Learning to Learn with Smooth Regularization

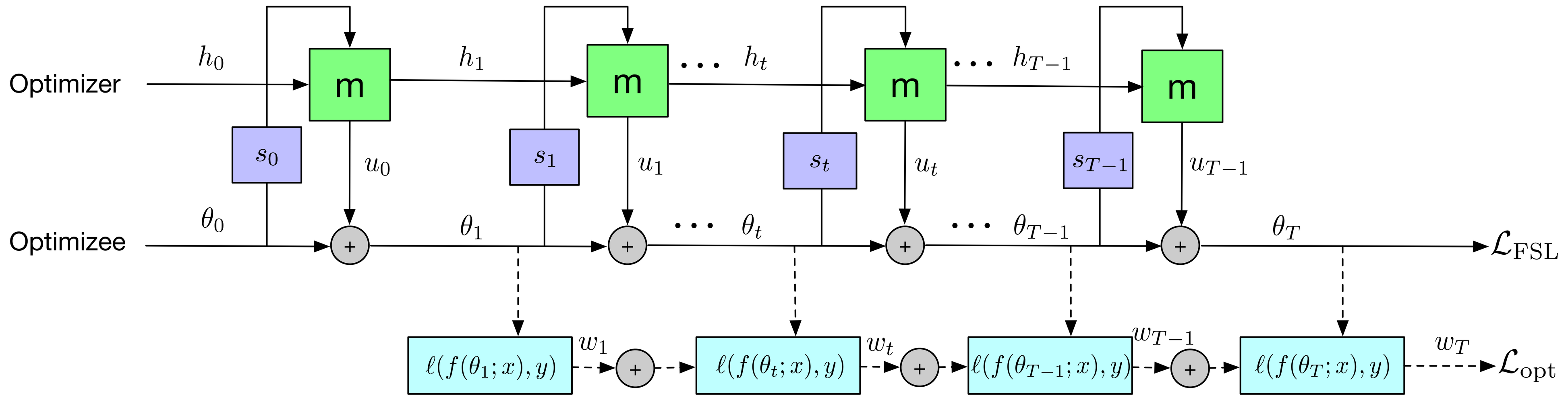
Yuanhao Xiong, Cho-Jui Hsieh

University of California, Los Angeles

TEL AVIV 2022

Background

- Learning to Learn (L2L) aims at an automatic optimization algorithm (optimizer) modeled by neural networks to learn rules for updating the target objective function.

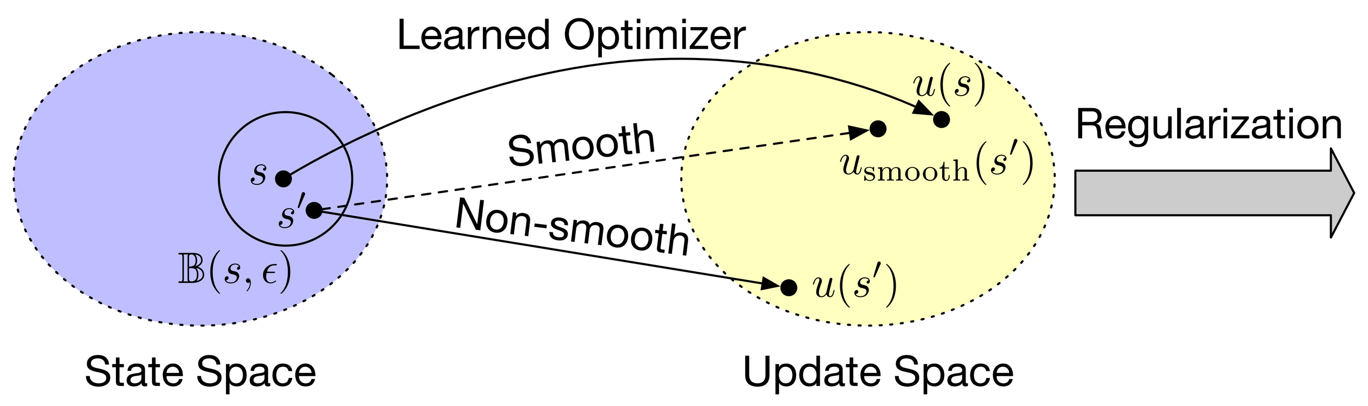


- Unlike hand-engineered algorithms, neural optimizers may suffer from the instability issue: under distinct but similar states, the same neural optimizer can produce quite different updates.

Our solution: Stabilize the neural optimizer with smooth regularization!

Framework

Motivation



$$R = \max_{s' \in \mathbb{B}(s,\epsilon)} d(u(s), u(s')) = \max_{s' \in \mathbb{B}(s,\epsilon)} \|u(s) - u(s')\|^2$$

Key Component

★ Smooth regularizer:
Minimize the gap under the worst-case

$$R = \max_{s' \in \mathbb{B}(s,\epsilon)} ||u(s_t) - u(s_t')||^2$$

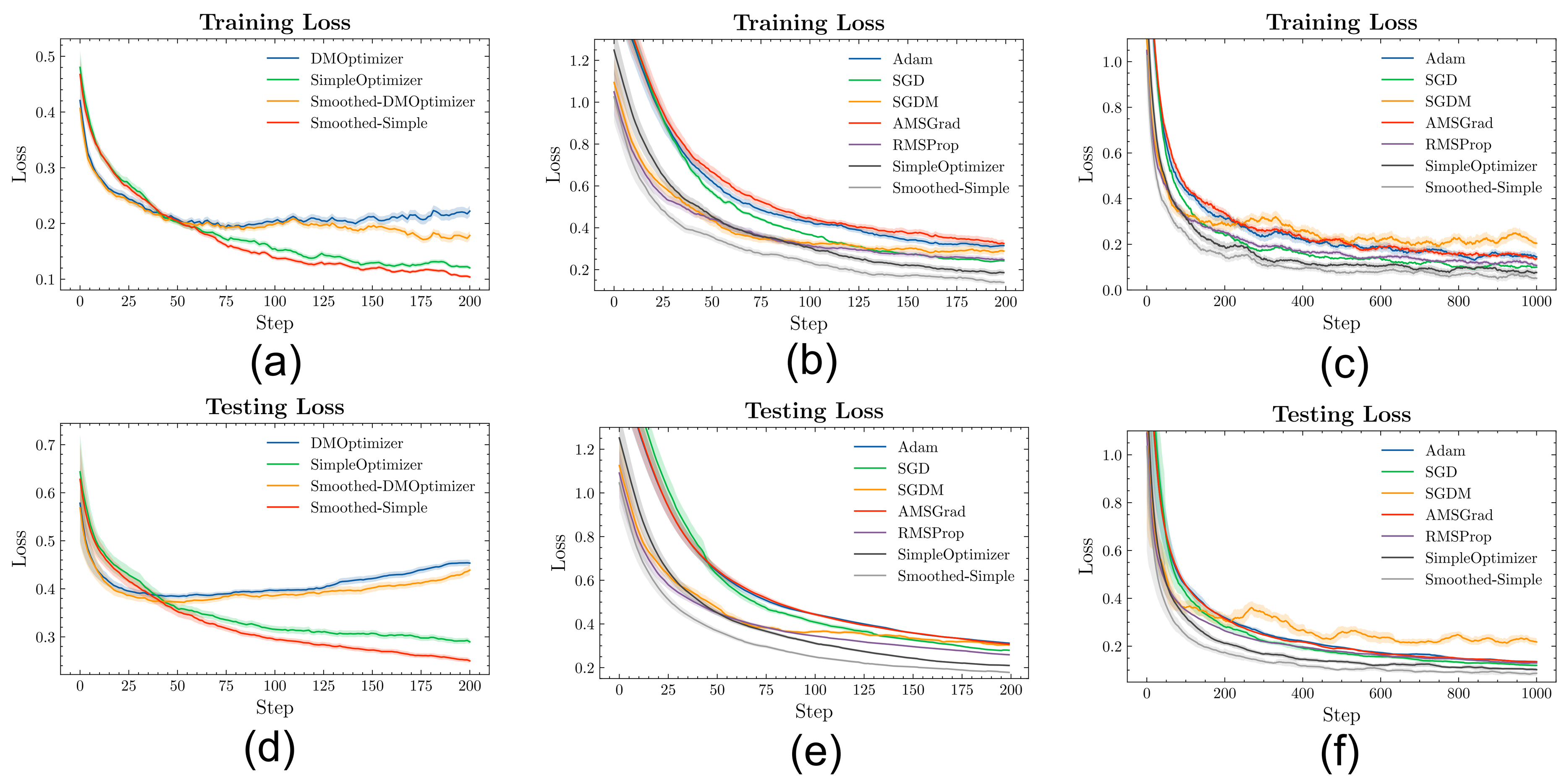
★ Training the optimizer:
Approximate the solution by Projected Gradient Descent

$$s' = \Pi_{\mathbb{B}(s,\epsilon)}(\text{sign}(\nabla_{s'} R) + s')$$

Experiments

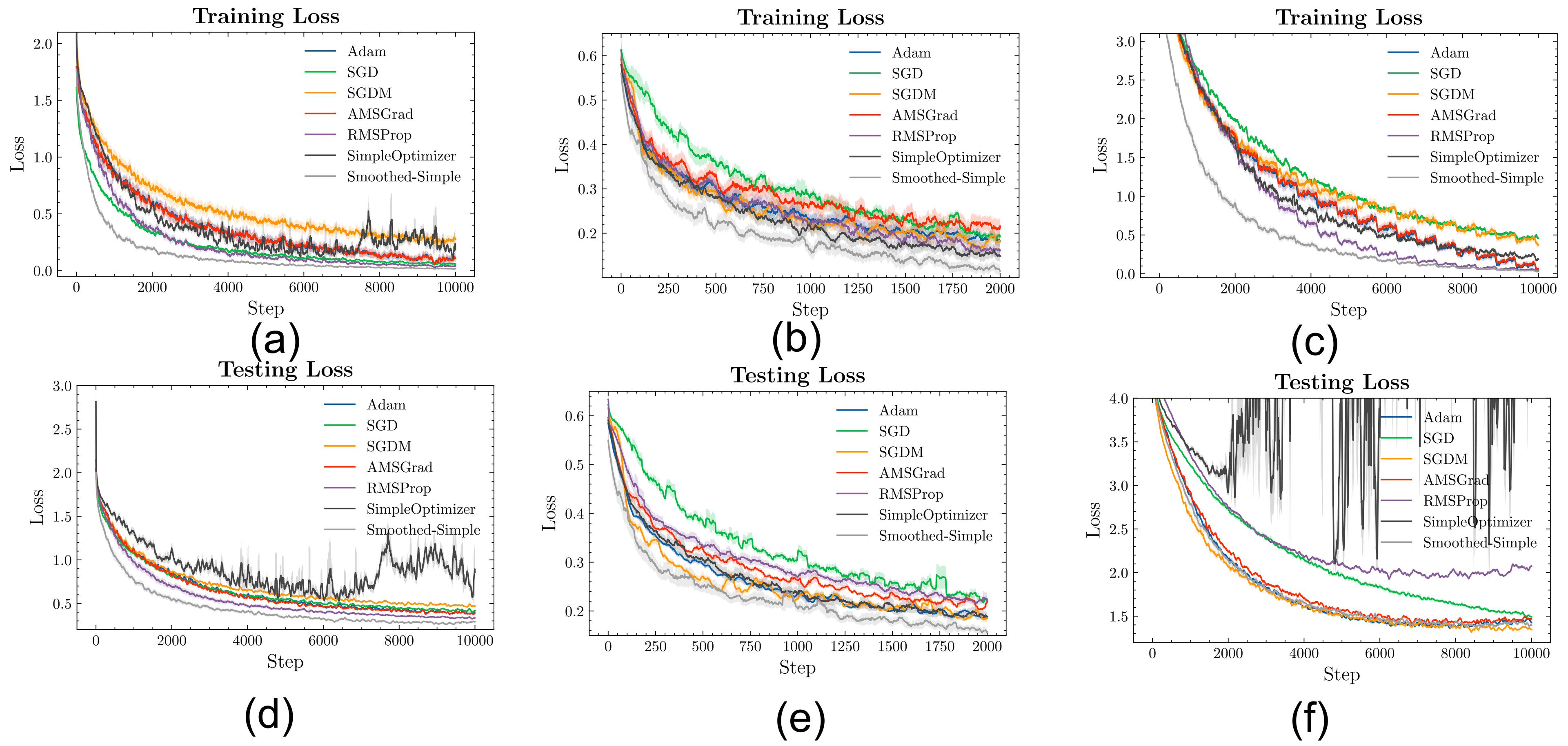
Image Classification

- MNIST



(a) & (d) show the compatibility of our proposed regularizer; (b) & (e) demonstrate performance for training LeNet of 200 steps; (c) & (f) extend it to 1000 steps.

- CIFAR10



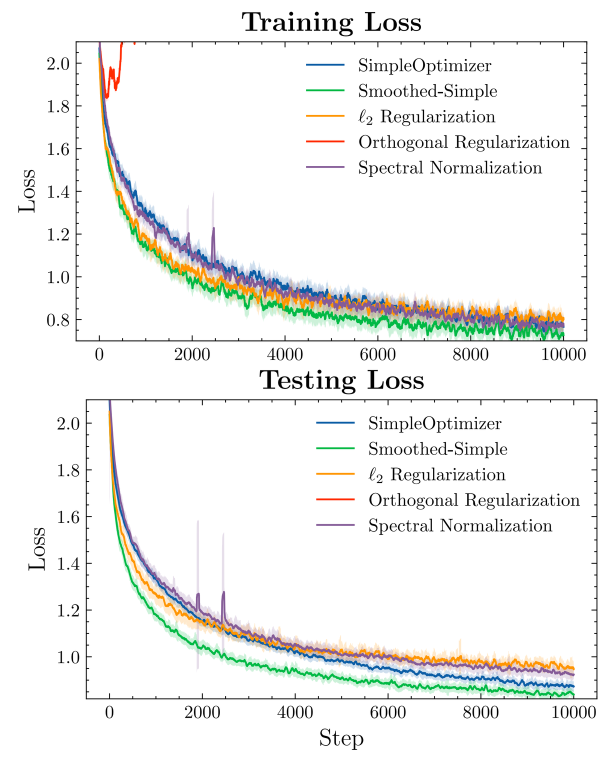
(a) & (d) show 10K-step optimization of ResNet-18; (b) & (e) show transferred binary classification; (c) & (f) train a ResNet-18 on CIFAR100.

Experiments

- Comparison with other regularizers

Table 1. Test accuracy of different regularizers.

Regularizer	**Test accuracy**
SimpleOptimizer	$69.02 \pm 0.58\%$
Simple-Smoothed	$\mathbf{72.50 \pm 0.49\%}$
ℓ_2 Regularization	$69.69 \pm 0.56\%$
Orthogonal Regularization	$10.00 \pm 0.04\%$
Spectral Normalization	$69.35 \pm 1.23\%$



Few-Shot Learning

- Comparison with Meta-LSTM

Table 2. Average accuracy of 5-way few shot learning on miniImageNet and tieredIamgeNet.

Model	miniImageNet 1-shot	miniImageNet 5-shot	tieredImageNet 1-shot	tieredImageNet 5-shot
Meta-LSTM	$38.20 \pm 0.73\%$	$56.56 \pm 0.65\%$	$36.43 \pm 0.65\%$	$53.45 \pm 0.61\%$
Smoothed Meta-LSTM	$\mathbf{40.42 \pm 0.68\%}$	$\mathbf{58.90 \pm 0.61\%}$	$\mathbf{36.74 \pm 0.76\%}$	$\mathbf{55.14 \pm 0.60\%}$

- Comparison with SIB

Table 3. Average accuracy of 5-way few shot learning problems on miniImageNet and CIFAR-FS.

Model	Backbone	miniImageNet 1-shot	miniImageNet 5-shot	CIFAR-FS 1-shot	CIFAR-FS 5-shot
SIB($\eta = 1e^{-3}$, $K = 3$)	WRN-28-10	$69.6 \pm 0.6\%$	$78.9 \pm 0.4\%$	$78.4 \pm 0.6\%$	$85.3 \pm 0.4\%$
Smoothed SIB	WRN-28-10	$\mathbf{70.0 \pm 0.5\%}$	$\mathbf{80.8 \pm 0.3\%}$	$\mathbf{79.2 \pm 0.4\%}$	$\mathbf{86.1 \pm 0.4\%}$

Conclusions

- We propose a regularization term for neural optimizers to enforce similar parameter updates given similar input states.
- Extensive experiments show that the regularizer can be combined with different L2L structures and consolidate its effectiveness on various tasks.
- Training a powerful optimizer that can generalize to longer horizon can be a potential future direction.