FORMATTING INSTRUCTIONS FOR ICLR 2023 CONFERENCE SUBMISSIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

TODO

1 INTRODUCTION

TODO

2 METHOD

2.1 BACKGROUND

We approach the problem of gradient-based hyperparameter optimization as follows. Given a bilevel optimization problem

$$\boldsymbol{\alpha}^* = \arg\min_{\boldsymbol{\alpha}} \mathcal{L}_2(\boldsymbol{w}^*, \boldsymbol{\alpha}), \tag{1}$$

$$\text{s.t.} \quad \boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} \mathcal{L}_1(\boldsymbol{w}, \boldsymbol{\alpha}). \tag{2}$$

TODO: introduce the functions and the variables.

Suppose that the lower level problem equation 1 is solved by an optimizer which takes the following form

$$\boldsymbol{w}_{t+1}(\boldsymbol{\alpha}) = \Phi(\boldsymbol{w}_t, \boldsymbol{\alpha}). \tag{3}$$

Now we derive a gradient of $\mathcal{L}_2(\boldsymbol{w}_T(\boldsymbol{\alpha}), \boldsymbol{\alpha})$ w.r.t. $\boldsymbol{\alpha}$. In fact, it could be done by using the chain rule:

$$\underbrace{\frac{d\mathcal{L}_2}{d\boldsymbol{\alpha}} = \frac{\partial \mathcal{L}_2}{\partial \boldsymbol{\alpha}}}_{\mathbf{g}_{\text{FO}}} + \underbrace{\sum_{t=1}^{T} \mathbf{B}_t \mathbf{A}_{t+1} \dots \mathbf{A}_T \frac{\partial \mathcal{L}_2(\mathbf{w}_T, \boldsymbol{\alpha})}{\partial \mathbf{w}}}_{\mathbf{g}_{\text{SO}}}, \tag{4}$$

where $\mathbf{B}_t = \frac{\partial \Phi(\mathbf{w}_{t-1}, \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}$, and $\mathbf{A}_t = \frac{\partial \Phi(\mathbf{w}_{t-1}, \boldsymbol{\alpha})}{\partial \mathbf{w}}$. However, true hypergradient computation is costly in terms of memory (citation).

2.2 HYPERGRADIENT COMPUTATION

In this section we provide an approximation of the true hypergradient and perform analysis of approximation exactness.

Consider the t-th step of the inner optimization performed by a gradient descent, i.e. $\Phi(\mathbf{w}_t, \alpha) = \mathbf{w}_t - \eta \frac{\partial \mathcal{L}_1(\mathbf{w}, \boldsymbol{\alpha})}{\partial \mathbf{w}}$. We motivate the approximation by the fact that we aimed to approximate the t-th term of $\mathbf{g}_{\text{SO}}$ equation 4. Since $\mathbf{A}_{t+1}$ are unknown at the timestamp t, we approximate the product $\mathbf{A}_{t+1} \dots \mathbf{A}_T \approx \gamma^{T-t}\mathbf{I}$, where $\gamma \in \mathbb{R}_+$. Additionally, for the same reason we approximate the gradient at the last timestamp $\frac{\partial \mathcal{L}_2(\mathbf{w}_T, \boldsymbol{\alpha})}{\partial \mathbf{w}} \approx \frac{\partial \mathcal{L}_2(\mathbf{w}_t, \boldsymbol{\alpha})}{\partial \mathbf{w}}$. Finally, the proposed approximation is as follows

$$\hat{\mathbf{g}}_{\text{SO}} = \sum_{t=1}^{T} \mathbf{B}_t \frac{\partial \mathcal{L}_2(\mathbf{w}_t, \boldsymbol{\alpha})}{\partial \mathbf{w}} \gamma^{T-t}. \tag{5}$$

Note that in effect the approximation equation 5 is a moving average of greedy hypergradients (cite).

Now we provide analysis of exactness of the provided approximation. Before doing this, we formulate a list of assumptions:

Assumption 1

1. *Let $\mathcal{L}_1(.,\boldsymbol{\alpha})$ and $\mathcal{L}_2(.,\boldsymbol{\alpha})$ be L-smooth and μ-strongly convex for any $\boldsymbol{\alpha}$*
2. *Let $\frac{\partial^2\mathcal{L}_1(.,\boldsymbol{\alpha})}{\partial\mathbf{w}\partial\mathbf{w}^\top}$ be H_w-Lipschitz for any $\boldsymbol{\alpha}$.*
3. *Let $1-\eta L \le \gamma \le 1-\eta\mu$*
4. *$\|\frac{\partial\mathcal{L}_1(\mathbf{w},\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}\partial\mathbf{w}^\top}\| \le B$ for any pair $(\mathbf{w},\boldsymbol{\alpha})$.*

Theorem 1 *Let $\mathcal{L}_1(.,.)$ and $\mathcal{L}_2(.,.)$ satisfy Assumption 1. Then the following is true*

$$\|\mathbf{g}_{SO}-\hat{\mathbf{g}}_{SO}\| \le \frac{2LB\|\mathbf{w}_0-\mathbf{w}_*\|\sqrt{1-\eta\mu}^T}{\sqrt{1-\eta\mu}^{-1}-1}+$$

$$B\left\|\frac{\mathcal{L}_2(\mathbf{w}_T,\boldsymbol{\alpha})}{\partial\mathbf{w}}\right\|\left\{\frac{1}{\eta}\left(\frac{1}{\mu}-\frac{1}{L}+\frac{1}{L}(1-\eta\mu)^T\right)+2\eta H_w\left[(T-1)\sqrt{1-\eta\mu}^T-\frac{\sqrt{1-\eta\mu}^{T-1}-(1-\eta\mu)^T}{\sqrt{1-\eta\mu}^{-1}-1}\right]\right\}$$

Note that in a general case the approximated gradient is inexact. Notably, when $L=\mu$ and thus $H_w=0$, we get a simplified upper bound:

$$\|\mathbf{g}_{\mathrm{SO}}-\hat{\mathbf{g}}_{\mathrm{SO}}\| \le \frac{2LB\|\mathbf{w}_0-\mathbf{w}_*\|\sqrt{1-\eta\mu}^T}{\sqrt{1-\eta\mu}^{-1}-1}. \tag{6}$$

Now we are trying to prove the fact that the bound vanishes if the number of solved inner optimizations tends to infinity. The proof has not been completed yet.

Theorem 2 *In the assumptions above $\|\mathbf{w}_0^{(k)}-\mathbf{w}_*^k\| \to_{k\to\infty} 0$, where k is the current epoch. Therefore, the proposed hypergradient is assimptotically exact under the mentioned assumptions.*

Furthermore, we provide analysis of a sufficient descent direction. We first formulate an additional list of assumptions.

Assumption 2

1. *Let $\mathcal{L}_1(.,\boldsymbol{\alpha})$ and $\mathcal{L}_2(.,\boldsymbol{\alpha})$ be M-Lipschitz for any $\boldsymbol{\alpha}$.*
2. *Let $\frac{\partial\mathcal{L}_1(.,\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}\partial\mathbf{w}^\top}$ is M_b-Lipschitz for any $\boldsymbol{\alpha}$*
3. *Let $(\frac{\partial\mathcal{L}_1(.,\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}\partial\mathbf{w}^\top})^\top\frac{\partial\mathcal{L}_1(.,\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}\partial\mathbf{w}^\top} \succeq \kappa\mathbf{I}$ for any $(\mathbf{w},\boldsymbol{\alpha})$*

Lemma 1 *Given an gradient descent update of the outer optimization with a learning rate η_{out}. Suppose that we start the whole optimization from $(\mathbf{w}_0,\boldsymbol{\alpha}_0)$. Suppose that Assumption 1 and the first point of Assumption 2 hold. Let also the following conditions are met:*

1. *There esist $\delta>0$ such that $\mathcal{L}_2(\mathbf{w}_*,\boldsymbol{\alpha})-\mathcal{L}_2^*-M\|\mathbf{w}_0-\mathbf{w}_*\|\sqrt{1-\eta\mu}^T \ge \delta$ for any $\boldsymbol{\alpha}$*
2. *$\sqrt{1-\eta\mu}^T \le 1/2$*
3. *$\eta_{out} \le \frac{\|\mathbf{w}_0-\mathbf{w}_*\|}{2}\left(\frac{LB^2}{(1-\eta L)^2}\right)^{-1}$*

Then the following is true for any $\boldsymbol{\alpha}$

$$\left\|\frac{\partial\mathcal{L}_2(\mathbf{w}_T,\boldsymbol{\alpha})}{\partial\mathbf{w}}\right\| \ge \sqrt{\mu\delta}. \tag{7}$$

Theorem 3 *Let Assumptions 1, 2 are satisfied. Additionally, let there exist large enough $\delta>0$. Then there exist $c>0$ such that*

$$\mathbf{g}_{SO}^\top\hat{\mathbf{g}}_{SO} \ge c\left\|\frac{\partial\mathcal{L}_2(\mathbf{w}_T,\boldsymbol{\alpha})}{\partial\mathbf{w}}\right\|^2. \tag{8}$$

In other words, if the outer function does not depend on the hyperparameters $\mathcal{L}_2 = \mathcal{L}_2(\mathbf{w})$, then the proposed approximation $\hat{\mathbf{g}}_{SO}$ is sufficient descent condition.

I was inspired by "Truncated Back-propagation for Bilevel Optimization" when formulating and proving this theorem.

3 EXPERIMENTS

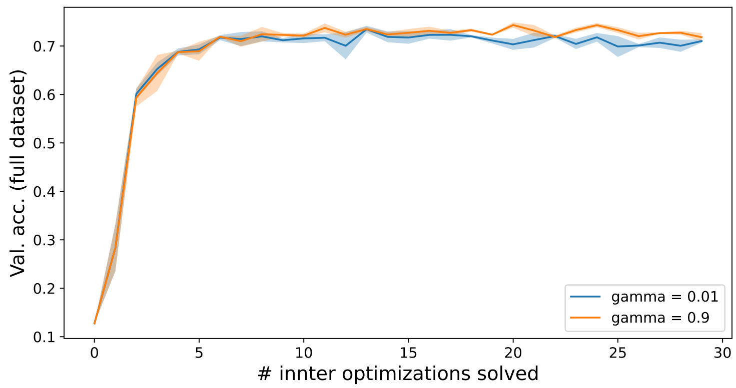

Figure 1: Experimental results on data hypercleaning.

Method	Valid. Acc.	#JVPs
Truncated backpropagation (Lukethina)	72.5	$1(1)$
DrMAD	69.8	$99(2T-1)$
IFT(9, 5)	70.3	$50((N+1)K)$
IFT(4, 10)	70.7	$50((N+1)K)$
Proposed ($\gamma = 0.99$)	**73.5***	$50(T)$

Table 1: Experimental results for data hypercleaning. * indicates that the results are statistically significant ($p < 0.05$). Note $\mathrm{IFT}(N, K)$ denotes that we performed K-step online optimization with N first terms of Neuman series.

REFERENCES

A APPENDIX

You may include other additional sections here.