Doctoral Dissertation

A Study Toward Practical Application of Domain Invariance Learning: Domain Invariance Estimation with Coarse Labels and its Hyperparameter Selection

Shoji Toyota

Department of Statistical Science
School of Multidisciplinary Sciences
The Graduate University for Advanced Studies, SOKENDAI

March, 2023

abstract

Machine learning models often inherit spurious correlations embedded in training data and hence may fail to predict desired labels on unseen domains, which have different distributions from the domain to provide training data. In response to the problem of spurious correlations, it is recognized that one of the important issues for the future of machine learning is generalization to data generated by a distribution outside training ones, which is often called *out-of-distribution (o.o.d.) generalization*. Domain Invariance Learning (DIL) is a rapidly developed approach for o.o.d. generalization; using training data in many domains, DIL estimates such a predictor that enables o.o.d. generalization. However, DIL has two drawbacks, which hinder the application of DIL to real-world problems. Firstly, DIL often involves expensive and exhausting annotations. In their estimation, DILs demand training data, consisting of the pairs of input data and its teacher labels, in multiple domains. Against the demands, teacher labels are not often attached to real-world data; for the estimation of DIL, labels must be annotated accurately at great financial or human expense. The second drawback is hyperparameter selection. Most DILs involve some hyperparameters to balance the classification accuracy and the degree of invariance. It is known that most DILs give high predictive performance only when a hyperparameter is selected by using unseen test data; without using them, simple methods of hyperparameter selection fail to find a preferable hyperparameter. The thesis aims to mitigate the two problems. Aiming to overcome the first drawback, we propose a novel DIL framework; assuming the availability of data from multiple domains for a classification task with *coarser* labels than those of the target classification, for which the labeling cost is lower, we estimate an invariant predictor for the target classification task with training data gathered in a *single* domain. Moreover, we propose two methods of cross-validation (CV) for hyperparameter selection in our new DIL framework. Since we assume training data of a single domain for the target task, it is impossible to estimate the deviation of the risks over the domains. Our CV methods mitigate the difficulty by using additional coarser labeled data from multiple domains. Theoretical analysis reveals that our framework can estimate the desirable invariant predictor with a hyperparameter fixed correctly, and that such a preferable hyperparameter is chosen by the proposed CV methods under some conditions. The effectiveness of the

proposed framework, including the cross-validation, is demonstrated empirically with various datasets.

Acknowledgments

First of all, I would like to express my sincere thanks to my adviser Prof. Kenji Fukumizu, for his continuous and passionate encouragement. He respected my decision on research topics and helped me search for my *own* Ph.D. course. Moreover, he devoted much time to discussions and give me critical and suggestive feedback. He also helped me improve my writing. What I learned from him has shaped my important foundation as a researcher. When I came to the Institute of Statistical Mathematics, I was thinking about quitting my career as a researcher. However, the three years I spent with him reminded me of how I had felt when I had met research: *research is exciting!*

Further, it is a pleasure for me to thank all people at the Institute of Statistical Mathematics. This thesis would not been accomplished without their help. I am grateful to Professors Hironori Fujisawa, Shiro Ikeda and Hideitsu Hino, who often listened to my problems, and gave useful advice. Especially, Prof. Hironori Fujisawa often sent me encouraging e-mails. Prof. Hideitsu Hino often gave me useful advice for my career. I am thankful to Professors Keisuke Yano and Akifumi Okuno, who often invited me to lunch and dinner and give me encouraging comments. We often discussed what was and would be important for the future of statistics and machine learning, which was valuable and exciting time for me. Prof. Keisuke Yano also introduced me to interesting papers and books on statistics. I wish to thank Professors Yoshiyuki Ninomiya, Yukito Iba, Ryo Yoshida, Hisashi Noma for their valuable comments. Especially, Prof. Ryo Yoshida and Prof. Hisashi Noma made time for discussion and taught me characteristics of material and medical data respectively. I would like to thank Professors Mirai Tanaka, Hideto Nakashima for their encouraging comments. Dr. Yoichi Mototake helped me with arranging experiment settings. I am grateful to the secretaries of our Fukumizu-Lab Ms. Shioko Itsumi and Ms. Kyoko Akatsuka, who taught me how to treat my budget properly. I would like to offer my special thanks to my college classmates. Especially, Dr. Kazuharu Harada and Dr.

Hibiki Kaibuchi often gave me some useful information about Ph.D. life. Mr. Shunya Minami and Ms. Niu Yuanyuan often devoted their time to chat with me.

Finally, I would like to thank my parents, Shinji and Mie, for their unwavering support. They allowed me the freedom in life and have respected my decisions for twenty-seven years.

The research was financially supported by Grant-in-Aid for JSPS Fellows 20J21396, JST CREST JPMJCR2015, and JSPS Grant-in-Aid for Transformative Research Areas (A) 22H05106.

" I heard reiteration of the following claim:

Complex theories do not work, simple algorithms do.

I would like to demonstrate that in the area of science a good old principle is valid:

Nothing is more practical than a good theory."

Vladimir N Vapnik. *Statistical Learning Theory.* Wiley, 1998.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Out-of-Distribution Generalization in Machine Learning

Machine learning has made remarkable progress. It gives a human-level performance on image recognition [He et al., 2015], beats humans in various games [Moravčík et al., 2017, Silver et al., 2016], translates texts like a human translator [Devlin et al., 2019], and generates photographs that look like real ones [Goodfellow et al., 2014, Sohl-Dickstein et al., 2015]. Thanks to the numerous successes, our lives have become dramatically richer than ever before.

Despite the rapid progress, machine learning is still having a serious problem; they often inherit *spurious correlations* in training. Training data may contain features that are spuriously correlated to the labels of data, and machine learning models often learn such spurious correlations embedded in training data. As a result, they may fail to predict desired labels of test data generated by a different distribution from one to provide training data. The phenomena is observed in many areas of machine learning. We show two examples:

Image Recognition In classification of animal images, Deep Neural Networks (DNNs) tend to misclassify cows on sandy beaches, since most training pictures are taken in green pastures and DNNs inherit context information in training [Beery et al., 2018, Shane, 2018]. Another example is detecting cancer from X-ray scans. Systems trained with X-ray data in one hospital do not generalize well to other hospitals; systems unintentionally extract factors specific to a particular hospital in training

[AlBadawy et al., 2018, Perone et al., 2019, Heaven, 2020].

Fairness Hiring tools for predicting candidates based on resumes, developed by Amazon, were found to prefer men [Dastin, 2018]. The unfair decision stems from spurious correlation embedded in previous human decisions: the model's decision rule may depend entirely on a spurious correlation "gender", once it is found by a model in training.

In response to the problem of spurious correlations, it is recognized that one of the important issues for the future of machine learning is generalization to data generated by distributions that have different correlations from ones on the training distribution. Recently, this kind of generalization is often called *out-of-distribution (o.o.d.) generalization*.

1.2 Domain Invariance Learning and Its Limitations

Domain Invariance Learning (DIL) is a rapidly developed approach for the out-of-distribution generalization [Arjovsky et al., 2020, Ahuja et al., 2020, Rothenhäusler et al., 2021, Heinze-Deml et al., 2018, Peters et al., 2015, Koyama and Yamaguchi, 2021, Krueger et al., 2021, Liu et al., 2021a,b, Creager et al., 2021, Parascandolo et al., 2022, Lu et al., 2022]. Their proposed estimator $f = w \circ \Phi : \mathcal{X} \to \mathcal{Y}$, which maps an input $x \in \mathcal{X}$ to its predictive class label $y \in \mathcal{Y}$, is composed of two maps: (i) a feature map $\Phi : \mathcal{X} \to \mathcal{H}$, which is called *a domain invariance* (defined in Chapter 2), from the input space $\mathcal{X}$ to the feature space $\mathcal{H}$, and (ii) a predictor $w : \mathcal{H} \to \mathcal{Y}$ which estimates the label of the feature $\Phi(x) \in \mathcal{H}$. The intuitive reason why f has high o.o.d. generalization performance is that Φ removes spurious features (*e.g.*, contexts of images) from $x \in \mathcal{X}$, and hence, f can predict labels with ignoring spurious correlations embedded in training data. Their training is implemented by training data from *multiple* domains[1].

While the DIL approaches have attracted much attention, they have two shortcomings in practice:

[1]In this thesis, we use the term *domain* to specify the distributions or random variables.

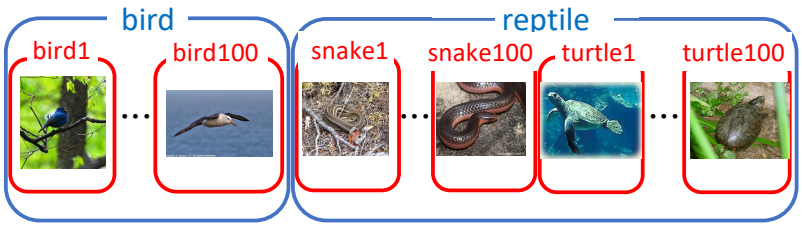

Figure 1.1: Example of Coarser Labels.

Problem 1: Expensive annotation Requiring training data from multiple domains may hinder wide applications; preparing training data in many domains often involves expensive data annotation, especially when the class number of target classification is large. In real-world data, labels may be missing [Pham et al., 2021, Zheng et al., 2017, Gu et al., 2020, Lakshminarayan et al., 1999, Tan et al., 2013] or incomplete; in some cases, data may only specify classes to which the image does *not* belong [Cour et al., 2011, Yan and Guo, 2020, Xu et al., 2019]. Such data with insufficient annotation are not directly applicable to the standard IL methods; they must be re-annotated accurately, often at great financial or human expense. The high cost drives a strong need to establish a new DIL framework with lower annotation costs.

Problem 2: Hyperparameter Selection The other important problem in DIL is hyperparameter selection. Most DIL methods involve some hyperparameters to balance the classification accuracy and the degree of invariance. As Krueger et al. [2021], Gulrajani and Lopez-Paz [2023] point out, in the literature of DIL, the best performances of invariance had often been achieved by selecting the hyperparameters using test data from unseen domains. Moreover, Gulrajani and Lopez-Paz [2023] numerically demonstrated that, without using test data, simple methods of hyperparameter selection fail to find a preferable hyperparameter. It demonstrates a strong need for establishing an appropriate method of hyperparameter selection for DILs.

1.3 Contribution

The present thesis tries to mitigate the two problems stated in the last section.

Toward aiming to solve the first problem, we propose a novel DIL framework for

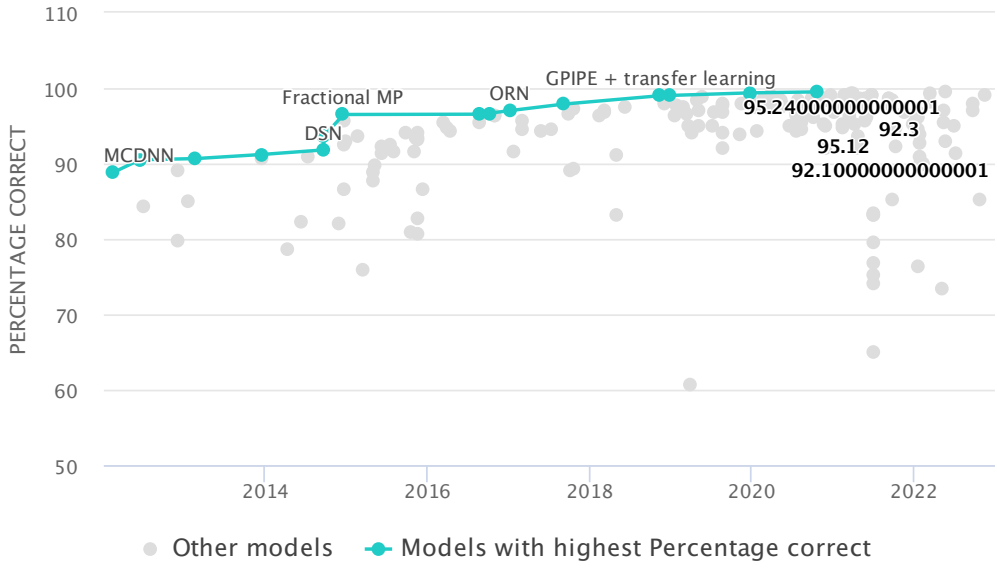

Figure 1.2: The Top Classification Accuracy for CIFAR-10 [Paperswithcode.com, 2023a].

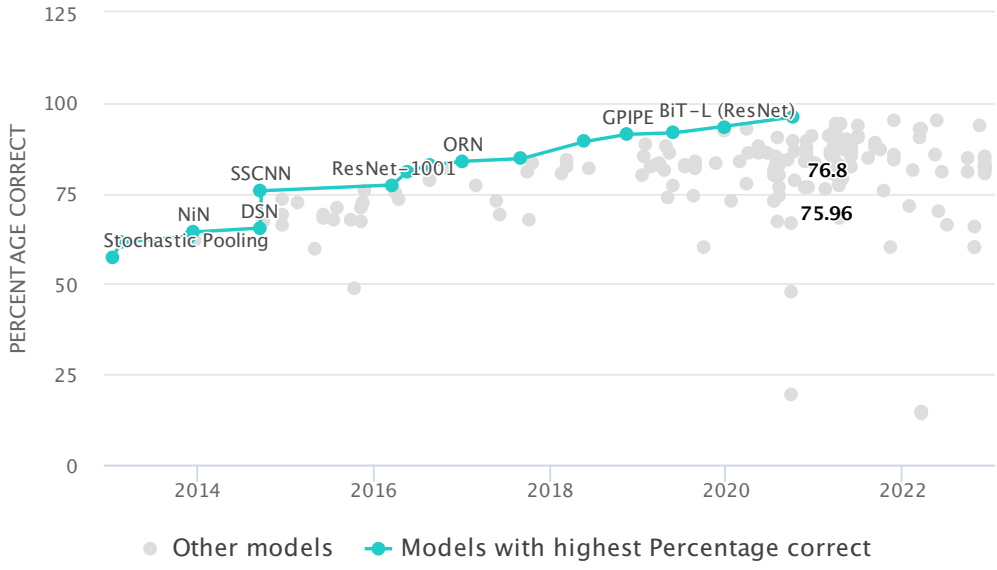

Figure 1.3: The Top Classification Accuracy for CIFAR-100 [Paperswithcode.com, 2023b].

the situation where the training data of target classification are given in only *one* domain, while the task with *coarser* labels than those of the target classification, which needs lower annotation cost, has data from multiple domains. Figure 1.1 shows an example of coarser labels. Consider the case where a target classification has 300 labels (colored red) $\{\text{bird}_1, ..\text{bird}_{100}, \text{snake}_1, ..., \text{snake}_{100}, \text{turtle}_1, ..., \text{turtle}_{100}\}$ corresponding to 300 species. Then, the binary labels (colored blue) $\{\text{bird}, \text{reptile}\}$ are an example of coarser labels. The annotation cost will be drastically reduced by changing the labels to much coarser ones. The following two examples are two such scenarios that can have the advantage of coarser labels in terms of annotation cost and quality.

Manual annotation In practice, annotation is done by humans through crowd-sourcing or asking annotation vendors. Then, the changes from original to coarser labels reduce the annotation cost from the following two viewpoints. At first, the decrease in the number of classes reduces annotation time per image. Let us consider the example in Fig 1.1. Then, annotation of coarser labels takes only a few seconds, since we may judge whether or not the image includes any birds. On the other hand, annotation of 300 labels will demand more time; it will take some time to correspond numbers to the 300 labels, judge which classes an image belong to, and attach class numbers. Secondly, coarser labels demand lower expert knowledge than ones needed in original labels. In the example in Fig. 1.1, annotating the sub-types of birds, snakes, and turtles would require expert knowledge. It is highly probable that an annotation vendor would charge very high fees or decline such a request. On the other hand, annotation of coarser labels (e.g., at the levels of bird or reptile) is much easier so that we can rely on non-experts or crowdsourcing at a lower cost to obtain annotated datasets for many domains.

Machine annotation Annotations of labels may be done by a pre-trained classifier on the Internet as well as by humans with crowd-sourcing. Recent progress in artificial intelligence enables us to access a high-quality, pre-trained classifier such as a ResNet [He et al., 2016] pre-trained with ImageNet. Note that classification ability is much higher for a task with a smaller number of classes. We can see the fact from Fig. 1.2 and Fig. 1.3; the top classification accuracy for CIFAR-10 (10 classes) attained almost 100% in 2016, while SOTA for CIFAR-100 (100 classes) at that time was about 75%. The figures show that classifiers can annotate labels more precisely, as the numbers

of classes become smaller, and therefore. as classifications become coarser.

From the above discussions, we can see that the new DIL framework significantly reduces the annotation cost in comparison with previous DIL methods; we need exhausting annotation of target classification only for one domain and just coarser labels, which demands lower cost, for other domains.

As for the second problem, we propose two methods of cross-validation (CV) for hyperparameter selection in our new DIL framework. Since we assume training data of a single domain for the target task, it is impossible to estimate the deviation of the risks over the domains. Our CV methods mitigate the difficulty by using additional coarser labeled data from multiple domains. Theoretical analysis proves that our methods select a hyperparameter correctly under some conditions.

1.4 Outline

The Ph.D. thesis is organized as follows. In Chapter 2, we review previous DILs and their shortcomings. Chapter 3 is the main part of the thesis. In Section 3.1, we establish a novel framework of DIL, which estimates an invariant predictor from single domain data, assuming additional data from multiple domains for a classification task with coarser labels. In Section 3.2, we propose two methods of cross-validation for selecting hyperparameters without accessing any samples from unseen target domains under the framework. In Section 3.3, we mathematically prove that our framework can estimate a correct invariant predictor with a hyperparameter fixed correctly and that such a preferable hyperparameter is selected by the proposed CV methods under some settings. Proofs of theorems in the thesis are contained in Section 4. In Chapter 5, we review some related works. In Chapter 6, we numerically demonstrate that the proposed framework extracts an invariant predictor more effectively than other existing methods. Finally, Chapter 7 is devoted to some concluding remarks.

Chapters 3, 4, and 6 are mostly based on the conference paper [Toyota and Fukumizu, 2022].

Chapter 2

Preliminaries

In this chapter, we mathematically formulate out-of-distribution generalization problem. Moreover, we review conventional Domain Invariance Learnings (DILs) and their limitations.

2.1 Notations

Throughout this thesis, the spaces of objective and response variables are denoted by $\mathcal{X}$ and $\mathcal{Y}$, respectively. For given predictor $f : \mathcal{X} \to \mathcal{Y}$ and random variable (X, Y) on $\mathcal{X} \times \mathcal{Y}$ with its probability $P_{X,Y}$, $\mathcal{R}^{(X,Y)}(f)$ denotes the risk of f on (X, Y); *i.e.*, $\mathcal{R}^{(X,Y)}(f) := \int l(f(x), y) dP_{X,Y}$, where $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is a loss function. For $m \in \mathbb{N}_{>0}$, $[m]$ denotes the set $\{1, ..., m\}$. For a finite set A, $|A| \in \mathbb{N}$ denotes the number of elements in A.

2.2 Mathematical Formulation of Out-of-Distribution Generalization

We mathematically formulate o.o.d. generalization problem following Arjovsky et al. [2020].

We assume that the joint distribution of data (X^e, Y^e) depends on the domain $e \in \mathcal{E}$, and consider the dependence of a predictor f on the domain variable e. Suppose we are given training datasets $\mathcal{D}^e := \{(x_i^e, y_i^e)\}_{i=1}^{n^e} \sim P_{X^e,Y^e}$ i.i.d. from domains $\mathcal{E}_{tr} \subset \mathcal{E}$. The final goal of the o.o.d. problem is to predict a desired label $Y^e \in \mathcal{Y}$ from $X^e \in \mathcal{X}$ for larger target domains $\mathcal{E} \supset \mathcal{E}_{tr}$. To address the issue

caused by spurious correlations mathematically, Arjovsky et al. [2020] introduced the o.o.d. risk

$$\mathcal{R}^{o.o.d.}(f) := \max_{e \in \mathcal{E}} \mathcal{R}^e(f), \tag{2.1}$$

where $\mathcal{R}^e(f) := \mathcal{R}^{(X^e, Y^e)}(f)$. This is the worst-case risk over $\mathcal{E}$, including unseen domains $\mathcal{E} \setminus \mathcal{E}_{tr}$. Through the concept o.o.d. risk, o.o.d. generalization is mathematically formulated as follows:

For a given parametric model $\{f_\theta\}_{\theta \in \Theta}$,
how can we find a model parameter $\theta^* \in \Theta$ which minimizes $\mathcal{R}^{o.o.d.}(f_\theta)$?

2.3 Domain Invariance Learning

DIL [Arjovsky et al., 2020, Ahuja et al., 2020, Rothenhäusler et al., 2021, Heinze-Deml et al., 2018, Koyama and Yamaguchi, 2021, Krueger et al., 2021, Liu et al., 2021a,b, Creager et al., 2021, Parascandolo et al., 2022, Lu et al., 2022] is a rapidly developed approach for o.o.d. generalization. The framework train a *domain invariance* defined as follows:

Definition 1. *We call $\Phi : \mathcal{X} \to \mathcal{H}$ a domain invarince or domain invariance feature when conditional distributions $P_{Y^{e_1}|\Phi(X^{e_1})}$ and $P_{Y^{e_2}|\Phi(X^{e_2})}$ satisfy $P_{Y^{e_1}|\Phi(X^{e_1})} = P_{Y^{e_2}|\Phi(X^{e_2})}$[1] for any $e_1, e_2 \in \mathcal{E}$.*

Here, this definition is a domain invariance based on conditional independence [Peters et al., 2015, Koyama and Yamaguchi, 2021, Rojas-Carulla et al., 2018], while Arjovsky et al. [2020], Ahuja et al. [2020] use a different type of domain invariances based on $\operatorname{argmin}_w \mathcal{R}^e(w \circ \Phi)$ instead of $P_{Y^e|\Phi(X^e)}$. Throughout the thesis, we carry an argument by adopting the definition based on conditional independence.

In the following section, we review when and why the concept *domain invariance* was proposed, and how it came to be used for out-of-distribution generalization problem.

[1] Throughout the thesis, we write $P_{Y^{e_1}|\Phi(X^{e_1})} = P_{Y^{e_2}|\Phi(X^{e_2})}$ when

$$P_{Y^{e_1}|\Phi(X^{e_1})=\Phi(x)} = P_{Y^{e_2}|\Phi(X^{e_2})=\Phi(x)}$$

holds as an equation between distributions, for any $x \in \mathcal{X}$.

2.3.1 Origin of Domain Invariance

The concept *domain invariance* is firstly proposed in statistical causal discovery [Peters et al., 2015]. In the paper, they address the case where $\mathcal{X} = \mathbb{R}^p$ and $\mathcal{Y} = \mathbb{R}$. Moreover, they assume that (X^e, Y^e) satisfies the following condition:

Assumption 2. *There exists a vector of coefficients* $\gamma^* = (\gamma_1^*, ..., \gamma_p^*)^t \in \mathbb{R}^p$ *with support* $S^* = \{k \mid \gamma_k^* \neq 0\} \subset \{1, ..., p\}$ *that satisfies*

$$\forall e \in \mathcal{E},\ X^e \text{ has an arbitrary distribution and}$$
$$Y^e = \mu + X^e \cdot \gamma^* + \varepsilon^e, \varepsilon^e \sim F_\varepsilon \text{ and } \varepsilon^e \perp\!\!\!\perp X^e_{S^*},$$

where $X^e_{S^*} \in \mathbb{R}^{|S^*|}$ *is* S^**-components of* X^e*,* μ *is an intercept term,* ε^e *is a random noise with mean zero, finite variance and the same distribution* F_ε *across* $e \in \mathcal{E}$*.*

Under the assumption, a projection Φ_{S^*}, which maps $x \in \mathcal{X}$ to the subset S^* of its component, becomes a domain invariance among $e \in \mathcal{E}$; namely, $P_{Y^{e_1}|\Phi_{S^*}(X^{e_1})} = P_{Y^{e_2}|\Phi_{S^*}(X^{e_2})}$ holds for any $e_1, e_2 \in \mathcal{E}$. Peters et al. [2015] show that *plausible causal predictors*, domain invariances which satisfy the following properties, are useful for causal discovery:

Definition 3. *We call the variables* $S \subset \{1, ..., p\}$ *plausible causal predictors under* $\mathcal{E}$ *if there exists* $\exists \gamma \in \mathbb{R}^p$ *such that the following null hypothesis is true:*

$$\gamma_k = 0 \text{ if } \gamma_k \notin S \text{ and } \begin{cases} \exists F_\varepsilon \text{ for all } e \in \mathcal{E} \\ Y^e = X^e \cdot \gamma + \varepsilon^e \text{ where } \varepsilon^e \perp\!\!\!\perp X^e_S \text{ and } \varepsilon^e \sim F_\varepsilon. \end{cases}$$

As shown in Peters et al. [2015], under the case where (X^{e_1}, Y^{e_1}) follows a Gaussian structural equation model and (X^e, Y^e) for $e \in \mathcal{E} - \{e_1\}$ have some conditions,

$$\bigcap_{S:\text{plausible causal predictors}} S$$

coincides with the parents of Y^{e_1}; in other words, the parents of Y^{e_1} can be specified if we can identify all plausible causal predictors. In detail, see Theorem 4 and 5 in Peters et al. [2015]. Peters et al. [2015] also proposed a method to estimate $\bigcap_{S:\text{plausible causal predictors}} S$ with confidence intervals, and demonstrate its effectiveness by a gene perturbation problem.

2.3.2 Domain Invariances for Out-of-Distribution Generalization

Recently, a domain invariance has become utilized for the o.o.d. generalization problem [Arjovsky et al., 2020]. Their proposed estimator $f = w \circ \Phi$ is composed of two maps: a domain invariance $\Phi : \mathcal{X} \to \mathcal{H}$, which realizes a feature of $x \in \mathcal{X}$ in the feature space $\mathcal{H}$, and a predictor $w : \mathcal{H} \to \mathcal{Y}$ of labels. Here, note that the domain invariance Φ is not necessarily a variable selection same as one in the last section; in an image recognition task, a feature map that removes contexts can not be necessarily represented by some variable selection. The estimation of an invariant predictor is implemented by solving the following optimization problem:

$$\min_{\Phi \in \mathcal{I}_{tr}, w:\mathcal{H}\to\mathcal{Y}} \sum_{e \in \mathcal{E}_{tr}} \mathcal{R}^e(w \circ \Phi), \tag{2.2}$$

where $\mathcal{I}_{tr}$ is the set of domain invariances among training domains $\mathcal{E}_{tr}$:

$$\mathcal{I}_{tr} := \left\{ \Phi : \mathcal{X} \to \mathcal{H} \mid P_{Y^{e_1}|\Phi(X^{e_1})} = P_{Y^{e_2}|\Phi(X^{e_2})} \text{ for any } e_1, e_2 \in \mathcal{E}_{tr} \right\}.$$

The following theorem ensures that the minimum of the bi-level optimization problem (2.2) also minimizes o.o.d. risk under some conditions and simplifications:

Theorem 4 (o.o.d. optimality of IRM optimization problem)**.** *Let $\mathcal{X} := \mathcal{X}_1 \times \mathcal{X}_2$ and $\mathcal{Y} := \mathbb{R}$ where $\mathcal{X}_1 := \mathbb{R}^{n_1}$ and $\mathcal{X}_2 := \mathbb{R}^{n_2}$ with $n_1, n_2 \in \mathbb{N}$. Let (X_1^I, Y^I) be a fixed random variable on $\mathcal{X}_1 \times \mathcal{Y}$. For simplicity of analysis, the domain set $\mathcal{E}$ is defined by all the random variables (X, Y) which satisfy $P_{Y|\Phi^{\mathcal{X}_1}(X)} = P_{Y^I|X_1^I}$, where $\Phi^{\mathcal{X}_1} : \mathcal{X} \to \mathcal{X}_1$ is a projection of $x \in \mathcal{X}$ onto $\mathcal{X}_1$; namely,*

$$\{(X^e, Y^e)\}_{e\in\mathcal{E}} := \left\{ (X, Y) : \textit{a random variable on } \mathcal{X} \times \mathcal{Y} \,\middle|\, P_{Y|\Phi^{\mathcal{X}_1}(X)} = P_{Y^I|X_1^I} \right\}.$$

Let a loss function l be the least square loss; for given $f : \mathcal{X} \to \mathcal{Y}$,

$$\mathcal{R}^e(f) := \int \|f(x) - y\|^2 dP_{X^e,Y^e}.$$

To avoid discussing the non-trivial effects of nonlinear domain invariance Φ, we focus on the simplified case of variable selections; namely, for the finite training domains $\{(X^e, Y^e)\}_{\mathcal{E}_{tr}} \subset \{(X^e, Y^e)\}_{\mathcal{E}}$, a domain invariance Φ in the optimization problem

(2.2) only runs among $\mathcal{I}_{tr}^{v.s.}$ *defined by*

$$\mathcal{I}_{tr} \supset \mathcal{I}_{tr}^{v.s.} := \left\{ \Phi : \textit{a variable selection} \;\middle|\; \begin{array}{l} P_{Y^{e_1}|\Phi(X^{e_1})} = P_{Y^{e_2}|\Phi(X^{e_2})} \textit{ for any} \\ (X^{e_1}, Y^{e_1}), (X^{e_2}, Y^{e_2}) \in \{(X^e, Y^e)\}_{\mathcal{E}_{tr}}. \end{array} \right\}.$$

For a projection Φ*, let* Φ_i *denote the* $\mathcal{X}_i$*-component of* Φ *(*$i = 1, 2$*). If* Φ *has or has not an* $\mathcal{X}_i$*-component, we write* $\mathrm{Im}\Phi_i \neq \emptyset$ *or* $\mathrm{Im}\Phi_i = \emptyset$ *respectively. For* Φ_i *(*$i = 1, 2$*),* $\Phi_i^{\perp}$ *denotes the projection onto orthogonal complements of* $\mathrm{Im}\Phi_i$ *with respect to* $\mathcal{X}_i$*; namely,* $\mathrm{Im}\Phi_i \bigotimes \mathrm{Im}\Phi_i^{\perp} \simeq \mathcal{X}_i$

Assume that $\mathcal{E}$ *and* $\mathcal{E}_{tr}$ *satisfy the following conditions:*

- $\mathcal{I}_{tr}^{v.s.} = \mathcal{I}^{v.s.}$ *holds, where*

$$\mathcal{I}^{v.s.} := \left\{ \Phi : \textit{a variable selection} \;\middle|\; \begin{array}{l} P_{Y^{e_1}|\Phi(X^{e_1})} = P_{Y^{e_2}|\Phi(X^{e_2})} \textit{ for any} \\ (X^{e_1}, Y^{e_1}), (X^{e_2}, Y^{e_2}) \in \{(X^e, Y^e)\}_{\mathcal{E}}. \end{array} \right\}.$$

- *For any variable selection* Φ *with* $\mathrm{Im}\Phi \subsetneq \mathcal{X}_1$*, there exist* $\bar{x}, \bar{\bar{x}}$ *and* $\bar{\bar{\bar{x}}} \in \mathcal{X}$ *with* $\Phi_1^{\perp}(\bar{\bar{x}}) \neq \Phi_1^{\perp}(\bar{\bar{\bar{x}}})$ *such that*

$$P_{Y^I|X^I=(\Phi_1(\bar{x}),\Phi_1^{\perp}(\bar{\bar{x}}))} \neq P_{Y^I|X^I=(\Phi_1(\bar{x}),\Phi_1^{\perp}(\bar{\bar{\bar{x}}}))}$$

Then, the inclusion

$$\mathrm{argmin}_{\Phi \in \mathcal{I}_{tr}^{v.s.}, w:\mathcal{H}\to\mathcal{Y}} \sum_{e\in\mathcal{E}_{tr}} \mathcal{R}^e(w \circ \Phi) \subset \operatorname*{argmin}_{f:\mathcal{X}\to\mathcal{Y}} \mathcal{R}^{o.o.d.}(f)$$

holds. Here, w *and* f *run among all measurable functions.*

Remark 1 In our variable selection setting, the feature map Φ is chosen from the projections of x to a subset of its components. For example, Φ may be $\Phi(x_1, x_2, x_3) = (x_1, x_3)$ when x is three-dimensional. This type of IL appears practically in causal inference [Peters et al., 2015, Heinze-Deml et al., 2018] and regression [Rojas-Carulla et al., 2018].

Remark 2 We will add some remarks about the first condition $\mathcal{I}_{tr}^{v.s.} = \mathcal{I}^{v.s.}$. In general, the inclusion $\mathcal{I}_{tr}^{v.s.} \subset \mathcal{I}^{v.s.}$ holds by definitions of $\mathcal{I}_{tr}^{v.s.}$ and $\mathcal{I}^{v.s.}$. The equality $\mathcal{I}_{tr}^{v.s.} = \mathcal{I}^{v.s.}$ does not necessarily hold. Arjovsky et al. [2020] investigated necessary conditions for the equality.

The theorem is proven in Section 4.1. While the theorem ensures the validness of (2.2), it is still a challenging optimization problem since each constraint calls an inner optimization routine. So, Arjovsky et al. [2020] introduce the following objective function:

$$\sum_{e\in\mathcal{E}_{tr}} \mathcal{R}^e(\Phi) + \lambda \cdot \|\nabla_{w=1.0}\mathcal{R}^e(w\cdot\Phi)\|^2. \tag{2.3}$$

Arjovsky et al. [2020] model Φ by DNNs and minimize (2.3) by conventional optimization procedure, such as Adam [Kingma and Ba, 2015].

2.3.3 Species of Domain Invariance Learning

The primary work [Arjovsky et al., 2020] inspired various research concerning DILs [Ahuja et al., 2020, Krueger et al., 2021, Liu et al., 2021a,b, Creager et al., 2021, Parascandolo et al., 2022, Lu et al., 2022, Rosenfeld et al., 2021, Kamath et al., 2021, Lin et al., 2022].

The theoretical properties of the optimization problem (2.2) and the objective function (2.3) were analyzed [Rosenfeld et al., 2021, Kamath et al., 2021]. Rosenfeld et al. [2021] reveal conditions of $\{(X^e, Y^e)\}_{e\in\mathcal{E}}$ under which the optimization problem (2.2) succeeds or fails to minimize the o.o.d. risk (2.1), assuming that data are generated from a simple linear structural equation model. Rosenfeld et al. [2021] also show that the objective function (2.3) fails to minimize the o.o.d. risk (2.1) under some non-linear structural equation models. Kamath et al. [2021] show that the objective function (2.3) fails to minimize the bi-leveled optimization problem (2.2) even when $\{(X^e, Y^e)\}_{e\in\mathcal{E}}$ follows a simple linear model. Moreover, Kamath et al. [2021] find a linear structural equation model where (2.2) fails to minimize the o.o.d. risk (2.1).

Another important direction is proposing new learning frameworks to improve the primary work [Arjovsky et al., 2020]. Ahuja et al. [2020] introduced a new objective function with the help of game theory. Krueger et al. [2021] used domain invariances based on risk $\mathcal{R}^e(f)$ instead of conditional independence $P(Y^e|\Phi(X^e))$. Parascandolo et al. [2022] utilized domain invariances based on loss landscape among domains $\mathcal{E}$. Lin et al. [2022] introduced Bayesian inference into conventional DILs and numerically shows that their new DIL framework prevents models from overfitting to training data. Lu et al. [2022] propose new objective function with the helps of variational autoencoders [Kingma and Ba, 2015, Rezende et al., 2014]. While common DIL methods assume that the training examples are partitioned into "domains", Liu et al.

[2021a,b], Creager et al. [2021] focus on the setting where such partitions are not provided. Creager et al. [2021] proposed the methods to attach domain labels that can then be used to apply an invariant learning algorithm. Liu et al. [2021a,b] proposed a novel domain invariance learning framework without domain labels.

2.4 Limitations of Domain Invariance Learning

Previous DILs have two shortcomings in practice:

Limitation I: Annotation cost problem Conventional DILs often demand expensive and exhausting annotation. Please consider an image classification task and let $\mathcal{Y}$ be a set of finite class labels. Previous DILs estimate a domain invariance based on the discrepancy of $P_{Y^e|\Phi(X^e)}$ among domains $e \in \mathcal{E}$, and hence, their estimation demands training data $\mathcal{D}^e := \{(x_i^e, y_i^e)\}$ from multiple domains $\mathcal{E}_{tr} \subset \mathcal{E}$. In practice, labels y_i^e are not attached to all images x_i^e generated by multiple domains $\mathcal{E}_{tr} \subset \mathcal{E}$; labels y_i^e may be missing [Pham et al., 2021, Zheng et al., 2017, Gu et al., 2020, Lakshminarayan et al., 1999, Tan et al., 2013] or in some cases, may only specify classes to which the image does *not* belong [Cour et al., 2011, Yan and Guo, 2020, Xu et al., 2019]. For the application of previous DILs, labels must be attached, often at great financial or human expense.

Limitation II: Hyperparamter selection problem Objective functions in most DILs have a hyperparameter λ to select, as (2.3) in Arjovsky et al. [2020]. The hyperparameter selection in DIL has special difficulty, however; because the o.o.d. problem needs to predict Y^e on unseen domains, λ must be chosen without accessing any data in such unseen domains. It was reported that the success of DIL methods depends strongly on the careful choice of hyperparameters, and some of the results even used data from unseen domains in the choice [Gulrajani and Lopez-Paz, 2023, Krueger et al., 2021]. Gulrajani and Lopez-Paz [2023] reported also experimental results of various DIL methods with two CV methods, training-domain validation (Tr-CV) and leave-one-domain-out validation (LOD-CV), and showed that the CV methods failed to select preferable hyperparameters. In the Colored MNIST experiment, for example, the accuracy of Arjovsky et al. [2020] is 52.0% at best, which is about a random guess level.

In the following chapter, we propose a new DIL framework to mitigate the annotation problem, and then propose two methods of cross-validation (CV) for hyperparameter selection in our new DIL framework.

Chapter 3

Proposed Method

In the chapter, we propose a novel DIL framework to mitigate the annotation cost problem of conventional DILs. In the new DIL framework, we consider the situation where the training data of target classification are given in only *one* domain e^*, while the task with coarser labels, which needs lower annotation cost, has data from multiple domains $\mathcal{E}_{ad} \subset \mathcal{E}$. Moreover, we propose two CV methods for the new DIL framework. In the remaining chapters, we consider an image classification task and let $\mathcal{X}$ and $\mathcal{Y}$ be spaces of input images and finite class labels.

3.1 Domain Invariance Estimation by Coarser Label Data

Our goal is to make a domain invariant predictor from a single training domain $\mathcal{E}_{tr} = \{e^*\}$. In this case, (2.2) is reduced to the empirical risk minimization $\min_f \mathcal{R}^{e^*}(f)$ on e^*, and therefore the standard DIL framework is not able to extract a domain invariance.

In this chapter, we introduce an assumption that additional data $\mathcal{D}^e_{ad}$ for another task (X^e, Z^e), which have *coarser* labels than those of (X^e, Y^e), is available with respect to multiple domains $\mathcal{E}_{ad} \subset \mathcal{E}$. Formally, Z^e is represented as $Z^e = g(Y^e)$ with a surjective label mapping $g : \mathcal{Y} \to \mathcal{Z}$ from the original to coarser labels. The example in Section 1.3 is formalized by a surjevtive function g as, setting $\mathcal{Y} := \{\text{bird}_1, .., \text{bird}_{100}, \text{turtle}_1, .., \text{turtle}_{100}, \text{snake}_1, .., \text{snake}_{100}\}$ and $\mathcal{Z} := \{\text{bird}, \text{reptile}\}$, $g(y) := \text{bird}$ if $y = \text{bird}_i$ $(i \in \{1, 2, ..., 100\})$ and $g(y) := \text{reptile}$ else.

By making use of $\{\mathcal{D}_{ad}^e\}_{e\in\mathcal{E}_{ad}}$, our objective for the domain invariance prediction is given by

$$\min_{\Phi\in\mathcal{I}_{ad},w:\mathcal{H}\to\mathcal{Y}} \mathcal{R}^{e^*}(w\circ\Phi), \tag{3.1}$$

where $\mathcal{I}_{ad}$ is the set of domain invariances:

$$\mathcal{I}_{ad} := \left\{\Phi:\mathcal{X}\to\mathcal{H} \mid P_{g(Y^{e_1})|\Phi(X^{e_1})} = P_{g(Y^{e_2})|\Phi(X^{e_2})} \text{ for any } e_1, e_2 \in \mathcal{E}_{ad}\right\}.$$

Note that (3.1) evaluates the risk with a single training domain while the domain invariances are given by additional data of multiple domains. The following theorem ensures that the minimum of (3.1) also minimizes o.o.d. risk under some settings.

Theorem 5. *Assume that the settings of $\mathcal{X}$, $\mathcal{Y}$, the loss function $l(\cdot,\cdot)$ and $\mathcal{E}$ and notations with respect to a variable selection Φ are the same as ones in Theorem 4. For $\{(X^e,Y^e)\}_{e\in\mathcal{E}_{ad}}\subset\{(X^e,Y^e)\}_{e\in\mathcal{E}}$, define $\mathcal{I}_{ad}^{v.s.}$ by*

$$\mathcal{I}_{ad}\supset\mathcal{I}_{ad}^{v.s.} := \left\{\Phi : \textit{a variable selection} \;\middle|\; \begin{array}{l} P_{g(Y^{e_1})|\Phi(X^{e_1})} = P_{g(Y^{e_2})|\Phi(X^{e_2})} \textit{ for any} \\ \quad (X^{e_1},Y^{e_1}),(X^{e_2},Y^{e_2})\in\{(X^e,Y^e)\}_{\mathcal{E}_{ad}} \end{array}\right\}.$$

Assume that $\mathcal{E}$ and $\mathcal{E}_{ad}$ satisfy the following conditions:

- *$\mathcal{I}_{ad}^{v.s.}=\mathcal{I}^{v.s.}$ holds, where*

$$\mathcal{I}^{v.s.} := \left\{\Phi : \textit{a variable selection} \;\middle|\; \begin{array}{l} P_{Y^{e_1}|\Phi(X^{e_1})} = P_{Y^{e_2}|\Phi(X^{e_2})} \textit{ for any} \\ \quad (X^{e_1},Y^{e_1}),(X^{e_2},Y^{e_2})\in\{(X^e,Y^e)\}_{\mathcal{E}} \end{array}\right\}.$$

- *For any variable selection Φ with $\mathrm{Im}\Phi\subsetneq\mathcal{X}_1$, there exist $\bar{x},\bar{\bar{x}}$ and $\bar{\bar{\bar{x}}}\in\mathcal{X}$ with $\Phi_1^\perp(\bar{\bar{x}})\neq\Phi_1^\perp(\bar{\bar{\bar{x}}})$ such that*

$$P_{Y^I|X^I=(\Phi_1(\bar{x}),\Phi_1^\perp(\bar{\bar{x}}))} \neq P_{Y^I|X^I=(\Phi_1(\bar{x}),\Phi_1^\perp(\bar{\bar{\bar{x}}}))}$$

Then, the inclusion

$$\mathrm{argmin}_{\Phi\in\mathcal{I}_{ad}^{v.s.},w:\mathcal{H}\to\mathcal{Y}}\mathcal{R}^{e^*}(w\circ\Phi)\subset\operatorname*{argmin}_{f:\mathcal{X}\to\mathcal{Y}}\mathcal{R}^{o.o.d.}(f)$$

holds. Here, w and f run among all measurable functions.

The theorem is proven in Section 4.2.

3.2 Construction of Objective Function

Among several candidates of the loss and model design, we focus a probabilistic output case and evaluate its error by the cross entropy loss; that is, we model w by $p_\theta : \mathcal{H} \to \mathcal{P}_\mathcal{Y}$, where $\mathcal{P}_\mathcal{Y}$ denotes the set of probabilities on $\mathcal{Y}$ and θ denotes a model parameter. The risk is then written by

$$\mathcal{R}^e(p_\theta \circ \Phi) = \int -\log p_\theta(Y^e|\Phi(X^e))dP_{X^e,Y^e}.$$

We aim to solve (3.1) by minimizing the following objective function:

$$\begin{aligned}\text{Objective}(\theta, \Phi) := \hat{\mathcal{R}}^{e^*}(p_\theta \circ \Phi) \\ + \lambda \cdot (\text{Dependence measure of } P_{g(Y^e)|\Phi(X^e)} \text{ on } e \in \mathcal{E}_{ad}).\end{aligned} \tag{3.2}$$

Here, $\hat{\mathcal{R}}^{e^*}(p_\theta \circ \Phi)$ denotes the empirical risk of $p_\theta \circ \Phi$ on the training domain $\mathcal{E}_{tr} = \{e^*\}$ evaluated by $\mathcal{D}^{e^*}$: $\hat{\mathcal{R}}^{e^*}(p_\theta \circ \Phi) := -\frac{1}{|\mathcal{D}^{e^*}|}\sum_{(x^{e^*},y^{e^*})\in\mathcal{D}^{e^*}} \log p_\theta(y^{e^*}|\Phi(x^{e^*}))$. While we can consider some variations of domain invariance regularization, we adopt the one used in Arjovsky et al. [2020] and construct an objective function as

$$\text{Objective}(\theta, \theta_{ad}, \Phi) := \hat{\mathcal{R}}^{e^*}(p_\theta \circ \Phi) + \lambda \cdot \sum_{e\in\mathcal{E}_{ad}} \|\nabla_{\hat{\theta}_{ad}=\theta_{ad}} \hat{\mathcal{R}}^{(X^e,Z^e)}(p^{\mathcal{Z}|\mathcal{H}}_{\hat{\theta}_{ad}} \circ \Phi)\|^2. \tag{3.3}$$

Here, $p^{\mathcal{Z}|\mathcal{H}}_\theta : \mathcal{H} \to \mathcal{P}_\mathcal{Z}$ and p_θ are the linear logistic regression model same as Arjovsky et al. [2020], Φ is a nonlinear neural network, and

$$\hat{\mathcal{R}}^{(X^e,Z^e)}(p^{\mathcal{Z}|\mathcal{H}}_{\theta_{ad}} \circ \Phi) := -\frac{1}{|\mathcal{D}^e_{ad}|}\sum_{(x^e,z^e)\in\mathcal{D}^e_{ad}} \log p^{\mathcal{Z}|\mathcal{H}}_{\theta_{ad}}(z^e|\Phi(x^e)).$$

It is not obvious if the regularization term in (3.3) is valid as a dependence measure of $P_{g(Y^e)|\Phi(X^e)}$ since it was proposed for another type of domain invariance based on $\text{argmin}_w \mathcal{R}^e(w \circ \Phi)$. The next lemma shows that these notions of domain invariance are the same in the current setting.

Lemma 6. *When modeling w by conditional probabilities, the following statements are equivalent:*

$$\begin{aligned}&P_{Z^e|\Phi(X^e)} \textit{ does not depend on } e \\ &\qquad \Leftrightarrow \text{argmin}_{p^{\mathcal{Z}|\mathcal{H}}_{\theta_{ad}}} \mathcal{R}^{(X^e,Z^e)}(p^{\mathcal{Z}|\mathcal{H}}_{\theta_{ad}} \circ \Phi) \textit{ does not depend on } e,\end{aligned}$$

where model $p_{\theta_{ad}}^{\mathcal{Z}|\mathcal{H}}$ *in* $\operatorname{argmin}_{p_{\theta_{ad}}^{\mathcal{Z}|\mathcal{H}}}$ *runs over all probability densities.*

Proof. Noting that $\operatorname{argmin}_{\theta_{ad}} \mathcal{R}^{(X^e,Z^e)}(p_{\theta_{ad}} \circ \Phi)$ coincides with the probability density function of $P_{Z^e|\Phi(X^e)}$, the above equivalence follows immediately. □

While our objective function (3.3) is similar to the ones in Arjovsky et al. [2020], Krueger et al. [2021] in that they are composed of an empirical risk and a domain invariance regularization, the correctness has not been fully discussed so far. In Section 3.4, we will mathematically prove the correctness of (3.3) under some settings.

3.3 Hyperparameter Selection Method

3.3.1 Difficulty in Hyperparameter Selection

The objective function (3.3) has a hyperparameter λ to select, as is often the case with DIL methods. The hyperparameter selection has difficulty as noted in Section 1.2 and 2.4. Gulrajani and Lopez-Paz [2023] reported that the success of DIL methods depended strongly on the careful choice of hyperparameters and that the existing two CV methods, training-domain validation (Tr-CV) and leave-one-domain-out validation (LOD-CV), failed to select preferable hyperparameters.

The failure of the CV methods is caused by the improper design of the objective function for CV; they do not simulate the o.o.d. risk, which is the maximum risk over the domains. Tr-CV splits data in each training domain into training and validation subsets, and takes the sum of the validated risks over the training domains. Obviously, this is not an estimate of the o.o.d. risk. LOD-CV holds out one domain among the training domains in turn and validates models with the average of the validated risks over the held-out domains. Again, this average does not correspond to the o.o.d. risk. In summary, the problem we need to solve is answering the following question: how can we construct an evaluation function of the o.o.d. risk from validation data? In the sequel, we will propose two methods of CV, which are summarized in Algorithm 1.

3.3.2 Method I: Using Coarser Label Data

We divide each of $\mathcal{D}^{e^*}, \mathcal{D}_{ad}^{e_1}, ..., \mathcal{D}_{ad}^{e_n}$ into K parts where $|\mathcal{E}_{ad}| = n$, and use the k-th sample $\{\mathcal{D}_{[k]}^{e^*}, \mathcal{D}_{ad,[k]}^{e_1}, ..., \mathcal{D}_{ad,[k]}^{e_n}\}$ and the rest $\{\mathcal{D}_{[-k]}^{e^*}, \mathcal{D}_{ad,[-k]}^{e_1}, ..., \mathcal{D}_{ad,[-k]}^{e_n}\}$ for validation

and training, respectively. To approximate the o.o.d. risk of the trained predictor $p_{\theta^\lambda_{[-k]}} \circ \Phi^\lambda_{[-k]}$, we wish to estimate $\mathcal{R}^e(p_{\theta^\lambda_{[-k]}} \circ \Phi^\lambda_{[-k]})$ for $e \in \mathcal{E}_{ad} \cup \{e^*\}$ by the validation set. For e^*, we use the standard empirical estimate $\hat{\mathcal{R}}^{e^*}_{[k]}(p_{\theta^\lambda_{[-k]}} \circ \Phi^\lambda_{[-k]})$. For $e \in \mathcal{E}_{ad}$, we substitute unavailable Y^e with Z^e and use $\hat{\mathcal{R}}^{(X^e,Z^e)}_{[k]}(p_{\theta^\lambda_{[-k]}} \circ \Phi^\lambda_{[-k]}) := \frac{1}{|\mathcal{D}^e_{ad,[k]}|} \sum_{(x^e,z^e)\in\mathcal{D}^e_{ad,[k]}} -\log p_{\theta^\lambda_{[-k]}}(z^e|\Phi^\lambda_{[-k]}(x^e))$.

3.3.3 Method II: Using Correction Term

Method I can be improved by correcting the replacement $\mathcal{R}^e = \mathcal{R}^{(X^e,Y^e)}$ with $\mathcal{R}^{(X^e,Z^e)}$ for $e \in \mathcal{E}_{ad}$. We use the following theorem for the correction:

Theorem 7. *Let $\mathcal{Z}^{\not\sim} := \{z \in \mathcal{Z} \,||g^{-1}(z)| > 1\}$. For any map $\Phi : \mathcal{X} \to \mathcal{H}$, $p_\theta : \mathcal{H} \to \mathcal{P}_\mathcal{Y}$, and random variable (X, Y) on $\mathcal{X} \times \mathcal{Y}$, the following equality holds:*

$$\mathcal{R}^{(X,Y)}(p_\theta \circ \Phi) = \mathcal{R}^{(X,g(Y))}(p_\theta \circ \Phi) + \sum_{z^{\not\sim}\in\mathcal{Z}^{\not\sim}} \Big\{ P(g(Y) = z^{\not\sim}) \times \mathcal{R}^{(X,Y)|z^{\not\sim}}(p_\theta \circ \Phi) \Big\}.$$

Here,

$$\mathcal{R}^{(X,Y)|z^{\not\sim}}(p_\theta \circ \Phi) := \int -\log p_\theta\big(Y|\Phi(X), g(Y) = z^{\not\sim}\big) dP_{(X,Y)|g(Y)=z^{\not\sim}}$$

where $P_{(X,Y)|g(Y)=z^{\not\sim}}$ denotes the conditional distribution of (X, Y) given the event $g(Y) = z^{\not\sim}$, and $p_\theta(y|\Phi(x), g(Y) = z^{\not\sim}) := \frac{p_\theta(y|\Phi(x))}{\sum_{y\in g^{-1}(z^{\not\sim})} p_\theta(y|\Phi(x))}$.

The proof is given in Section 4.3. The theorem shows that, to estimate the correction term, we need to estimate (i)$P(g(Y^e) = z^{\not\sim})$ and (ii) $\mathcal{R}^{(X^e,Y^e)|z^{\not\sim}}(p_{\theta^\lambda_{[-k]}} \circ \Phi^\lambda_{[-k]})$ for every $z^{\not\sim} \in \mathcal{Z}^{\not\sim}$.

(i) is naturally estimated even on $e \in \mathcal{E}_{ad}$: $\hat{P}(Z^e = z^{\not\sim}) := \frac{|\mathcal{D}^e_{ad,z^{\not\sim}}|}{|\mathcal{D}^e_{ad}|}$, where $\mathcal{D}^e_{ad,z^{\not\sim}} := \{(x, z) \in \mathcal{D}^e_{ad} \,|z = z^{\not\sim}\}$. (ii) is not easily estimable; while a direct simulation of the integration $\int dP_{(X^e,Y^e)|g(Y^e)=z^{\not\sim}}$ demands data from $(X^e, Y^e) \sim P_{X^e,Y^e}$, our available data $\mathcal{D}^e_{ad}$ on $e \in \mathcal{E}_{ad}$ is from $P_{X^e,g(Y^e)}$, not from P_{X^e,Y^e}. To solve the non-availability of data from P_{X^e,Y^e}, we use the training data $\mathcal{D}^{e^*} \sim P_{X^{e^*},Y^{e^*}}$ instead. Namely, (ii) is estimated by

$$\hat{\mathcal{R}}^{(X^{e^*},Y^{e^*})|z^{\not\sim}}_{[k]}(p_{\theta^\lambda_{[-k]}} \circ \Phi^\lambda_{[-k]}) := \frac{1}{|\mathcal{D}^{e^*}_{[k],z^{\not\sim}}|} \sum_{(x,y)\in\mathcal{D}^{e^*}_{[k],z^{\not\sim}}} -\log p_{\theta^\lambda_{[-k]}}(y|\Phi^\lambda_{[-k]}(x), g(Y) = z^{\not\sim}),$$

where $\mathcal{D}^{e^*}_{[k],z^{\not\sim}} := \Big\{(x, y) \in \mathcal{D}^{e^*}_{[k]} \,|g(y) = z^{\not\sim}\Big\} \subset \mathcal{D}^{e^*}_{[k]}$. In Algorithm 1, the above risk estimate is abbreviated by $\hat{\mathcal{R}}^{e^*|z^{\not\sim}}_{[k]}(\lambda)$ for notation simplicity.

Algorithm 1 CV methods. If CORRECTION = True, λ is selected by method II and if False, I.

Require: : Split $\mathcal{D}^{e^*}, \mathcal{D}^{e_1}_{ad}, ..., \mathcal{D}^{e_n}_{ad}$ into K parts.

Require: : Set the hyperparameter candidates Λ.

Require: :$\hat{P}^e(z^\curlywedge) \leftarrow \frac{|\mathcal{D}^e_{ad,z^\curlywedge}|}{|\mathcal{D}^e_{ad}|}$, where $\mathcal{D}^e_{ad,z^\curlywedge} := \{(x,z) \in \mathcal{D}^e_{ad} \,|z = z^\curlywedge\}$ for all $e \in \mathcal{E}_{ad}$ and $z^\curlywedge \in \mathcal{Z}^\curlywedge$.

for $\lambda \in \Lambda$ **do**
  for $k = 1$ to K **do**
    Learn $\theta^\lambda_{[-k]}, \Phi^\lambda_{[-k]}$ by using $\mathcal{D}^{e^*}_{[-k]}, \mathcal{D}^{e_1}_{ad,[-k]}, ..., \mathcal{D}^{e_n}_{ad,[-k]}$.
    $\hat{\mathcal{R}}^{e^*}_k(\lambda) \leftarrow \frac{1}{|\mathcal{D}^{e^*}_{[k]}|} \sum_{(x^{e^*}, y^{e^*}) \in \mathcal{D}^{e^*}_{[k]}} - \log p_{\theta^\lambda_{[-k]}}(y^{e^*} | \Phi^\lambda_{[-k]}(x^{e^*}))$
    //Risk estimation on e^*.
    $\hat{\mathcal{R}}^{e^*|z^\curlywedge}_k(\lambda) \leftarrow \frac{1}{|\mathcal{D}^{e^*}_{[k],z^\curlywedge}|} \sum_{(x,y) \in \mathcal{D}^{e^*}_{[k],z^\curlywedge}} - \log p_{\theta^\lambda_{[-k]}}(y | \Phi^\lambda_{[-k]}(x), g(Y) = z^\curlywedge)$ for $z^\curlywedge$ in $\mathcal{Z}^\curlywedge$.
    for $e \in \mathcal{E}_{ad}$ **do**
      $\hat{\mathcal{R}}^e_k(\lambda) \leftarrow \frac{1}{|\mathcal{D}^e_{ad,[k]}|} \sum_{(x^e, z^e) \in \mathcal{D}^e_{ad,[k]}} - \log p_{\theta^\lambda_{[-k]}}(z^e | \Phi^\lambda_{[-k]}(x^e))$.
      // Risk estimation on e.
      if CORRECTION **then**
        $\hat{\mathcal{R}}^e_k(\lambda) \; + \leftarrow \; \sum_{z^\curlywedge \in \mathcal{Z}^\curlywedge} \hat{P}^e(z^\curlywedge) \cdot \hat{\mathcal{R}}^{e^*|z^\curlywedge}_k(\lambda)$ // Correction term addition.
      end if
    end for
    $\hat{\mathcal{R}}^{o.o.d.}_k(\lambda) \leftarrow \max_{e \in \mathcal{E}_{ad} \cup \{e^*\}} \hat{\mathcal{R}}^e_k(\lambda)$ // o.o.d. risk estimation.
  end for
  $\hat{\mathcal{R}}^{o.o.d.}(\lambda) \leftarrow \frac{1}{K} \sum_{k=1}^K \hat{\mathcal{R}}^{o.o.d.}_k(\lambda)$ // Final o.o.d. risk estimation.
end for
Select $\lambda^* := \text{argmin}_{\lambda \in \Lambda} \hat{\mathcal{R}}^{o.o.d.}(\lambda)$

3.4 Theoretical Analysis

Throughout this section, to avoid discussing the non-trivial effects of nonlinear Φ, we focus on the simplified case of variable selections, as Theorems 4 and 5. Let $\mathcal{X} := \mathcal{X}_1 \times \mathcal{X}_2$ where $\mathcal{X}_1 := \mathbb{R}^{n_1}$ and $\mathcal{X}_2 := \mathbb{R}^{n_2}$ with $n_1, n_2 \in \mathbb{N}$. For a projection Φ, let Φ_i denote the $\mathcal{X}_i$-component of Φ $(i = 1, 2)$. If Φ has a $\mathcal{X}_2$-component, we write $\mathrm{Im}\Phi_2 \neq \emptyset$. Let (X_1^I, Y^I) be a fixed random variable on $\mathcal{X}_1 \times \mathcal{Y}$. For simplicity of analysis, the domain set $\mathcal{E}$ is defined by all the random variables (X, Y) with their distributions $P_{\Phi^{\mathcal{X}_1}(X),Y}$ are equal to $P_{X_1^I,Y^I}$, where $\Phi^{\mathcal{X}_1} : \mathcal{X} \to \mathcal{X}_1$ denotes the projection onto $\mathcal{X}_1$; namely,

$$\{(X^e, Y^e)\}_{e\in\mathcal{E}} := \left\{(X, Y) : \text{a random variable on } \mathcal{X} \times \mathcal{Y} \,\middle|\, P_{\Phi^{\mathcal{X}_1}(X),Y} = P_{X_1^I,Y^I}\right\}. \tag{※}$$

In this case, for any $e \in \mathcal{E}$ the variable (X^e, Y^e) satisfies (i) $P_{Y^e|\Phi^{\mathcal{X}_1}(X^e)}$ equals to $P_{Y^I|X_1^I}$, and (ii) the marginal distribution $P_{\Phi^{\mathcal{X}_1}(X)}$ of the invariant feature $\Phi^{\mathcal{X}_1}(X)$ equals to $P_{X_1^I}$. The above setting and definition persist through Section 3.4.

3.4.1 Theoretical Analysis of Objective Function

The following theorem ensures that, neglecting estimations and under some conditions, a minimum of our objective function (3.3) with careful hyperparameter choice also minimizes the o.o.d. risk (2.1):

Theorem 8 (o.o.d. optimality of our objective function, Setting I)**.** *Under the setting (※), additionally assume that the following condition holds:*

(A) For any variable selection Φ with $\mathrm{Im}\Phi_2 \neq \emptyset$, there exist two domains $\{e_1, e_2\} \subset \mathcal{E}_{ad}$ such that $P_{g(Y^{e_1})|\Phi(X^{e_1})} \neq P_{g(Y^{e_2})|\Phi(X^{e_2})}$.

Then, there exists $\lambda^* \in \mathbb{R}$ such that any minimizer $(\theta^\dagger, \theta_{ad}^\dagger, \Phi^\dagger)$ of (3.3),

$$(\theta^\dagger, \theta_{ad}^\dagger, \Phi^\dagger) \in \operatorname*{argmin}_{\theta,\theta_{ad},\Phi}\left\{\mathcal{R}^{e^*}(p_\theta \circ \Phi) + \lambda^* \cdot \sum_{e\in\mathcal{E}_{ad}} \|\nabla_{\hat{\theta}_{ad}=\theta_{ad}} \mathcal{R}^{(X^e,Z^e)}(p_{\hat{\theta}_{ad}}^{\mathcal{Z}|\mathcal{H}} \circ \Phi)\|^2\right\},$$

is o.o.d. optimal, i.e.,

$$p_{\theta^\dagger} \circ \Phi^\dagger \in \operatorname{argmin}_{p_\theta:\mathcal{X}\to\mathcal{P}_\mathcal{Y}} \mathcal{R}^{o.o.d.}(p_\theta),$$

where models p_θ and $p_{\theta_{ad}}^{\mathcal{Z}|\mathcal{H}}$ in $\min_{\theta,\theta_{ad},\Phi}$ run all the probability density functions, and Φ runs all the variable selections. The gradient $\nabla_{\theta_{ad}}$ should be understood as the functional derivative on the space of probability density functions.

For the proof, see Section 4.4.1. Condition (A) means that $\mathcal{E}_{ad}$ has sufficient variation to capture the desirable domain invariance $\Phi^{\mathcal{X}_1}$.

While Theorem 8 assumes $p_{\theta_{ad}}^{\mathcal{Z}|\mathcal{H}}$ runs all the probability density functions, our method is implemented with $p_{\theta_{ad}}^{\mathcal{Z}|\mathcal{H}}$ running all linear logistic functions (see, Subsection 3.2). To see the effectiveness under the linear logistic case, we deduce the following theorem:

Theorem 9 (o.o.d. optimality of our objective function, Setting II)**.** *Under the setting (※), additionally assume that the following condition holds:*

(A)' For any variable selection Φ with $\mathrm{Im}\Phi_2 \neq \emptyset$, there exist two domains $\{e_1, e_2\} \subset \mathcal{E}_{ad}$ such that $P_{g(Y^{e_1})|\Phi(X^{e_1})} \neq P_{g(Y^{e_2})|\Phi(X^{e_2})}$ and both $P_{g(Y^{e_1})|\Phi(X^{e_1})}$ and $P_{g(Y^{e_2})|\Phi(X^{e_2})}$ are in the linear logistic model.

(B) $P_{Y^I|\Phi^{\mathcal{X}_1}(X)}$ is in the linear logistic model.

Then, there exists $\lambda^* \in \mathbb{R}$ such that any minimizer $(\theta^\dagger, \theta_{ad}^\dagger, \Phi^\dagger)$ of (3.3),

$$(\theta^\dagger, \theta_{ad}^\dagger, \Phi^\dagger) \in \underset{\theta, \theta_{ad}, \Phi}{\mathrm{argmin}} \Big\{ \mathcal{R}^{e^*}(p_\theta \circ \Phi) + \lambda^* \cdot \sum_{e \in \mathcal{E}_{ad}} \|\nabla_{\hat{\theta}_{ad}=\theta_{ad}} \mathcal{R}^{(X^e, Z^e)}(p_{\hat{\theta}_{ad}}^{\mathcal{Z}|\mathcal{H}} \circ \Phi)\|^2 \Big\},$$

is o.o.d. optimal, i.e.,

$$p_{\theta^\dagger} \circ \Phi^\dagger \in \mathrm{argmin}_{p_\theta : \mathcal{X} \to \mathcal{P}_\mathcal{Y}} \mathcal{R}^{o.o.d.}(p_\theta),$$

where models p_θ runs all the probability density functions, $p_{\theta_{ad}}^{\mathcal{Z}|\mathcal{H}}$ runs all linear logistic functions, and Φ runs all the variable selections.

For the proof, see Section 4.4.2.

3.4.2 Theoretical Analysis of Cross Validation Methods

In Sections 3.3.2 and 3.3.3, we approximate $\mathcal{R}^{(X^e, Y^e)}$ using coarser labels Z^e. While the approximation is not exact, we will prove that the proposed CV methods still select a correct hyperparameter under some conditions. We will also elucidate the difference of the two CV methods. Given hyperparameter λ, minimizing (3.3) over the model yields the feature map (variable selection) denoted by $\Phi^\lambda : \mathcal{X} \to \mathbb{R}^{n_\lambda}$ ($n_\lambda \leq n_1 + n_2$). For simplicity of theoretical analysis, we assume that the minimization of (3.3) achieves perfectly the conditional probability density function of $P_{Y^{e^*}|\Phi^\lambda(X^{e^*})}$, denoted by $p^{*,\lambda}(y|\Phi^\lambda(x))$. Then, neglecting estimation errors, the approximated

o.o.d. risk of $p^{*,\lambda} \circ \Phi^\lambda$ used in Methods I and II are represented by the following $\mathcal{R}^I(\lambda)$ and $\mathcal{R}^{II}(\lambda)$, respectively:

$$\mathcal{R}^I(\lambda) := \max\left\{ \max_{e \in \mathcal{E}_{ad}} \mathcal{R}^{(X^e, g(Y^e))}(p^{*,\lambda} \circ \Phi^\lambda), \mathcal{R}^{(X^{e^*}, Y^{e^*})}(p^{*,\lambda} \circ \Phi^\lambda) \right\}, \tag{3.4}$$

$$\mathcal{R}^{II}(\lambda) := \max_{e \in \mathcal{E}_{ad} \cup \{e^*\}} \left\{ \mathcal{R}^{(X^e, g(Y^e))}(p^{*,\lambda} \circ \Phi^\lambda) + \sum_{z^{\mathcal{Y}} \in \mathcal{Z}^{\mathcal{Y}}} P(Z^e = z^{\mathcal{Y}}) \cdot \mathcal{R}^{(X^{e^*}, Y^{e^*})|z^{\mathcal{Y}}}(p^{*,\lambda} \circ \Phi^\lambda) \right\}. \tag{3.5}$$

We have the following theoretical justification of our CV methods: the chosen λ gives a minimizer of the correct CV criterion. For the proofs, see Sections 4.4.3 and 4.4.4.

Theorem 10 (Correctness of Method I)**.** *Under the setting of variable selection (※), assume further that the following conditions (i) and (ii) hold:*

(i) Among a set Λ of hyperparameter candidates, there exists $\lambda^I \in \Lambda$ such that $\Phi^{\lambda^I} = \Phi^{\mathcal{X}_1}$.

(ii) Let p^{e^*} be the probability density function of $P_{X^{e^*}, g(Y^{e^*})}$. Then, for any λ with $\mathrm{Im}\Phi_2^\lambda \neq \emptyset$, there is $e_\lambda \in \mathcal{E}_{ad}$ such that

$$(x, z) \sim P_{X^{e_\lambda}, g(Y^{e_\lambda})} \text{ satisfies } p^{e^*}(z|\Phi^\lambda(x)) \leq e^{-\beta} - \varepsilon \text{ with probability 1.}$$

Here, $\varepsilon \in \mathbb{R}_{>0}$ is some sufficient small positive real number (that is, $0 < \varepsilon \ll 1$) and $\beta := H(Y^{e^*}|\Phi^{\mathcal{X}_1}(X^{e^*}))$ is the conditional entropy of $(\Phi^{\mathcal{X}_1}(X^{e^*}), Y^{e^*})$.

Then, we have

$$\mathrm{argmin}_{\lambda \in \Lambda} \mathcal{R}^I(\lambda) \subset \mathrm{argmin}_{\lambda \in \Lambda} \mathcal{R}^{o.o.d.}(p^{*,\lambda} \circ \Phi^\lambda).$$

Theorem 11 (Correctness of Method II)**.** *Under the setting of variable selection (※), assume that, in addition to (i) in Theorem 10, the following condition (iii) holds:*

(ii)' for any λ with $\mathrm{Im}\Phi_2^\lambda \neq \emptyset$, there is $e_\lambda \in \mathcal{E}_{ad}$ such that

$(x,z) \sim P_{X^{e_\lambda}, g(Y^{e_\lambda})}$ *satisfies* $p^{e^*}(z|\Phi^\lambda(x)) \leq e^{-\beta_\lambda} - \varepsilon$ *holds with probability 1.*

Here, ε *is some sufficiently small positive real number and*

$$\beta_\lambda := H(Y^{e^*}|\Phi^{\mathcal{X}_1}(X^{e^*})) - \sum_{z^\curlywedge \in \mathcal{Z}^\curlywedge} \Big\{ P(g(Y^{e^*}) = z^\curlywedge) \times \mathcal{R}^{(X^{e^*}, Y^{e^*})|z^\curlywedge}(p^{*,\lambda} \circ \Phi^\lambda) \Big\}.$$

Then, under the setting (※), we have

$$\mathrm{argmin}_{\lambda \in \Lambda} \mathcal{R}^{II}(\lambda) \subset \mathrm{argmin}_{\lambda \in \Lambda} \mathcal{R}^{o.o.d.}(p^{*,\lambda} \circ \Phi^\lambda).$$

The conditions (ii) and (ii)' impose that, for at least one $e_\lambda \in \mathcal{E}_{ad}$, the two domains e_λ and e^* are *different* in the following meaning. If λ fails to remove domain-specific factors (*i.e.*, $\mathrm{Im}\Phi_2^\lambda \neq \emptyset$), for some $e_\lambda \in \mathcal{E}_{ad}$, $(x,z) \sim P_{X^{e_\lambda}, g(Y^{e_\lambda})}$ yields low $p^{e^*}(z|\Phi^\lambda(x))$ with high probability. On the other hand, $(x,z) \sim P_{X^{e^*}, g(Y^{e^*})}$ yields high $p^{e^*}(z|\Phi^\lambda(x))$ with high probability: that is, e^* and e_λ are *different*.

The theoretical analysis shows, while Method I is simpler to implement than Method II, Method II is more applicable. Noting that $\beta \geq \beta_\lambda$ and hence, $e^{-\beta} - \varepsilon \leq e^{-\beta_\lambda} - \varepsilon$, the condition (ii)' is milder than (ii). Recalling that (ii) and (ii)' impose the discrepancy between $\mathcal{E}_{ad}$ and e^* as discussed in the last paragraph, relaxation of conditions from (ii) to (ii)' implies that *method II can be applied even when domains* $\mathcal{E}_{ad}$ *and* e^* *have smaller discrepancy than the condition for Method I.* The difference of these two methods will be demonstrated in Section 6.

We discuss the feasibility of (ii) and (ii)', and show these conditions are not necessarily strong. First, we discuss the Condition (ii). Since $\beta = H(Y^e|\Phi^{\mathcal{X}_1}(X^e))$ is the conditional entropy, we have

$$0 \leq \beta \leq \log|\mathcal{Y}|$$

and hence

$$\frac{1}{|\mathcal{Y}|} - \varepsilon \leq e^{-\beta} - \varepsilon \leq 1 - \varepsilon$$

holds. We can see that Condition (ii) is weak if $e^{-\beta} - \varepsilon$ approaches 1, or if β is small. Recall that $\Phi^{\mathcal{X}_1}(X^e)$ is the bias-removed feature of X^e (digit of CMNIST, or object of ImageNet, for example). We can then expect that, in many real-world settings, $\beta = H(Y^e|\Phi^{\mathcal{X}_1}(X^e))$ is often small, since the bias-removed feature $\Phi^{\mathcal{X}_1}(X^e)$ should

have a large amount of information on the labels. Condition (ii) is satisfied if the likelihood $p^{e^*}(z|\Phi^\lambda(x))$ evaluated at a random point $(x, z) \sim P_{X^e, g(Y^e)}$ is bounded by the large value $e^{-\beta} - \varepsilon$ for at least one $e \in \mathcal{E}_{ad}$, so that the inequality in (ii) is likely to hold. Noting that (ii)' is weaker than (ii), the feasibility of (ii)' is concluded from one of (ii).

3.4.3 Sufficient Conditions of Theorem 10 and 11

In the section, we reveal sufficient conditions of e^* for there to exist $(X^{e_\lambda}, Y^{e_\lambda})$ that satisfies (ii) and (ii)' in Theorems 10 and 11, respectively.

Theorem 12. *Assume that (X^{e^*}, Y^{e^*}) satisfies the following condition:*

(A2) For a sufficiently small $\varepsilon \ll 1$, any λ with $\mathrm{Im}\Phi_2^\lambda \neq \emptyset$, any $a \in \mathrm{Im}\Phi_1^\lambda$, and any $b \in \mathcal{Y}$, there exists $c(\lambda, a, b)$[1] such that

$$P(Y^{e^*} = b | \Phi_1^\lambda(X^{e^*}) = a, \Phi_2^\lambda(X^{e^*}) = c) \geq (1 - e^{-\beta}) + \varepsilon.$$

Then, for any λ with $\mathrm{Im}\Phi_2^\lambda \neq \emptyset$, there exists $(X^{e_\lambda}, Y^{e_\lambda}) \in \{(X^e, Y^e)\}_{e \in \mathcal{E}}$ such that the inequality in Theorem 10 (ii) holds.

Theorem 13. *(X^{e^*}, Y^{e^*}) satisfies the following condition:*

(A2)' For a sufficiently small $\varepsilon \ll 1$, the following statement holds:
$\forall \lambda$ with $\mathrm{Im}\Phi_2^\lambda \neq \emptyset$, $\forall \alpha \in \mathrm{Im}\Phi_1^\lambda$, $\forall b \in \mathcal{Y}$, $\exists c(\lambda, a, b)$ s.t.

$$P(Y^{e^*} = b | \Phi_1^\lambda(X^{e^*}) = a, \Phi_2^\lambda(X^{e^*}) = c) \geq (1 - e^{-\beta_\lambda}) + \varepsilon.$$

Then, $\forall \lambda$ with $\mathrm{Im}\Phi_2^\lambda \neq \emptyset$, there exists $(X^{e_\lambda}, Y^{e_\lambda}) \in \{(X^e, Y^e)\}_{e \in \mathcal{E}}$ such that the inequality in (ii)' holds.

For the proofs of Theorems 12 and 13, see Sections 4.4.5 and 4.4.6. The conditions (A2) and (A2)' means that, in the domain $e = e^*$, the affection of domain-specific factors ($= \mathcal{X}_2$) to the response variable Y^{e^*} is large; indeed, the inequality in (A2) and (A2)' means that, if λ fails to remove domain-specific factors (*i.e.*, $\mathrm{Im}\Phi_2^\lambda \neq \emptyset$), we can control the probability of $Y^{e^*} = b$ by selecting c for any $b \in \mathcal{Y}$. Note also that the inequality (A2) and (A2)' is a lower bound of the likelihood, while the condition

[1] $c(\lambda, a, b)$ means $c \in \mathcal{X}_2$ is determined by given $\lambda \in \Lambda$, $a \in \mathcal{X}_1$, $b \in \mathcal{Y}$.

in (ii) and (ii)', Theorem 10 and 11, is an upper bound of the likelihood. Although imposing an upper bound might look reasonable to reflect non-fitting of the projection Φ^{λ}, Theorem 12 shows that we can use a lower bound as a sufficient condition.

Chapter 4

Proofs

4.1 Proof of Theorem 4

Throughout this section, for given X^e and $x \in \mathcal{X}$ $(= \mathcal{X}_1 \times \mathcal{X}_2)$, $\mathcal{X}_1$- and $\mathcal{X}_2$-components of X^e and x are often abbreviated X_1^e and X_2^e, or x_1 and x_2 respectively. For a variable selection Φ, let Φ_i denote the $\mathcal{X}_i$-component of Φ $(i = 1, 2)$. If Φ has or has not an $\mathcal{X}_i$-component, we write $\mathrm{Im}\Phi_i \neq \emptyset$ or $\mathrm{Im}\Phi_i = \emptyset$ respectively. For Φ_i $(i = 1, 2)$, $\Phi_i^{\perp}$ denotes the projection onto orthogonal complements of $\mathrm{Im}\Phi_i$ with respect to $\mathcal{X}_i$; namely, $\mathrm{Im}\Phi_i \bigotimes \mathrm{Im}\Phi_i^{\perp} \simeq \mathcal{X}_i$.

For the proof of Theorem 4, we prepare two lemmas.

Lemma 14. *$\mathcal{I}_{tr}^{v.s} = \{\Phi^{\mathcal{X}_1}\}$ holds. Here, recall that $\Phi^{\mathcal{X}_1}$ is the projecton onto $\mathcal{X}_1$.*

Lemma 15.

$$(w^*, \Phi^*) \in \mathrm{argmin}_{\Phi \in \mathcal{I}_{tr}^{v.s.}, w:\mathcal{H} \to \mathcal{Y}} \sum_{e \in \mathcal{E}_{tr}} \mathcal{R}^e(w \circ \Phi)$$

coincides with $(w^{\Phi^{\mathcal{X}_1}}, \Phi^{\mathcal{X}_1})$, where $w^{\Phi^{\mathcal{X}_1}}$ is a conditional expectation $\mathbb{E}[Y^e|\Phi^{\mathcal{X}_1}(X^e)]$ of Y^e given $\Phi^{\mathcal{X}_1}(X^e)$ on P_{X^e,Y^e}; that is,

$$w^{\Phi^{\mathcal{X}_1}}(\Phi^{\mathcal{X}_1}(x)) := \mathbb{E}[Y^e|\Phi^{\mathcal{X}_1}(X^e) = \Phi^{\mathcal{X}_1}(x)].$$

Since $P_{Y^e|\Phi^{\mathcal{X}_1}(X^e)} = P_{Y^e|X_1^e}$ does not depend on a choice of $((X^e, Y^e)) \in \{X^e, Y^e\}_{e \in \mathcal{E}}$, $\mathbb{E}[Y^e|\Phi^{\mathcal{X}_1}(X^e)] = \mathbb{E}[Y^e|X_1^e]$ also does not depend on a choice of $(X^e, Y^e) \in \{(X^e, Y^e)\}_{e \in \mathcal{E}}$.

We prove Theorem 4 based on the above lemmas, before proving them.

Proof of Theorem 4 By Lemma 15, we may prove that

$$w^{\Phi^{\mathcal{X}_1}} \circ \Phi^{\mathcal{X}_1} \in \operatorname*{argmin}_{f:\mathcal{X}\to\mathcal{Y}} \mathcal{R}^{o.o.d}(f).$$

To prove it, it suffices to prove the following statement:

$$\text{For any } f : \mathcal{X} \to \mathcal{Y} \text{ and } (X^{e_1}, Y^{e_1}) \in \{(X^e, Y^e)\}_{e\in\mathcal{E}}, \text{ there exists } (X^{e_2}, Y^{e_2}) \in \{(X^e, Y^e)\}_{e\in\mathcal{E}} \text{ such that}$$

$$\int \|w^{\Phi_{\mathcal{X}_1}} \circ \Phi^{\mathcal{X}_1}(x) - y\|^2 dP_{X^{e_1},Y^{e_1}}(x,y) \leq \int \|f(x) - y\|^2 dP_{X^{e_2},Y^{e_2}}(x,y). \tag{4.1}$$

Take arbitrary $f : \mathcal{X} \to \mathcal{Y}$ and $(X^{e_1}, Y^{e_1}) \in \{(X^e, Y^e)\}_{e\in\mathcal{E}}$. Define $(X^{e_2}, Y^{e_2}) \in \{(X^e, Y^e)\}_{e\in\mathcal{E}}$ such that its distribution is the direct product $P_{X_1^{e_1},Y^{e_1}} \otimes P_{X_2}$, where $P_{X_1^{e_1},Y^{e_1}}$ is the marginal distribution of $P_{X_1^{e_1},Y^{e_1}}$ on $\mathcal{X}_1 \times \mathcal{Y}$ and P_{X_2} is an arbitrary distribution on $\mathcal{X}_2$.

Then, the right-hand side of the inequality (4.1) is given by

$$\begin{aligned}\int \|f(x) - y\|^2 dP_{X^{e_2},Y^{e_2}}(x,y) &= \int \|f(x) - y\|^2 d(P_{X_1^{e_1},Y^{e_1}} \otimes P_{X_2})(x,y) \\ &= \int P_{X_2}(x_2) \int \|f(x_1,x_2) - y\|^2 dP_{X_1^{e_1},Y^{e_1}}(x_1,y).\end{aligned}$$

We can see that, for any $x_2^* \in \mathcal{X}_2$, the inequality

$$\begin{aligned}\int \|f(x_1,x_2^*) - y\|^2 dP_{X_1^{e_1},Y^{e_1}}(x_1,y) &\geq \int \|\mathbb{E}[Y|\Phi^{\mathcal{X}_1}(X^e) = x_1] - y\|^2 dP_{X_1^{e_1},Y^{e_1}}(x_1,y) \\ &= \int \|w^{\Phi_{\mathcal{X}}} \circ \Phi^{\mathcal{X}_1}(x_1) - y\|^2 dP_{X_1^{e_1},Y^{e_1}}(x_1,y)\end{aligned}$$

holds, since the minimum of a risk on the least square loss is attained at the conditional

expectation $\mathbb{E}[Y^e|\Phi^{\mathcal{X}_1}(X^e)]$. Hence, we obtain

$$\begin{aligned}
\int \|f(x)-y\|^2 dP_{X^{e_2},Y^{e_2}}(x,y) &= \int P_{X_2}(x_2)\int \|f(x_1,x_2)-y\|^2 dP_{X_1^{e_1},Y^{e_1}}(x_1,y)\\
&\geq \int P_{X_2}(x_2)\int \|w^{\Phi_{\mathcal{X}}}\circ\Phi^{\mathcal{X}_1}(x)-y\|^2 dP_{X_1^{e_1},Y^{e_1}}(x_1,y)\\
&= \int \|w^{\Phi_{\mathcal{X}}}\circ\Phi^{\mathcal{X}_1}(x)-y\|^2 dP_{X_1^{e_1},Y^{e_1}}(x_1,y)\\
&= \int P_{X_2^{e_1}|X_1^{e_1},Y^{e_1}}(x_2)\int \|w^{\Phi_{\mathcal{X}}}\circ\Phi^{\mathcal{X}_1}(x)-y\|^2 dP_{X_1^{e_1},Y^{e_1}}(x_1,y)\\
&= \int \|w^{\Phi_{\mathcal{X}}}\circ\Phi^{\mathcal{X}_1}(x)-y\|^2 d(P_{X_1^{e_1},Y^{e_1}}\otimes P_{X_2^{e_1}|X_1^{e_1},Y^{e_1}})(x,y)\\
&= \int \|w^{\Phi_{\mathcal{X}}}\circ\Phi^{\mathcal{X}_1}(x)-y\|^2 dP_{X^{e_1},Y^{e_1}}(x,y),
\end{aligned}$$

which concludes the proof. □

Proof of Lemma 14 Since $\mathcal{I}_{tr}^{v.s}=\mathcal{I}^{v.s}$, we may prove that $\mathcal{I}^{v.s}=\{\Phi^{\mathcal{X}_1}\}$. We prove $\mathcal{I}^{v.s}=\{\Phi^{\mathcal{X}_1}\}$ by following three steps:

Step 1

Take any variable selection Φ with $\mathrm{Im}\Phi_1\neq\emptyset$ and $\mathrm{Im}\Phi_2\neq\emptyset$. Then $\Phi\notin\mathcal{I}^{v.s.}$.

Step 2

Take any variable selection Φ with $\mathrm{Im}\Phi_1=\emptyset$ and $\mathrm{Im}\Phi_2\neq\emptyset$. Then $\Phi\notin\mathcal{I}^{v.s.}$.

Step 3

Take any variable selection Φ with $\mathrm{Im}\Phi\subsetneq\mathcal{X}_1$. Then $\Phi\notin\mathcal{I}^{v.s.}$.

Poof of Step 1 It suffices to prove the following statement.

For any variable selection Φ with $\mathrm{Im}\Phi_1\neq\emptyset$ and $\mathrm{Im}\Phi_2\neq\emptyset$, there exist two distributions (X^{e_1},Y^{e_1}) and $(X^{e_2},Y^{e_2})\in\{(X^e,Y^e)\}_{e\in\mathcal{E}}$ which satisfy $P_{Y^{e_1}|\Phi(X^{e_1})}\neq P_{Y^{e_2}|\Phi(X^{e_2})}$.

Take any variable selection Φ with $\mathrm{Im}\Phi_1\neq\emptyset$ and $\mathrm{Im}\Phi_2\neq\emptyset$. Fix $x^*\in\mathcal{X}$, y^*, $y^{**}\in\mathcal{Y}$ with $y^*\neq y^{**}$ and $p^I(y^*|x_1^*)>0$. Here, recall that p^I denotes the p.d.f. of $P_{Y^I|X_1^I}$. Define two maps $g^i:\mathrm{Im}\Phi_1\times\mathcal{Y}\to\mathrm{Im}\Phi_2$ $(i=1,2)$ by

$$g^1(\Phi_1(x), y) = \begin{cases} \Phi_2(x^*) & (\Phi_1(x), y) = \big(\Phi_1(x^*), y^*\big) \\ \Phi_2(x^*) - 1 & (\text{ else }) \end{cases}$$

$$g^2(\Phi_1(x), y) = \begin{cases} \Phi_2(x^*) & (\Phi_1(x), y) = (\Phi_1(x^*), y^{**}) \\ \Phi_2(x^*) - 1 & (\text{ else }) \end{cases}$$

Take two distributions $(X^{e_1}, Y^{e_1}), (X^{e_2}, Y^{e_2}) \in \{(X^e, Y^e)\}_{e \in \mathcal{E}}$ such that their distributions $P_{X^{e_1}, Y^{e_1}}$ and $P_{X^{e_2}, Y^{e_2}}$ coincide with

$$\begin{aligned} P_{X^{e_1}, Y^{e_1}} &= P_{\mathrm{Im}\Phi_2^\perp} \otimes P^{e_1}_{\mathrm{Im}\Phi_2|\mathrm{Im}\Phi_1, \mathcal{Y}} \otimes P_{Y^I|X_1^I} \otimes P_{\mathrm{Im}\Phi_1} \otimes P_{\mathrm{Im}\Phi_1^\perp} \\ P_{X^{e_2}, Y^{e_2}} &= P_{\mathrm{Im}\Phi_2^\perp} \otimes P^{e_2}_{\mathrm{Im}\Phi_2|\mathrm{Im}\Phi_1, \mathcal{Y}} \otimes P_{Y^I|X_1^I} \otimes P_{\mathrm{Im}\Phi_1} \otimes P_{\mathrm{Im}\Phi_1^\perp}. \end{aligned}$$

Here,

- $P_{\mathrm{Im}\Phi_2^\perp}$ denotes an arbitrary distribution on $\mathrm{Im}\Phi_2^\perp$,
- $P_{\mathrm{Im}\Phi_1}$ denotes an arbitrary distributions on $\mathrm{Im}\Phi_1$ where its p.d.f. $p_{\mathrm{Im}\Phi_1}(\Phi_1(x))$ satisfies $p_{\mathrm{Im}\Phi_1}(\Phi_1(x^*)) \neq 0$,
- $P_{\mathrm{Im}\Phi_1^\perp}$ is a distribution on $\mathrm{Im}\Phi_1^\perp$ with its p.d.f. coincides with a delta function $\delta_{\Phi_1^\perp(x^*)}(\Phi_1^\perp(x))$ on $\Phi_1^\perp(x^*)$,
- $P^{e_i}_{\mathrm{Im}\Phi_2|\mathrm{Im}\Phi_1, \mathcal{Y}}$ $(i = 1, 2)$ denotes a conditional distribution on $\mathrm{Im}\Phi_2$ given $\mathrm{Im}\Phi_1 \times \mathcal{Y}$, with its p.d.f. $p^{e_i}_{\mathrm{Im}\Phi_2|\mathrm{Im}\Phi_1, \mathcal{Y}}$ coincides with a delta function $\delta_{g^i(\Phi_1(x), y)}(\Phi_2(x))$ on $g^i(\Phi_1(x), y)$.

Let the p.d.f of $P_{\mathrm{Im}\Phi_2^\perp}$ be $p_{\mathrm{Im}\Phi_2^\perp}$. Then, the p.d.f. of the conditional distribution $P_{Y^{e_1}|\Phi(X^{e_1})=\Phi(x^*)}$ is represented as

$$\frac{\begin{aligned} \int p_{\mathrm{Im}\Phi_2^\perp}(\Phi_2^\perp(x)) \times \delta_{g^1(\Phi_1(x^*), y)}(\Phi_2(x^*)) \times p^I(y|\Phi_1(x^*), \Phi_1^\perp(x)) \\ \times \delta_{\Phi_1^\perp(x^*)}(\Phi_1^\perp(x)) \times p_{\mathrm{Im}\Phi_1}(\Phi_1(x^*)) d\Phi_2^\perp(x) d\Phi_1^\perp(x) \end{aligned}}{\begin{aligned} \int p_{\mathrm{Im}\Phi_2^\perp}(\Phi_2^\perp(x)) \times \delta_{g^1(\Phi_1(x^*), y)}(\Phi_2(x^*)) \times p^I(y|\Phi_1(x^*), \Phi_1^\perp(x)) \\ \times \delta_{\Phi_1^\perp(x^*)}(\Phi_1^\perp(x)) \times p_{\mathrm{Im}\Phi_1}(\Phi_1(x^*)) d\Phi_2^\perp(x) d\Phi_1^\perp(x) dy \end{aligned}} \tag{4.2}$$

Note that, for fixed x^*, $\delta_{g^1(\Phi_1(x^*),y)}(\Phi_2(x^*))$ coincides with $\delta_{y^*}(y)$; indeed, by the definitions of g^1 ,

$$\delta_{g^1(\Phi_1(x^*),y)}(\Phi_2(x^*)) = \begin{cases} \infty & (\text{ if } \Phi_2(x^*) = g^1(\Phi_1(x^*), y)) \\ 0 & (\text{ else }) \end{cases}$$

and noting that $\Phi_2(x^*) = g^1(\Phi_1(x^*), y)$ hold i.f.f. $y = y^*$, we can see that $\delta_{g^1(\Phi_1(x^*),y)}(\Phi_2(x^*)) = \delta_{y^*}(y)$ holds as functions of $y \in \mathcal{Y}$.

Hence, noting the fact, the numerator of (4.2) is rewritten as

$$\begin{aligned} &\int p_{\mathrm{Im}\Phi_2^\perp}(\Phi_2^\perp(x)) \times \delta_{g^1(\Phi_1(x^*),y)}(\Phi_2(x^*)) \times p^I(y|\Phi_1(x^*), \Phi_1^\perp(x)) \\ &\qquad \times \delta_{\Phi_1^\perp(x^*)}(\Phi_1^\perp(x)) \times p_{\mathrm{Im}\Phi_1}(\Phi_1(x^*)) d\Phi_2^\perp(x) d\Phi_1^\perp(x) \\ &= \int \delta_{g^1(\Phi_1(x^*),y)}(\Phi_2(x^*)) \times p^I(y|\Phi_1(x^*), \Phi_1^\perp(x)) \\ &\qquad \times \delta_{\Phi_1^\perp(x^*)}(\Phi_1^\perp(x)) \times p_{\mathrm{Im}\Phi_1}(\Phi_1(x^*)) d\Phi_1^\perp(x) \\ &= \delta_{g^1(\Phi_1(x^*),y)}(\Phi_2(x^*)) \times p^I(y|\Phi_1(x^*), \Phi_1^\perp(x^*)) \times p_{\mathrm{Im}\Phi_1}(\Phi_1(x^*)) \\ &= \delta_{y^*}(y) \times p^I(y|x_1^*) \times p_{\mathrm{Im}\Phi_1}(\Phi_1(x^*)). \end{aligned}$$

On the other hand, the denominator of (4.2) is rewritten as

$$\begin{aligned} &\int p_{\mathrm{Im}\Phi_2^\perp}(\Phi_2^\perp(x)) \times \delta_{g^1(\Phi_1(x^*),y)}(\Phi_2(x^*)) \times p^I(y|\Phi_1(x^*), \Phi_1^\perp(x)) \\ &\qquad \times \delta_{\Phi_1^\perp(x^*)}(\Phi_1^\perp(x)) \times p_{\mathrm{Im}\Phi_1}(\Phi_1(x^*)) d\Phi_2^\perp(x) d\Phi_1^\perp(x) dy \\ &= \int \delta_{g^1(\Phi_1(x^*),y)}(\Phi_2(x^*)) \times p^I(y|\Phi_1(x^*), \Phi_1^\perp(x)) \\ &\qquad \times \delta_{\Phi_1^\perp(x^*)}(\Phi_1^\perp(x)) \times p_{\mathrm{Im}\Phi_1}(\Phi_1(x^*)) d\Phi_1^\perp(x) dy \\ &= \int \delta_{g^1(\Phi_1(x^*),y)}(\Phi_2(x^*)) \times p^I(y|\Phi_1(x^*), \Phi_1^\perp(x^*)) \times p_{\mathrm{Im}\Phi_1}(\Phi_1(x^*)) dy \\ &= \int \delta_{y^*}(y) \times p^I(y|x_1^*) \times p_{\mathrm{Im}\Phi_1}(\Phi_1(x^*)) dy \end{aligned}$$

$$= p^I(y^*|x_1^*) \times p_{\mathrm{Im}\Phi_1}(\Phi_1(x^*)).$$

Combining the two transformations, (4.3) is represented by

$$(4.3) = \frac{\delta_{y^*}(y) \times p^I(y|x_1^*) \times p_{\mathrm{Im}\Phi_1}(\Phi_1(x^*))}{p^I(y^*|x_1^*) \times p_{\mathrm{Im}\Phi_1}(\Phi_1(x^*))} = \frac{\delta_{y^*}(y) \times p^I(y|x_1^*)}{p^I(y^*|x_1^*)}$$

Noting that

$$\frac{\delta_{y^*}(y) \times p^I(y|x_1^*)}{p^I(y^*|x_1^*)} = \begin{cases} \infty & (y = y^*) \\ \frac{0 \times p^I(y|x_1^*)}{p^I(y^*|x_1^*)} = 0 & (\text{ else }) \end{cases},$$

we can see that

$$(4.3) = \delta_{y^*}(y).$$

By the same procedure, we can also show that the p.d.f. of $P_{Y^{e_2}|\Phi(X^{e_2})}$ is $\delta_{y^{**}}(y)$. Recalling that

$$y^{**} \neq y^*,$$

we can see that

$$P_{Y^{e_1}|\Phi(X^{e_1})=\Phi(x^*)} \neq P_{Y^{e_2}|\Phi(X^{e_2})=\Phi(x^*)},$$

which concludes the proof of Step 1.

Poof of Step 2 It suffices to prove the following statement.

> For any variable selection Φ with $\mathrm{Im}\Phi_1 = \emptyset$ and $\mathrm{Im}\Phi_2 \neq \emptyset$, there exist two distribution (X^{e_1}, Y^{e_1}) and $(X^{e_2}, Y^{e_2}) \in \{(X^e, Y^e)\}_{e \in \mathcal{E}}$ which satisfy $P_{Y^{e_1}|\Phi(X^{e_1})} \neq P_{Y^{e_2}|\Phi(X^{e_2})}$.

Take any variable selection Φ with $\mathrm{Im}\Phi_1 = \emptyset$ and $\mathrm{Im}\Phi_2 \neq \emptyset$. Fix $x^* \in \mathcal{X}$, y^*, $y^{**} \in \mathcal{Y}$ which satisfy $y^* \neq y^{**}$ and $p^I(y^*|x_1^*) > 0$. Define two maps $g^i : \mathcal{Y} \to \mathrm{Im}\Phi_2$ $(i = 1, 2)$ by

$$g^1(y) = \begin{cases} \Phi_2(x^*) & (y = y^*) \\ \Phi_2(x^*) - 1 & (\text{ else }) \end{cases}$$

$$g^2(y) = \begin{cases} \Phi_2(x^*) & (y = y^{**}) \\ \Phi_2(x^*) - 1 & (\text{ else }) \end{cases}$$

Take two distributions $(X^{e_1}, Y^{e_1}), (X^{e_2}, Y^{e_2}) \in \{(X^e, Y^e)\}_{e\in\mathcal{E}}$ such that their distributions $P_{X^{e_1},Y^{e_1}}$ and $P_{X^{e_2},Y^{e_2}}$ coincide with

$$P_{X^{e_1},Y^{e_1}} = P_{\mathrm{Im}\Phi_2^\perp} \otimes P^{e_1}_{\mathrm{Im}\Phi_2|\mathcal{Y}} \otimes P_{Y^I|X_1^I} \otimes P_{\mathcal{X}_1}$$
$$P_{X^{e_2},Y^{e_2}} = P_{\mathrm{Im}\Phi_2^\perp} \otimes P^{e_2}_{\mathrm{Im}\Phi_2|\mathcal{Y}} \otimes P_{Y^I|X_1^I} \otimes P_{\mathcal{X}_1}.$$

Here,

- $P_{\mathrm{Im}\Phi_2^\perp}$ denotes an arbitrary distribution on $\mathrm{Im}\Phi_2^\perp$,
- $P_{\mathcal{X}_1}$ is a distribution on $\mathcal{X}_1$ with its p.d.f. coincides with a delta function $\delta_{x^*}(x)$ on x^*,
- $P^{e_i}_{\mathrm{Im}\Phi_2|\mathcal{Y}}$ $(i = 1, 2)$ denotes conditional distributions on $\mathrm{Im}\Phi_2$ given $\mathcal{Y}$, with their p.d.f. $p^{e_i}_{\mathrm{Im}\Phi_2|\mathcal{Y}}$ coincides with delta function $\delta_{g^i(y)}(\Phi_2(x))$ on $g^i(y)$.

Let the p.d.f. of $P_{\mathrm{Im}\Phi_2^\perp}$ be $p_{\mathrm{Im}\Phi_2^\perp}$. Then, the p.d.f. of the conditional distribution $P_{Y^{e_1}|\Phi(X^{e_1})=\Phi(x^*)}$ is represented as

$$\frac{\int p_{\mathrm{Im}\Phi_2^\perp}(\Phi_2^\perp(x))\times\delta_{g^1(y)}(\Phi_2(x^*)) \times p^I(y|x_1) \times p^I(y|x_1) \times \delta_{x_1^*}(x_1) d\Phi_2^\perp(x) dx_1}{\int p_{\mathrm{Im}\Phi_2^\perp}(\Phi_2^\perp(x))\times\delta_{g^1(y)}(\Phi_2(x^*)) \times p^I(y|x_1) \times p^I(y|x_1) \times \delta_{x_1^*}(x_1) d\Phi_2^\perp(x) dx_1 dy} \tag{4.3}$$

Noting that, for fixed x^*, $\delta_{g^1(y)}(\Phi(x^*))$ coincides with $\delta_{y^*}(y)$, the numerator of (4.3) is rewritten as

$$\int p_{\mathrm{Im}\Phi_2^\perp}(\Phi_2^\perp(x)) \times \delta_{g^1(y)}(\Phi_2(x^*)) \times p^I(y|x_1) \times p^I(y|x_1) \times \delta_{x_1^*}(x_1) d\Phi_2^\perp(x) dx_1$$

$$= \int \delta_{g^1(y)}(\Phi_2(x^*)) \times p^I(y|x_1) \times \delta_{x_1^*}(x_1) dx_1$$

$$= \delta_{y^*}(y) \times p^I(y|x_1^*)$$

On the other hand, the denominator of (4.3) is rewritten as

$$\int p_{\mathrm{Im}\Phi_2^{\perp}}(\Phi_2^{\perp}(x)) \times \delta_{g^1(y)}(\Phi_2(x^*)) \times p^I(y|x_1) \\ \times p^I(y|x_1) \times \delta_{x_1^*}(x_1) d\Phi_2^{\perp}(x) dx_1 dy$$

$$= \int \delta_{g^1(y)}(\Phi_2(x^*)) \times p^I(y|x_1) \times \delta_{x_1^*}(x_1) dx_1 dy$$

$$= \int \delta_{y^*}(y) \times p^I(y|x^*) dy = p^I(y^*|x_1^*)$$

Combining the two transformations, (4.3) is represented by

$$(4.3) = \frac{\delta_{y^*}(y) \times p^I(y|x_1^*)}{p^I(y^*|x_1^*)} = \delta_{y^*}(y).$$

By the same procedure, we can also show that the p.d.f. of $P_{Y^{e_2}|\Phi(X^{e_2})}$ is $\delta_{y^{**}}(y)$. Recalling that

$$y^{**} \neq y^*,$$

we can see that

$$P_{Y^{e_1}|\Phi(X^{e_1})=\Phi(x^*)} \neq P_{Y^{e_2}|\Phi(X^{e_2})=\Phi(x^*)},$$

which concludes the proof of Step 2.

Poof of Step 3 It suffices to prove the following statement.

For any variable selection Φ with $\mathrm{Im}\Phi \subsetneq \mathcal{X}_1$, there exist two distribution (X^{e_1}, Y^{e_1}) and $(X^{e_2}, Y^{e_2}) \in \{(X^e, Y^e)\}_{e\in\mathcal{E}}$ which satisfy $P_{Y^{e_1}|\Phi(X^{e_1})} \neq P_{Y^{e_2}|\Phi(X^{e_2})}$.

Take any variable selection Φ with $\mathrm{Im}\Phi \subsetneq \mathcal{X}_1$. Take $\bar{x}, \bar{\bar{x}}$ and $\bar{\bar{\bar{x}}} \in \mathcal{X}$ which satisfy

$$P_{Y^I|X_1^I=(\Phi_1(\bar{x}),\Phi_1^{\perp}(\bar{\bar{x}}))} \neq P_{Y^I|X_1^I=(\Phi_1(\bar{x}),\Phi_1^{\perp}(\bar{\bar{\bar{x}}}))}$$

Here, there exist such $\bar{x}, \bar{\bar{x}}$ and $\bar{\bar{\bar{x}}}$ by the assumption of Theorem 4. Take two distributions $(X^{e_1}, Y^{e_1}), (X^{e_2}, Y^{e_2}) \in \{(X^e, Y^e)\}_{e\in\mathcal{E}}$ such that their distributions

$P_{X^{e_1},Y^{e_1}}$ and $P_{X^{e_2},Y^{e_2}}$ coincide with

$$P_{X^{e_1},Y^{e_1}} = P_{\mathcal{X}_2} \otimes P_{Y^I|X_1^I} \otimes P_{\mathrm{Im}\Phi_1} \otimes P^{e_1}_{\mathrm{Im}\Phi_1^\perp}$$
$$P_{X^{e_2},Y^{e_2}} = P_{\mathcal{X}_2} \otimes P_{Y^I|X_1^I} \otimes P_{\mathrm{Im}\Phi_1} \otimes P^{e_2}_{\mathrm{Im}\Phi_1^\perp},$$

where

- $P_{\mathcal{X}_2}$ denotes an arbitrary distribution on $\mathcal{X}_2$,
- $P_{\mathrm{Im}\Phi_1}$ denotes an arbitrary distribution on $\mathrm{Im}\Phi_1$ where its p.d.f. $p_{\mathrm{Im}\Phi_1}(\Phi_1(x))$ satisfies $p_{\mathrm{Im}\Phi_1}(\Phi_1(\bar{x})) \neq 0$,
- $P^{e_1}_{\mathrm{Im}\Phi_1^\perp}$ and $P^{e_2}_{\mathrm{Im}\Phi_1^\perp}$ are distributions on $\mathrm{Im}\Phi_1^\perp$ with their p.d.f.s coincide with delta functions $\delta_{\Phi_1^\perp(\bar{\bar{x}})}(\Phi_1^\perp(x))$ on $\Phi_1^\perp(\bar{\bar{x}})$ and a delta function $\delta_{\Phi_1^\perp(\bar{\bar{\bar{x}}})}(\Phi_1^\perp(x))$ on $\Phi_1^\perp(\bar{\bar{\bar{x}}})$ respectively.

Here, the two distributions are included in $\mathcal{I}^{v.s.}_{tr}$ since the equality $\mathcal{I}^{v.s.}_{tr} = \mathcal{I}^{v.s.}$ holds by the assumption. Let $p_{\mathcal{X}_2}(x_2)$ be the p.d.f. of $P_{\mathcal{X}_2}(x_2)$. Then the p.d.f. of conditional probability $P_{Y^{e_1}|\Phi(X^{e_1})=\Phi(\bar{x})}$ is represented as

$$\frac{\begin{array}{l}\int p_{\mathcal{X}_2}(x_2) \times p^I(y|\Phi_1(x^*), \Phi_1^\perp(x)) \times \\ \qquad\qquad \delta_{\Phi_1^\perp(x^*)}(\Phi_1^\perp(x)) \times p_{\mathrm{Im}\Phi_1}(\Phi_1(\bar{x}))dx_2 d\Phi_1^\perp(x)\end{array}}{\begin{array}{l}\int p_{\mathcal{X}_2}(x_2)) \times p^I(y|\Phi_1(x^*), \Phi_1^\perp(x)) \times \\ \qquad\qquad \delta_{\Phi_1^\perp(x^*)}(\Phi_1^\perp(x)) \times p_{\mathrm{Im}\Phi_1}(\Phi_1(\bar{x}))dx_2 d\Phi_1^\perp(x)dy\end{array}}$$

$$= \frac{\int p^I(y|\Phi_1(\bar{x}), \Phi_1^\perp(x)) \times \delta_{\Phi_1^\perp(\bar{\bar{x}})}(\Phi_1^\perp(x)) \times p_{\mathrm{Im}\Phi_1}(\Phi_1(\bar{x}))d\Phi_1^\perp(x)}{\int p^I(y|\Phi_1(\bar{x}), \Phi_1^\perp(x)) \times \delta_{\Phi_1^\perp(\bar{\bar{x}})}(\Phi_1^\perp(x)) \times p_{\mathrm{Im}\Phi_1}(\Phi_1(\bar{x}))d\Phi_1^\perp(x)dy}$$

$$= \frac{p^I(y|\Phi_1(\bar{x}), \Phi_1^\perp(\bar{\bar{x}})) \times p_{\mathrm{Im}\Phi_1}(\Phi_1(\bar{x}))}{\int p^I(y|\Phi_1(\bar{x}), \Phi_1^\perp(\bar{\bar{x}})) \times p_{\mathrm{Im}\Phi_1}(\Phi_1(\bar{x}))dy}$$

$$= \frac{p^I(y|\Phi_1(\bar{x}), \Phi_1^\perp(\bar{\bar{x}})) \times p_{\mathrm{Im}\Phi_1}(\Phi_1(\bar{x}))}{p_{\mathrm{Im}\Phi_1}(\Phi_1(\bar{x}))}$$

$$= p^I(y|\Phi_1(\bar{x}), \Phi_1^{\perp}(\bar{\bar{x}}))$$

Conducting the same procedure, we can see that the p.d.f. of the conditional probability $P_{Y^{e_2}|\Phi(X^{e_2})=\Phi(\bar{x})}$ is represented as

$$p^I(y|\Phi_1(\bar{x}), \Phi_1^{\perp}(\bar{\bar{\bar{x}}})).$$

Recalling that

$$P_{Y^I|X_1^I=(\Phi_1(\bar{x}),\Phi_1^{\perp}(\bar{\bar{x}}))} \neq P_{Y^I|X_1^I=(\Phi_1(\bar{x}),\Phi_1^{\perp}(\bar{\bar{\bar{x}}}))},$$

we can see that,

$$P_{Y^{e_1}|\Phi(X^{e_1})=\Phi(\bar{x})} \neq P_{Y^{e_2}|\Phi(X^{e_2})=\Phi(\bar{x})},$$

which concludes the proof. □

Proof of Lemma 15 Take any

$$(w, \Phi^{\mathcal{X}_1}) \in \mathrm{argmin}_{\Phi \in \mathcal{I}_{tr}^{v.s.}, w:\mathcal{H}\to\mathcal{Y}} \sum_{e\in\mathcal{E}_{tr}} \mathcal{R}^e(w \circ \Phi).$$

Note that

$$\mathcal{R}^e(w \circ \Phi^{\mathcal{X}_1}) \geq \mathcal{R}^e(w^{\Phi_{\mathcal{X}_1}} \circ \Phi^{\mathcal{X}_1})$$

for any $e \in \mathcal{E}_{tr}$ and the lower bound is attained i.f.f. $w = w^{\Phi_{\mathcal{X}_1}}$. Summarizing the inequality with respect to $e \in \mathcal{E}_{tr}$, we obtain

$$\sum_{w\in\mathcal{E}_{tr}} \mathcal{R}^e(w \circ \Phi^{\mathcal{X}_1}) \geq \sum_{w\in\mathcal{E}_{tr}} \mathcal{R}^e(w^{\Phi_{\mathcal{X}_1}} \circ \Phi^{\mathcal{X}_1}),$$

and the lower bound is attained i.f.f. $w = w^{\Phi_{\mathcal{X}_1}}$. It concludes the proof. □

4.2 Proof of Theorem 5

For the proof of Theorem 5, we prepare two lemmas.

Lemma 16. $\mathcal{I}_{ad}^{v.s} = \{\Phi^{\mathcal{X}_1}\}$ *holds.*

Lemma 17.

$$(w^*, \Phi^*) \in \mathrm{argmin}_{\Phi\in\mathcal{I}_{ad}^{v.s.}, w:\mathcal{H}\to\mathcal{Y}} \mathcal{R}^{e^*}(w \circ \Phi)$$

coincides with $(w^{\Phi^{\mathcal{X}_1}}, \Phi^{\mathcal{X}_1})$*, where* $w^{\Phi^{\mathcal{X}_1}}$ *is a conditional expectation* $\mathbb{E}[Y^e|\Phi^{\mathcal{X}_1}(X^e)]$ *of* Y^e *given* $\Phi^{\mathcal{X}_1}(X^e)$ *on* P_{X^e,Y^e}*; that is,*

$$w^{\Phi^{\mathcal{X}_1}}(\Phi^{\mathcal{X}_1}(x)) := \mathbb{E}[Y^e|\Phi^{\mathcal{X}_1}(X^e) = \Phi^{\mathcal{X}_1}(x)].$$

Proof of Theorem 5 We omit the proof since it is essentially same as the one of Theorem 4. □

Proof of Lemma 16 We omit the proof since it is essentially same as the one of Lemma 14. □

Proof of Lemma 17 Noting that

$$\mathcal{R}^e(w \circ \Phi^{\mathcal{X}_1}) \geq \mathcal{R}^e(w^{\Phi_{\mathcal{X}_1}} \circ \Phi^{\mathcal{X}_1})$$

for any $e \in \mathcal{E}$ and $w : \mathcal{X}_1 \to \mathcal{Y}$ and the lower bound is attained i.f.f. $w = w^{\Phi_{\mathcal{X}_1}}$, the conclusion follows immediately. □

4.3 Proof of Theorem 7

$$\begin{aligned}
\mathcal{R}^{(X,Y)}(p_\theta \circ \Phi) - \mathcal{R}^{(X,g(Y))}(p_\theta \circ \Phi) &= \int -\log p_\theta(Y|\Phi(X)) dP_{Y,\Phi(X)} \\
&\qquad + \int \log p_\theta(g(Y)|\Phi(X)) dP_{g(Y),\Phi(X)} \\
&= -\int \log \frac{p_\theta(Y|\Phi(X))}{p_\theta(g(Y)|\Phi(X))} dP_{(Y,\Phi(X))} \\
&= -\int dP_{g(Y)} \int \log \frac{p_\theta(Y|\Phi(X))}{p_\theta(g(Y)|\Phi(X))} dP_{(Y,\Phi(X))|g(Y)}
\end{aligned} \tag{4.4}$$

By the definition of $p_\theta(y|\Phi(x), g(Y) = z)$ in Theorem 7,

$$\frac{p_\theta(y|\Phi(x))}{p_\theta(g(y)|\Phi(x))} = p_\theta(y|\Phi(x), g(Y) = z)$$

holds, where $z = g(y)$. Therefore, we obtain

$$
\begin{aligned}
(6) &= -\int dP_{g(Y)} \int \log \frac{p_\theta(Y|\Phi(X))}{p_\theta(g(Y)|\Phi(X))} dP_{(Y,\Phi(X))|g(Y)} \\
&= -\int dP_{g(Y)} \int \log p_\theta(Y|\Phi(X), g(Y) = z) dP_{(Y,\Phi(X))|g(Y)=z} \\
&= -\sum_{z\in\mathcal{Z}} P(g(Y) = z) \int \log p_\theta(Y|\Phi(X), g(Y) = z) dP_{(Y,\Phi(X))|g(Y)=z} \\
&= -\sum_{z^{\not\hookrightarrow}\in\mathcal{Z}\not\hookrightarrow} P(g(Y) = z^{\not\hookrightarrow}) \int \log p_\theta(Y|\Phi(X), g(Y) = z^{\not\hookrightarrow}) dP_{(Y,\Phi(X))|g(Y)=z^{\not\hookrightarrow}} \\
&\qquad + \sum_{z^{\hookrightarrow}\in\mathcal{Z}-\mathcal{Z}\not\hookrightarrow} P(g(Y) = z^{\hookrightarrow}) \int \log p_\theta(Y|\Phi(X), g(Y) = z^{\hookrightarrow}) dP_{(Y,\Phi(X))|g(Y)=z^{\hookrightarrow}}.
\end{aligned}
\tag{4.5}
$$

Noting that, for any $z^{\hookrightarrow} \in \mathcal{Z} - \mathcal{Z}\not\hookrightarrow$ and $y := g^{-1}(z^{\hookrightarrow})$[1], $p_\theta(y|\Phi(x), g(Y) = z^{\hookrightarrow}) = 1$ holds, we can see that

$$\log p_\theta(y|\Phi(x), g(Y) = z^{\hookrightarrow}) = 0.$$

The second term in the last line thus equals to zero, which concludes the proof. □

[1] $z^{\hookrightarrow} \in \mathcal{Z} - \mathcal{Z}\not\hookrightarrow$ implies that $|g^{-1}(z^{\hookrightarrow})| = 1$ and therefore, $g^{-1}(z^{\hookrightarrow})$ is determined uniquely. Note that there is no chance that $|g^{-1}(z^{\hookrightarrow})| = 0$ by the surjectivity of g.

4.4 Proofs of Theorems in Section 3.4

We rephrase the problem simplification (※) in Section 3.4 with some notation arrangements. Let $\mathcal{X} := \mathcal{X}_1 \times \mathcal{X}_2$ where $\mathcal{X}_1 := \mathbb{R}^{n_1}$ and $\mathcal{X}_2 := \mathbb{R}^{n_2}$ with $n_1, n_2 \in \mathbb{N}$. Let (X_1^I, Y^I) be a fixed random variable on $\mathcal{X}_1 \times \mathcal{Y}$. Throughout our theoretical analysis, the domain set $\mathcal{E}$ is defined by all the probability distributions with the fixed marginal distribution $P_{X_1^I, Y^I}$ of (X_1, Y); namely, all domains $T_{all} := \{(X^e, Y^e)\}_{e \in \mathcal{E}}$ are defined by

$$T_{all} := \left\{ (X, Y) : \text{a random variable on } \mathcal{X} \times \mathcal{Y} \,\middle|\, P_{\Phi^{\mathcal{X}_1}(X), Y} = P_{X_1^I, Y^I} \right\}, \qquad (※)$$

where $\Phi^{\mathcal{X}_1} : \mathcal{X} \to \mathcal{X}_1$ is the projection onto $\mathcal{X}_1$. The above setting and definition persist through Section 4.4.

For a projection Φ, let Φ_i denote the $\mathcal{X}_i$-component of Φ $(i = 1, 2)$. If Φ has a $\mathcal{X}_i$-component, we write $\mathrm{Im}\Phi_i \neq \emptyset$ $(i = 1, 2)$.

We prepare some additional notations to state Theorem 8 and its proof more clearly and briefly. Recall that the single training domain e^* for the target task and the domains $\mathcal{E}_{ad}$ for the additional task play important roles in our problem setting (see Section 3.1). Throughout the section, the domains are abbreviated as follows. The single training domain $(X^{e^*}, Y^{e^*}) \in T_{all}$ for the target task is abbreviated by (X^*, Y^*). For the domains $\mathcal{E}_{ad}$ of the additional task with coarser labels, $\{(X^e, Y^e)\}_{e \in \mathcal{E}_{ad}}$ is abbreviated by a subclass $T_{ad} \subset T_{all}$. For a projection $\Phi : \mathcal{X} \to \mathbb{R}^{n_\Phi}$ with its range n_Φ variables, let $p^{*,\Phi} : \mathbb{R}^{n_\Phi} \to \mathcal{P}_\mathcal{Y}$ denote the conditional probability density functions (p.d.f.) of $P(Y^*|\Phi(X^*))$. With a slight abuse of notation, for any probability P_θ on $\mathcal{X} \times \mathcal{Y}$ and a projection Φ, the density function of the conditional distribution $P_\theta(Y|\Phi(X))$ is denoted by $p_\theta \circ \Phi$.

We add some additional explanations and interpretations about the definition (※). From the condition of T_{all}, for the projection $\Phi^{\mathcal{X}_1}$, the conditional probability $P_{Y|\Phi^{\mathcal{X}_1}(X)}$ for *any* random variable $(X, Y) \in T_{all}$ is the same; namely, letting $p^I : \mathcal{X}_1 \to \mathcal{P}_\mathcal{Y}$ denote the conditional p.d.f. of the invariant predictor $P_{Y^I|X_1^I}$, we have

$$p^e \circ \Phi^{\mathcal{X}_1} = p^I \qquad (4.6)$$

for any $(X^e, Y^e) \in T_{all}$, where p^e is the conditional p.d.f. of $P_{Y^e|\Phi^{\mathcal{X}_1}(X^e)}$.

4.4.1 Proof of Theorem 8

We restate Theorem 8 as follows.

Theorem 18 (Theorem 8 in the main body, with some notation arrangements)**.** *Assume that all domains $T_{all} := \{(X^e, Y^e)\}_{e\in\mathcal{E}}$ are fixed as (※); namely,*

$$T_{all} := \left\{ (X,Y) : \textit{a random variable on } \mathcal{X}\times\mathcal{Y} \,\middle|\, P_{\Phi^{\mathcal{X}_1}(X),Y} = P_{X_1^I, Y^I} \right\}. \tag{4.7}$$

Additionally, assume that the following condition holds:

(A) For any projection Φ with $\mathrm{Im}\Phi_2 \neq \emptyset$, there exist $(X^{e_1}, Y^{e_1}), (X^{e_2}, Y^{e_2}) \in T_{ad}$ such that $P_{g(Y^{e_1})|\Phi(X^{e_1})} \neq P_{g(Y^{e_2})|\Phi(X^{e_2})}$.

Then, there exists $\lambda^* \in \mathbb{R}$ such that a minimizer $(\theta^\dagger, \theta^\dagger_{ad}, \Phi^\dagger)$ of the objective function

$$\min_{\theta,\theta_{ad},\Phi} \left\{ \mathcal{R}^{(X^*,Y^*)}(p_\theta \circ \Phi) + \lambda^* \cdot \sum_{(X^e,Y^e)\in T_{ad}} \|\nabla_{\hat{\theta}_{ad}=\theta_{ad}} \mathcal{R}^{(X^e, g(Y^e))}(p^{\mathcal{Z}|\mathcal{H}}_{\hat{\theta}_{ad}} \circ \Phi)\|^2 \right\} \tag{4.8}$$

is o.o.d. optimial, i.e.,

$$p_{\theta^\dagger} \circ \Phi^\dagger \in \operatorname*{argmin}_{p_\theta : \mathcal{X}\to\mathcal{P}_\mathcal{Y}} \mathcal{R}^{o.o.d.}(p_\theta),$$

where p_θ and $p^{\mathcal{Z}|\mathcal{H}}_{\theta_{ad}}$ in $\min_{\theta,\theta_{ad},\Phi}$ run all the p.d.f.s, and Φ runs all the variable selections. The gradient $\nabla_{\theta_{ad}}$ should be understood as the functional derivative on the space of p.d.f.

Before proving Theorem 18, we prepare one lemma, which asserts that, if $\mathrm{Im}\Phi_2 \neq \emptyset$, at least one domain in T_{ad} has non-trivial gradient:

Lemma 19.

$$\min_{\theta_{ad},\Phi:\mathrm{Im}\Phi_2\neq\emptyset} \sum_{(X^e,Y^e)\in T_{ad}} \|\nabla_{\hat{\theta}_{ad}=\theta_{ad}} \mathcal{R}^{(X^e, g(Y^e))}(p^{\mathcal{Z}|\mathcal{H}}_{\hat{\theta}_{ad}} \circ \Phi)\|^2 \;>\; 0.$$

Proof. It suffices to prove that, for any projection Φ with $\mathrm{Im}\Phi_2 \neq \emptyset$ and $p^{\mathcal{Z}|\mathcal{H}}_{\theta_{ad}}$, there is $(X^e, Y^e) \in T_{ad}$ such that $\|\nabla_{\hat{\theta}_{ad}=\theta_{ad}} \mathcal{R}^{(X^e, g(Y^e))}(p^{\mathcal{Z}|\mathcal{H}}_{\hat{\theta}_{ad}} \circ \Phi)\|^2 \neq 0$. We prove this by contradiction. Suppose that there exist a projection Φ with $\mathrm{Im}\Phi_2 \neq \emptyset$ and $p^{\mathcal{Z}|\mathcal{H}}_{\theta_{ad}}$ which satisfy

$$\|\nabla_{\hat{\theta}_{ad}=\theta_{ad}} \mathcal{R}^{(X^e, g(Y^e))}(p^{\mathcal{Z}|\mathcal{H}}_{\hat{\theta}_{ad}} \circ \Phi)\|^2 = 0 \qquad (\forall (X^e, Y^e) \in T_{ad}).$$

From Assumption (A), take (X^{e_1}, Y^{e_1}) and (X^{e_2}, Y^{e_2}) in T_{ad} such that $P(g(Y^{e_1})|\Phi(X^{e_1})) \neq P(g(Y^{e_2})|\Phi(X^{e_2}))$.

Note that the risk is defined by the cross-entropy loss:

$$\mathcal{R}^{(X^e,g(Y^e))}(p^{\mathcal{Z}|\mathcal{H}}_{\hat{\theta}_{ad}} \circ \Phi) = -\int \log p^{\mathcal{Z}|\mathcal{H}}_{\hat{\theta}_{ad}}(g(Y^e)|\Phi(X^e))dP_{X^e,Y^e}.$$

It is well known that this is minimized in the space of probability distributions if and only if $p^{\mathcal{Z}|\mathcal{H}}_{\hat{\theta}_{ad}}$ equals to $P(Y^e|\Phi(X^e))$. From $\|\nabla_{\hat{\theta}_{ad}=\theta_{ad}}\mathcal{R}^{(X^e,g(Y^e))}(p^{\mathcal{Z}|\mathcal{H}}_{\hat{\theta}_{ad}} \circ \Phi)\|^2 = 0$ for (X^{e_1}, Y^{e_1}) and (X^{e_2}, Y^{e_2}), we can conclude that $p^{\mathcal{Z}|\mathcal{H}}_{\theta_{ad}}$ should equal the p.d.f. both of $P_{g(Y^{e_1})|\Phi(X^{e_1})}$ and $P_{g(Y^{e_2})|\Phi(X^{e_2})}$. This contradicts with the assumption $P_{g(Y^{e_1})|\Phi(X^{e_1})} \neq P_{g(Y^{e_2})|\Phi(X^{e_2})}$. □

Proof of Theorem 18

Let Φ^{id} denote the identity map of $\mathcal{X}$. Define the constants C_1, C_2, and C_3 by

$$\begin{aligned}
C_1 &:= \mathcal{R}^{(X^*,Y^*)}(p^{*,\Phi^{id}} \circ \Phi^{id}) = H(Y^*|X^*),\\
C_2 &:= \mathcal{R}^{(X^*,Y^*)}(p^{*,\Phi^I} \circ \Phi^I) = H(Y^*|X^*_1) = H(Y^I|X^I_1),\\
C_3 &:= \frac{C_2 - C_1}{\min_{\theta_{ad},\Phi:\mathrm{Im}\Phi_2\neq\emptyset} \sum_{(X^e,Y^e)\in T_{ad}} \|\nabla_{\hat{\theta}_{ad}=\theta_{ad}}\mathcal{R}^{(X^e,g(Y^e))}(p^{\mathcal{Z}|\mathcal{H}}_{\hat{\theta}_{ad}} \circ \Phi)\|^2},
\end{aligned}$$

where $H(Y^I|X^I_1)$ and $H(Y^*|X^*)$ denote the conditional entropy. Note that C_3 is well-defined because of the positivity result of Lemma 19.

Take λ^* such that $\lambda^* > C_3$. For notational simplicity, the objective function (4.13) is denoted by $O(\theta, \theta_{ad}, \Phi)$; namely,

$$O(\theta, \theta_{ad}, \Phi) := \mathcal{R}^{(X^*,Y^*)}(p_\theta \circ \Phi) + \lambda^* \cdot \sum_{(X^e,Y^e)\in T_{ad}} \|\nabla_{\hat{\theta}_{ad}=\theta_{ad}}\mathcal{R}^{(X^e,g(Y^e))}(p^{\mathcal{Z}|\mathcal{H}}_{\hat{\theta}_{ad}} \circ \Phi)\|^2.$$

We prove the theorem in three steps.

Step 1

$$\min_{p:\mathcal{X}\to\mathcal{P}_\mathcal{Y}} \mathcal{R}^{o.o.d.}(p) = H(Y^I|X^I_1)$$

.

proof of Step 1

We will prove $p^I \in \operatorname{argmin}_{p:\mathcal{X}\to\mathcal{P}_\mathcal{Y}} \mathcal{R}^{o.o.d.}(p)$. From the definition

$$\mathcal{R}^{o.o.d.}(p) = \max_{(X^e,Y^e)\in T_{all}} -\int \log p(Y^e|X^e)dP_{Y^e,X^e},$$

$p^I \in \operatorname*{argmin}_{p:\mathcal{X}\to\mathcal{P}_\mathcal{Y}} \mathcal{R}^{o.o.d.}(p)$ holds if and only if

$$\max_{(X^e,Y^e)\in T_{all}} -\int \log p_\theta(Y^e|X^e) dP_{Y^e,X^e} \geq \max_{(X^e,Y^e)\in T_{all}} -\int \log p^I(Y^e|X_1^e) dP_{Y^e,X^e}$$

for any $p_\theta : \mathcal{X} \to \mathcal{P}_\mathcal{Y}$. Note that, as discussed before Theorem 18, for any $(X^e, Y^e) \in T_{all}$, we have $P_{Y^e|X_1^e} = P_{Y^I|X_1^I}$. Then, it suffices to prove that for any p_θ there exists $(X^{e'}, Y^{e'}) \in T_{all}$ such that

$$\int -\log p_\theta(Y^{e'}|X^{e'}) dP_{Y^{e'},X^{e'}} \geq \int -\log p^I(Y^I|X_1^I) dP_{X^e,Y^e}. \tag{4.9}$$

Define $(X^{e'}, Y^{e'}) \in T_{all}$ such that its distribution is the direct product $P_{X_1^I,Y^I} \otimes P_{X_2^{e'}}$, where $P_{X_2^{e'}}$ is an arbitrary distribution on $\mathcal{X}_2$. In this case, the left hand side of (4.9) is given by

$$\begin{aligned}\int -\log p_\theta(Y^{e'}|X^{e'}) dP_{Y^{e'},X^{e'}} &= \int -\log p_\theta(Y^{e'}|X_1^{e'}, X_2^{e'}) dP_{Y^{e'},X^{e'}} \\ &= \int dP_{X_2^{e'}} \int -\log p_\theta(Y^I|X_1^I, X_2^{e'}) dP_{X_1^I,Y^I}. \end{aligned} \tag{4.10}$$

We can see that, for any $x_2 \in \mathcal{X}_2$, the inequality

$$\int -\log p_\theta(Y^I|X_1^I, X_2^{e'} = x_2)) dP_{X_1^I,Y^I} \geq \int -\log p^I(Y^I|X_1^I) dP_{X_1^I,Y^I}$$

holds, since the minimum of the cross entropy loss is attained at the conditional p.d.f. p^I. Integrating this inequality with $P_{X_2^{e'}}$, we have

$$\int dP_{X_2^e} \int -\log p_\theta(Y^I|X_1^I, X_2^e) dP_{X_1^I,Y^I} \geq \int -\log p^I(Y^I|X_1^I) dP_{X_1^I,Y^I}. \tag{4.11}$$

Eqs. (4.10) and (4.11) show (4.9), from which the assertion is obtained by

$$-\int \log p^I(Y^I|X_1^I) dP_{X_1^I,Y^I} = H(Y^I|X_1^I)$$

.

Step 2 Any minimizer of the objective function,

$$(\theta^\dagger, \theta^\dagger_{ad}, \Phi^\dagger) \in \operatorname*{argmin}_{\theta,\theta_{ad},\Phi} O(\theta, \theta_{ad}, \Phi),$$

satisfies $\mathrm{Im}\Phi^\dagger_2 = \emptyset$.

proof of Step 2
It suffices to prove that $\min_{\Phi:\mathrm{Im}\Phi_2=\emptyset,\theta,\theta_{ad}} O(\theta, \theta_{ad}, \Phi) < \min_{\Phi:\mathrm{Im}\Phi_2\neq\emptyset,\theta,\theta_{ad}} O(\theta, \theta_{ad}, \Phi)$. First, we have

$$\begin{aligned}
&\min_{\Phi:\mathrm{Im}\Phi_2\neq\emptyset,\theta,\theta_{ad}} O(\theta, \theta_{ad}, \Phi) \\
&= \min_{\Phi:\mathrm{Im}\Phi_2\neq\emptyset,\theta,\theta^{ad}} \left\{ \mathcal{R}^{(X^*,Y^*)}(p_\theta \circ \Phi) + \lambda^* \cdot \sum_{(X^e,Y^e)\in T_{ad}} \|\nabla_{\hat\theta_{ad}=\theta_{ad}} \mathcal{R}^{(X^e,g(Z^e))}(p^{\mathcal{Z}|\mathcal{H}}_{\hat\theta_{ad}} \circ \Phi)\|^2 \right\}. \\
&> \min_{\Phi:\mathrm{Im}\Phi_2\neq\emptyset,\theta,\theta_{ad}} \Bigg\{ \mathcal{R}^{(X^*,Y^*)}(p_\theta \circ \Phi) \\
&\qquad + \frac{C_2 - C_1}{\min_{\theta_{ad},\Phi:\mathrm{Im}\Phi_2\neq\emptyset} \sum_{(X^e,Y^e)\in T_{ad}} \|\nabla_{\hat\theta_{ad}=\theta_{ad}} \mathcal{R}^{(X^e,g(Y^e))}(p^{\mathcal{Z}|\mathcal{H}}_{\hat\theta_{ad}} \circ \Phi)\|^2} \\
&\qquad\qquad \times \sum_{(X^e,Y^e)\in T_{ad}} \|\nabla_{\hat\theta_{ad}=\theta_{ad}} \mathcal{R}^{(X^e,g(Y^e))}(p^{\mathcal{Z}|\mathcal{H}}_{\hat\theta_{ad}} \circ \Phi)\|^2 \Bigg\}
\end{aligned}$$

$$\begin{aligned}
&\geq \min_{\Phi:\mathrm{Im}\Phi_2\neq\emptyset,\theta,\theta_{ad}} \{\mathcal{R}^{(X^*,Y^*)}(p_\theta \circ \Phi) + C_2 - C_1\} \\
&= \min_{\Phi:\mathrm{Im}\Phi_2\neq\emptyset,\theta,\theta_{ad}} \{\mathcal{R}^{(X^*,Y^*)}(p_\theta \circ \Phi)\} + C_2 - C_1 \\
&\geq \mathcal{R}^{(X^*,Y^*)}(p^{*,\Phi^{id}} \circ \Phi^{id}) + C_2 - C_1 = C_2.
\end{aligned}$$

On the other hand, by taking $\Phi = \Phi^I$, we obtain

$$\begin{aligned}
&\min_{\Phi:\mathrm{Im}\Phi_2=\emptyset,\theta,\theta_{ad}} O(\theta, , \theta_{ad}, \Phi) \\
&\qquad \leq \mathcal{R}^{(X^*,Y^*)}(p^I) + \lambda^* \cdot \sum_{(X^e,Y^e)\in T_{ad}} \|\nabla_{\hat\theta_{ad}=\theta^*_{ad}} \mathcal{R}^{(X^e,g(Y^e))}(p^{\mathcal{Z}|\mathcal{H}}_{\hat\theta_{ad}} \circ \Phi^I)\|^2.
\end{aligned}$$

Since $p^{\mathcal{Z}|\mathcal{H}}_{\hat\theta_{ad}} \circ \Phi^I = p^I(g(Y^I)|X^I_1)$ does not depend on θ_{ad}, the gradient is zero, and

therefore

$$\min_{\Phi:\mathrm{Im}\Phi_2=\emptyset,\theta,\theta_{ad}} O(\theta,,\theta_{ad},\Phi) \leq \mathcal{R}^{(X^*,Y^*)}(p^I) = C_2.$$

We thus obtain

$$\min_{\Phi:\mathrm{Im}\Phi_2=\emptyset,\theta,\theta_{ad}} O(\theta,\theta_{ad},\Phi) \leq C_2 < \min_{\Phi:\mathrm{Im}\Phi_2\neq\emptyset,\theta,\theta_{ad}} O(\theta,\theta_{ad},\Phi),$$

which completes the proof.

Step 3 If $(p_{\theta^\dagger}, p^{\mathcal{Z}|\mathcal{H}}_{\theta^\dagger_{ad}}, \Phi^\dagger) \in \operatorname*{argmin}_{\theta,\theta_{ad},\Phi} O(\theta,\theta^{ad},\Phi)$, then

$$\mathcal{R}^{o.o.d.}(p_{\theta^\dagger}\circ\Phi^\dagger) = H(Y^I|X_1^I)$$

.

proof of Step 3
From Step 1, we have $H(Y^I|X_1^I) \leq \mathcal{R}^{o.o.d.}(p_{\theta^\dagger}\circ\Phi^\dagger)$. We will probe the converse inequality.

From Step 2, we have $\mathrm{Im}\Phi_2^\dagger = \emptyset$. This tells $\mathcal{R}^{o.o.d.}(p_{\theta^\dagger}\circ\Phi^\dagger) = \mathcal{R}^{e^*}(p_{\theta^\dagger}\circ\Phi^\dagger)$, since $P_{X_1,Y}$ are the same for all elements in T_{all}. Therefore,

$$\begin{aligned}
&\mathcal{R}^{o.o.d.}(p_{\theta^\dagger}\circ\Phi^\dagger) = \mathcal{R}^{(X^*,Y^*)}(p_{\theta^\dagger}\circ\Phi^\dagger)\\
&\leq \min_{\theta_{ad}} \left\{ \mathcal{R}^{(X^*,Y^*)}(p_{\theta^\dagger}\circ\Phi^\dagger) + \lambda^* \cdot \sum_{(X^e,Y^e)\in T_{ad}} \|\nabla_{\hat{\theta}_{ad}=\theta^*_{ad}} \mathcal{R}^{(X^e,g(Y^e))}(p^{\mathcal{Z}|\mathcal{H}}_{\hat{\theta}_{ad}}\circ\Phi)\|^2 \right\}\\
&= \min_{\Phi,\theta,\theta_{ad}} \left\{ \mathcal{R}^{(X^*,Y^*)}(p_{\theta}\circ\Phi) + \lambda^* \cdot \sum_{(X^e,Y^e)\in T_{ad}} \|\nabla_{\hat{\theta}_{ad}=\theta_{ad}} \mathcal{R}^{(X^e,g(Y^e))}(p^{\mathcal{Z}|\mathcal{H}}_{\hat{\theta}_{ad}}\circ\Phi)\|^2 \right\}\\
&\leq C_2 = H(Y^I|X_1^I).
\end{aligned}$$

Final step for the proof of Theorem 18
For $(\theta^\dagger,\theta^\dagger_{ad},\Phi^\dagger) \in \operatorname*{argmin}_{\theta,\theta_{ad},\Phi} O(\theta,\theta_{ad},\Phi)$, Step 1 and Step 3 show

$$\mathcal{R}^{o.o.d.}(p_{\theta^\dagger}\circ\Phi^\dagger) = H(Y^I|X_1^I) = \min_{p:\theta\to\mathcal{P}_\mathcal{Y}} \mathcal{R}^{o.o.d.}(p),$$

which completes the proof.

4.4.2 Proof of Theorem 9

Theorem 20 (Theorem 9 in the main body, with some notation arrangements)**.** *Assume that all domains $T_{all} := \{(X^e, Y^e)\}_{e\in\mathcal{E}}$ are fixed as (※); namely,*

$$T_{all} := \left\{ (X,Y) : \textit{a random variable on } \mathcal{X}\times\mathcal{Y} \,\middle|\, P_{\Phi^{\mathcal{X}_1}(X),Y} = P_{X_1^I,Y^I} \right\}. \tag{4.12}$$

Additionally, assume that the following condition holds:

(A) For any projection Φ with $\mathrm{Im}\Phi_2 \neq \emptyset$, there exist $(X^{e_1}, Y^{e_1}), (X^{e_2}, Y^{e_2}) \in T_{ad}$ such that $P_{g(Y^{e_1})|\Phi(X^{e_1})} \neq P_{g(Y^{e_2})|\Phi(X^{e_2})}$ and the p.d.f.s of both $P_{g(Y^{e_1})|\Phi(X^{e_1})}$ and $P_{g(Y^{e_2})|\Phi(X^{e_2})}$ are in the linear logistic model.

(B) The p.d.f. of $P_{Y^I|\Phi^{\mathcal{X}_1}(X)}$ is in the linear logistic model.

Then, there exists $\lambda^* \in \mathbb{R}$ such that a minimizer $(\theta^\dagger, \theta_{ad}^\dagger, \Phi^\dagger)$ of the objective function

$$\min_{\theta,\theta_{ad},\Phi} \left\{ \mathcal{R}^{(X^*,Y^*)}(p_\theta \circ \Phi) + \lambda^* \cdot \sum_{(X^e,Y^e)\in T_{ad}} \|\nabla_{\hat{\theta}_{ad}=\theta_{ad}} \mathcal{R}^{(X^e,g(Y^e))}(p^{\mathcal{Z}|\mathcal{H}}_{\hat{\theta}_{ad}} \circ \Phi)\|^2 \right\} \tag{4.13}$$

is o.o.d. optimial, i.e.,

$$p_{\theta^\dagger} \circ \Phi^\dagger \in \operatorname*{argmin}_{p_\theta : \mathcal{X}\to\mathcal{P}_\mathcal{Y}} \mathcal{R}^{o.o.d.}(p_\theta),$$

where p_θ runs all the p.d.f.s, $p^{\mathcal{Z}|\mathcal{H}}_{\theta_{ad}}$ runs all linear logistic functions, and Φ runs all the variable selections.

Before proving Theorem 20, we prepare one lemma:

Lemma 21.

$$\min_{\theta_{ad},\Phi:\mathrm{Im}\Phi_2\neq\emptyset} \sum_{(X^e,Y^e)\in T_{ad}} \|\nabla_{\hat{\theta}_{ad}=\theta_{ad}} \mathcal{R}^{(X^e,g(Y^e))}(p^{\mathcal{Z}|\mathcal{H}}_{\hat{\theta}_{ad}} \circ \Phi)\|^2 \;>\; 0.$$

Proof. It suffices to prove that, for any projection Φ with $\mathrm{Im}\Phi_2 \neq \emptyset$ and $p^{\mathcal{Z}|\mathcal{H}}_{\theta_{ad}}$, there is $(X^e, Y^e) \in T_{ad}$ such that $\|\nabla_{\hat{\theta}_{ad}=\theta_{ad}} \mathcal{R}^{(X^e,g(Y^e))}(p^{\mathcal{Z}|\mathcal{H}}_{\hat{\theta}_{ad}} \circ \Phi)\|^2 \neq 0$. We prove this by contradiction. Suppose that there exist a projection Φ with $\mathrm{Im}\Phi_2 \neq \emptyset$ and $p^{\mathcal{Z}|\mathcal{H}}_{\theta_{ad}}$ which satisfy

$$\|\nabla_{\hat{\theta}_{ad}=\theta_{ad}} \mathcal{R}^{(X^e,g(Y^e))}(p^{\mathcal{Z}|\mathcal{H}}_{\hat{\theta}_{ad}} \circ \Phi)\|^2 = 0 \qquad (\forall (X^e, Y^e) \in T_{ad}).$$

From Assumption (A), take (X^{e_1}, Y^{e_1}) and (X^{e_2}, Y^{e_2}) in T_{ad} such that $P_{g(Y^{e_1})|\Phi(X^{e_1})} \neq P_{g(Y^{e_2})|\Phi(X^{e_2})}$ and the p.d.f.s of both $P_{g(Y^{e_1})|\Phi(X^{e_1})}$ and $P_{g(Y^{e_2})|\Phi(X^{e_2})}$ are in the logistic model.

Note that the risk is defined by the cross-entropy loss:

$$\mathcal{R}^{(X^e, g(Y^e))}(p^{\mathcal{Z}|\mathcal{H}}_{\hat{\theta}_{ad}} \circ \Phi) = -\int \log p^{\mathcal{Z}|\mathcal{H}}_{\hat{\theta}_{ad}}(g(Y^e)|\Phi(X^e)) dP_{X^e, Y^e}.$$

Then this is minimized in the space of linear logistic functions if and only if $p^{\mathcal{Z}|\mathcal{H}}_{\hat{\theta}_{ad}}$ equals to the p.d.f.s both of $P_{Y^e|\Phi(X^e)}$ on e_1 and e_2. From $\|\nabla_{\hat{\theta}_{ad}=\theta_{ad}} \mathcal{R}^{(X^e, g(Y^e))}(p^{\mathcal{Z}|\mathcal{H}}_{\hat{\theta}_{ad}} \circ \Phi)\|^2 = 0$ for (X^{e_1}, Y^{e_1}) and (X^{e_2}, Y^{e_2}), we can conclude that $p^{\mathcal{Z}|\mathcal{H}}_{\theta_{ad}}$ should equal to the p.d.f.s both of $P_{g(Y^{e_1})|\Phi(X^{e_1})}$ and of $P_{g(Y^{e_2})|\Phi(X^{e_2})}$. This contradicts with the assumption $P_{g(Y^{e_1})|\Phi(X^{e_1})} \neq P_{g(Y^{e_2})|\Phi(X^{e_2})}$. □

The rest of its proof is essentially same as one of Theorem 18, and hence we omit.

4.4.3 Proof of Theorem 10

Before the proof, let us rearrange some notations introduced in Section 3.4.2. Notations are the same as in Section 4.4.1. Recall that we assume, given hyperparameter λ, the minimization of (3.3) achieves the global optimum perfectly, which yields the projection (variable selection) $\Phi^\lambda(x) : \mathcal{X} \to \mathbb{R}^{n_\lambda}$ $(n_\lambda \leq n_1 + n_2)$ and the conditional p.d.f. of $P_{Y^{e^*}|\Phi^\lambda(X^{e^*})}$, denoted by $p^{*,\lambda}(y|\Phi^\lambda(x))$. The $\mathcal{X}_1$ and $\mathcal{X}_2$ components of $\Phi^\lambda(X)$ are denoted by $\Phi^\lambda_1(X)$ and $\Phi^\lambda_2(X)$, respectively.

We rephrase the o.o.d. risk (2.1) and its evaluation (3.4) by Method I with some notational rearrangements. For $\lambda \in \Lambda$ and the training variable (X^*, Y^*) for the target task, the conditional p.d.f. of $P(Y^*|\Phi^\lambda(X^*))$ given the selected variables is denoted by $p^{*,\lambda} : \mathbb{R}^{n_\lambda} \to \mathcal{P}_\mathcal{Y}$. Then, the the o.o.d. risk $\mathcal{R}^{o.o.d.}(\lambda)$ of $p^{*,\lambda} \circ \Phi^\lambda$ and its evaluation $\mathcal{R}^I(\lambda)$ ((3.4) in the main body) are represented as

$$\mathcal{R}^{o.o.d.}(\lambda) := \max_{(X,Y) \in T_{all}} \mathcal{R}^{(X,Y)}(p^{*,\lambda} \circ \Phi^\lambda),$$
$$\mathcal{R}^I(\lambda) := \max\left\{ \max_{(X,Y) \in T_{ad}} \mathcal{R}^{(X, g(Y))}(p^{*,\lambda} \circ \Phi^\lambda), \mathcal{R}^{(X^*, Y^*)}(p^{*,\lambda} \circ \Phi^\lambda) \right\},$$

respectively. We restate Theorem 10 with some notation arrangements:

Theorem 22 (Theorem 10 in the main body, with some notational arrangements)**.** *Assume that all domains* $T_{all} := \{(X^e, Y^e)\}_{e\in\mathcal{E}}$ *are fixed as (※) in Section 4.4.1; namely,*

$$T_{all} := \left\{ (X,Y) : \textit{a random variable on } \mathcal{X}\times\mathcal{Y} \,\middle|\, P_{\Phi^{\mathcal{X}_1}(X),Y} = P_{X_1^I,Y^I} \right\}.$$

Additionally, assume the following two conditions:

(I) there is $\lambda^I \in \Lambda$ *such that* $\Phi^{\lambda^I} = \Phi^{\mathcal{X}_1}$*, where* $\Phi^{\mathcal{X}_1}$ *is the projection to the* $\mathcal{X}_1$*-components.*

(II) Let p^* *be the p.d.f of* $P_{X^*,g(Y^*)}$*. For any* λ *with* $\mathrm{Im}\Phi_2^\lambda \neq \emptyset$*, there is* $(X^{e_\lambda}, Y^{e_\lambda}) \in T_{ad}$ *such that*

$$(x,z) \sim P_{X^{e_\lambda},g(Y^{e_\lambda})} \textit{ satisfies } p^*(z|\Phi^\lambda(x)) \leq e^{-\beta-\varepsilon} \textit{ with probability } 1 \textit{ in } P_{X^{e_\lambda},g(Y^{e_\lambda})}.$$

Here, $\varepsilon \in \mathbb{R}_{>0}$ *is a sufficiently small positive real number (that is,* $0 < \varepsilon \ll 1$*) and* $\beta := H(Y^*|(X_1^*))$ *is the conditional entropy of* $((X_1^*), Y^*)$*. Then, we have*

$$\operatorname*{argmin}_{\lambda\in\Lambda} \mathcal{R}^I(\lambda) \subset \operatorname*{argmin}_{\lambda\in\Lambda} \mathcal{R}^{o.o.d.}(\lambda).$$

To prove Theorem 22, we prepare three lemmas, in which the notations are the same as in Theorem 22 and conditions (I) and (II) in Theorem 22 are also imposed.

Lemma 23. $\lambda^I \in \operatorname*{argmin}_{\lambda\in\Lambda} \mathcal{R}^{o.o.d.}(\lambda)$.

Lemma 24. *If* $\hat{\lambda} \in \operatorname*{argmin}_{\lambda\in\Lambda} \mathcal{R}^I(\lambda)$*, then* $\mathrm{Im}\Phi_2^{\hat{\lambda}} = \emptyset$.

Lemma 25. *If* $\hat{\lambda} \in \Lambda$ *satisfies* $\mathrm{Im}\Phi_2^{\hat{\lambda}} = \emptyset$*, then* $\mathcal{R}^I(\hat{\lambda}) = \mathcal{R}^{o.o.d.}(\hat{\lambda})$.

We prove Theorem 22 based on the above lemmas, before proving them.

proof of Theorem 22
Take $\hat{\lambda} \in \operatorname{argmin} \mathcal{R}^I(\lambda)$. Then, $\mathrm{Im}\Phi_2^{\hat{\lambda}} = \emptyset$ holds by Lemma 24 and therefore, $\mathcal{R}^I(\hat{\lambda}) = \mathcal{R}^{o.o.d.}(\hat{\lambda})$ holds by Lemma 25. Moreover, $\mathcal{R}^{o.o.d.}(\hat{\lambda}) \geq \mathcal{R}^{o.o.d.}(\lambda^I)$ holds by Lemma 23 and $\mathcal{R}^{o.o.d.}(\lambda^I) = \mathcal{R}^I(\lambda^I)$ holds by Lemma 25 (since Φ^{λ^I} is the projection onto $\mathcal{X}_1$, $\mathrm{Im}\Phi_2^{\lambda^I} = \emptyset$). By the assumption $\hat{\lambda} \in \operatorname*{argmin}_{\lambda\in\Lambda} \mathcal{R}^I(\lambda)$, $\mathcal{R}^I(\lambda^I) \geq \mathcal{R}^I(\hat{\lambda})$ holds. Arranging

these inequalities, we obtain

$$\mathcal{R}^I(\hat{\lambda}) = \mathcal{R}^{o.o.d.}(\hat{\lambda}) \geq \mathcal{R}^{o.o.d.}(\lambda^I) = \mathcal{R}^I(\lambda^I) \geq \mathcal{R}^I(\hat{\lambda}), \tag{4.14}$$

in which the inequalities must be equalities. Hence, we obtain $\mathcal{R}^{o.o.d.}(\hat{\lambda}) = \mathcal{R}^{o.o.d.}(\lambda^I)$. Because λ^I achieves the minimum of $\mathcal{R}^{o.o.d.}$ (Lemma 23), so does $\hat{\lambda}$, which concludes the proof. □

Since Lemma 23 is proven in the proof of Theorem 18 (especially, the proof in Step 1), we may prove the others.

proof of Lemma 24 .
Let us prove the contraposition of Lemma 24. Take $\hat{\lambda} \in \Lambda$ with $\mathrm{Im}\Phi_2^{\hat{\lambda}} \neq \emptyset$. To prove that $\hat{\lambda} \notin \operatorname{argmin} \mathcal{R}^I(\lambda)$, we may prove that $\mathcal{R}^I(\hat{\lambda}) > \mathcal{R}^I(\lambda^I)$ since $\lambda^I \in \Lambda$ (Assumption (I) in the statement). It then suffices to prove the following:

$$\text{there exists } (\bar{X}, \bar{Y}) \in T_{ad} \text{ such that } \int -\log p^{*,\hat{\lambda}}(g(\bar{Y})|\Phi^{\hat{\lambda}}(\bar{X}))dP_{\bar{X},g(\bar{Y})} > \mathcal{R}^I(\lambda^I). \tag{4.15}$$

From Condition (II), we can take $(X^{e_{\hat{\lambda}}}, Y^{e_{\hat{\lambda}}}) \in T_{ad}$ such that

$$(x, z) \sim P_{X^{e_{\hat{\lambda}}},g(Y^{e_{\hat{\lambda}}})} \text{ satisfies } p^*(z|\Phi^{\hat{\lambda}}(x)) \leq e^{-\beta} - \varepsilon \text{ with probability 1.}$$

To prove (4.15), we prepare one supplementary inequality:

Supplementary Inequality

$$\int -\log p^{*,\hat{\lambda}}\big(g(Y^{e_{\hat{\lambda}}})|\Phi^{\hat{\lambda}}(X^{e_{\hat{\lambda}}})\big)dP_{X^{e_{\hat{\lambda}}},g(Y^{e_{\hat{\lambda}}})} \geq -\log\left\{e^{-\beta} - \epsilon\right\}.$$

This inequality can be easily seen; from the way of taking $e_{\hat{\lambda}}$, we have

$$-\log p^*(z|\Phi^{\hat{\lambda}}(x)) \geq -\log\{e^{-\beta} - \varepsilon\}$$

with probability 1 with respect to $(x, z) \sim P_{X^{e_{\hat{\lambda}}},g(Y^{e_{\hat{\lambda}}})}$, and thus the integration proves the inequality.

Proof of Inequality (4.15)

It follows from the above supplementary inequality that

$$\int -\log p^{*,\hat{\lambda}}(g(Y^{e_{\hat{\lambda}}})|\Phi^{\hat{\lambda}}(X^{e_{\hat{\lambda}}}))dP_{X^{e_{\hat{\lambda}}},g(Y^{e_{\hat{\lambda}}})} \geq -\log\left\{e^{-\beta}-\epsilon\right\} > \beta = H(Y^*|X_1^*). \tag{4.16}$$

Since $\Phi^{\lambda^I} = \Phi^{\mathcal{X}_1}$ by Condition (I), the discussion at (4.6) tells that $\mathcal{R}^I(\lambda^I) = H(Y^I|X_1^I) = H(Y^*|X_1^*)$, which concludes (4.15) and the proof.

Proof of Lemma 25

Take $\hat{\lambda} \in \Lambda$ that satisfies $\mathrm{Im}\Phi_2^{\hat{\lambda}} = \emptyset$. Then, $P_{\Phi^{\hat{\lambda}}(X),Y} = P_{\Phi^{\hat{\lambda}}(X^I),Y^I}$ holds for any $(X,Y) \in T_{all}$ because of $P_{X_1,Y} = P_{X_1^I,Y}$, and therefore, $\mathcal{R}^{(X,g(Y))}(p^{*,\hat{\lambda}} \circ \Phi^{\hat{\lambda}}) = \mathcal{R}^{(X^I,g(Y^I))}(p^{*,\hat{\lambda}} \circ \Phi^{\hat{\lambda}})$ and $\mathcal{R}^{(X^*,Y^*)}(p^{*,\hat{\lambda}} \circ \Phi^{\hat{\lambda}}) = \mathcal{R}^{(X^I,Y^I)}(p^{*,\hat{\lambda}} \circ \Phi^{\hat{\lambda}})$ hold. These two equalities lead the following equality:

$$\begin{aligned}\mathcal{R}^I(\hat{\lambda}) &= \max\left\{\max_{(X,Y)\in T_{ad}} \mathcal{R}^{(X,g(Y))}(p^{*,\hat{\lambda}} \circ \Phi^{\hat{\lambda}}), \mathcal{R}^{(X^*,Y^*)}(p^{*,\hat{\lambda}} \circ \Phi^{\hat{\lambda}})\right\} \\ &= \max\left\{\mathcal{R}^{(X^I,g(Y^I))}(p^{*,\hat{\lambda}} \circ \Phi^{\hat{\lambda}}), \mathcal{R}^{(X^I,Y^I)}(p^{*,\hat{\lambda}} \circ \Phi^{\hat{\lambda}})\right\}\end{aligned} \tag{4.17}$$

It follows from Theorem 7 that

$$\begin{aligned}&R^{(X^I,Y^I)}(p^{*,\hat{\lambda}} \circ \Phi^{\hat{\lambda}}) \\ &= \mathcal{R}^{(X^I,g(Y^I))}(p^{*,\hat{\lambda}} \circ \Phi^{\hat{\lambda}}) \\ &\quad + \sum_{z^{\not\prec} \in \not{\mathcal{Z}}} P(Y^I = g^{-1}(z^{\not\prec})) \int -\log p^{*,\hat{\lambda}}(Y^I|\Phi^{\hat{\lambda}}(X^I), g(Y^I) = z^{\not\prec}) dP_{X^I,Y^I|g(Y^I)=z^{\not\prec}} \\ &\geq \mathcal{R}^{(X^I,g(Y^I))}(p^{*,\hat{\lambda}} \circ \Phi^{\hat{\lambda}})\end{aligned}$$

Therefore, from (4.17), we have $\mathcal{R}^I(\hat{\lambda}) = \mathcal{R}^{(X^I,Y^I)}(p^{*,\hat{\lambda}} \circ \Phi^{\hat{\lambda}})$. Since $P_{\Phi^{\hat{\lambda}}(X),Y}$ are the same for any elements in T_{all}, we obtain

$$\mathcal{R}^{(X^I,Y^I)}(p^{*,\hat{\lambda}} \circ \Phi^{\hat{\lambda}}) = \max_{(X,Y)\in T_{all}} \mathcal{R}^{(X,Y)}(p^{*,\hat{\lambda}} \circ \Phi^{\hat{\lambda}}) = \mathcal{R}^{o.o.d.}(p^{*,\hat{\lambda}} \circ \Phi^{\hat{\lambda}}),$$

which concludes the proof.

4.4.4 Proof of Theorem 11

Before proving Theorem 11, we rephrase the evaluation (3.5) of the o.o.d. risk by Method II with some notation rearrangements. By using notation simplifications in Sections 4.4.1 and 4.4.3, the evaluation $\mathcal{R}^{II}(\lambda)$ by method II (corresponding to (3.5) in the main body) is represented as

$$\mathcal{R}^{II}(\lambda) := \max_{(X,Y)\in T_{ad}\cup\{(X^*,Y^*)\}} \left\{ \mathcal{R}^{(X,g(Y))}(p^{*,\lambda}\circ\Phi^\lambda) + D_\lambda(Y) \right\},$$

where the correction term $D_\lambda(Y)$ is defined by

$$D_\lambda(Y) := \sum_{z^{\not\prec}\in\mathcal{Z}^{\not\prec}} \left\{ P(g(Y)=z^{\not\prec}) \int -\log p^{*,\lambda}(Y^*|\Phi^\lambda(X^*), g(Y^*)=z^{\not\prec}) dP_{(X^*,Y^*)|g(Y^*)=z^{\not\prec}} \right\}$$

Note that, in $D_\lambda(Y)$, although the random variable Y is given by $(X,Y)\in T_{all}$, the marginal distributions of Ys are the same by the assumption of T_{all}. Thus, hereafter, we use D_λ for the notation, and

$$D_\lambda = \sum_{z^{\not\prec}\in\mathcal{Z}^{\not\prec}} \left\{ P(g(Y^*)=z^{\not\prec}) \int -\log p^{*,\lambda}(Y^*|\Phi^\lambda(X^*), g(Y^*)=z^{\not\prec}) dP_{(X^*,Y^*)|g(Y^*)=z^{\not\prec}} \right\}.$$

Note also that $\beta_\lambda = H(Y^*|X_1^*) - D_\lambda$. We restate Theorem 11 with some notation arrangements:

Theorem 26 (Theorem 11 in the main body, with some notation arrangements)**.** *Assume that all domains $T_{all} := \{(X^e,Y^e)\}_{e\in\mathcal{E}}$ are fixed as (※) in Section 4.4.1; namely,*

$$T_{all} := \left\{ (X,Y) : \textit{a random variable on } \mathcal{X}\times\mathcal{Y} \,\middle|\, P_{\Phi^{\mathcal{X}_1}(X),Y} = P_{X_1^I,Y^I} \right\}.$$

Notations are the same as in the statement of Theorem 22. In addition to the condition (I), assume the following condition (II)':

(II)' Let p^* be the p.d.f. of $P_{X^*,g(Y^*)}$. For any λ with $\mathrm{Im}\Phi_2^\lambda \neq \emptyset$, there is $(X^{e_\lambda},Y^{e_\lambda})\in$

T_{ad} *such that the following statement holds.*

$$(x,z) \sim P_{X^{e_\lambda},g(Y^{e_\lambda})} \text{ satisfies}$$
$$p^*(z|\Phi^\lambda(x)) \leq e^{-\beta_\lambda} - \varepsilon \text{ with probability } 1 \text{ in } P_{X^{e_\lambda},g(Y^{e_\lambda})}.$$

Here, ε *is some positive real number and*

$$\beta_\lambda := H(Y^*|X_1^*) - D_\lambda(Y^*).$$

Then, we have

$$\operatorname*{argmin}_{\lambda\in\Lambda} \mathcal{R}^{II}(\lambda) \subset \operatorname*{argmin}_{\lambda\in\Lambda} \mathcal{R}^{o.o.d.}(\lambda).$$

We first show lemmas before the proof of the theorem.

Lemma 27. *If* $\hat{\lambda} \in \operatorname*{argmin}_{\lambda\in\Lambda} \mathcal{R}^{II}(\lambda)$*, then* $\mathrm{Im}\Phi_2^{\hat{\lambda}} = \emptyset$*.*

Lemma 28. *If* $\hat{\lambda} \in \Lambda$ *satisfies* $\mathrm{Im}\Phi_2^{\hat{\lambda}} = \emptyset$*, then* $\mathcal{R}^{II}(\hat{\lambda}) = \mathcal{R}^{o.o.d.}(\hat{\lambda})$*.*

proof of Theorem 26
Combining the above two lemmas and Lemma 23, we can derive the required assertion in essentially the same manner as in the proof of Theorem 22.

proof of Lemma 27
Let us prove the contraposition of Lemma 27. Take $\hat{\lambda} \in \Lambda$ with $\mathrm{Im}\Phi_2^{\hat{\lambda}} \neq \emptyset$. To prove that $\hat{\lambda} \notin \operatorname{argmin} \mathcal{R}^{II}(\lambda)$, we may prove that $\mathcal{R}^{II}(\hat{\lambda}) > \mathcal{R}^{II}(\lambda^I)$ since $\lambda^I \in \Lambda$ (Assumption (I) in the statement). To show this, it suffices to prove the following statement:

$$\text{there is } (\bar{X},\bar{Y}) \in T_{ad} \text{ such that } \mathcal{R}^{(\bar{X},g(\bar{Y}))}(\hat{\lambda}) + D_{\hat{\lambda}} > \mathcal{R}^{II}(\lambda^I). \tag{4.18}$$

Take $(X^{e_{\hat{\lambda}}}, Y^{e_{\hat{\lambda}}}) \in T_{ad}$ as in Condition (II)'. Then, in the same way as the proof of Lemma 24, we have the following inequality:

$$\int -\log p^{*,\hat{\lambda}}\big(g(Y^{e_{\hat{\lambda}}})|\Phi^{\hat{\lambda}}(X^{e_{\hat{\lambda}}})\big) dP_{X^{e_{\hat{\lambda}}},g(Y^{e_{\hat{\lambda}}})} \geq -\log\left\{e^{-\beta_{\hat{\lambda}}} - \epsilon\right\},$$

which leads us to obtain

$$\mathcal{R}^{(X^{e_{\hat{\lambda}}},g(Y^{e_{\hat{\lambda}}}))}(\hat{\lambda}) + D_{\hat{\lambda}} > \beta_{\hat{\lambda}} + D_{\hat{\lambda}} = H(Y^*|X_1^*) = \mathcal{R}^{(Y^*,X^*)}(p^{*,\lambda^I} \circ \Phi^{\lambda^I}). \tag{4.19}$$

On the other hand, for any $(X, Y) \in T_{all}$ the marginal distribution of $(Y, \Phi^I(X))$ is the same as that of (Y^*, X_1^*). Noting that $\mathcal{R}^{(X,g(Y))}(p^{*,\lambda^I} \circ \Phi^{\lambda^I}) + D_{\lambda^I}$ depends only on (Y, X_1), we have

$$\mathcal{R}^{II}(\lambda^I) = \mathcal{R}^{(X^*,g(Y^*))}(p^{*,\lambda^I} \circ \Phi^{\lambda^I}) + D_{\lambda^I}. \tag{4.20}$$

Now, Lemma 7 implies

$$\mathcal{R}^{(Y^*,X^*)}(p^{*,\lambda^I} \circ \Phi^{\lambda^I}) = \mathcal{R}^{(X^*,g(Y^*))}(p^{*,\lambda^I} \circ \Phi^{\lambda^I}) + D_{\lambda^I}. \tag{4.21}$$

From (4.19), (4.20), and (4.21), we thus have

$$\mathcal{R}^{(X^{e_{\hat{\lambda}}},g(Y^{e_{\hat{\lambda}}}))}(p^{*,\lambda^I} \circ \Phi^{\lambda^I}) + D_{\hat{\lambda}} > \mathcal{R}^{II}(\lambda^I),$$

which shows (4.18) and completes the proof.

proof of Lemma 28 Take $\hat{\lambda} \in \Lambda$ such that $\mathrm{Im}\Phi_2^{\hat{\lambda}} = \emptyset$. It follows from $\mathrm{Im}\Phi_2^{\hat{\lambda}} = \emptyset$ that $P_{\Phi^{\hat{\lambda}}(X),Y} = P_{\Phi^{\hat{\lambda}}(X^*),Y^*}$ holds for all $(X, Y) \in T_{all}$. Therefore,

$$\mathcal{R}^{o.o.d.}(\hat{\lambda}) = \max_{(X,Y)\in T_{all}} \mathcal{R}^{(X,Y)}(p^{*,\hat{\lambda}} \circ \Phi^{\hat{\lambda}}) = \mathcal{R}^{(X^*,Y^*)}(p^{*,\hat{\lambda}} \circ \Phi^{\hat{\lambda}}).$$

Likewise, from the condition of $\hat{\lambda}$, the definition of $\mathcal{R}^{II}(\hat{\lambda})$ involves the same distribution for $(Y, \Phi^{\hat{\lambda}}(X))$, and thus

$$\mathcal{R}^{II}(\hat{\lambda}) = \mathcal{R}^{(X^*,g(Y^*))}(p^{*,\hat{\lambda}} \circ \Phi^{\hat{\lambda}}) + D_{\hat{\lambda}}.$$

In a similar way to the proof of Lemma 27, Theorem 7 tells

$$\mathcal{R}^{(X^*,Y^*)}(p^{*,\hat{\lambda}} \circ \Phi^{\hat{\lambda}}) = \mathcal{R}^{(X^*,g(Y^*))}(p^{*,\hat{\lambda}} \circ \Phi^{\hat{\lambda}}) + D_{\hat{\lambda}}.$$

This completes the proof.

4.4.5 Proof of Theorem 12

We rephrase Theorem 12 with some notation arrangements.

Theorem 29. *Notations are the same as in Theorem 22. Assume that* (X^*, Y^*) *satisfies the following condition:*

(A2) For a sufficiently small $\varepsilon \ll 1$*, any* λ *with* $\mathrm{Im}\Phi_2^\lambda \neq \emptyset$*, any* $a \in \mathrm{Im}\Phi_1^\lambda$*, and any* $b \in \mathcal{Y}$*, there exists* $c(\lambda, a, b)$[2] *such that*

$$P(Y^* = b | \Phi_1^\lambda(X^*) = a, \Phi_2^\lambda(X^*) = c) \geq (1 - e^{-\beta}) + \varepsilon.$$

Then, for any λ *with* $\mathrm{Im}\Phi_2^\lambda \neq \emptyset$*, there exists* $(X^{e_\lambda}, Y^{e_\lambda}) \in T_{all}$ *such that the inequality in Theorem 22 (ii) holds.*

Proof. Fix λ with $\mathrm{Im}\Phi_2^\lambda \neq \emptyset$. Take $(\bar{X}, \bar{Y}) \in T_{all}$ such that its probability measure corresponds to $\bar{P}_{X_2|Y,X_1} \times P_{Y^I, X_1^I}$, where $\bar{P}_{X_2|Y,X_1}$ is defined by, setting $\hat{c}(\hat{\lambda}, a, b)$ by

$$\hat{c}(\hat{\lambda}, a, b) \in \operatorname*{argmin}_{c \in \mathcal{X}_2} P(g(Y^*) = g(b) | \Phi_1^{\hat{\lambda}}(X^*) = \Phi_1^{\hat{\lambda}}(a), \Phi_2^{\hat{\lambda}}(X^*) = \Phi_2^{\hat{\lambda}}(c)),$$

$\bar{P}_{X_2|Y=b,X_1=a} := \delta_{X_2 = \hat{c}(\hat{\lambda}, a, b)}$. Here, for $c \in \mathcal{X}_2$, the probability measure $\delta_{X_2=c}$ on $\mathcal{X}_2$ denotes a Dirac measure at $c \in \mathcal{X}_2$. Before proving Theorem 12, we prepare the following inequalities:

Supplementary Inequality 1

$$\forall a \in \mathcal{X}_1, \forall b \in \mathcal{Y},$$
$$P\left(g(Y^*) = g(b) \,\middle|\, \Phi_1^{\hat{\lambda}}(X^*) = \Phi_1^{\hat{\lambda}}(a), \Phi_2^{\hat{\lambda}}(X^*) = \Phi_2^{\hat{\lambda}}\big(\hat{c}(\hat{\lambda}, a, b)\big)\right) \leq e^{-\beta} - \epsilon.$$

To see the fact, take $b^* \in \mathcal{Y}$ such that $g(b^*) \neq g(b)$ Note that such b^* always exists if $|\mathcal{Z}| \geq 2$ by the following reason. Take $\mathcal{Z} \ni z^* \neq g(b)$. By the surjectivity of g, $g^{-1}(z^*) \neq \emptyset$. Taking $b^* \in g^{-1}(z^*)$, $g(b^*) = z^* \neq g(b)$. Then, by the condition (ii) of Theorem 11 and $\mathrm{Im}\Phi_2^{\hat{\lambda}} \neq \emptyset$, there exists $c(\hat{\lambda}, a, b) \in \mathcal{X}_2$ such that

$$P\left(Y^* = b^* \,\middle|\, \Phi_1^{\hat{\lambda}}(X^*) = \Phi_1^{\hat{\lambda}}(a), \Phi_2^{\hat{\lambda}}(X^*) = \Phi_2^{\hat{\lambda}}\big(c(\hat{\lambda}, a, b)\big)\right) \geq 1 - e^{-\beta} + \epsilon.$$

[2] $c(\lambda, a, b)$ means $c \in \mathcal{X}_2$ is determined by given $\lambda \in \Lambda$, $a \in \mathcal{X}_1$, $b \in \mathcal{Y}$.

Therefore,

$$
\begin{aligned}
&P\left(g(Y^*)=g(b)\,\middle|\,\Phi_1^{\hat{\lambda}}(X^*)=\Phi_1^{\hat{\lambda}}(a),\Phi_2^{\hat{\lambda}}(X^*)=\Phi_2^{\hat{\lambda}}\big(\hat{c}(\hat{\lambda},a,b)\big)\right)\\
&=\min_{c\in\mathcal{X}_2}P\big(g(Y^*)=g(b)|\Phi_1^{\hat{\lambda}}(X^*)=\Phi_1^{\hat{\lambda}}(a),\Phi_2^{\hat{\lambda}}(X^*)=\Phi_2^{\hat{\lambda}}(c)\big)\\
&\leq P\left(g(Y^*)=g(b)\,\middle|\,\Phi_1^{\hat{\lambda}}(X^*)=\Phi_1^{\hat{\lambda}}(a),\Phi_2^{\hat{\lambda}}(X^*)=\Phi_2^{\hat{\lambda}}\big(c(\hat{\lambda},a,b)\big)\right)\\
&=1-\sum_{\bar{z}\neq g(b)}P\left(g(Y^*)=\bar{z}\,\middle|\,\Phi_1^{\hat{\lambda}}(X^*)=\Phi_1^{\hat{\lambda}}(a),\Phi_2^{\hat{\lambda}}(X^*)=\Phi_2^{\hat{\lambda}}\big(c(\hat{\lambda},a,b)\big)\right)\\
&\leq 1-P\left(g(Y^*)=g(b^*)\,\middle|\,\Phi_1^{\hat{\lambda}}(X^*)=\Phi_1^{\hat{\lambda}}(a),\Phi_2^{\hat{\lambda}}(X^*)=\Phi_2^{\hat{\lambda}}\big(c(\hat{\lambda},a,b)\big)\right)
\end{aligned}
$$

$$
\begin{aligned}
&\leq 1-P\left(Y^*=b^*\,\middle|\,\Phi_1^{\hat{\lambda}}(X^*)=\Phi_1^{\hat{\lambda}}(a),\Phi_2^{\hat{\lambda}}(X^*)=\Phi_2^{\hat{\lambda}}\big(c(\hat{\lambda},a,b)\big)\right)\\
&\leq 1-(1-e^{-\beta}+\epsilon)\\
&\leq e^{-\beta}-\epsilon.
\end{aligned}
$$

Proof of Theorem 29

We may prove that $P_{\bar{X},\bar{Y}}(A)=1$ where

$$
\left\{(x,y)\in\mathcal{X}\times\mathcal{Y}\,\middle|\,P\left(g(Y^*)=g(b)\,\middle|\,\Phi^{\hat{\lambda}}(X^*)=\Phi^{\hat{\lambda}}(x)\right)\leq e^{-\beta}-\epsilon\right\}.
$$

Then,

$$
\begin{aligned}
P_{\bar{X},\bar{Y}}(A)&=\int 1_A dP_{\bar{X},\bar{Y}}=\int 1_A d(\bar{P}_{X_2|Y,X_1}\times P_{Y^I,X_1^I})\\
&=\int dP_{Y^I,X_1^I}\int 1_A d\bar{P}_{X_2|Y,X_1}=\int dP_{Y^I,X_1^I}(x_1,y)\delta_{X_2=\hat{c}(\hat{\lambda},x_1,y)}(A_{(x_1,y)})
\end{aligned}
$$

holds where $A_{(x_1,y)}:=\{x_2\in\mathcal{X}_2|((x_1,x_2),y)\in\mathcal{X}\times\mathcal{Y}\}$. By the Supplementary Inequality 1, $\hat{c}(\hat{\lambda},x_1,y)\in A_{(x_1,y)}$ holds and therefore, $\delta_{X_2=\hat{c}(\hat{\lambda},x_1,y)}(A_{(x_1,y)})=1$, which leads us to the equation $\int dP_{Y^I,X_1^I}(x_1,y)\delta_{X_2=\hat{c}(\hat{\lambda},x_1,y)}(A_{(x_1,y)})=1$.

□

4.4.6 Proof of Theorem 13

The proof of Theorem 13 is essentially same as the one of Theorem 12 and therefore, we omit.

Chapter 5

Related Works

5.1 Transfer Learning

The proposed framework uses additional data from multiple domains as well as the training data for the target domains. The setting is relevant to Transfer Learning (TL) [Pan and Yang, 2010, Yang et al., 2020, Yosinski et al., 2014]. TLs try to improve the predictive performance on target domains with limited data supply, with the help of a large amount of data in additional domains. Its effectiveness is demonstrated in various real-world problems, including computer vision [Krizhevsky et al., 2012, Csurka, 2017], natural language processing [Ruder et al., 2019, Devlin et al., 2019], and reinforcement learning [Taylor and Stone, 2009].

The usual approach of TLs is to train a base network with a large amount of additional data, and then, copy its first n layers ($n \in \mathbb{N}_{>0}$) with the first n layers of neural networks used in the target domain prediction. The remaining layers of the target network are then randomly initialized and trained toward the target domain. The transferred feature layers can be *fine-tuned*, meaning that they are trained by samples from the target domains, or can be left *frozen*, meaning that they do not change during training on the new domain. Whether or not to fine-tune the first n layers of the target network depends on the size of the target dataset and the number of parameters in the first layers [Yosinski et al., 2014].

Although they show advantages in many learning problems, they may not work effectively in the current setting. When the transferred feature is fine-tuned, the model tends to learn spurious correlation in $\mathcal{D}^{e^*} \sim P_{X^{e^*},Y^{e^*}}$ and does not generalize to unseen domains $\mathcal{E} - \{e^*\}$. Even when frozen, the transferred feature often fails

to remove spurious correlation compared to our proposed approach. The fact is demonstrated empirically in Chapter 6.

5.2 Meta Learning

Meta-learning methods are related to the problem addressed in the thesis [Snell et al., 2017, Vinyals et al., 2016, Finn et al., 2017, Yoon et al., 2018]. The goal of mete-learning is training models (which are often called mete-learners) that can solve new target tasks using only a small number of training samples; it can be said that mete-learning is *learning to learn*. Such meta-learners are often trained by easily accessible samples generated by different tasks from target ones.

Some methods of few-shot and zero-shot learning [Snell et al., 2017, Vinyals et al., 2016] are successful methods for meta-learning. They try to generalize to new classes not seen in the training set, given only a small number of examples of each new class or given no examples of each new class. Snell et al. [2017], Vinyals et al. [2016] train prototype representations of each class, which enable us to generalize to new classes not seen in the training set. These approaches have generated some of the most successful results. Model-agnostic meta-learning (MAML) is a gradient-based meta-learning framework [Finn et al., 2017, Yoon et al., 2018]. In the frameworks, the parameters of the base network are explicitly trained by additional data such that a small number of gradient steps with a small amount of training data from a target task will produce good generalization performance on that task. MAML is known to be *model-agnostic*, in the sense that it is compatible with any model trained with gradient descent and hence, is applicable to a variety of different learning problems including classifications, regressions, and reinforce learnings [Finn et al., 2017].

The meta-learning framework is also unsuitable in our problem setting. Meta-learning framework improves predictive performance only on the target domain where any samples are available. In our problem settings, we can access samples of the target task only from e^*, not from $\mathcal{E} - \{e^*\}$; the meta-learning framework will not train any models which generalize well on unseen target domains $\mathcal{E} - \{e^*\}$.

5.3 Domain Adaptation by Deep Feature Learning

Unsupervised domain adaptation methods try to train a classifier that works well on a target domain on the condition that we are provided labeled source samples

and unlabeled target samples during training [Ganin et al., 2016, Ben-David et al., 2006, Louppe et al., 2017, Stojanov et al., 2021, Zhang et al., 2019, Long et al., 2015, Sun and Saenko, 2016]. Most of the previous deep domain adaptation methods try to obtain data representation $\Phi(X^e)$ that follows the same distribution for the training and test domains. Their training is done by minimizing the divergence between domains as well as a training loss on the source domain, such as maximum mean discrepancy [Long et al., 2015], correlation distance [Sun and Saenko, 2016], or adversarial discriminator accuracy [Ganin et al., 2016].

While the strategies sometimes lead to high predictive performance on a test domain similar to a training domain, such Φ does not function by discarding domain-specific factors from $X^e \in \mathcal{X}$ as theoretically noted in Arjovsky et al. [2020]. Experimental comparisons will be shown in Chapter 6.

5.4 Distributionally Robust Supervised Learning

Distributionally Robust Supervised Learning (DRSL) frameworks [Hu et al., 2018, Namkoong and Duchi, 2016, Sinha et al., 2019, Sagawa et al., 2020] introduce the concept o.o.d. risk in advance of Arjovsky et al. [2020], and proposed methods to minimize it. For a single training domain $\{e_{tr}\}$, DRSL tries to generalized on a small ε-ball centered at the training distribution $P_{X^{e_{tr}},Y^{e_{tr}}}$; more formally,

$$\{(X^e, Y^e)\}_{e\in\mathcal{E}} := \{(X, Y) : \text{a random variable on } \mathcal{X} \times \mathcal{Y} \,|D\left(P_{X^{e_{tr}},Y^{e_{tr}}} || P_{X,Y}\right) < \varepsilon\}.$$

Here, D denotes some divergence among distributions, including f-divergence [Hu et al., 2018, Namkoong and Duchi, 2016] and Wasserstein distance [Sinha et al., 2019]. Recently, Sagawa et al. [2020] considered different DRSL settings, which are often called *group Distributionally Robust Optimizations*. They set $\mathcal{E}$ as a small probability simplex which includes the training domain e_{tr}, instead of ε-balls.

DRSL methods also can not be applicable to our problem. Probability distributions that have different spurious correlations from ones in a training distribution are not necessarily included on a small ε-ball or a probability simplex which includes e_{tr}; for example, the distance between two distributions that generate images of cows on sandy beaches and green postures may be large. Domain Invariance Learning framework explicitly or implicitly imposes the following assumption on $\{(X^e, Y^e)\}_{e\in\mathcal{E}}$, which is different from the one on DRSL;

(#) There exist some space $\mathcal{H}$ and feature map $\Phi : \mathcal{X} \to \mathcal{H}$ which satisfy
" $P_{Y^{e_1}|\Phi(X^{e_1})} = P_{Y^{e_2}|\Phi(X^{e_2})}$ for any $e_1, e_2 \in \mathcal{E}$. "

Recall that, in Chapters 2, 3, and 4, the effectiveness of DILs is investigated on

$$\{(X^e, Y^e)\}_{e\in\mathcal{E}} := \left\{(X, Y) : \text{a random variable on } \mathcal{X} \times \mathcal{Y} \,\middle|\, P_{Y|\Phi^{\mathcal{X}_1}(X)} = P_{Y^I|X_1^I}\right\},$$

or

$$\{(X^e, Y^e)\}_{e\in\mathcal{E}} := \left\{(X, Y) : \text{a random variable on } \mathcal{X} \times \mathcal{Y} \,\middle|\, P_{\Phi^{\mathcal{X}_1}(X),Y} = P_{X_1^I,Y^I}\right\}.$$

They are examples of the distributions $\{(X^e, Y^e)\}_{e\in\mathcal{E}}$ which satisfy (#); $\mathcal{H}$ and Φ in (#) correspond to $\mathcal{X}_1$ and the projection $\Phi^{\mathcal{X}_1}$ onto $\mathcal{X}_1$ respectively.

5.5 Other Strategies

Bahng et al. [2020] try to obtain a de-biased feature Φ following the independence $\Phi(X) \perp\!\!\!\perp E$, seeing $\mathcal{E}$ as a random variable E. Recently, Wang et al. [2022] consider the setting where there exists some f in the model that $f \neq f^{o.o.d.}$, where $f^{o.o.d.}$ is an estimator with high prediction performance on both training and test domain, and that $f(x) = f^{o.o.d.}(x)$ for a sample x from training domains. Under the setting, they derive an upper bound of the risk on a test domain and propose a method for decreasing the upper bound. As a de-bias method, Nam et al. [2020] use two NNs; the first NN learns a biased mapping by the standard ERM, while the second one is trained with the samples that have large errors by the first NN. This method is based on the idea that the training with samples with large errors by the first NN mitigates data bias.

Chapter 6

Experiments

We study the effectiveness of the proposed framework and CVs through experiments, comparing them with several existing methods: empirical risk minimization (ERM), transfer learning (TL) methods, and deep domain adaptation strategies. We implement two kinds of ERM: its objective functions are evaluated only by $\mathcal{D}^{e^*}$ (ERM1), and by both $\mathcal{D}^{e^*}$ and coarser labeled data $\mathcal{D}^{e}$ (ERM2). For TLs, we employ two typical methods: *fine tune* (FT) and *frozen feature* (FF) [Pan and Yang, 2010, Yang et al., 2020, Yosinski et al., 2014]. As a deep domain adaptation technique, we adopt the state-of-the-art method *Domain-Specific Adversarial Network* (DSAN) [Stojanov et al., 2021]. We also compare our two CVs (CVI and CVII) with conventional CVs: training-domain validation (Tr-CV) and leave-one-domain-out cross-validation (LOD-CV) [Gulrajani and Lopez-Paz, 2023]. We have two hyperparameters to be selected by CV. In the training with (3.3), we set 1 when the training epoch is less than a certain threshold t, and $\lambda := \lambda_{\text{after}}$ if the epoch is larger than t; namely,

$$\lambda = \begin{cases} 1 & (\text{ epoch } \leq t) \\ \lambda_{\text{after}} & (\text{ epoch } > t) \end{cases}.$$

It is known that these two hyperparameters are critical for DIL methods to achieve good results. From a set of candidates, each of the CV methods selects a pair $(t, \lambda_{\text{after}})$. To know the best possible performance among the candidates, we also apply the test-domain validation (TDV) [Gulrajani and Lopez-Paz, 2023], which selects the hyperparameters with the unseen test domain, and thus is not applicable in practical situations. Additional experiments and experimental details can be found in Appendices A and B, respectively.

Figure 6.1: Visulization of Synthesized Data.

6.1 Synthesized Data

We compared the proposed method with the other approaches using synthesized data with $\mathcal{X} = \mathbb{R}^2$, $\mathcal{Y} = [3]$ and $\mathcal{Z} := [2]$. We used distributions

$$N_0 := \mathcal{N}(0, 10^2) \times \mathcal{N}(e, 10^2),$$
$$N_1 := \mathcal{N}(30, 10^2) \times \mathcal{N}(-4e, 10^2),$$
$$N_2 := \mathcal{N}(-30, 10^2) \times \mathcal{N}(-e, 10^2),$$

where $\mathcal{N}(a, b)$ denotes a normal distribution with its (mean, variance) $= (a, b)$. Given $x \sim N_i$, the task is to predict N_i among $i = 0, 1, 2$. The aim of DIL is to ignore the second component of x, as it works as a domain-specific factor. Given $e^* \in \mathbb{N}_{\geq 0}$ ranging from 0 to 50, each experiment draws $\mathcal{D}^{e^*} \sim P_{X^{e^*}, Y^{e^*}}$ with its sample size $n^{e^*} = 2000$, and then predicts Y^{-e^*} from X^{-e^*}. Setting g by $g(0) = 0$ and $g(1) = g(2) = 1$, we draw $\mathcal{D}^e_{ad} \sim P_{X^e, Z^e}$ from $\mathcal{E}_{ad} = \{-100, -50, 0, 50, 100\}$ with its sample size $n^e = 1000$ ($\forall e \in \mathcal{E}_{ad}$). These data are visualized as Fig. 6.1. Left and middle figures illustrate training and test data on $e^* = 5$ and 50, respectively. As e^* increases, the test data and train data are more different, and therefore ERM will yield lower performance. Right figure illustrates $\mathcal{D}^e_{ad}$. We model Φ by a 3-layer neural net. Setting the maximum epoch 500, we select (t, λ_{after}) from 3×5 candidates with $t \in \{0, 100, 200\}$ and $\lambda_{after} \in \{10^0, 10^1, ..., 10^4\}$ by each of the CV methods.

Table 6.1 shows the test accuracy of the estimates for $e = -e^*$ over 2000 random samples $(x, y) \sim P_{X^{-e^*}, Y^{-e^*}}$. *Oracle* shows the results of the experiments with the first component. The best scores are **bolded**. When $e^* = 0$ and 5, the distribution of training domain (e^*) are similar to the one of test ($-e^*$), and hence, the TL methods

	$e^*=0$	$e^*=5$	$e^*=10$	$e^*=15$	$e^*=20$	$e^*=25$	$e^*=30$	$e^*=35$	$e^*=40$	$e^*=45$	$e^*=50$
Oracle	906 (.007)										
ERM1	.789 (.218)	.791 (.174)	.637 (.188)	.329 (.201)	.324 (.328)	.311 (.260)	.159 (.193)	.140 (.171)	.132 (.161)	.166 (.147)	.051 (.101)
ERM2	.868 (.043)	.849 (.101)	.741 (.159)	.690(.141)	.591 (.138)	.651(.118)	.613 (.150)	.539 (.096)	.565(..017)	.600 (.035)	.689 (.177)
FT	**.899 (.000)**	**.863 (.001)**	.575 (.002)	.568 (.001)	.673 (.103)	.583 (.088)	.402 (.004)	.350 (.001)	.003 (.000)	.000 (.000)	.000 (.000)
FF	**.899 (.000)**	.861 (.002)	.540 (.102)	.568 (.001)	.673 (.102)	.628 (.001)	.401 (.001)	.351 (.002)	.066 (.132)	.000 (.000)	.000 (.000)
DSAN	.684 (.008)	.367 (.016)	.195 (.015)	.112 (.008)	.045 (.008)	.013 (.003)	.006 (.001)	.001(.001)	.000 (.000)	000 (.000)	.000 (.000)
Ours + Our CV I	.799 (.232)	.784 (.231)	**.884 (.021)**	**.875 (.044)**	**.815 (.098)**	**.738 (.209)**	**.865 (.047)**	**.659 (.233)**	**.666 (.285)**	**.776 (.080)**	**.699 (.255)**
Ours + Our CV II	.799 (.232)	.783 (.231)	**.884 (.021)**	**.875 (.044)**	**.815 (.098)**	**.738 (.209)**	**.865 (.047)**	**.659 (.233)**	.563 (.291)	**.776 (.080)**	**.699 (.255)**
Ours + Tr-CV	.790 (.230)	.776 (.225)	.609 (.163)	.491 (.095)	.366 (.147)	.248 (.192)	.376 (.033)	.215 (.168)	.148 (.127)	.189 (.108)	.031 (.138)
Ours + LOD-CV	.662 (.180)	.521 (.145)	.569 (.204)	.538 (.168)	.450 (.158)	.371 (.213)	.641 (.221)	.571 (.221)	.380 (.196)	.423 (.218)	.316 (.127)
Ours + TDV	.915 (.005)	.905 (.006)	.896 (.002)	.895 (.010)	.848 (.059)	.849 (.069)	.887 (.030)	.764 (.152)	.796 (.174)	.848 (.055)	.775 (.179)

Table 6.1: Average Test ACCs and SEs of Synthesized Data (5 runs)

yield high performances. As e^* increases, the difference between the training (e^*) and test ($-e^*$) distributions becomes larger, and the previous methods fail to achieve high accuracy. The proposed methods (Ours) keep higher performance than the others even for large e^*. Among the CV methods, our two methods (CVI, CVII) significantly outperform the others for larger e^*. For this data set, CVI and CVII show almost the same performance.

6.2 Colored MNIST

We apply our framework to *Colored MNIST* [Arjovsky et al., 2020] with $\mathcal{Y} = [10]$ and $\mathcal{Z} := [2]$. We aim to predict $Y^e \in \mathcal{Y}$ from digit image data $X^e \in \mathbb{R}^{2\times 24\times 24}$. The label Y^e is changed randomly to one of the rest uniformly with a probability of 15%. All digits in images are colored red or green. The domain $e \in [0, 1]$ controls the color of digits; the digits $Y^e > 4$ and $Y^e \leq 4$ are colored in red and green, respectively, with probability e. In training, $\mathcal{D}^{e^*} \sim P_{X^{0.1},Y^{0.1}}$ is drawn with sample size $n^{e^*} = 5000$, and in testing, Y^e is predicted from X^e for $e_2 := 0.9$. Regarding the coarser labels Z^e, the task is to predict $Z^e = 0$ for X^e in $0-4$ and $Z^e = 1$ for $5-9$ (that is, $g(Y^e) = 1$ if $Y^e > 4$ and else, $g(Y^e) = 0$). The label Z^e is swapped randomly with 15%. We set $\mathcal{E}_{ad} = \{0.1, 0.3, 0.5, 0.7, 0.9\}$ with $n^e = 5000$ for each $e \in \mathcal{E}_{ad}$. We model Φ by a 3-layer neural net. Setting the maximum epoch 500 and $\lambda_{\text{before}} := 1.0$, we select $(t, \lambda_{\text{after}})$ from 4×7 candidates with $t \in \{0, 100, 200, 300\}$, $\lambda_{\text{after}} \in \{10^0, 10^1, ..., 10^6\}$ by each of the CVs.

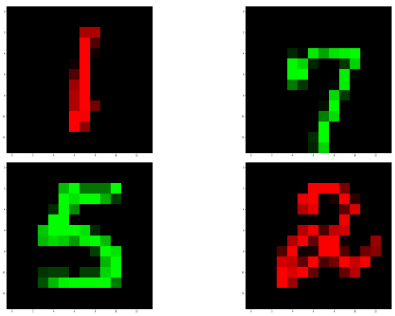

Figure 6.2: Colored MNIST Dataset.

Dataset	CMNIST	ImageNet $\mathcal{Y} = [3]$	ImageNet $\mathcal{Y} = [7]$	ImageNet $\mathcal{Y} = [17]$,
Best possible	.850			
random guess	.100	.333	.143	.059
Oracle	.822 (.000)	.743 (.018)	.749 (.008)	.708 (.010)
ERM 1	.630 (.006)	.417 (.016)	.507 (.020)	.357 (.020)
ERM 2	.751. (.002)	.606 (.014)	.535 (.005)	.465 (.008)
FT	.493 (.038)	.463(.030)	.409 (.020)	.361 (.011)
FF	.512 (.019)	.482 (.127)	.226 (.046)	.162 (.011)
DSAN	.091 (.005)	.278 (.004)	.293 (.008)	.060 (.007)
Ours + CV I	.673 (.006)	.652 (.028)	**.622 (.011)**	**.556 (.004)**
Ours + CV II	**.774 (.006)**	**.666 (.027)**	**.622 (.011)**	**.556 (.004)**
Ours + Tr-CV	.678 (.008)	.641 (.033)	.612 (.012)	.544 (.013)
Ours + LOD CV	**.774 (.006)**	.525(.028)	.572 (.022)	.528 (.019)
Ours + TDV	.774 (.006)	.673 (.035)	.634 (.033)	.556 (.004)

Table 6.2: Average Test Accuracies and SEs of Colored MNIST and ImageNet (5 runs)

Dataset	CVI	CVII	Tr-CV	LOD-CV
CMNIST	.102 (.006)	**.000 (.000)**	.102 (.007)	.003 (.002)
ImageNet: $\mathcal{Y} = [3]$	.027 (.029)	**.013 (.020)**	.025 (.021)	.170 (.041)
ImageNet: $\mathcal{Y} = [7]$	**.012 (.001)**	**.012 (.001)**	.018 (.015)	.054 (.024)
ImageNet: $\mathcal{Y} = [17]$	**.000 (.000)**	**.000 (.000)**	.001 (.002)	.025 (.021)

Table 6.3: Means and SEs of {(Accuracy of TDV on e_2) − (Accuracy of Each CV on e_2) } (5runs).

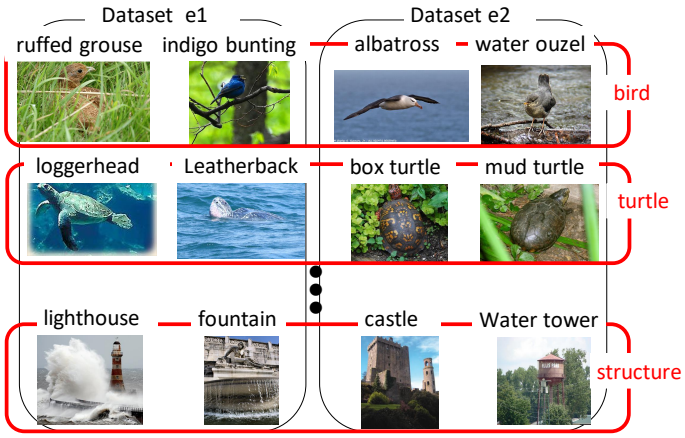

Figure 6.3: ImageNet Experiment Dataset.

Table 6.2 shows the test accuracies for 2000 random samples in the domain e_2. *Oracle* show the result of ERM with grayscale MNIST. The best scores are **bolded**. The results, together with additional ones in Appendices A.1 and A.2, demonstrate that the proposed methods outperform the others for e_2. Among the two proposed methods, CV II and LODCV yields higher test accuracies on e_2. Table 6.3 shows the accuracy gain of each CV from TDV with the same data sets for domain e_2. The lowest errors are **bolded**. These results, together with Appendices A.1 and A.2, concur with the theory in Section 3.4.2 suggesting that CVII succeeds in wider situations, resulting in smaller errors.

6.3 ImageNet

To see the performance of the proposed methods for more practical data, they are applied to the ImageNet [Deng et al., 2009] with its label re-annotated imitating BREEDS [Santurkar et al., 2022], which proposes a method for re-annotating ImageNet to create an o.o.d. benchmark. The target task here is to predict labels $Y^e \in \mathcal{Y}$ of images $X^e \in \mathbb{R}^{3\times 224\times 224}$. We conduct three experiments with $|\mathcal{Y}| = 3, 7, 17$. For each experiment, we prepare image datasets in different two manners e_1 and e_2. The datasets consist of images belonging to one of the classes $\mathcal{Y}$. 2, 4, and 8 classes out of 3, 7, and 17 classes, respectively, are composed of different subtypes between e_1 and e_2; for example, the images of class *bird* in e_1 are composed of ruffed grouse and indigo bunting, and the bird images on e_2 are composed of albatross and water ouzel

(Figure 6.3). In detail, show Appendix B.1. In training, $\mathcal{D}^{e^*} \sim P_{X^{e_1},Y^{e_1}}$ is drawn, and in testing, Y^e is predicted from X^e on e_2. The coarser label Z^e is binary (that is, $\mathcal{Z} = [2]$), and the sample with coarser labels $\mathcal{D}^e_{ad}$ of (X^e, Z^e) is drawn from both e_1 and e_2. Here, $\mathcal{D}^{e_1}_{ad}$ is the same as $\mathcal{D}^{e^*}$ but with labels re-annotated by g. We model Φ by ResNet50 [He et al., 2016]. Setting the maximum epoch 32 and $\lambda_{\text{before}} := 0.1$, we select $(t, \lambda_{\text{after}})$ from 3×4 candidates with $t \in \{10, 20, 30\}, \lambda_{\text{after}} \in \{0, 1, 10, 100\}$ by each of the CVs.

Table 6.2 shows the test accuracies on e_2. *Oracle* show the result of training with both e_1 and e_2. The best scores are **bolded**. We can see that the proposed framework succeeded in predicting on e_2, while the other methods failed. Table 6.3 shows the accuracy gain of each CV from TDV with the same data sets for domain e_2. The lowest errors are **bolded**. This result verifies that CVI and II select λ with the smallest error.

6.4 Comparison of Two CV Methods

To highlight the difference between the proposed two CVs, we compare them regarding the discrepancy between the additional domains $\mathcal{E}_{ad}$ and the domain for training of the target task e^*. We used synthesized data with $\mathcal{X} = \mathbb{R}^2$, $\mathcal{Y} = [10]$ and $\mathcal{Z} := [2]$, preparing ten distributions $\{N_i\}_{i=1}^{10}$ on $\mathbb{R}^2$, which include a domain-specific factor in the second component depending on $e \in \mathbb{Z}$. Explicit representations of $\{N_i\}_{i=1}^{10}$ are as follows:

$$
\begin{aligned}
N_1 &= \mathcal{N}(-180, 20^2) \times \mathcal{N}(-5e, 30^2), \\
N_2 &= \mathcal{N}(-100, 20^2) \times \mathcal{N}(-3e, 30^2), \\
N_3 &= \mathcal{N}(-20, 20^2) \times \mathcal{N}(-1e, 30^2), \\
N_4 &= \mathcal{N}(60, 20^2) \times \mathcal{N}(-2e, 30^2), \\
N_5 &= \mathcal{N}(140, 20^2) \times \mathcal{N}(-4e, 30^2), \\
N_6 &= \mathcal{N}(-140, 20^2) \times \mathcal{N}(4e, 30^2), \\
N_7 &= \mathcal{N}(-60, 20^2) \times \mathcal{N}(2e, 30^2), \\
N_8 &= \mathcal{N}(20, 20^2) \times \mathcal{N}(1e, 30^2), \\
N_9 &= \mathcal{N}(100, 20^2) \times \mathcal{N}(3e, 30^2), \\
N_{10} &= \mathcal{N}(180, 20^2) \times \mathcal{N}(5e, 30^2).
\end{aligned}
$$

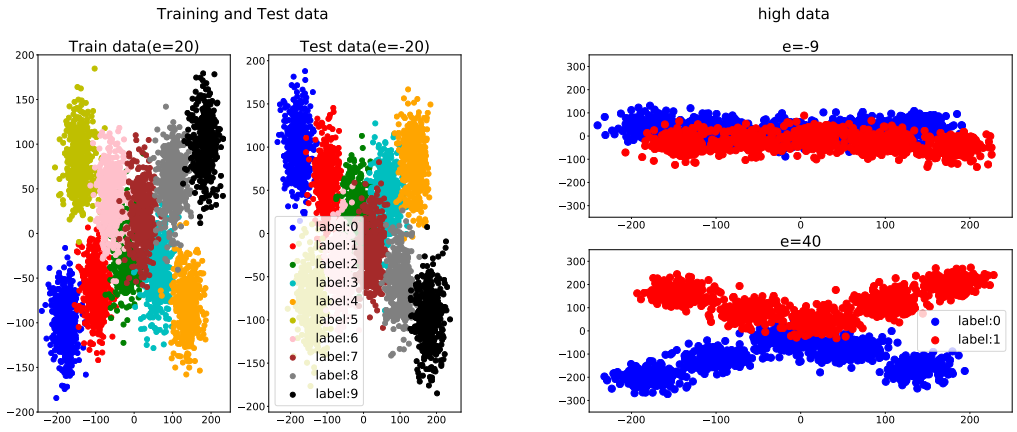

Figure 6.4: Data Visualization of Comparison of Two CV Methods.

	$e_{ad}=-9$	$e_{ad}=-8$	$e_{ad}=-7$	$e_{ad}=-6$	$e_{ad}=-5$	$e_{ad}=-4$	$e_{ad}=-3$	$e_{ad}=-2$	$e_{ad}=-1$	$e_{ad}=0$	$e_{ad}=1$
TDV	.596 (.078)	.621 (.046)	.630 (.041)	.595 (.061)	.590 (.087)	.621 (.059)	.564 (.071)	.582 (.056)	.535 (.093)	.520 (.121)	.575 (.107)
CV I	**.529 (.128)**	.555 (.111)	.562 (.086)	.566 (.109)	.375 (.145)	.346 (.172)	.372 (.176)	.358 (.167)	.300 (.146)	.173 (.143)	.218 (.087)
CV II	.527 (.152)	**.573 (.089)**	**.565 (.085)**	**.572 (.072)**	**.522 (.110)**	**.523 (.102)**	**.482 (.113)**	**.506 (.153)**	**.430 (.146)**	**.437 (.157)**	**.502 (.149)**

Table 6.4: Comparison of Two CVs: Average Test ACCs and SEs of the Estimates (10runs).

The task is to predict the distribution label $i \in \{1, \ldots, 10\}$. Setting $e^* := 20$ with $n^{e^*} = 60000$, the test task is to predict the label for domain $e = -20$. Regarding the task with coarser labels, we use $g(y) = 0$ if $y > 4$ and $g(y) = 0$ else. We draw $\mathcal{D}^e_{ad} \sim P_{X^e,Z^e}$ $(n^e = 20000)$ from $\mathcal{E}_{ad} = \{e_{ad}, 40\}$, where e_{ad} ranges from -9 to 1. As e_{ad} increases, e_{ad} approaches to e^*. Fig. 6.4 visualizes the data generating process. Left figure illustrates the training and test data of second experiment. Right figure illustrates $\mathcal{D}^{40}_{ad}$ and $\mathcal{D}^{e_{ad}}_{ad}$ with $e_{ad} = -9$.

The model Φ is a 3-layer neural net. We set the maximum epoch 500 and $t = 0$, and select λ_{after} from 4 candidates $\lambda_{\text{after}} \in \{0, 0.001, 80, 100\}$ by each CV method.

Table 6.4 shows the test accuracy on $e = -e^*$ with 2000 random samples $(x, y) \sim P_{X^{-e^*},Y^{-e^*}}$. From the results, we can see that CVII tends to select better hyperparameters than CVI, especially in the case where the variation among the domains is smaller as e_{ad} approaches to e^*. This accords with the theoretical results in Theorems 10 and 11, which show that CVII finds a correct hyperparameter in smaller discrepancy between $\mathcal{E}_{ad}$ and e^* than CVI.

	annotation	CVI	CVII	TDV
$\mathcal{Y} = [7]$	complete	.622 (.008)	.622 (.008)	.634 (.033)
	automatic	.629 (.013)	.630 (.011)	.631 (.011)
$\mathcal{Y} = [17]$	complete	.556 (.004)	.556 (.004)	.556 (.004)
	automatic	.552 (.013)	.552 (.013)	.554 (.009)

Table 6.5: Means and SEs of Pre-trained Classifier Experiment (5runs).

	$\mathcal{Y} = [7]$	$\mathcal{Y} = [17]$
acc.	.995 (.000)	.995 (.000)

Table 6.6: Means and SEs of Pre-trained Classifier Experiment (5runs).

6.5 Coarser Label Annotation with Pre-trained Classifiers

We demonstrate the performance of the proposed methods with coarser labels $Z^e = g(Y^e)$ annotated by pre-trained classifiers available on the Internet. In the previous experiments, we assume that the binary coarser labels Z^e are given in advance. In practice, its annotation may be done by a binary classifier on the Internet as well as by humans with crowd-sourcing. Recent progress in artificial intelligence enables us to access a high-quality, pre-trained classifier such as a ResNet pre-trained with ImageNet. Noting that classification ability is much higher for a task of a smaller number of classes as noted in Chapter 1, pre-trained classifiers will enable us to annotate precisely binary labels.

We prepare image datasets in Section 6.3 with $|\mathcal{Y}| = 7, 17$. In training, $\mathcal{D}^{e^*} = \{(x_i^{e_1}, y_i^{e_1})\} \sim P_{X^{e_1},Y^{e_1}}$ is drawn on e_1, and $\mathcal{D}^{e^2} = \{(x_i^{e_2})\} \sim P_{X^{e_2}}$, dataset of images without any labels, is drawn on e_2. We prepare ResNet50 [He et al., 2016] with its hyperparameter fixed following Vryniotis [20121] (the pre-trained parameter is available on Pytorch). After annotating Z^e by the pre-trained classifier, we apply our proposed method.

Table 6.5 shows the result of test accuracy on $e = e_2$. Here, the row "complete" shows the results with Z^e given in advance (that is, the same results as the ones in Section 5.2), and the row "automatic" shows the results with Z^e annotated by the pre-trained classifier. The result shows that our framework with automatic annotations gives the almost same result as one with completely coarser labels. Table 6.6 shows the accuracies of automatic annotation by a pre-trained classifier. The

result shows that automatic annotation of coarser labels Z achieves good accuracy so that we can use it reliably as the coarser label to extract invariance in the proposed methods.

Chapter 7

Conclusions

Out-of-distribution generalization is an important problem for the future of machine learning. Domain Invariance Learning framework is a rapidly developed approach for the out-of-distribution generalization problem. Their proposed estimator is composed of two maps, (i) a domain invariance Φ defined in Chapter 2 and (ii) a predictor of labels from a featured image $\Phi(x)$.

DIL approaches have two shortcomings in practice. The first shortcoming is annotation cost. While the conventional estimation of domain invariances demands training data from multiple domains, it often involves expensive and exhausting annotation. The second limitation is hyperparameter selection. While most DIL methods involve some hyperparameters to balance the classification accuracy and the degree of invariance, its selection from training data is known to have special difficulty.

The Ph.D. thesis has two contributions. Firstly, we have proposed a new domain invariance framework to reduce annotation costs: assuming the availability of datasets for another relevant task with coarser labels, we obtain a domain invariant predictor for the target classification task using training data in a *single* domain. Since the additional task with coarser labels involves lower annotation cost, our novel DIL demands lower and cheaper costs than ones needed in conventional DIL. Secondly, we have also proposed two cross-validation methods for hyperparameter selection. The key idea is to use additional coarser labeled data from multiple domains, in addition to training data of a single domain for the target task, for the derivation of o.o.d. risk. Theoretical analysis has revealed the correctness of our methods, including cross-validation methods, and the experimental results have demonstrated the effectiveness of the proposed framework and cross-validation methods.

There may be some limitations of the proposed methods, in that our method removes important information for a prediction as well as unnecessary ones. There are fewer domain invariance features among domains as the number of domains becomes larger; more formally, for given two domains $\mathcal{E}_1$ and $\mathcal{E}_2$ with $\mathcal{E}_1 \subset \mathcal{E}_2$,

$$\begin{aligned}&\left\{\Phi : \mathcal{X} \to \mathcal{H} \,\middle|\, P_{Y^{e_1}|\Phi(X^{e_1})} = P_{Y^{e_2}|\Phi(X^{e_2})} \text{ for all } e_1, e_2 \in \mathcal{E}_2 \right\} \\ &\qquad \subset \left\{\Phi : \mathcal{X} \to \mathcal{H} \,\middle|\, P_{Y^{e_1}|\Phi(X^{e_1})} = P_{Y^{e_2}|\Phi(X^{e_2})} \text{ for all } e_1, e_2 \in \mathcal{E}_1 \right\}\end{aligned}$$

holds. The inclusion shows that, in return for the domain invariance estimation, domain invariances estimated by our method may remove important factors on some domains especially when $|\mathcal{E}|$ is large.

7.1 Suggestions for Future Research

This section suggests possible extensions and developments of our analysis.

7.1.1 Discrepancy among Training Distributions

In Theorem 8, 9, 10 and 11, the effectiveness of the proposed objective function and CV methods are ensured only if at least two training domains have enough discrepancy in distributions. In general, different domains do not necessarily have a discrepancy in distributions; judging it is a further important problem.

7.1.2 Application to Medical Data

The proposed method should be applied to other problems in addition to the ones addressed in Chapter 6. Disease detection from X-ray scans is one of the important examples. The X-ray scan problem struggled with out-of-distribution generalization as noted in Chapter 1, and hence, it is important to apply DILs to the problem. Against the importance, its application is often difficult because of annotation costs; annotating the sub-types of diseases would require expert knowledge. Our new framework is expected to mitigate the expensive cost; as classification becomes coarser, its annotation demands less expert knowledge.

7.1.3 Application to Problems in Fairness

Moreover, our methods may be useful for fairness problems in machine learning [Dastin, 2018]. Several models trained by conventional methods have been pointed out as making decisions based on discriminatory factors, for example, gender or nationality, and estimation without them is known to be important. Our domain invariance estimation will be available beyond the image recognition task, and hence, will enable us to train models which predict labels without using any discriminatory factors.

7.1.4 Scope of Proposed CV Methods

The applicable scope of the proposed CV methods should be investigated. In the thesis, we only apply the proposed CV methods to hyperparameter selection on our proposed objective function. As noted in Chapter 2, various objective functions for DILs are proposed, and they also include hyperparameters to be selected from training data. It is expected that the two CVs can also select these hyperparameters. Moreover, our CVs may select other parameters, such as neural network architectures or running rates. This should be also investigated.

7.1.5 Improvement of the Proposed CVs via Inequalities in Theorem 7 and 8

The inequalities (ii) and (ii)' in Theorem 7 and 8 define the notion of *how good* our CV methods are; the second method is *better* than the first one since $\beta_\lambda \leq \beta$. Through inequalities (ii) and (ii)', the notion of *improvement* can be also considered; new methods which attain $\beta^* \leq \beta_\lambda$ are better than the second CV method. the new concept of improvement is expected to open new doors to the development of hyperparameter selection for o.o.d. generalization problem.

References

K. Ahuja, K. Shanmugam, K. Varshney, and A. Dhurandhar. Invariant Risk Minimization Games. In *Proceedings of the 37th International Conference on Machine Learning*, pages 145–155, 2020.

E. A. AlBadawy, A. Saha, and M. A. Mazurowski. Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing. *Medical Physics*, 45(3):1150–1158, 2018.

M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant Risk Minimization. *arXiv:1907.02893*, 2020.

H. Bahng, S. Chun, S. Yun, J. Choo, and S. J. Oh. Learning De-biased Representations with Biased Representations. In *Proceedings of the 37th International Conference on Machine Learning*, pages 528–539, 2020.

S. Beery, G. Van Horn, and P. Perona. Recognition in Terra Incognita. In *European Conference on Computer Vision*, pages 472–489, 2018.

S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. Analysis of Representations for Domain Adaptation. In *Advances in Neural Information Processing Systems*, volume 19, 2006.

T. Cour, B. Sapp, and B. Taskar. Learning from Partial Labels. *Journal of Machine Learning Research*, 12(42):1501–1536, 2011.

E. Creager, J.-H. Jacobsen, and R. Zemel. Environment Inference for Invariant Learning. In *Proceedings of the 38th International Conference on Machine Learning*, pages 2189–2200, 2021.

G. Csurka. A Comprehensive Survey on Domain Adaptation for Visual Applications. In *Domain Adaptation in Computer Vision Applications*, Advances in Computer

Vision and Pattern Recognition, pages 1–35. Springer International Publishing, 2017.

J. Dastin. Amazon scraps secret AI recruiting tool that showed bias against women, say sources. https://www.japantimes.co.jp/news/2018/10/11/business/tech/amazon-scraps-secret-ai-recruiting-tool-showed-bias-women-sources/, 2018.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

C. Finn, P. Abbeel, and S. Levine. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1126–1135, 2017.

Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky. Domain-Adversarial Training of Neural Networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

X. Gu, J. Sun, and Z. Xu. Spherical Space Domain Adaptation With Robust Pseudo-Label Loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9101–9110, 2020.

I. Gulrajani and D. Lopez-Paz. In Search of Lost Domain Generalization. In *International Conference on Learning Representations*, 2023.

K. He, X. Zhang, S. Ren, and J. Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.

K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

W. D. Heaven. Google's medical AI was super accurate in a lab. Real life was a different story. https://www.technologyreview.com/2020/04/27/1000658/google-medical-ai-accurate-lab-real-life-clinic-covid-diabetes-retina-disease/, 2020.

C. Heinze-Deml, J. Peters, and N. Meinshausen. Invariant Causal Prediction for Nonlinear Models. *Journal of Causal Inference*, 6(2), 2018.

W. Hu, G. Niu, I. Sato, and M. Sugiyama. Does Distributionally Robust Supervised Learning Give Robust Classifiers? In *Proceedings of the 35th International Conference on Machine Learning*, pages 2029–2037, 2018.

P. Kamath, A. Tangella, D. Sutherland, and N. Srebro. Does Invariant Risk Minimization Capture Invariance? In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pages 4069–4077, 2021.

D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*, 2015.

M. Koyama and S. Yamaguchi. When is invariance useful in an Out-of-Distribution Generalization problem ? *arXiv:2008.01883*, 2021.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, D. Zhang, R. L. Priol, and A. Courville. Out-of-Distribution Generalization via Risk Extrapolation (REx). In *Proceedings of the 38th International Conference on Machine Learning*, pages 5815–5826, 2021.

K. Lakshminarayan, S. A. Harp, and T. Samad. Imputation of Missing Data in Industrial Databases. *Applied Intelligence*, 11(3):259–275, 1999.

Y. Lin, H. Dong, H. Wang, and T. Zhang. Bayesian Invariant Risk Minimization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

J. Liu, Z. Hu, P. Cui, B. Li, and Z. Shen. Heterogeneous Risk Minimization. In *Proceedings of the 38th International Conference on Machine Learning*, pages 6804–6814, 2021a.

J. Liu, Z. Hu, P. Cui, B. Li, and Z. Shen. Kernelized Heterogeneous Risk Minimization. *arXiv:2110.12425*, 2021b.

M. Long, Y. Cao, J. Wang, and M. I. Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, pages 97–105, 2015.

G. Louppe, M. Kagan, and K. Cranmer. Learning to Pivot with Adversarial Networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

C. Lu, Y. Wu, J. M. Hernández-Lobato, and B. Schölkopf. Invariant Causal Representation Learning for Out-of-Distribution Generalization. In *International Conference on Learning Representations*, 2022.

M. Moravčík, M. Schmid, N. Burch, V. Lisý, D. Morrill, N. Bard, T. Davis, K. Waugh, M. Johanson, and M. Bowling. DeepStack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 356(6337):508–513, 2017.

J. Nam, H. Cha, S. Ahn, J. Lee, and J. Shin. Learning from Failure: De-biasing Classifier from Biased Classifier. In *Advances in Neural Information Processing Systems*, volume 33, pages 20673–20684, 2020.

H. Namkoong and J. C. Duchi. Stochastic Gradient Methods for Distributionally Robust Optimization with f-divergences. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

S. J. Pan and Q. Yang. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

Paperswithcode.com. CIFAR-10 Benchmark (Image Classification) — Papers With Code. https://paperswithcode.com/sota/image-classification-on-cifar-10, 2023a.

Paperswithcode.com. Papers with Code - CIFAR-100 Benchmark (Image Classification). https://paperswithcode.com/sota/image-classification-on-cifar-100, 2023b.

G. Parascandolo, A. Neitz, A. Orvieto, L. Gresele, and B. Schölkopf. Learning explanations that are hard to vary. In *International Conference on Learning Representations*, 2022.

C. S. Perone, P. Ballester, R. C. Barros, and J. Cohen-Adad. Unsupervised domain adaptation for medical imaging segmentation with self-ensembling. *NeuroImage*, 194:1–11, 2019.

J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference using invariant prediction: Identification and confidence intervals. *arXiv:1501.01332*, 2015.

H. Pham, Z. Dai, Q. Xie, and Q. V. Le. Meta Pseudo Labels. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11552–11563, 2021.

D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1278–1286, 2014.

M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters. Invariant Models for Causal Transfer Learning. *Journal of Machine Learning Research*, 19(36):1–34, 2018.

E. Rosenfeld, P. K. Ravikumar, and A. Risteski. The Risks of Invariant Risk Minimization. In *International Conference on Learning Representations*, 2021.

D. Rothenhäusler, N. Meinshausen, P. Bühlmann, and J. Peters. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(2):215–246, 2021.

S. Ruder, M. E. Peters, S. Swayamdipta, and T. Wolf. Transfer Learning in Natural Language Processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, 2019.

S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally Robust Neural Networks. In *International Conference on Learning Representations*, 2020.

S. Santurkar, D. Tsipras, and A. Madry. BREEDS: Benchmarks for Subpopulation Shift. In *International Conference on Learning Representations*, 2022.

J. Shane. Do neural nets dream of electric sheep? - AI WeirdnessCommentShareCommentShare. https://www.aiweirdness.com/do-neural-nets-dream-of-electric-18-03-02/, 2018.

D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

A. Sinha, H. Namkoong, and J. Duchi. Certifying Some Distributional Robustness with Principled Adversarial Training. In *International Conference on Learning Representations*, 2019.

J. Snell, K. Swersky, and R. Zemel. Prototypical Networks for Few-shot Learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2256–2265, 2015.

P. Stojanov, Z. Li, M. Gong, R. Cai, J. Carbonell, and K. Zhang. Domain Adaptation with Invariant Representation Learning: What Transformations to Learn? In *Advances in Neural Information Processing Systems*, volume 34, pages 24791–24803, 2021.

B. Sun and K. Saenko. Deep CORAL: Correlation Alignment for Deep Domain Adaptation. In *European Conference on Computer Vision Workshops*, pages 443–450, 2016.

H. Tan, G. Feng, J. Feng, W. Wang, Y.-J. Zhang, and F. Li. A tensor-based method for missing traffic data completion. *Transportation Research Part C: Emerging Technologies*, 28:15–27, 2013.

M. E. Taylor and P. Stone. Transfer Learning for Reinforcement Learning Domains: A Survey. *Journal of Machine Learning Research*, 10(56):1633–1685, 2009.

S. Toyota and K. Fukumizu. Invariance Learning based on Label Hierarchy. In *Advances in Neural Information Processing Systems*, volume 36, 2022.

O. Vinyals, C. Blundell, T. Lillicrap, k. kavukcuoglu, and D. Wierstra. Matching Networks for One Shot Learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

V. Vryniotis. Improve the accuracy of Classification models by using SOTA recipes and primitives · Issue #3995 · pytorch/vision. https://github.com/pytorch/vision/issues/3995, 20121.

H. Wang, Z. Huang, H. Zhang, Y. J. Lee, and E. P. Xing. Toward learning human-aligned cross-domain robust models by countering misaligned features. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, pages 2075–2084, 2022.

P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. https://resolver.caltech.edu/CaltechAUTHORS:20111026-155425465, 2010.

N. Xu, J. Lv, and X. Geng. Partial Label Learning via Label Enhancement. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):5557–5564, 2019.

Y. Yan and Y. Guo. Partial Label Learning with Batch Label Correction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):6575–6582, 2020.

Q. Yang, Y. Zhang, W. Dai, and S. J. Pan. *Transfer Learning*. Cambridge University Press, 2020.

J. Yoon, T. Kim, O. Dia, S. Kim, Y. Bengio, and S. Ahn. Bayesian Model-Agnostic Meta-Learning. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Y. Zhang, H. Tang, K. Jia, and M. Tan. Domain-Symmetric Networks for Adversarial Domain Adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5026–5035, 2019.

Z. Zheng, L. Zheng, and Y. Yang. Unlabeled Samples Generated by GAN Improve the Person Re-identification Baseline in Vitro. In *IEEE International Conference on Computer Vision*, pages 3774–3782, 2017.

B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 Million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018.

Appendix A

Additional Experiment

A.1 Additional Experiments: Colored MNIST in Section 6.2

Although the Colored MNIST experiment in Chapter 6 fixes its flip rate to 15%, we additionally demonstrate by changing its flip rate among $\{10\%, 15\%, 20\%, 25\%\}$.

flip rate	Test Acc. on	Oracle	ERM1	ERM2	FT	FF	DSAN	Ours + CV1	Ours + CV2	Ours+TDV
0.25	$e = 0.1$	.715(.001)	.693(.001)	.697(.001)	.676(.003)	.677(.002)	.593(.007)	.706(.005)	.664 (.013)	.690 (.008)
	$e = 0.9$		.433 (.004)	.633(.002)	.250 (.020)	.248 (.015)	.073(.003)	.753(.011)	.618 (.018)	.657(.008)
0.20	$e = 0.1$	.769(.001)	.800(.001)	.734(.001)	.727(.002)	.725(.004)	.639(.003)	.752(.006)	.721 (.015)	.745 (.007)
	$e = 0.9$		.525 (.004)	.697(.002)	.368 (.029)	.364(.011)	.080(.004)	.576(.014)	.685 (.019)	.719 (.004)
0.15	$e = 0.1$	.822(.000)	.802(.002)	.786(.001)	.782(.006)	.786(.003)	.682(.002)	.806(.006)	.794 (.008)	.794 (.008)
	$e = 0.9$		.630 (.006)	.751(.002)	.493 (.038)	.512(.019)	.091(.005)	.673(.006)	.774 (.006)	.774 (.006)
0.10	$e = 0.1$	.872(.001)	.848(.002)	.829(.002)	.827(.004)	.829(.003)	.593(.007)	.857(.005)	.842 (.008)	.834 (.001)
	$e = 0.9$		.719 (.004)	.808 (.002)	.611(.016)	.623 (.021)	.073(.003)	.756(.007)	.800 (.007)	.821 (.006)

Table A.1: Test Acc. of Colored MNIST (5runs)

	Tr-CV	LOD-CV
0.25	.702 (.002) .597 (.006)	.590 (.004) .460 (.197)
0.20	.754 (.004) .678 (.008)	.716 (.018) .692 (.010)
0.15	.801 (.016) .678 (.008)	.787 (.004) .774 (.006)
0.10	.854 (.005) .751 (.013)	.836 (.004) .819 (.008)

Table A.2: Baselines of CV Methods

	CV I	CV II	Tr-CV	LOD-CV
0.25	.051 (.053)	.040 (.017)	.163 (.006)	.197 (.205)
0.20	.143 (.012)	.034 (.017)	.132 (.008)	.023 (.018)
0.15	.102 (.006)	.000 (.000)	.102 (.007)	.003 (.002)
0.10	.065 (005)	.021 (.010)	.075 (.010)	.005 (.002)

Table A.3: Means and SEs of {(Accuracy of TDV on $e = 0.9$) -(Accuracy of Each CV on $e = 0.9$) } (5runs).

Table A.1 and A.2 show that, among several CV methods, our method II keeps a high predictive performance regardless of flipping rates. Table A.3 shows the difference between accuracies by TDV and each CV for the same data set with $e = 0.9$. The result verifies that CVII selects preferable hyperparameters with smaller errors.

A.2 Additional Experiments: Colored MNIST II

We conduct an additional Colored MNIST experiment, changing annotation and coloring rules from ones in Section 6.2. Setting $\mathcal{Y} = [3]$ and $\mathcal{Z} := [2]$, we aim to predict Y^e from digit image data X^e, which are in the three categories $0-2$ ($Y^e = 0$), 3 or 4 ($Y^e = 1$) and $5-9$ ($Y^e = 2$). The label is changed randomly to one of the rest with a some probability ranging from $\{10\%, 15\%, 20\%, 25\%\}$. The domain index $e \in [0.0, 1.0]$ controls the color of the digit; for $Y^e = 0, 1$, the digit is colored in red with probability e and for $Y^e = 2$ colored in green with probability e. In the experiment, $\mathcal{D}^{e^*} \sim P_{X^{0.1},Y^{0.1}}$ is drawn with sample size $n^{e^*} = 5000$, and Y^e is predicted based on X^e for $e = 0.1$ and 0.9. Regarding Z^e, we consider the task where we predict $Z^e = 0$ for X^e in $0-2$ and $Z^e = 1$ for $3-9$ (that is, $g(0) = 0$ and $g(1) = g(2) = 1$).

We obtain the final label by flipping with some probability. As the domain-specific factor, we color the digit red for $Z^e = 0$ with probability e and green for $Z^e = 1$ with probability e. We set $\mathcal{E}_{ad} = \{0.1, 0.3, 0.5, 0.7, 0.9\}$ with $n^e = 5000$ for $\forall e \in \mathcal{E}_{ad}$. We model Φ by a 3-layer neural net. With the maximum epoch 500, we select (t, λ_{after}) from 3×10 candidates with $t \in \{0, 100, 200\}, \lambda_{after} \in \{10^0, 10^1, ..., 10^9\}$ by each CV method.

flip rate	Test Acc. on	Oracle	ERM1	ERM2	FT	FF	DSAN	Ours + CVI	Ours + CVII	Ours+ TDV
0.25	$e = 0.1$	.729 (.004)	.771 (.001)	.776(.002)	.771 (.001)	..771 (.001)	.767 (.004)	.727 (.004)	.714 (.013)	.673 (.006)
	$e = 0.9$		.125 (.003)	.277(.002)	.128 (.002)	.131 (.002)	.085 (.003)	.622 (.015)	.644 (.019)	.690 (.009)
0.20	$e = 0.1$	.780 (.002)	.796 (.000)	.801 (.001)	.800 (.001)	.796 (.001)	.789 (.004)	.773 (.003)	.745 (.008)	.738 (.018)
	$e = 0.9$		.177 (.006)	.353(.004)	.201 (.004)	.200(.007)	.091 (.005)	.644 (.011)	.707 (.012)	.732 (.008)
0.15	$e = 0.1$	.828 (.004)	.822 (.000)	.830 (.001)	.823 (.001)	.824 (.002)	.815 (.002)	.814 (.007)	.797 (.011)	.822 (.001)
	$e = 0.9$		.277 (.007)	.453 (.004)	.323 (.006)	.312 (.012)	.091 (.002)	.724 (.037)	.743 (.020)	.782 (.012)
0.10	$e = 0.1$	.880 (.004)	.852 (.002)	.861 (.001)	.855(.001)	.856 (.001)	.833(.003)	.848 (.005)	.848 (.005)	.857 (.005)
	$e = 0.9$		.468 (.002)	.450 (.018)	.497 (.005)	.500 (.007)	.106 (.010)	.792 (.005)	.792 (.005)	.829 (.005)

Table A.4: Test Accuracy for Colored MNIST (5runs)

	Tr-CV	LOD-CV
0.25	.759 (.008)	.362 (.059)
	.459 (.012)	.372 (.037)
0.20	.794 (.004)	.338 (.048)
	.541 (.007)	.334 (.029)
0.15	.834 (.002)	.348 (.031)
	.634 (.008)	.358 (.024)
0.10	.876 (.003)	.502 (.196)
	.708 (.006)	.497 (.194)

Table A.5: Baselines of CV Methods

	CV I	CV II	Tr-CV	LOD-CV
0.25	.068 (.007)	.046 (.023)	.231 (.013)	.319 (.033)
0.20	.088 (.004)	.025 (.006)	.191 (.014)	.398 (.025)
0.15	.059 (.038)	.039 (.022)	.148 (.019)	.430 (.028)
0.10	.037 (.010)	.037 (.010)	.121 (.008)	.332 (.196)

Table A.6: Means and SEs of {(Accuracy of TDV on $e = 0.9$) -(Accuracy of Each CV on $e = 0.9$) } (5runs).

Table A.4 and A.5 show test accuracies for 2000 random samples in the domains $e = 0.1$ and $e = 0.9$. The results demonstrate that the proposed methods significantly outperform the others for $e = 0.9$. Among the two proposed methods, CV II yields

the higher test accuracy. Table A.6 shows the difference between accuracies by TDV and each CV for the same data set with $e = 0.9$. The results verify that CVII selects preferable hyperparameters with smaller errors.

A.3 Additional Experiment: Bird Recognition

Our method is applied to the Bird recognition problem [Sagawa et al., 2020], which aims to predict three labels Y^e of images X^e: *waterbird* (Y^e= 0), *landbird* (Y^e= 1) and *no bird* (Y^e= 2). The dataset is made by combining background images from the Places dataset [Zhou et al., 2018] and bird images from the CUB dataset [Welinder et al., 2010] in two different ways $\mathcal{E} := \{e_1, e_2\}$. In domain e_1, we prepare three types of images: landbird image with land background, waterbird image with water background, and no bird with land background (Figure A.1, left). In domain e_2, we have landbird images with water background, waterbird images with land background, and no bird with water background (Figure A.1, right). For the sample of the target task, we used the domain $e^* = e_1$ and generated $n^{e^*} = 8649$ data $\mathcal{D}^{e^*} \sim P_{X^{e_1},Y^{e_1}}$. The sample with coarser labels $\mathcal{D}^e_{ad}$ of (X^e, Z^e), whose label is *landbird* ($Z^e = 0$) and *no landbird* ($Z^e = 1$) (*i.e.*, $g(1) = 0$ and $g(0) = g(2) = 1$), is drawn from both e_1 and e_2 with $n^{e_1} = n^{e_2} = 8649$. Here, we use $\mathcal{D}^{e^*}$ as $\mathcal{D}^{e_1}_{ad}$ with labels of $\mathcal{D}^{e^*}$ re-annotated by g. We made a predictor of Y^e based on X^e, and evaluated the test accuracy in the two domains $e = e_1, e_2$. We model Φ by ResNet50 [He et al., 2016]. Setting the maximum epoch 5, we select (t, λ_{after}) from 5×5 candidates with $t \in [5], \lambda_{after} \in \{10^0, 10^1, ..., 10^4\}$ by each CV method.

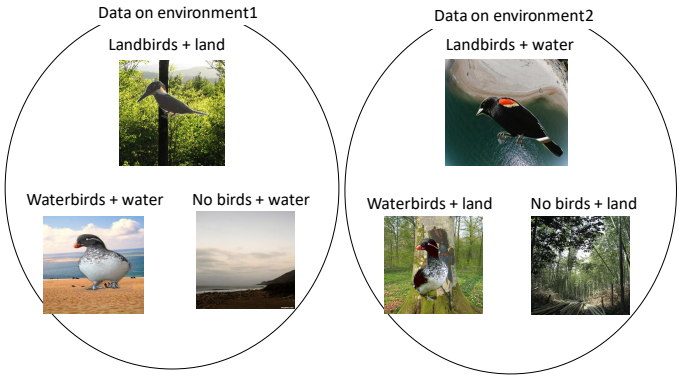

Figure A.1: Visualization of Bird Recognition Problem

Table A.7 shows test accuracies with 2162 random samples for e_1 and e_2. *Oracle* shows a result of ERM with samples from both e_1 and e_2 given. TDV selects λ which yields the highest performance on e_2. Best scores are **bolded**. We can see that the proposed framework together with CV methods succeeded in capturing the predictor invariant to the change of background, while the other methods failed. ERM and FT

	Test Acc. on e_1	Test Acc. on e_2
Oracle	.875 (.018)	
ERM1	.902 (.008)	.317 (.044)
ERM2	.904(.112)	.465(.008)
FT	.909 (.012)	.364 (.028)
FE	.767 (.024)	.052 (.013)
Ours +Our CV I	.897 (.020)	**.727 (.062)**
Ours +Our CV II	.897 (.020)	**.727 (.062)**
Ours +Tr-CV	.919 (.006)	.651 (.031)
Ours +LOD CV	.338 (.048)	.334 (.029)
Ours +TDV	.886 (.035)	.782 (.020)

Table A.7: Average Test Accuracies and SEs of Bird Recognition Problem (5 runs).

show much higher accuracy for e_1 than Oracle and worst results for e_2, which implies that these methods learn spurious correlation in $\mathcal{D}^{e^*}$.

A.4 Additional Experiment: ImageNet

In the main body, only test accuracies on e_2 are shown. The result adding test accuracies on the training domain e_1 are as follows:

ImageNet: $\mathcal{Y} = [3], \mathcal{Z} = [2]$.

	Test Acc. on e_1	Test Acc. on e_2
random guess	.333	
Oracle	.743 (.018)	
ERM1	.750 (.016)	.417 (.016)
ERM2	.713 (.009)	.606 (.014)
FT	.793 (.018)	.463 (.030)
FF	.439 (.002)	.482 (.117)
DSAN	.288 (.012)	.278 (.004)
Ours + CV I	.843 (.024)	.652 (.028)
Ours + CV II	.852 (.009)	**.666 (.027)**
Ours + Tr-CV	.873 (.009)	.641 (.033)
Ours + LOD CV	.857 (.012)	.525 (.028)
Ours + TDV	.857 (.012)	.673 (.035)

ImageNet: $\mathcal{Y} = [7], \mathcal{Z} = [2]$.

	Test Acc. on e_1	Test Acc. on e_2
random guess	.143	
Oracle	.749 (.008)	
ERM1	.740 (.017)	.507 (.020)
ERM2	.683(.006)	.535(.005)
FT	.626 (.028)	.409 (.020)
FF	.191 (.004)	.226 (.046)
DSAN	.184 (.012)	.293 (.008)
Ours + CV I	.853 (.006)	**.622 (.011)**
Ours + CV II	.853 (.006)	**.622 (.011)**
Ours + Tr-CV	.850 (.004)	.612 (.012)
Ours + LOD CV	.825 (.017)	.572 (.022)
Ours + TDV	.837 (.019)	.634 (.003)

ImageNet: $\mathcal{Y} = [17], \mathcal{Z} = [2]$.

	Test Acc. on e_1	Test Acc. on e_2
random guess	.059	
Oracle	.708 (.010)	
ERM1	.577 (.003)	.357 (.020)
ERM2	.610(.015)	.450(.018)
FT	.545 (.009)	.361 (.011)
FF	.201 (.004)	.162 (.008)
DSAN	.058 (.008)	.060 (.007)
Ours + CV I	.776 (.006)	**.556 (.004)**
Ours + CV II	.776 (.006)	**.556 (.004)**
Ours + Tr-CV	.767 (.005)	.544 (.013)
Ours + LOD CV	.742 (.027)	.527 (.019)
Ours + TDV	.776 (.006)	.556 (.004)

Appendix B

Experimental Details

B.1 Detail of ImageNet Experiment Dataset

In the ImageNet experiment in Section 6.3, $\mathcal{Y}$ is set as follows:

- $\mathcal{Y} = [3]$:{**bird**, **turtle**, snake}
- $\mathcal{Y} = [7]$: {**bird**, **turtle**, snake, cat, food, **vehicle**, **building**},
- $\mathcal{Y} = [17]$: {**bird**, **turtle**, snake, cat, **dog**, monkey, spider, butterfly, **food**, **vehicle**, **building**, **shoes**, hat, instrument, tellephone, **bottle**, chair}.

Images of **bolded** labels are composed of different species among e_1 and e_2. Explicitly, dataset are composed as follows:

$$\mathcal{Y} = [3]$$

label	e_1	e_2
bird	ruffed grouse, indigo bunting	albatross, water ouzel
turtle	loggerhead, leathback	box turtle, mud turtle
snake	thunder snake, garther snake, ringneck. snake	

$\mathcal{Y} = [7]$

label	e_1	e_2
bird	ruffed grouse, indigo bunting	albatross, water ouzel
turtle	loggerhead, leathback	box turtle, mud turtle
snake	thunder snake, garther snake, ringneck. snake	
cat	persian cat, siamese cat, egyptian cat	
food	cucumber, strawberry, pizza	
vechicle	submarine, container ship	golfcart, jeep
building	lighthouse, fountaink	castle, water tower

$\mathcal{Y} = [17]$

label	e_1	e_2
bird	ruffed grouse, indigo bunting	albatross, water ouzel
turtle	loggerhead, leathback	box turtle, mud turtle
snake	thunder snake, garther snake, ringneck. snake	
cat	persian cat, siamese cat, egyptian cat	
dog	eskimo dog, dalmatian	newfoundland, German shepherd
monkey	guenon, colobus, titi	
spider	wolf spider, garden spider, barn spider	
butterfly	ringlet, monarch, cabbage butterfly	
food	pizza, strawberry	cucumber, broccoli
vechicle	submarine, container ship	golfcart, jeep
building	lighthouse, fountaink	castle, water tower
shoes	clog, sandal	running shoe, loafer
hat	pickelhaube, crash helmet, hat with a wide brim	
instrument	acoustic guitar, electric guitar, violin	
tellephone	cellular telephone, dial telephone, pay-phone	
bottle	pill bottle, pop bottle	beer bottle, wine bottle
chair	barber chair, folding chair, rocking chair	

B.2 Model Architectures and Optimization Procedures

Through the experiment in the present thesis, all models of competitors are composed of neural networks where its loss function, activation function, and optimizer are cross entropy, Relu Networks and Adam [Kingma and Ba, 2015]. In the following explanation, NN with its model architecture $a \to h_1 \to \cdots h_k \to h_n \to \mathcal{P}_{[m]}$ means that its input and hidden dimensions are a and $(h_1, ..., h_n)$ respectively, and its output is probability density functions on $[m]$. NN with its model architecture $a \to h_1 \to \cdots h_k \to h_n \to b$ means that its input, hidden and output dimensions are a, $(h_1, ..., h_n)$ and b respectively. All the experiment, we add L^2-reguralized term to our objective function.

We add explanations of previous CV methods. Tr-CV implements cross-validation with using only $\mathcal{D}^*$. In LOD-CV, a model is learnt with excluding one of the $\mathcal{D}^e \subset \mathcal{D}_{ad}$ from $\mathcal{D}_{ad}$, and evaluate its performance by $\mathcal{D}^e$. Changing the role of $e \in \mathcal{E}_{ad}$, and taking their mean, we evaluate final CV-value.

Synthesized Data

We set model architecture of Φ used in our method $2 \to 20 \to 20 \to 1$. We set model architecture of ERM $2 \to 20 \to 20 \to \mathcal{P}_{[3]}$. When we use FT and FF, its model architecture on pre-train phase and retraining phase are $2 \to 20 \to 20 \to \mathcal{P}_{[2]}$ and $2 \to 20 \to 20 \to \mathcal{P}_{[3]}$ respectively. We set running rate and hyperparameters of L^2-regularized term 0.0115 and 0.01 respectively. When we use $DSAN$ [Stojanov et al., 2021], we inherit learning condition in the *Amazon Review dataset* experiment. When training, we use batch learning. We set $K = 10$ of each CV method.

Colored MNIST

We set model architecture of Φ used in our method $2 \to 440 \to 440 \to 440$. We set model architecture of and ERM $2 \to 440 \to 440 \to \mathcal{P}_{[|\mathcal{Y}|]}$. When we use FT and FF, its model architecture on pre-train phase and retraining phase are $2 \to 440 \to 440 \to \mathcal{P}_{[2]}$ and $2 \to 440 \to 440 \to \mathcal{P}_{[|\mathcal{Y}|]}$ respectively. We set running rate and hyperparameter of L^2-regularized term 0.0004 and 0.002 respectively. When we use $DSAN$, we inherit learning condition in the *Amazon Review dataset* experiment. When training, we use batch learning. We set $K = 10$ of our CV method.

ImageNet

We set model architecture of Φ used in our method ResNet50 [He et al., 2016] with changing its output dimension 256. We set model architecture of and ERM ResNet50 [He et al., 2016] with changing its output $\mathcal{P}_{[3]}$. When we use FT and FF, its model architecture on pre-train phase and retraining phase are ResNet50 [He et al., 2016] with changing its output dimension 2 and 3 respectively. We set running rate and hyperparameter of L^2-regularized term 0.00004 and 0.001 respectively. When training, we use minibatch learning with a minibatch size 56. We set $K = 5$ of each CV method.

CV comparison experiment

We set model architecture of Φ used in our method $2 \to 8 \to 8 \to 1$. We set running rate and hyperparameters of L^2-regularized term 0.05 and 0.001 respectively. When training, we use minibatch learning with dividing $\mathcal{D}^*$, $\mathcal{D}^{e^{ad}}$ and $\mathcal{D}^{40}$ into 50 equal parts respectively. We set $K = 10$ of each CV method.

Appendix: Birds recognition

We set model architecture of Φ used in our method ResNet50 [He et al., 2016] with changing its output dimension 256. We set model architecture of ERM ResNet50 [He et al., 2016] with changing its output $\mathcal{P}_{[3]}$. When we use FT and FF, its model architecture on pre-train phase and retraining phase are ResNet50 [He et al., 2016] with changing its output dimension 2 and 3 respectively. We set running rate and hyperparameter of L^2-regularized term 0.00004 and 0.001 respectively. When training, we use minibatch learning with a minibatch size 56. We set $K = 5$ of each CV method.