Careful with that Scalpel: Improving Gradient Surgery with an EMA

Yu-Guan Hsieh, James Thornton, Eugene Ndiaye, Michal Klein, Marco Cuturi, Pierre Ablin
ICML 2024 · Apple

Introduction

Summary

1. We revisit mixed optimization (i.e., with main + auxiliary losses) as a **simple bilevel problem** and propose **Bloop,** a method closely related to gradient surgery, for solving it.
2. We provide **theoretical justification** and **empirical evidence** to support Bloop, using an important variant that relies on **EMA** in the stochastic setting.

Optimization with Two Losses

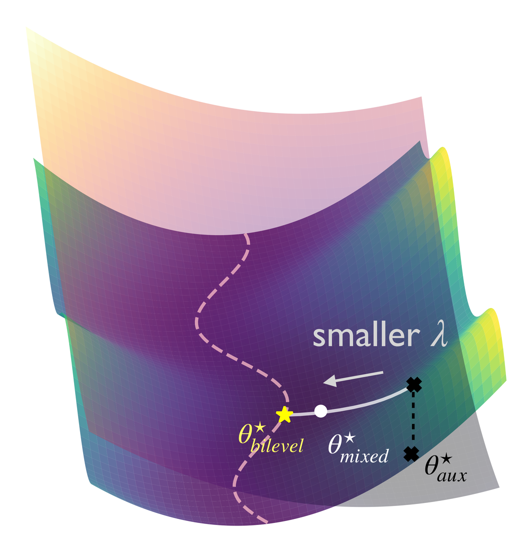
- Main Loss L_{main} : classification, regression, next-token prediction, denoising ...
- Auxiliary Loss L_{aux} : explicit bias (regularization), different dataset, calibration ...

The regularization approach

$$\theta^\star_{mixed}(\lambda) = \operatorname{argmin}_\theta L^\lambda_{mixed}(\theta) = L_{main}(\theta) + \lambda L_{aux}(\theta)$$

The simple bilevel approach ($\lambda \to 0$)

$$\theta^\star_{bilevel} = \operatorname{argmin}_\theta L_{aux}(\theta) \ s.t.\ \theta \in \operatorname{argmin} L_{main}(\theta)$$



Preferred when there is a hierarchy between losses

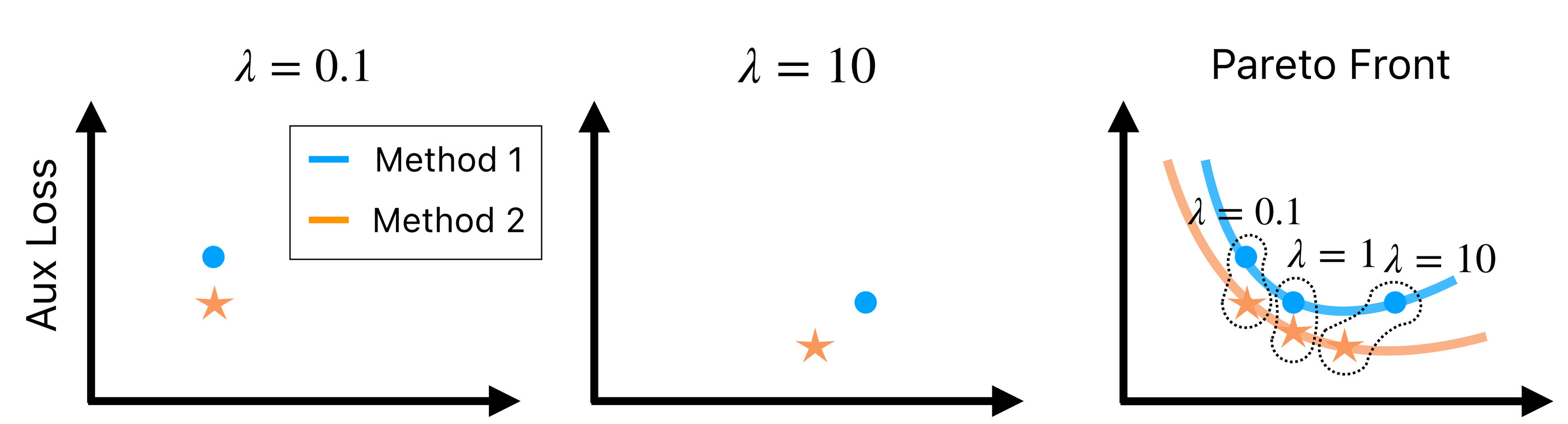
Regularized problems can be hard to optimize

$$\theta = (a, b), \quad L_{main}(\theta) = \frac{1}{2}a^2,\ L_{aux} = \frac{1}{2}((a-1)^2 + b^2)$$

The Hessian of L^λ_{mixed} is $\begin{pmatrix} 1+\lambda & 0 \\ 0 & \lambda \end{pmatrix}$, with conditioning $1 + 1/\lambda \xrightarrow{\lambda \to 0} \infty$

Pareto Front

Goal: Find methods that better trade-off the two losses
We use a hyperparameter λ to control the trade off

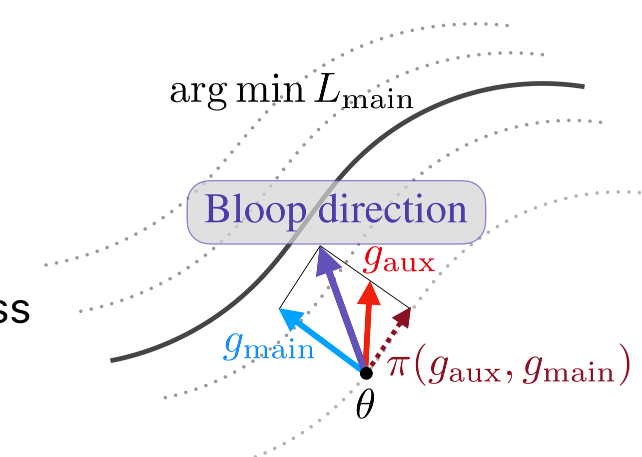


Algorithm

Bloop: BiLevel Optimization with Orthogonal Projection

Consider the following direction

$$d = g_{main} + \lambda\pi(g_{aux}, g_{main})$$
$$where \quad \pi(g_{aux}, g_{main}) = g_{aux} - \frac{\langle g_{aux}, g_{main}\rangle}{\| g_{main} \|^2} g_{main}$$



The projection guarantees descent of the main loss function when learning rate is small

$$L_{main}(\theta - \eta d) \simeq L_{main}(\theta) - \eta \| g_{main} \|^2$$

Theorem [Small Bloop direction iff Near-Stationary point]

- If d is small, there exists vector v such that both $\| g_{main} \|$ and $\| g_{aux} - \nabla^2 L_{main}(\theta)\ v \|$ are small
- For any first-order stationary point $\theta^\star$, we have $\lim_{\varepsilon\to 0} d(\theta^\star + \varepsilon v) = 0$ where v is the Lagrange multiplier

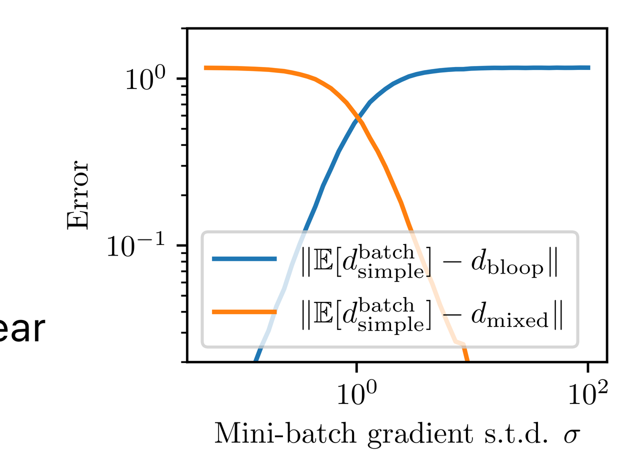
Extension to the Stochastic Setting

In practice, $\mathbb{E}[d^{batch}_{main}] = d_{main}$ and $\mathbb{E}[d^{batch}_{aux}] = d_{aux}$

The naive solution

$$d^{batch}_{simple} = d^{batch}_{main} + \lambda\pi(g^{batch}_{aux}, g^{batch}_{main})$$

Issue: $\pi(g^{batch}_{aux}, g^{batch}_{main})$ is *biased*: in expectation collinear with g_{aux} in infinite noise regime



The EMA solution

$$d^{batch} = d^{batch}_{main} + \lambda\pi(g^{batch}_{aux}, g^{EMA}_{main}) \quad where \quad g^{EMA}_{main} \leftarrow (1-\rho) g^{EMA}_{main} + \rho g^{batch}_{main}$$

Theorem [Convergence of average gradient norm]

Fix any time horizon T, by choosing $\eta \simeq T^{-\frac{3}{4}}$ and $\rho \simeq \eta^{\frac{2}{3}}$, we have

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\| \nabla L_{main}(\theta^t) \|^2] = O(T^{-\frac{1}{4}})$$

Experiments

Baselines

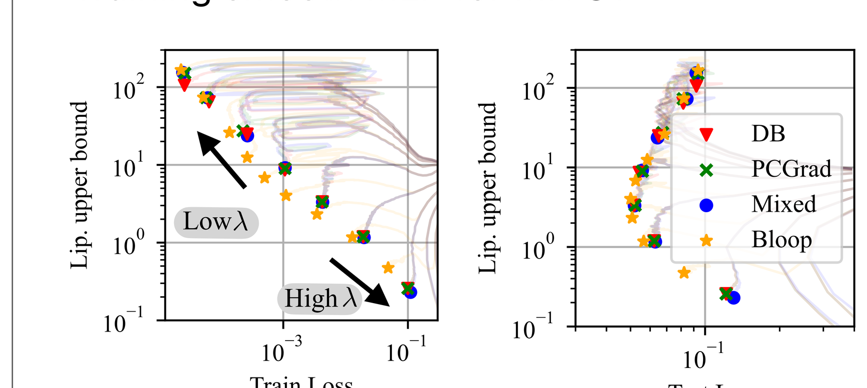
1. Mixed (i.e., SGD on the mixed loss)
   $d = g^{batch}_{main} + \lambda g^{batch}_{aux}$
2. Dynamic Barrier (DB) [Gong & Liu, 2021]
   $d = \max(1 - \lambda(\langle g^{batch}_{main}, g^{batch}_{aux}\rangle / \| g^{batch}_{main} \|^2), 0)\ g^{batch}_{main} + \lambda g^{batch}_{aux}$
3. PCGrad [Yu et al., 2020] : Project only when gradients are in conflict
   - $d = g^{batch}_{main} + \lambda g^{batch}_{aux}$ if $\langle g^{batch}_{main}, g^{batch}_{aux}\rangle > 0$
   - $d = \pi(g^{batch}_{main}, g^{batch}_{aux}) + \lambda\pi(g^{batch}_{aux}, g^{batch}_{main})$ if $\langle g^{batch}_{main}, g^{batch}_{aux}\rangle \leq 0$

Setup

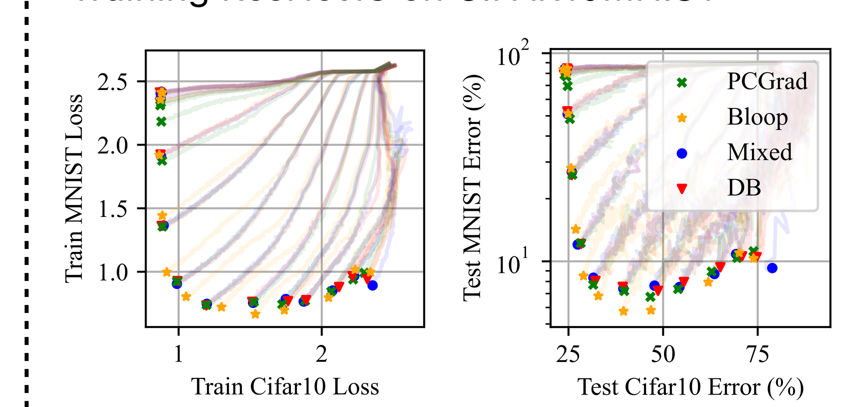
Task	Main Loss / Dataset	Auxiliary Loss / Dataset
Training smooth MLP For MNIST	Classification loss	Logarithm of Lipschitz upper bound of network
Training ResNet18 on CIFAR10MNIST	Classification loss for background CIFAR-10 image	Classification loss for foreground MNIST digit
LM pre-training	30M examples from c4	20k examples from RCV-1
EN-DE translation	36M samples from Paracrawl	10k examples WMT 09 - 19

Results

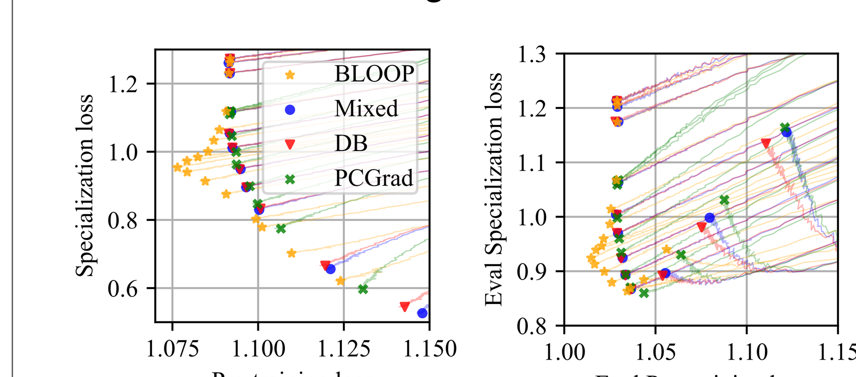
Training smooth MLP for MNIST



Training ResNet18 on CIFAR10MNIST



Joint dataset training for EN-DE translation



Joint dataset training for LM pre-training