

A Simple Guard for Learned Optimizers

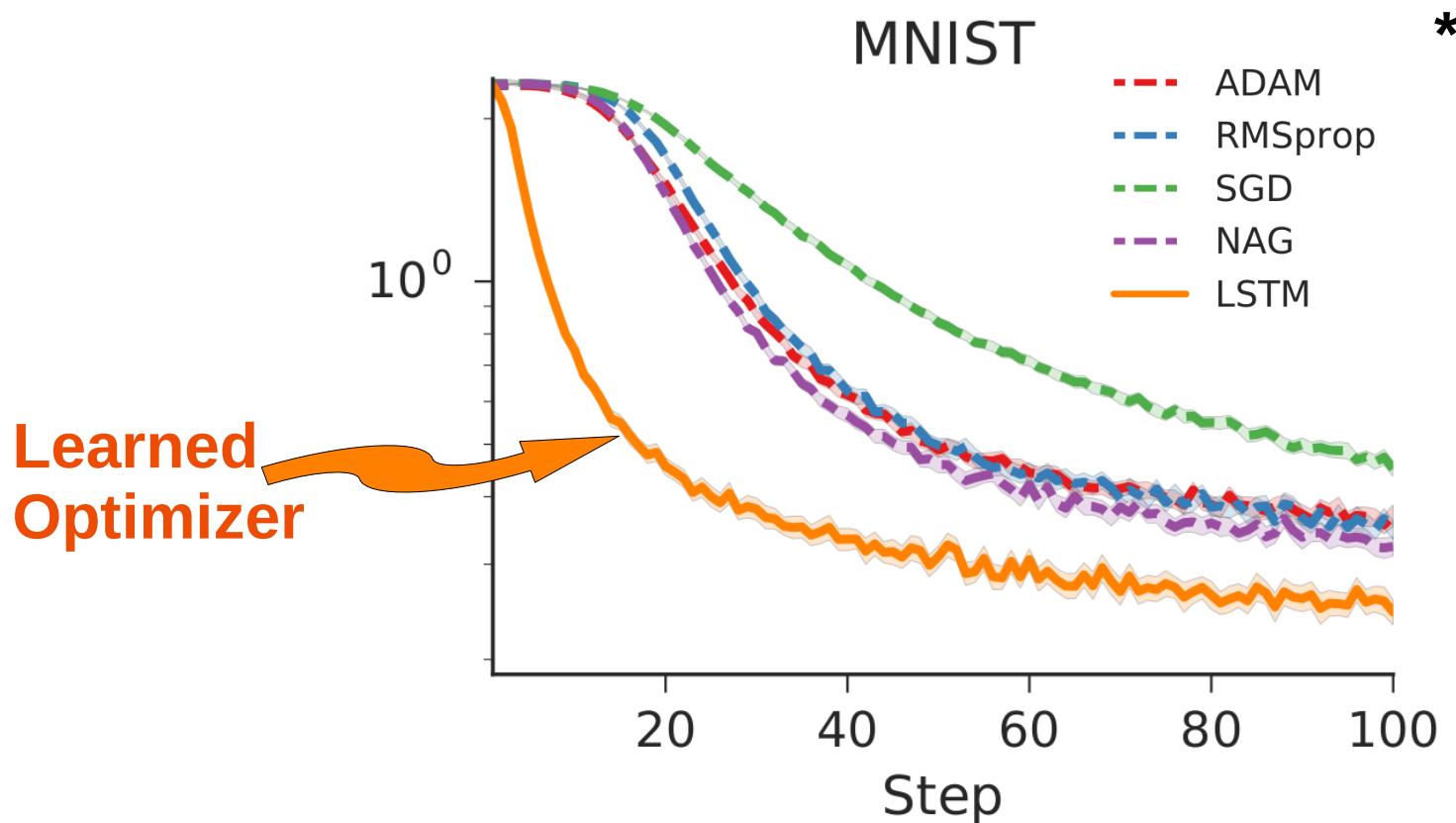


Jaroslav Vítků, Isabeau Prémont-Schwarz, Jan Feyereisl



The Hope

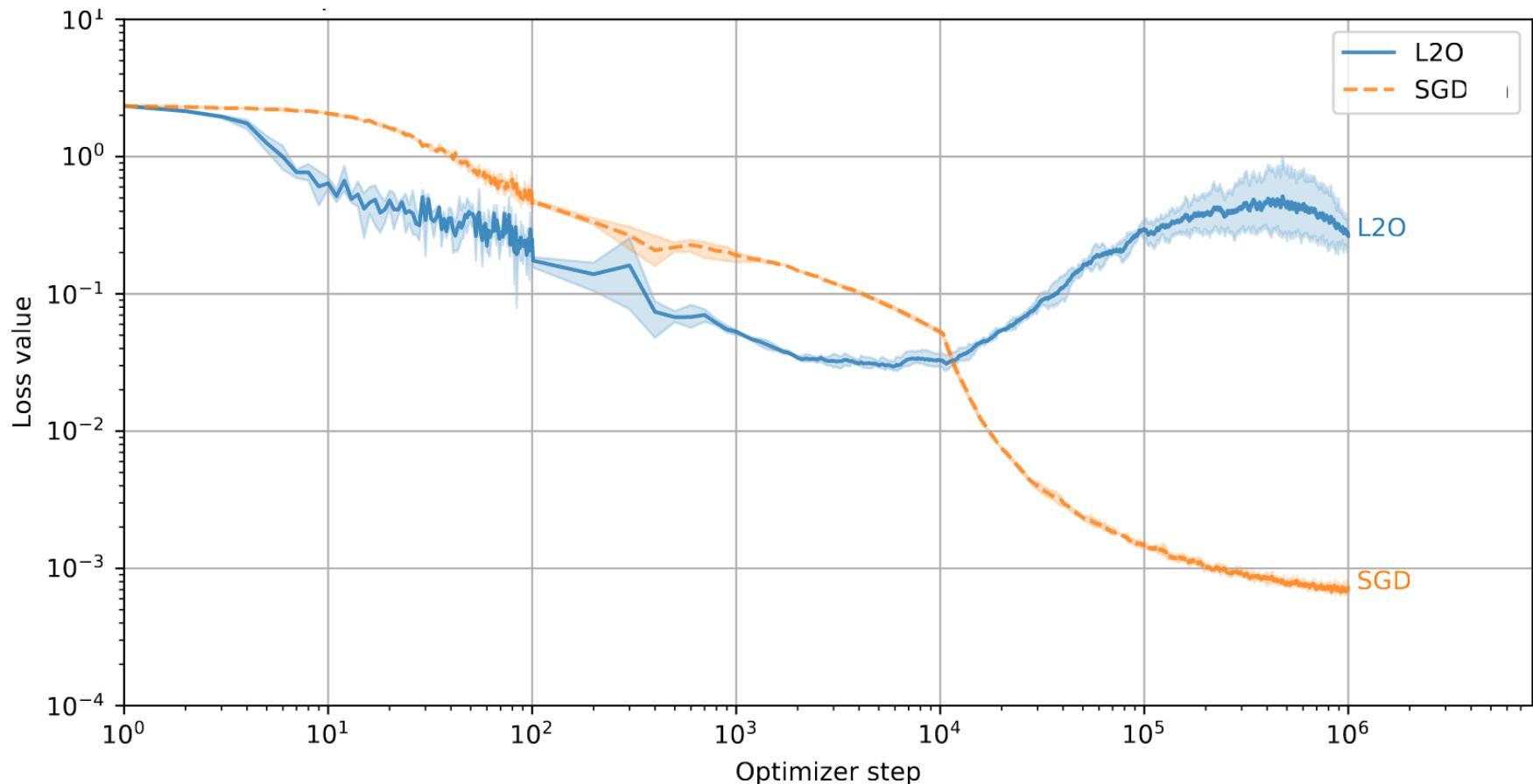
The Hope



* Andrychowicz et. al, *Learning to Learn by Gradient Descent by Gradient Descent*, Neurips 2016

The Problem

The Problem



The Problem

THEORETICAL

- Convergence
Guarantee

The Problem

THEORETICAL

- Convergence
Guarantee

PRACTICAL

The Problem

THEORETICAL

- Convergence Guarantee

PRACTICAL

- Longer Horizon

The Problem

THEORETICAL

- Convergence Guarantee

PRACTICAL

- Longer Horizon
- Different Data / Distributional Shift

The Problem

THEORETICAL

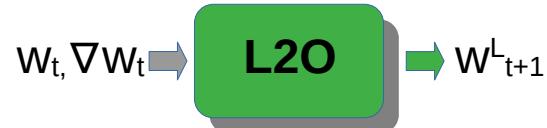
- Convergence Guarantee

PRACTICAL

- Longer Horizon
- Different Data / Distributional Shift
- Change in NeuralNet architecture

The Solution

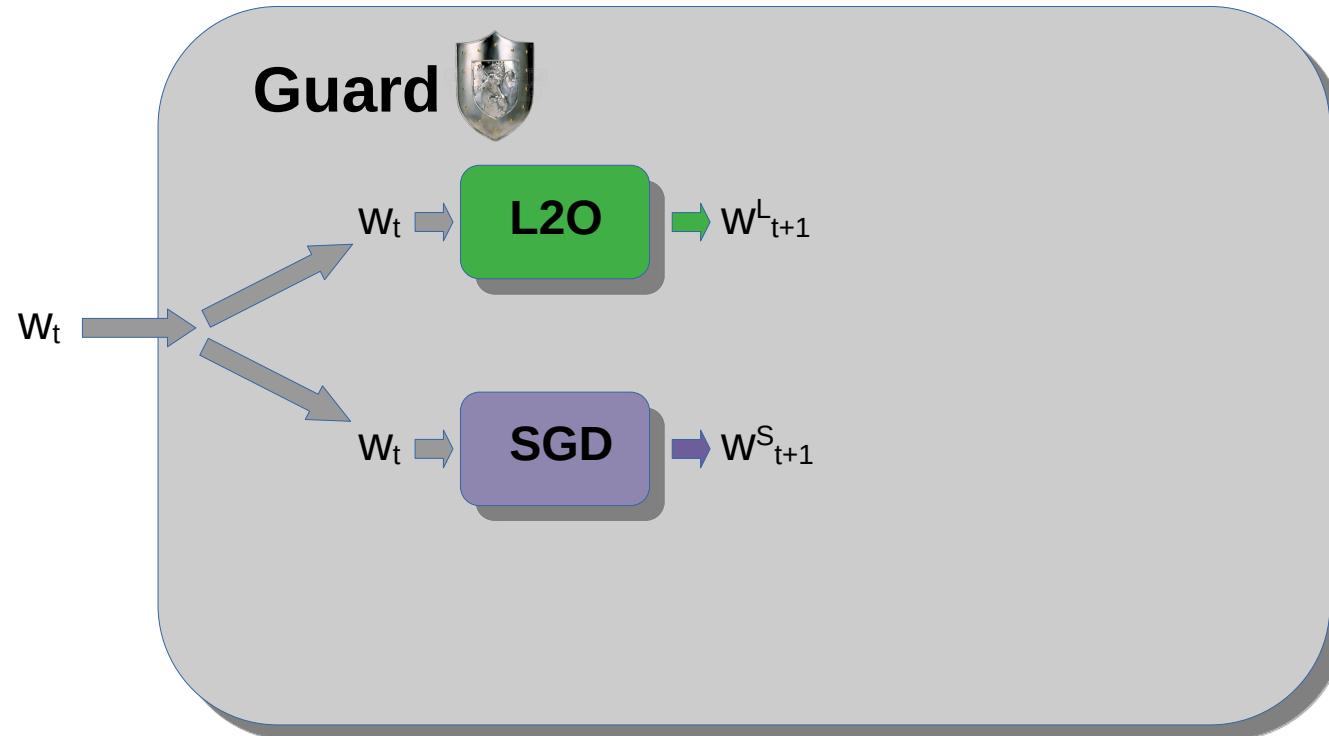
The Solution



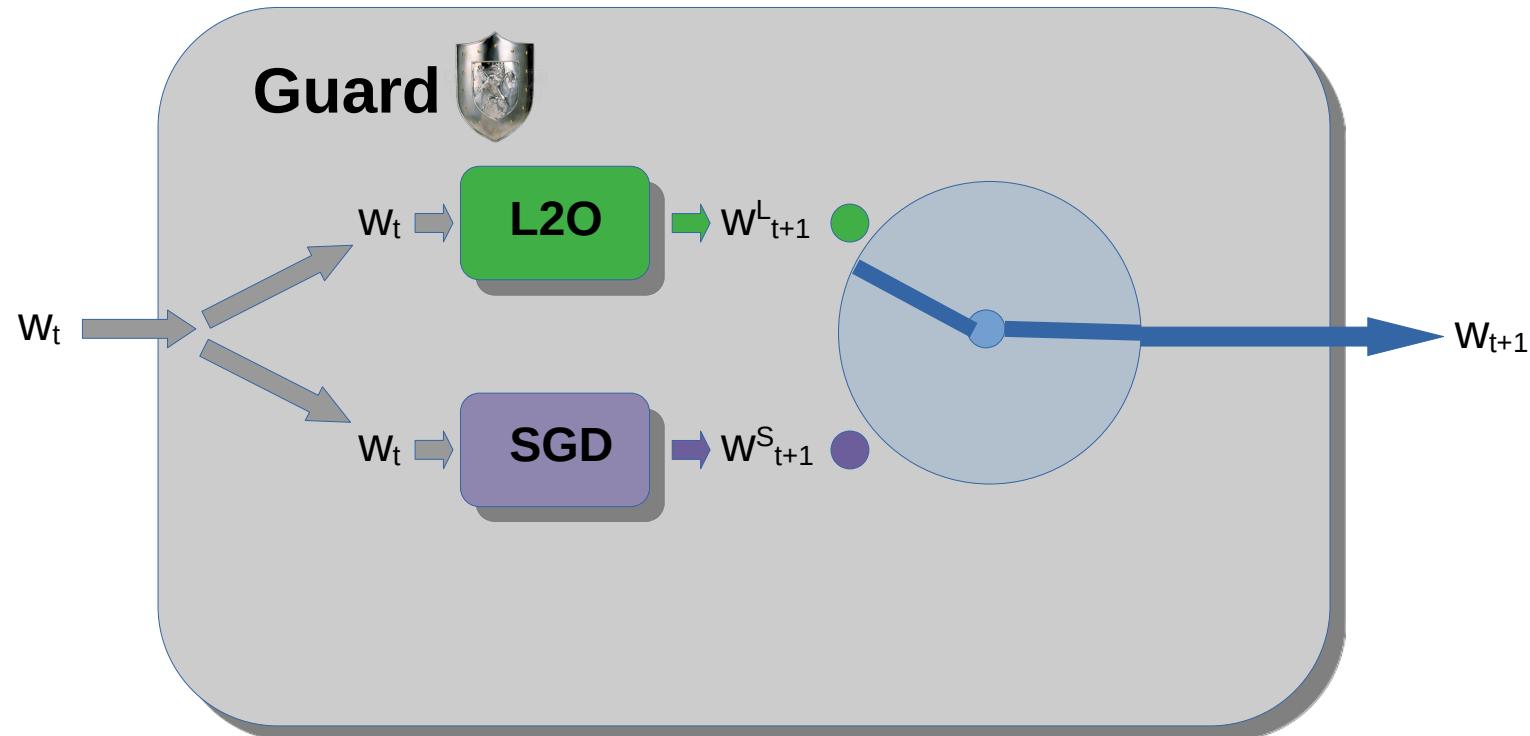
The Solution



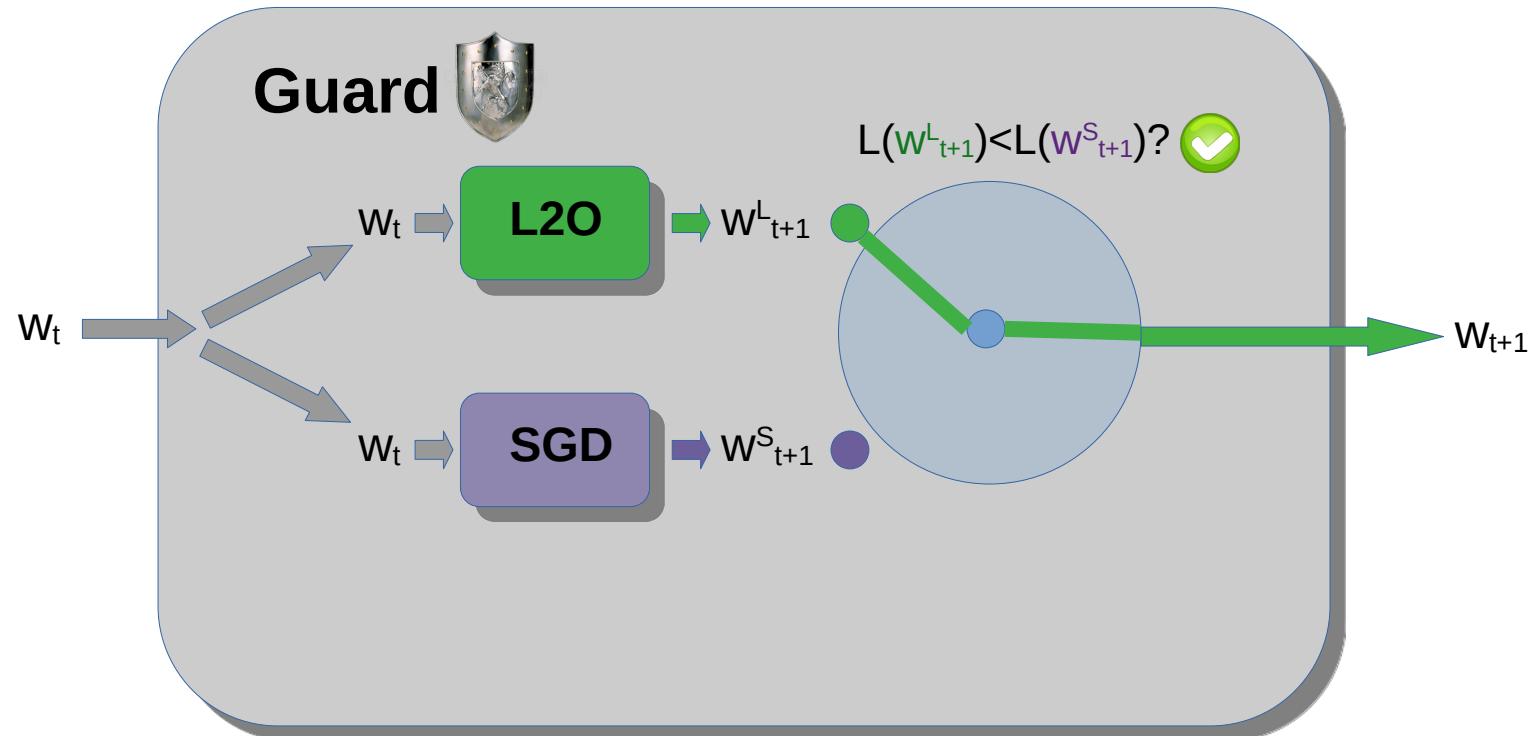
The Solution



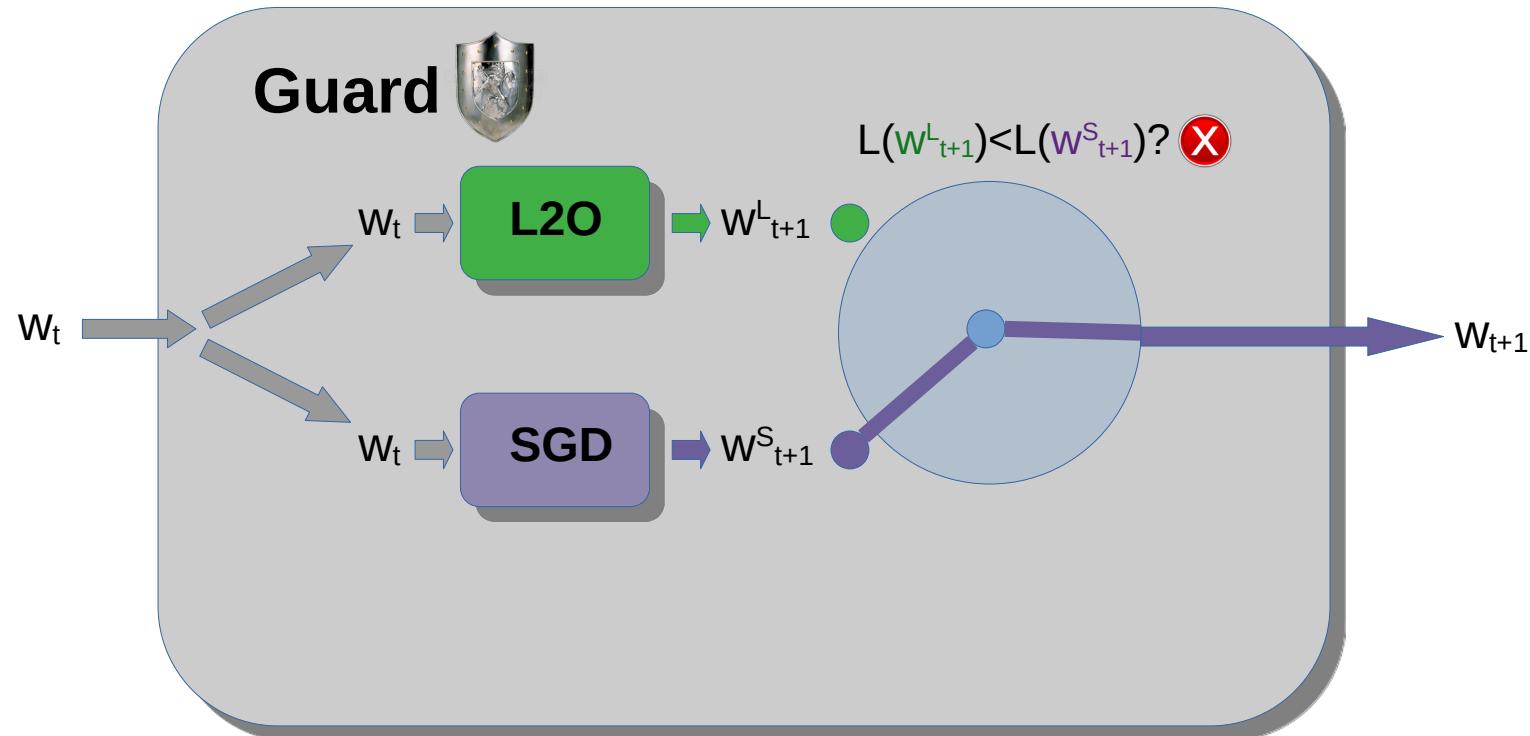
The Solution



The Solution



The Solution



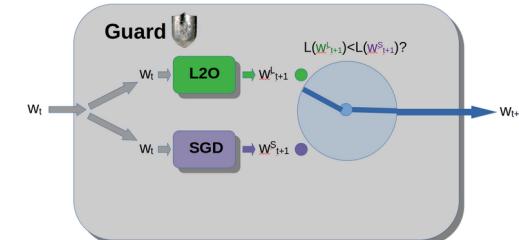
The Solution

Theorem 1 Let \mathcal{F} be a continuous loss function which is μ -strongly convex¹ and L -smooth and let w^* be its global minimum. Let $w_i \in \mathbb{N}$ be a sequence of points obtained from applying the Loss-Guarded L2O algorithm with gradient descent or stochastic gradient as the guarding algorithm. In the case of stochastic gradient descent, we assume that in expectation, the stochastic gradient $\nabla_{mb}\mathcal{F}(w)$ is equal to the true gradient,

$$\mathbb{E}(\nabla_{mb}\mathcal{F}(w)) = \nabla\mathcal{F}(w),$$

and that the variance of the stochastic gradient around the true gradient is bounded. Then given a constant learning rate $0 < \lambda < \min(\frac{2}{L}, 2\mu)$ for gradient descent or a decaying learning rate $\lambda_i \propto \frac{1}{i_0+i}$ for SGD, the sequence converges to the minimum, i.e.

$$\lim_{i \rightarrow \infty} w_i = w^*.$$



Compared To Alternative*

*Heaton, et. al, *Safeguarded Learned Convex Optimization*, arXiv:2003.01880

Compared To Alternative*

- Simpler Algorithm

*Heaton, et. al, *Safeguarded Learned Convex Optimization*, arXiv:2003.01880

Compared To Alternative*

- Simpler Algorithm
- Fewer Computation of Gradients

*Heaton, et. al, *Safeguarded Learned Convex Optimization*, arXiv:2003.01880

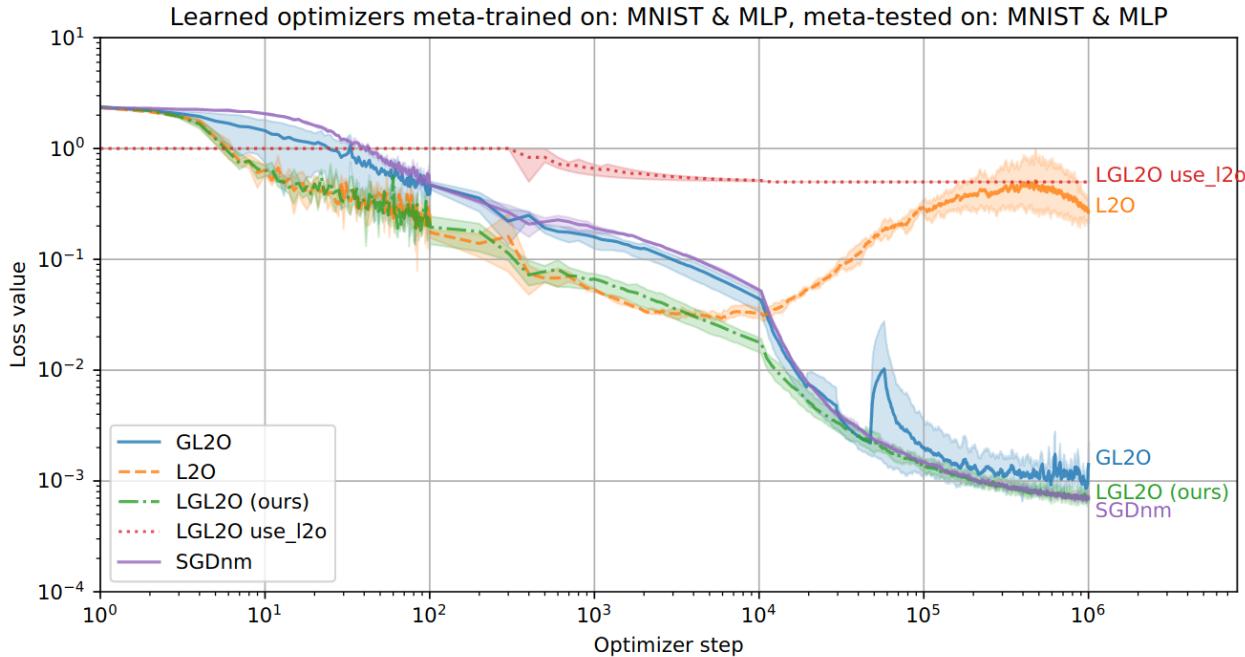
Compared To Alternative*

- Simpler Algorithm
- Fewer Computation of Gradients
- Works better in Practice

*Heaton, et. al, *Safeguarded Learned Convex Optimization*, arXiv:2003.01880

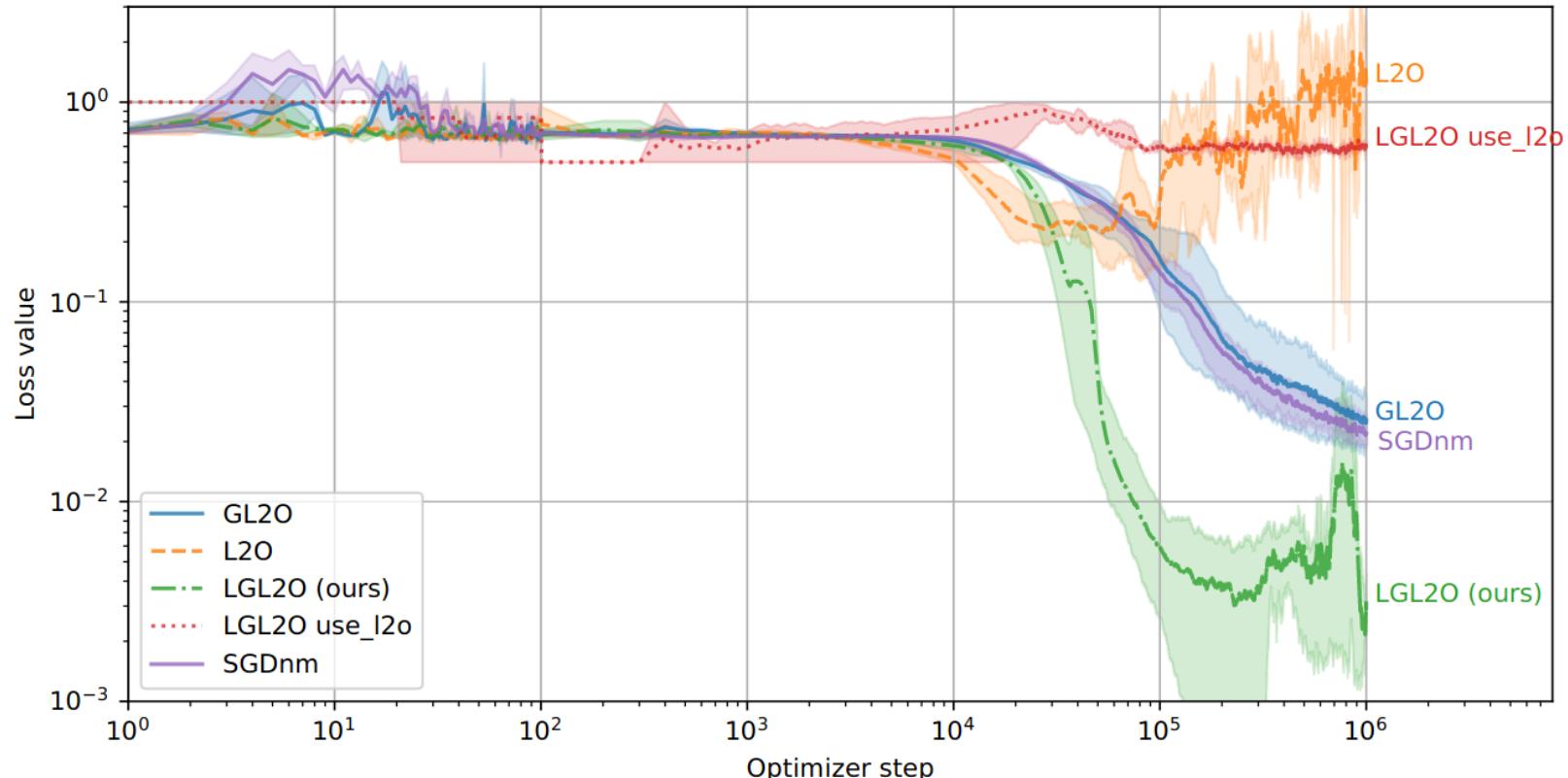
The Results

In Distribution



Out of Distribution

Learned optimizers meta-trained on: MNIST & MLP, meta-tested on: Spirals & MLP



Concluding Remarks

- Guard can be used with any learned optimizer and fallback optimizer

Concluding Remarks

- Guard can be used with any learned optimizer and fallback optimizer
- Inherits the convergence guarantee of the fallback optimizer

Concluding Remarks

- Guard can be used with any learned optimizer and fallback optimizer
- Inherits the convergence guarantee of the fallback optimizer
- Empirically retains the performance of learned optimizer

The End

Thank you

- Come see our poster (#806)
Today 18:00-20:00
- arXiv:2201.12426

