Escaping limit cycles:
Recent advances in first-order methods for structured nonmonotone games

Thomas Pethick

What are we interested in?

Multiplayer games appears in a lot of places:

- *Adversarial training (minimax): e.g. generative adversarial networks (GANs)*
- *Anywhere with multiple (possibly opposing) objectives*

Goal of today

- Make minimax (hopefully) less intimidating
- Provide an intuitive geometry approach
- We get some nonconvex cases "for free"

Convince you that halfspace projections are very useful!

Content

- Start with **deterministic** case
- Show that the **stochastic** case is not so different
- See how we can leverage the ideas for **federated learning**

Escaping limit cycles: Global convergence for constrained nonconvex-nonconcave minimax problems

Solving stochastic weak Minty variational inequalities without increasing batch size

Efficient interpolation between extragradient and proximal methods for weak MVIs *(under review)*

iFedDR: Auto-tuning local computation with inexact Douglas-Rachford splitting in federated learning *(under review)*

Volkan Cevher
Oliver Fercoq
Puya Latafat
Panagiotis Patrinos
Peter Kairouz
Ioannis Mavrothalassitis
Zheng Xu
Basile Lewandowski

Why is (convex) minimization easy

Minimization

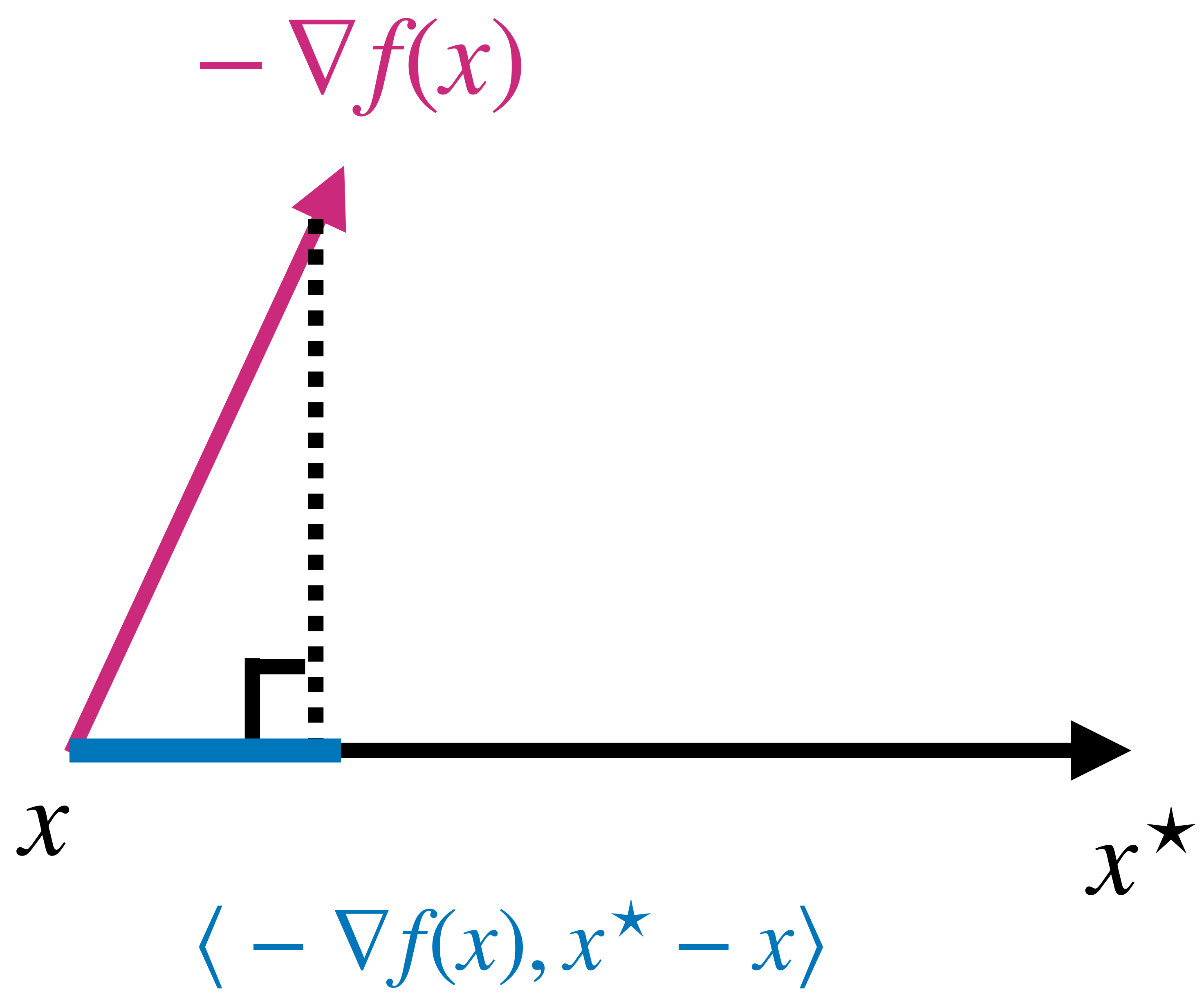
$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \;\; f(x)$$

(Star)-convexity

$$\langle \nabla f(x), x - x^\star \rangle \geqslant \boxed{f(x) - f(x^\star),} \quad \text{for all} \quad x \in \mathbb{R}^n$$

(Star)-convexity + L-Lipschitz gradients

$$\langle \nabla f(x), x - x^\star \rangle \geqslant \boxed{\frac{1}{L}\|\nabla f(x)\|^2,} \quad \text{for all} \quad x \in \mathbb{R}^n$$



The **gradient direction** is always guaranteed to make **progress** towards the solution

Suggests that the **gradient method** suffice:

$$x^{k+1} = x^k - \gamma_k \nabla f(x^k)$$

for some stepsize $\gamma_k > 0$.

The difficulty of minimax problems

Minimax

$$\underset{x \in \mathbb{R}^n}{\text{minimize}}\ \underset{y \in \mathbb{R}^m}{\text{maximize}}\ f(x, y)$$

Operator view:

$$z = (x, y) \in \mathbb{R}^d \qquad d = n + m$$

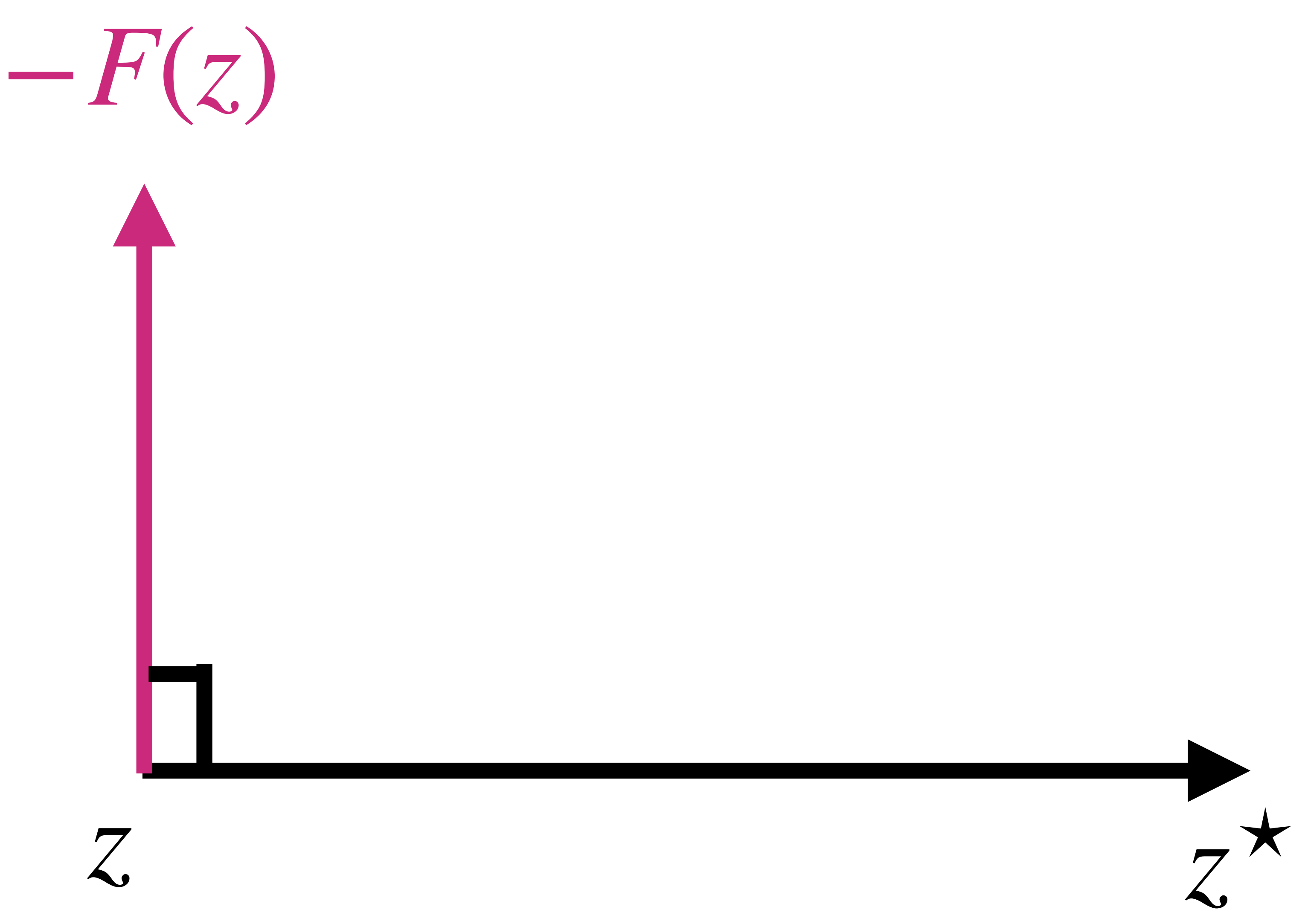
$$F(z) = (\nabla_x f(x, y), -\nabla_y f(x, y))$$

First order stationarity conditions: Find $z \in \mathbb{R}^d$ such that

$$F(z) = 0$$

Monotonicity equivalent to convex in x and concave in y

$$\langle F(z), z - z^\star \rangle \geqslant 0, \quad \text{for all} \quad z \in \mathbb{R}^d$$



$\langle -F(z), z^\star - z \rangle = 0$ possibly

The **gradient direction** is **not** guaranteed to make **progress** towards the solution

In fact: **Never** the case for bilinear problems:

Example (Bilinear): $f(x, y) = \langle x, Ay \rangle$

Linear interactive terms are very common, e.g.

- Lagrange (re)formulations
- Game theory

Nonmonotone operators

Find $z \in \mathbb{R}^d$ **such that**

$$0 = F(z)$$

first order condition of

$$\underset{x \in \mathbb{R}^n}{\text{minimize}}\ \underset{y \in \mathbb{R}^m}{\text{maximize}}\ f(x, y)$$

Operator view:

$$z = (x, y)$$

$$F(z) = (\nabla_x f(x, y), -\nabla_y f(x, y))$$

Assumption 1 $F : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ satisfies the **weak Minty variational inequality (MVI)**, i.e. for all $z^\star \in \mathcal{Z}^\star \subseteq \operatorname{zer} S$ (nonempty $\mathcal{Z}^\star$) and some $\rho \in (-\frac{1}{L}, \infty)$

$$\boxed{\langle F(z), z - z^\star \rangle \geqslant \rho \|F(z)\|^2} \quad \text{for all } z \in \mathbb{R}^d$$

Nonmonotone when negative!

$\langle -F(z), z^\star - z \rangle \leq 0$ possibly

Assumption 2 $F : \mathbb{R}^d \to \mathbb{R}^d$ is L-Lipschitz

$$\|F(z) - F(z')\| \leqslant L\|z - z'\| \quad \text{for all } z, z' \in \mathbb{R}^d$$

For simplicity: but everything generalizes to (proximal) non-smooth terms

The **gradient direction** can point **away** from the solutions

Weak MVI: Why do we care?

Assumption (Weak MVI)

$$\langle F(z), z - z^\star \rangle \geqslant \rho \|F(z)\|^2$$

Two main reasons:

- The counterexample [Hsieh et al.,2021,Ex. 5.2] is solvable
- Turns out to be fundamental (pops out of the analysis)

We should expect structure to be needed:

- *Minimization*: local solutions can be found efficiently for nonconvex
- *Minimax*: even a local solution is in general intractable
  [Hirsch & Vavasis, 1987, Papadimitriou, 1994, Daskalakis et al., 2021]

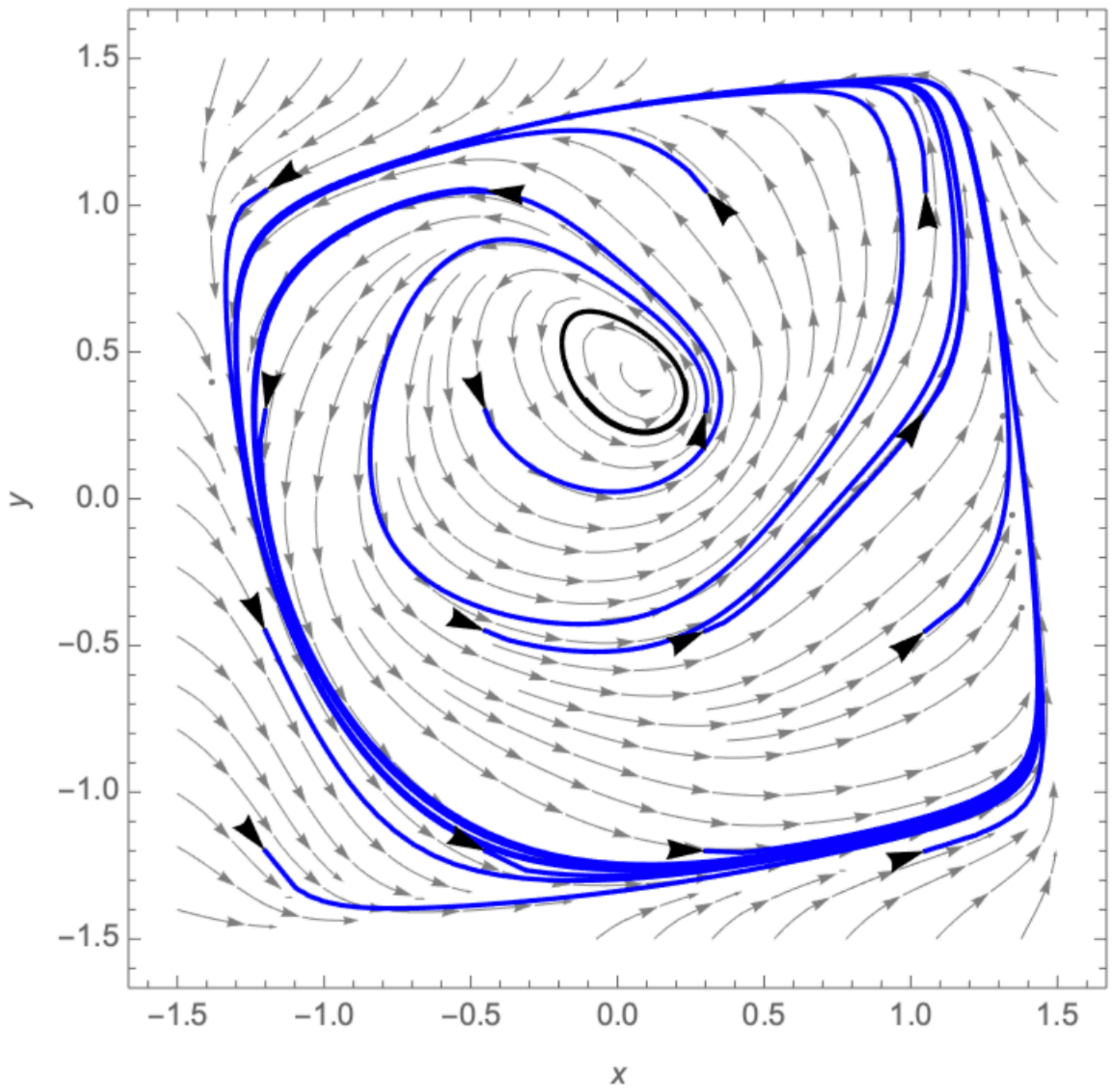


Figure 1. [Hsieh et al., 2021, Ex. 5.2]

How do we ensure progress?

Assumption (Weak MVI)

$\langle F(z), z - z^\star \rangle \geqslant \rho \|F(z)\|^2$

Halfspace construction: Given $\bar{z} \in \mathbb{R}^d$ and $\bar{v} = \gamma F(\bar{z})$ (to be chosen)

$$\mathscr{D} = \left\{ w \mid \langle \bar{v}, \bar{z} - w \rangle \geq \frac{\rho}{\gamma} \|\bar{v}\|^2 \right\}$$

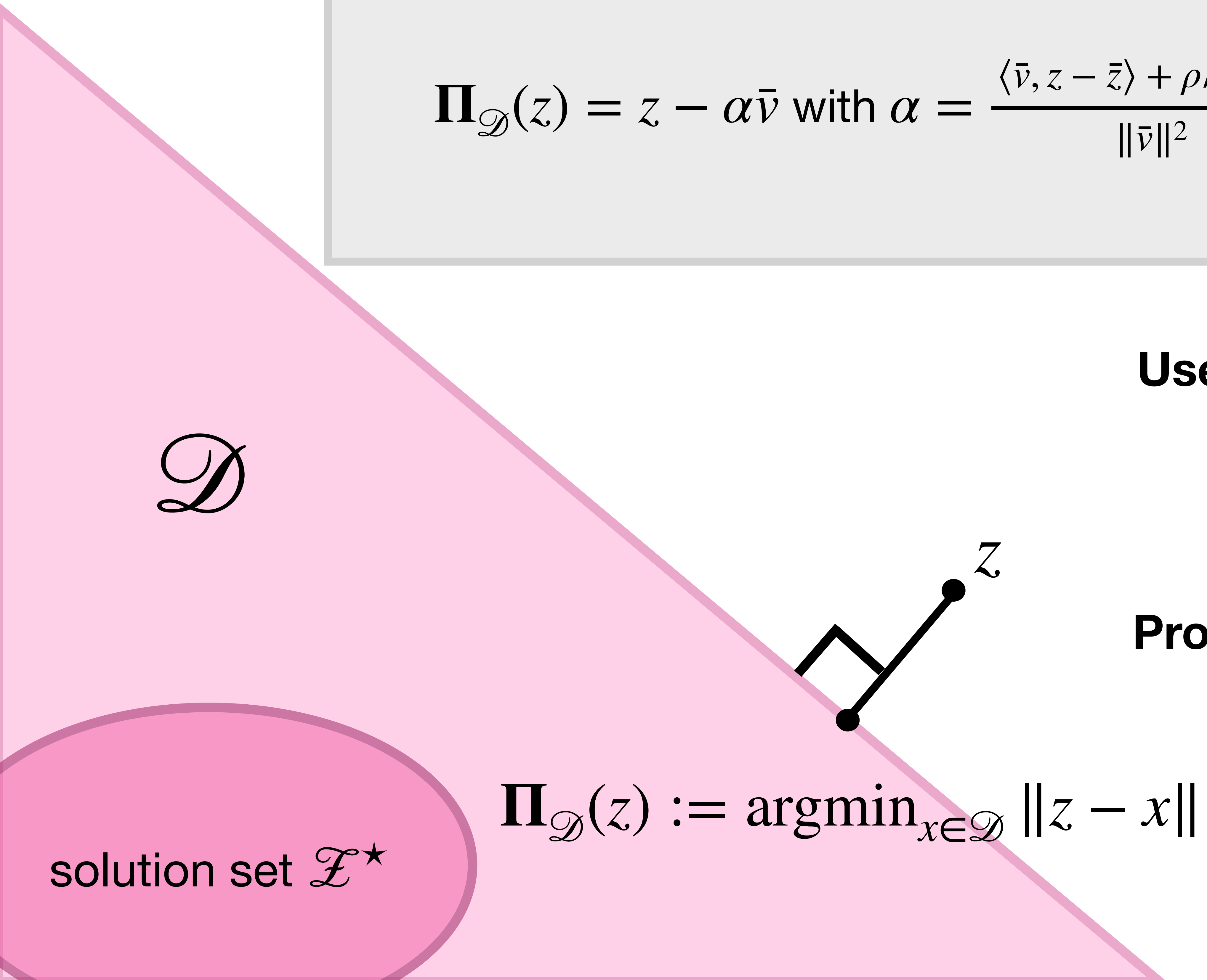
Fact For $z \notin \mathscr{D}$ the projection is given as

$$\mathbf{\Pi}_{\mathscr{D}}(z) = z - \alpha \bar{v} \text{ with } \alpha = \frac{\langle \bar{v}, z - \bar{z} \rangle + \rho/\gamma \|\bar{v}\|^2}{\|\bar{v}\|^2}$$

(i) Contains the solution set, i.e. $\mathscr{Z}^\star \subseteq \mathscr{D}$ ✓

(ii) The projection $\mathbf{\Pi}_{\mathscr{D}}(z)$ moves z towards a fixed point ✓

(iii) We just need to choose $\bar{z}$ such that $\mathrm{fix}\, \mathbf{\Pi}_{\mathscr{D}} \subseteq \mathrm{zer}\, F$!



Use current gradient: Given $z \in \mathbb{R}^d$

$$\bar{z} = z \quad \text{such that} \quad \bar{v} = \gamma F(z) \;\Rightarrow\; \alpha = \rho/\gamma \not> 0$$ ✗

Proximal point: Given $z \in \mathbb{R}^d$ find

$$\bar{z} = z - \bar{v} \quad \text{and} \quad \bar{v} = \gamma F(\bar{z}) \;\Rightarrow\; \alpha = 1 + \rho/\gamma > 0$$ ✓

How do we ensure progress?

Assumption (Weak MVI)

$$\langle F(z), z - z^\star \rangle \geqslant \rho \|F(z)\|^2$$

Inexact proximal point: Given $z \in \mathbb{R}^d$ find, for some error $\varepsilon \in \mathbb{R}^d$,

$$\bar{z} = z - (\bar{v} + \varepsilon) \quad \text{and} \quad \bar{v} = \boxed{\gamma F(\bar{z})}$$

requires extragradient *after* computing $\bar{z}$

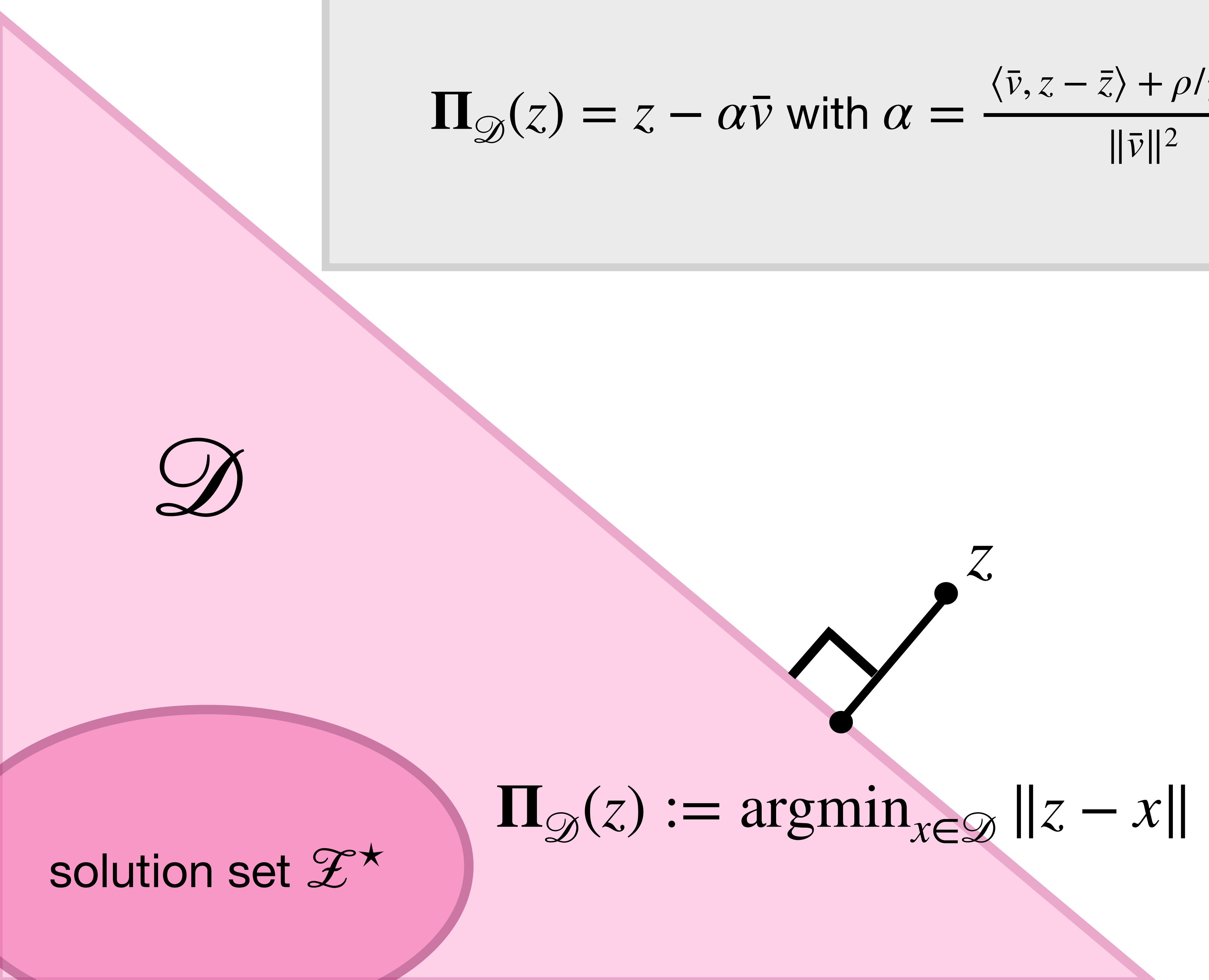
Fact For $z \notin \mathcal{D}$ the projection is given as

$$\mathbf{\Pi}_{\mathcal{D}}(z) = z - \alpha \bar{v} \text{ with } \alpha = \frac{\langle \bar{v}, z - \bar{z} \rangle + \rho/\gamma \|\bar{v}\|^2}{\|\bar{v}\|^2}$$

(i) Contains the solution set, i.e. $\mathcal{Z}^\star \subseteq \mathcal{D}$ ✓

(ii) The projection $\mathbf{\Pi}_{\mathcal{D}}(z)$ moves z towards a fixed point ✓

(iii) We just need to choose ε such that $\operatorname{fix} \mathbf{\Pi}_{\mathcal{D}} \subseteq \operatorname{zer} F$! ✓



Lemma Suppose the following relative error condition is satisfied

$$-\langle \varepsilon, \bar{v} \rangle \leq \sigma \|\bar{v}\|^2, \qquad \sigma \in [0, 1 + \tfrac{\rho}{\gamma})$$

Then

(i) $\mathcal{Z}^\star \subseteq \operatorname{fix} \mathbf{\Pi}_{\mathcal{D}} \subseteq \operatorname{zer} S$

(ii) $\mathbf{\Pi}_{\mathcal{D}}$ is firmly quasi-nonexpansive.

(iii) $\mathbf{\Pi}_{\mathcal{D}}(z) = z - \alpha \bar{v}$ with $\alpha = \frac{\langle \bar{v}, z - \bar{z} \rangle + \rho/\gamma \|\bar{v}\|^2}{\|\bar{v}\|^2} \geq 1 + \frac{\rho}{\gamma} - \sigma$

Dictates the range of ρ

A Hybrid proximal extragradient method

Assumption (Weak MVI)

$$\langle F(z), z - z^\star \rangle \geqslant \rho \|F(z)\|^2$$

Implicit method

$$
\begin{array}{rll}
\text{find} & \bar{z}^k \in \mathbb{R}^d \quad \text{and} \quad \bar{v}^k = \gamma F(\bar{z}^k) \\
\text{s.t.} & \bar{z}^k = z^k - (\bar{v}^k + \varepsilon^k) \quad \text{and} \quad -\langle \varepsilon^k, \bar{v}^k \rangle \leq \sigma \|\bar{v}^k\|^2 \\
\text{update} & z^{k+1} = z^k - \lambda_k \alpha_k \bar{v}^k \quad \alpha_k = \dfrac{\langle \bar{v}^k, z^k - \bar{z}^k \rangle + \delta/\gamma \|\bar{v}^k\|^2}{\|\bar{v}^k\|^2}
\end{array}
$$

Parameters

$\sigma \in [0, 1+\frac{\delta}{\gamma})$

$\delta \leq \rho$

$\lambda_k \in (0,2)$

Special cases:

- $\varepsilon^k = 0 \Rightarrow$ **Relaxed proximal point algorithm (rPPA)**

$$
\begin{aligned}
\bar{z}^k &= z^k - \gamma F(\bar{z}^k) \\
z^{k+1} &= z^k - \alpha\gamma F(\bar{z}^k)
\end{aligned}
\quad \Leftrightarrow \quad
\begin{aligned}
\bar{z}^k &= z^k - \gamma F(\bar{z}^k) \\
z^{k+1} &= (1-\alpha) z^k + \alpha \bar{z}^k
\end{aligned}
$$

for $\alpha \in (0, 2(1+\frac{\rho}{\gamma})) \quad \Rightarrow \quad \rho > -\gamma$

- $\varepsilon^k = \gamma(Fz^k - F\bar{z}^k) \Rightarrow$ **Relaxed extragradient (EG+)**

$$
\begin{aligned}
\bar{z}^k &= z^k - \gamma F(z^k) \\
z^{k+1} &= z^k - \alpha\gamma F(\bar{z}^k)
\end{aligned}
\quad \Leftrightarrow \quad
\begin{aligned}
\bar{z}^k &= z^k - \gamma F(z^k - \gamma F(z^k)) \\
z^{k+1} &= (1-\alpha) z^k + \alpha \bar{z}^k
\end{aligned}
$$

for $\alpha > 1+\frac{2\rho}{\gamma} > 0 \quad \Rightarrow \quad \rho > -\frac{\gamma}{2}$

Absorbing adaptive α_k into the relaxation parameter λ_k

Error correction property: The hyperplane projection makes the scheme behave like exact rPPA *despite error*

A Hybrid method (explicit)

Assumption (Weak MVI)

$$\langle F(z), z - z^\star \rangle \geqslant \rho \|F(z)\|^2$$

Proximal point: Given $z \in \mathbb{R}^d$ find

$$z' = z - \gamma F(z')$$

Approximate with fixed point iteration

$$Q_z : \bar{z} \mapsto z - \gamma F(\bar{z})$$

Algorithm (Explicit method)

For $k = 0,1,...$

1. $\bar{z}^k \leftarrow z^k$
2. **repeat** $\bar{z}^k \leftarrow z^k - \gamma F(\bar{z}^k)$
3. **until** $\langle z^k - \bar{z}^k, \bar{v}^k \rangle \geq (1-\sigma)\|\bar{v}^k\|^2$ where $\bar{v}^k = \gamma F(\bar{z}^k)$
4. $z^{k+1} = z^k - \lambda_k \alpha_k \bar{v}^k$ with $\alpha_k = \dfrac{\langle \bar{v}^k, z^k - \bar{z}^k \rangle + \frac{\delta}{\gamma}\|\bar{v}^k\|^2}{\|\bar{v}^k\|^2}$

error condition

when error condition passes **immediately** $\Rightarrow$

$$\bar{z}^k = z^k - \gamma F(z^k)$$
$$\bar{z}^k = z^k - \lambda_k \alpha_k \gamma F(\bar{z}^k)$$

(relaxed extragradient)

How quickly can the inner loop satisfy the ***error condition****?*

- When $\rho > -\gamma/2$ the error condition can pass **immediately**!
- More inner iteration leads to more relaxed condition on ρ through σ

Theorem (informal) Suppose Assumption 1 & 2 hold. Then

$$\min_{k \in \{0,\ldots,K-1\}} \|F(\bar{z}^k)\|^2 \leq \frac{\|z^0 - z^\star\|^2}{\kappa \gamma^2 (1 + \frac{\delta}{\gamma} - \sigma)^2 K}$$

The sufficient number of inner iteration n:

(i) for $\rho > -\frac{\gamma}{2}$ error condition can pass immediately ($n = 1$)

(ii) for $\rho > -\gamma$ there exists a finite n for which the error condition passes

In contrast with the monotone case:
the extragradient approximation is not for free!
(the approximation error trades off the ρ range)

What have we learned?

Assumption (Weak MVI)

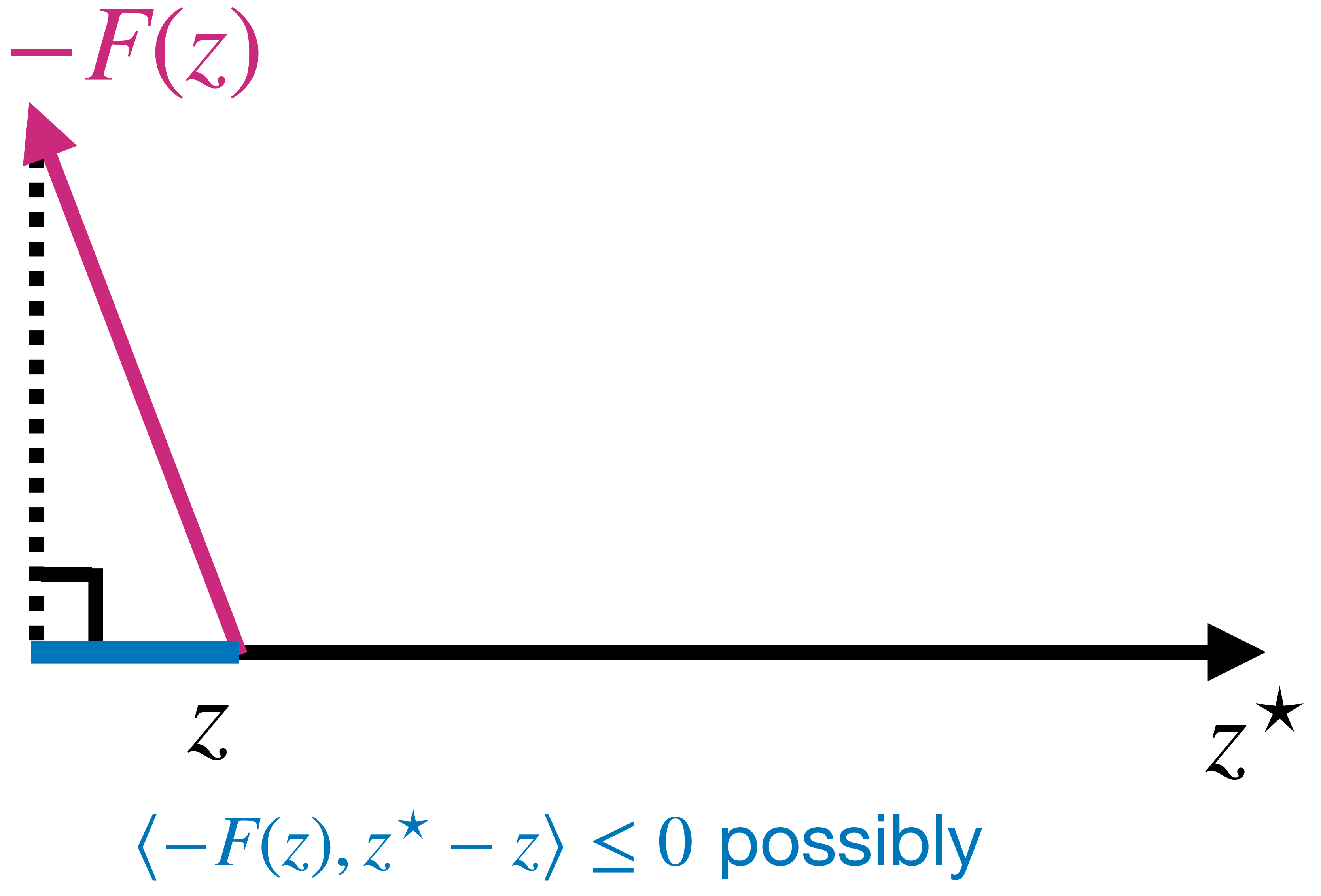
$$\langle F(z), z - z^\star \rangle \geqslant \rho \|F(z)\|^2$$

- Halfspace projections gives convergence by construction!

- Large extrapolation stepsize γ is important, since e.g. $\rho > -\frac{\gamma}{2}$ for relaxed extragradient:

$$\bar{z}^k = z^k - \gamma F(z^k)$$
$$z^{k+1} = z^k - \alpha\gamma F(\bar{z}^k)$$

Provides a challenge in the stochastic case



The **gradient direction** can point **away** from the solutions

- We have a stopping criterion for the proximal solver (the error condition)

Can we leverage it for the client solver in a federated learning setting?

Stochastic case

The stochastic case: A naive attempt

Assumption 1 $F : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ satisfies the weak MVI, i.e.

$$\langle F(z), z - z^\star \rangle \geqslant \rho \|F(z)\|^2 \quad \text{for all } z \in \mathbb{R}^d$$

Assumption 2a $F : \mathbb{R}^d \to \mathbb{R}^d$ is L-Lipschitz

$$\|F(z) - F(z')\| \leqslant L\|z - z'\| \quad \text{for all } z, z' \in \mathbb{R}^d$$

A stochastic oracle:

Assumption 3 The stochastic oracle $\hat{F}(\,\cdot\,, \xi) : \mathbb{R}^d \to \mathbb{R}^d$:

(i) Unbiased $\qquad \mathbb{E}_\xi[\hat{F}(z,\xi)] = F(z)$

(ii) Bounded variance $\qquad \mathbb{E}_\xi[\|\hat{F}(z,\xi) - F(z)\|^2] \le \sigma^2$

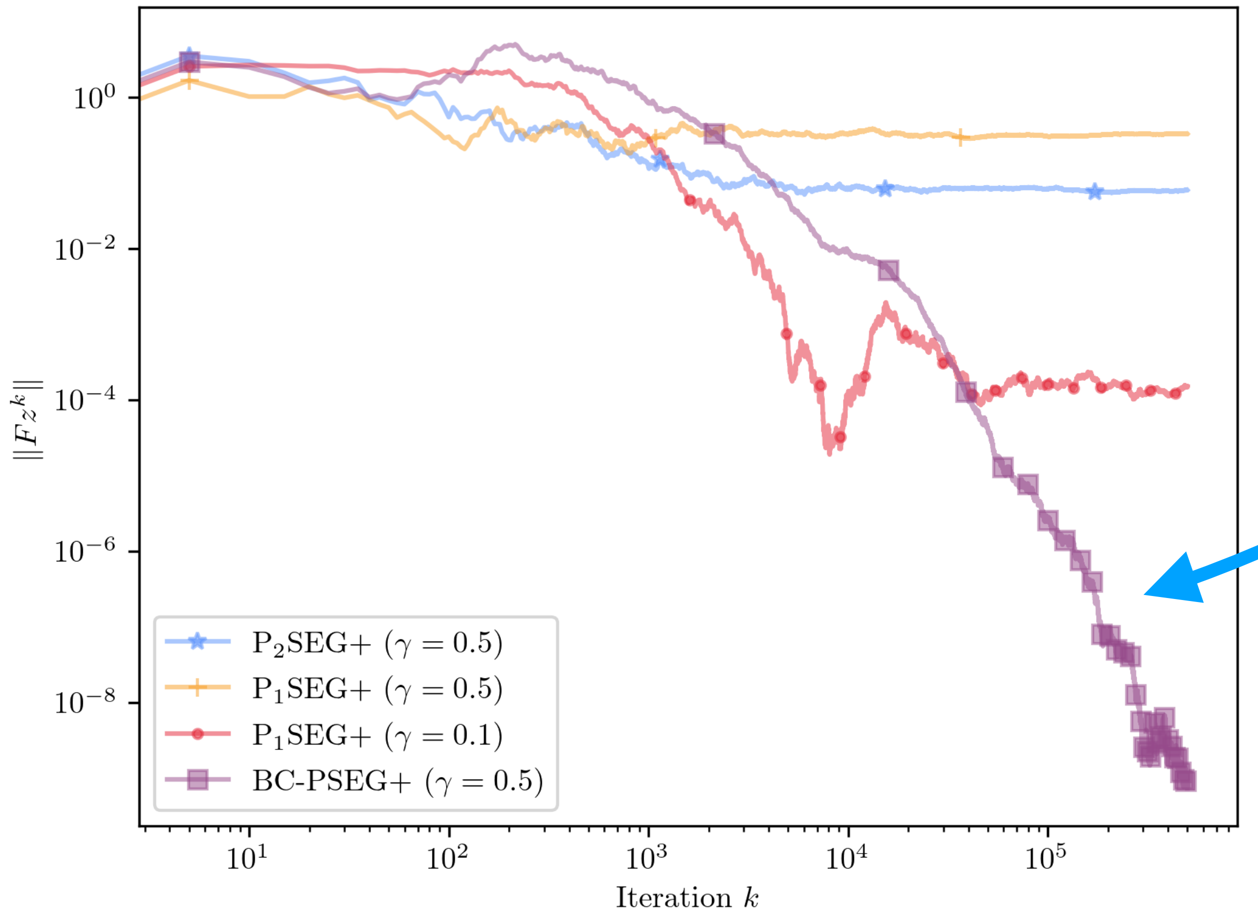
Stochastic extragradient+ (SEG+)

$$\bar{z}^k = z^k - \gamma \hat{F}(z^k, \xi_k) \qquad \xi_k \sim \mathscr{P}$$

$$z^{k+1} = z^k - \alpha_k \gamma \hat{F}(\bar{z}^k, \bar{\xi}_k) \qquad \bar{\xi}_k \sim \mathscr{P}$$

Problem!

- $\hat{F}(z^k, \xi_k)$ is unbiased
- … but $\bar{z}^k$ still has variance
- … so $\hat{F}(\bar{z}^k, \bar{\xi}_k)$ is biased!
- (even when monotone!)



We will see how to fix it

The stochastic case: Bias-correction

Assumption 1 $F : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ satisfies the weak MVI, i.e.

$$\langle F(z), z - z^\star \rangle \geqslant \rho \|F(z)\|^2 \quad \text{for all } z \in \mathbb{R}^d$$

Assumption 2a $F : \mathbb{R}^d \to \mathbb{R}^d$ is L-Lipschitz

$$\|F(z) - F(z')\| \leqslant L\|z - z'\| \quad \text{for all } z, z' \in \mathbb{R}^d$$

Assumption 2b: mean square smoothness

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

<u>**A stochastic oracle:**</u>

Assumption 3 The stochastic oracle $\hat{F}(\,\cdot\,,\xi) : \mathbb{R}^d \to \mathbb{R}^d$:

(i) Unbiased $\quad \mathbb{E}_\xi[\hat{F}(z,\xi)] = F(z)$

(ii) Bounded variance $\quad \mathbb{E}_\xi[\|\hat{F}(z,\xi) - F(z)\|^2] \leq \sigma^2$

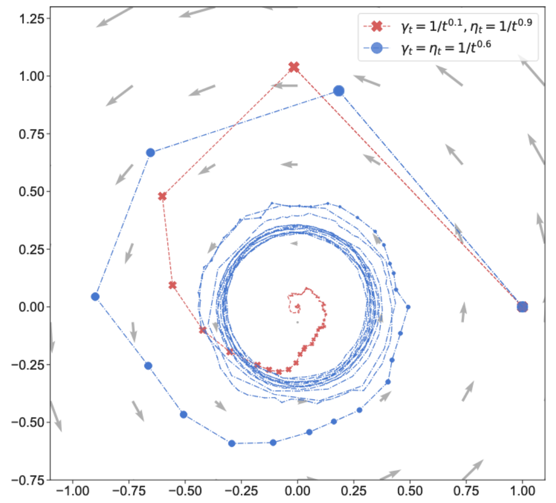
Bias-corrected stochastic extragradient+ (BC-SEG+)

$$\bar{z}^k = z^k - \gamma \hat{F}(z^k, \xi_k) \qquad \xi_k \sim \mathcal{P}$$

$$\boxed{+(1-\alpha_k)(\bar{z}^{k-1} - z^{k-1} + \gamma \hat{F}(z^{k-1}, \xi_k))}$$

$$z^{k+1} = z^k - \alpha_k \gamma \hat{F}(\bar{z}^k, \bar{\xi}_k) \qquad \bar{\xi}_k \sim \mathcal{P}$$

- No additional hyperparameters
- One additional gradient call



[Hsieh et al.,2020]

BC-SEG+ enjoys

interesting even in the monotone case where iterates of SEG cycles!

- $\mathcal{O}(1/\sqrt{K})$ rates for best iterate $\min\limits_{k=0,\dots,K-1} \|F(z^k)\|^2$
- Almost sure convergence (last iterate result)
- As $\alpha_k \to 0$ we recover the deterministic range $\rho \in (-\frac{1}{2L}, \infty)$

The stochastic case: Bias-correction

Algorithm Bias-corrected stochastic extragradient+ (BC-SEG+)

$$\bar{z}^k = z^k - \gamma \hat{F}(z^k, \xi_k) \quad +(1-\alpha_k)(\bar{z}^{k-1} - z^{k-1} + \gamma \hat{F}(z^{k-1}, \xi_k)) \qquad \xi_k \sim \mathscr{P}$$

$$z^{k+1} = z^k - \alpha_k \gamma \hat{F}(\bar{z}^k, \bar{\xi}_k) \qquad \bar{\xi}_k \sim \mathscr{P}$$

Why this correction term?

- We want to approximate the deterministic update:

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

- The error decomposes:

recursively controlled

$$u^k = \boxed{\alpha_k \gamma (F(z^k) - \hat{F}(z^k, \xi_k))} + (1-\alpha_k)\boxed{u^{k-1}} + (1-\alpha_k)\boxed{\left(\gamma(F(z^k) - F(z^{k-1})) + \gamma(\hat{F}(z^{k-1}, \xi_k) - \hat{F}(z^k, \xi_k))\right)}$$

bounded variance: controlled through α_k small

Lipschitz condition converts to $\|z^k - z^{k-1}\|^2$

- Potential function:

$$\mathscr{U}_k := \|z^k - z^\star\|^2 + A_k \boxed{\|u^k\|^2} + B_k \boxed{\|z^k - z^{k-1}\|^2}$$

The stochastic case: What can we conclude?

Algorithm Bias-corrected stochastic extragradient+ (BC-SEG+)

$$\bar{z}^k = z^k - \gamma \hat{F}(z^k, \xi_k) \quad +(1-\alpha_k)(\bar{z}^{k-1} - z^{k-1} + \gamma \hat{F}(z^{k-1}, \xi_k)) \qquad \xi_k \sim \mathscr{P}$$

$$z^{k+1} = z^k - \alpha_k \gamma \hat{F}(\bar{z}^k, \bar{\xi}_k) \qquad \bar{\xi}_k \sim \mathscr{P}$$

- The **stochastic case** tries to approximate the deterministic case
- The bias-correction allows us to take **large stepsizes** γ for the extrapolation (important for nonmonotone)

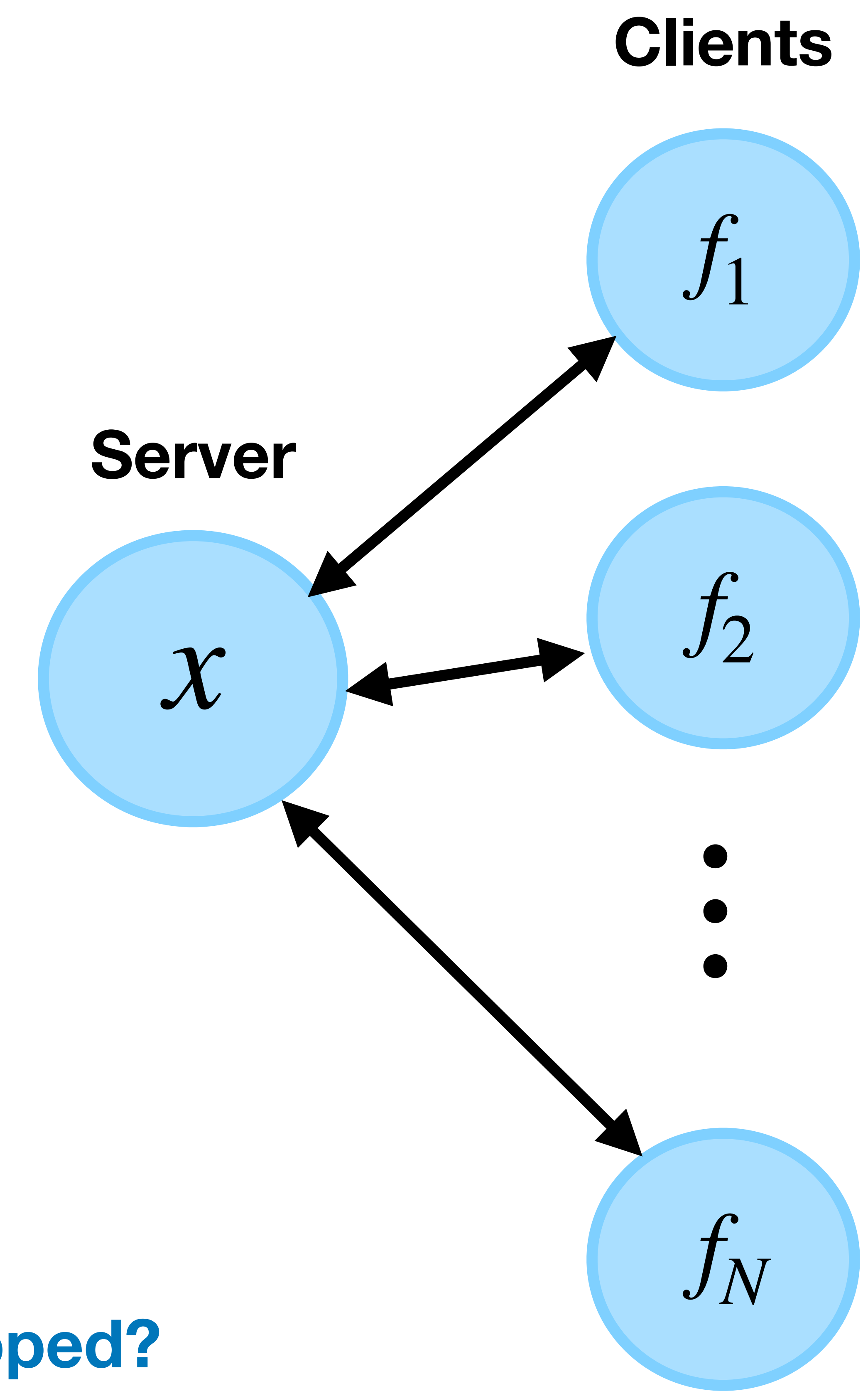
Federated learning

What is the problem?

Finite-sum minimization

$$\text{minimize}_{x\in\mathbb{R}^n} \sum_{i=1}^{N} f_i(x)$$

- f_i is only accessed locally
- We need to specify a local solver
- **Problem**: How do we determine the accuracy needed of the local?



Maybe we can provide a stopping criterion with the inexact prox we developed?

Inexact Federated Douglas-Rachford splitting

Algorithm (iFedDR)

For minimization: $F_i = \nabla f_i$

For k=0,1,…

1. Client computes the approximate prox

$$\bar{x}_i^k = s_i^k - \gamma\big(F_i(\bar{x}_i^k) + \varepsilon_i^k\big)$$

   and sends to server

$$\big(\bar{x}_i^k, F_i(\bar{x}_i^k), s_i^k\big)$$

2. Server computes the (corrected) average

$$\hat{p}^k = \tfrac{1}{N}\textstyle\sum_{i=1}^N (\bar{x}_i^k - \gamma F_i(\bar{x}_i^k))$$

   Check error condition:

$$\textstyle\sum_{i=1}^N \|s_i^k - \gamma F_i(\bar{x}_i^k) - \bar{x}_i^k\|^2 \leq \sigma^2 \max\{\xi_k, \zeta_k\}$$

   If not passed refine step 1. **else** send back to client:

$$(\hat{p}^k, \bar{\alpha}_k) \quad \text{where} \quad \bar{\alpha}_k = \tfrac{\mu_k}{\xi_k}\,.$$

3. Client steps

$$s_i^{k+1} = s_i^k - \lambda\bar{\alpha}_k(\bar{x}_i^k - \hat{p}^k)$$

Scalar quantities

$$\xi_k = \textstyle\sum_{i=1}^N \|\bar{x}_i^k - \hat{p}^k\|^2,$$

$$\zeta_k = \tfrac{1}{\gamma^2}\textstyle\sum_{i=1}^N \|\gamma F_i(\bar{x}_i^k) - s_i^k + \hat{p}^k\|^2,$$

$$\mu_k = \textstyle\sum_{i=1}^N \langle \bar{x}_i^k - \hat{p}^k, s_i^k - \gamma F_i(\bar{x}_i^k) - \hat{p}^k\rangle\,.$$

Experiments

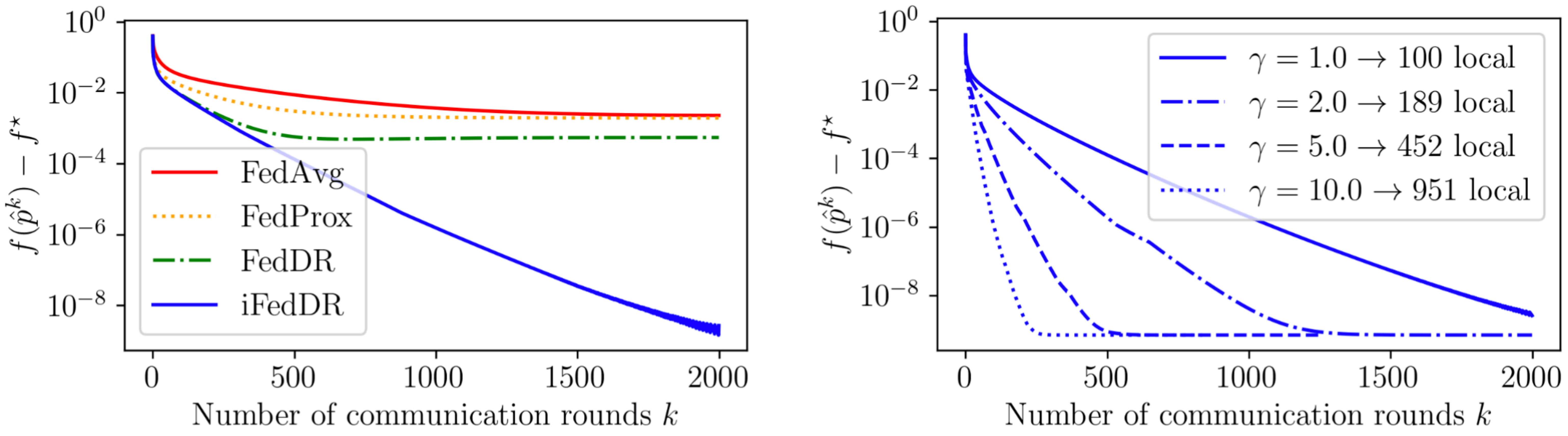


Figure 1: *Logistic regression on the heterogeneous vehicle dataset*

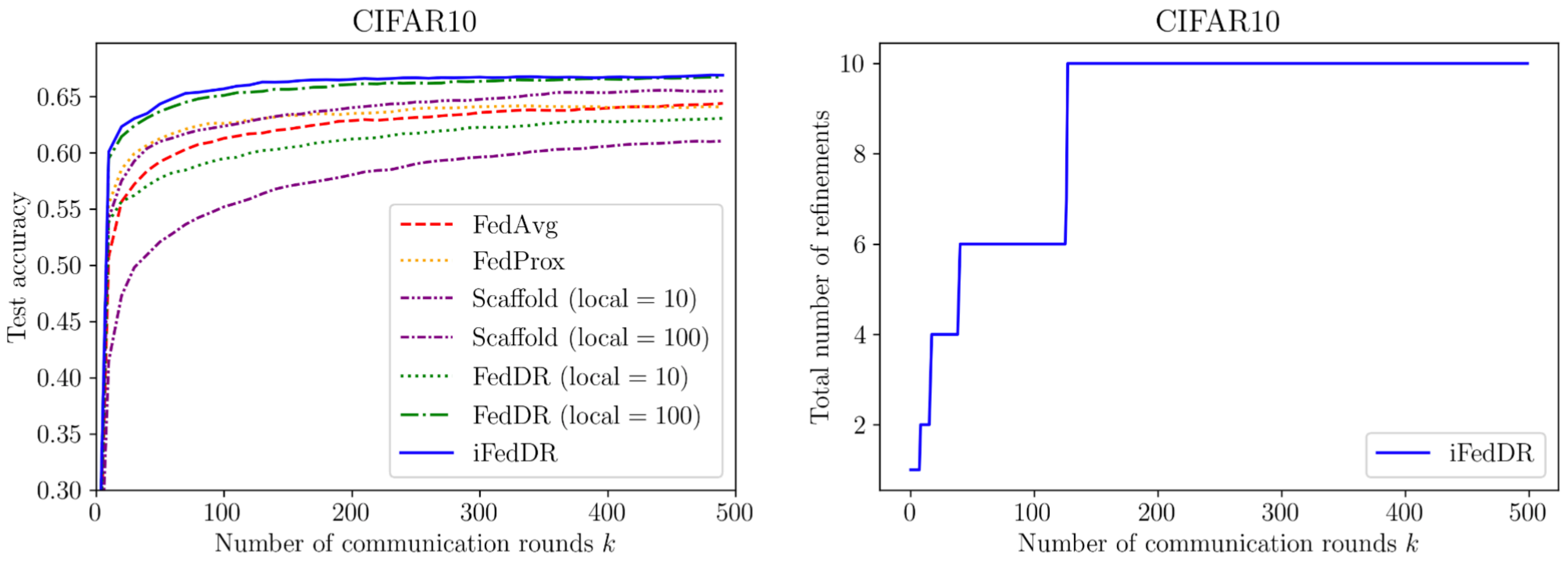


Figure 2: *Linear probing on CIFAR10 under heterogeneous data split*

Conclusion

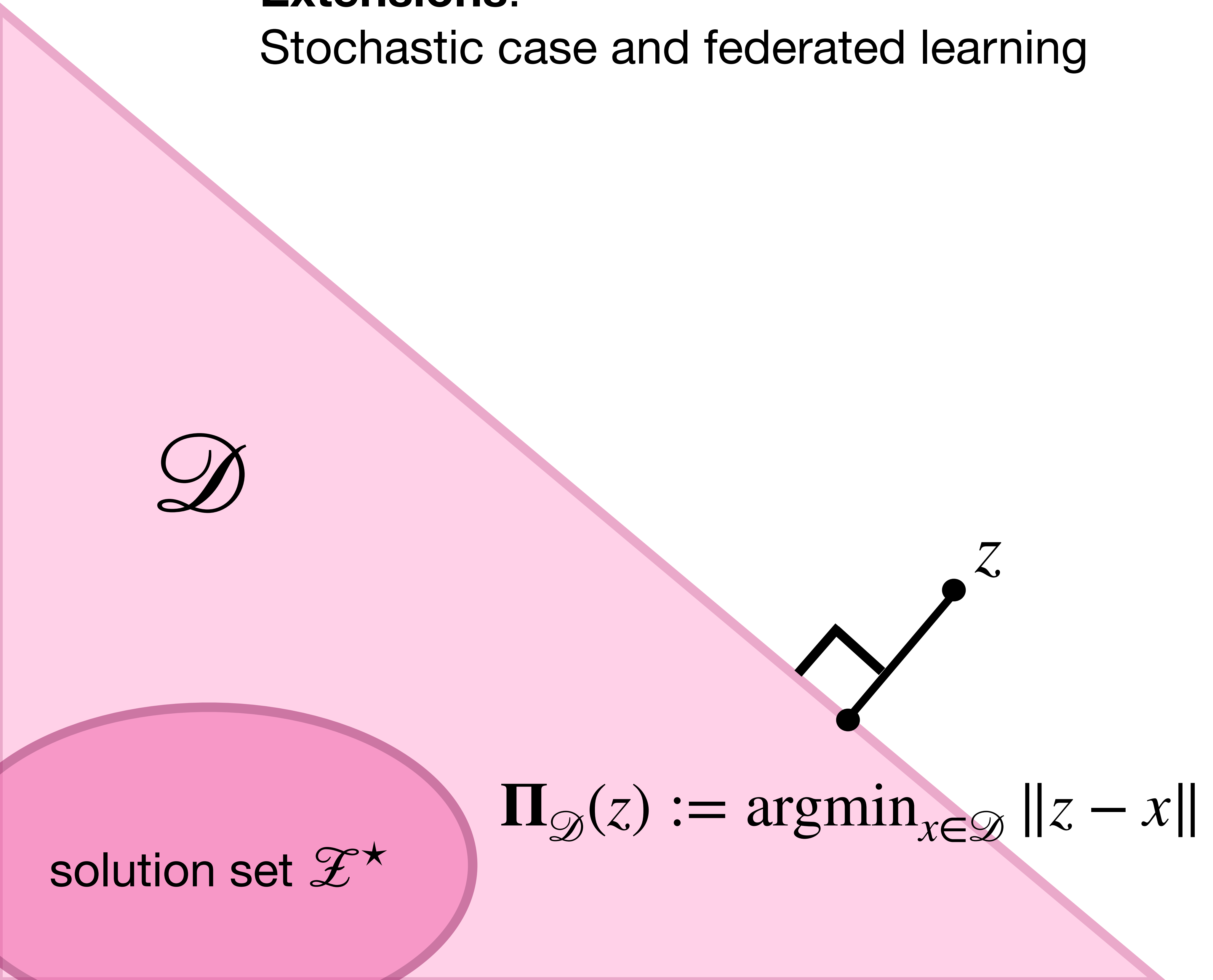
What have we learned?

- **Nonmonotone**:
  The operator direction can point away from the solutions
- **Halfspace projection**:
  Convergence by constructing a halfspace containing the solution
- **Extensions**:
  Stochastic case and federated learning

$\langle -F(z), z^\star - z\rangle \leq 0$ possibly

Assumption (Weak MVI)

$\langle F(z), z - z^\star\rangle \geqslant \rho \|F(z)\|^2$



Open problems

Last iterate

- Extragradient can converge (but only $\rho > -1/8L$)
  [Gorbunov et al., 2022]
- Relaxed inexact prox converges (but suffers log factor)
  [Alacaoglu et al., 2024, P. et al., 2023]

Can relaxed extra gradient enjoy last iterate guarantees (under cohypomonotonicity)?

Stochastic

- Increasing batch size
  [Diakonikolas et al., 2021, Alacaoglu et al., 2024]
- Lipschitz continuous in mean
  [P. et al., 2023]

Can increasing batch size be avoided without additional assumptions?

Single-call with constraints

- Unrelaxed method (so restricted ρ)
  [Cai & Zheng 2023]
- Relaxed method (but unconstrained)
  [Böhm, 2022]

Can a single-call method converge for $\rho > -1/2L$ with constraints?

Halpern / Accelerated methods

- Anchoring (but only for $\rho > -1/2L$)
  [Lee & Kim, 2021]
- Inexact Halpern (but suffers logarithmic)
  [Alacaoglu et al., 2024]

Can we shave off the logarithmic factor for inexact Halpern?

Appendix

Oracle complexity

Key takeaway We can satisfy the error condition in a *finite* number of inner iterations

Corollary (informal)

The explicit scheme (Algorithm 1) achieves $\min_{k \in \{0,\ldots,K-1\}} \operatorname{dist}(0, S\bar{z}^k)^2 \leq \epsilon$ after at most

$$\#(\textbf{oracle calls}) \leq \frac{c\|z^0 - z^\star\|^2}{\gamma^2(1 + \frac{\delta}{\gamma} - \sigma)^2 \epsilon}$$

No log factor in ϵ !

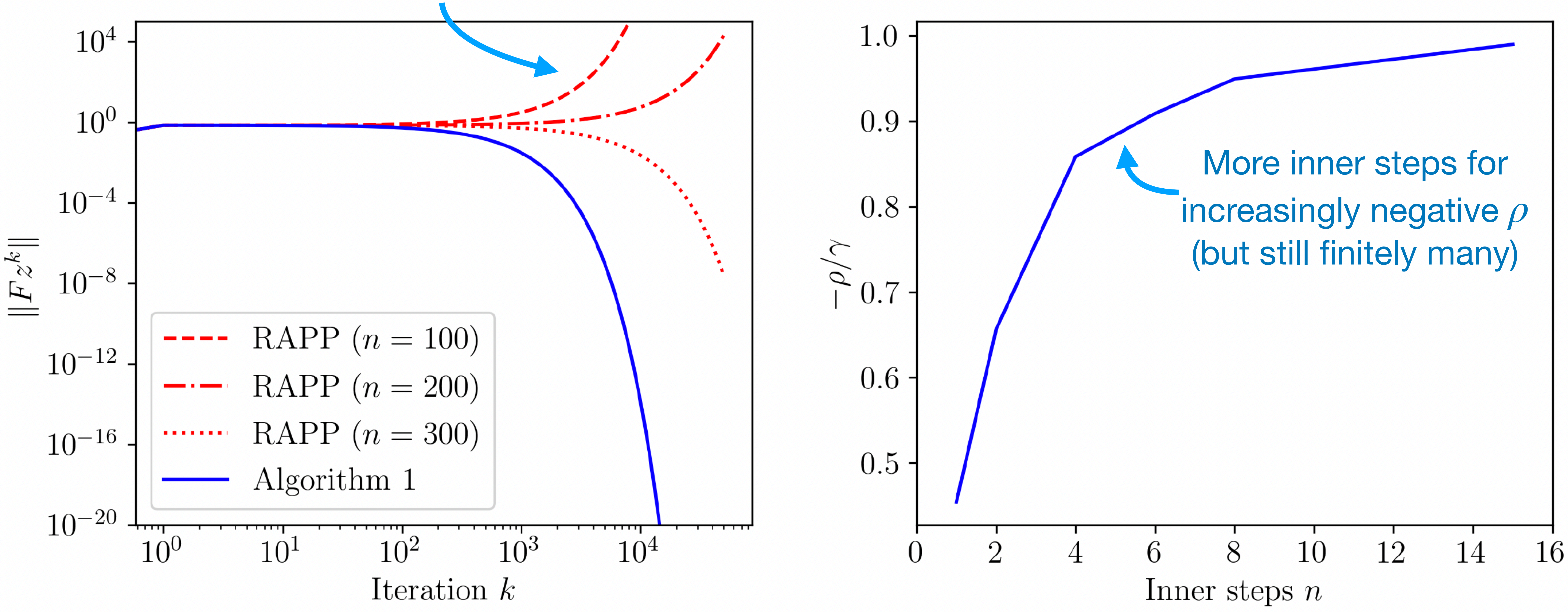
to both the operator F and the resolvent $(\mathrm{id} + \gamma A)^{-1}$ where $c = \tilde{\mathcal{O}}(\frac{1}{1+\rho L})$.

- Improves the complexity from $\mathcal{O}(\frac{1}{\epsilon} \ln \frac{1}{\epsilon})$ to $\mathcal{O}(\frac{1}{\epsilon})$
- Removes the need for prespecifying the number of inner steps n and the stepsize α_k.

Experiments

Lower bound example

Too few inner iterations diverges



More inner steps for increasingly negative ρ (but still finitely many)

A Hybrid method

Assumption (Weak MVI)

$\langle v, z - z^\star \rangle \geqslant \rho \|v\|^2$ for all $(v, z) \in \operatorname{grph} S$

Implicit method

$$\begin{array}{ll} \text{find} & \bar{z}^k \in \mathbb{R}^d \quad \text{and} \quad \bar{v}^k \in \gamma S \bar{z}^k \\ \text{s.t.} & \bar{z}^k = z^k - (\bar{v}^k + \varepsilon^k) \quad \text{and} \quad -\langle \varepsilon^k, \bar{v}^k \rangle \leq \sigma \|\bar{v}^k\|^2 \\ \text{update} & z^{k+1} = z^k - \bar{\alpha} \bar{v}^k \end{array}$$

Nonadaptive stepsize

Theorem (informal) The scheme converges if $\rho > -(1 + \sigma + \bar{\alpha}/2)\gamma$.

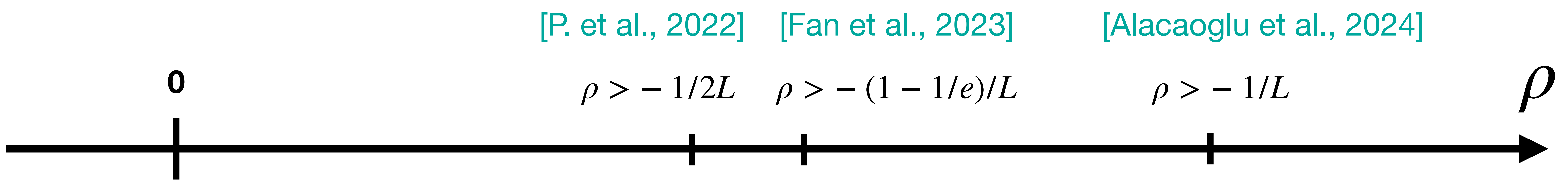
Proof.

$$\begin{aligned} \|z^{k+1} - z^\star\|^2 &= \|z^k - z^\star\|^2 + \bar{\alpha}^2 \|\bar{v}^k\|^2 - 2\bar{\alpha}\langle \bar{v}^k, z^k - z^\star \rangle \\ &= \|z^k - z^\star\|^2 + \bar{\alpha}^2 \|\bar{v}^k\|^2 - 2\bar{\alpha}\langle \bar{v}^k, z^k - \bar{z}^k \rangle - 2\bar{\alpha}\langle \bar{v}^k, \bar{z}^k - z^\star \rangle \\ &= \|z^k - z^\star\|^2 - 2\bar{\alpha}(1 - \tfrac{\bar{\alpha}}{2})\|\bar{v}^k\|^2 - 2\bar{\alpha}\langle \bar{v}^k, \varepsilon^k \rangle - 2\bar{\alpha}\langle \bar{v}^k, \bar{z}^k - z^\star \rangle \\ \text{(error condition)} &\leq \|z^k - z^\star\|^2 - 2\bar{\alpha}(1 - \sigma - \tfrac{\bar{\alpha}}{2})\|\bar{v}^k\|^2 - 2\bar{\alpha}\langle \bar{v}^k, \bar{z}^k - z^\star \rangle \\ \text{(weak MVI)} &\leq \|z^k - z^\star\|^2 - 2\bar{\alpha}\underbrace{(1 - \sigma - \tfrac{\bar{\alpha}}{2} + \tfrac{\rho}{\gamma})}_{\text{Needs to be positive}}\|\bar{v}^k\|^2 \end{aligned}$$

$\square$

Existing literature

Assumption (Weak MVI)
$\langle v, z - z^\star \rangle \geqslant \rho \|v\|^2$ for all $(v, z) \in \operatorname{grph} S$



Monotone (w.r.t. solution)

Negative $\rho \Rightarrow$ increasingly nonmonotone

relaxed update: $z^{k+1} = (1-\alpha) z^k + \alpha \bar{z}^k$
($A \equiv 0$)

$$\begin{cases} \bar{z}^k = z^k - \gamma F(z^k - \gamma F z^k) & \text{(extragradient)} \quad \Rightarrow \quad \rho > -1/2L \\ \bar{z}^k = z^k - \gamma (F \bar{z}^k - \varepsilon^k) & \text{(proximal point)} \quad \Rightarrow \quad \rho > -1/L \end{cases}$$

Method	Minimum ρ	Complexity[1]	Interpolates	Stopping criteria[2]	Constraints	Fejér monotone
Pethick et al. [2022]	$-\frac{1}{2L}$	$\mathcal{O}(\frac{1}{\varepsilon})$	✗	-	✓	✓
Fan et al. [2023]	$-\frac{1-1/e}{L}$	$\mathcal{O}(\frac{1}{\varepsilon})$	✓	✗	✗	✓
Alacaoglu et al. [2024]	$-\frac{1}{L}$	$\mathcal{O}(\frac{1}{\varepsilon}\ln\frac{1}{\varepsilon})$	✗	✗	✓	✗
This paper	$-\frac{1}{L}$	$\mathcal{O}(\frac{1}{\varepsilon})$	✓	✓	✓	✓

Main contribution $\rho > -1/L$ without suffering a logarithmic factor!

Key insight
- Inaccuracy in the halfspace projection can both correct for a proximal approximation and enlarge the problem class
- We will extend ideas from Solodov and Svaiter [1999] and P. et al. [2022]