^{B} \hat{F}\left(\bar{z}^t, \bar{\xi}_{t,i}\right). \tag{EG+}$$

Oblivious extension of EG+?

- Extension of EG+ with $A \equiv 0$ to the stochastic setting:

$$\bar{z}^t = z^t - \gamma \frac{1}{B} \sum_{i=1}^{B} \hat{F}\left(z^t, \xi_{t,i}\right), \quad z^{t+1} = z^t - \alpha_t \gamma \frac{1}{B} \sum_{i=1}^{B} \hat{F}\left(\bar{z}^t, \bar{\xi}_{t,i}\right). \tag{EG+}$$

- Let $\tilde{\bar{z}}^t := z^t - \gamma F z^t$ and $\tilde{z}^{t+1} := z^t - \alpha_t \gamma F \tilde{\bar{z}}^t$.
- Due to the unbiased assumption of $\hat{F}$, we have $\mathbb{E}_{\xi_t}(\bar{z}^t) = \tilde{\bar{z}}^t$.

Oblivious extension of EG+?

- Extension of EG+ with $A \equiv 0$ to the stochastic setting:

$$\bar{z}^t = z^t - \gamma \frac{1}{B} \sum_{i=1}^{B} \hat{F}\left(z^t, \xi_{t,i}\right), \quad z^{t+1} = z^t - \alpha_t \gamma \frac{1}{B} \sum_{i=1}^{B} \hat{F}\left(\bar{z}^t, \bar{\xi}_{t,i}\right). \tag{EG+}$$

- Let $\tilde{\bar{z}}^t := z^t - \gamma F z^t$ and $\tilde{z}^{t+1} := z^t - \alpha_t \gamma F \tilde{\bar{z}}^t$.
- Due to the unbiased assumption of $\hat{F}$, we have $\mathbb{E}_{\xi_t}(\bar{z}^t) = \tilde{\bar{z}}^t$.

Observation

When F is linear, we have,

$$\mathbb{E}_{\xi_t, \bar{\xi}_t}(z^{t+1}) = z^t - \alpha_t \gamma \mathbb{E}_{\xi_t}\left(\frac{1}{B} \sum_{i=1}^{B} \mathbb{E}_{\bar{\xi}_t | \xi_t} \hat{F}\left(\bar{z}^t, \bar{\xi}_{t,i}\right)\right) = z^t - \alpha_t \gamma \mathbb{E}_{\xi_t} F \bar{z}^t = z^t - \alpha_t \gamma F(\mathbb{E}_{\xi_t} \bar{z}^t) = \tilde{z}^{t+1} \tag{Linear F}$$

Remark: The linearity of F leads to unbiased update of z and **convergence result** when $A \equiv 0$.

What if F is nonlinear or $A \not\equiv 0$?

Observation

In general,

$$\mathbb{E}_{\xi_t, \bar{\xi}_t}(z^{t+1}) = z^t - \alpha_t \gamma \mathbb{E}_{\xi_t} F \bar{z}^t \neq z^t - \alpha_t \gamma F(\mathbb{E}_{\xi_t} \bar{z}^t) = \tilde{z}^{t+1} \qquad \text{(Nonlinear } F\text{)}$$

- **Bias** appears with nonlinear F.

What if F is nonlinear or $A \not\equiv 0$?

Observation

In general,

$$\mathbb{E}_{\xi_t,\bar{\xi}_t}(z^{t+1}) = z^t - \alpha_t \gamma \mathbb{E}_{\xi_t} F \bar{z}^t \neq z^t - \alpha_t \gamma F(\mathbb{E}_{\xi_t} \bar{z}^t) = \tilde{z}^{t+1} \qquad \text{(Nonlinear } F\text{)}$$

- **Bias** appears with nonlinear F.
- How to reduce the bias in z^{t+1}?
  - Ideally, the above equality holds with deterministic extrapolation point $\bar{z}^t$.
  - The smaller the variance of $\bar{z}^t$, the better.

What if F is nonlinear or $A \not\equiv 0$?

Observation

In general,

$$\mathbb{E}_{\xi_t,\bar{\xi}_t}(z^{t+1}) = z^t - \alpha_t \gamma \mathbb{E}_{\xi_t} F \bar{z}^t \neq z^t - \alpha_t \gamma F(\mathbb{E}_{\xi_t} \bar{z}^t) = \tilde{z}^{t+1} \qquad \text{(Nonlinear } F\text{)}$$

- **Bias** appears with nonlinear F.
- How to reduce the bias in z^{t+1}?
  - Ideally, the above equality holds with deterministic extrapolation point $\bar{z}^t$.
  - The smaller the variance of $\bar{z}^t$, the better.
- Two obvious approaches:
  - Increasing batchsize B. But can be too expensive. [**?** , Thm. 4.5]
  - Reduce the stepsize γ. But contradicts the requirement of large enough stepsize γ for the convergence in weak MVI. [**?**]

Bias-corrected Approach

- Idea: leveraging **two-point** oracle.

Algorithm (BC-SEG+) Stochastic algorithm for problem (**??**) when $A \equiv 0$

REQUIRE $z^{-1} = \bar{z}^{-1} = z^0 \in \mathbb{R}^n, \alpha_t \in (0,1), \gamma \in (\lfloor -2\rho \rfloor_+, 1/L_F)$
for $t = 0, 1, \ldots$ until convergence **do**
  Sample $\xi_t \sim \mathcal{P}$
    $\bar{z}^t = z^t - \gamma \hat{F}\left(z^t, \xi_t\right) + (1-\alpha_t)\left(\bar{z}^{t-1} - z^{t-1} + \gamma \hat{F}\left(z^{t-1}, \xi_t\right)\right)$
  Sample $\bar{\xi}_t \sim \mathcal{P}$
    $z^{t+1} = z^t - \alpha_t \gamma \hat{F}\left(\bar{z}^t, \bar{\xi}_t\right)$
end for
RETURN z^{t+1}

Bias-corrected Approach

- Idea: leveraging **two-point** oracle.

Algorithm (BC-SEG+) Stochastic algorithm for problem (**??**) when $A \equiv 0$

1: REQUIRE $\quad z^{-1} = \bar{z}^{-1} = z^0 \in \mathbb{R}^n, \alpha_t \in (0,1), \gamma \in (\lfloor -2\rho \rfloor_+, 1/L_F)$
2: **for** $t = 0, 1, \ldots$ until convergence **do**
3: $\quad$ Sample $\xi_t \sim \mathcal{P}$
4: $\qquad \bar{z}^t = z^t - \gamma \hat{F}\left(z^t, \xi_t\right) + (1-\alpha_t)\left(\bar{z}^{t-1} - z^{t-1} + \gamma \hat{F}\left(z^{t-1}, \xi_t\right)\right)$
5: $\quad$ Sample $\bar{\xi}_t \sim \mathcal{P}$
6: $\qquad z^{t+1} = z^t - \alpha_t \gamma \hat{F}\left(\bar{z}^t, \bar{\xi}_t\right)$
7: **end for**
8: RETURN z^{t+1}

- Line 4 is the key step.
- Idea: add a 'correction' term to reduce the variance of $\bar{z}^t$. (Similar idea in STORM algorithm [**?**].)
- Correction term = Real Preceding Extrapolation Point - Imagined Stochastic Extrapolation Point with ξ_t

Why intuitively the algorithm works

◦ Approximate error of the stochastic extrapolation point.

$$u^t := \bar{z}^t - \left(z^t - \gamma F\left(z^t\right)\right) \tag{29}$$

Why intuitively the algorithm works

- Approximate error of the stochastic extrapolation point.

$$u^t := \bar{z}^t - \left(z^t - \gamma F\left(z^t\right)\right) \tag{29}$$

- The error u^t can be decomposed as,

$$\begin{aligned}\left\|u^t\right\| &= \left\|\gamma F z^t - \gamma \hat{F}\left(z^t, \xi_t\right) + (1-\alpha_t)\left(u^{t-1} - \gamma F z^{t-1} + \gamma \hat{F}\left(z^{t-1}, \xi_t\right)\right)\right\| && (30)\\ &= \left\|(1-\alpha_t)\, u^{t-1} + \alpha_t \gamma\left(F z^t - \hat{F}\left(z^t, \xi_t\right)\right) + (1-\alpha_t)\,\gamma\left(\left(F z^t - F z^{t-1}\right) + \left(\hat{F}\left(z^{t-1}, \xi_t\right) - \hat{F}\left(z^t, \xi_t\right)\right)\right)\right\| && (31)\end{aligned}$$

Why intuitively the algorithm works

- Approximate error of the stochastic extrapolation point.

$$u^t := \bar{z}^t - \left(z^t - \gamma F\left(z^t\right)\right) \tag{29}$$

- The error u^t can be decomposed as,

$$\left\|u^t\right\| = \left\|\gamma F z^t - \gamma \hat{F}\left(z^t, \xi_t\right) + (1-\alpha_t)\left(u^{t-1} - \gamma F z^{t-1} + \gamma \hat{F}\left(z^{t-1}, \xi_t\right)\right)\right\| \tag{30}$$

$$= \left\|(1-\alpha_t)\, u^{t-1} + \alpha_t \gamma \left(F z^t - \hat{F}\left(z^t, \xi_t\right)\right) + (1-\alpha_t)\,\gamma\left(\left(F z^t - F z^{t-1}\right) + \left(\hat{F}\left(z^{t-1}, \xi_t\right) - \hat{F}\left(z^t, \xi_t\right)\right)\right)\right\| \tag{31}$$

$$\leq (1-\alpha_t)\left\|u^{t-1}\right\| + \alpha_t \gamma \underbrace{\left\|F z^t - \hat{F}\left(z^t, \xi_t\right)\right\|}_{\approx \mathcal{O}(\sigma_F)} + (1-\alpha_t)\,\gamma \underbrace{\left(\left\|F z^t - F z^{t-1}\right\| + \left\|\hat{F}\left(z^{t-1}, \xi_t\right) - \hat{F}\left(z^t, \xi_t\right)\right\|\right)}_{\approx \mathcal{O}\left((L_F + L_{\hat{F}})\|z^t - z^{t-1}\|\right)} \tag{32}$$

  - Control the second term by $\alpha_t \to 0$.
  - Control the third term by Lipschitz property.

Why intuitively the algorithm works

- Approximate error of the stochastic extrapolation point.

$$u^t := \bar{z}^t - \left(z^t - \gamma F\left(z^t\right)\right) \tag{29}$$

- The error u^t can be decomposed as,

$$\begin{aligned}
\left\|u^t\right\| &= \left\|\gamma F z^t - \gamma \hat{F}\left(z^t, \xi_t\right) + (1-\alpha_t)\left(u^{t-1} - \gamma F z^{t-1} + \gamma \hat{F}\left(z^{t-1}, \xi_t\right)\right)\right\| && (30)\\
&= \left\|(1-\alpha_t)\, u^{t-1} + \alpha_t \gamma \left(F z^t - \hat{F}\left(z^t, \xi_t\right)\right) + (1-\alpha_t)\,\gamma\left(\left(F z^t - F z^{t-1}\right) + \left(\hat{F}\left(z^{t-1}, \xi_t\right) - \hat{F}\left(z^t, \xi_t\right)\right)\right)\right\| && (31)\\
&\leq (1-\alpha_t)\left\|u^{t-1}\right\| + \alpha_t \gamma \underbrace{\left\|F z^t - \hat{F}\left(z^t, \xi_t\right)\right\|}_{\approx \mathcal{O}(\sigma_F)} + (1-\alpha_t)\,\gamma \underbrace{\left(\left\|F z^t - F z^{t-1}\right\| + \left\|\hat{F}\left(z^{t-1}, \xi_t\right) - \hat{F}\left(z^t, \xi_t\right)\right\|\right)}_{\approx \mathcal{O}\left((L_F + L_{\hat{F}})\|z^t - z^{t-1}\|\right)} && (32)
\end{aligned}$$

  - Control the second term by $\alpha_t \to 0$.
  - Control the third term by Lipschitz property.
- $\|u^t\|$ can be controlled.

Convergence guarantee in unconstrained case

Theorem (Random Iterate Convergence)

Suppose the three assumptions on $F, \hat{F}, A$ hold. Suppose in addition that $\gamma \in (\lfloor -2\rho \rfloor_+, 1/L_F)$ and $(\alpha_t)_{t\in[T]} \subset (0,1)$ is a diminishing sequence such that

$$2\gamma L_{\hat{F}}\sqrt{\alpha_0} + \left(1 + \left(\frac{1+\gamma^2 L_F^2}{1-\gamma^2 L_F^2}\gamma^2 L_F^2\right)\gamma^2 L_{\hat{F}}^2\right)\alpha_0 \leq 1 + \frac{2\rho}{\gamma}$$

Then, the following estimate holds for all $z^\star \in \mathcal{S}^\star$

$$\mathbb{E}\left[\left\|F\left(z^{t_\star}\right)\right\|^2\right] \leq \frac{\left(1+\eta\gamma^2 L_F^2\right)\left\|z^0 - z^\star\right\|^2 + C\sigma_F^2\gamma^2\sum_{t=0}^{T}\alpha_t^2}{\mu\sum_{t=0}^{T}\alpha_t}$$

where $C = 1 + 2\eta\left(\left(\gamma^2 L_{\hat{F}}^2 + 1\right) + 2\alpha_0\right), \eta = \frac{1}{2}\frac{1+\gamma^2 L_F^2}{1-\gamma^2 L_F^2}\gamma^2 L_F^2 + \frac{1}{\gamma L_{\hat{F}}\sqrt{\alpha_0}}, \mu = \gamma^2\left(1-\gamma^2 L_F^2\right)/2$ and $t_\star$ is chosen from $\{0,1,\ldots,T\}$ according to probability $\mathcal{P}[t_\star = t] = \frac{\alpha_t}{\sum_{t=0}^{T}\alpha_t}$.

Convergence guarantee in unconstrained case

Theorem (Random Iterate Convergence)

Suppose the three assumptions on $F, \hat{F}, A$ hold. Suppose in addition that $\gamma \in (\lfloor -2\rho \rfloor_+, 1/L_F)$ and $(\alpha_t)_{t\in[T]} \subset (0,1)$ is a diminishing sequence such that

$$2\gamma L_{\hat{F}}\sqrt{\alpha_0} + \left(1 + \left(\frac{1+\gamma^2 L_F^2}{1-\gamma^2 L_F^2}\gamma^2 L_F^2\right)\gamma^2 L_{\hat{F}}^2\right)\alpha_0 \leq 1 + \frac{2\rho}{\gamma}$$

Then, the following estimate holds for all $z^\star \in \mathcal{S}^\star$

$$\mathbb{E}\left[\left\|F\left(z^{t_\star}\right)\right\|^2\right] \leq \frac{\left(1+\eta\gamma^2 L_F^2\right)\left\|z^0 - z^\star\right\|^2 + C\sigma_F^2\gamma^2\sum_{t=0}^T \alpha_t^2}{\mu\sum_{t=0}^T \alpha_t}$$

where $C = 1 + 2\eta\left(\left(\gamma^2 L_{\hat{F}}^2 + 1\right) + 2\alpha_0\right)$, $\eta = \frac{1}{2}\frac{1+\gamma^2 L_F^2}{1-\gamma^2 L_F^2}\gamma^2 L_F^2 + \frac{1}{\gamma L_{\hat{F}}\sqrt{\alpha_0}}$, $\mu = \gamma^2\left(1-\gamma^2 L_F^2\right)/2$ and $t_\star$ is chosen from $\{0,1,\ldots,T\}$ according to probability $\mathcal{P}[t_\star = t] = \frac{\alpha_t}{\sum_{t=0}^T \alpha_t}$.

- Set $\alpha_t = \Theta\left(1/\sqrt{t}\right)$. $O(\log T/\sqrt{T})$ convergence rate for $\mathbb{E}[\text{best } \|Fz^t\|^2]$ can be achieved.

Proof

◦ Introduce the potential function,

$$\mathcal{U}_t := \underbrace{\left\|z^t - z^\star\right\|^2}_{\text{distance to } \mathbf{zer}T} + A_t \left\|u^{t-1}\right\|^2 + B_t \left\|z^t - z^{t-1}\right\|^2, \tag{33}$$

where the first term is used in the deterministic case.

Proof

- Introduce the potential function,

$$\mathcal{U}_t := \underbrace{\left\|z^t - z^\star\right\|^2}_{\text{distance to } \mathbf{zer} T} + A_t \left\|u^{t-1}\right\|^2 + B_t \left\|z^t - z^{t-1}\right\|^2, \tag{33}$$

  where the first term is used in the deterministic case.

- **Goal**: show that $\mathbb{E}[\mathcal{U}_t] \leq \mathbb{E}[\mathcal{U}_{t-1}] - \Omega(\mathbb{E}[\alpha_t \|F(z^t)\|^2]) -$ positive coefficient $\cdot\, \mathbb{E}[\|F(\bar{z}^t)\|^2] +$ small error

Proof

- Introduce the potential function,

$$\mathcal{U}_t := \underbrace{\left\|z^t - z^\star\right\|^2}_{\text{distance to } \mathbf{zer}T} + A_t \left\|u^{t-1}\right\|^2 + B_t \left\|z^t - z^{t-1}\right\|^2, \tag{33}$$

  where the first term is used in the deterministic case.

- **Goal**: show that $\mathbb{E}[\mathcal{U}_t] \le \mathbb{E}[\mathcal{U}_{t-1}] - \Omega(\mathbb{E}[\alpha_t \|F(z^t)\|^2]) -$ positive coefficient $\cdot\, \mathbb{E}[\|F(\bar{z}^t)\|^2] +$ small error

- **Idea**: Reduce everything into some terms in the potential function, $\|Fz^t\|$ (what we want to bound), or $F\bar{z}^t$(which is $\mathbb{E}_{\bar{\xi}_t}[\hat{F}(\bar{z}^t, \bar{\xi}_t)] = \mathbb{E}_{\bar{\xi}_t}[\frac{z^{t+1}-z^t}{\alpha_t \gamma}]$).

Proof

- Introduce the potential function,

$$\mathcal{U}_t := \underbrace{\left\|z^t - z^\star\right\|^2}_{\text{distance to } \mathbf{zer}T} + A_t \left\|u^{t-1}\right\|^2 + B_t \left\|z^t - z^{t-1}\right\|^2, \tag{33}$$

  where the first term is used in the deterministic case.

- **Goal**: show that $\mathbb{E}[\mathcal{U}_t] \leq \mathbb{E}[\mathcal{U}_{t-1}] - \Omega(\mathbb{E}[\alpha_t \|F(z^t)\|^2]) -$ positive coefficient $\cdot\, \mathbb{E}[\|F(\bar{z}^t)\|^2] +$ small error

- **Idea**: Reduce everything into some terms in the potential function, $\|Fz^t\|$ (what we want to bound), or $F\bar{z}^t$(which is $\mathbb{E}_{\bar{\xi}_t}[\hat{F}(\bar{z}^t, \bar{\xi}_t)] = \mathbb{E}_{\bar{\xi}_t}[\frac{z^{t+1} - z^t}{\alpha_t \gamma}]$).

Fenchel-Young Inequality

$\forall a, b \in \mathbb{R}^n$ *and* $e > 0$, *we have,*

$$2\langle a, b\rangle \leq e\|a\|^2 + \tfrac{1}{e}\|b\|^2 \tag{34}$$

$$\|a+b\|^2 \leq (1+e)\|a\|^2 + (1 + \tfrac{1}{e})\|b\|^2 \tag{35}$$

Proof

- Introduce the potential function,

$$\mathcal{U}_t := \underbrace{\left\|z^t - z^\star\right\|^2}_{\text{distance to } \mathbf{zer} T} + A_t \left\|u^{t-1}\right\|^2 + B_t \left\|z^t - z^{t-1}\right\|^2, \tag{33}$$

  where the first term is used in the deterministic case.

- **Goal**: show that $\mathbb{E}[\mathcal{U}_t] \leq \mathbb{E}[\mathcal{U}_{t-1}] - \Omega(\mathbb{E}[\alpha_t \|F(z^t)\|^2]) -$ positive coefficient $\cdot\, \mathbb{E}[\|F(\bar{z}^t)\|^2] +$ small error

- **Idea**: Reduce everything into some terms in the potential function, $\|Fz^t\|$ (what we want to bound), or $F\bar{z}^t$(which is $\mathbb{E}_{\bar{\xi}_t}[\hat{F}(\bar{z}^t, \bar{\xi}_t)] = \mathbb{E}_{\bar{\xi}_t}[\frac{z^{t+1} - z^t}{\alpha_t \gamma}]$).

Fenchel-Young Inequality

$\forall a, b \in \mathbb{R}^n$ *and* $e > 0$, *we have,*

$$2\langle a, b\rangle \leq e\|a\|^2 + \tfrac{1}{e}\|b\|^2 \tag{34}$$

$$\|a + b\|^2 \leq (1 + e)\|a\|^2 + (1 + \tfrac{1}{e})\|b\|^2 \tag{35}$$

Proof.

$(1 + e)\|a\|^2 + (1 + \frac{1}{e})\|b\|^2 - \|a + b\|^2 = e\|a\|^2 + \frac{1}{e}\|b\|^2 - 2\langle a, b\rangle = \|\sqrt{e}a - \frac{1}{\sqrt{e}}b\|^2 \geq 0.$ □

Proof

- By (**??**), we can derive

$$\left\|u^t\right\|^2 = \left\|\gamma F\left(z^t\right) - \gamma \hat{F}\left(z^t, \xi_t\right) + (1-\alpha_t)\left(\gamma \hat{F}\left(z^{t-1}, \xi_t\right) - \gamma F\left(z^{t-1}\right)\right)\right\|^2 + (1-\alpha_t)^2 \left\|u^{t-1}\right\|^2 + \tag{36}$$

$$2(1-\alpha_t)\left\langle \bar{z}^{t-1} - z^{t-1} + \gamma F\left(z^{t-1}\right), \underbrace{\gamma\left(F\left(z^t\right) - \hat{F}\left(z^t, \xi_t\right)\right) + (1-\alpha_t)\gamma\left(\hat{F}\left(z^{t-1}, \xi_t\right) - F\left(z^{t-1}\right)\right)}_{\text{Conditioned on } \mathcal{F}_t \text{ (generated by random variables up to } z^t\text{), has expectation of zero.}}\right\rangle \tag{37}$$

Proof

- Conditioned on $\mathcal{F}_t$ (generated by random variables up to z^t)

$$
\begin{aligned}
& \mathbb{E}\left[\left\|u^t\right\|^2 \mid \mathcal{F}_t\right] && (38)\\
= & \gamma^2 \mathbb{E}\left[\left\|(1-\alpha_t)\left(F\left(z^t\right)-\hat{F}\left(z^t,\xi_t\right)+\hat{F}\left(z^{t-1},\xi_t\right)-F\left(z^{t-1}\right)\right)+\alpha_t\left(F\left(z^t\right)-\hat{F}\left(z^t,\xi_t\right)\right)\right\|^2 \mid \mathcal{F}_t\right] && (39)\\
& +(1-\alpha_t)^2\left\|u^{t-1}\right\|^2 && (40)\\
\overset{\text{Fenchel-Young}}{\leq} & (1-\alpha_t)^2\left\|u^{t-1}\right\|^2+2(1-\alpha_t)^2\gamma^2\mathbb{E}\left[\left\|\hat{F}\left(z^t,\xi_t\right)-\hat{F}\left(z^{t-1},\xi_t\right)\right\|^2 \mid \mathcal{F}_t\right] && (41)\\
& +2\alpha_t^2\gamma^2\mathbb{E}\left[\left\|F\left(z^t\right)-\hat{F}\left(z^t,\xi_t\right)\right\|^2 \mid \mathcal{F}_t\right] && (42)\\
\underset{\text{Bounded variance}}{\overset{\text{Lipschitz}}{\leq}} & (1-\alpha_t)^2\left\|u^{t-1}\right\|^2+2(1-\alpha_t)^2\gamma^2L_{\hat{F}}^2\left\|z^t-z^{t-1}\right\|^2+2\alpha_t^2\gamma^2\sigma_F^2 && (43)
\end{aligned}
$$

Proof (Cont.)

- By the step $z^{t+1} = z^t - \alpha_t \gamma \hat{F}(\bar{z}^t, \bar{\xi}_t)$, we have

$$\left\| z^{t+1} - z^\star \right\|^2 = \left\| z^t - z^\star \right\|^2 - 2\alpha_t \gamma \left\langle \hat{F}\left(\bar{z}^t, \bar{\xi}_t\right), z^t - z^\star \right\rangle + \alpha_t^2 \gamma^2 \left\| \hat{F}\left(\bar{z}^t, \bar{\xi}_t\right) \right\|^2 \tag{44}$$

Proof (Cont.)

- By the step $z^{t+1} = z^t - \alpha_t \gamma \hat{F}(\bar{z}^t, \bar{\xi}_t)$, we have

$$\left\|z^{t+1} - z^\star\right\|^2 = \left\|z^t - z^\star\right\|^2 - 2\alpha_t\gamma\left\langle \hat{F}\left(\bar{z}^t, \bar{\xi}_t\right), z^t - z^\star\right\rangle + \alpha_t^2\gamma^2\left\|\hat{F}\left(\bar{z}^t, \bar{\xi}_t\right)\right\|^2 \tag{44}$$

- Conditioned on $\overline{\mathcal{F}}_t$ generated by random variables up to $\bar{z}^t$,

$$\mathbb{E}\left[\left\langle -\gamma\hat{F}\left(\bar{z}^t, \bar{\xi}_t\right), z^t - z^\star\right\rangle \mid \overline{\mathcal{F}}_t\right] \tag{45}$$

$$\overset{\text{unbiased}}{=} -\gamma^2\left\langle F\left(\bar{z}^t\right), F\left(z^t\right)\right\rangle + \gamma\left\langle F\left(\bar{z}^t\right), \bar{z}^t - z^t + \gamma F\left(z^t\right)\right\rangle - \gamma\left\langle F\left(\bar{z}^t\right), \bar{z}^t - z^\star\right\rangle \tag{46}$$

Proof (Cont.)

- By the step $z^{t+1} = z^t - \alpha_t \gamma \hat{F}(\bar{z}^t, \bar{\xi}_t)$, we have

$$\left\| z^{t+1} - z^\star \right\|^2 = \left\| z^t - z^\star \right\|^2 - 2\alpha_t \gamma \left\langle \hat{F}\left(\bar{z}^t, \bar{\xi}_t\right), z^t - z^\star \right\rangle + \alpha_t^2 \gamma^2 \left\| \hat{F}\left(\bar{z}^t, \bar{\xi}_t\right) \right\|^2 \tag{44}$$

- Conditioned on $\overline{\mathcal{F}}_t$ generated by random variables up to $\bar{z}^t$,

$$\mathbb{E}\left[\left\langle -\gamma \hat{F}\left(\bar{z}^t, \bar{\xi}_t\right), z^t - z^\star \right\rangle \mid \overline{\mathcal{F}}_t\right] \tag{45}$$

$$\overset{\text{unbiased}}{=} -\gamma^2 \left\langle F\left(\bar{z}^t\right), F\left(z^t\right) \right\rangle + \gamma \left\langle F\left(\bar{z}^t\right), \bar{z}^t - z^t + \gamma F\left(z^t\right) \right\rangle - \gamma \left\langle F\left(\bar{z}^t\right), \bar{z}^t - z^\star \right\rangle \tag{46}$$

$$\underset{\text{Weak MVI}}{\overset{\text{Fenchel-Young}}{\leq}} \gamma^2 \left(\frac{1}{2} \left\| F\left(\bar{z}^t\right) - F\left(z^t\right) \right\|^2 - \frac{1}{2} \left\| F\left(\bar{z}^t\right) \right\|^2 - \frac{1}{2} \left\| F\left(z^t\right) \right\|^2 \right) \tag{47}$$

$$+ \frac{\gamma^2 \varepsilon_t}{2} \left\| F\left(\bar{z}^t\right) \right\|^2 + \frac{1}{2\varepsilon_t} \left\| \bar{z}^t - z^t + \gamma F\left(z^t\right) \right\|^2 - \gamma \rho \left\| F\left(\bar{z}^t\right) \right\|^2 \tag{48}$$

Proof (Cont.)

- By the step $z^{t+1} = z^t - \alpha_t \gamma \hat{F}(\bar{z}^t, \bar{\xi}_t)$, we have

$$\left\|z^{t+1} - z^\star\right\|^2 = \left\|z^t - z^\star\right\|^2 - 2\alpha_t\gamma \left\langle \hat{F}\left(\bar{z}^t, \bar{\xi}_t\right), z^t - z^\star \right\rangle + \alpha_t^2\gamma^2 \left\|\hat{F}\left(\bar{z}^t, \bar{\xi}_t\right)\right\|^2 \tag{44}$$

- Conditioned on $\overline{\mathcal{F}}_t$ generated by random variables up to $\bar{z}^t$,

$$\mathbb{E}\left[\left\langle -\gamma \hat{F}\left(\bar{z}^t, \bar{\xi}_t\right), z^t - z^\star \right\rangle \mid \overline{\mathcal{F}}_t\right] \tag{45}$$

$$\overset{\text{unbiased}}{=} -\gamma^2 \left\langle F\left(\bar{z}^t\right), F\left(z^t\right)\right\rangle + \gamma \left\langle F\left(\bar{z}^t\right), \bar{z}^t - z^t + \gamma F\left(z^t\right)\right\rangle - \gamma \left\langle F\left(\bar{z}^t\right), \bar{z}^t - z^\star\right\rangle \tag{46}$$

$$\underset{\text{Weak MVI}}{\overset{\text{Fenchel-Young}}{\leq}} \gamma^2 \left(\frac{1}{2}\left\|F\left(\bar{z}^t\right) - F\left(z^t\right)\right\|^2 - \frac{1}{2}\left\|F\left(\bar{z}^t\right)\right\|^2 - \frac{1}{2}\left\|F\left(z^t\right)\right\|^2\right) \tag{47}$$

$$+ \frac{\gamma^2 \varepsilon_t}{2}\left\|F\left(\bar{z}^t\right)\right\|^2 + \frac{1}{2\varepsilon_t}\left\|\bar{z}^t - z^t + \gamma F\left(z^t\right)\right\|^2 - \gamma\rho \left\|F\left(\bar{z}^t\right)\right\|^2 \tag{48}$$

$$\underset{\text{Fenchel-Young}}{\overset{\text{Lipschitz}}{\leq}} \gamma^2 L_F^2 \frac{1+b}{2}\left\|u^t\right\|^2 + \frac{1+b^{-1}}{2}\gamma^4 L_F^2 \left\|F\left(z^t\right)\right\|^2 - \frac{\gamma^2}{2}\left\|F\left(\bar{z}^t\right)\right\|^2 \tag{49}$$

$$- \frac{\gamma^2}{2}\left\|F\left(z^t\right)\right\|^2 + \frac{\gamma^2\varepsilon_t}{2}\left\|F\left(\bar{z}^t\right)\right\|^2 + \frac{1}{2\varepsilon_t}\left\|u^t\right\|^2 - \gamma\rho\left\|F\left(\bar{z}^t\right)\right\|^2 \tag{50}$$

Proof (Cont.)

- By the step $z^{t+1} = z^t - \alpha_t \gamma \hat{F}(\bar{z}^t, \bar{\xi}_t)$, we have

$$\left\|z^{t+1} - z^\star\right\|^2 = \left\|z^t - z^\star\right\|^2 - 2\alpha_t\gamma \left\langle \hat{F}\left(\bar{z}^t, \bar{\xi}_t\right), z^t - z^\star \right\rangle + \alpha_t^2\gamma^2 \left\|\hat{F}\left(\bar{z}^t, \bar{\xi}_t\right)\right\|^2 \tag{44}$$

- Conditioned on $\overline{\mathcal{F}}_t$ generated by random variables up to $\bar{z}^t$,

$$\begin{aligned}
&\mathbb{E}\left[\left\langle -\gamma \hat{F}\left(\bar{z}^t, \bar{\xi}_t\right), z^t - z^\star \right\rangle \mid \overline{\mathcal{F}}_t\right] && (45)\\
\overset{\text{unbiased}}{=} & -\gamma^2 \left\langle F\left(\bar{z}^t\right), F\left(z^t\right)\right\rangle + \gamma\left\langle F\left(\bar{z}^t\right), \bar{z}^t - z^t + \gamma F\left(z^t\right)\right\rangle - \gamma\left\langle F\left(\bar{z}^t\right), \bar{z}^t - z^\star\right\rangle && (46)\\
\underset{\text{Weak MVI}}{\overset{\text{Fenchel-Young}}{\leq}} & \gamma^2\left(\frac{1}{2}\left\|F\left(\bar{z}^t\right) - F\left(z^t\right)\right\|^2 - \frac{1}{2}\left\|F\left(\bar{z}^t\right)\right\|^2 - \frac{1}{2}\left\|F\left(z^t\right)\right\|^2\right) && (47)\\
& + \frac{\gamma^2\varepsilon_t}{2}\left\|F\left(\bar{z}^t\right)\right\|^2 + \frac{1}{2\varepsilon_t}\left\|\bar{z}^t - z^t + \gamma F\left(z^t\right)\right\|^2 - \gamma\rho\left\|F\left(\bar{z}^t\right)\right\|^2 && (48)\\
\underset{\text{Fenchel-Young}}{\overset{\text{Lipschitz}}{\leq}} & \gamma^2 L_F^2 \frac{1+b}{2}\left\|u^t\right\|^2 + \frac{1+b^{-1}}{2}\gamma^4 L_F^2\left\|F\left(z^t\right)\right\|^2 - \frac{\gamma^2}{2}\left\|F\left(\bar{z}^t\right)\right\|^2 && (49)\\
& - \frac{\gamma^2}{2}\left\|F\left(z^t\right)\right\|^2 + \frac{\gamma^2\varepsilon_t}{2}\left\|F\left(\bar{z}^t\right)\right\|^2 + \frac{1}{2\varepsilon_t}\left\|u^t\right\|^2 - \gamma\rho\left\|F\left(\bar{z}^t\right)\right\|^2 && (50)\\
= & \left(\gamma^2 L_F^2\frac{1+b}{2} + \frac{1}{2\varepsilon_t}\right)\left\|u^t\right\|^2 + \gamma^2\frac{\gamma^2 L_F^2\left(1+b^{-1}\right) - 1}{2}\left\|F\left(z^t\right)\right\|^2 + \left(\frac{\gamma^2\left(\varepsilon_t - 1\right)}{2} - \gamma\rho\right)\left\|F\left(\bar{z}^t\right)\right\|^2. && (51)
\end{aligned}$$

Proof (Cont.)

- $$\mathbb{E}\left[\left\|\hat{F}\left(\bar{z}^t,\bar{\xi}_t\right)\right\|^2 \mid \mathcal{F}_t\right] = \mathbb{E}\left[\mathbb{E}\left[\left\|\hat{F}\left(\bar{z}^t,\bar{\xi}_t\right) - F(\bar{z}^t) + F(\bar{z}^t)\right\|^2 \mid \overline{\mathcal{F}}_t\right] \mid \mathcal{F}_t\right] \underset{\text{Bounded variance}}{\overset{\text{Unbiased}}{\leq}} \left\|F\left(\bar{z}^t\right)\right\|^2 + \sigma_F^2. \tag{52}$$

Proof (Cont.)

- $$\mathbb{E}\left[\left\|\hat{F}\left(\bar{z}^t, \bar{\xi}_t\right)\right\|^2 \mid \mathcal{F}_t\right] = \mathbb{E}\left[\mathbb{E}\left[\left\|\hat{F}\left(\bar{z}^t, \bar{\xi}_t\right) - F(\bar{z}^t) + F(\bar{z}^t)\right\|^2 \mid \overline{\mathcal{F}}_t\right] \mid \mathcal{F}_t\right] \underset{\text{Bounded variance}}{\overset{\text{Unbiased}}{\leq}} \left\|F\left(\bar{z}^t\right)\right\|^2 + \sigma_F^2. \tag{52}$$

- The bound on the term $\|\hat{F}(\bar{z}^t, \bar{\xi}_t)\|^2$ can further imply,

$$\mathbb{E}\left[\|z^{t+1} - z^t\|^2 \mid \mathcal{F}_t\right] = \alpha_t^2\gamma^2\mathbb{E}\left[\|\hat{F}(\bar{z}^t, \bar{\xi}_t)\|^2 \mid \mathcal{F}_t\right] \leq \alpha_t^2\gamma^2\mathbb{E}\left[\|F\bar{z}^t\|^2 \mid \mathcal{F}_t\right] + \alpha_t^2\gamma^2\sigma_F^2 \tag{53}$$

Proof (Cont.)

- $$\mathbb{E}\left[\left\|\hat{F}\left(\bar{z}^t, \bar{\xi}_t\right)\right\|^2 \mid \mathcal{F}_t\right] = \mathbb{E}\left[\mathbb{E}\left[\left\|\hat{F}\left(\bar{z}^t, \bar{\xi}_t\right) - F(\bar{z}^t) + F(\bar{z}^t)\right\|^2 \mid \overline{\mathcal{F}}_t\right] \mid \mathcal{F}_t\right] \underset{\text{Bounded variance}}{\overset{\text{Unbiased}}{\leq}} \left\|F\left(\bar{z}^t\right)\right\|^2 + \sigma_F^2. \tag{52}$$

- The bound on the term $\|\hat{F}(\bar{z}^t, \bar{\xi}_t)\|^2$ can further imply,

$$\mathbb{E}\left[\|z^{t+1} - z^t\|^2 \mid \mathcal{F}_t\right] = \alpha_t^2\gamma^2\mathbb{E}\left[\|\hat{F}(\bar{z}^t, \bar{\xi}_t)\|^2 \mid \mathcal{F}_t\right] \leq \alpha_t^2\gamma^2\mathbb{E}\left[\|F\bar{z}^t\|^2 \mid \mathcal{F}_t\right] + \alpha_t^2\gamma^2\sigma_F^2 \tag{53}$$

- Combining the above bounds yields,

$$\begin{aligned} &\mathbb{E}[\|z^{t+1} - z^\star\|^2 + A_{t+1}\|u^t\|^2 + B_{t+1}\|z^{t+1} - z^t\|^2 \mid \mathcal{F}_t] \\ &\quad \leq \|z^t - z^\star\|^2 + \Big(\underbrace{A_{t+1} + \alpha_t\left(\gamma^2 L_F^2(1+b) + \frac{1}{\varepsilon_t}\right)}_{X_1^t}\Big)\mathbb{E}[\|u^t\|^2 \mid \mathcal{F}_t] - \alpha_t\mu\|F(z^t)\|^2 \\ &\qquad + \Big(\underbrace{\alpha_t\left(\gamma^2(\varepsilon_t - 1) - 2\gamma\rho\right) + \alpha_t^2\gamma^2}_{X_2^t} + B_{t+1}\alpha_t^2\gamma^2\Big)\mathbb{E}[\|F(\bar{z}^t)\|^2 \mid \mathcal{F}_t] + (1 + B_{t+1})\alpha_t^2\gamma^2\sigma_F^2. \end{aligned} \tag{54}$$

Proof (Cont.)

- Further using the bound on $\|u^t\|^2$.

$$\mathbb{E}\left[\left\|u^t\right\|^2 \mid \mathcal{F}_t\right] \leq (1-\alpha_t)^2 \left\|u^{t-1}\right\|^2 + 2(1-\alpha_t)^2 \gamma^2 L_{\hat{F}}^2 \left\|z^t - z^{t-1}\right\|^2 + 2\alpha_t^2\gamma^2\sigma_F^2 \tag{55}$$

Proof (Cont.)

- Further using the bound on $\|u^t\|^2$.

$$\mathbb{E}\left[\left\|u^t\right\|^2 \mid \mathcal{F}_t\right] \leq (1-\alpha_t)^2 \left\|u^{t-1}\right\|^2 + 2\,(1-\alpha_t)^2\,\gamma^2 L_{\hat{F}}^2 \left\|z^t - z^{t-1}\right\|^2 + 2\alpha_t^2\gamma^2\sigma_F^2 \tag{55}$$

- We have,

$$\begin{aligned}\mathbb{E}[\mathcal{U}_{t+1} \mid \mathcal{F}_t] - \mathcal{U}_t \leq\ & -\alpha_t\mu\|F(z^t)\|^2 + \left(X_1^t(1-\alpha_t)^2 - A_t\right)\|u^{t-1}\|^2 \\ & + 2\gamma^2 L_{\hat{F}}^2 \left(X_1^t(1-\alpha_t)^2 - \frac{B_t}{2\gamma^2 L_{\hat{F}}^2}\right)\|z^t - z^{t-1}\|^2 + \left(X_2^t + B_{t+1}\alpha_t^2\gamma^2\right)\mathbb{E}[\|F(\bar{z}^t)\|^2 \mid \mathcal{F}_t] \\ & + \left(B_{t+1}\alpha_t^2 + \alpha_t^2 + 2X_1^t\alpha_t^2\right)\gamma^2\sigma_F^2.\end{aligned} \tag{56}$$

Proof (Cont.)

- Further using the bound on $\|u^t\|^2$.

$$\mathbb{E}\left[\left\|u^t\right\|^2 \mid \mathcal{F}_t\right] \leq (1-\alpha_t)^2 \left\|u^{t-1}\right\|^2 + 2(1-\alpha_t)^2 \gamma^2 L_{\hat{F}}^2 \left\|z^t - z^{t-1}\right\|^2 + 2\alpha_t^2\gamma^2\sigma_F^2 \tag{55}$$

- We have,

$$\begin{aligned}\mathbb{E}[\mathcal{U}_{t+1} \mid \mathcal{F}_t] - \mathcal{U}_t \leq & -\alpha_t\mu\|F(z^t)\|^2 + \left(X_1^t(1-\alpha_t)^2 - A_t\right)\|u^{t-1}\|^2 \\ & + 2\gamma^2 L_{\hat{F}}^2 \left(X_1^t(1-\alpha_t)^2 - \frac{B_t}{2\gamma^2 L_{\hat{F}}^2}\right)\|z^t - z^{t-1}\|^2 + \left(X_2^t + B_{t+1}\alpha_t^2\gamma^2\right)\mathbb{E}[\|F(\bar{z}^t)\|^2 \mid \mathcal{F}_t] \\ & + \left(B_{t+1}\alpha_t^2 + \alpha_t^2 + 2X_1^t\alpha_t^2\right)\gamma^2\sigma_F^2. \end{aligned} \tag{56}$$

Having established (**??**), set $A_t = A$, $B_t = 2A\gamma^2 L_{\hat{F}}^2$, and $\varepsilon_t = \varepsilon$ to obtain by the law of total expectation that

$$\begin{aligned}\mathbb{E}[\mathcal{U}_{t+1}] - \mathbb{E}[\mathcal{U}_t] \leq & -\alpha_t\mu\mathbb{E}[\|F(z^t)\|^2] + \left(X_1^t(1-\alpha_t)^2 - A\right)\mathbb{E}[\|u^{t-1}\|^2] \\ & + 2\gamma^2 L_{\hat{F}}^2 \left(X_1^t(1-\alpha_t)^2 - A\right)\mathbb{E}[\|z^t - z^{t-1}\|^2] + \left(X_2^t + 2A\gamma^4 L_{\hat{F}}^2\alpha_t^2\right)\mathbb{E}[\|F(\bar{z}^t)\|^2] \\ & + \left(2A\gamma^2 L_{\hat{F}}^2 + 1 + 2X_1^t\right)\alpha_t^2\gamma^2\sigma_F^2. \end{aligned} \tag{57}$$

Proof (Cont.)

- By carefully choosing the constant A and ε, we can enforce the following negative coefficients,

$$X_1^t(1-\alpha_t)^2 - A \leq 0 \quad \text{and} \quad X_2^t + 2A\gamma^4 L_{\hat{F}}^2 \alpha_t^2 \leq 0.$$

  Recall that,

$$X_1^t = \alpha_t\left(\gamma^2 L_F^2(1+b) + \tfrac{1}{\varepsilon}\right) + A, \quad X_2^t = \alpha_t\left(\gamma^2(\varepsilon-1) - 2\rho\gamma + \alpha_t\ \gamma^2\right).$$

  Pick $A = \frac{1}{2}\left((b+1)\gamma^2 L_F^2 + \frac{1}{\varepsilon}\right)$ and $\varepsilon = \gamma L_{\hat{F}}\sqrt{\alpha_0}$, the negative requirements are satisfied.

Proof (Cont.)

- By carefully choosing the constant A and ε, we can enforce the following negative coefficients,

$$X_1^t(1-\alpha_t)^2 - A \leq 0 \quad \text{and} \quad X_2^t + 2A\gamma^4 L_{\hat{F}}^2 \alpha_t^2 \leq 0.$$

  Recall that,

$$X_1^t = \alpha_t \left(\gamma^2 L_F^2 (1+b) + \tfrac{1}{\varepsilon}\right) + A, \quad X_2^t = \alpha_t \left(\gamma^2(\varepsilon - 1) - 2\rho\gamma + \alpha_t \, \gamma^2\right).$$

  Pick $A = \frac{1}{2}\left((b+1)\gamma^2 L_F^2 + \frac{1}{\varepsilon}\right)$ and $\varepsilon = \gamma L_{\hat{F}} \sqrt{\alpha_0}$, the negative requirements are satisfied.
- Thus we can derive the recursion,

$$\mathbb{E}[\mathcal{U}_{t+1}] - \mathbb{E}[\mathcal{U}_t] \leq \, -\alpha_t \mu \mathbb{E}[\|F(z^t)\|^2] + C\alpha_t^2 \gamma^2 \sigma_F^2.$$

Proof (Cont.)

- By carefully choosing the constant A and ε, we can enforce the following negative coefficients,

$$X_1^t(1-\alpha_t)^2 - A \leq 0 \quad \text{and} \quad X_2^t + 2A\gamma^4 L_{\hat{F}}^2 \alpha_t^2 \leq 0.$$

  Recall that,

$$X_1^t = \alpha_t \left(\gamma^2 L_F^2 (1+b) + \tfrac{1}{\varepsilon}\right) + A, \quad X_2^t = \alpha_t \left(\gamma^2(\varepsilon - 1) - 2\rho\gamma + \alpha_t \, \gamma^2\right).$$

  Pick $A = \frac{1}{2}\left((b+1)\gamma^2 L_F^2 + \frac{1}{\varepsilon}\right)$ and $\varepsilon = \gamma L_{\hat{F}} \sqrt{\alpha_0}$, the negative requirements are satisfied.

- Thus we can derive the recursion,

$$\mathbb{E}[\mathcal{U}_{t+1}] - \mathbb{E}[\mathcal{U}_t] \leq \; -\alpha_t \mu \mathbb{E}[\|F(z^t)\|^2] + C\alpha_t^2 \gamma^2 \sigma_F^2.$$

- Telescoping the above inequality completes the proof.

Almost Sure Convergence

- Random iterate convergence is weak.
  - Hard to check and report the result, since evaluating the deterministic F is expensive.
  - The result can be volatile.

Almost Sure Convergence

- Random iterate convergence is weak.
  - Hard to check and report the result, since evaluating the deterministic F is expensive.
  - The result can be volatile.
- Almost sure convergence is more desirable.
  - No variance to be taken care of.
  - Can just report the final solution after running long enough.

Almost Sure Convergence

- Random iterate convergence is weak.
  - Hard to check and report the result, since evaluating the deterministic F is expensive.
  - The result can be volatile.
- Almost sure convergence is more desirable.
  - No variance to be taken care of.
  - Can just report the final solution after running long enough.

Theorem (Almost Sure Convergence)

Suppose that the three assumptions on $F, \hat{F}, A$ hold. Suppose $\gamma \in (\lfloor -2\rho \rfloor_+, 1/L_F)$, $\alpha_t = \frac{1}{t+r}$ for any positive natural number r and

$$(\gamma L_{\hat{F}} + 1)\alpha_t + 2\left(\frac{1+\gamma^2 L_F^2}{1-\gamma^2 L_F^2}\gamma^4 L_F^2 L_{\hat{F}}^2 \alpha_{t+1} + \gamma L_{\hat{F}}\right)(\alpha_{t+1} + 1)\,\alpha_{t+1} \leq 1 + \frac{2\rho}{\gamma}. \tag{58}$$

*Then, the sequence $(z^t)_{k\in\mathbb{N}}$ generated by the Alg. **??** converges almost surely to some $z^\star \in \operatorname{zer} T$.*

Almost Sure Convergence

- Random iterate convergence is weak.
  - Hard to check and report the result, since evaluating the deterministic F is expensive.
  - The result can be volatile.
- Almost sure convergence is more desirable.
  - No variance to be taken care of.
  - Can just report the final solution after running long enough.

Theorem (Almost Sure Convergence)

Suppose that the three assumptions on $F, \hat{F}, A$ hold. Suppose $\gamma \in (\lfloor -2\rho \rfloor_+, {}^1/L_F)$, $\alpha_t = \frac{1}{t+r}$ for any positive natural number r and

$$(\gamma L_{\hat{F}} + 1)\alpha_t + 2\left(\frac{1+\gamma^2 L_F^2}{1-\gamma^2 L_F^2}\gamma^4 L_F^2 L_{\hat{F}}^2 \alpha_{t+1} + \gamma L_{\hat{F}}\right)(\alpha_{t+1}+1)\,\alpha_{t+1} \leq 1 + \frac{2\rho}{\gamma}. \tag{58}$$

Then, the sequence $(z^t)_{k\in\mathbb{N}}$ generated by the Alg. **??** *converges almost surely to some $z^\star \in \operatorname{zer} T$.*

Remark:

- Step size $\alpha_t = \frac{1}{t+r} = \Theta(\frac{1}{t})$, diminishes faster than $\Theta(\frac{1}{\sqrt{t}})$.
- Loses $\tilde{\mathcal{O}}(\frac{1}{\sqrt{T}})$ random iterate convergence rate, but obtain almost sure convergence.

Constrained case

- Using the prox/resolvent to give the BC-PSEG+ algorithm,

Algorithm (BC-PSEG+) Stochastic algorithm for constrained problem.

REQUIRE $z^{-1} = z^0 \in \mathbb{R}^n$, $h^{-1} \in \mathbb{R}^n$, $\alpha_t \in (0,1)$, $\gamma \in (\lfloor -2\rho \rfloor_+, 1/L_F)$
for $t = 0, 1, \ldots$ until convergence **do**
  Sample $\xi_t \sim \mathcal{P}$
  $h^t = \left(z^t - \gamma \hat{F}(z^t, \xi_t)\right) + (1-\alpha_t)\left(h^{t-1} - \left(z^{t-1} - \gamma \hat{F}(z^{t-1}, \xi_t)\right)\right)$
  $\bar{z}^t = (\mathrm{id} + \gamma A)^{-1} h_t$
  Sample $\bar{\xi}_t \sim \mathcal{P}$
  $z^{t+1} = z^t - \alpha_t\left(h^t - \bar{z}^t + \gamma \hat{F}(\bar{z}^t, \bar{\xi}_t)\right)$
end for
Return z^{t+1}

Constrained case

- Using the prox/resolvent to give the BC-PSEG+ algorithm,

Algorithm (BC-PSEG+) Stochastic algorithm for constrained problem.

REQUIRE $z^{-1} = z^0 \in \mathbb{R}^n$, $h^{-1} \in \mathbb{R}^n$, $\alpha_t \in (0,1)$, $\gamma \in (\lfloor -2\rho \rfloor_+, 1/L_F)$
for $t = 0, 1, \ldots$ until convergence **do**
  Sample $\xi_t \sim \mathcal{P}$
  $h^t = \left(z^t - \gamma \hat{F}(z^t, \xi_t)\right) + (1 - \alpha_t)\left(h^{t-1} - \left(z^{t-1} - \gamma \hat{F}(z^{t-1}, \xi_t)\right)\right)$
  $\bar{z}^t = (\mathrm{id} + \gamma A)^{-1} h_t$
  Sample $\bar{\xi}_t \sim \mathcal{P}$
  $z^{t+1} = z^t - \alpha_t \left(h^t - \bar{z}^t + \gamma \hat{F}(\bar{z}^t, \bar{\xi}_t)\right)$
end for
Return z^{t+1}

- **Key idea**: apply similar bias-corrected technique to the extrapolation point before the resolvent.
- **Properties:**
  - Converges to $\mathrm{zer} T$.

Nonlinear preconditioned primal dual extragradient (NP-PDEG)

$$\min_{x\in\mathbb{R}^n}\max_{y\in\mathbb{R}^r} \quad f(x) + \varphi(x,y) - g(y). \tag{59}$$

where $\varphi(x,y) := \mathbb{E}_\xi[\hat{\varphi}(x,y,\xi)]$.

Nonlinear preconditioned primal dual extragradient (NP-PDEG)

$$\min_{x\in\mathbb{R}^n}\max_{y\in\mathbb{R}^r} \quad f(x)+\varphi(x,y)-g(y). \tag{59}$$

where $\varphi(x,y) := \mathbb{E}_\xi[\hat{\varphi}(x,y,\xi)]$.

Algorithm Nonlinearly preconditioned primal dual extragradient (NP-PDEG)

1: REQUIRE $z^{-1}=z^0=(x^0,y^0)$ with $x^0,x^{-1},\hat{x}^{-1},\bar{x}^{-1}\in\mathbb{R}^n$, $y^0,y^{-1}\in\mathbb{R}^r$, $\theta\in[0,\infty)$, $\Gamma_1\succ 0$, $\Gamma_2\succ 0$
2: **for** $t=0,1,\ldots$ until convergence **do**
3: $\quad\hat{x}^t = x^t - \Gamma_1\nabla_x\hat{\varphi}(z^t,\xi_t) + (1-\alpha_t)\left(\hat{x}^{t-1}-x^{t-1}+\Gamma_1\nabla_x\hat{\varphi}(x^{t-1},y^{t-1},\xi_t)\right)$, $\xi_t\sim\mathcal{P}$
4: $\quad\bar{x}^t = \operatorname{prox}_f^{\Gamma_1^{-1}}\left(\hat{x}^t\right)$
5: $\quad\hat{y}^t = y^t + \Gamma_2\nabla_y\hat{\varphi}(\bar{x}^t,y^t,\xi_t') + (1-\alpha_t)\left(\hat{y}^{t-1}-y^{t-1}-\Gamma_2\nabla_y\hat{\varphi}(\bar{x}^{t-1},y^{t-1},\xi_t')\right)$, $\xi_t'\sim\mathcal{P}$
6: $\quad\bar{y}^t = \operatorname{prox}_g^{\Gamma_2^{-1}}\left(\hat{y}^t\right)$
7: $\quad\bar{\xi}_t\sim\mathcal{P}$
8: $\quad x^{t+1} = x^t + \alpha_t\left(\bar{x}^t-\hat{x}^t-\Gamma_1\nabla_x\hat{\varphi}(\bar{z}^t,\bar{\xi}_t)\right)$, $y^{t+1}=y^t+\alpha_t\left(\bar{y}^t-\hat{y}^t+\Gamma_2\nabla_y\hat{\varphi}(\bar{z}^t,\bar{\xi}_t)\right)$
9: **end for**
10: **Return** $z^{t+1}=(x^{t+1},y^{t+1})$

Nonlinear preconditioned primal dual extragradient (NP-PDEG)

$$\min_{x \in \mathbb{R}^n} \max_{y \in \mathbb{R}^r} \quad f(x) + \varphi(x, y) - g(y). \tag{59}$$

where $\varphi(x, y) := \mathbb{E}_\xi[\hat{\varphi}(x, y, \xi)]$.

Algorithm Nonlinearly preconditioned primal dual extragradient (NP-PDEG)

REQUIRE $z^{-1} = z^0 = (x^0, y^0)$ with $x^0, x^{-1}, \hat{x}^{-1}, \bar{x}^{-1} \in \mathbb{R}^n$, $y^0, y^{-1} \in \mathbb{R}^r$, $\theta \in [0, \infty)$, $\Gamma_1 \succ 0$, $\Gamma_2 \succ 0$
for $t = 0, 1, \ldots$ until convergence **do**
  $\hat{x}^t = x^t - \Gamma_1 \nabla_x \hat{\varphi}(z^t, \xi_t) + (1 - \alpha_t)\left(\hat{x}^{t-1} - x^{t-1} + \Gamma_1 \nabla_x \hat{\varphi}(x^{t-1}, y^{t-1}, \xi_t)\right)$, $\xi_t \sim \mathcal{P}$
  $\bar{x}^t = \operatorname{prox}_f^{\Gamma_1^{-1}}\left(\hat{x}^t\right)$
  $\hat{y}^t = y^t + \Gamma_2 \nabla_y \hat{\varphi}(\bar{x}^t, y^t, \xi_t') + (1 - \alpha_t)\left(\hat{y}^{t-1} - y^{t-1} - \Gamma_2 \nabla_y \hat{\varphi}(\bar{x}^{t-1}, y^{t-1}, \xi_t')\right)$, $\xi_t' \sim \mathcal{P}$
  $\bar{y}^t = \operatorname{prox}_g^{\Gamma_2^{-1}}\left(\hat{y}^t\right)$
  $\bar{\xi}_t \sim \mathcal{P}$
  $x^{t+1} = x^t + \alpha_t\left(\bar{x}^t - \hat{x}^t - \Gamma_1 \nabla_x \hat{\varphi}(\bar{z}^t, \bar{\xi}_t)\right)$, $y^{t+1} = y^t + \alpha_t\left(\bar{y}^t - \hat{y}^t + \Gamma_2 \nabla_y \hat{\varphi}(\bar{z}^t, \bar{\xi}_t)\right)$
end for
Return $z^{t+1} = (x^{t+1}, y^{t+1})$

- Alternatively sampling relaxes the Lipschitz assumption.

$$\|\nabla_x \varphi(z') - \nabla_x \varphi(z)\| \le L_x \|z' - z\|, \quad \|\nabla_y \varphi(z') - \nabla_y \varphi(x', y)\| \le L_y \|y' - y\|.$$

Experiment 1

Example (Unconstrained quadratic game [? , Ex. 5])

Consider,

$$\min_{x\in\mathbb{R}} \max_{y\in\mathbb{R}} \phi(x,y) := axy + \frac{b}{2}x^2 - \frac{b}{2}y^2, \tag{60}$$

where $a \in \mathbb{R}_+$ and $b \in \mathbb{R}$.

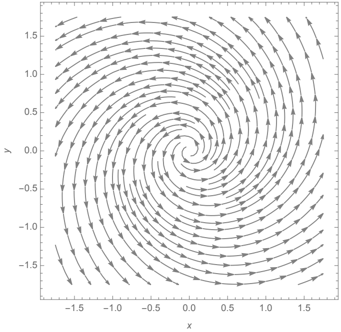


Figure: Unstable dynamics!

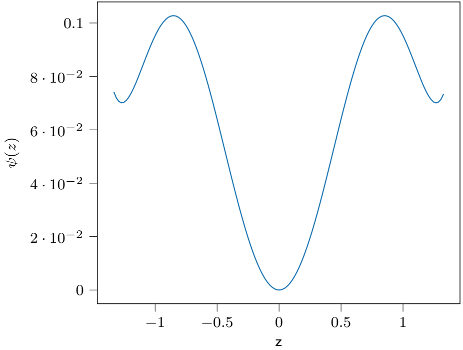
Experiment 2

Example (Constrained minimax [? , Ex. 4])

Consider

$$\min_{|x|\le 4/3} \max_{|y|\le 4/3} \phi(x,y) := xy + \psi(x) - \psi(y), \qquad \text{(GlobalForsaken)}$$

where $\psi(z) = \frac{2z^6}{21} - \frac{z^4}{3} + \frac{z^2}{3}$.



Experiment 2

Example (Constrained minimax [? , Ex. 4])

Consider

$$\min_{|x|\le 4/3}\ \max_{|y|\le 4/3}\ \phi(x,y) := xy + \psi(x) - \psi(y), \qquad \text{(GlobalForsaken)}$$

where $\psi(z) = \frac{2z^6}{21} - \frac{z^4}{3} + \frac{z^2}{3}$.

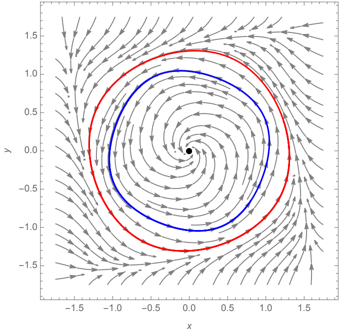
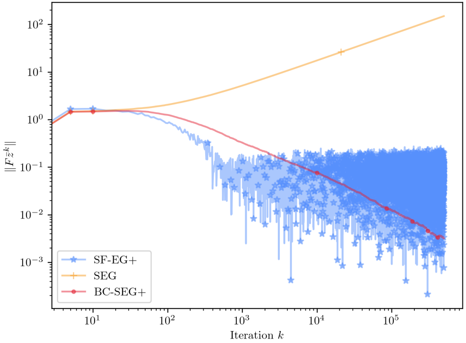


Figure: Limit cycles.

Experimental Results



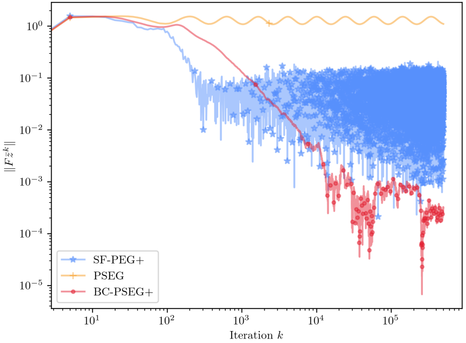


Figure: Comparison of methods in the unconstrained setting and the constrained setting.

Remark:

- SEG diverges (As expected, since the dynamics is unstable.) and PSEG cycles;
- Both oblivious extensions (SF-EG+) and (SF-PEG+) only converge to a neighborhood.
- Only BC-SEG+ and BC-PSEG+ converges properly, with probability 1 as established;

References I