Generalized Greedy Gradient-Based Hyperparameter Optimization

Konstantin Yakovlev
MIPT
Moscow, Russia
iakovlev.kd@phystech.edu

Abstract

TODO

1 Introduction

Bilevel optimization has become an essential component of machine learning, which includes Neural Architecture Search [16, 23, 31], Hyperparameter Optimization [12], and Meta-Learning [11, 21, 6]. In the hierarchical optimization framework, the outer-level objective is aimed to be minimize given the optimality in the inner level. Solving the bilevel problem is challenging due to the intricate dependency of the optimal inner parameters given the outer parameters.

Naive approaches such as random search and grid search [3] become impractical with the growing number of hyperparameters to be optimized due to the curse of dimensionality. Another approach that has proven effective is Bayesian Optimization [29].

In the current work we develop a novel gradient-based algorithm [2]. The challenge is that the exact hypergradient calculation is computationally demanding [7]. Specifically, Forward-Mode differentiation is memory demanding, since it increases linearly with the number of hyperparameters. This limits the application of the method for large-scale problems with millions of hyperparameters, such as meta-learning. By contrast, Revers-Mode Differentiation perfectly scales to problems with millions of hyperparameters, but it requires the full inner optimization trajectory of model parameters to be saved, which is computationally costly. Moreover, RMD suffers from gradient vanishing or explosion [1], which leads to training instability. Truncation of the optimization trajectory was proposed to alleviate high memory consumption [28] while calculating an approximate hypergradient. TODO: state that this family is not suitable for online optimization. However, this approach suffers from short horizon bias [30]. TODO: few words about it.

Alternatively, an implicit differentiation may be used to compute the hypergradient [17, 18, 22]. This approach mitigates the need for unrolling, but it heavily relies on Implicit Function Theorem, which requires the convergence of the inner optimization [8, 4]. The challenge of this family of methods is computing inverse hessian-vector product. This computation may be approximated with Neumann series [17] or conjugate gradients [22].

In this paper, we propose an alternative approach to hypergradient computation. Namely, the proposed approach resolves the following issues simultaneously: short horizon bias, high memory requirements, applicability to large-scale problems with millions of hyperparameters, and independence of inner optimization convergence. Overall, our contributions are as follows:

1. we introduce a procedure that aggregates the greedy gradients calculated at each iteration of the inner objective, which satisfies the requirements above.
2. We provide a theoretical analysis of the proposed approach. Under some assumptions, a sufficient descent condition holds.

Preprint. Under review.

3. We empirically prove the effectiveness of the proposed approach on Meta-Learning and Data Hyper-Cleaning tasks.

2 Related Work

Gradient-Based Hyperaprameter Optimization. Differentiation through optimization [5] was successfully applied to hyperparameter optimization at a large-scale [19]. The unrolled differentiation could be categorized into Forward-Mode and Reverse-Mode differentiation [7]. The former one suits best for the cases when a handful of hyperparameters is needed to be optimized [20], for instance, learning rate and weight dacay. The latter is suitable for the setup with millions of hyperparameters while sacrificing the memory consumption when the number of inner optimization steps increases, except for the cases when SGD with momentum is used [19]. Additionally, truncated unrolled differentiation [28] introduces a trade-off between computational complexity and hypergradient accuracy. However, computations done on truncated trajectories suffer from short horizon bias [30].

Alternatively, impicit differentiation, inspired by the Implicit Function Theorem, is used to compute the hypergradient [2, 17, 22, 18]. In [2] an exact inverse hessian is computed, which is computationally intractable in huge-scale scenario with millions of model parameters. To sidestep this issue, an approximation is needed. Specifically, the Neumann series approximation [17], conjugate gradients [22], GMRES [4] for solving linear systems, Nyström method [10], and Broyden's method [9]. The major limitation is that the near-optimality of the inner optimization is crucial for accurate approximation of the true hypergradient [8, 4]. Moreover, the method is inapplicable to tackling the optimizer hyperaprameters such as learning rate.

Meta-Learning. Another fundamental application of bilevel optimization is meta-learning [27] (or learning to learn). It aims to train a model that generalizes well over the distribution of tasks [26]. In the context of gradient-based model-agnostic meta-learning [6], the task is to learn an initialization of model parameters such that gradient-based fine-tuning shows good generalization. MAML optimization successfully inherts the methods for hypergradient computation. Specifically, [15] successfully employed [18], [25] used implicit differentiation with conjugate gradient algorithm.

3 Background

TODO: describe the section

3.1 Hypergradient computation

Given a vector of model parameters $\mathbf{w} \in \mathbb{R}^P$ and a vector of hyperparameters $\boldsymbol{\alpha} \in \mathbb{R}^H$. The dynamic of model parameters $\{\mathbf{w}_t\}_{t=0}^T$ for some $T \in \mathbb{N}$ and some $\boldsymbol{\alpha}$ is defined as follows $\mathbf{w}_{t+1} = \boldsymbol{\Phi}(\mathbf{w}_t, \boldsymbol{\alpha})$, where $\boldsymbol{\Phi}(.,.)$ is a smooth mapping. For instance, a vanilla gradient descent with stepsize $\eta > 0$ could be written as $\boldsymbol{\Phi}(\mathbf{w}_t, \boldsymbol{\alpha}) = \mathbf{w}_t - \eta \nabla_{\mathbf{w}} \mathcal{L}_{\text{train}}(\mathbf{w}_t, \boldsymbol{\alpha})$, where $\mathcal{L}_{\text{train}}$ is a training loss function. Given also a validation loss function $\mathcal{L}_{\text{val}}(\mathbf{w}, \boldsymbol{\alpha})$. Under the notations above we formulate a hyperparameter optimization problem as follows:

$$\boldsymbol{\alpha}^* - \arg\min_{\boldsymbol{\alpha} \in \mathbb{R}^H} \mathcal{L}_{\text{val}}(\mathbf{w}_T, \boldsymbol{\alpha}), \tag{1}$$

$$\text{s.t.} \quad \mathbf{w}_t = \boldsymbol{\Phi}(\mathbf{w}_{t-1}, \boldsymbol{\alpha}), \quad t \in \overline{1, T}. \tag{2}$$

Now the goal is to derive a hypergdadient $d_{\boldsymbol{\alpha}} \mathcal{L}_{\text{val}}(\mathbf{w}_T, \boldsymbol{\alpha})$, viewing $\mathbf{w}_T$ as a function of $\boldsymbol{\alpha}$:

$$d_{\boldsymbol{\alpha}} \mathcal{L}_{\text{val}}(\mathbf{w}_T, \boldsymbol{\alpha}) = \nabla_{\boldsymbol{\alpha}} \mathcal{L}_{\text{val}}(\mathbf{w}_T, \boldsymbol{\alpha}) + \nabla_{\mathbf{w}_T} \mathcal{L}_{\text{val}}(\mathbf{w}_T, \boldsymbol{\alpha}) \frac{d\mathbf{w}_T}{d\boldsymbol{\alpha}}. \tag{3}$$

Here $\nabla_{\boldsymbol{\alpha}} \mathcal{L}_{\text{val}}(\mathbf{w}_T, \boldsymbol{\alpha})$ is a row-vector. The chain rule suggests that $d\mathbf{w}_T / d\boldsymbol{\alpha}$ is computed in the following way [7]:

$$\frac{d\mathbf{w}_T}{d\boldsymbol{\alpha}} = \sum_{t=1}^{T} \left(\prod_{k=t+1}^{T} \mathbf{A}_k \right) \mathbf{B}_t, \quad \mathbf{A}_k = \frac{\partial \boldsymbol{\Phi}(\mathbf{w}_{k-1}, \boldsymbol{\alpha})}{\partial \mathbf{w}_{k-1}}, \quad \mathbf{B}_t = \frac{\partial \boldsymbol{\Phi}(\mathbf{w}_{t-1}, \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}. \tag{4}$$

Therefore, the hypergradient is calculated as follows:

$$d_{\boldsymbol{\alpha}}\mathcal{L}_{\text{val}}(\mathbf{w}_T, \boldsymbol{\alpha}) = \nabla_{\boldsymbol{\alpha}}\mathcal{L}_{\text{val}}(\mathbf{w}_T, \boldsymbol{\alpha}) + \sum_{t=1}^{T} \nabla_{\mathbf{w}_T}\mathcal{L}_{\text{val}}(\mathbf{w}_T, \boldsymbol{\alpha}) \left(\prod_{k=t+1}^{T} \mathbf{A}_k \right) \mathbf{B}_t. \tag{5}$$

The computation of (4) could be implemented with a Reverse-Mode Differentiation (RMD) or Forward-Mode Differentiation (FMD) [7]. However, the aforementioned method is computationally expensive in terms of either latency (FMD) or memory (RMD). Note that RMD may not need to store the trajectory $\mathbf{w}_1, \ldots, \mathbf{w}_T$ in case of SGD with momentum. However, this would require $2T - 1$ Jacobian-vector products (JVPs), which is computationally demanding. So, we develop the method that performs only $T - 1$ JVPs for the hypergradient computation.

4 The Method

4.1 Hypergradient approximation

In this section we will introduce a computationally efficient approximation to (5). Specifically, consider the t-th step of the inner optimization. The challenge is that the computation of $\prod_{k=t+1}^{T} \mathbf{A}_k$ requires the tail of the trajectory $\mathbf{w}_t, \ldots, \mathbf{w}_T$. To this end, we introduce an approximation of the product with γ^{T-t}, where $0 < \gamma \leq 1$. Additionally, we replace the gradient of the validation loss $\nabla_{\mathbf{w}_T}\mathcal{L}_{\text{val}}(\mathbf{w}_T, \boldsymbol{\alpha})$ with the gradient from the current iteration $\nabla_{\mathbf{w}_t}\mathcal{L}_{\text{val}}(\mathbf{w}_t, \boldsymbol{\alpha})$ due to the same reason. Write down the proposed approximation:

$$\hat{d}_{\boldsymbol{\alpha}}\mathcal{L}_{\text{val}}(\mathbf{w}_T, \boldsymbol{\alpha}) = \nabla_{\boldsymbol{\alpha}}\mathcal{L}_{\text{val}}(\mathbf{w}_T, \boldsymbol{\alpha}) + \sum_{t=1}^{T} \gamma^{T-t} \nabla_{\mathbf{w}_t}\mathcal{L}_{\text{val}}(\mathbf{w}_t, \boldsymbol{\alpha}) \mathbf{B}_t. \tag{6}$$

Note that the intuition from (6) was previously used in [14]. However, it was used as an intermediate step in the reasoning. Moreover, the approximation of the gradient of the validation loss function w.r.t. model parameters was not considered.

4.2 Convergence Analysis

Assumption 4.1. *Suppose that the following assumptions on the functions $\mathcal{L}_{\text{train}}(.,.)$, $\mathcal{L}_{\text{val}}(.,.)$, and the optimization operator $\boldsymbol{\Phi}(.,.)$ are satisfied:*

- *$\mathcal{L}_{\text{val}}(., \boldsymbol{\alpha})$ is L-smooth and μ-strongly convex for any $\boldsymbol{\alpha}$.*
- *$\nabla^2_{\mathbf{w}}\mathcal{L}_{\text{train}}(., \boldsymbol{\alpha}) = \mathbf{I}$ for any $\boldsymbol{\alpha}$.*
- *$\nabla_{\boldsymbol{\alpha}}\mathcal{L}_{\text{val}}(\mathbf{w}, \boldsymbol{\alpha}) = \mathbf{0}$ for any $\mathbf{w}$.*
- *$\mathbf{B}_t\mathbf{B}_t^{\top} \succeq \kappa\mathbf{I}$ for some $\kappa > 0$.*
- *$\mathbf{B}_t(., \boldsymbol{\alpha})$ is C_B-Lipschitz for any $\boldsymbol{\alpha}$.*
- *$\|\mathbf{B}_t\| \leq B$ for any pair $(\mathbf{w}, \boldsymbol{\alpha})$ for some $B \geq 0$.*
- *$\boldsymbol{\Phi}(\mathbf{w}, \boldsymbol{\alpha}) = \mathbf{w} - \eta\nabla_{\mathbf{w}}\mathcal{L}_{\text{train}}(\mathbf{w}, \boldsymbol{\alpha})$ for some $\eta \in (0, 1)$.*
- *$\mathbf{w}$ belongs to a bounded convex set with diameter D.*
- *Define $\mathbf{w}_\infty := \arg\min_{\mathbf{w}} \mathcal{L}_{\text{train}}(\mathbf{w}, \boldsymbol{\alpha})$, $\mathbf{w}_2^* := \arg\min_{\mathbf{w}} \mathcal{L}_{\text{val}}(\mathbf{w}, \boldsymbol{\alpha})$. Assume that $\|\mathbf{w}_\infty - \mathbf{w}_2^*\| \geq 2De^{-\mu\eta T} + \delta$, for some $\delta > 0$. Intuitively, this requirements asserts that an overfitting takes place.*

Lemma 4.2. *([28]) In the assumptions above the sequence $\{\mathbf{w}_t\}_{t\geq 0}$ satisfies:*

$$\|\mathbf{w}_t - \mathbf{w}_\infty\|_2 \leq \|\mathbf{w}_0 - \mathbf{w}_\infty\|e^{-\eta t}. \tag{7}$$

Lemma 4.3. *Let the assumptions 4.1 hold. Then the following is true:*

$$\|\nabla_{\mathbf{w}_T}\mathcal{L}_{\text{val}}(\mathbf{w}_T, \boldsymbol{\alpha})\|_2 \geq \mu\delta. \tag{8}$$

Proof. First, use the Polyak-Lojasiewicz condition, since $\mathcal{L}_{\text{val}}(.,.)$ is μ-strongly convex in the first argument. Second, use the strong convexity. Third, use the convergence of $\mathbf{w}_T$, and finally the overfitting condition: TODO: convergence lemma.

$$\|\nabla_{\mathbf{w}_T}\mathcal{L}_{\text{val}}(\mathbf{w}_T,\boldsymbol{\alpha})\|_2^2 \geq 2\mu(\mathcal{L}_{\text{val}}(\mathbf{w}_T,\boldsymbol{\alpha}) - \mathcal{L}_{\text{val}}(\mathbf{w}_2^*,\boldsymbol{\alpha})) \geq \mu^2\|\mathbf{w}_T - \mathbf{w}_2^*\|^2 \geq$$
$$\mu^2(\|\mathbf{w}_T - \mathbf{w}_\infty\|_2^2 + \|_2\mathbf{w}_2^* - \mathbf{w}_\infty\|_2^2 - 2\|\mathbf{w}_T - \mathbf{w}_\infty\|_2 \cdot \|\mathbf{w}_2^* - \mathbf{w}_\infty\|_2) \geq$$
$$\mu^2(\|\mathbf{w}_2^* - \mathbf{w}_\infty\|_2 - 2De^{-\mu\eta T})\|\mathbf{w}_2^* - \mathbf{w}_\infty\|_2 \geq \mu^2\delta^2.$$

□

The following theorem guarantees that the proposed hypergradient is a sufficient descent direction.

Theorem 4.4. *Suppose that $\gamma = 1 - \eta \in (0,1)$. Then, under the assumptions above there exists a sufficiently large T and a universal constant $c > 0$ such that:*

$$d_{\boldsymbol{\alpha}}\mathcal{L}_{\text{val}}(\mathbf{w}_T,\boldsymbol{\alpha})\hat{d}_{\boldsymbol{\alpha}}\mathcal{L}_{\text{val}}(\mathbf{w}_T,\boldsymbol{\alpha})^\top \geq c\|d_{\boldsymbol{\alpha}}\mathcal{L}_{\text{val}}(\mathbf{w}_T,\boldsymbol{\alpha})\|_2^2.$$

Proof. Define $\mathbf{g}_j := \nabla_{\mathbf{w}_j}\mathcal{L}_{\text{val}}(\mathbf{w}_j,\boldsymbol{\alpha})$ for $j \in \{1,\dots,T\}$. Write down the dot product:

$$d_{\boldsymbol{\alpha}}\mathcal{L}_{\text{val}}(\mathbf{w}_T,\boldsymbol{\alpha})\hat{d}_{\boldsymbol{\alpha}}\mathcal{L}_{\text{val}}(\mathbf{w}_T,\boldsymbol{\alpha})^\top =$$
$$\sum_{j=1}^{T}\sum_{t=1}^{T}(1-\eta)^{2T-t-j}\nabla_{\mathbf{w}_T}\mathcal{L}_{\text{val}}(\mathbf{w}_T,\boldsymbol{\alpha})\mathbf{B}_t\mathbf{B}_j^\top\nabla_{\mathbf{w}_j}\mathcal{L}_{\text{val}}(\mathbf{w}_j,\boldsymbol{\alpha})^\top =$$
$$\sum_{j=1}^{T}\sum_{t=1}^{T}(1-\eta)^{2T-j-t}\mathbf{g}_T\mathbf{B}_t\mathbf{B}_j^\top\mathbf{g}_j.$$

Now estimate each term from below

$$\mathbf{g}_T\mathbf{B}_t\mathbf{B}_j^\top\mathbf{g}_j = \mathbf{g}_T\mathbf{B}_t\mathbf{B}_t^\top\mathbf{g}_j + \mathbf{g}_T\mathbf{B}_t(\mathbf{B}_j - \mathbf{B}_t)^\top\mathbf{g}_j \geq$$
$$\mathbf{g}_T\mathbf{B}_t\mathbf{B}_t^\top\mathbf{g}_j - C_B\|\mathbf{w}_j - \mathbf{w}_t\| \cdot \|\mathbf{g}_j\| \cdot \|\mathbf{g}_T\| \cdot \|\mathbf{B}_t\| \geq$$
$$\mathbf{g}_T\mathbf{B}_t\mathbf{B}_t^\top\mathbf{g}_j - C_BB\|\mathbf{w}_j - \mathbf{w}_t\| \cdot \|\mathbf{g}_j - \nabla_{\mathbf{w}_\infty}\mathcal{L}_{\text{train}}(\mathbf{w}_\infty,\boldsymbol{\alpha})\| \cdot \|\mathbf{g}_T\| \geq$$
$$\mathbf{g}_T\mathbf{B}_t\mathbf{B}_t^\top\mathbf{g}_j - C_BB\|\mathbf{w}_j - \mathbf{w}_t\| \cdot L\|\mathbf{w}_j - \mathbf{w}_t\| \cdot \|\mathbf{g}_T\| =$$
$$\mathbf{g}_T\mathbf{B}_t\mathbf{B}_t^\top\mathbf{g}_j - C_BBL\|\mathbf{w}_j - \mathbf{w}_\infty + \mathbf{w}_\infty - \mathbf{w}_t\|^2\|\mathbf{g}_T\| \geq$$
$$\mathbf{g}_T\mathbf{B}_t\mathbf{B}_t^\top\mathbf{g}_j - C_BBL(\|\mathbf{w}_0 - \mathbf{w}_\infty\|e^{-\eta t} + \|\mathbf{w}_0 - \mathbf{w}_\infty\|e^{-\eta j})^2\|\mathbf{g}_T\| \geq$$
$$\mathbf{g}_T\mathbf{B}_t\mathbf{B}_t^\top\mathbf{g}_j - 2C_BBL\|\mathbf{w}_0 - \mathbf{w}_\infty\|^2(e^{-2\eta t} + e^{-2\eta j})\|\mathbf{g}_T\|.$$

Now bound $\mathbf{g}_T\mathbf{B}_t\mathbf{B}_t^\top\mathbf{g}_j$ from below:

$$\mathbf{g}_T\mathbf{B}_t\mathbf{B}_t^\top\mathbf{g}_j = \mathbf{g}_T\mathbf{B}_t\mathbf{B}_t^\top\mathbf{g}_T + \mathbf{g}_T\mathbf{B}_t\mathbf{B}_t^\top(\mathbf{g}_j - \mathbf{g}_T) \geq$$
$$\kappa\|\mathbf{g}_T\|^2 - L\|\mathbf{g}_T\|B^2\|\mathbf{w}_j - \mathbf{w}_T\| \geq$$
$$\kappa\|\mathbf{g}_T\|^2 - L\|\mathbf{g}_T\|B^2\|\mathbf{w}_0 - \mathbf{w}_\infty\|(e^{-\eta T} + e^{-\eta j}).$$

Combining together the above bounds, we have:

$$\sum_{j=1}^{T}\sum_{t=1}^{T}(1-\eta)^{2T-j-t}\mathbf{g}_T\mathbf{B}_t\mathbf{B}_j^\top\mathbf{g}_j \geq$$
$$\kappa T^2\|\mathbf{g}_T\|^2 - 2C_BBL\|\mathbf{w}_0 - \mathbf{w}_\infty\|^2\|\mathbf{g}_T\|\sum_{j=1}^{T}\sum_{t=1}^{T}[e^{-2\eta t} + e^{-2\eta j}] -$$
$$LB^2\|\mathbf{w}_0 - \mathbf{w}_\infty\| \cdot \|\mathbf{g}_T\|(T^2e^{-\eta T} + T\sum_{j=1}^{T}e^{-\eta j}) \geq$$
$$\kappa T^2\|\mathbf{g}_T\|^2 - 2C_BBL\|\mathbf{w}_0 - \mathbf{w}_\infty\|^2\|\mathbf{g}_T\|T\eta^{-1} - LB^2\|\mathbf{w}_0 - \mathbf{w}_\infty\| \cdot \|\mathbf{g}_T\|(T^2e^{-\eta T} + T\eta^{-1}) \geq$$
$$\kappa T^2\|\mathbf{g}_T\|^2 - 2C_BBLD^2\|\mathbf{g}_T\|T\eta^{-1} - LB^2D \cdot \|\mathbf{g}_T\|(T^2e^{-\eta T} + T\eta^{-1}).$$

Using Lemma 4.3 we make the following statement. Since the first term of the bound is $\Theta(T^2)$ and the second and the third are $\Theta(T)$, then there exists sufficiently large T and a universal constant c such that the expression is bounded from below with $c\|\mathbf{g}_T\|^2$ for $\|\mathbf{g}_T\| \geq \mu\delta$. □

Method	$p_{\text{noise}} = 0.3$	$p_{\text{noise}} = 0.5$	$p_{\text{noise}} = 0.7$
w/o HPO	28.94	25.17	19.23
IFT(2)	28.97	25.81	20.89
Luketina	30.36	26.76	21.38
Ours ($\gamma = 0.99$)	30.39	26.77	21.47

Table 1: The results for data hyper-cleaning experiment. Validation accuracy is reported.

Method	3-way,10-shot	4-way, 10-shot	5-way, 10-shot
FO	43.48 +- 0.69	34.15 +- 0.53	29.66 +- 0.42
Luketina	42.96 +- 0.79	33.95 +- 0.64	28.78 +- 0.78
IFT(4)	41.64 +- 0.74	32.42 +- 1.11	26.4 +- 0.83
Ours ($\gamma = 0.99$)	46.1 +- 0.82	36.94 +- 1.07	30.62 +- 0.99

Table 2: Caption

5 Experiments

In this section we present numerical experiments that validate the effectiveness and efficiency of the proposed approach. Upon acceptance, we will make the source codes available.

TODO: describe the baselines

5.1 Data hyper-cleaning

Following [7], the task is formulated as follows. Given a training dataset $\mathfrak{D}_{\text{train}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n_{\text{train}}}$, where $\mathbf{x}_i$ is an object and y_i is a class label. Similary, define a validation dataset $\mathfrak{D}_{\text{val}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n_{\text{val}}}$. We assume that the labels of the training dataset are corrupted. More precisely, the label is replaced by a random class with probability p_{noise}. To mitigate the influence of noisy labels we introduce a vector of weights for each training object $\boldsymbol{\alpha} \in \mathbb{R}^{n_{\text{train}}}$. The task is to find a vector such that the model trained on the reweighted samples achives the optimal validation performance on clan data. Given a model parameters $\mathbf{w}$. The training loss function is $\mathcal{L}_{\text{train}}(\mathbf{w}, \boldsymbol{\alpha}) = \sum_{(\mathbf{x}_i, y_i) \in \mathfrak{D}_{\text{train}}} \sigma(\alpha_i) \ell(\mathbf{w}, \mathbf{x}_i, y_i)$, where $\sigma(.)$ is a sigmoid function, $\ell(.)$ is a cross-entropy loss function for the training pair $(\mathbf{x}_i, y_i)$. The validation loss function is $\mathcal{L}_{\text{val}}(\mathbf{w}, \boldsymbol{\alpha}) = \sum_{(\mathbf{x}_i, y_i) \in \mathfrak{D}_{\text{val}}} \ell(\mathbf{w}, \mathbf{x}_i, y_i)$.

TODO: hyperparameters, setup, and mention JAX)))

We report the validation accuracy of the compared baselines and the proposed method in Table 1. The results suggest that the proposed method outperforms the baselines in terms of validation accuracy, having comparable computational cost. TODO: report the number of JVPs for each method.

5.2 Gradient-based Meta-Learning

We consider gradient-base Meta-Learning task for few-shot image classification task [6]. TODO: describe the setup and hyperparameters. Inspired by [24, 13], we treat the logits head as a model parameters and the backbone of the convolutional network as hyperparameters. The dataset contains 100 classes, which are ranomly splited into 50 for meta-training and 50 for meta-validation. The accuracy on meta-validation split is depicted in Table 2. We report mean and standard deviation among 10 runs (TODO: stat tests and 95% intervals + describe FO baseline in the exp. section). It could be clearly seen that the proposed approach shows substantial improvements over the baselines in terms of accuracy on the meta-validation split.

5.3 Data Reweighting

Do we need this experiment? The results are on par with IFT, but we are faster. Maybe we need to report the plot: computational complexity in terms of JVPs vs accuracy.

6 Conclusion

TODO

References

[1] Antreas Antoniou, Harrison Edwards, and Amos Storkey. How to train your maml. In *International conference on learning representations*, 2018.

[2] Yoshua Bengio. Gradient-based optimization of hyperparameters. *Neural computation*, 12(8):1889–1900, 2000.

[3] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2), 2012.

[4] Mathieu Blondel, Quentin Berthet, Marco Cuturi, Roy Frostig, Stephan Hoyer, Felipe Llinares-López, Fabian Pedregosa, and Jean-Philippe Vert. Efficient and modular implicit differentiation. *Advances in neural information processing systems*, 35:5230–5242, 2022.

[5] Justin Domke. Generic methods for optimization-based modeling. In *Artificial Intelligence and Statistics*, pages 318–326. PMLR, 2012.

[6] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[7] Luca Franceschi, Michele Donini, Paolo Frasconi, and Massimiliano Pontil. Forward and reverse gradient-based hyperparameter optimization. In *International Conference on Machine Learning*, pages 1165–1173. PMLR, 2017.

[8] Riccardo Grazzi, Luca Franceschi, Massimiliano Pontil, and Saverio Salzo. On the iteration complexity of hypergradient computation. In *International Conference on Machine Learning*, pages 3748–3758. PMLR, 2020.

[9] Zhongkai Hao, Chengyang Ying, Hang Su, Jun Zhu, Jian Song, and Ze Cheng. Bi-level physics-informed neural networks for pde constrained optimization using broyden's hypergradients. In *The Eleventh International Conference on Learning Representations*, 2022.

[10] Ryuichiro Hataya and Makoto Yamada. Nyström method for accurate and scalable implicit differentiation. In *International Conference on Artificial Intelligence and Statistics*, pages 4643–4654. PMLR, 2023.

[11] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5149–5169, 2021.

[12] Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren. *Automated machine learning: methods, systems, challenges*. Springer Nature, 2019.

[13] Khurram Javed and Martha White. Meta-learning representations for continual learning. *Advances in neural information processing systems*, 32, 2019.

[14] Hae Beom Lee, Hayeon Lee, Jaewoong Shin, Eunho Yang, Timothy Hospedales, and Sung Ju Hwang. Online hyperparameter meta-learning with hypergradient distillation. *arXiv preprint arXiv:2110.02508*, 2021.

[15] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[16] Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*, 2018.

[17] Jonathan Lorraine, Paul Vicol, and David Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. In *International conference on artificial intelligence and statistics*, pages 1540–1552. PMLR, 2020.

[18] Jelena Luketina, Mathias Berglund, Klaus Greff, and Tapani Raiko. Scalable gradient-based tuning of continuous regularization hyperparameters. In *International conference on machine learning*, pages 2952–2960. PMLR, 2016.

[19] Dougal Maclaurin, David Duvenaud, and Ryan Adams. Gradient-based hyperparameter optimization through reversible learning. In *International conference on machine learning*, pages 2113–2122. PMLR, 2015.

[20] Paul Micaelli and Amos Storkey. Non-greedy gradient-based hyperparameter optimization over long horizons. 2020.

[21] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

[22] Fabian Pedregosa. Hyperparameter optimization with approximate gradient. In *International conference on machine learning*, pages 737–746. PMLR, 2016.

[23] Hieu Pham, Melody Guan, Barret Zoph, Quoc Le, and Jeff Dean. Efficient neural architecture search via parameters sharing. In *International conference on machine learning*, pages 4095–4104. PMLR, 2018.

[24] Aniruddh Raghu, Maithra Raghu, Samy Bengio, and Oriol Vinyals. Rapid learning or feature reuse? towards understanding the effectiveness of maml. *arXiv preprint arXiv:1909.09157*, 2019.

[25] Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. Meta-learning with implicit gradients. *Advances in neural information processing systems*, 32, 2019.

[26] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *International conference on learning representations*, 2016.

[27] Jürgen Schmidhuber. *Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-... hook*. PhD thesis, Technische Universität München, 1987.

[28] Amirreza Shaban, Ching-An Cheng, Nathan Hatch, and Byron Boots. Truncated back-propagation for bilevel optimization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1723–1732. PMLR, 2019.

[29] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, 25, 2012.

[30] Yuhuai Wu, Mengye Ren, Renjie Liao, and Roger Grosse. Understanding short-horizon bias in stochastic meta-optimization. *arXiv preprint arXiv:1803.02021*, 2018.

[31] Barret Zoph and Quoc V Le. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*, 2016.

A Appendix / supplemental material

Optionally include supplemental material (complete proofs, additional experiments and plots) in appendix. All such materials **SHOULD be included in the main submission.**

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and precede the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: **[TODO]**

   Justification: **[TODO]**

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: **[TODO]**

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: **[TODO]**

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: **[TODO]**

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: **[TODO]**

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: **[TODO]**

    Justification: **[TODO]**

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: **[TODO]**

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.