MAP: Towards Balanced Generalization of IID and OOD through Model-Agnostic Adapters

Min Zhang¹, Junkun Yuan¹, Yue He², Wenbin Li³, Zhengyu Chen¹, Kun Kuang¹*
¹Zhejiang University ²Tsinghua University ³Nanjing University

heyuethu@mail.tsinghua.edu.cn, liwenbin@nju.edu.cn,
{zhangmin.milab, yuanjk, chenzhengyu, kunkuang}@zju.edu.cn

Abstract

Deep learning has achieved tremendous success in recent years, but most of these successes are built on an independent and identically distributed (IID) assumption. This somewhat hinders the application of deep learning to the more challenging out-of-distribution (OOD) scenarios. Although many OOD methods have been proposed to address this problem and have obtained good performance on testing data that is of major shifts with training distributions, interestingly, we experimentally find that these methods achieve excellent OOD performance by making a great sacrifice of the IID performance. We call this finding the IID-OOD dilemma. Clearly, in real-world applications, distribution shifts between training and testing data are often uncertain, where shifts could be minor, and even close to the IID scenario, and thus it is truly important to design a deep model with the balanced generalization ability between IID and OOD. To this end, in this paper, we investigate an intriguing problem of balancing IID and OOD generalizations and propose a novel Model-Agnostic Adapters (MAP) method, which is more reliable and effective for distribution-shift-agnostic real-world data. Our key technical contribution is to use auxiliary adapter layers to incorporate the inductive bias of IID into OOD methods. To achieve this goal, we apply a bilevel optimization to explicitly model and optimize the coupling relationship between the OOD model and auxiliary adapter layers. We also theoretically give a first-order approximation to save computational time. Experimental results on six datasets successfully demonstrate that MAP can greatly improve the performance of IID while achieving good OOD performance.

1. Introduction

Deep learning has achieved unprecedented success in various applications of computer vision, e.g., image classification [15, 17, 18], but most of these successes are based on an independent and identically distributed (IID) assumption, i.e., training and testing data are drawn from the same distribution [41, 16]. However, out-of-distribution (OOD) shifts between training and testing data are usually inevitable in the real world due to the widespread existence of unobserved confounders or data bias [46, 8]. Under such circumstances, deep models trained by empirical risk minimization (ERM) [51] with the IID assumption usually suffer from poor performance on OOD data. Therefore, it is important to improve the OOD generalization of deep models.

Recently, many OOD methods have been proposed to learn representations or predictors that are invariant to different distributions (or named environments) by introducing various regularizers [3, 4, 44, 1, 26, 2, 34, 64, 63, 50]. Although these methods achieve good OOD performance on testing data that is of major distribution shifts with training data, we experimentally found that they would significantly damage the performance on IID (with nearly no shift discrepancy) or minor shift data. We implement some representative OOD methods in both OOD and IID scenarios on ColoredMNIST and show results in Figure 1. We have an interesting observation that these methods have significant OOD accuracy but lower IID performance compared with the IID method (e.g., ERM) with higher IID performance.

Figure 1. Comparison of OOD and IID accuracy of OOD, IID and our proposed MAP methods on ColoredMNIST. HM is the harmonic mean of IID and OOD accuracy. All OOD methods achieve high OOD performance with the sacrifice of IID accuracy compared with ERM (i.e., an IID method). MAP achieves balanced generalization by incorporating IID inductive bias into OOD generalization learning. More results are shown in Table 1.
but lower OOD accuracy. A possible reason for causing this phenomenon is that many OOD methods extract invariant features while possibly losing some information that contributes to IID generalization. Next, we explore this phenomenon from the perspective of inductive bias learned by different IID and OOD methods inspired by [36].

In Figure 2, following [53], we visualize channel-wise BatchNorm (BN) statistics of the 4-th layer as the inductive bias (other layers have similar observations). The inductive bias, i.e., running means and variances, of IID and OOD methods are significantly different, i.e., IID (OOD) methods are routed to BN_{IID} (BN_{OOD}). As pointed out by [2, 32], ERM extracts easy-to-learn variant features (e.g., color on ColoredMNIST) in training distributions and generalizes well to testing data with the same distribution. In comparison, these OOD methods adopt regularizers to encourage the model to extract hard-to-learn invariant features (e.g., digit on ColoredMNIST) and improve the performance of testing data that differ significantly from training distributions. Moreover, the regularizer guides the OOD model in different optimization directions compared with the IID method, which causes good OOD accuracy with low IID results. The finding is named as the IID-OOD dilemma. Both IID and OOD methods can only perform well in a specific scenario (IID or OOD), which limits their real-world applications with uncertain distribution shifts. Therefore, this observation motivates us to ask: is it possible to design a model with a balanced performance between IID and OOD generalizations in the IID-OOD dilemma?

In this paper, we take a step forward to propose a novel Model Agnostic adaPters (MAP) method that achieves balanced generalization performance in both IID and OOD evaluations. Specifically, we insert auxiliary adapter layers (AALs) in the OOD model to learn variant features with the inductive bias of the IID scenario, while keeping the ability to extract invariant features with the inductive bias of the OOD scenario. Training processes of the OOD model and AALs are viewed as two kinds of tasks: the OOD model learns OOD knowledge and AALs extract IID information. To achieve this, we formulate the learning into a bilevel optimization (BLO) problem. In the inner level, we optimize the OOD model with AALs by using an OOD loss. In the outer level, we utilize the IID criterion evaluated on the validation set based on the optimized OOD model in the inner level as the outer objective to guide the training of AALs. We alternatively perform the inner level and outer level and finally obtain a set of optimal parameters for the adapter and OOD model. To save computational time and memory, we theoretically give a first-order approximation of BLO. Note that AALs are model-agnostic and can be plugged into an arbitrary OOD method. Experiments evaluate the effectiveness of MAP, which improves the trade-off ability of OOD methods (see the HM metric in Figure 1) by capturing the inductive bias of both IID and OOD scenarios in Figure 2. Our main contributions are summarized as follows:

- We investigate a problem called the IID-OOD dilemma, i.e., most OOD (or IID) methods achieve good OOD (or IID) performance with a sacrifice of IID (or OOD) accuracy, which is beyond the capability of these methods in real-world data with uncertain shifts.

- We propose a simple yet effective Model Agnostic adaPters (MAP) method to simultaneously learn inductive biases of both IID and OOD. To achieve this, a bilevel optimization (BLO) is used to train our MAP. Unlike the computationally intensive BLO solver, we theoretically give a first-order approximation.

- We conduct extensive experiments across six datasets, three model architectures, and sixteen baselines. We show that (1) MAP balances the performance of IID and OOD. (2) MAP is model-agnostic and can be plugged into any OOD method. (3) MAP is able to achieve reliable performance under various settings.

2. Related Work

Out-of-distribution generalization. To enable deep learning models to generalize to unknown data distributions, the task of out-of-distribution (OOD) generalization aims to train a generalizable model from one or multiple source domains and make it predict well on the previously unseen target domains. To get rid of the independent and identically distributed (IID) assumption employed by the conventional algorithms like empirical risk minimization (ERM) [51], a variety of OOD strategies have been proposed to overcome distribution shifts, including regularized training [60, 26, 38, 47, 23, 54, 57], meta-learning [27, 7, 59, 22, 62], data augmentation [55, 42], domain alignment [3, 28], causal
Given a dataset \( D \), we detail the optimization processes of IID and OOD. Domain-specific gradients. These methods aim to improve the correlation between domains to encourage the alignment between the training and testing data from source domains. ARM \([60]\) optimizes to unseen domains by simulating distribution shifts with virtual target data. MLDG \([27]\) makes the model learn how to generalize shifts \([27, 31, 40, 61, 6, 48, 52, 39, 58, 14]\). As a pioneer work, MLDG \([27]\) makes the model learn how to generalize to unseen domains by simulating distribution shifts with virtual target data from source domains. ARM \([60]\) optimizes the model for effective adaptation to shift by learning to adapt on training domains. Fish \([47]\) augments the loss with an auxiliary term that maximizes the gradient inner product between domains to encourage the alignment between the domain-specific gradients. These methods aim to improve the OOD performance of the model by using bilevel optimization (BLO), while our goal is to optimize MAP so that it can balance the performance of IID and OOD data.

3. Preliminaries

In this section, we first formulate the problem definition. Then, we detail the optimization processes of IID and OOD.

Problem definition. Given a dataset \( D := \{(x_i, y_i)\}_{i=1}^n \) with \( n \) samples \( (x_i, y_i) \) is drawn from a joint space of \( \mathcal{X} \times \mathcal{Y} \).

In general, \( D \) contains the training data \( D_{tr} \) sampled from the training distribution \( P_{tr}(\mathcal{X}, \mathcal{Y}) \) and the testing data \( D_{te} \) drawn from the testing distribution \( P_{te}(\mathcal{X}, \mathcal{Y}) \). Supervised learning methods aim to predict labels \( y_i \) of \( x_i \) originate from a featurizer \( f(\cdot; \theta) \) parameterized by \( \theta \) and the classifier \( g(\cdot; \phi) \) parameterized by \( \phi \), i.e., \( \theta : \mathcal{X} \rightarrow \mathcal{Z} \) and \( \phi : \mathcal{Z} \rightarrow \mathcal{Y} \). \( Z \) represents the sample feature space.

IID learning. Deep learning usually assumes that training and testing data are both IID realizations from a common underlying distribution, i.e., \( P_{tr}(\mathcal{X}, \mathcal{Y}) = P_{te}(\mathcal{X}, \mathcal{Y}) \). Based on such a hypothesis, empirical risk minimization (ERM) which minimizes the average loss on training samples could optimize the model for the testing distribution. Specifically, ERM minimizes the following objective:

\[
\mathcal{L}_{ERM}(D_{tr}, \theta, \phi) = \frac{1}{n} \sum_{i=1}^{n} \ell(g_\phi(f_\theta(x_i)), y_i),
\]

where \( \ell(\cdot, \cdot) \) is the loss function, e.g., cross entropy for image classification. \( z_i = f_\theta(x_i) \) is the feature representation and \( y_i = g_\phi(z_i) = g_\phi(f_\theta(x_i)) \) is the prediction.

OOD learning. There is an OOD problem when the testing distribution is unseen and different from the training distribution, i.e., \( P_{tr}(\mathcal{X}, \mathcal{Y}) \neq P_{te}(\mathcal{X}, \mathcal{Y}) \). Specifically, following \([4, 1]\), we have multiple environments (or distributions) \( \mathcal{E} = \{e_1, e_2, \ldots, e_E\} \) in the sample space \( \mathcal{X} \times \mathcal{Y} \) with the training distribution. The correlation between the non-semantic information (i.e., variant features) and labels (i.e., invariant features) is unstable among different environments. Existing OOD methods learn invariant representations or predictors by introducing the regularizer. The optimization process of OOD methods (IRM \([4]\) and VERx \([26]\) as an example due to its simple yet effective) is as below:

\[
\mathcal{R}^{ERM} = \sum_e \mathcal{L}_{ERM}(D_{tr}^e, \theta, \phi) + \lambda \mathcal{V}_e[\mathcal{L}(D^e, \phi)],
\]

where \( \mathcal{V}_e[\mathcal{L}(D^e, \phi)] \) is the variance of the loss across different environments. Equation (2) is the optimization objective of IRM using the fixed “dummy” classifier. Equation (3) is the VERx optimization objective. \( \lambda \in [0, \infty) \) is a hyperparameter to balance between the ERM and regularizer loss.

4. Methodology

To address the interesting problem of the IID-OOD dilemma, in this section, we propose a model-agnostic adapters (MAP) method by introducing auxiliary adapter layers (AALs) in the OOD model. Specifically, as illustrated in Figure 3, we incorporate AALs into the OOD model to learn the inductive bias of IID and OOD data by using a bilevel alternating way. In the following subsections, we detail the bilevel optimization (BLO) process and the specific form of our proposed auxiliary adapter layers.

4.1. Bilevel Optimization

As shown in Figure 2, OOD and IID models learn different inductive biases from training data to improve the generalization in testing data. Under such circumstances, these models cannot achieve good performance under both IID and OOD. To improve the trade-off of IID and OOD, we design MAP to help OOD models learn IID knowledge by
integrating AALs into OOD models. Our goal is to let OOD models extract invariant features (e.g., object information) that help OOD generalization while also focusing on variant features (e.g., background) that improve IID performance.

To this end, the learning paradigm of AALs and OOD models is viewed as two kinds of tasks: 1) OOD models learn OOD inductive bias, and 2) AALs recognize IID inductive bias. Moreover, we interpret the OOD-learning task (i.e., 1) and the AALs-learning task (i.e., 2) as two optimization levels, where the latter is formulated as an outer-level optimization problem, and it relies on the optimization of the inner-level OOD-learning task. To the best of our knowledge, the bilevel optimization (BLO) framework has not been considered for balancing IID and OOD in-depth knowledge, the bilevel optimization (BLO) framework has been used to solve the IG challenge, we propose an alternating approximation algorithm to save computational time and memory.

**4.2. BLO with Gradient Approximation**

The computation of implicit gradient (IG) is the key challenge of optimizing Equation (4). In this section, to solve the IG challenge, we propose an alternating approximation algorithm to save computational time and memory.

**Updating ω in the inner level.** In each outer iteration, instead of completely solving the inner level problem, we fix auxiliary adapter layers α and only consider gradient steps of the model parameters ω at the t-th iteration as follows:

$$\omega(t) = \omega(t-1) - \eta_\omega \nabla_\omega \mathcal{R}(B_\omega, \omega(t-1), \alpha^{(t-1)})$$

where \( \nabla_\omega \) is partial derivatives of \( \omega \). \( \eta_\omega \) is the learning rate for model parameters \( \omega \). \( B_\omega \) is batch sampled from \( D_{tr} \).

**Updating α in the outer level.** After obtaining the parameters \( \omega(t) \) (a reasonable approximation of \( \omega^*(\alpha) \)), we update \( \alpha \) by calculating the outer level optimization objective as:

$$\alpha(t) = \alpha(t-1) - \eta_\alpha \nabla_\alpha \mathcal{L}_{ERM}(B_\alpha, \omega(t), \alpha^{(t-1)})$$

where \( \nabla_\alpha \) means partial derivatives of \( \alpha \). \( \eta_\alpha \) is the learning rate for adapter parameters \( \alpha \). \( B_\alpha \) is batch sampled from validation data \( D_{val} \) flipping by \( D_{tr} \). When we directly backpropagate the gradient, the IG problem occurs because \( \omega(t) \) nested inside \( \alpha^{(t)} \). Therefore, we propose a method to approximate the gradient \( \nabla_\alpha \mathcal{L}_{ERM}(B_\alpha, \omega(t), \alpha^{(t-1)}) \) of \( \alpha^{(t)} \) (see supplementary for detailed derivations) as below:

$$\nabla_\alpha \mathcal{L}_{ERM}(\omega(t), \alpha^{(t-1)}) = \nabla_\omega \mathcal{L}_{ERM}(\omega(t), \alpha^{(t-1)}) \cdot \nabla_\alpha \omega^*(\alpha) \quad \text{(IG)}$$

**Gradient of AALs**

$$= -\eta_\omega \frac{1}{\epsilon} (\nabla_\omega \mathcal{R}(\omega(t-1) + \epsilon v, \alpha^{(t-1)})) - \nabla_\alpha \mathcal{L}_{ERM}(\omega(t-1), \alpha^{(t-1)})$$

where \( v = \nabla_\omega \mathcal{L}_{ERM}(\omega(t), \alpha^{(t-1)}) \) with small \( \epsilon > 0 \). For ease of notation, we omit \( B_\omega \) and \( B_\alpha \) in loss \( \mathcal{R} \) and
\[ L_{ERM}, \] respectively. Equation (8) can be easily implemented by maintaining \( \omega^{(t-1)} \) at last iteration to catch the OOD loss \( \mathcal{R}(\omega^{(t-1)}, \alpha^{(t-1)}) \) and compute the new loss \( \mathcal{R}(\omega^{(t-1)} + \alpha^{(t-1)}) \). When \( \eta_\omega \) is set to 0 in Equation (8), the second-order derivative will disappear, resulting in a first-order approximate. The complexity of the first order is the same as OOD methods, and the performance is as efficient as the second order (see results in Section 5.4).

### 4.3. Auxiliary Adapter Layers

In this subsection, we discuss how to connect the OOD featurizer with AALs and how to design the form of AALs. For the first problem, we insert the adapter between the conv layer and BN, as shown in the right of Figure 5 in each OOD module (left of Figure 5). Concretely, The \( l \in [1, \ldots, L] \)-th layer of the featurizer \( f_\theta \) is denoted as \( f_\theta_\alpha \) with the weights \( \theta_\alpha \). And the adapter in each module is denoted as \( A_{\alpha_l} \) parameterized by \( \alpha_l \). The information extracted by \( A_{\alpha_l} \) can be incorporated into the output of the \( l \)-th layer as the input of \( l + 1 \) layer. The formulation is as follows:

\[
f_{(\theta_\alpha, \alpha_\alpha)}(z_l) = A_{\alpha_l}(f_{\theta_l}(z_l), z_l),
\]

where \( z_l \in \mathbb{R}^{W_l \times H_l \times C_l} \) is the feature tensor that is the input of the \( l \)-th module \( \theta_l \). \( W_l, H_l \) and \( C_l \) are the width, height and channel of the \( l \)-th convolutional layer, respectively. Motivated by [29], we consider two connection ways for incorporating adapter \( A_{\alpha_\alpha} \) into OOD featurizer \( f_{\theta_l} \):

1. Serial connection by subsequently applying it to the output:

\[
f_{(\theta, \alpha_\alpha)}(z_l) = A_{\alpha_l} \circ f_{\theta_l}(z_l),
\]

which is illustrated in Figure 4 (a), and

2. Parallel connection by a residual addition and is illustrated in Figure 4 (b):

\[
f_{(\theta, \alpha_\alpha)}(z_l) = A_{\alpha_l} + f_{\theta_l}(z_l).
\]

For the second problem, we consider two options for \( A_{\alpha_l} \):

- Matrix multiplication with \( \alpha_l \in \mathbb{R}^{C_l \times C_{l+1}} \) in Figure 4 (c), where \( C_l \) and \( C_{l+1} \) are the number of input and output channels, respectively. The formulation is as below:

\[
A_{\alpha_l}(f_{\theta_l}(z_l)) = z_l \odot \alpha_l,
\]

where \( \odot \) denotes a convolutional operation with \( 1 \times 1 \) kernels in our code.

### 5. Experiments

In this section, we evaluate the performance of the proposed MAP, aiming to answer the following questions: Q1: Could MAP balance the robustness of IID and OOD generalization compared with prior IID and OOD methods? (Section 5.2) Q2: Could MAP be the model-agnostic adapters? (Section 5.3) Q3: How effective is the proposed MAP in different settings (or ablation study)? (Section 5.4)\(^1\)

#### 5.1. Experimental Setup

**Dataset.** We use six OOD classification datasets, including three toy, i.e., ColoredMNIST [4], ColoredCOCO [1], COCOPlaces [1] and three real, i.e., NICO [19], CelebA [35], WILDS-Camelyon [24] (more details in the supplementary).

**Baseline methods.** We compare our MAP to a large number of algorithms that span different learning strategies, including (1) IID learning: ERM [51] (2) OOD learning (fifteen methods): IRM [4], VREx [26], ARM [60], GroupDRO [44], MLDG [27], MMD [28], IGA [25], SANDMask [45], Fish [47], CDANN [30], TRM [54] IBERM [2], IBIRM [2], CondCAD [42], CausIRL, CORAL [9] where ARM, MLDG and Fish also use the bilevel optimization (see more details in Section 2).

**Backbone.** We use the four-layer convolutional neural network (Conv4) for ColoredMNIST and ResNet18 pretrained by ImageNet [43] for three real datasets. Following [1], a

\(^1\)The code is available at: https://github.com/remiMZ/MAP-ICCV23.
<table>
<thead>
<tr>
<th>Methods</th>
<th>ColoredMNIST OOD</th>
<th>ColoredMNIST IID</th>
<th>ColoredCOCO OOD</th>
<th>ColoredCOCO IID</th>
<th>COCOPlaces OOD</th>
<th>COCOPlaces IID</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ERM [51]</td>
<td>29.7 ± 0.2</td>
<td>86.0 ± 0.2</td>
<td>44.2</td>
<td>45.4 ± 0.9</td>
<td>77.0 ± 0.6</td>
<td>57.1</td>
</tr>
<tr>
<td>IRM [4]</td>
<td>60.3 ± 2.8</td>
<td>32.6 ± 7.0</td>
<td>42.3</td>
<td>49.2 ± 0.3</td>
<td>70.9 ± 1.7</td>
<td>58.1</td>
</tr>
<tr>
<td>VREX [26]</td>
<td>52.9 ± 1.2</td>
<td>14.6 ± 0.3</td>
<td>22.9</td>
<td>48.8 ± 0.7</td>
<td>73.6 ± 0.9</td>
<td>58.7</td>
</tr>
<tr>
<td>GroupDRO [44]</td>
<td>38.5 ± 1.5</td>
<td>51.5 ± 0.3</td>
<td>44.1</td>
<td>49.1 ± 0.6</td>
<td>74.8 ± 1.8</td>
<td>59.3</td>
</tr>
<tr>
<td>MLDG [27]</td>
<td>29.4 ± 0.6</td>
<td>50.3 ± 0.0</td>
<td>34.6</td>
<td>11.9 ± 0.8</td>
<td>20.5 ± 0.1</td>
<td>15.1</td>
</tr>
<tr>
<td>MMD [28]</td>
<td>50.6 ± 0.1</td>
<td>51.3 ± 0.6</td>
<td>51.0</td>
<td>50.4 ± 0.8</td>
<td>73.8 ± 1.0</td>
<td>59.9</td>
</tr>
<tr>
<td>IGA [25]</td>
<td>50.5 ± 0.1</td>
<td>25.0 ± 7.9</td>
<td>33.4</td>
<td>11.0 ± 0.6</td>
<td>17.5 ± 2.7</td>
<td>13.5</td>
</tr>
<tr>
<td>SANDMask [45]</td>
<td>58.6 ± 6.5</td>
<td>42.2 ± 7.2</td>
<td>49.1</td>
<td>49.2 ± 1.2</td>
<td>74.0 ± 0.7</td>
<td>59.1</td>
</tr>
<tr>
<td>Fish [47]</td>
<td>28.0 ± 1.5</td>
<td>46.4 ± 3.2</td>
<td>34.9</td>
<td>41.7 ± 0.5</td>
<td>71.7 ± 1.0</td>
<td>52.7</td>
</tr>
<tr>
<td>CDANN [30]</td>
<td>41.7 ± 3.5</td>
<td>22.6 ± 1.5</td>
<td>29.3</td>
<td>38.4 ± 1.5</td>
<td>70.1 ± 1.3</td>
<td>49.6</td>
</tr>
<tr>
<td>TRM [54]</td>
<td>44.2 ± 5.0</td>
<td>32.1 ± 9.5</td>
<td>37.2</td>
<td>47.5 ± 0.6</td>
<td>72.8 ± 0.2</td>
<td>57.5</td>
</tr>
<tr>
<td>IB_ESRM [2]</td>
<td>50.2 ± 0.2</td>
<td>51.7 ± 1.7</td>
<td>50.9</td>
<td>45.4 ± 0.1</td>
<td>72.4 ± 2.5</td>
<td>55.8</td>
</tr>
<tr>
<td>CausIRL_CORAL[9]</td>
<td>28.7 ± 1.3</td>
<td>50.6 ± 0.2</td>
<td>36.6</td>
<td><strong>51.5 ± 1.1</strong></td>
<td>73.9 ± 0.9</td>
<td>60.7</td>
</tr>
<tr>
<td>CondCAD [42]</td>
<td>49.2 ± 0.5</td>
<td>21.1 ± 2.6</td>
<td>29.5</td>
<td>41.2 ± 0.7</td>
<td>67.2 ± 1.3</td>
<td>51.1</td>
</tr>
<tr>
<td>IB_IRM [2]</td>
<td>53.8 ± 1.8</td>
<td>37.9 ± 10.0</td>
<td>44.5</td>
<td>33.9 ± 0.6</td>
<td>67.3 ± 1.4</td>
<td>45.1</td>
</tr>
<tr>
<td>ARM [60]</td>
<td>28.1 ± 0.0</td>
<td>49.9 ± 0.1</td>
<td>36.0</td>
<td>33.0 ± 0.6</td>
<td>63.3 ± 0.6</td>
<td>43.4</td>
</tr>
</tbody>
</table>

**Table 1.** Experiments of three toy (top) and three real (bottom) datasets. Here, we show average accuracy (%) of IID and OOD. **HM** is the harmonic mean as a trade-off metric. We repeat experiments three times across 20 hyperparameter seeds by following DomainBed [13].

residual network trained from scratch is used for ColoredCOCO and COCOPlaces and is called as ResNet8. For Conv4, AALs are placed behind the convolutional layer. For residual networks, we only use AALs in each block.

**Model selection and implementation details.** To evaluate the performance of IID and OOD, we split each environment for each dataset into two subsets of $d_1$ and $d_2$, where the number of samples of $d_1$ and $d_2$ is 9:1. The subset $d_1$ of training environments is used to train the model, and $d_2$ is used to evaluate IID accuracy. While the subset $d_1$ of testing environments is used to evaluate OOD accuracy, and $d_2$ is used to select the best model (or named oracle selection) by following the standard protocol of DomainBed [13, 56].

Then IID and OOD accuracy are calculated based on the selected model. Note that MAP uses the OOD loss of the VREX [26] method in the inner level.

### 5.2. Evaluating the Balance of IID and OOD Data

In Table 1, we report the overall performance of MAP and sixteen baselines under IID and OOD evaluations on six datasets. We further reported the harmonic mean \((HM = \frac{1}{2} \times \frac{1}{\frac{1}{d_1} + \frac{1}{d_2}})\) of accuracy on IID and OOD data. Following [36], we use this metric to evaluate the trade-off between IID and OOD performance. According to Table 1, we have the following findings: (1) Compared with the IID method (i.e., ERM), these fifteen OOD methods have good OOD perfor-
mance, but a significant drop in IID accuracy on all datasets, which demonstrates our motivation, i.e., most OOD methods might lose some information that helps the IID learning in the OOD generalization. Furthermore, since the spurious correlation might be weaker on real datasets, the performance gap between IID and OOD methods becomes smaller. This phenomenon demonstrates that the IID model extracts easy-to-learn variant features to learn the inductive bias of training data, in comparison, the OOD model learns hard-to-learn invariant features to improve the performance of unseen testing distributions. (2) According to the HM metric, MAP achieves the balanced generalization ability on all datasets. This is because MAP has the advantage of simultaneously capturing the inductive bias between IID and OOD data, which demonstrates the effectiveness of our method. (3) Surprisingly, MAP even increases the OOD performance of OOD methods in some settings and the IID performance in ColoredCOCO outperforms ERM by 1.1%. Similar conclusions can be obtained based on Table 2. One possible reason is that the bilevel optimization can find suitable model parameters for IID and OOD generalizations (see more discussion in Sections 5.3).

5.3. Evaluating the Flexibility of MAP

We study how the proposed MAP improves the trade-off ability of many OOD methods. In Table 2, we show results by using eight OOD methods with or without MAP on three datasets. (1) We can observe that all OOD methods with MAP outperform those without MAP, especially in using VREx on ColoredMNIST bring a 37% improvement. It not only indicates that MAP is a model-agnostic framework, but also significantly improves the trade-off ability of these OOD methods. (2) With a deep look at the OOD evaluation, MAP shows its comparable performance. Surprisingly, MAP even increases the OOD performance in some OOD methods, especially on ColoredMNIST. We think that the iterative learning of the OOD model and AALs using bilevel optimization may help the OOD model to explore more knowledge that is helpful for OOD generalization.

In Figure 6, we visualize the inductive bias of five OOD methods with or without using MAP to observe the learned statistics information. The inductive bias learned by MAP can capture both IID and OOD scenarios. An interesting observation is that compared with CDANN, MAP not only uses adapters to capture IID generalization information but also explores more OOD generalization knowledge. A similar observation is made in the ARM method. This phenomenon shows that MAP has the ability to simultaneously improve the performance of IID and OOD evaluations.

5.4. Ablation Study

We conduct extensive ablation studies to evaluate the robustness of the proposed MAP under various settings.

Comparison of different structural designs of MAP. In Table 3, we analyze the impact of different connections (i.e., serial or residual in Figure 4 (a) and (b)), different forms
which shows that MAP can learn the knowledge lost by ant features while possibly losing some information that demonstrates that these OOD methods extract invariance.

In Table 4, we construct various distribution shifts, i.e., matrix or channel in Figure 4 (c) and (d)) and different initializations (i.e., random or eye). In all settings, a combination of residual, matrix and random has the best performance and this combination is also used in other experimental settings. Although different combinations bring different performances, in most cases, MAP is far superior to the baseline (close to 40% improvement), especially in ColoredMNIST with strong spurious correlations. We will further study the best combination form in future work.

**Could MAP perform well under different distribution shifts?** In Table 4, we construct various distribution shifts, i.e., from major shifts to minor shifts, to simulate uncertain real-world data. The performance of most OOD methods degrades as the shifts get smaller or closer to IID data, which demonstrates that these OOD methods extract invariant features while possibly losing some information that contributes to IID generalization. On the contrary, our MAP has good performance under different distribution shifts, which shows that MAP can learn the knowledge lost by OOD methods. More results are in the supplementary.

**Is the bilevel optimization effective?** We analyze the training strategy of first-order, second-order bilevel optimization (BLO) and without BLO (we optimize $\omega$ simultaneously with $\alpha$ on training data without validation and two levels).

![Figure 7](image)

**Figure 7.** Experiments using different optimization algorithms. VREx, FBO, SBL0 and WBO denote vanilla VREx, first-order BLO, second-order BLO and without BLO, respectively.

Figure 7 shows results on the ColoredMNIST and NICO datasets. We can observe that the performance of first-order and second-order BLO is similar and better than the vanilla VREx, but WBO has poor performance in NICO. This demonstrates that the BLO can significantly improve generalization capability, and the first-order approximate is sufficient as the second order with a good performance.

**Model parameters.** In Table 5, we compare model parameters with or without MAP based on three featureizers. The form of adapters is residual and matrix, because this combination brings the largest parameters compared with other forms in Table 2. A few model parameters can bring significant performance, especially in Conv4, which shows the practical availability of the proposed MAP in applications.

### 6. Conclusion

In this paper, we investigate a problem of the IID-OOD dilemma and propose an effective Model Agnostic adaptors (MAP) method to achieve the balanced generalization performance between IID and OOD evaluations for uncertain real-world distribution shifts. Specifically, the proposed MAP method inserts auxiliary adapter layers (AALs) in the OOD model to simultaneously learn the inductive bias of IID and OOD data. To achieve this, a bilevel optimization (BLO) is used, which optimizes the OOD model in the inner level using an OOD loss and updates AALs in the outer.
level with an ERM loss based on the optimized OOD model in the inner level. Extensive experiments on six datasets demonstrate that our MAP is able to balance both IID and OOD performance compared with sixteen IID and OOD methods. In future work, we may extend our method to other tasks, e.g., adversarial attacks, to improve the trade-off ability of clean accuracy and adversarial robustness.

Acknowledgement

This work was supported in part by National Key Research and Development Program of China (2022YFC3340900), National Natural Science Foundation of China (62006207, U20A20387), Young Elite Scientists Sponsorship Program by CAST (2021QNRC001), Zhejiang Province Natural Science Foundation (No. LQ21F020020), the StarryNight Science Fund of Zhejiang University of China (62006207, U20A20387), Young Elite Scientists Sponsorship Program by CAST (2021QNRC001), Zhejiang University of China (62006207, U20A20387), National Natural Science Foundation of China (62006207, U20A20387), Young Elite Scientists Sponsorship Program by CAST (2021QNRC001), Zhejiang University of China (62006207, U20A20387), and the Fundamental Research Funds for the Central Universities (226-2022-00142, 226-2022-00051).

References


[22] Yinjie Jiang, Zhengyu Chen, Kun Kuang, Luotian Yuan, Xinhai Ye, Zhihua Wang, Fei Wu, and Ying Wei. The role


