EPFL

Efficient Interpolation between Extragradient and Proximal Methods for Weak MVIs

Thomas Pethick, Ioannis Mavrothalassitis, Volkan Cevher

HASLERSTIFTUNG Google AI SDSC ZEISS FNSNF Swiss National Science Foundation

Motivation

Minimax

nonsmooth (i.e. indicator of a constraint set)

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\ \underset{y\in\mathbb{R}^m}{\text{maximize}}\ g(x) + \varphi(x,y) - h(y)$$

Smooth

- **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]
  - exemplified by the extragradient method converging to a limit cycle (Fig. 1)

In todays talk:

- Fundamental question of what nonmonotonicity **first order methods** can handle
- **Separation** between extragradient and prox-based methods (in contrast with monotone)
- We will shave off a **logarithmic factor** in the literature

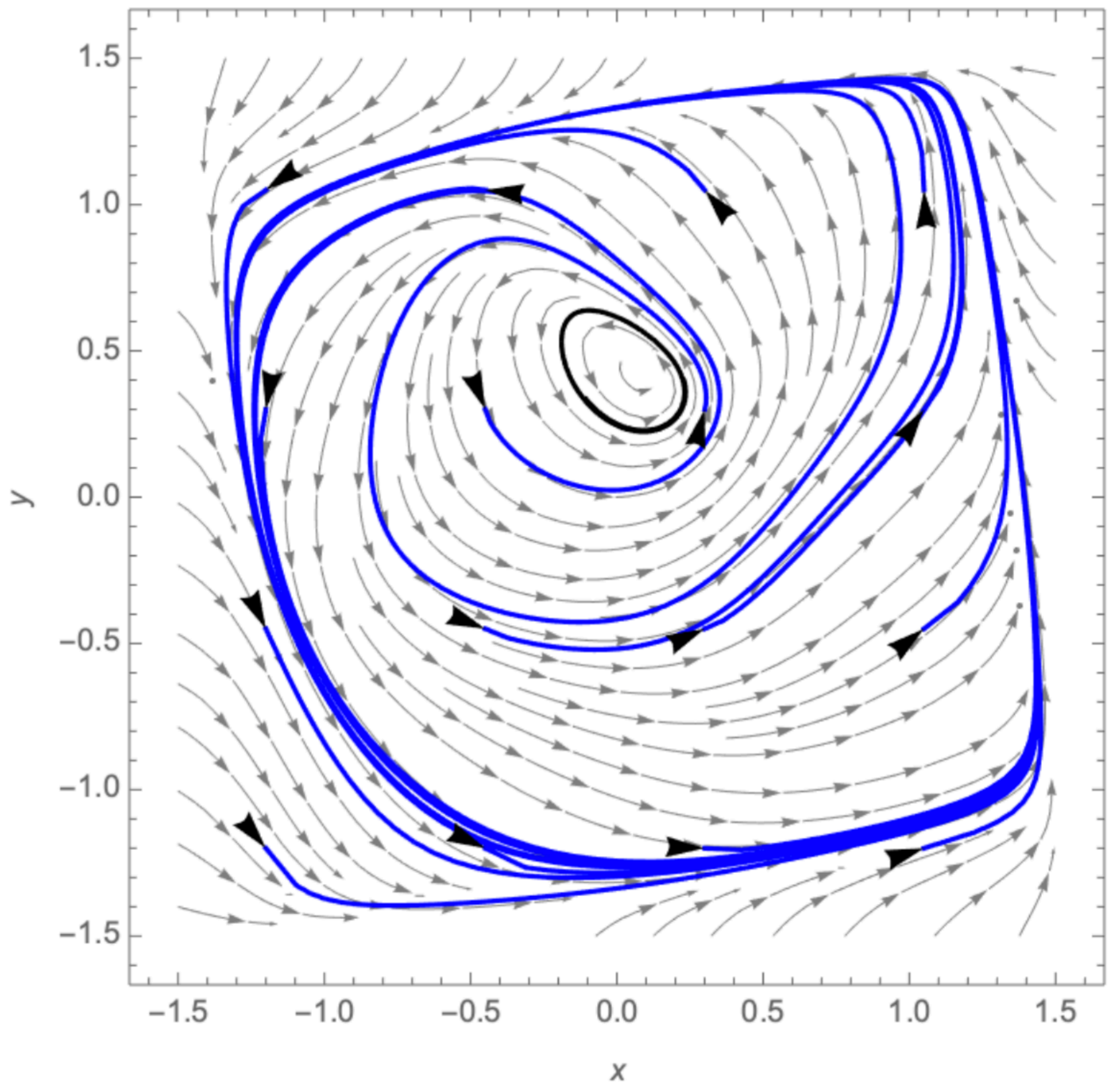


Figure 1. [Hsieh et al., 2021, Ex. 5.2]

Problem formulation

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\ \underset{y\in\mathbb{R}^m}{\text{maximize}}\ g(x) + \varphi(x,y) - h(y)$$

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

Captures first order condition of

$$0 \in Fz + Az =: Sz$$

Operator view: $z = (x, y)$

$$Fz = (\nabla_x \varphi(x,y), -\nabla_y \varphi(x,y))$$

$$Az = (\partial g(x), \partial h(y))$$

$$(\mathrm{id} + \gamma A)^{-1}(z) = (\mathrm{prox}_{\gamma g}(x), \mathrm{prox}_{\gamma h}(x))$$

Assumption 1 $F : \mathbb{R}^d \to \mathbb{R}^d$ is L-Lipschitz

Assumption 2 $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is maximally monotone

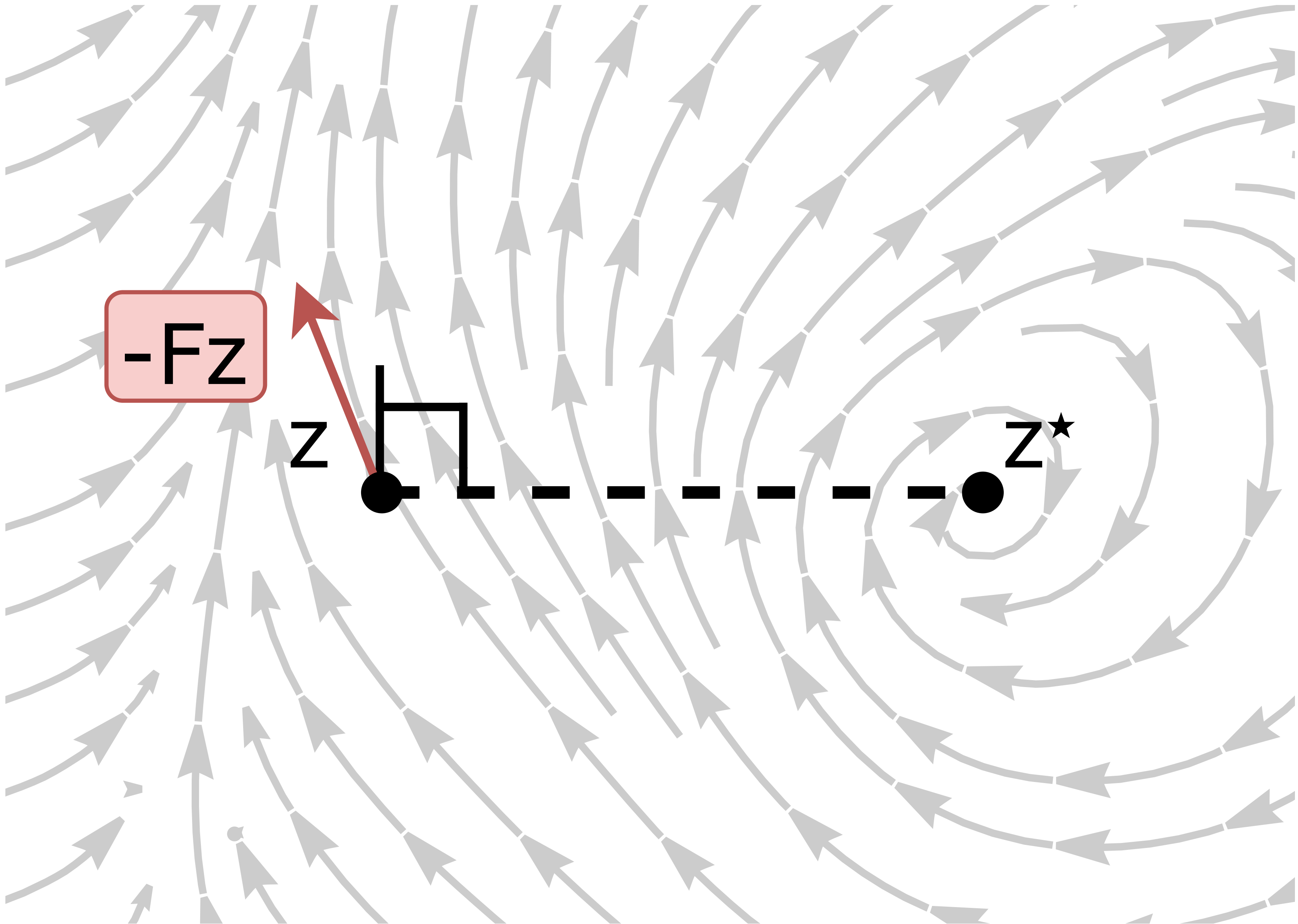
Assumption 3 $S : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ satisfies the weak Minty variational inequality (MVI), i.e. for all $z^\star \in \mathcal{Z}^\star \subseteq \operatorname{zer} S$ (nonempty $\mathcal{Z}^\star$) and some $\rho \in (-\frac{1}{L}, \infty)$

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

Nonmonotone when negative!

$$\Downarrow (A \equiv 0)$$

$$\langle Fz, z - z^\star \rangle \geqslant \rho \|Fz\|^2$$



Why do we care?

- The structure can capture [Hsieh et al.,2021,Ex. 5.2]
- Turns out to be fundamental (pops out of the analysis)

Existing literature

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

[P. et al., 2022] [Fan et al., 2023] [Alacaoglu et al., 2024]

0 $\quad \rho > -1/2L \quad \rho > -(1-1/e)/L \quad \rho > -1/L \quad \rho$

Monotone (w.r.t. solution)

Negative $\rho \Rightarrow$ increasingly nonmonotone

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

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

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]

A Hybrid method

Assumption (Weak MVI)

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

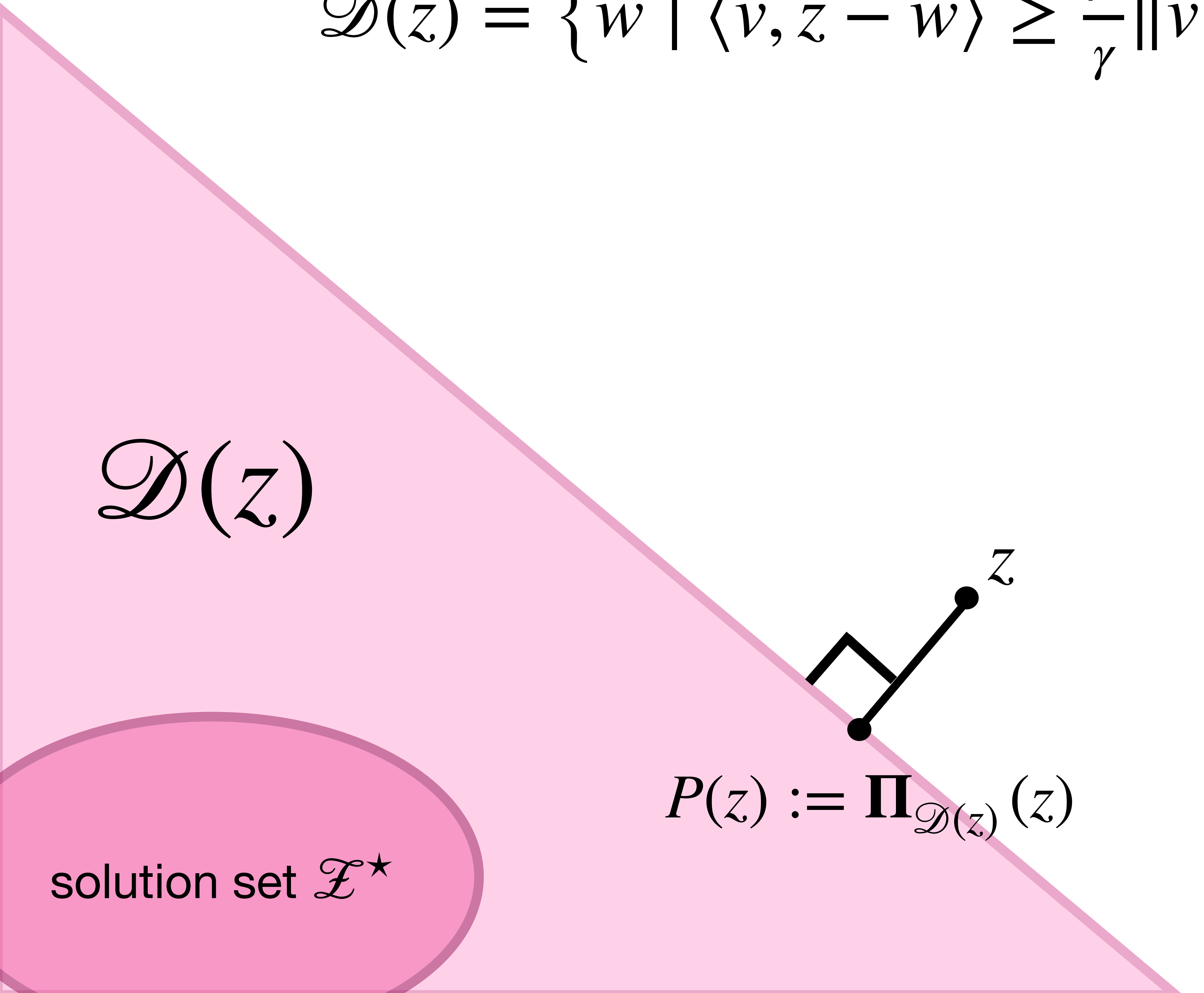
Inexact proximal point: Given $z \in \mathbb{R}^d$ find, for some error $\varepsilon \in \mathbb{R}^d$,

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

Halfspace construction:

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

- Contains the solution set, i.e. $\mathscr{Z}^\star \subseteq \mathscr{D}(z)$
- The projection $P(z) := \mathbf{\Pi}_{\mathscr{D}(z)}(z)$ moves z towards the solutions
- We just need to convince ourselves that $\operatorname{fix} P \subseteq \operatorname{zer} S$!



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) P is firmly quasi-nonexpansive.

(ii) $P(z) = z - \alpha \bar{v}$ with $\alpha = \frac{\langle \bar{v}, z - \bar{z} \rangle + \rho/\gamma \|\bar{v}\|^2}{\|\bar{v}\|^2} \geq 1 + \frac{\rho}{\gamma} - \sigma$

(iii) $\mathscr{Z}^\star \subseteq \operatorname{fix} P \subseteq \operatorname{zer} S$

Dictates the range of ρ

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

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

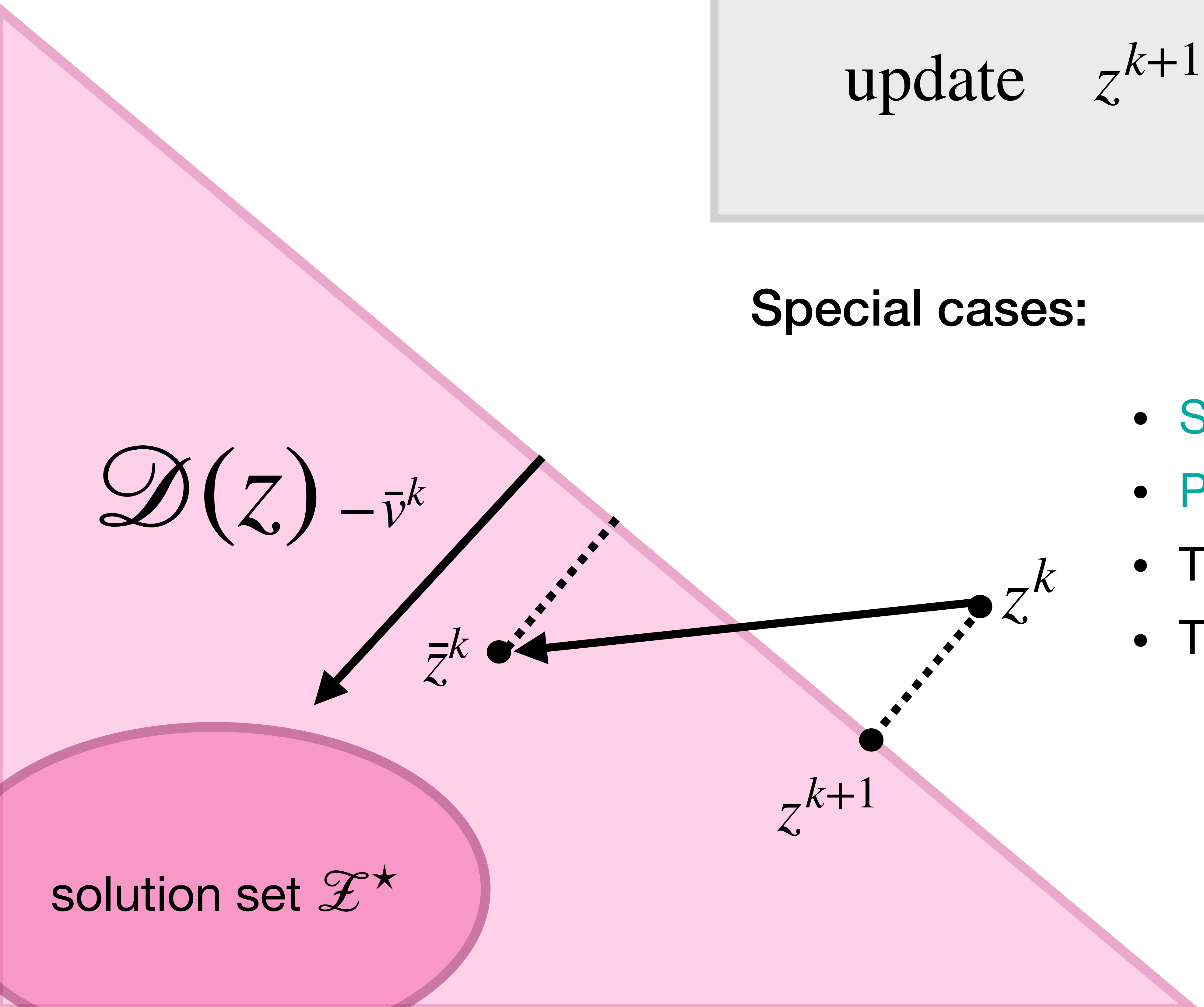
Parameters

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

$\delta \leq \rho$

$\lambda_k \in (0,2)$

Special cases:



- Solodov and Svaiter [1999] ($\rho = 0$, apply Cauchy-Schwarz on error condition)
- P. et al. [2022] ($\varepsilon^k = \gamma(Fz^k - F\bar{z}^k)$)
- The relaxed proximal point algorithm ($\varepsilon^k = 0$)
- The relaxed extragradient method (by absorbing α_k into λ_k)

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.

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

□

A Hybrid method (explicit)

Find $z \in \mathbb{R}^d$ such that $0 \in Fz + Az =: Sz$

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

$$z' = (\mathrm{id} + \gamma S)^{-1} z = (\mathrm{id} + \gamma A)^{-1}(z - \gamma F z')$$

Approximate with fixed point iteration

$$Q_z : \bar{z} \mapsto (\mathrm{id} + \gamma A)^{-1}(z - \gamma F \bar{z})$$

Algorithm 1 (Explicit method)

Repeat for $k = 0, 1, \ldots$
1. $\bar{z}^k \leftarrow z^k$
2. **repeat**
3. $\quad h^k \leftarrow z^k - \gamma F \bar{z}^k$
4. $\quad \bar{z}^k \leftarrow (\mathrm{id} + \gamma A)^{-1} h^k$
5. **until** $\langle z^k - \bar{z}^k, \bar{v}^k \rangle \geq (1-\sigma)\|\bar{v}^k\|^2$ where $\bar{v}^k = h^k - \bar{z}^k + \gamma F \bar{z}^k$
6. $z^{k+1} = z^k - \lambda_k \alpha_k \bar{v}^k$ 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

$\bar{v}^k \in \gamma S \bar{z}^k$

For $A \equiv 0$ 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 $\rho > -1/2L$ the error condition can pass **immediately**!
- More inner iteration leads to more relaxed condition on ρ through σ

Theorem (informal) Suppose Assumption 1-3 holds. Then

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

The sufficient number of inner iteration n:

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

(ii) for $\rho > -\frac{1}{L}$ 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)

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

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

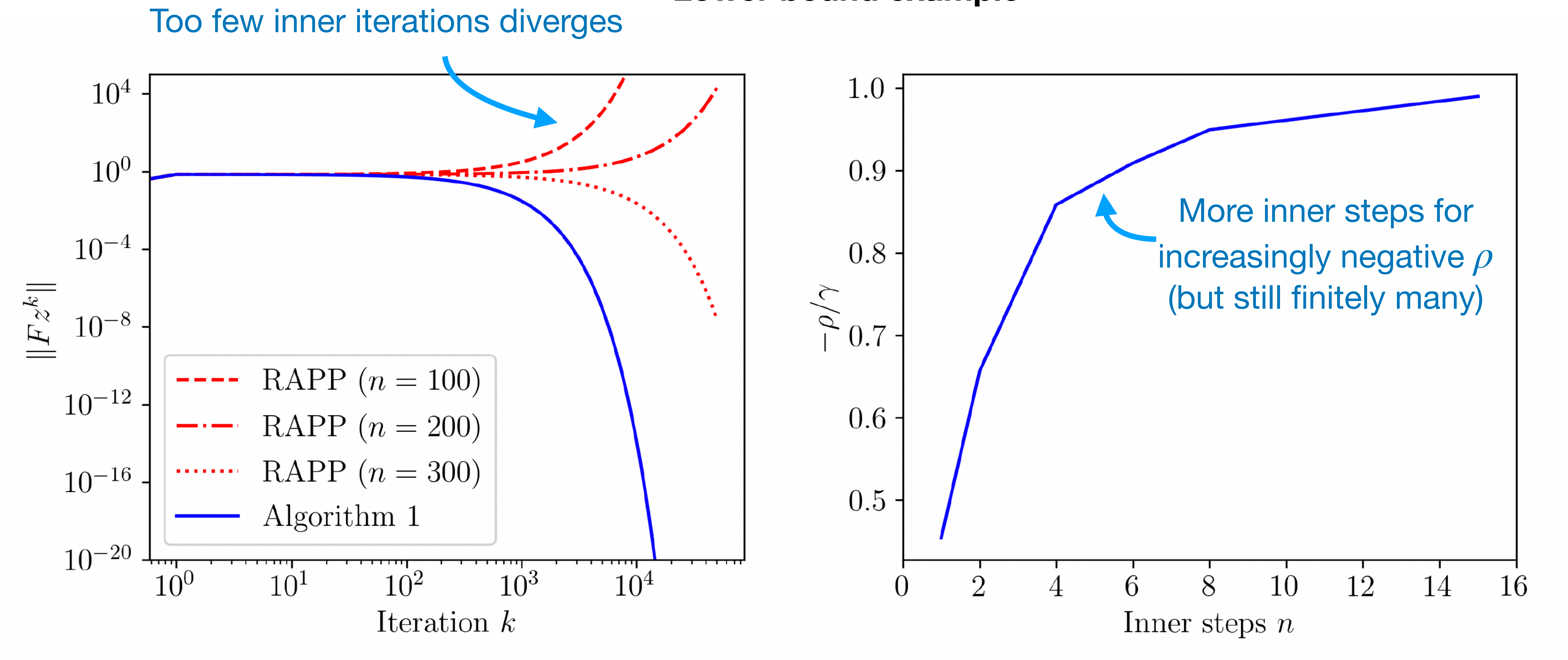
No log factor in ϵ !

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



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

- 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?