SHINE: SHaring the INverse Estimate for bi-level optimization

Thomas Moreau
Université Paris-Saclay, Inria, CEA, Palaiseau, 91120, France
thomas.moreau@inria.fr

In recent years, bi-level optimization has raised much interest in the machine learning community, in particular for hyper-parameters optimization [4] and implicit deep learning [1]. Bilevel optimization aims at minimizing a function whose value depends on the result of another optimization problem, that is:

$$\begin{aligned} &\min_{x\in\mathbb{R}^d} h(x) = F(z^*(x), x) \ , \\ &\text{such that } z^*(x) \in \arg\min_{z\in\mathbb{R}^p} G(z, x) \ , \end{aligned} \tag{1}$$

where F and G are two real valued functions defined on $\mathbb{R}^p \times \mathbb{R}^d$. This type of problems is often tackled using first-order that requires the computation of the gradient of h, whose expression can be obtained using the implicit function theorem: $\nabla h(x) = \nabla_2 F(z^*(x), x) - \nabla^2_{2,1} G(z^*(x), x) \left[\nabla^2_{1,1} G(z^*(x), x)\right]^{-1} \nabla_1 F(z^*(x), x)$. The computation of this gradient requires the computation of matrix-vector products involving the inverse of a large matrix $\nabla^2_{1,1} G$, which is computationally demanding.

In our work [5], we propose a novel strategy coined SHINE to tackle this computational bottleneck when the inner problem G can be solved with a quasi-Newton algorithm. The main idea is to use the quasi-Newton matrices estimated from the resolution of the inner problem to efficiently approximate the inverse matrix in the direction needed for the gradient computation $\left[\nabla^2_{1,1} G(z^*(x), x)\right]^{-1} \nabla_1 F(z^*(x), x)$. We prove that under some restrictive conditions, this strategy gives a consistent estimate of the true gradient. In addition, by modifying the quasi-Newton updates, we provide theoretical guarantees that our method asymptotically estimates the true implicit gradient under weaker hypothesis.

Figure 1 shows on a classical hyperparameter optimization benchmark [4] that our method accelerate the resolution of the bi-level problem compare to HOAG [4] and the Jacobian-Free method that replace the inverse by the identity. Experiments for multi-scale Deep-Equilibrium networks (DEQ [2]) in [5] applied to CIFAR10 and ImageNet show that SHINE reduces the computational cost of the backward pass by up to two orders of magnitude, while retaining performances close to the original training methods. While these results are encouraging, our method still suffer from small performance drop on DEQ for ImageNet, leaving room for further improvement.

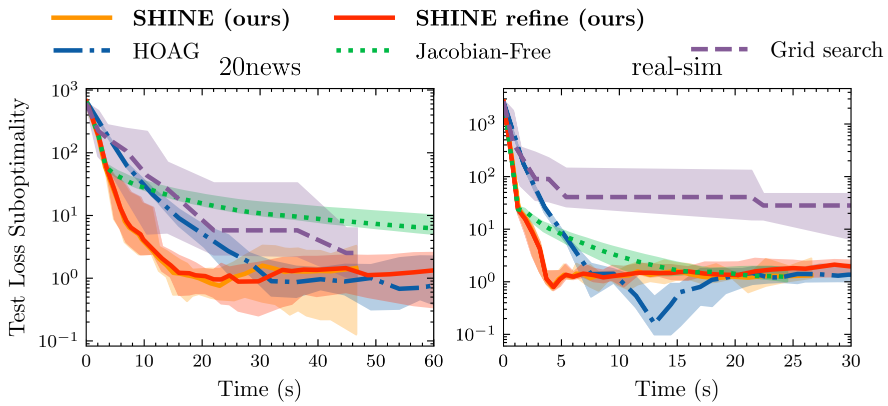

Figure 1: Convergence of test loss for different hyperparameter optimization methods on the ℓ_2-regularized logistic regression problem for the 2 datasets (20news and real-sim).

Joint work with: Z. Ramzi, S. Bai, F. Mannel, J.-L. Stark and P. Ciuciu.

References

[1] S. Bai, J. Kolter and V. Koltun. Deep Equilibrium Models. *NeurIPS*, 2019.

[2] S. Bai, V. Koltun and J. Kolter. Multiscale deep equilibrium models. *NeurIPS*, 2020.

[3] S. Fung, H. Heaton, Q. Li, D. Mckenzie, S. Osher, and W. Yin. Fixed Point Networks: Implicit Depth Models with Jacobian-Free Backprop. preprint ArXiv, 2021

[4] F. Pedregosa. Hyperparameter optimization with approximate gradient. *ICML*, 2016.

[5] Z. Ramzi, F. Mannel, S. Bai, J.-L. Starck, P. Ciuciu, T. Moreau. SHINE: SHaring the INverse Estimate from the forward pass for bi-level optimization and implicit models. *ICLR*, 2022.