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

Bartolomeo Stellato

PRINCETON
UNIVERSITY

Vinit Ranjan
Princeton

Jisun Park
Princeton

Andrea Lodi
CornellTech

Stefano Gualandi
Univ. of Pavia

Discrete Optimization and Machine Learning (DOxML) — May 15 2025

Most applications require fast and effective decisions in real-time

We use first-order methods…

decisions

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

parameters

same problem with varying parameters

$$x \sim \mathbf{P}$$

(unknown distribution)

example
projected gradient descent

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

projection

gradient step

benefits of first-order methods

✓ cheap iterations

✓ easy to warm-start

embedded optimization

large-scale optimization

many general-purpose solvers available today

PDLP

SCS

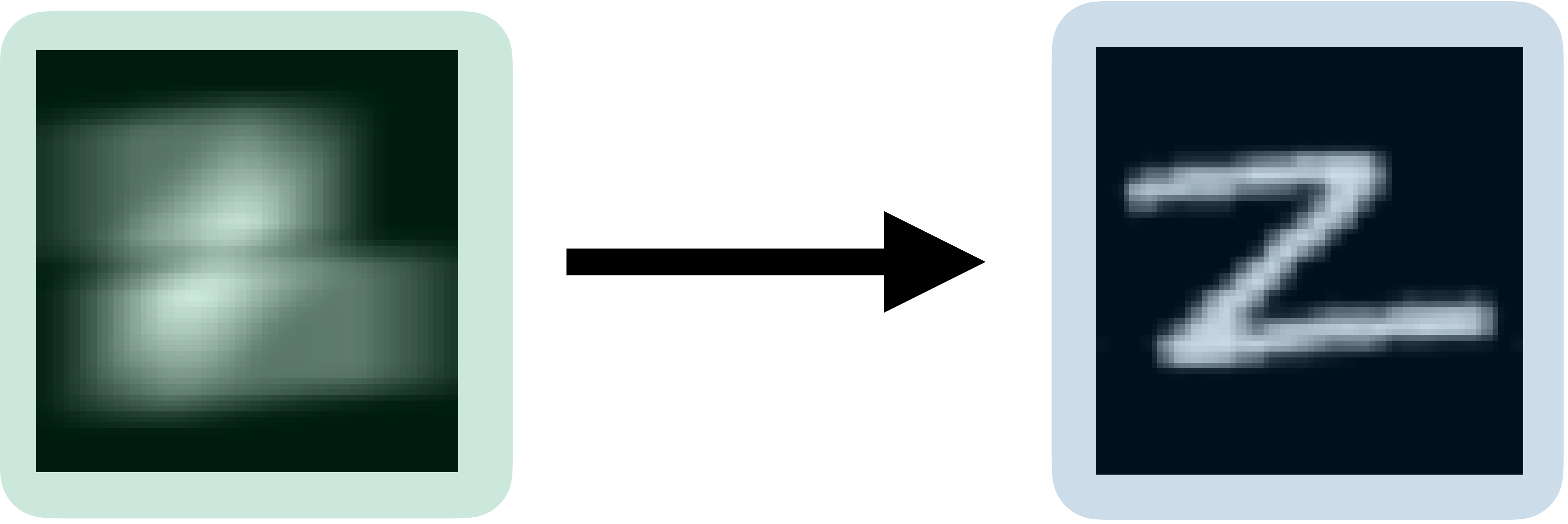
OSQP

SDPNAL+, cuOPT, …

…but we still do not fully understand their convergence!

image deblurring problem

emnist dataset

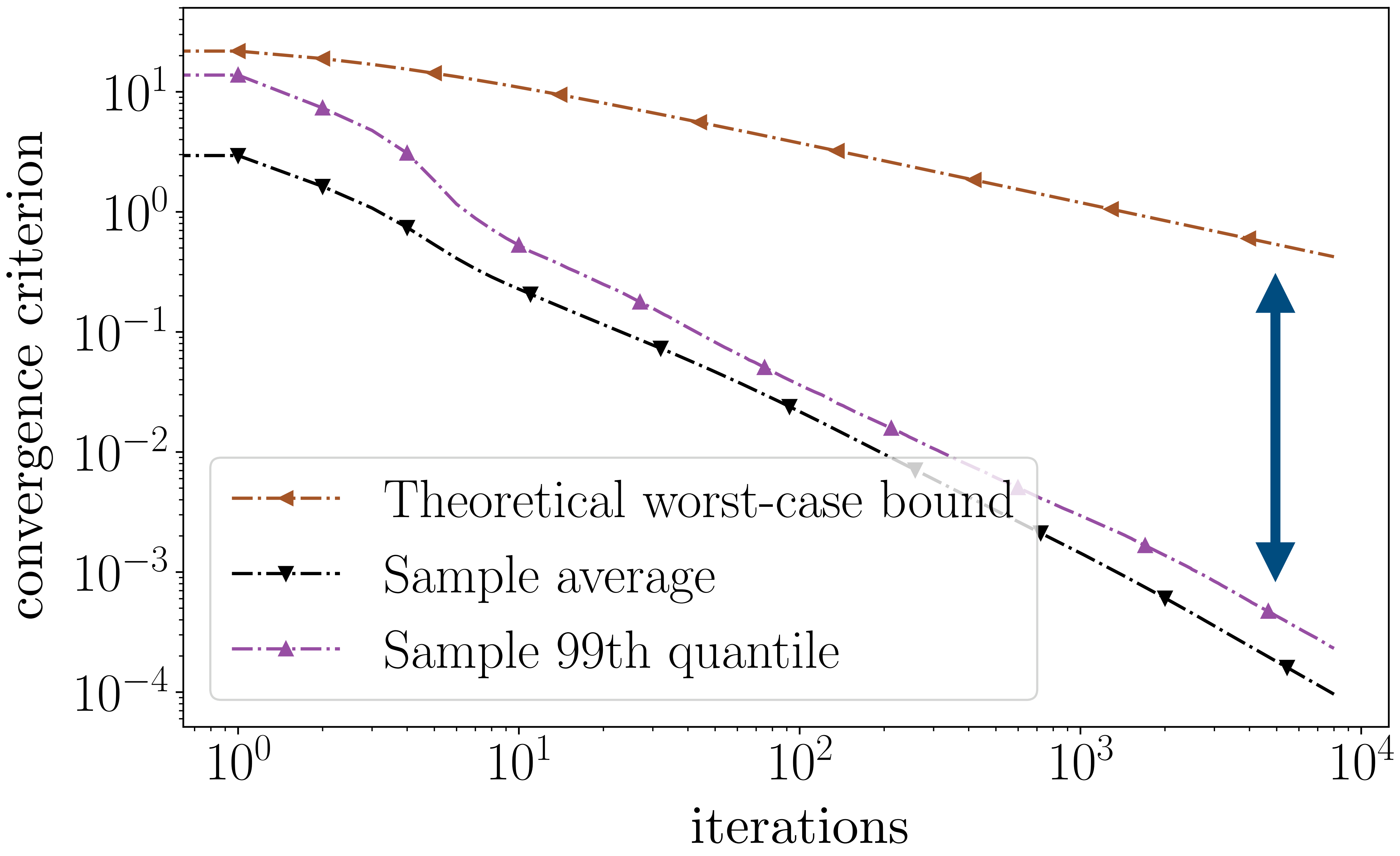


minimize $\|Az - x\|_2^2 + \lambda\|z\|_1$

subject to $0 \le z \le 1$

blurred image

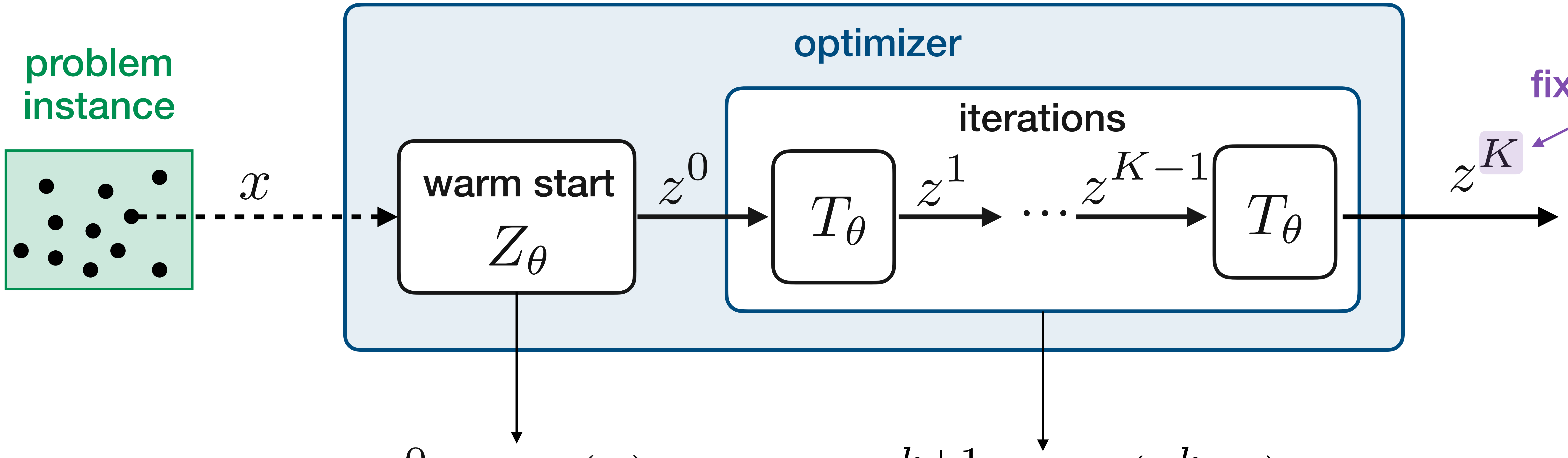
deblurred image



How can we compute tighter bounds?

Theoretical bounds from Performance Estimation Problem (PEP). Drori and Teboulle (2014), Taylor, Hendrickx, Gilneur (2017), and many others…

Algorithms as fixed-length computational graphs



problem instance

optimizer

iterations

fixed-length

x

warm start Z_θ

z^0 T_θ z^1 $\cdots$ z^{K-1} T_θ z^K

$$z^0 = Z_{\theta}(x) \qquad z^{k+1} = T_{\theta}(z^k, x)$$

algorithm hyperparameters
(e.g., step-sizes, accelerations, warm-starts…)

goal: find
fixed-points ↔ optimal solutions

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

performance metric

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

$\uparrow$
fixed-point residual
(converges to 0)

Verifying the algorithm performance after K iterations

goal

estimate norm of fixed-point residual

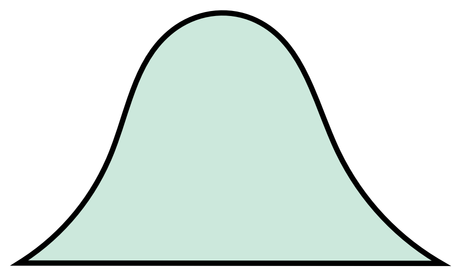
$$r^K(x) = \|z^K - z^{K-1}\|$$

worst-case

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

problem instances

convergence tolerance



probabilistic

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

probability bound

problem instances

convergence tolerance

Worst-case algorithm verification

parametric quadratic optimization

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

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!

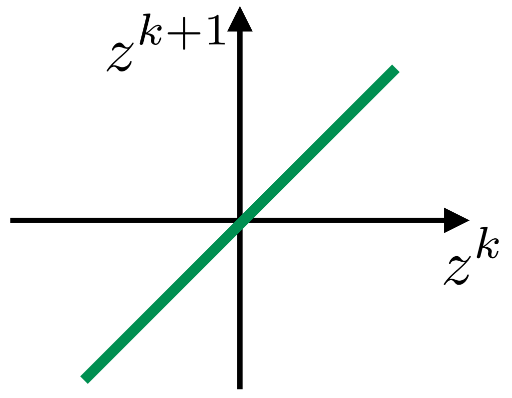
Verification of First-Order Methods for Parametric Quadratic Optimization
V. Ranjan and B. Stellato
arXiv e-prints:2403.03331 (2025)
github.com/stellatogrp/sdp_algo_verify

Algorithm steps as mixed-integer linear constraints

Linear steps → Linear constraints

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

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

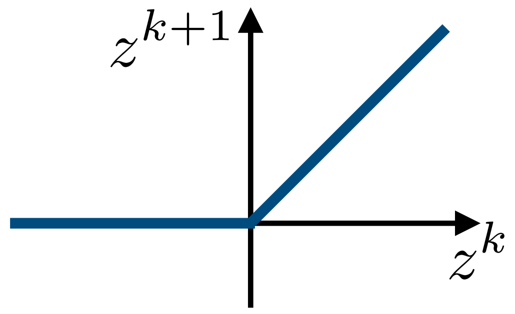


Piecewise affine steps → Mixed-integer constraints

Elementwise maximum (ReLU)

e.g., one-sided projections

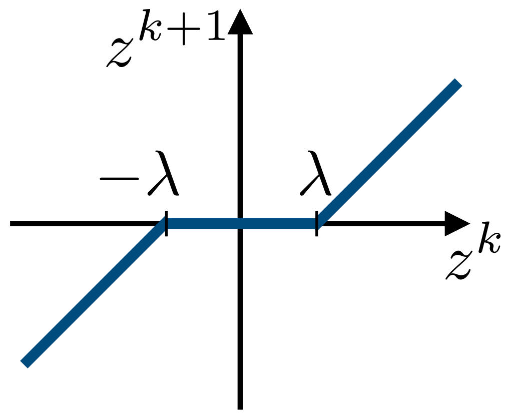
$$z^{k+1} = (z^k)_+ = \max\{z^k, 0\}$$



Soft-thresholding

e.g., prox of 1-norm function

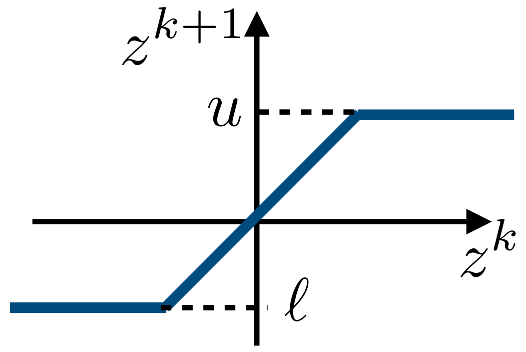
$$z^{k+1} = \phi_\lambda(z^k) = \max\{z^k, \lambda\} - \max\{-z^k, \lambda\}$$



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

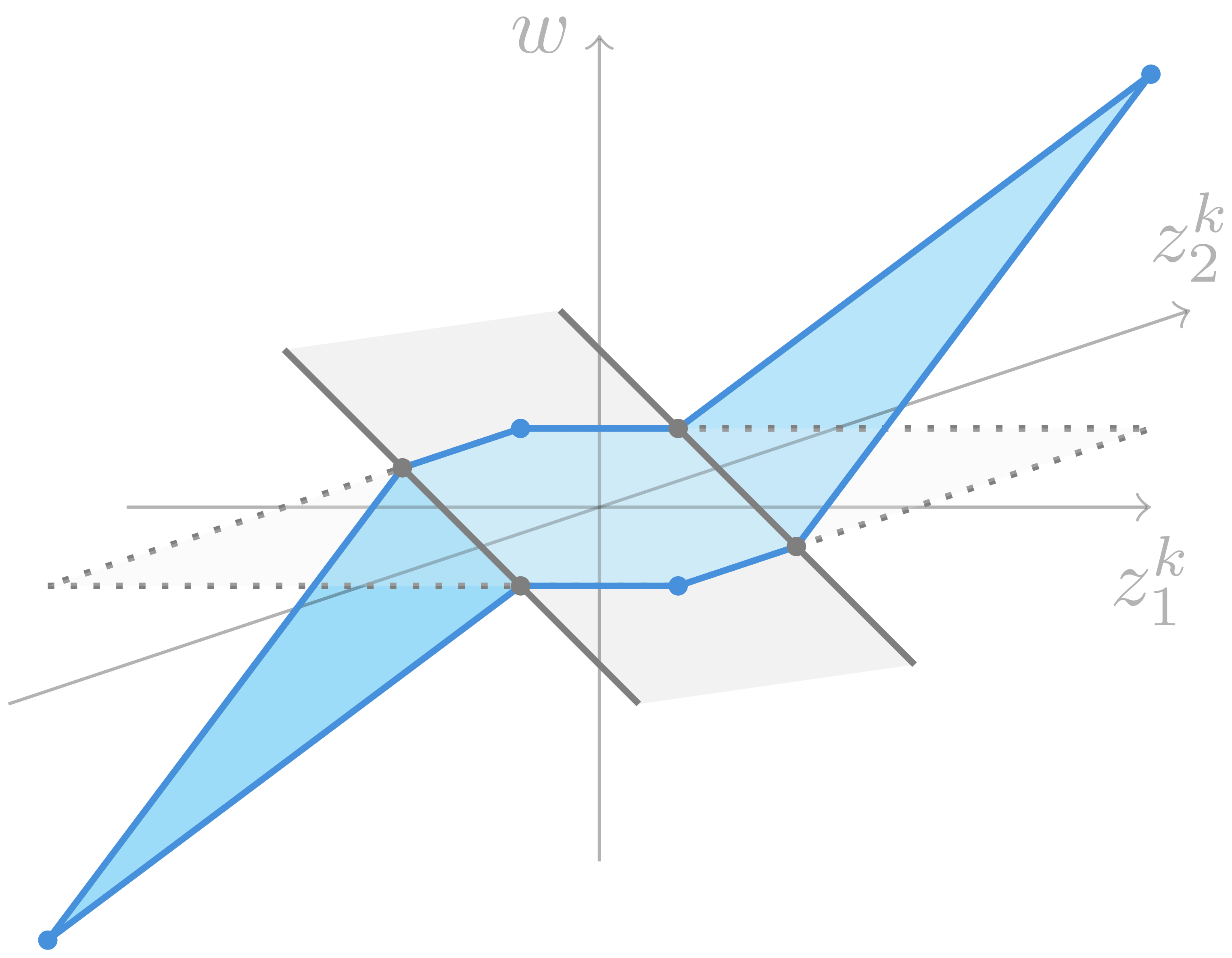
Liu et al. (2021), Albarghouthi (2021), Ceccon et al. (2022), Fischetti and Jo (2018), Tjeng et al. (2019)

Constructing strong MIP formulations

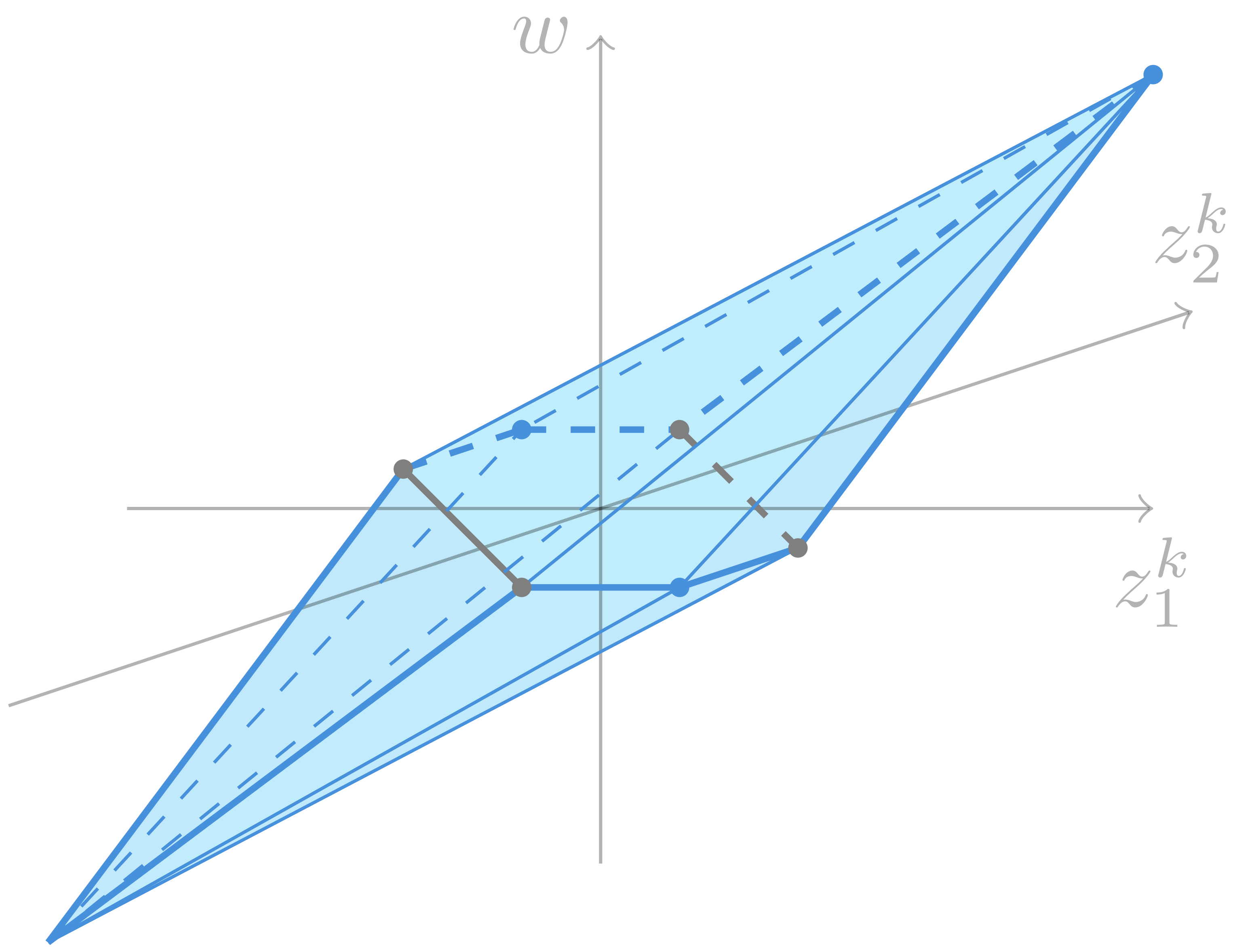
piecewise affine steps

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

$$\max_{x \in \mathcal{X}} r^K(x) = \quad \text{maximize} \quad \|z^K - z^{K-1}\|_\infty$$

$$\text{subject to} \quad z^{k+1} = T_\theta(z^k, x), \quad k = 0, \dots, K-1$$

$$z^0 = Z_\theta(x), \quad x \in \mathcal{X}$$

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)

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

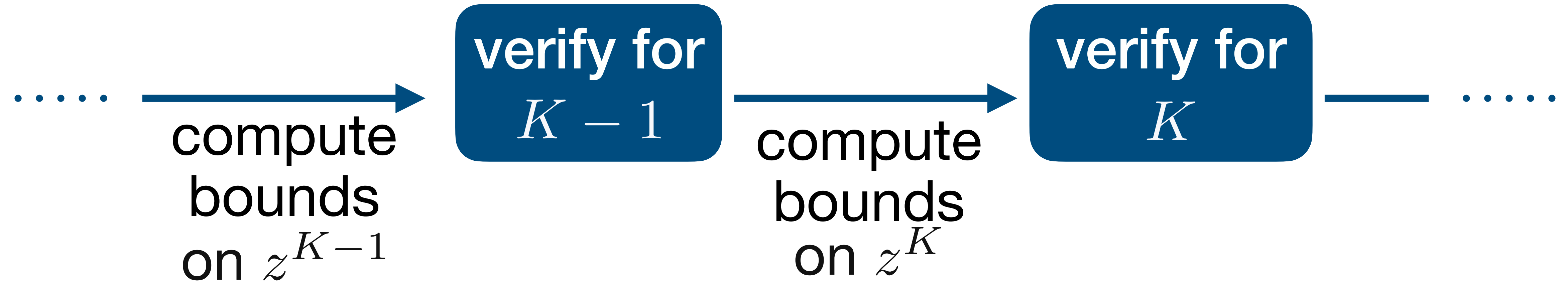
combine

bounds on latest iterate

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

main idea

solve verification problem for increasing K



compute bounds on z^{K-1}

verify for $K-1$

compute bounds on z^K

verify for K

Examples

Sparse coding for signal reconstruction

reconstructed
signal

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

known
dictionary

noisy
signal

Iterative Soft-Thresholding Algorithm (ISTA)

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

soft-thresholding
operator

Fast Iterative Soft-Thresholding Algorithm (FISTA)

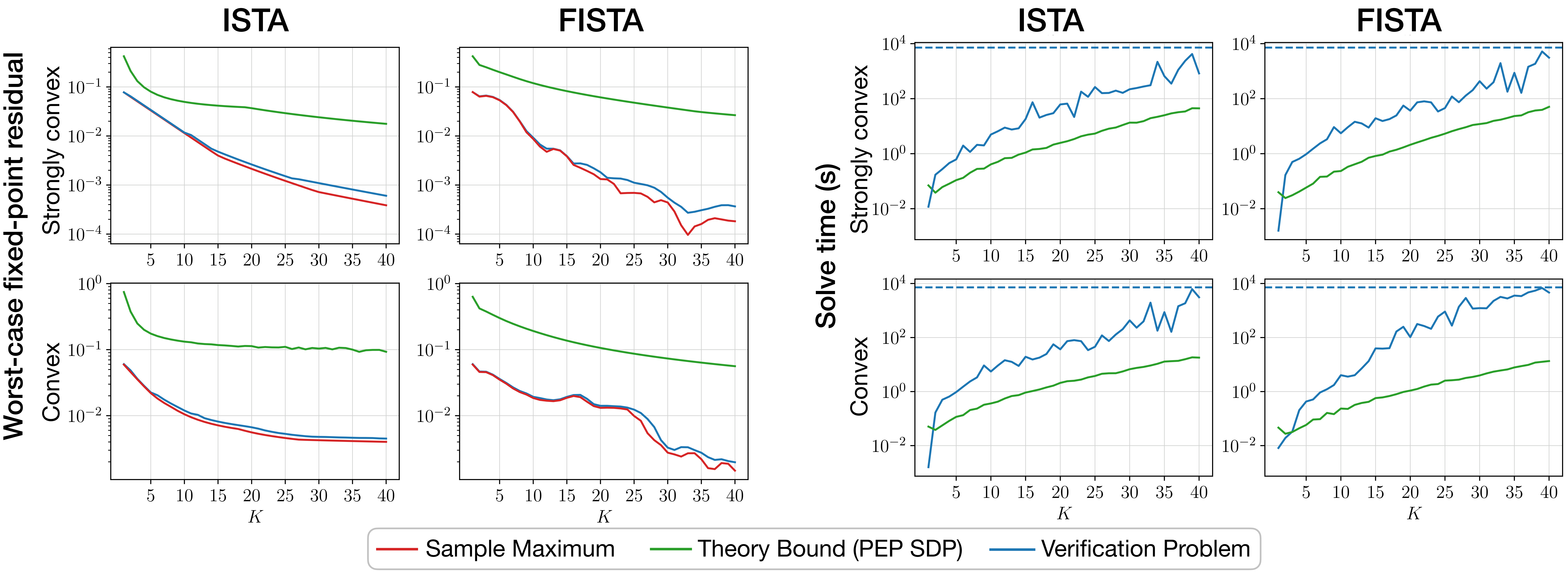
$$w^{k+1} = \phi_{\lambda\theta}\left((I - \theta D^T D)z^k + \theta D^T x\right)$$

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

soft-thresholding
operator

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

Optimal control

optimal sequence of states and controls

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

linear dynamics

initial state

$$\longrightarrow$$

condensed formulation

$$z = (u_1, \ldots, u_T)$$

$$\begin{array}{ll} \text{maximize} & (1/2) z^T P z + q(x)^T z \\ \text{subject to} & l(x) \leq Mz \leq u(x) \end{array}$$

OSQP ADMM splitting

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

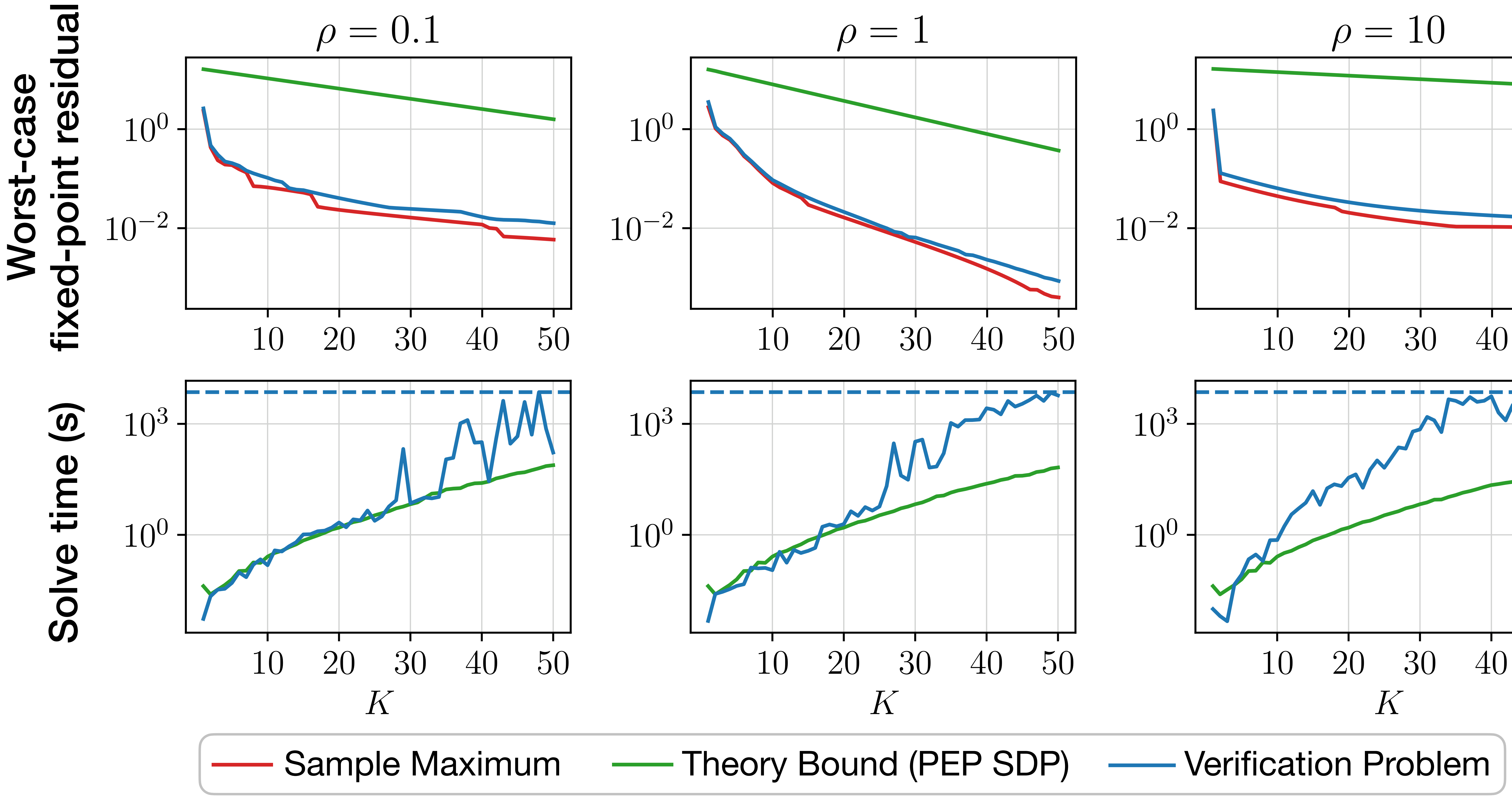
$$\text{Solve } (P + \sigma I + \rho M^T M) z^{k+1} = \sigma z^k - q(x) + \rho M^T (2w^{k+1} - v^k)$$

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

saturated linear unit

linear system

Verification results for optimal control problem

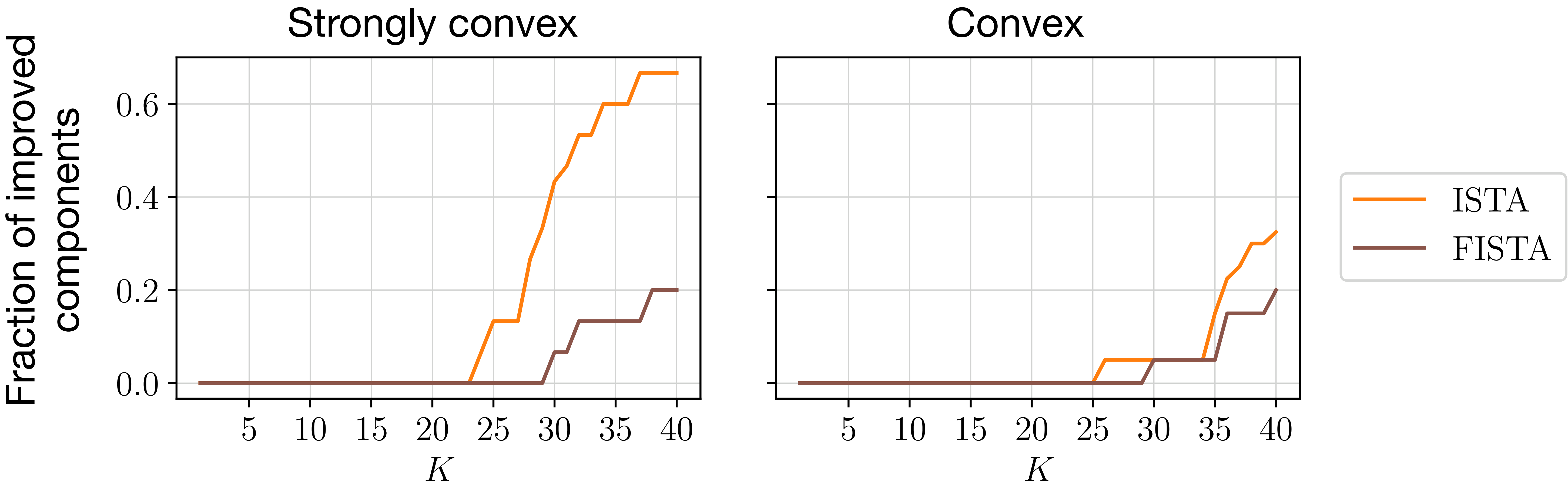


can be used to design algorithms!

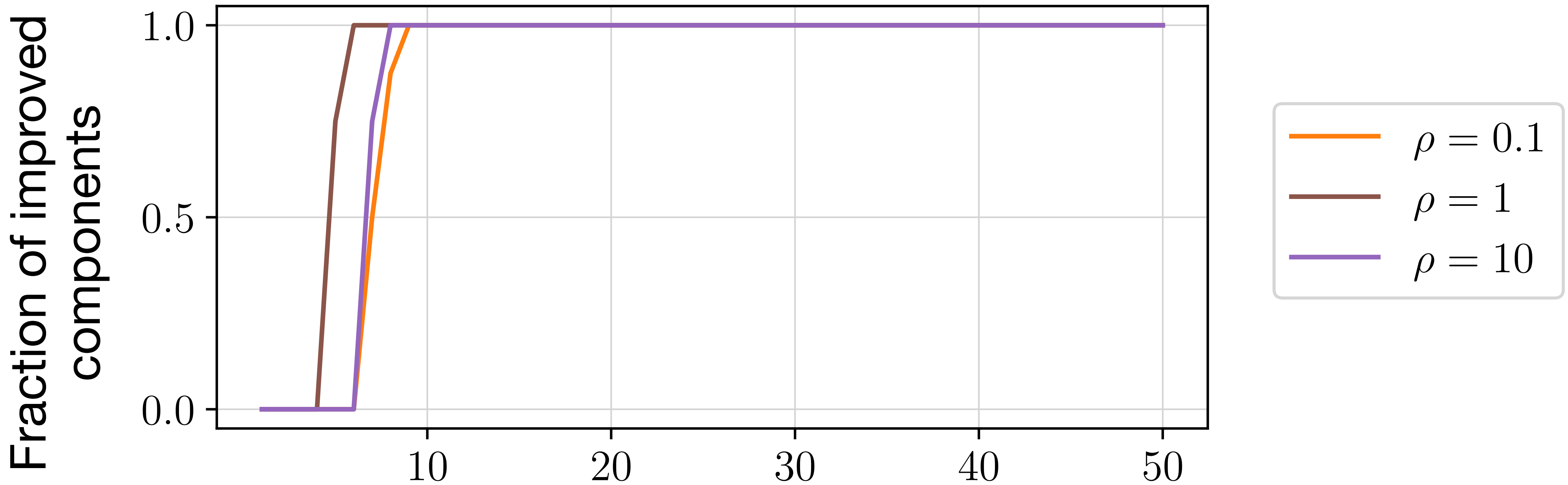
exactly quantifies the iterations required → crucial in real-time applications!

Operator theory tightens bounds on the iterates

Sparse coding



Optimal control



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

Network flow optimization

network flow

$$\begin{array}{ll} \text{minimize} & c^T z \\ \text{subject to} & A_s z \leq b_s \\ & A_d z = x \\ & 0 \leq z \leq u \end{array}$$

supplies

demands

arc-node matrices
(supply and demand)

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)

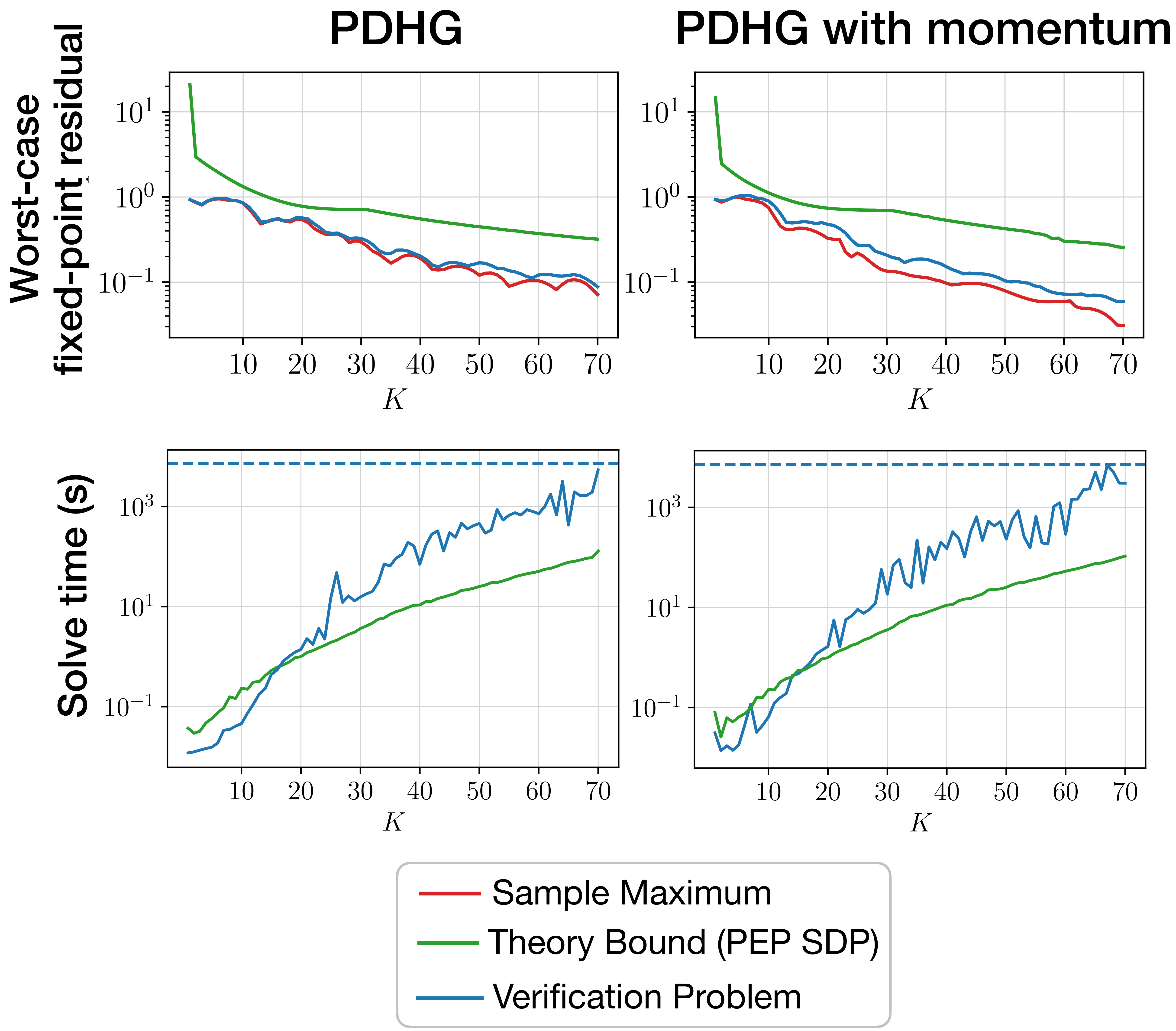
$$\begin{aligned} z^{k+1} &= \mathcal{S}_{[0,u]}(z^k - \eta(c + A_s^T v^k - A_d^T w^k)) \\ \tilde{z}^{k+1} &= z^k + k/(k+3)(z^{k+1} - z^k) \\ v^{k+1} &= (v^k + \eta(-b_s + A_s(2\tilde{z}^{k+1} - z^k)))_+ \\ w^{k+1} &= w^k + \eta(x - A_d(2\tilde{z}^{k+1} - z^k)) \end{aligned}$$

saturated
linear unit

momentum

one-sided
projection

Verification results for network flow optimization example



directly quantify PDHG ripples

can be faster than SDP for few iterations

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

1. parametric structure matters
2. MIP-based 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

minimize $(1/2)z^T Pz + x^T z$

parameters

verification problem

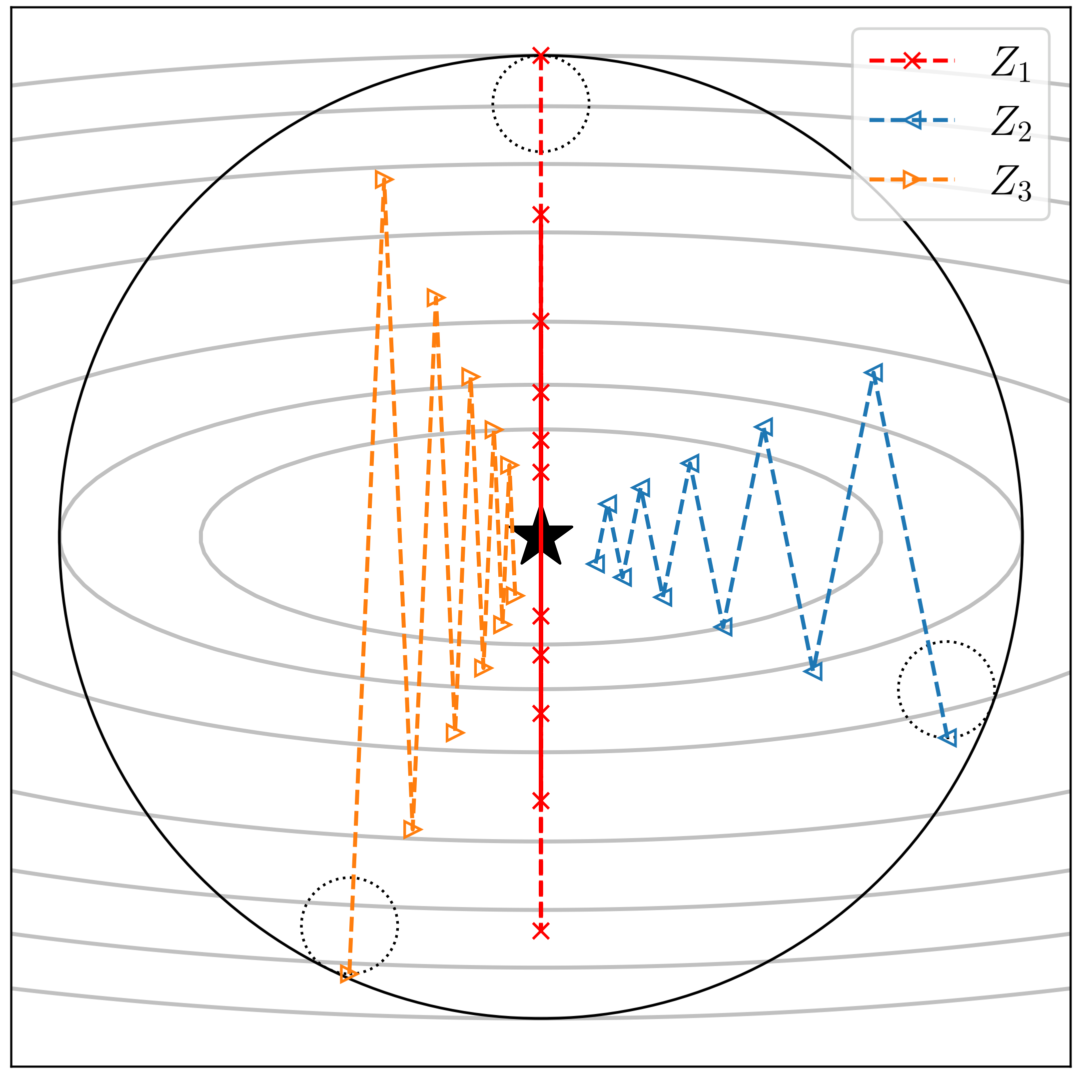
$$\begin{array}{ll} \text{maximize} & \|z^K - z^{K-1}\| \\ \text{subject to} & z^{k+1} = z^k - \theta(Pz^k + x), \quad k = 0, \dots, K-1 \\ & z^0 = Z_\theta(x), \quad x \in \mathcal{X} \end{array}$$

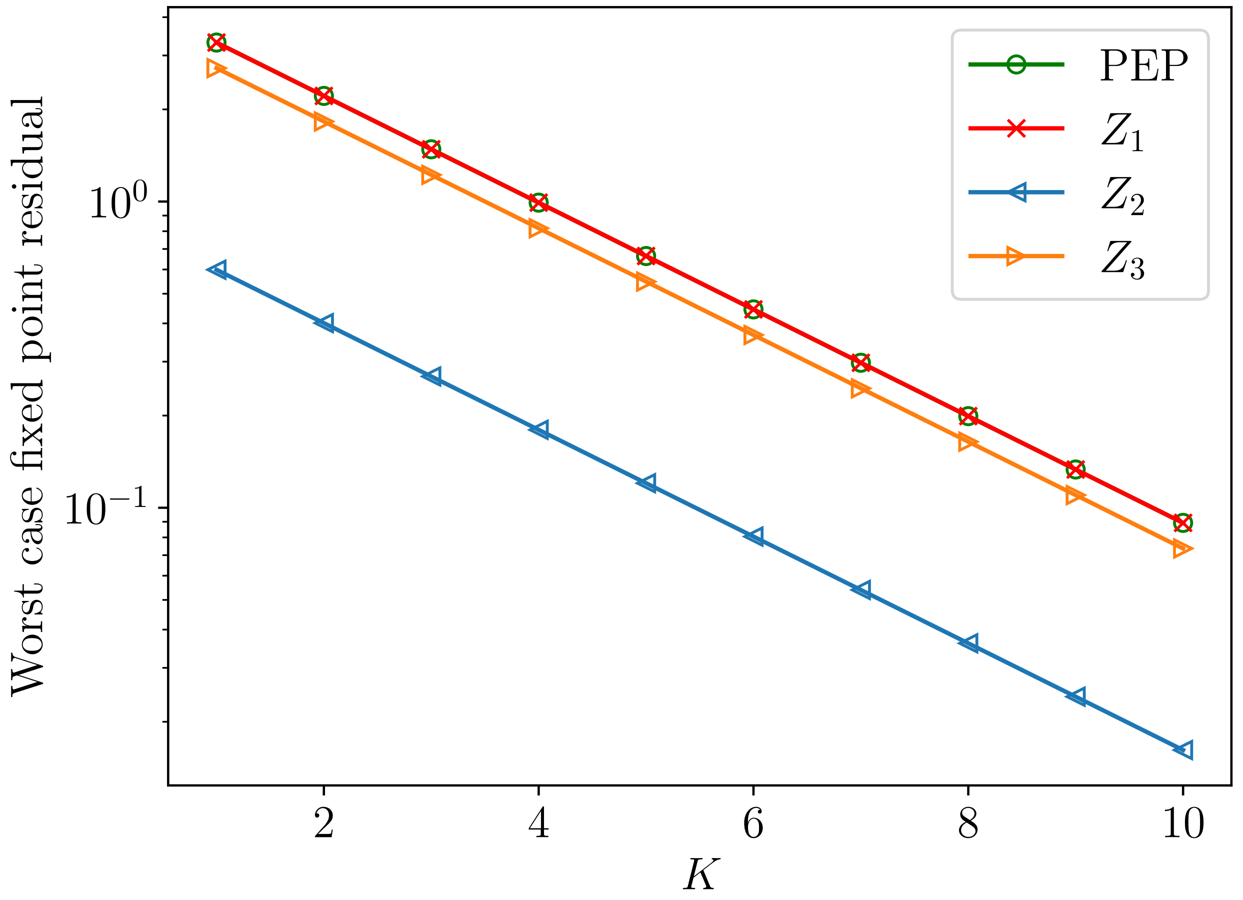
gradient descent

warm-starts

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

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





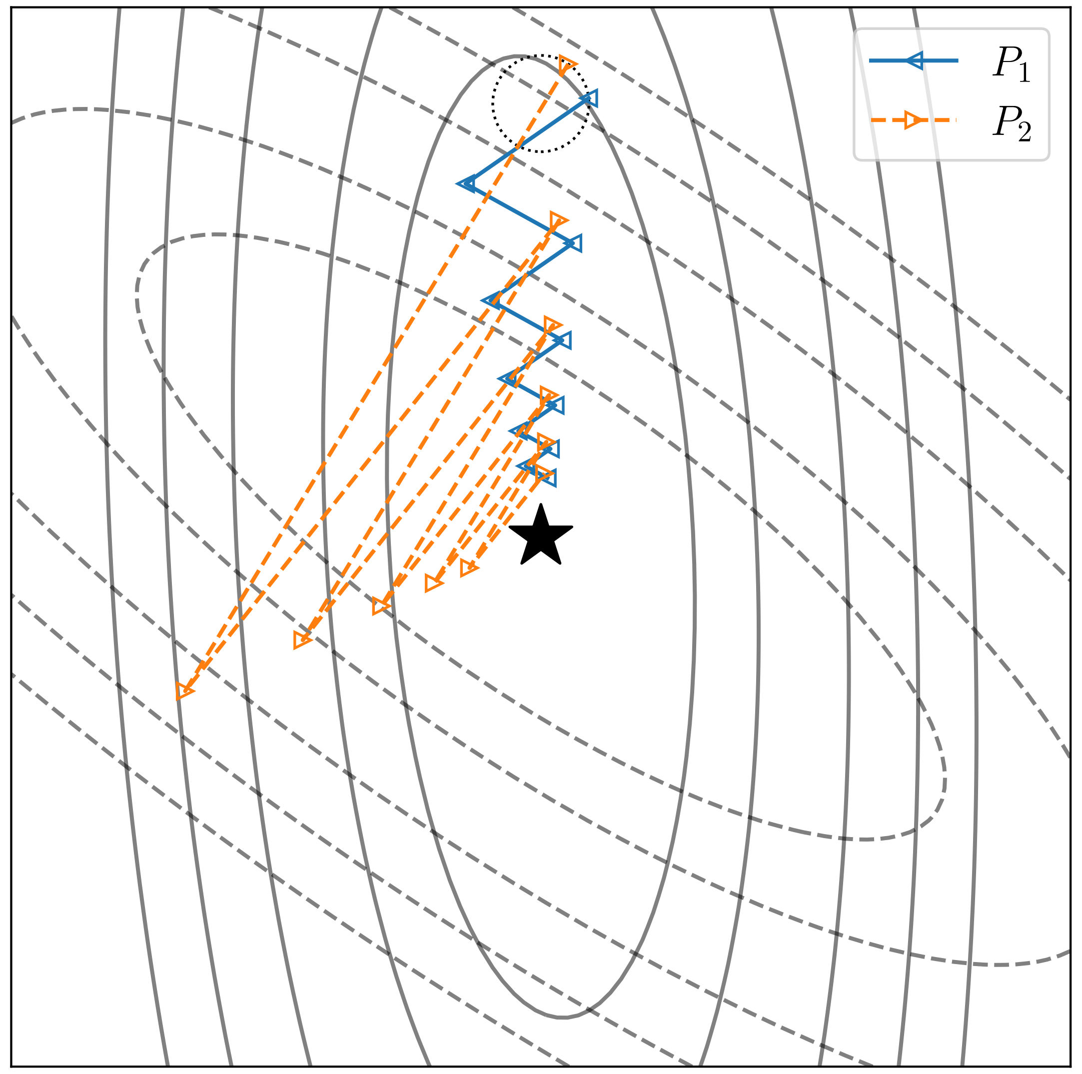
PEP-SDP cannot distinguish warm-starts

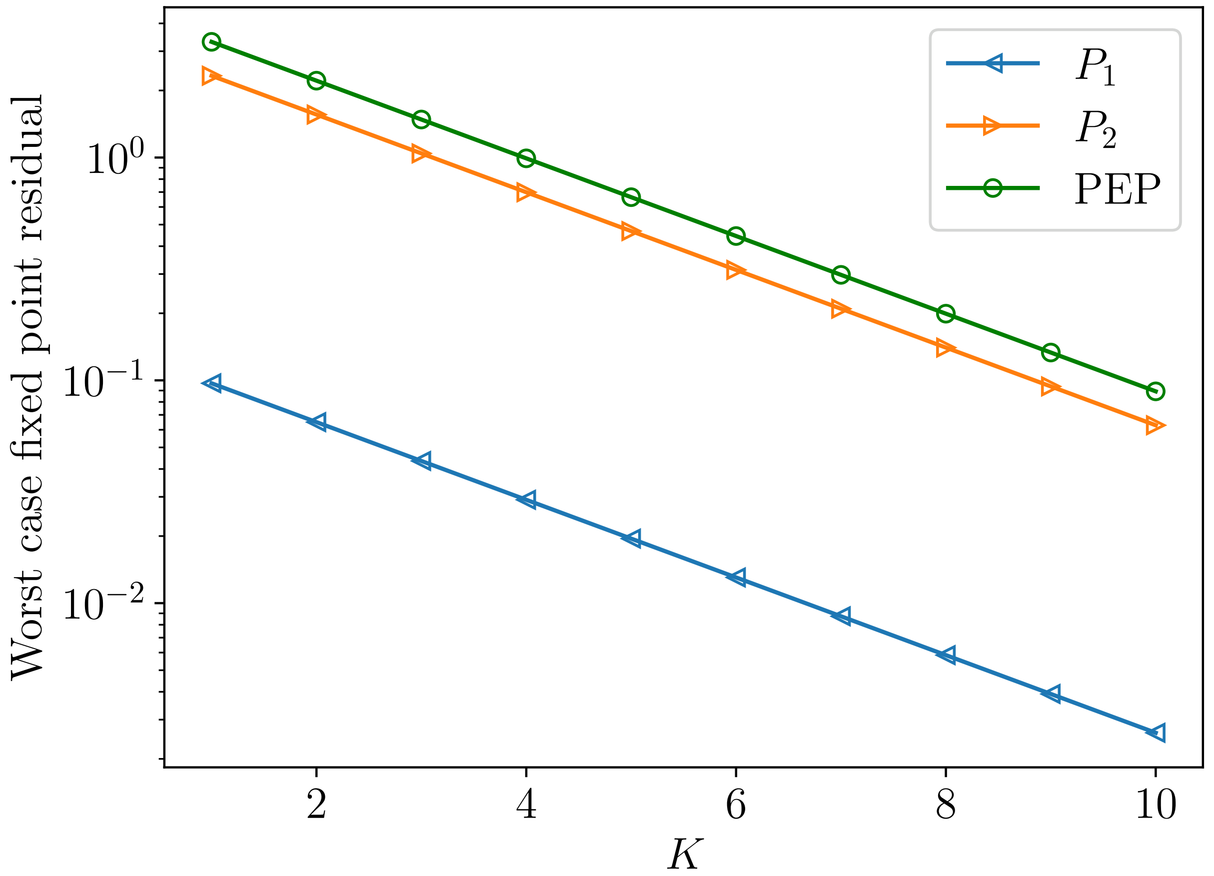
rotated functions

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

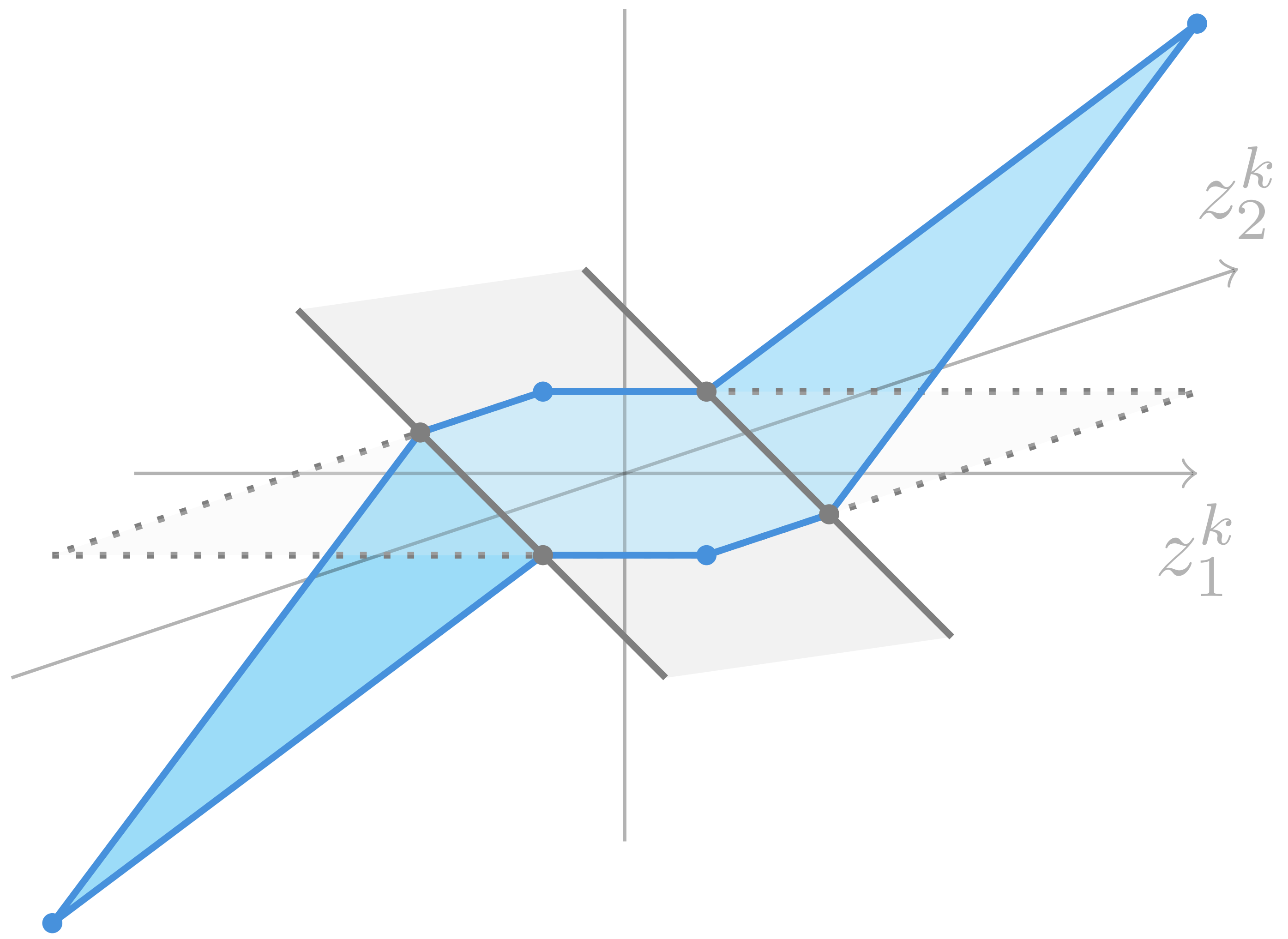
$$t = t^+ - t^-, \quad t^+ \leq \bar{t} \odot w, \quad t^- \leq -\underline{t} \odot (\mathbf{1} - w)$$
$$t^+ + t^- \leq \delta_K \leq t^+ + t^- + \max\{\bar{t}, -\underline{t}\} \odot (\mathbf{1} - \gamma)$$
$$\mathbf{1}^T\gamma = 1, \quad t^+ \geq 0, \quad t^- \geq 0$$

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

Soft-thresholding operator

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

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)

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

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

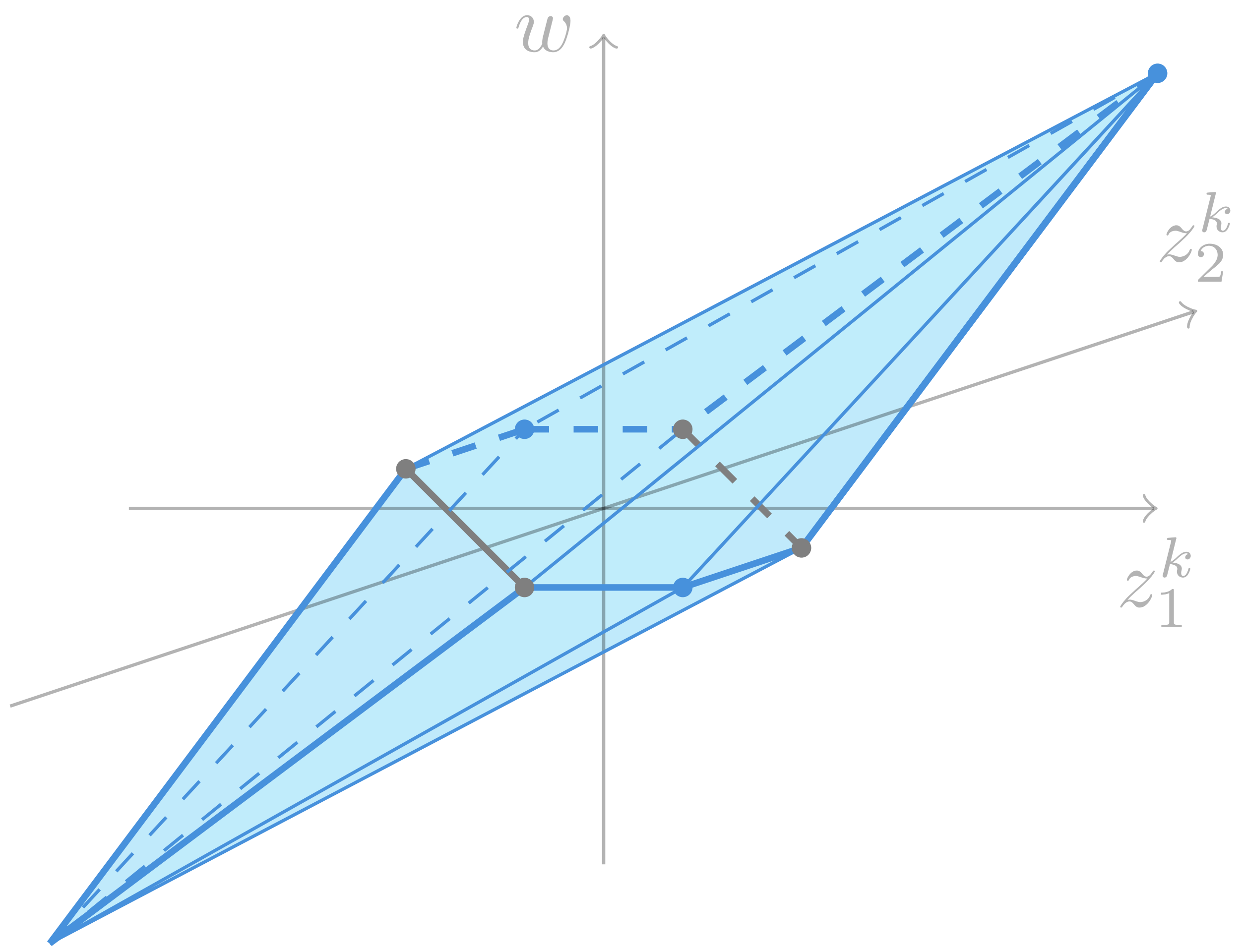
$$u_i^0 = \begin{cases} \bar{z}_i & a_i \geq 0 \\ \underline{z}_i & \text{otherwise} \end{cases} \qquad \ell_i^0 = \begin{cases} \underline{z}_i & a_i \geq 0 \\ \bar{z}_i & \text{otherwise} \end{cases}$$

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

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

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

Computing initial radius

PEP

$$
\begin{aligned}
R^2 = \quad & \text{maximize} && \|z^0 - z^\star\|_2^2 \\
& \text{subject to} && \boxed{z^\star = T(z^\star, x)} \longleftarrow \text{fixed-points} \\
& && x \in \mathcal{X}
\end{aligned}
$$

example
linear convergence

$$\alpha_K = 2\beta^{k-1}R$$

Verification

upper-bounds
∞-norm

$$
\begin{aligned}
R = \quad & \text{maximize} && \|z^0 - z^\star\|_{\boxed{1}} \\
& \text{subject to} && \boxed{z^\star = T(z^\star, x)} \longleftarrow \text{fixed-points} \\
& && x \in \mathcal{X}
\end{aligned}
$$