Exact Verification of First-order Methods via Mixed-Integer Linear Programming

Bartolomeo Stellato

Department of Operations Research and Financial Engineering
Department of Electrical and Computer Engineering
Department of Computer Science

PRINCETON UNIVERSITY

CRM Workshop: Combinatorial optimization and data science — May 7 2025


Most applications require fast and effective decisions in real-time

Real-time optimization can help us

decisions

$$
\begin{array}{ll}
\text{minimize} & f(z, x) \\
\text{subject to} & z \in C(x)
\end{array}
$$

parameters

objective f: energy consumption, costs
constraints C: dynamics, physical limits

re-planning in real-time is the key to effective decision-making

How do we solve such problems?

First-order methods are now widely popular…

use only first-order information (e.g., gradients)
to solve optimization problems

example
projected gradient descent

$$\begin{array}{ll} \text{minimize} & f(z, x) \\ \text{subject to} & z \in C(x) \end{array}$$

$$z^{k+1} = \Pi_{C(x)}(z^k - \theta \nabla f(z^k, x))$$

projection ↑ gradient step ↑

projection **gradient step**

benefits of first-order methods

✓ cheap iterations

✓ easy to warm-start

embedded optimization

large-scale optimization

…and they can solve many constrained convex problems!

Linear Programs

PDLP

Applegate, Díaz, Hinder, Lu, Lubin, O'Donoghue, Schudy (2021)

Quadratic Programs

OSQP

Stellato, Banjac, Goulart, Bemporad, Boyd (2020)

Conic Programs

SCS

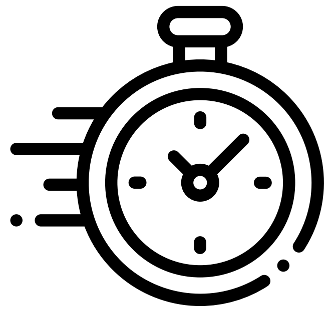
O'Donoghue, Chu, Parikh, Boyd (2016)

many more solvers available by the day: cuOPT, PDCS, ...

But they can converge slowly

major issue in safety-critical applications with

real-time requirements



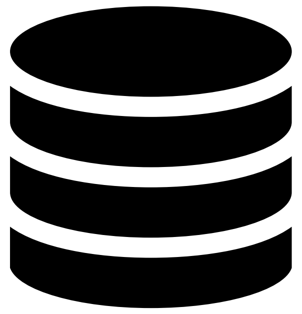
limited computing power

main idea

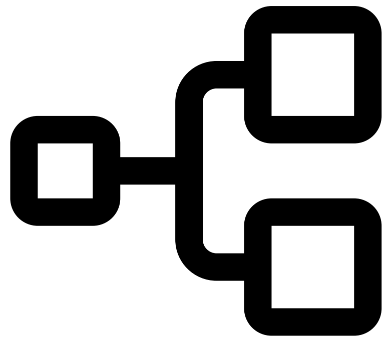
in most applications we repeatedly solve the same problem with varying parameters

$$\begin{array}{ll} \text{minimize} & f(z, x) \\ \text{subject to} & z \in C(x) \end{array}$$

→



large amount of data
(e.g., instances, solutions)



structured problems
(e.g., parameters -> solutions)

Performance verification of first-order methods

Convergence of first-order methods

iterations

$$z^{k+1} = T(z^k, x) \quad \text{for } k = 0, 1, \ldots$$

operator
(e.g., contractive, averaged)

goal: find fixed-points

$$z^\star = T(z^\star, x)$$

example
gradient descent

problem

minimize $f(z, x)$ $\longrightarrow$

optimality conditions

$$\nabla f(z^\star, x) = 0$$

iterations

$$z^{k+1} = z^k - \theta \nabla f(z^k, x)$$

fixed-points

$$z^\star = z^\star - \theta \nabla f(z^\star, x)$$

same as

performance metric

$$r^k(x) = \|T(z^{k-1}) - z^{k-1}\| = \|z^k - z^{k-1}\|$$

fixed-point residual
(converges to 0)

Best known convergence bounds via Performance Estimation

problem

minimize $\quad f(z)$

algorithm

$$z^{k+1} = T(z^k)$$

$$\begin{array}{ll} \text{maximize} & \|z^K - z^{K-1}\| \\ \text{subject to} & f \in \mathcal{F} \\ & z^0, z^1, \ldots, z^K \quad \text{iterates of algorithm } T \text{ on } f \\ & \|z^1 - z^0\| \le r \end{array}$$

performance metric

write using interpolation conditions

function class
(e.g., strongly-convex, smooth functions)

initial condition

convex SDP
Gram matrix reformulation

$$G = \begin{bmatrix} \|z^1 - z^0\|_2^2 & (z^1 - z^0)^T g^1 & (z^1 - z^0)^T g^0 & \\ (z^1 - z^0)^T g^1 & \|g^1\|_2^2 & (g^1)^T g^0 & \\ (z^1 - z^0)^T g^0 & (g^1)^T g^0 & \|g^0\|_2^2 & \\ & & & \ddots \end{bmatrix}$$

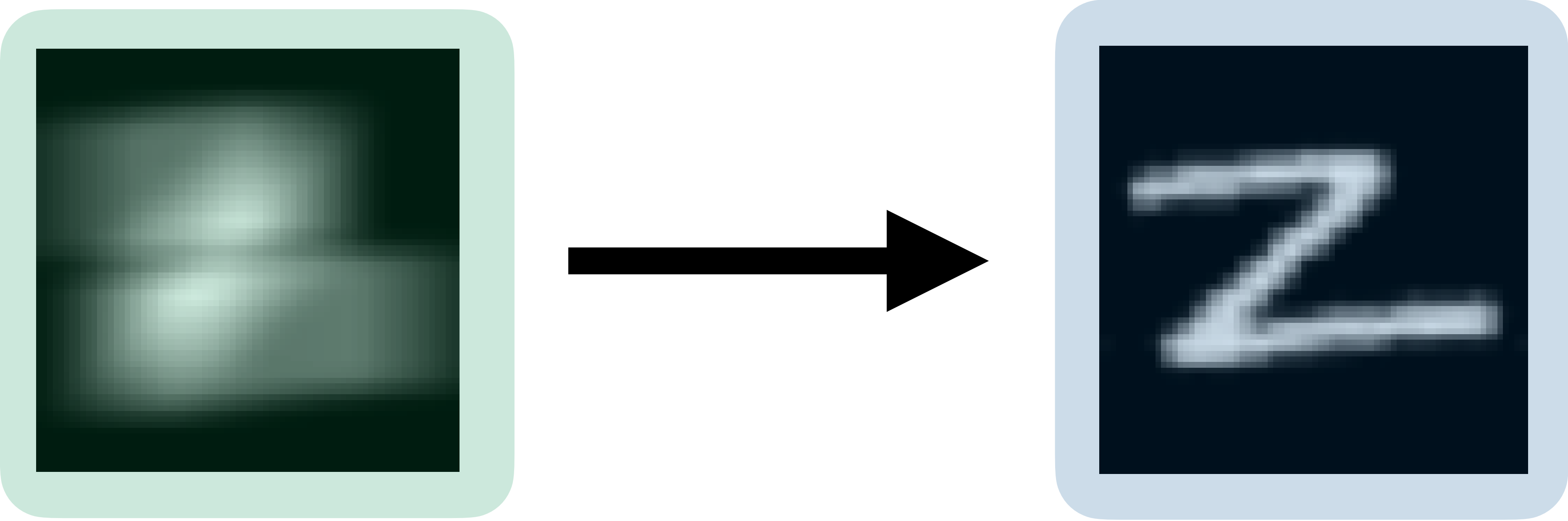
gradients

independent from iterate dimensions

Drori and Teboulle (2014), Taylor, Hendrickx, Gilneur (2017), and many others including Ryu, De Klerk, Altschuler, Grimmer…

Classical worst-case convergence bounds can be very loose

image deblurring problem
emnist dataset

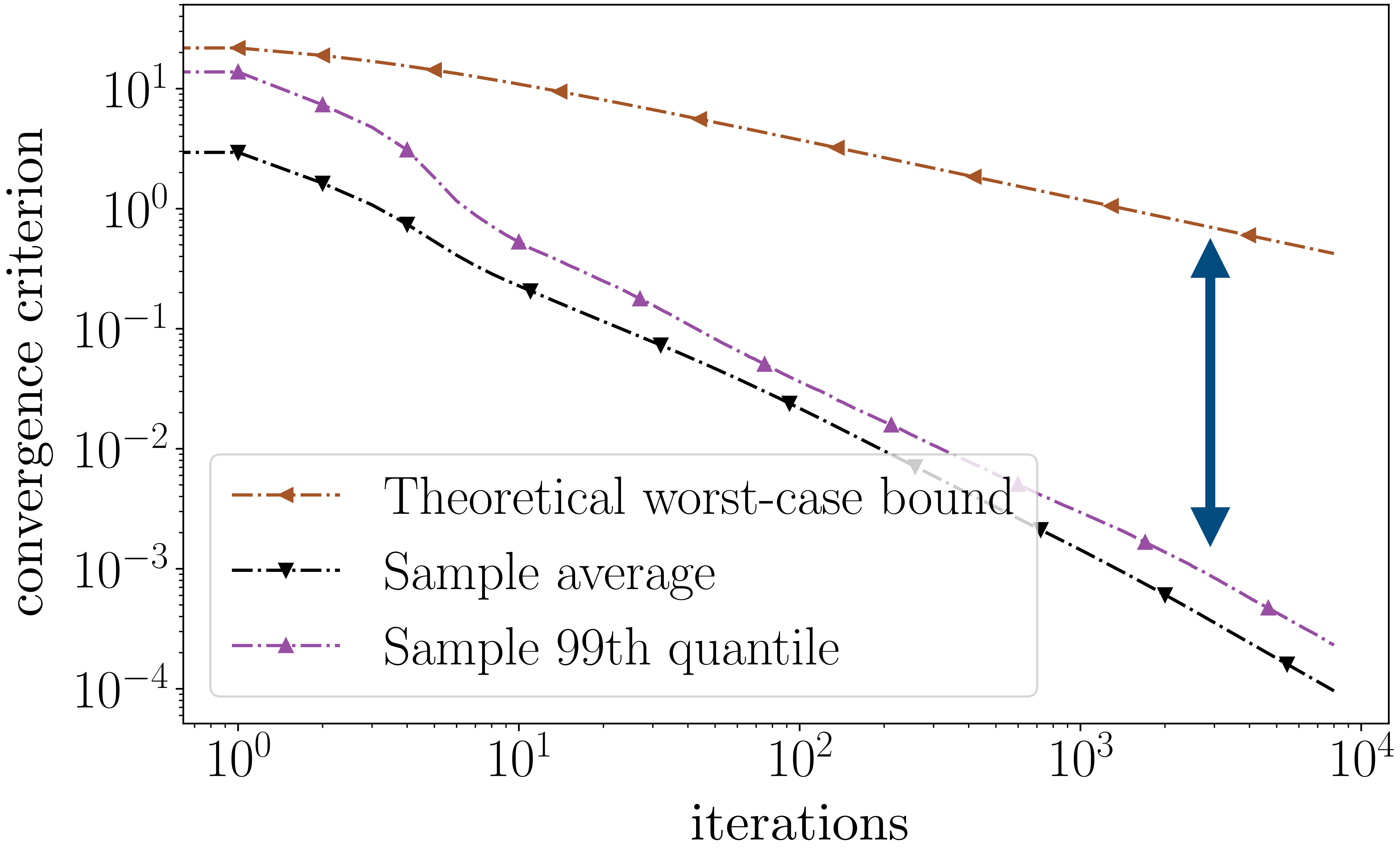


minimize $\|Az - x\|_2^2 + \lambda\|z\|_1$

subject to $0 \leq z \leq 1$

blurred image

deblurred image

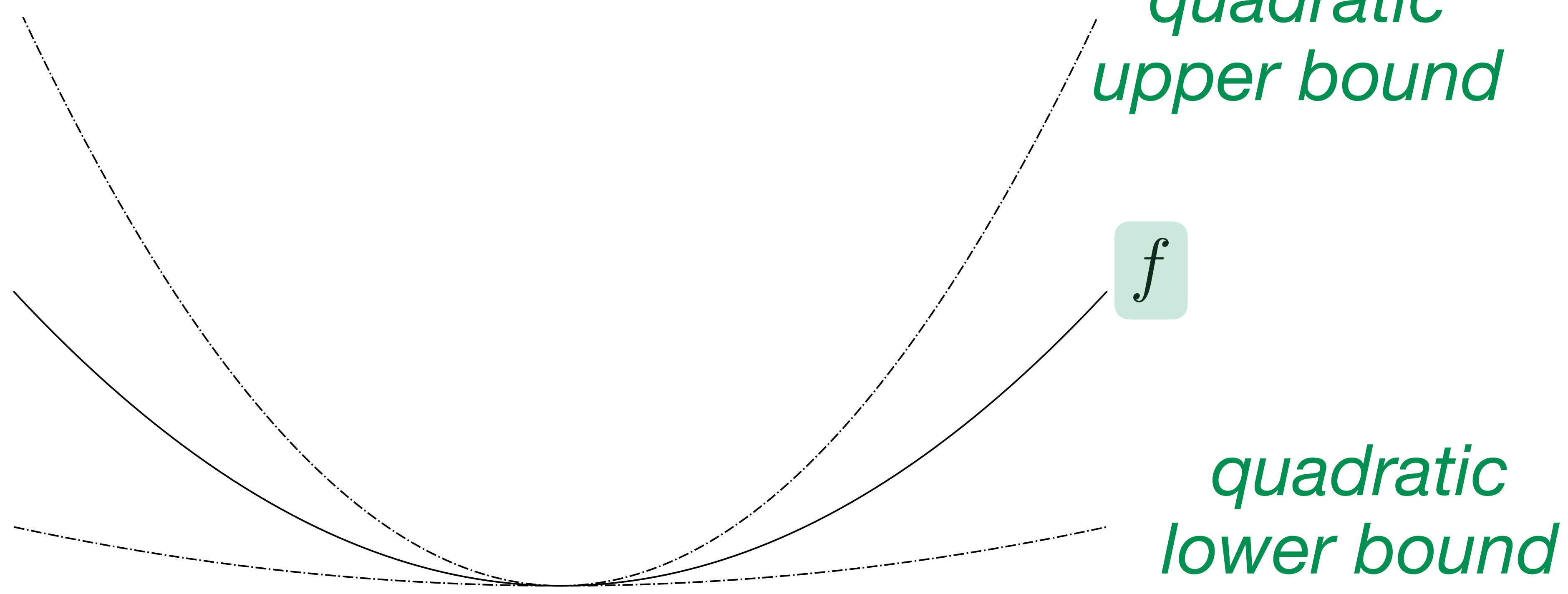


why are worst-case bounds pessimistic?

Issues with classical convergence analysis

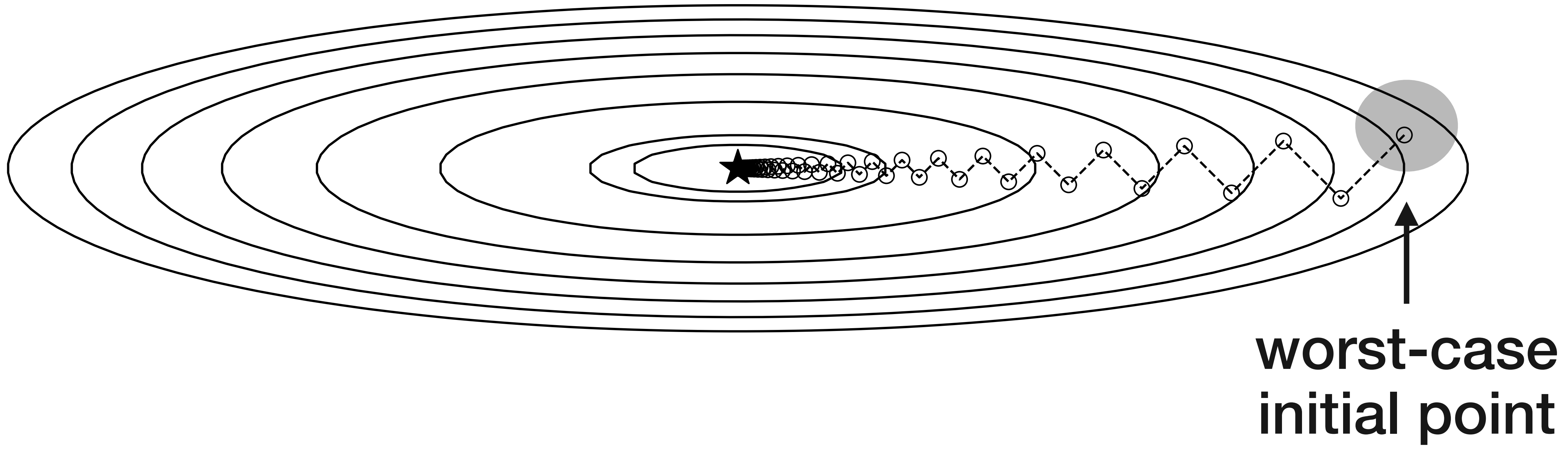
general function classes

(f is strongly convex and smooth...)



we may never encounter that function

pessimistic bounds



we may never start from that point

practical settings

minimize $\quad f(z, x)$

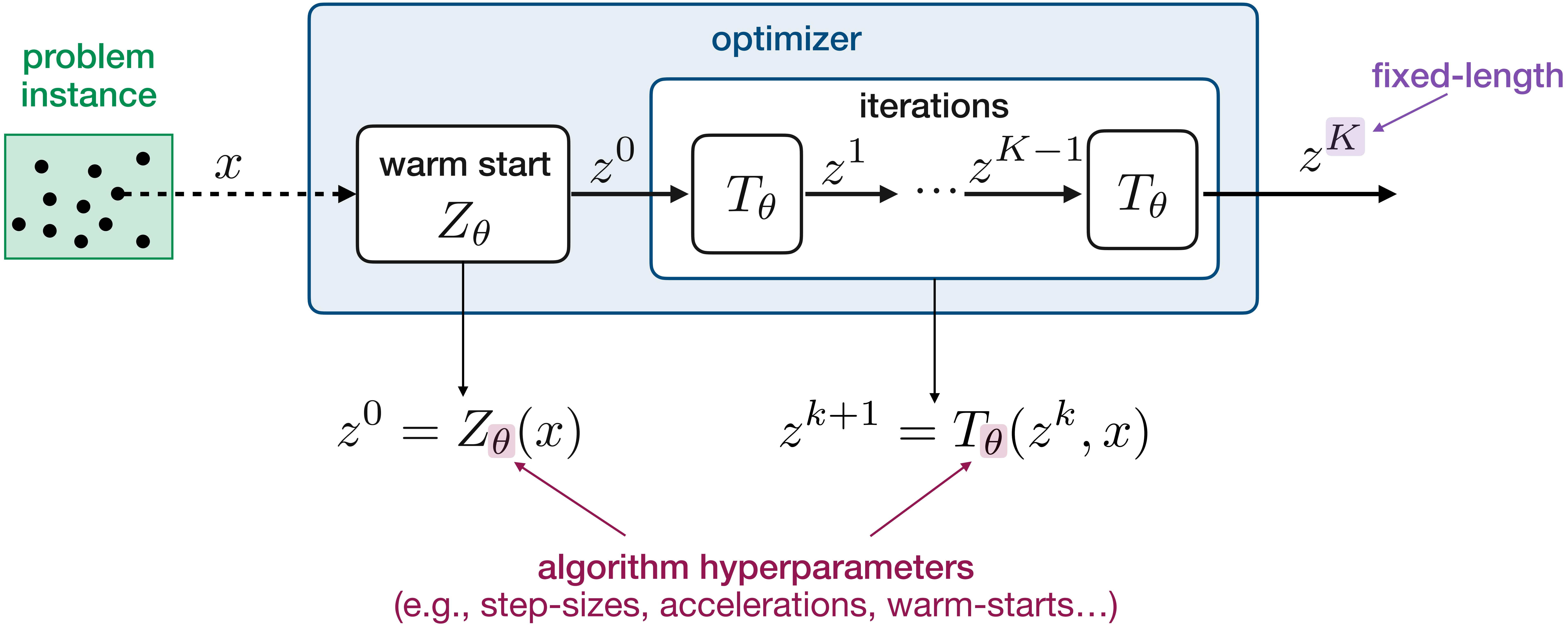
subject to $\quad z \in C(x)$

$\longrightarrow$ same problem with varying parameters

$x \sim \mathbf{P}$

(unknown distribution)

Algorithms as fixed-length computational graphs



example

projected gradient descent

$$z^{k+1} = \Pi_{C(x)}(z^k - \theta \nabla_z f(z^k, x))$$

Verifying the algorithm performance after K iterations

goal

estimate norm of fixed-point residual

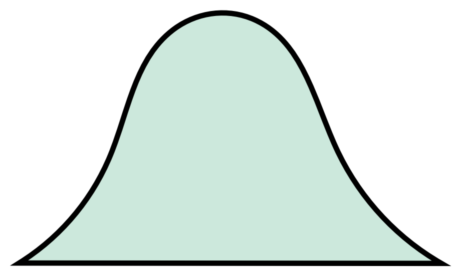
$$r^K(x) = \|z^K - z^{K-1}\|$$

worst-case

$$\max_{x \in \mathcal{X}} r^K(x) \leq \epsilon$$

problem instances → $x \in \mathcal{X}$

convergence tolerance → ϵ



probabilistic

$$\mathbf{P}\left(r^K(x) > \epsilon\right) \leq \eta$$

problem instances → $\mathbf{P}$

convergence tolerance → ϵ

probability bound → η

Verification via mixed-integer linear programming

Worst-case algorithm verification

parametric quadratic optimization

$$\begin{array}{ll} \text{minimize} & (1/2)z^T P z + q(x)^T z \\ \text{subject to} & Az \leq b(x) \end{array}$$

problem instances

algorithm

(ADMM, PDHG,…)

$$z^{k+1} = T_\theta(z^k, x)$$

verification problem

performance metric

$$\max_{x \in \mathcal{X}} r^K(x) = \begin{array}{ll} \text{maximize} & \|z^K - z^{K-1}\|_\infty \\ \text{subject to} & z^{k+1} = T_\theta(z^k, x), \quad k = 0, \dots, K-1 \\ & z^0 = Z_\theta(x), \quad x \in \mathcal{X} \end{array}$$

problem instances

NP-hard problem!

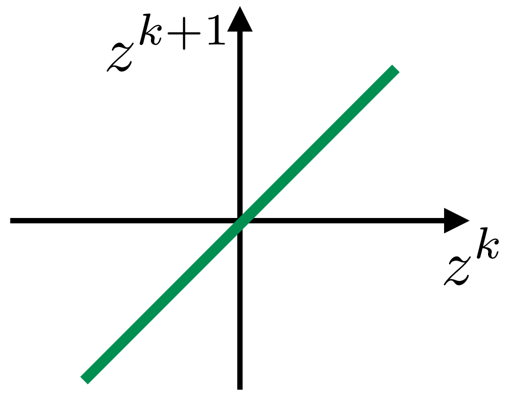
Verification of First-Order Methods for Parametric Quadratic Optimization
V. Ranjan and B. Stellato
arXiv e-prints:2403.03331 (2025)
github.com/stellatogrp/sdp_algo_verify

Algorithm steps as mixed-integer linear constraints

Linear steps → Linear constraints

e.g., gradient, momentum, restarts, anchors, prox of quadratic functions…

$$Mz^{k+1} = Az^k + Bx$$

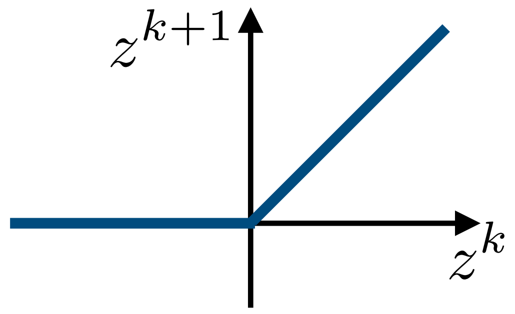


Piecewise affine steps → Mixed-integer constraints

Elementwise maximum (ReLU)

e.g., one-sided projections

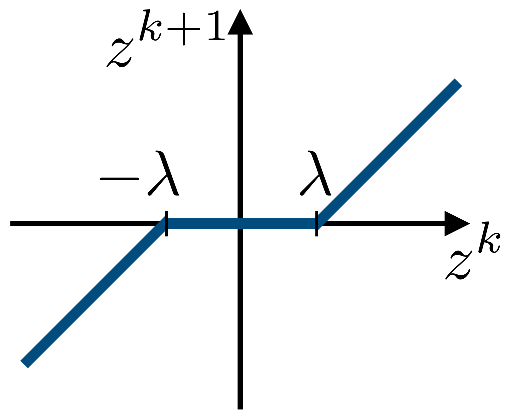
$$z^{k+1} = (z^k)_+ = \max\{z^k, 0\}$$



Soft-thresholding

e.g., prox of 1-norm function

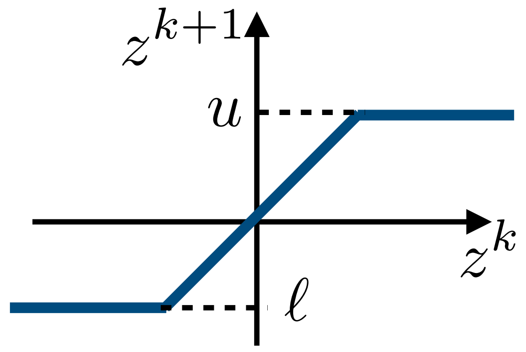
$$z^{k+1} = \phi_\lambda(z^k) = \max\{z^k, \lambda\} - \max\{-z^k, \lambda\}$$



Saturated linear unit (SatLin)

e.g., box projections

$$z^{k+1} = \mathcal{S}_{\ell,u}(z^k) = \min\{\max\{z^k, \ell\}, u\}$$



Example:
Nonnegative least squares

minimize $\quad (1/2)\|Dz - x\|_2^2$

subject to $\quad z \geq 0$

gradient step

projection step

Projected Gradient Descent

$$w^{k+1} = (I - \theta D^T D)z^k + \theta D^T x$$

$$z^{k+1} = \max\{w^{k+1}, 0\}$$

similar MIP constraints in neural network verification

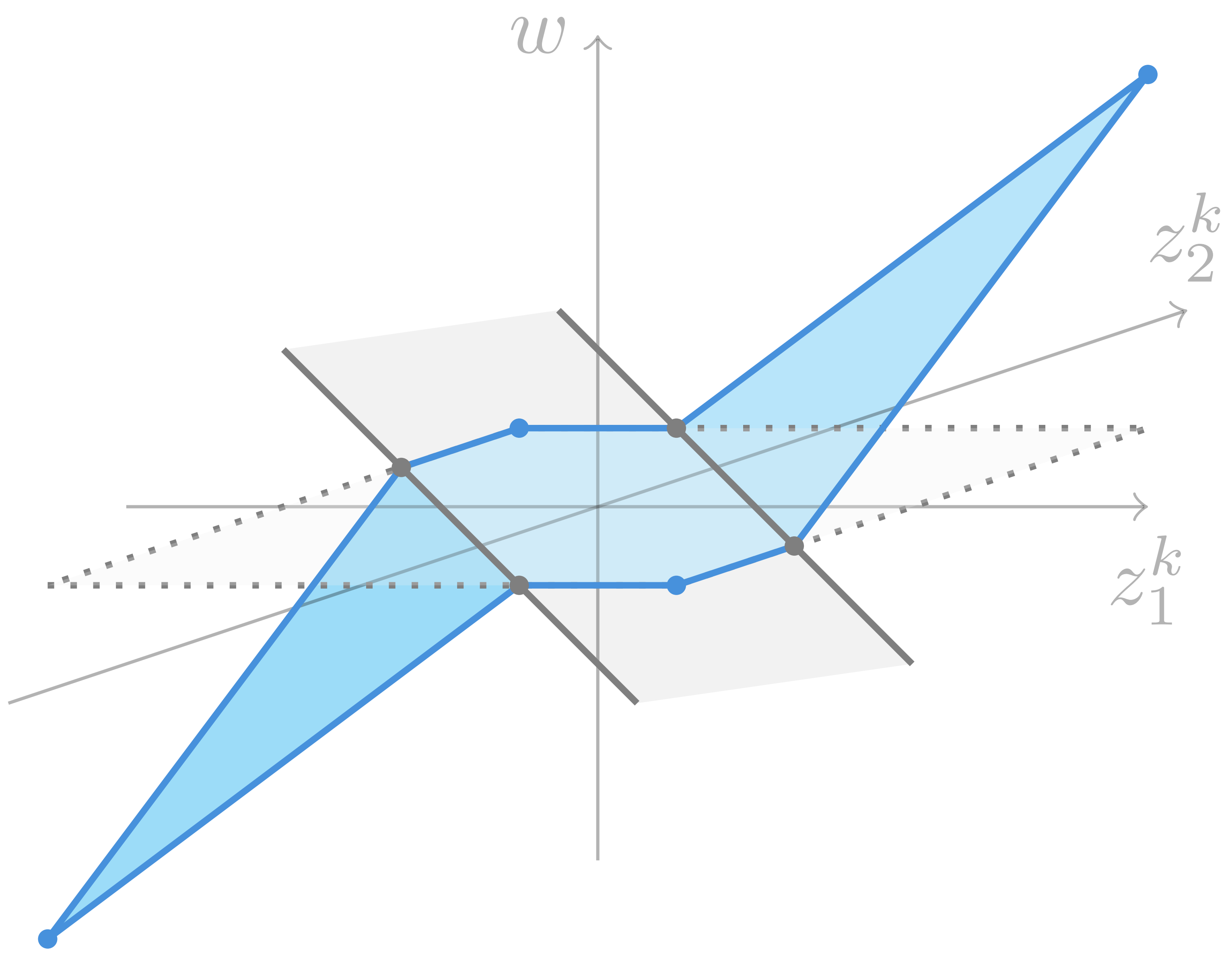
Liu et al. (2021), Albarghouthi (2021), Ceccon et al. (2022), Fischetti and Jo (2018), Tjeng et al. (2019)

Constructing strong MIP formulations

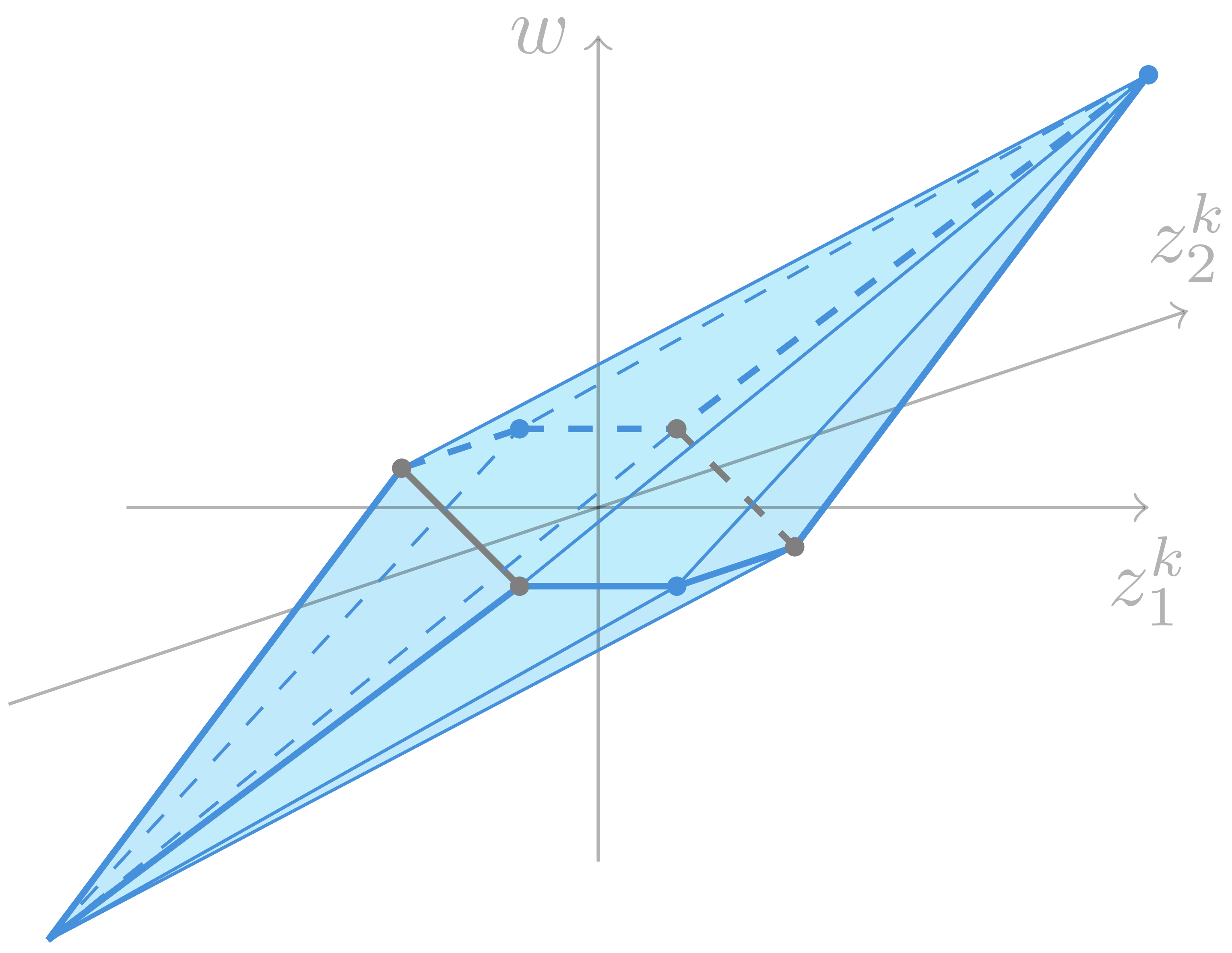
piecewise affine steps

soft-thresholding operator

$$w = \phi_\gamma(z_1^k + z_2^k)$$



convex hull



Inspired by: Anderson et al. (2020), Tjandraatmadja et al. (2020), Tsay et al. (2021), Hojny et al. (2024), Huchette et al. (2025)

exponential number of inequalities!

✓ separation problem solvable in linear time

Using operator theory to tighten MIP formulations

$$\max_{x \in \mathcal{X}} r^K(x) = \begin{array}{ll} \text{maximize} & \|z^K - z^{K-1}\|_\infty \\ \text{subject to} & z^{k+1} = T_\theta(z^k, x), \quad k = 0, \dots, K-1 \\ & z^0 = Z_\theta(x), \quad x \in \mathcal{X} \end{array}$$

← **performance metric**

operator theory bound

$$\|z^K - z^{K-1}\|_\infty \leq \alpha_K$$

e.g., linear convergence

$$\alpha_K = C\tau^K$$

← *rate*

previous iterate bounds
(bound tightening, interval propagation, etc)

$$\underline{z}^{K-1} \leq z^{K-1} \leq \bar{z}^{K-1}$$

combine →

bounds on latest iterate

$$\underline{z}^{K-1} - \alpha_K \leq z^K \leq \bar{z}^{K-1} + \alpha_K$$

main idea
solve verification problem for increasing K and exploit bounds

Examples

Sparse coding for signal reconstruction

$$\text{minimize} \quad (1/2)\|Dz - x\|_2^2 + \lambda\|z\|_1$$

- D: known dictionary
- z: reconstructed signal
- x: noisy signal

Iterative Soft-Thresholding Algorithm (ISTA)

$$z^{k+1} = \phi_{\lambda\theta}\left((I - \theta D^T D)z^k + \theta D^T x\right)$$

$\phi_{\lambda\theta}$: soft-thresholding operator

Fast Iterative Soft-Thresholding Algorithm (FISTA)

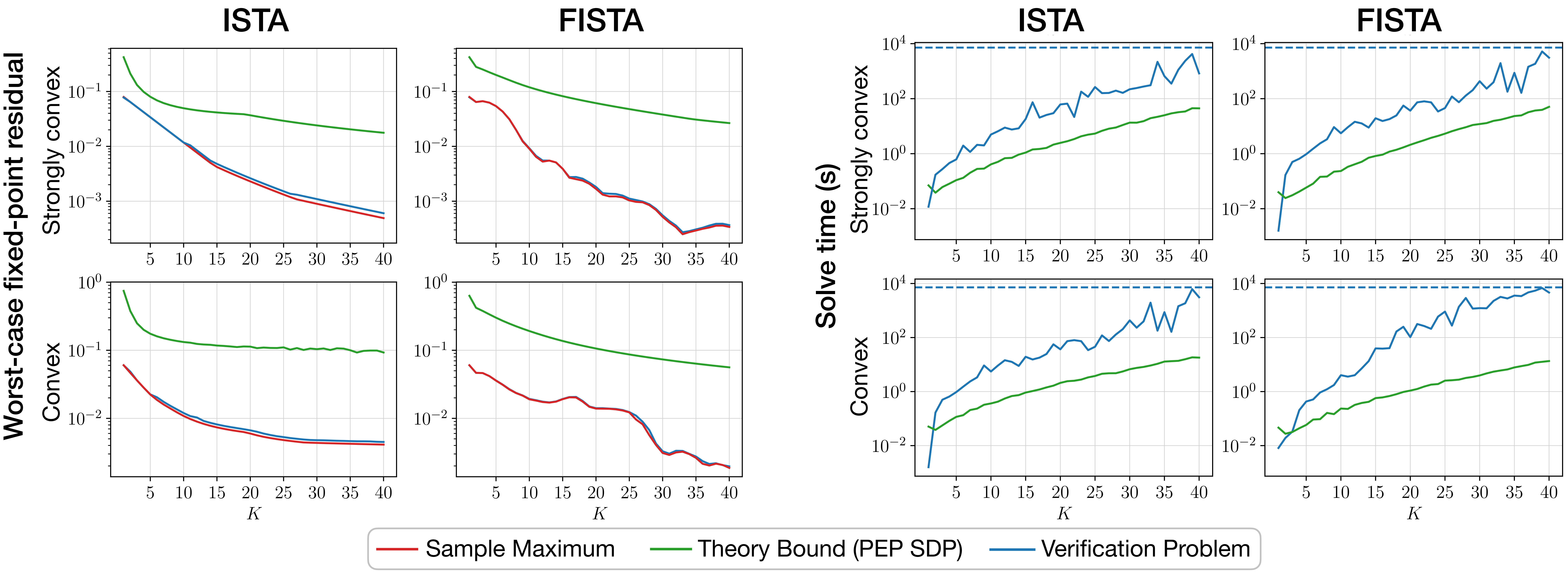
$$w^{k+1} = \phi_{\lambda\theta}\left((I - \theta D^T D)z^k + \theta D^T x\right)$$

$$z^{k+1} = w^{k+1} + (\beta_k - 1)/\beta_{k+1}(w^{k+1} - w^k)$$

$\phi_{\lambda\theta}$: soft-thresholding operator

$(\beta_k - 1)/\beta_{k+1}$: momentum

Verification results for sparse coding example



10x-100x reduction in worst-case fixed-point residual *(exploiting parametric structure)*

exactly captures the ripples of the FISTA acceleration

Network flow optimization

network flow

$$\begin{aligned} \text{minimize} \quad & c^T z \\ \text{subject to} \quad & A_s z \le b_s \\ & A_d z = x \\ & 0 \le z \le u \end{aligned}$$

arc-node matrices (supply and demand)

supplies

demands

Primal-dual hybrid gradient (PDHG)

$$\begin{aligned} z^{k+1} &= \mathcal{S}_{[0,u]}(z^k - \eta(c + A_s^T v^k - A_d^T w^k)) \\ v^{k+1} &= (v^k + \eta(-b_s + A_s(2z^{k+1} - z^k)))_+ \\ w^{k+1} &= w^k + \eta(x - A_d(2z^{k+1} - z^k)) \end{aligned}$$

saturated linear unit

one-sided projection

Primal-dual hybrid gradient with momentum (mPDHG)

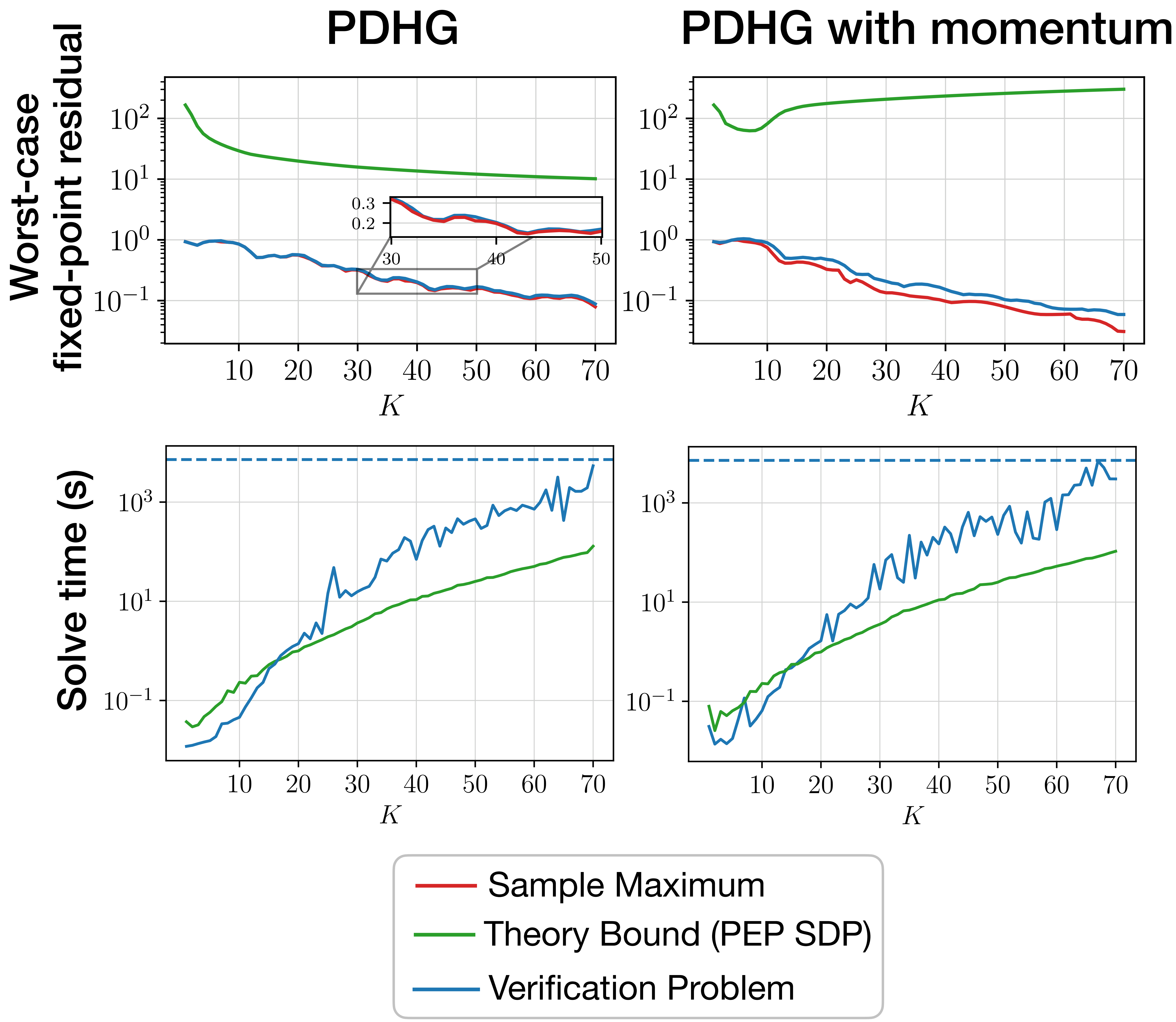
$$\begin{aligned} z^{k+1} &= \mathcal{S}_{[0,u]}(z^k - \eta(c + A_s^T v^k - A_d^T w^k)) \\ \tilde{z}^{k+1} &= z^k + k/(k+3)(z^{k+1} - z^k) \\ v^{k+1} &= (v^k + \eta(-b_s + A_s(2\tilde{z}^{k+1} - z^k)))_+ \\ w^{k+1} &= w^k + \eta(x - A_d(2\tilde{z}^{k+1} - z^k)) \end{aligned}$$

saturated linear unit

momentum

one-sided projection

Verification results for network flow optimization example



we verify convergence even when PEP doesn't!

known behavior in momentum/heavy-ball methods
[L. Lessard, B. Recht, and A. Packard, (2016)]
[Goujaud, Taylor, Dieuleveut (2023)]

can be faster than SDP for few iterations

Optimal control

optimal sequence of states and controls

$$
\begin{array}{lll}
\text{minimize} & \sum_{t=0}^{T} s_t^T Q s_t + u_t^T R u_t & \\
\text{subject to} & s_{t+1} = A^{\text{dyn}} s_t + B^{\text{dyn}} u_t, & t = 1, \dots, T-1 \\
& s_{\min} \le s_t \le s_{\max}, & t = 1, \dots, T \\
& u_{\min} \le u_t \le u_{\max}, & t = 1, \dots, T \\
& s_0 = x &
\end{array}
$$

linear dynamics

initial state

OSQP ADMM splitting

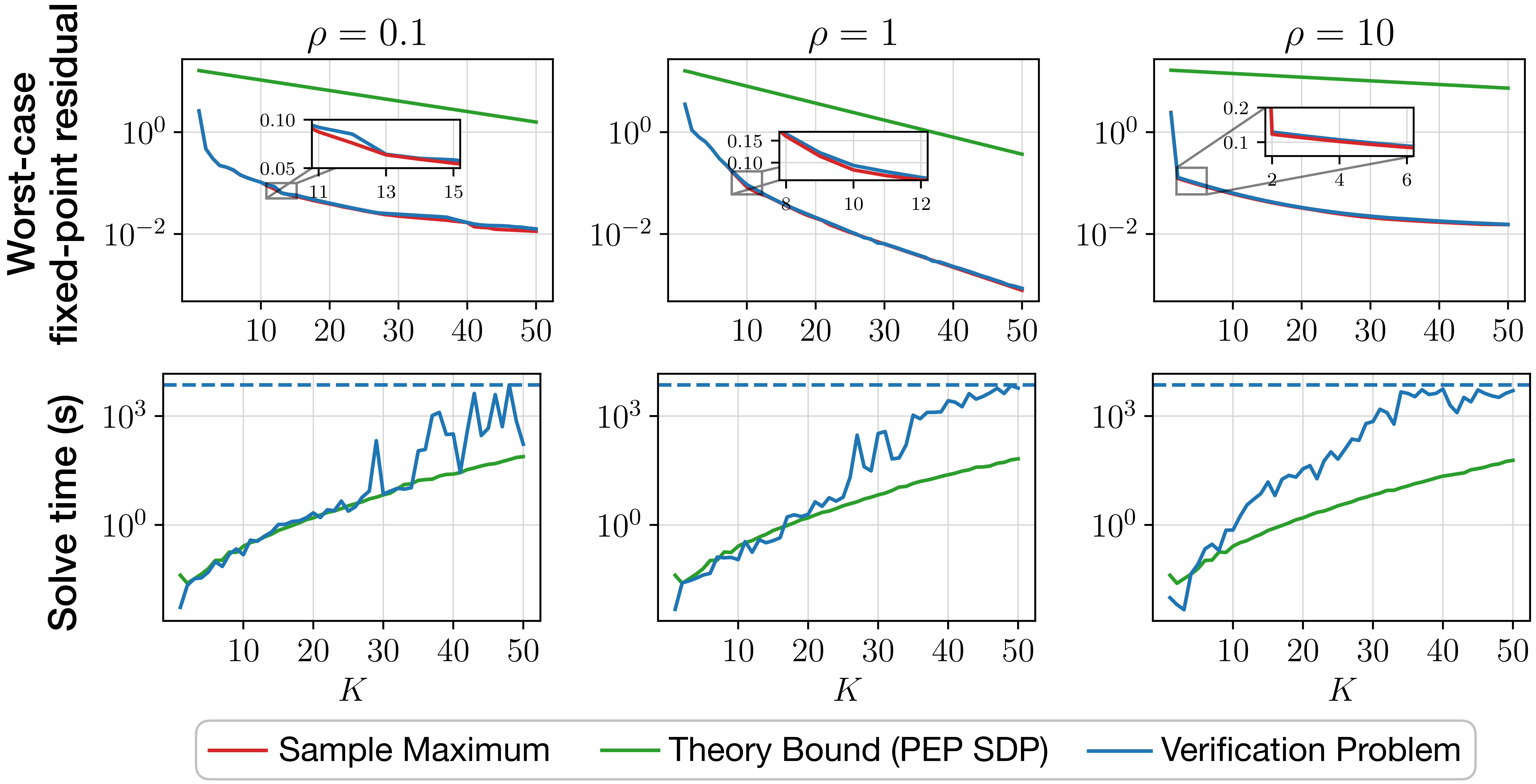
$$w^{k+1} = \mathcal{S}_{[l(x),u(x)]}(v^k)$$

Solve $(P + \sigma I + \rho M^T M) z^{k+1} = \sigma z^k - q(x) + \rho M^T (2w^{k+1} - v^k)$

$$v^{k+1} = v^k + M z^{k+1} - w^{k+1}$$

saturated linear unit

Verification results for optimal control problem

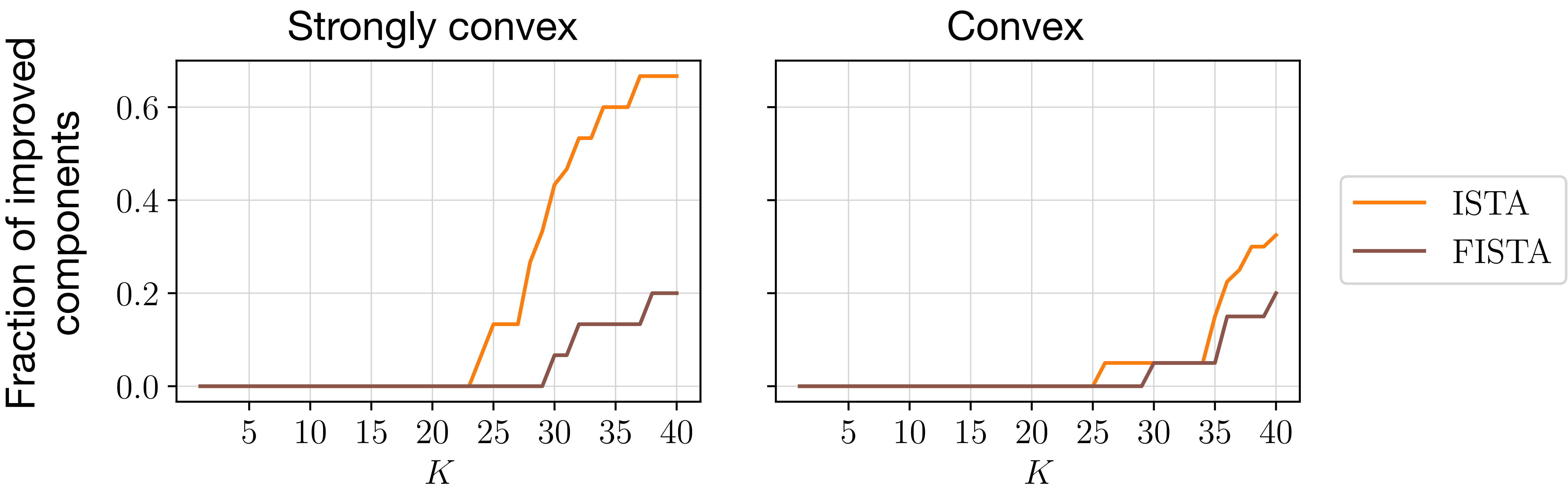


can be used to design (tune) custom algorithms!

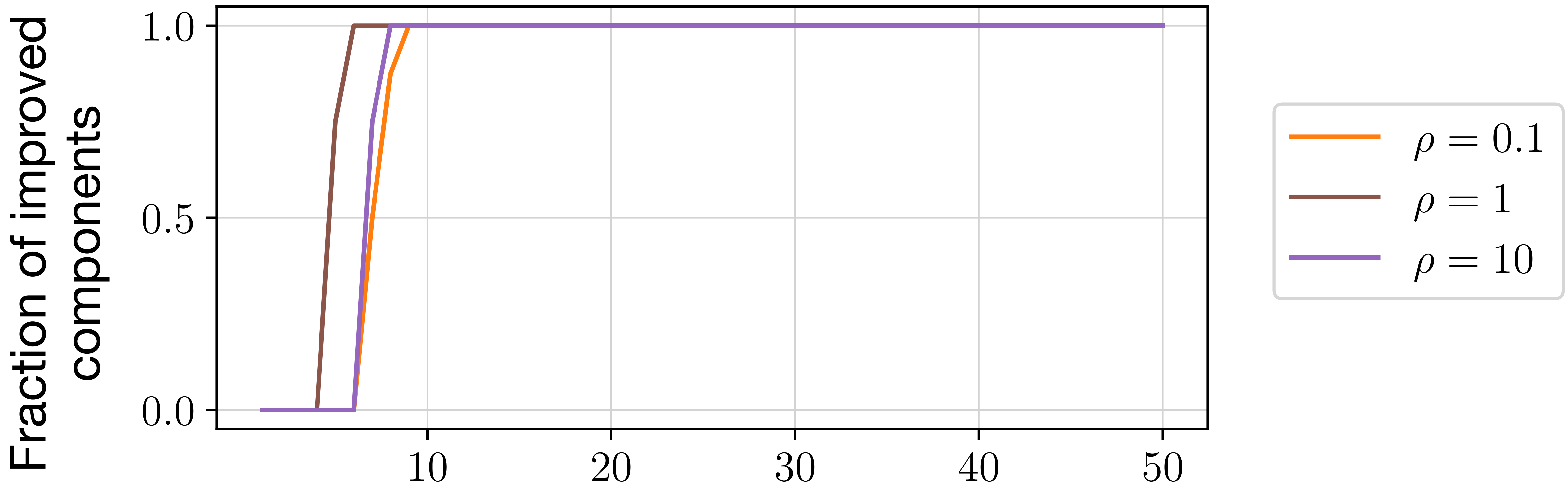
exactly quantifies the iterations required → crucial in real-time settings!

Operator theory can tighten bounds on the iterates

Sparse coding



Optimal control



strongly convex problems have the largest benefits
(because of linear convergence)

operator theory bounds did not help PDHG
(usually requires thousands of iterations)

Acknowledgements

Vinit Ranjan
Princeton

Jisun Park
Princeton

Andrea Lodi
CornellTech

Stefano Gualandi
Univ. of Pavia

Verification of First-order Methods via Mixed-Integer Linear Programming

1. **parametric** structure matters
2. **algorithm verification** 🤝 **operator theory**
3. useful to **design new algorithms**

traditional view

- general-purpose
- one-size-fits all

→

new view

- task-specific
- tunable/trainable
- deployable anywhere

Exact Verification of First-Order Methods via Mixed-Integer Linear Programming
V. Ranjan, J. Park, S. Gualandi, A. Lodi, and B. Stellato
arXiv e-prints:2412.11330 (2025)
github.com/stellatogrp/mip_algo_verify

Verification of First-Order Methods for Parametric Quadratic Optimization
V. Ranjan and B. Stellato
arXiv e-prints:2403.03331 (2025)
github.com/stellatogrp/sdp_algo_verify

@stellato.io bstellato@princeton.edu stellato.io

Backup

Unconstrained QP

$$\text{minimize} \quad (1/2)z^T P z + x^T z$$

parameters

verification problem

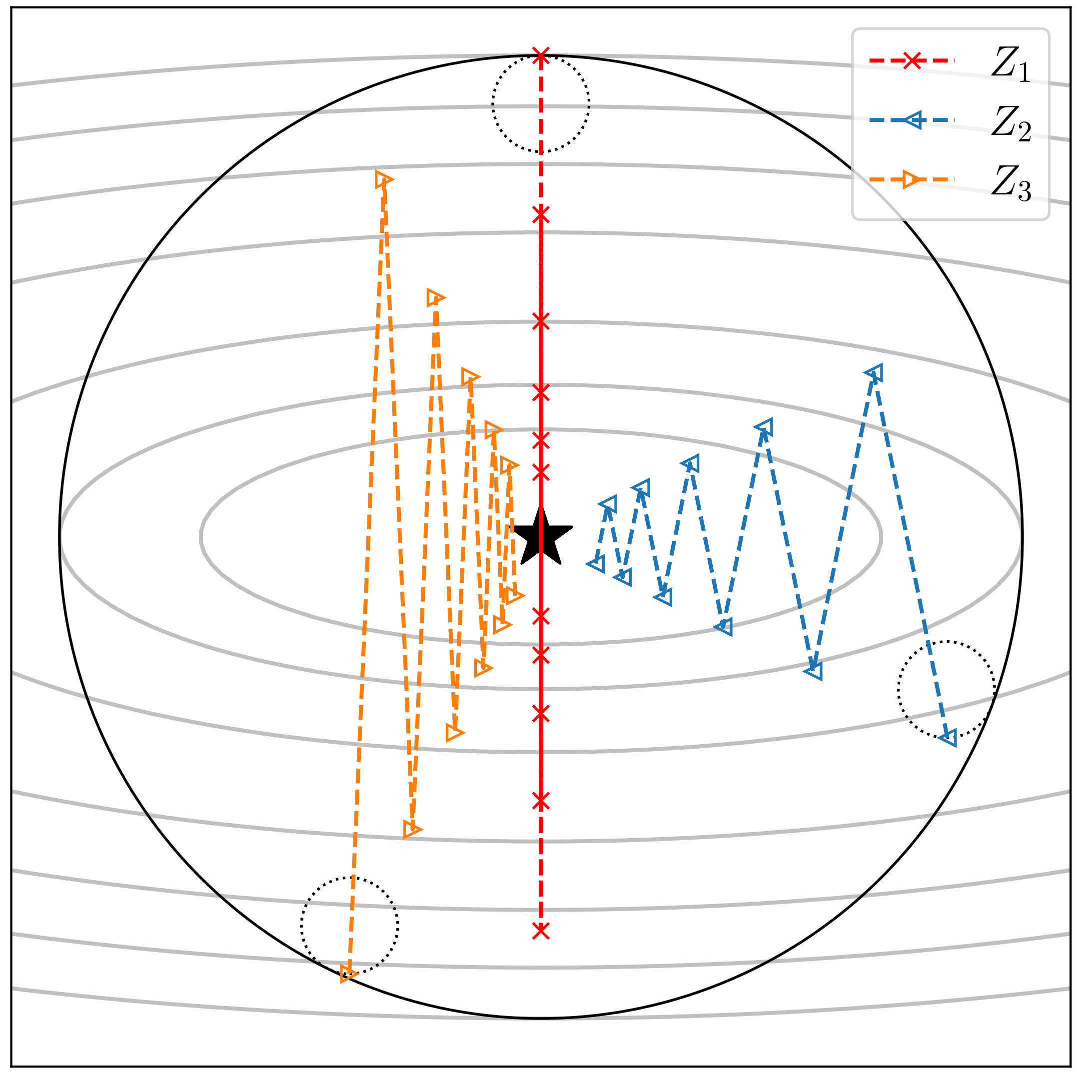
$$\begin{array}{ll} \text{maximize} & \|z^K - z^{K-1}\| \\ \text{subject to} & z^{k+1} = z^k - \theta(Pz^k + x), \quad k = 0, \dots, K-1 \\ & z^0 = Z_\theta(x), \quad x \in \mathcal{X} \end{array}$$

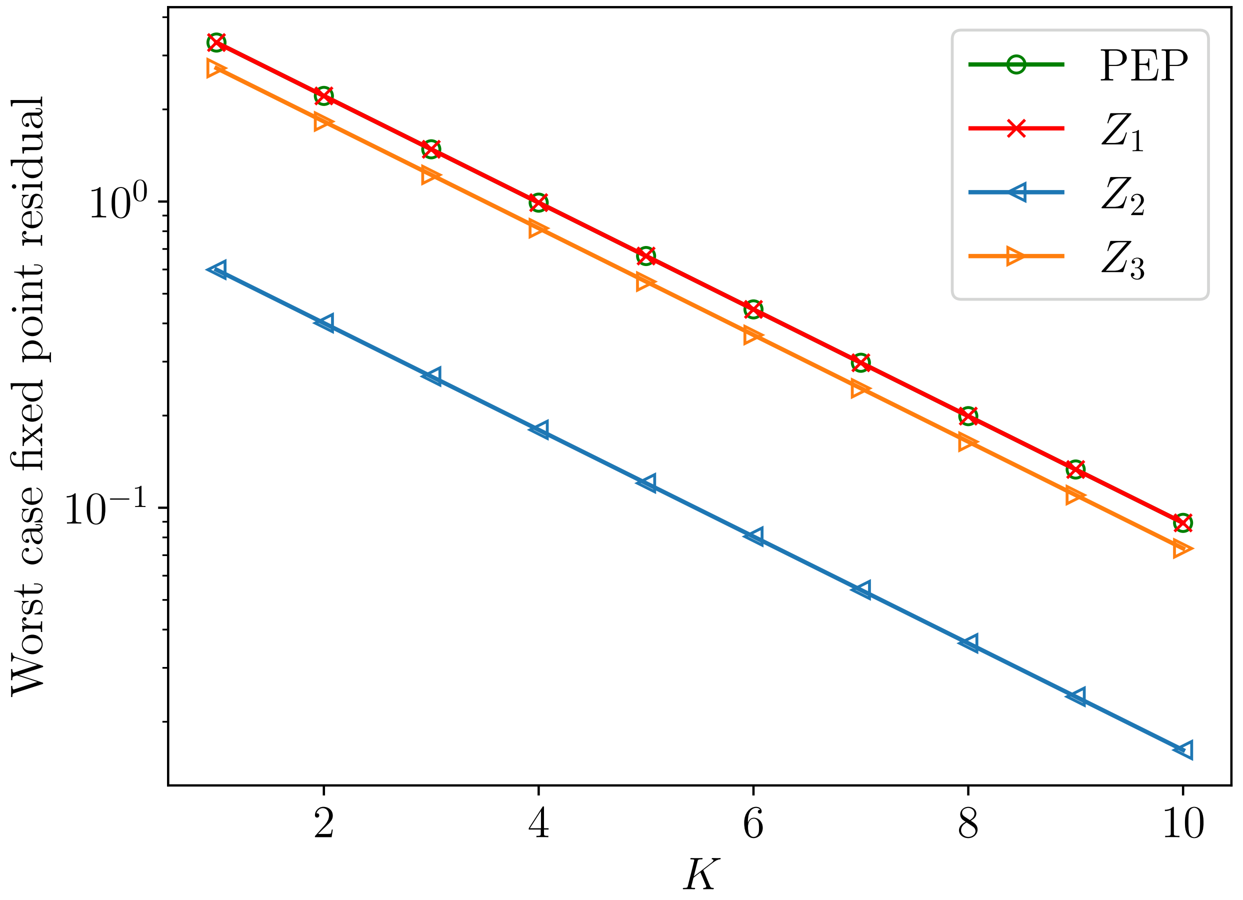
gradient descent

warm-starts

case I $Z_\theta(x) = Z_1, Z_2,$ or Z_3

$x \in \mathcal{X} = \{0\}$





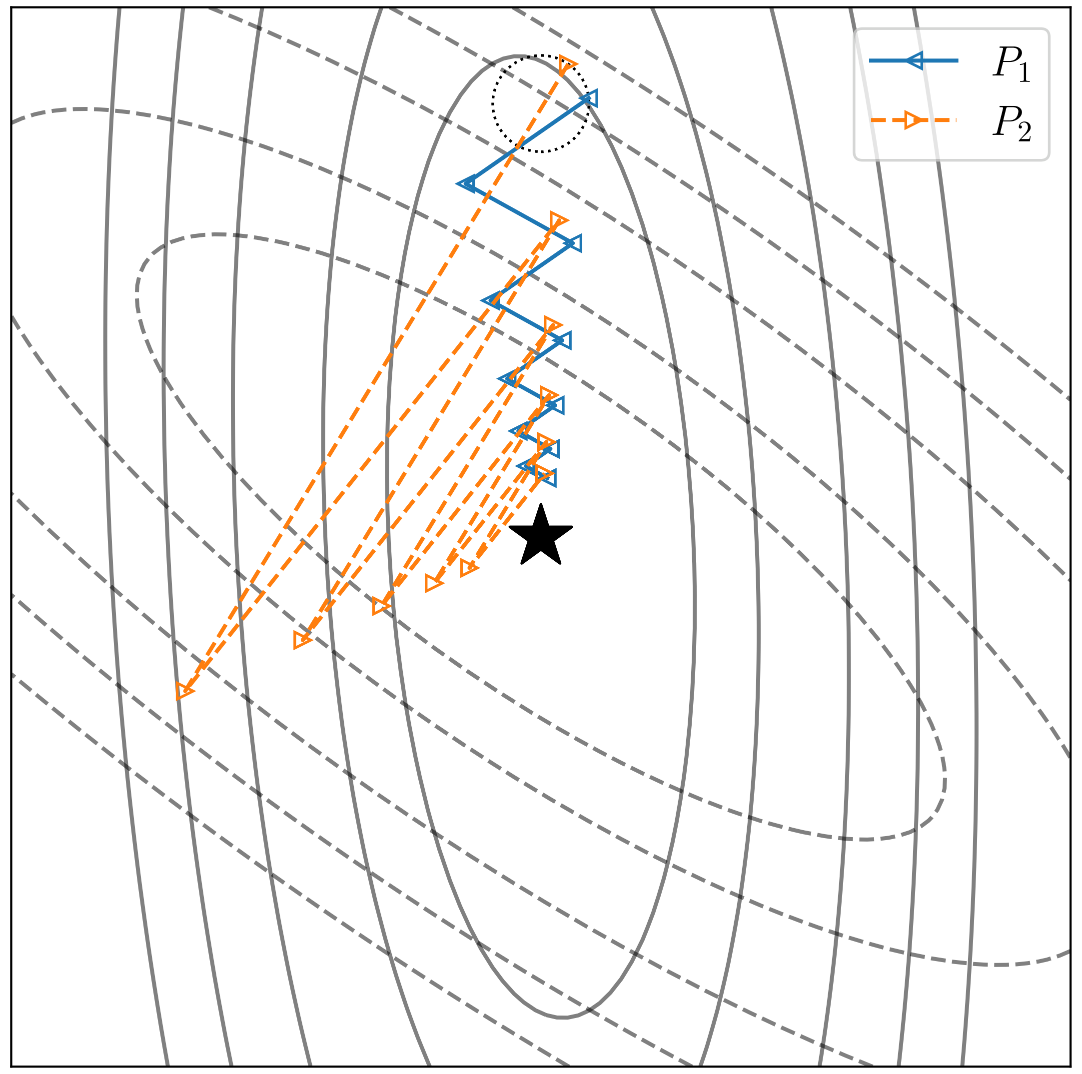
PEP-SDP cannot distinguish warm-starts

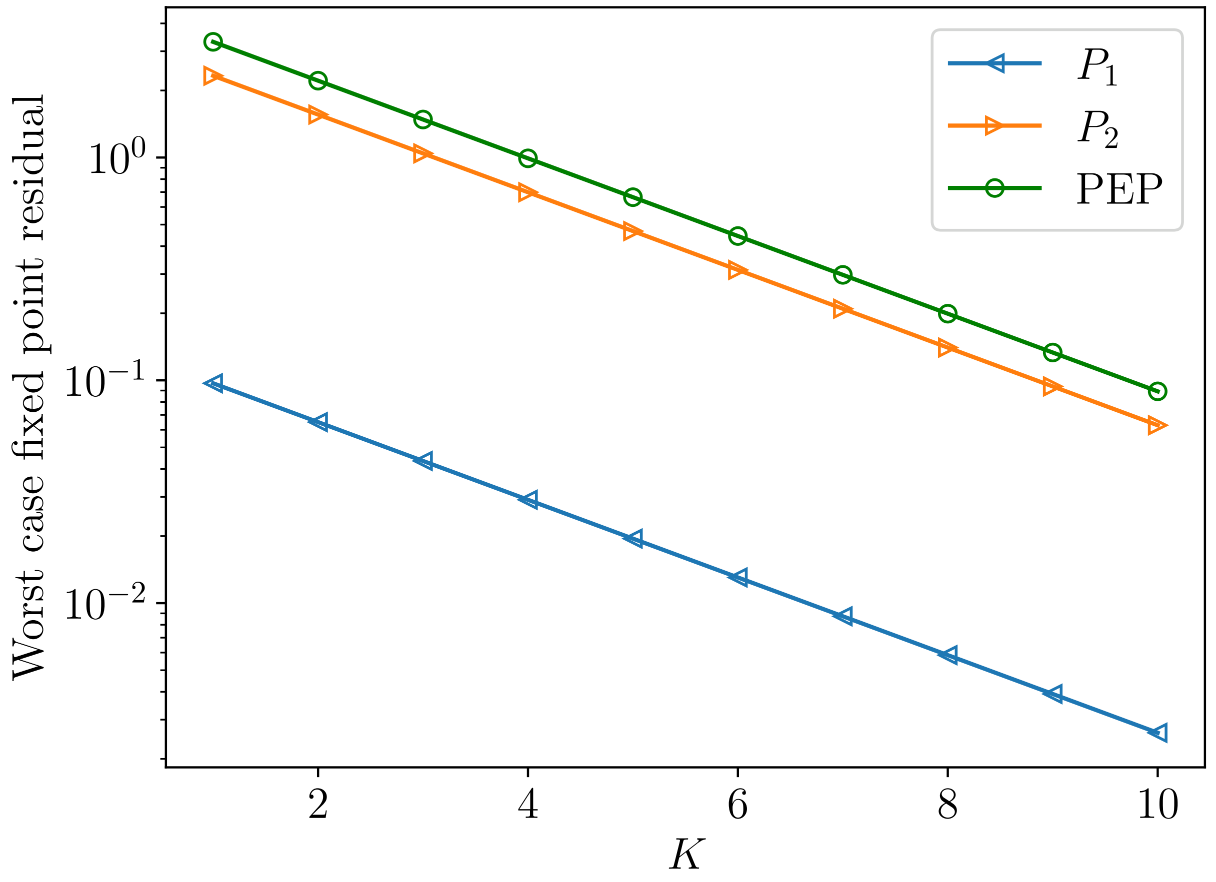
rotated functions

case II $Z_\theta(x) = \{z \mid \|z - 0.9 \cdot \mathbf{1}\| \leq 0.1\}$

$x \in \mathcal{X} = \{0\}$

P_1, P_2 rotations of P





PEP-SDP cannot distinguish quadratic functions

Objective of verification problem as MIP

$$\|s^K - s^{K-1}\|_\infty = \|t\|_\infty = \delta_K$$

- lower bounds $\underline{s}^{K-1}, \underline{s}^K$
- upper bounds, $\bar{s}^{K-1}$ and $\bar{s}^K$

→

- lower bound $\underline{t} = \underline{s}^K - \bar{s}^{K-1}$
- upper bound $\bar{t} = \bar{s}^K - \underline{s}^{K-1}$

exact reformulation

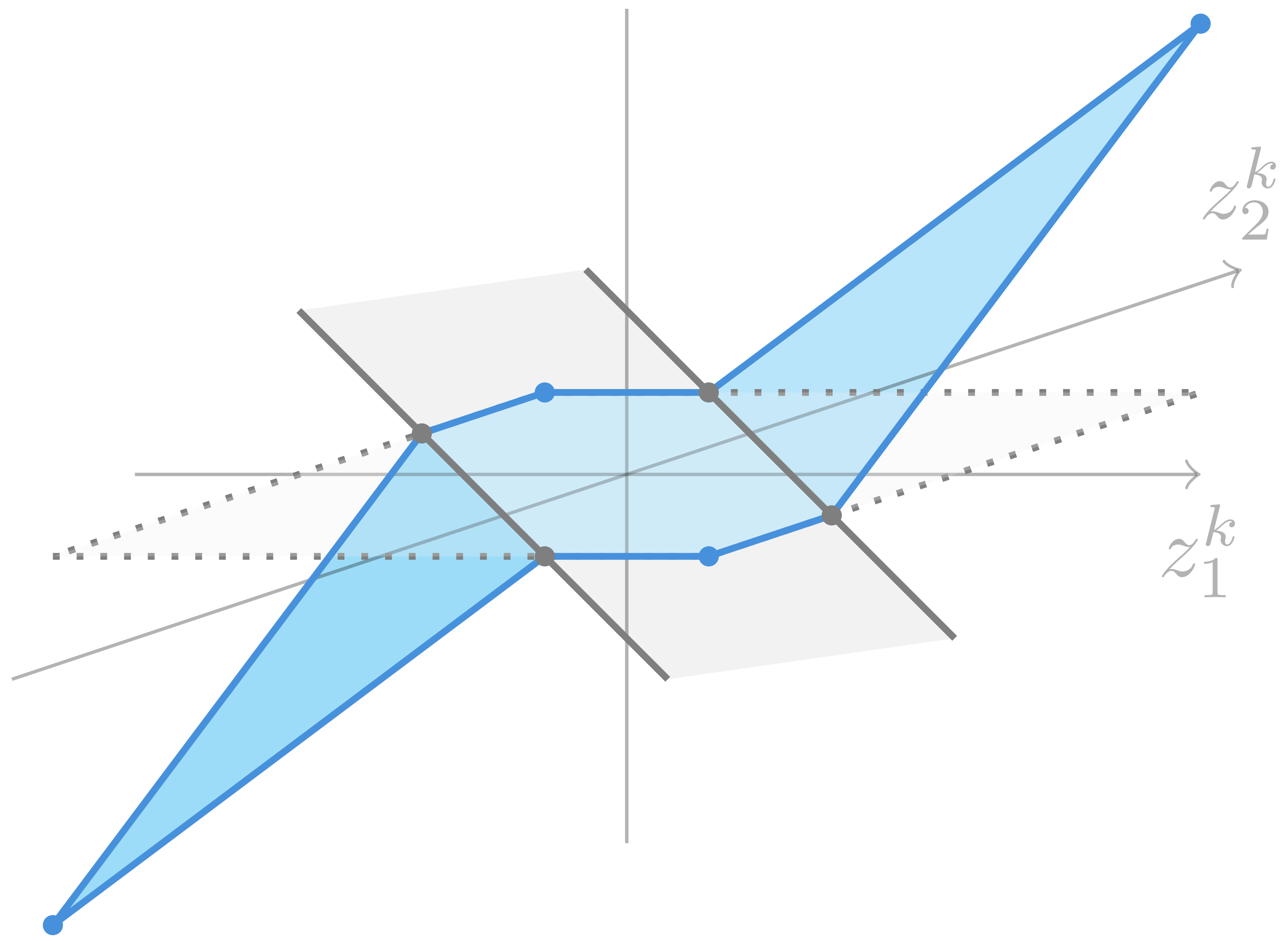
$$t = t^+ - t^-, \quad t^+ \leq \bar{t} \odot w, \quad t^- \leq -\underline{t} \odot (\mathbf{1} - w)$$
$$t^+ + t^- \leq \delta_K \leq t^+ + t^- + \max\{\bar{t}, -\underline{t}\} \odot (\mathbf{1} - \gamma)$$
$$\mathbf{1}^T \gamma = 1, \quad t^+ \geq 0, \quad t^- \geq 0$$

$w \in \{0,1\}^n$ (absolute values of the components of t)
$\gamma \in \{0,1\}^d$ (maximum inside the ℓ_∞-norm)

Soft-thresholding operator

$$w = \phi_\lambda(a^T z) = \begin{cases} a^T z - \lambda & a^T z > \lambda \\ 0 & |a^T z| \le \lambda \\ a^T z + \lambda & a^T z < -\lambda \end{cases}$$

example: $\phi_\lambda(z_1^k + z_2^k)$



region

$$\Phi = \{(z, w) \in [\underline{z}, \bar{z}] \times \mathbf{R} \mid w = \phi_\lambda(a^T z)\}$$

lower and upper bounds
(needed for convex hull)

$$\ell_I = \sum_{i \in I} a_i \ell_i^0 + \sum_{i \notin I} a_i u_i^0$$

$$u_I = \sum_{i \in I} a_i u_i^0 + \sum_{i \notin I} a_i \ell_i^0$$

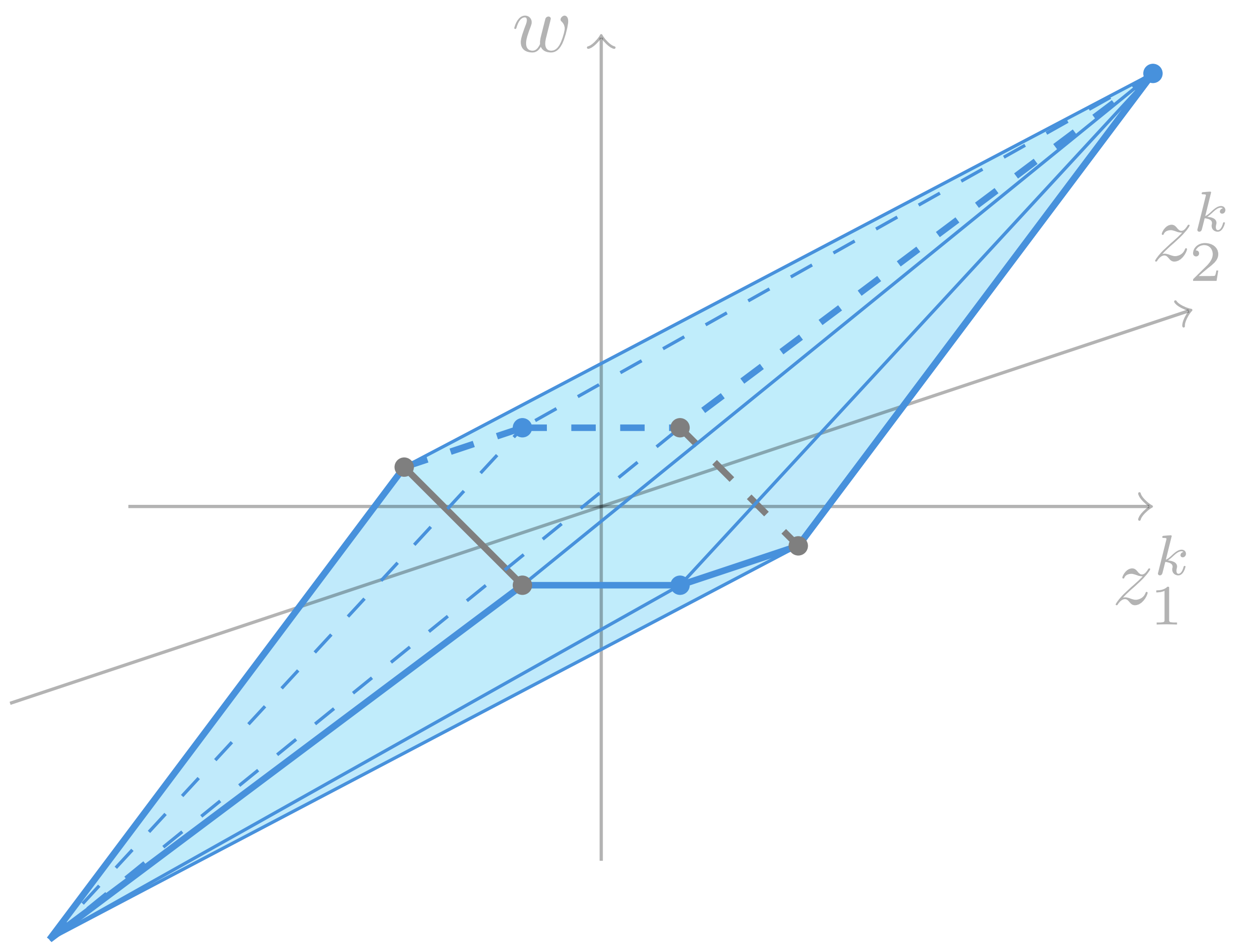
$$u_i^0 = \begin{cases} \bar{z}_i & a_i \ge 0 \\ \underline{z}_i & \text{otherwise} \end{cases} \qquad \ell_i^0 = \begin{cases} \underline{z}_i & a_i \ge 0 \\ \bar{z}_i & \text{otherwise} \end{cases}$$

Convex hull of soft-thresholding operator

convex hull

$$\mathbf{conv}(\Phi) = \left\{ (z,w) \in [\underline{z}, \bar{z}] \times \mathbf{R} \;\middle|\; \begin{cases} w = a^T z - \lambda & \ell_{\{1,\ldots,n\}} > \lambda \\ w = a^T z + \lambda & u_{\{1,\ldots,n\}} < -\lambda \\ w = 0 & -\lambda \le \ell_{\{1,\ldots,n\}} \le u_{\{1,\ldots,n\}} \le \lambda \\ (z,w) \in Q & \text{otherwise} \end{cases} \right\}$$

example: $\phi_\lambda(z_1^k + z_2^k)$



$$Q = \left\{ \begin{array}{l} a^T z - \lambda \le w \le a^T z + \lambda \\ \frac{\ell_J + \lambda}{\ell_J - \lambda}(a^T z - \lambda) \le w \le \frac{u_J - \lambda}{u_J + \lambda}(a^T z + \lambda) \\ w \le \sum_{i \in I} a_i (z_i - \ell_i^0) + \frac{\ell_I - \lambda}{u_o^0 - \ell_o^0}(z_o - \ell_o^0), \ \forall (I,o) \in \mathcal{I}^+ \\ w \ge \sum_{i \in I} a_i (z_i - u_i^0) + \frac{u_I + \lambda}{\ell_o^0 - u_o^0}(z_o - u_o^0), \ \forall (I,o) \in \mathcal{I}^- \end{array} \right\}$$

separation problem can be solved in linear time (by sorting)

exponential number of inequalities

$$\mathcal{I}^- = \{(I,o) \in 2^{\{1,\ldots,n\}} \times \{1,\ldots,n\} \mid u_I \le -\lambda < u_{I \cup \{o\}}, \ w_I \ne 0\}$$

$$\mathcal{I}^+ = \{(I,o) \in 2^{\{1,\ldots,n\}} \times \{1,\ldots,n\} \mid \ell_{I \cup \{o\}} < \lambda \le \ell_I, \ w_I \ne 0\}$$