Data-Driven Performance Guarantees for Classical and Learned Optimizers

PRINCETON UNIVERSITY

Rajiv Sambharya, Bartolomeo Stellato

Context and Motivation

- In real-world optimization we often repeatedly solve similar instances of the same parametric problem.
- Worst-case bounds for classical optimizers can be loose since they do not take advantage of the parametric structure.
- Learned optimizers use machine learning to accelerate optimizers over the parametric family, but lack generalization guarantees.


Robotics


Machine learning


Energy


Finance

Parametric problem
minimize $f(z, \theta)$

decision variable z

Fixed-point algorithm
$z^{k+1}(\theta) = T(z^k(\theta), \theta)$

parameter $\theta \sim \mathcal{X}$

Contributions

- We use a sample convergence bound to provide probabilistic guarantees for classical optimizers over a parametric distribution of problems.
- We construct generalization bounds for learned optimizers using PAC-Bayes theory and directly optimize the bounds themselves.
- We show the strength of our guarantees with numerical examples.

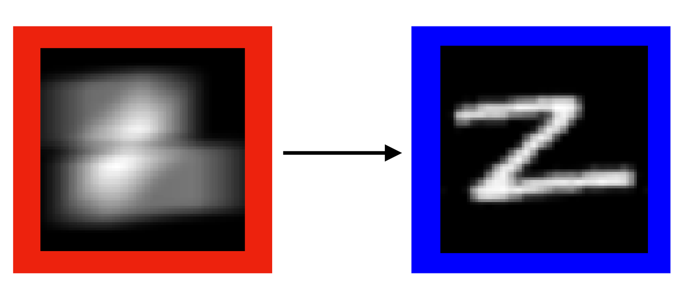
Part I: Guarantees for Classical Optimizers

Motivation

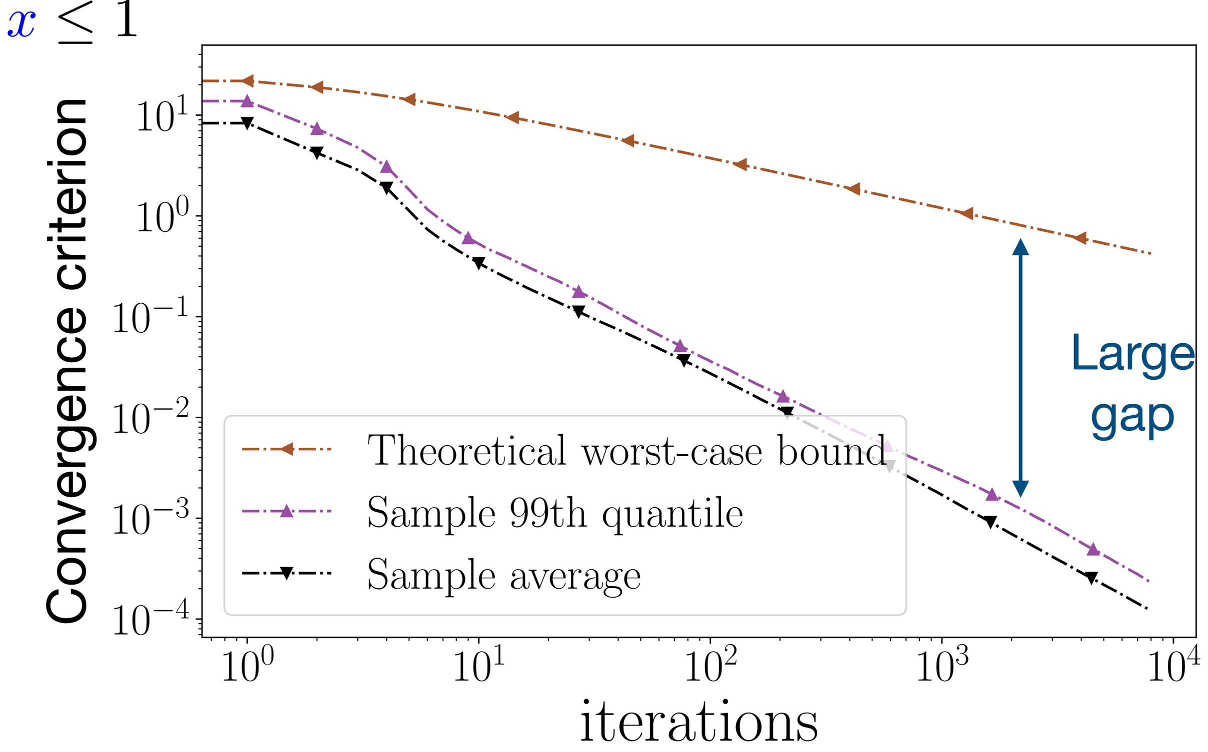
Example: image deblurring

minimize $\|Ax - b\|_2^2 + \lambda \|x\|_1$

subject to $0 \leq x \leq 1$



1000 EMNIST image deblurring problems solved w/ OSQP



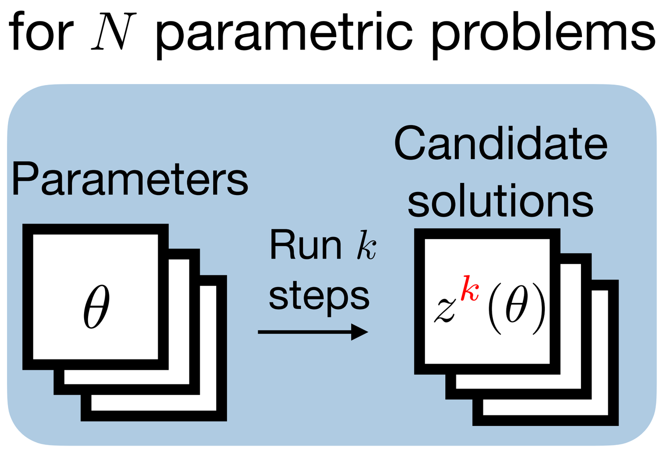
Recipe for probabilistic guarantees

algorithm steps → $e(\theta) = \mathbf{1}(\ell^k(\theta) > \epsilon)$ ← tolerance

Any metric (*e.g.*, fixed-point residual)

Step 1
Run k steps
for N parametric problems

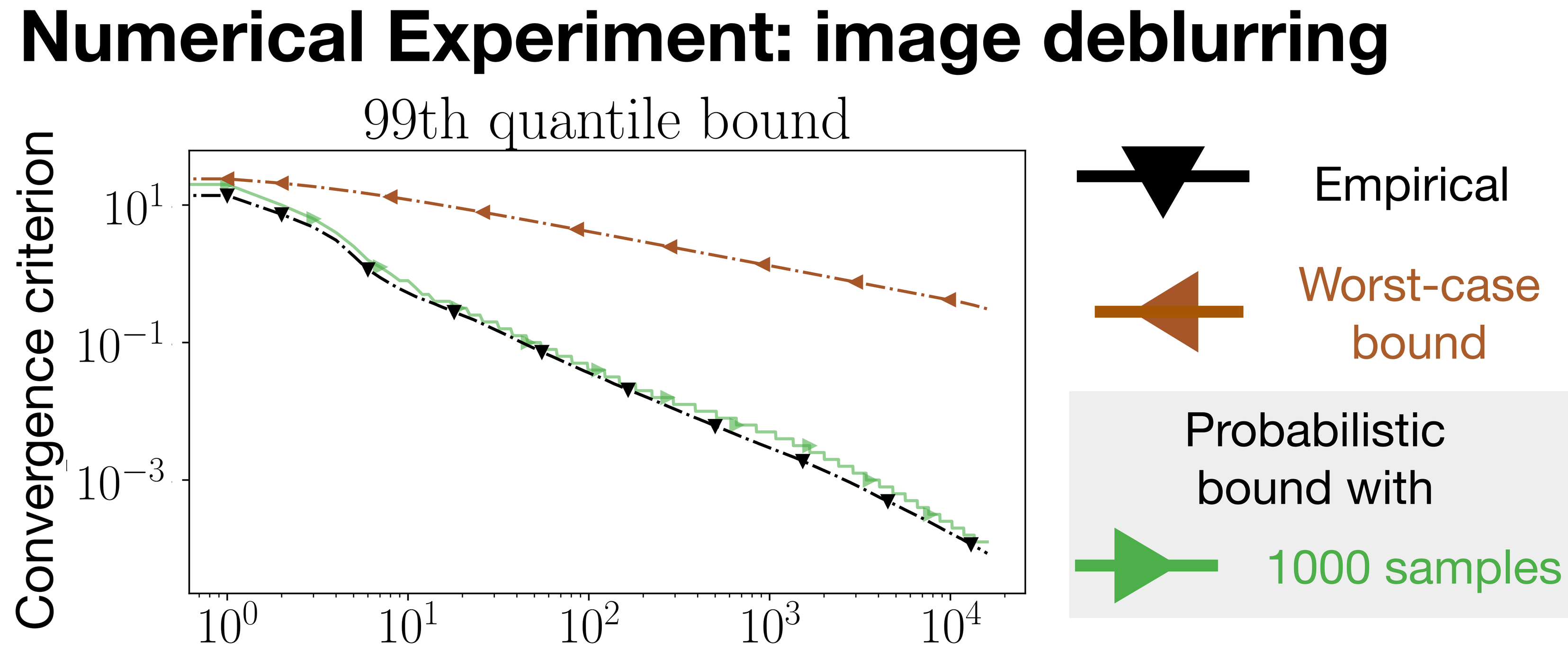
Step 2
Bound the risk w.p. $1 - \delta$



Risk, Empirical risk, Regularizer

$$\mathbf{E}_{\theta \sim \mathcal{X}} e(\theta) \leq \mathrm{KL}^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} e(\theta_i) \,\middle|\, \frac{2/\delta}{N}\right)$$

Numerical Experiment: image deblurring



With 1000 samples, we provide strong probabilistic guarantees on the 99th quantile

Part II: Guarantees for Learned Optimizers

Motivation

- Learning to optimize is a paradigm that uses machine learning to accelerate optimizers over a parametric family of problems.
- Learned optimizers lack generalization guarantees to unseen data and can fail to converge to reasonable solutions since the algorithm steps are replaced with learned variants.

Recipe for generalization guarantees

algorithm steps, tolerance, learnable weights

$$e_w(\theta) = \mathbf{1}(\ell_w^k(\theta) > \epsilon)$$

McAllester bound: given posterior and prior distributions P and P_0, with probability $1 - \delta$ [McAllester et. al 2003]

$$\mathbf{E}_{\theta \sim \mathcal{X}} \mathbf{E}_{w \sim P} e_w(\theta) \leq \mathrm{KL}^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{E}_{w \sim P} e_w(\theta_i) \,\middle|\, \frac{1}{N}(\mathrm{KL}(P \,\|\, P_0) + \log(N/\delta))\right)$$

risk $\leq \mathrm{KL}^{-1}$ (empirical risk | regularizer)

Optimize the bounds directly

Numerical Experiment: sparse coding

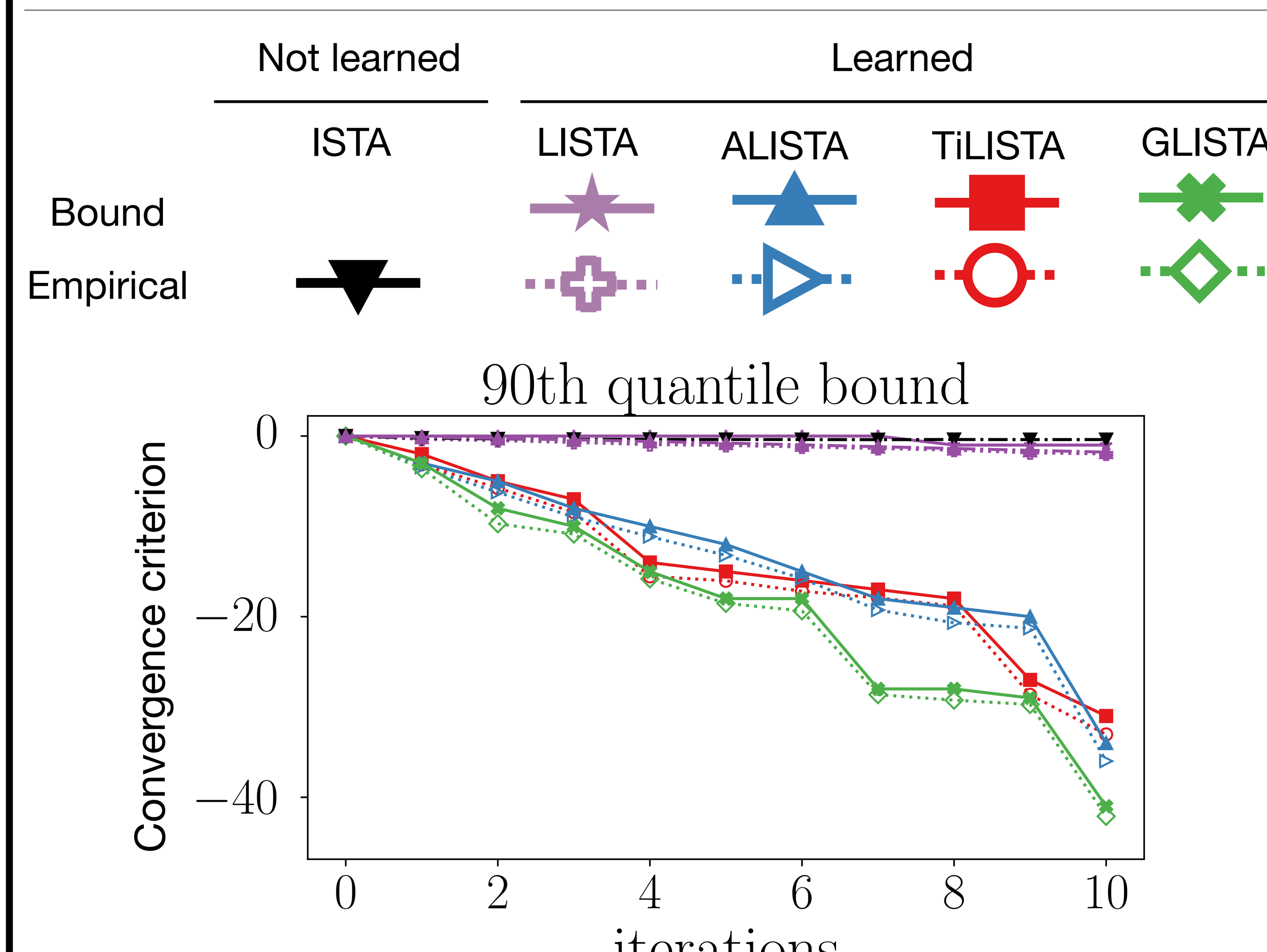
Noisy measurements $\theta = b$ → **Sparse coding**: Recover sparse $z^\star$ from $b = Dz^\star + \sigma$ → Ground truth sparse signal $z^\star$

D: dictionary, σ: noise

Standard technique
minimize $\|Dz - b\|_2^2 + \lambda \|z\|_1$

Classical optimizer — ISTA
$z^{j+1} = \text{soft threshold}_{\frac{\lambda}{L}}\left(z^j - \frac{1}{L}(Dz^j - b)\right)$

Learned optimizer — Learned ISTA (LISTA)
$z^{j+1} = \text{soft threshold}_{\psi^j}\left(W_1^j z^j + W_2^j b\right)$



Learned optimizers provably perform well in just 10 steps

Our bounds are close to empirical performance