Towards Robust Learning to Optimize with Theoretical Guarantee

Qingyu Song†, Wei Lin†, Juncheng Wang‡, Hong Xu†

†The Chinese University of Hong Kong, ‡Hong Kong Baptist University

What is learning to optimize (L2O)?

- Optimization Problem

$$\min_{x} f(x),$$
$$x \in R^n, f: R^n \to R$$

Human-Designed Algorithm

Gradient Descent

Learning to Optimize

ML/DL Model + Training

- Benefits
  - Better optimality (potential).
  - Better convergence/efficiency [1].

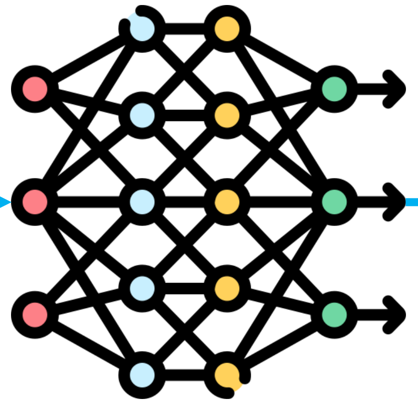
[1]. Jialin Liu, Xiaohan Chen, Zhangyang Wang, Wotao Yin, and HanQin Cai. Towards Constituting Mathematical Structures for Learning to Optimize. In ICML, 2023.

How does L2O work?

- Workflow of L2O (Inference)

Given initial point x_i^0

$$x_i^0 \in \boxed{N(0,1)}, \boxed{f\left(x_i^0\right)}$$



x_i^*

A distribution

A class of objective

L2O’s Failure in Out-of-Distribution (OOD) Scenarios

① x OOD

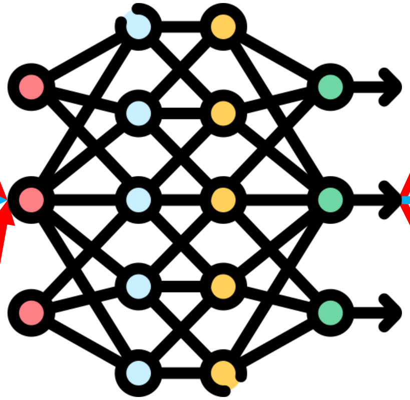
$x^0 \in N(10,1),\ f(x)$

In-Distribution (InD)

$x^0 \in N(0,1),\ f(x)$

② f OOD

$x^0 \in N(0,1),\ f = f(x-10)$



$x^* \in D'$

$x^* \neq x^*$

$x^* \in D$

$x^* \neq x^* + 10$

$x^* \in D''$

Theoretical Convergence Analysis

- Convergence of Single-Iteration (Smooth Case)

$$
\begin{aligned}
&F'(x_k + s_k) - F'(x_{k-1} + s_{k-1}) \\
&\le \boxed{-\frac{\|\nabla f'(x_{k-1} + s_{k-1})\|^2}{2L}} \qquad \text{Convergence of Gradient-Descent} \\
&\left.\begin{aligned}
&+ L\| \operatorname{diag}(\mathbf{J}_{1,k-1} s') \nabla f'(x_{k-1} + s_{k-1})\|^2 \\
&+ L\|\frac{\nabla f'(x_{k-1} + s_{k-1}) - \nabla f(x_{k-1})}{2L} - \mathbf{J}_{2,k-1} s'\|^2.
\end{aligned}\right\} \text{Deterioration w.r.t. OOD}
\end{aligned}
$$

Theoretical Convergence Rate Analysis

- Convergence Rate (Smooth Case)

$$\min_{k=1,\ldots,K} F'(x_k + s_k) - F'(x^* + s^*)$$

$$\le \boxed{\frac{L}{2}\|x_0 - x^* + s_0 - s^*\|^2 - \frac{L}{2}\|x_K - x^* + s_K - s^*\|^2}$$

Convergence Rate of [Gradient-Descent]

$$+ \frac{L}{K}\sum_{k=1}^{K}(x_k + s_k - x^* - s^*)^\top$$

$$\Big(x_k + s_k - \big(x_{k-1} + s_{k-1} - \frac{\nabla f'(x_{k-1} + s_{k-1})}{L}\big)\Big).$$

Deterioration w.r.t. OOD

Convergence Improvement

- Upper Bound Relaxation

$$\begin{aligned}
&F'(x_k + s_k) - F'(x_{k-1} + s_{k-1}) \\
&\le -\frac{\|\nabla f'(x_{k-1} + s_{k-1})\|^2}{2L} \\
&+ \frac{\|\nabla f'(x_{k-1} + s_{k-1}) - \nabla f(x)\|^2}{2L} \\
&+ \left(LC_1^2 n \|\nabla f'(x_{k-1} + s_{k-1})\|^2 + 2LC_2^2 n\right) \boxed{\|s'\|^2}.
\end{aligned}$$

OOD vector,
NN's input feature

- Improve upper bound: Magnitude reduction.
  - Our approach: Input Feature Simplification.

A New L2O Model with Gradient-Only Input

- New Model Formulation Based on [1]

$$
\begin{aligned}
x_k =& x_{k-1} - \mathbf{R}_k \nabla f(x_{k-1}) - \mathbf{R}_k g_k - \mathbf{Q}_k v_{k-1} - b_{1,k}, \\
v_k =& (\mathbf{I} - \mathbf{B}_k) G_k + \mathbf{B}_k G_{k-1} - b_{2,k}, \\
G_k :=& \mathbf{R}_k^{-1} (x_{k-1} - x_k - \mathbf{Q}_k v_{k-1} - b_{1,k}),
\end{aligned}
$$

- Learn $\mathbf{R}, \mathbf{Q}, \mathbf{B}$. Details at

[1]. Jialin Liu, Xiaohan Chen, Zhangyang Wang, Wotao Yin, and HanQin Cai. Towards Constituting Mathematical Structures for Learning to Optimize. In ICML, 2023.

Empirical Outperformance

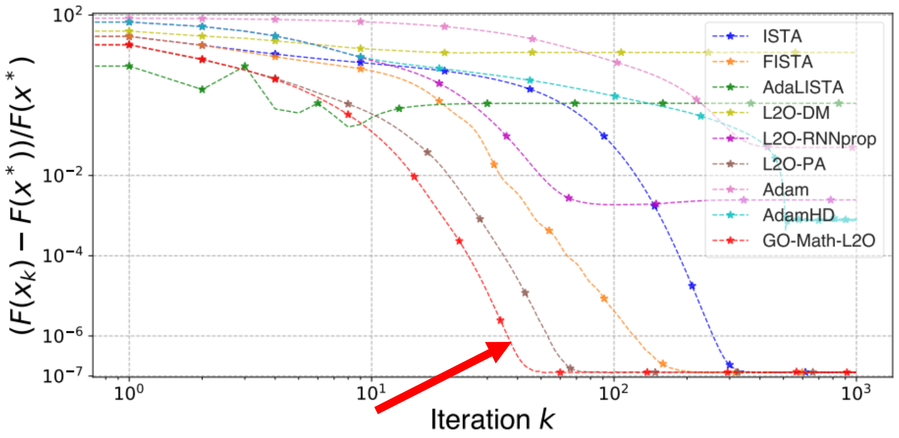


Figure 1. LASSO Regression: InD.

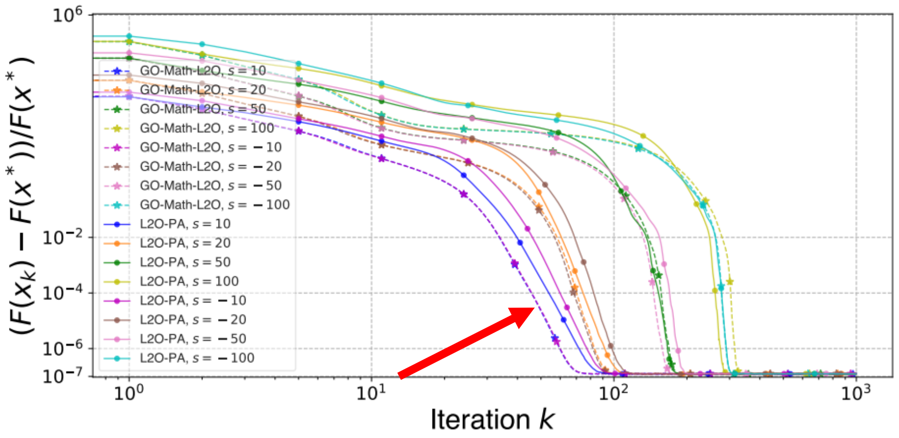


Figure 3. LASSO Regression: OOD by Trigger 1.

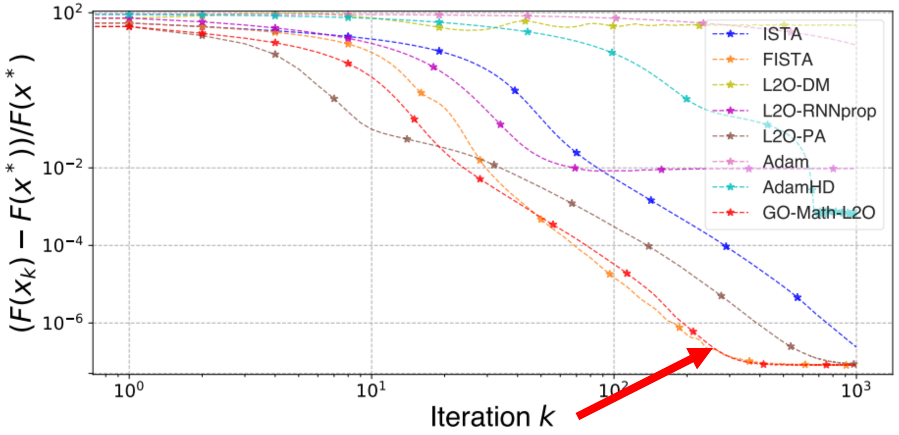


Figure 2. LASSO Regression: Real-World OOD.

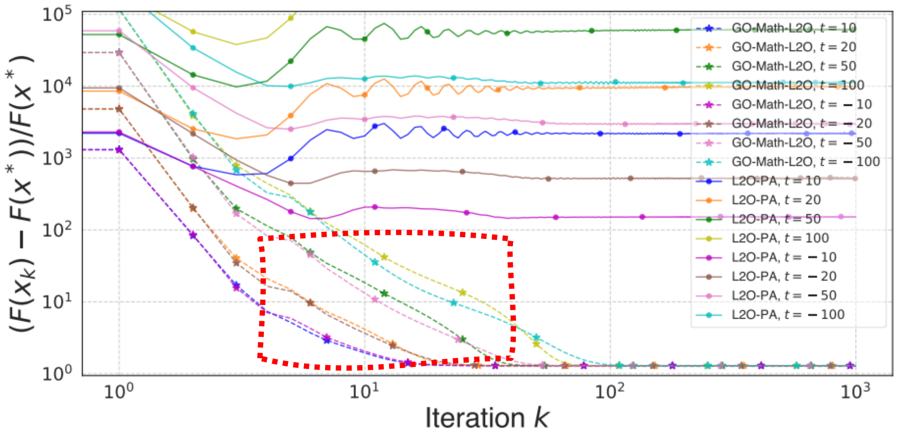


Figure 4. LASSO Regression: OOD by Trigger 2.

Empirical Outperformance

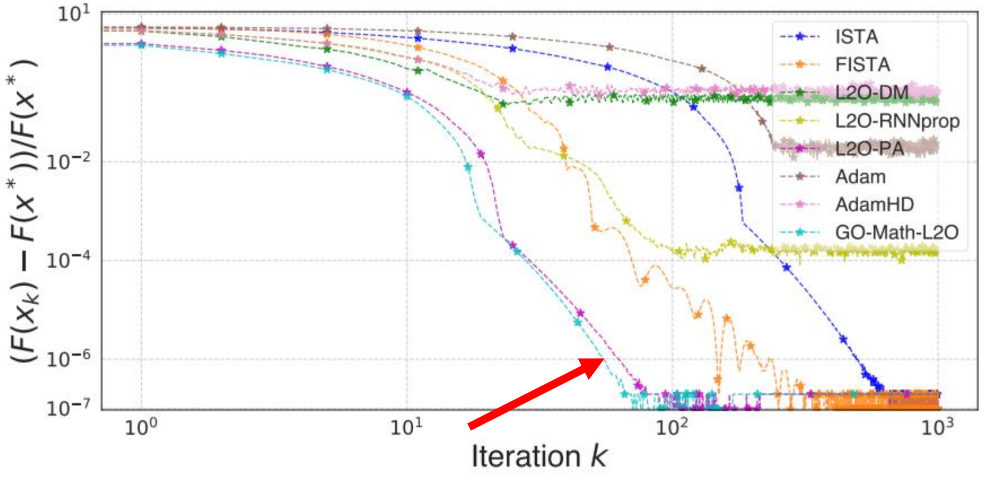


Figure 14. Logistic Regression: Real-World Ionoshpere Dataset.

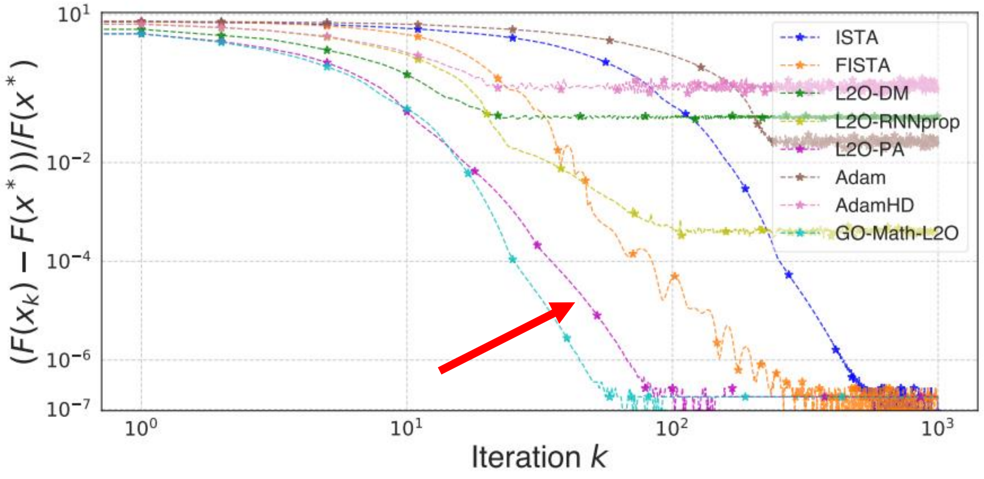


Figure 15. Logistic Regression: Real-World Spambase Dataset.

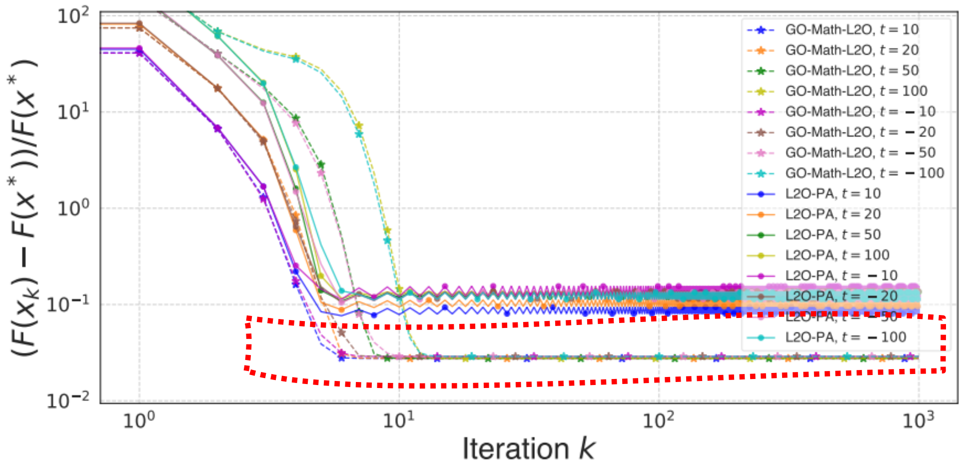


Figure 17. Logistic Regression: OOD by Trigger 2.

Thank You!

Project at NetX-lab/GoMathL2O-Official (github.com)

Paper

Code

香港中文大學 The Chinese University of Hong Kong

香港浸會大學 HONG KONG BAPTIST UNIVERSITY