Is Out-of-distribution Detection Learnable?

Zhen Fang[1], **Yixuan Li**[2], **Jie Lu**[1,*], **Jiahua Dong**[3,4], **Bo Han**[5], **Feng Liu**[1,6*]
[1]Australian Artificial Intelligence Institute, University of Technology Sydney.
[2]Department of Computer Sciences, University of Wisconsin Madison.
[3]State Key Laboratory of Robotics, Shenyang Institute of Automation,
Chinese Academy of Sciences. [4]ETH Zurich, Switzerland.
[5]Department of Computer Science, Hong Kong Baptist University.
[6]School of Mathematics and Statistics, University of Melbourne.
{zhen.fang,jie.lu}@uts.edu.au, sharonli@cs.wisc.edu,
dongjiahua1995@gmail.com,bhanml@comp.hkbu.edu.hk,feng.liu1@unimelb.edu.au

Abstract

Supervised learning aims to train a classifier under the assumption that training and test data are from the same distribution. To ease the above assumption, researchers have studied a more realistic setting: *out-of-distribution* (OOD) detection, where test data may come from classes that are unknown during training (*i.e.*, OOD data). Due to the unavailability and diversity of OOD data, good generalization ability is crucial for effective OOD detection algorithms. To study the generalization of OOD detection, in this paper, we investigate the *probably approximately correct* (PAC) learning theory of OOD detection, which is proposed by researchers as an *open problem*. First, we find a necessary condition for the learnability of OOD detection. Then, using this condition, we prove several impossibility theorems for the learnability of OOD detection under some scenarios. Although the impossibility theorems are frustrating, we find that some conditions of these impossibility theorems may not hold in some practical scenarios. Based on this observation, we next give several necessary and sufficient conditions to characterize the learnability of OOD detection in some practical scenarios. Lastly, we also offer theoretical supports for several representative OOD detection works based on our OOD theory.

1 Introduction

The success of supervised learning is established on an implicit assumption that training and test data share a same distribution, *i.e.*, *in-distribution* (ID) [1, 2, 3, 4]. However, test data distribution in many real-world scenarios may violate the assumption and, instead, contain *out-of-distribution* (OOD) data whose labels have not been seen during the training process [5, 6]. To mitigate the risk of OOD data, researchers have considered a more practical learning scenario: OOD detection which determines whether an input is ID/OOD, while classifying the ID data into respective classes. OOD detection has shown great potential to ensure the reliable deployment of machine learning models in the real world. A rich line of algorithms have been developed to empirically address the OOD detection problem [6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]. However, very few works study theory of OOD detection, which hinders the rigorous path forward for the field. This paper aims to bridge the gap.

In this paper, we provide a theoretical framework to understand the learnability of the OOD detection problem. We investigate the probably approximately correct (PAC) learning theory of OOD detection, which is posed as an open problem to date. Unlike the classical PAC learning theory in a supervised setting, our problem setting is fundamentally challenging due to the *absence of OOD data* in training.

*Corresponding author

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

In many real-world scenarios, OOD data can be diverse and priori-unknown. Given this, we study whether there exists an algorithm that can be used to detect various OOD data instead of merely some specified OOD data. Such is the significance of studying the learning theory for OOD detection [4]. This motivates our question: *is OOD detection PAC learnable? i.e., is there the PAC learning theory to guarantee the generalization ability of OOD detection?*

To investigate the learning theory, we mainly focus on two basic spaces: domain space and hypothesis space. The domain space is a space consisting of some distributions, and the hypothesis space is a space consisting of some classifiers. Existing agnostic PAC theories in supervised learning [21, 22] are distribution-free, *i.e.*, the domain space consists of all domains. Yet, in Theorem 4, we shows that the learning theory of OOD detection is not distribution-free. In fact, we discover that OOD detection is learnable only if the domain space and the hypothesis space satisfy some special conditions, *e.g.,* Conditions 1 and 3. Notably, there are many conditions and theorems in existing learning theories and many OOD detection algorithms in the literature. Thus, it is very difficult to analyze the relation between these theories and algorithms, and explore useful conditions to ensure the learnability of OOD detection, especially when we have to explore them *from the scratch*. Thus, the main aim of our paper is to study these essential conditions. From these essential conditions, we can know *when* OOD detection can be successful in practical scenarios. We restate our question and goal in following:

> *Given hypothesis spaces and several representative domain spaces, what are the conditions to ensure the learnability of OOD detection? If possible, we hope that these conditions are necessary and sufficient in some scenarios.*

Main Results. We investigate the learnability of OOD detection starting from the largest space—the total space, and give a necessary condition (Condition 1) for the learnability. However, we find that the overlap between ID and OOD data may result in that the necessary condition does not hold. Therefore, we give an impossibility theorem to demonstrate that OOD detection fails in the total space (Theorem 4). Next, we study OOD detection in the separate space, where there are no overlaps between the ID and OOD data. Unfortunately, there still exists impossibility theorem (Theorem 5), which demonstrates that OOD detection is not learnable in the separate space under some conditions.

Although the impossibility theorems obtained in the separate space are frustrating, we find that some conditions of these impossibility theorems may not hold in some practical scenarios. Based on this observation, we give several necessary and sufficient conditions to characterize the learnability of OOD detection in the separate space (Theorems 6 and 10). Especially, when our model is based on *fully-connected neural network* (FCNN), OOD detection is learnable in the separate space if and only if the feature space is finite. Furthermore, we investigate the learnability of OOD detection in other more practical domain spaces, *e.g.*, the finite-ID-distribution space (Theorem 8) and the density-based space (Theorem 9). By studying the finite-ID-distribution space, we discover a compatibility condition (Condition 3) that is a necessary and sufficient condition for this space. Next, we further investigate the compatibility condition in the density-based space, and find that such condition is also the necessary and sufficient condition in some practical scenarios (Theorem 11).

Implications and Impacts of Theory. Our study is not of purely theoretical interest; it has also practical impacts. First, when we design OOD detection algorithms, we normally only have finite ID datasets, corresponding to the finite-ID-distribution space. In this case, Theorem 8 gives the necessary and sufficient condition to the success of OOD detection. Second, our theory provides theoretical support (Theorems 10 and 11) for several representative OOD detection works [7, 8, 23]. Third, our theory shows that OOD detection is learnable in image-based scenarios when ID images have clearly different semantic labels and styles (*far-OOD*) from OOD images. Fourth, we should not expect a universally working algorithm. It is necessary to design different algorithms in different scenarios.

2 Learning Setups

We start by introducing the necessary concepts and notations for our theoretical framework. Given a feature space $\mathcal{X} \subset \mathbb{R}^d$ and a label space $\mathcal{Y} := \{1, ..., K\}$, we have an ID joint distribution $D_{X_{\mathrm{I}}Y_{\mathrm{I}}}$ over $\mathcal{X} \times \mathcal{Y}$, where $X_{\mathrm{I}} \in \mathcal{X}$ and $Y_{\mathrm{I}} \in \mathcal{Y}$ are random variables. We also have an OOD joint distribution $D_{X_{\mathrm{O}}Y_{\mathrm{O}}}$, where X_{O} is a random variable from $\mathcal{X}$, but Y_{O} is a random variable whose outputs do not belong to $\mathcal{Y}$. During testing, we will meet a mixture of ID and OOD joint distributions: $D_{XY} := (1 - \pi^{\mathrm{out}})D_{X_{\mathrm{I}}Y_{\mathrm{I}}} + \pi^{\mathrm{out}}D_{X_{\mathrm{O}}Y_{\mathrm{O}}}$, and can only observe the marginal distribution $D_X := (1 - \pi^{\mathrm{out}})D_{X_{\mathrm{I}}} + \pi^{\mathrm{out}}D_{X_{\mathrm{O}}}$, where the constant $\pi^{\mathrm{out}} \in [0, 1)$ is an unknown class-prior probability.

Problem 1 (OOD Detection [4]). *Given an ID joint distribution $D_{X_{\mathrm{I}}Y_{\mathrm{I}}}$ and a training data $S := \{(\mathbf{x}^1, y^1), ..., (\mathbf{x}^n, y^n)\}$ drawn independent and identically distributed from $D_{X_{\mathrm{I}}Y_{\mathrm{I}}}$, the aim of OOD detection is to train a classifier f by using the training data S such that, for any test data $\mathbf{x}$ drawn from the mixed marginal distribution D_X: 1) if $\mathbf{x}$ is an observation from $D_{X_{\mathrm{I}}}$, f can classify $\mathbf{x}$ into correct ID classes; and 2) if $\mathbf{x}$ is an observation from $D_{X_{\mathrm{O}}}$, f can detect $\mathbf{x}$ as OOD data.*

According to the survey [4], when $K > 1$, OOD detection is also known as the open-set recognition or open-set learning [24, 25]; and when $K = 1$, OOD detection reduces to one-class novelty detection and semantic anomaly detection [26, 27, 28].

OOD Label and Domain Space. Based on Problem 1, we know it is not necessary to classify OOD data into the correct OOD classes. Without loss of generality, let all OOD data be allocated to one big OOD class, *i.e.*, $Y_{\mathrm{O}} = K + 1$ [24, 29]. To investigate the PAC learnability of OOD detection, we define a domain space $\mathscr{D}_{XY}$, which is a set consisting of some joint distributions D_{XY} mixed by some ID joint distributions and some OOD joint distributions. In this paper, the joint distribution D_{XY} mixed by ID joint distribution $D_{X_{\mathrm{I}}Y_{\mathrm{I}}}$ and OOD joint distribution $D_{X_{\mathrm{O}}Y_{\mathrm{O}}}$ is called ***domain***.

Hypothesis Spaces and Scoring Function Spaces. A hypothesis space $\mathcal{H}$ is a subset of function space, *i.e.*, $\mathcal{H} \subset \{h : \mathcal{X} \to \mathcal{Y} \cup \{K+1\}\}$. We set $\mathcal{H}^{\mathrm{in}} \subset \{h : \mathcal{X} \to \mathcal{Y}\}$ to the ID hypothesis space. We also define $\mathcal{H}^{\mathrm{b}} \subset \{h : \mathcal{X} \to \{1, 2\}\}$ as the hypothesis space for binary classification, where 1 represents the ID data, and 2 represents the OOD data. The function h is called the hypothesis function. A scoring function space is a subset of function space, *i.e.*, $\mathcal{F}_l \subset \{\mathbf{f} : \mathcal{X} \to \mathbb{R}^l\}$, where l is the output's dimension of the vector-valued function $\mathbf{f}$. The function $\mathbf{f}$ is called the scoring function.

Loss and Risks. Let $\mathcal{Y}_{\mathrm{all}} = \mathcal{Y} \cup \{K+1\}$. Given a loss function ℓ[2] $: \mathcal{Y}_{\mathrm{all}} \times \mathcal{Y}_{\mathrm{all}} \to \mathbb{R}_{\geq 0}$ satisfying that $\ell(y_1, y_2) = 0$ if and only if $y_1 = y_2$, and any $h \in \mathcal{H}$, then the *risk* with respect to D_{XY} is

$$R_D(h) := \mathbb{E}_{(\mathbf{x},y)\sim D_{XY}} \ell(h(\mathbf{x}), y). \tag{1}$$

The α-risk $R_D^{\alpha}(h) := (1-\alpha)R_D^{\mathrm{in}}(h) + \alpha R_D^{\mathrm{out}}(h), \forall \alpha \in [0,1]$, where the risks $R_D^{\mathrm{in}}(h)$, $R_D^{\mathrm{out}}(h)$ are

$$R_D^{\mathrm{in}}(h) := \mathbb{E}_{(\mathbf{x},y)\sim D_{X_{\mathrm{I}}Y_{\mathrm{I}}}} \ell(h(\mathbf{x}), y), \qquad R_D^{\mathrm{out}}(h) := \mathbb{E}_{\mathbf{x}\sim D_{X_{\mathrm{O}}}} \ell(h(\mathbf{x}), K+1).$$

Learnability. We aim to select a hypothesis function $h \in \mathcal{H}$ with approximately minimal risk, based on finite data. Generally, we expect the approximation to get better, with the increase in sample size. Algorithms achieving this are said to be consistent. Formally, we introduce the following definition:

Definition 1 (Learnability of OOD Detection). *Given a domain space $\mathscr{D}_{XY}$ and a hypothesis space $\mathcal{H} \subset \{h : \mathcal{X} \to \mathcal{Y}_{\mathrm{all}}\}$, we say OOD detection is* ***learnable*** *in $\mathscr{D}_{XY}$ for $\mathcal{H}$, if there exists an algorithm $\mathbf{A}$[3] $: \cup_{n=1}^{+\infty}(\mathcal{X} \times \mathcal{Y})^n \to \mathcal{H}$ and a monotonically decreasing sequence $\epsilon_{\mathrm{cons}}(n)$, such that $\epsilon_{\mathrm{cons}}(n) \to 0$, as $n \to +\infty$, and for any domain $D_{XY} \in \mathscr{D}_{XY}$,*

$$\mathbb{E}_{S\sim D_{X_{\mathrm{I}}Y_{\mathrm{I}}}^n}\left[R_D(\mathbf{A}(S)) - \inf_{h\in\mathcal{H}} R_D(h)\right] \leq \epsilon_{\mathrm{cons}}(n), \tag{2}$$

An algorithm $\mathbf{A}$ for which this holds is said to be consistent with respect to $\mathscr{D}_{XY}$.

Definition 1 is a natural extension of agnostic PAC learnability of supervised learning [30]. If for any $D_{XY} \in \mathscr{D}_{XY}$, $\pi^{\mathrm{out}} = 0$, then Definition 2 is the agnostic PAC learnability of supervised learning. Although the expression of Definition 1 is different from the normal definition of agnostic PAC learning in [21], one can easily prove that they are equivalent when ℓ is bounded, see Appendix D.3.

Since OOD data are unavailable, it is impossible to obtain information about the class-prior probability π^{out}. Furthermore, in the real world, it is possible that π^{out} can be any value in $[0, 1)$. Therefore, the imbalance issue between ID and OOD distributions, and the priori-unknown issue (*i.e.*, π^{out} is unknown) are the core challenges. To ease these challenges, researchers use AUROC, AUPR and FPR95 to estimate the performance of OOD detection [18, 31, 32, 33, 34, 35]. It seems that there is a gap between Definition 1 and existing works. To eliminate this gap, we revise Eq. (2) as follows:

$$\mathbb{E}_{S\sim D_{X_{\mathrm{I}}Y_{\mathrm{I}}}^n}\left[R_D^{\alpha}(\mathbf{A}(S)) - \inf_{h\in\mathcal{H}} R_D^{\alpha}(h)\right] \leq \epsilon_{\mathrm{cons}}(n), \ \forall \alpha \in [0,1]. \tag{3}$$

If an algorithm $\mathbf{A}$ satisfies Eq. (3), then the imbalance issue and the prior-unknown issue disappear. That is, $\mathbf{A}$ can simultaneously classify the ID data and detect the OOD data well. Based on the above discussion, we define the strong learnability of OOD detection as follows:

[2] Note that $\mathcal{Y}_{\mathrm{all}} \times \mathcal{Y}_{\mathrm{all}}$ is a finite set, therefore, ℓ is bounded.

[3] Similar to [30], in this paper, we regard an algorithm as a mapping from $\cup_{n=1}^{+\infty}(\mathcal{X} \times \mathcal{Y})^n$ to $\mathcal{H}$.

Definition 2 (Strong Learnability of OOD Detection)**.** *Given a domain space $\mathscr{D}_{XY}$ and a hypothesis space $\mathcal{H} \subset \{h : \mathcal{X} \to \mathcal{Y}_{\text{all}}\}$, we say OOD detection is* ***strongly learnable*** *in $\mathscr{D}_{XY}$ for $\mathcal{H}$, if there exists an algorithm $\mathbf{A} : \cup_{n=1}^{+\infty}(\mathcal{X} \times \mathcal{Y})^n \to \mathcal{H}$ and a monotonically decreasing sequence $\epsilon_{\text{cons}}(n)$, such that $\epsilon_{\text{cons}}(n) \to 0$, as $n \to +\infty$, and for any domain $D_{XY} \in \mathscr{D}_{XY}$,*

$$\mathbb{E}_{S\sim D^n_{X_{\text{I}}Y_{\text{I}}}}\left[R^{\alpha}_{D}(\mathbf{A}(S)) - \inf_{h\in\mathcal{H}} R^{\alpha}_{D}(h)\right] \leq \epsilon_{\text{cons}}(n),\ \forall\alpha \in [0,1].$$

In Theorem 1, we have shown that the strong learnability of OOD detection is equivalent to the learnability of OOD detection, if the domain space $\mathscr{D}_{XY}$ is a *prior-unknown space* (see Definition 3). In this paper, we mainly discuss the learnability in the prior-unknown space. Therefore, *when we mention that OOD detection is learnable, we also mean that OOD detection is strongly learnable*.

Goal of Theory. Note that the agnostic PAC learnability of supervised learning is distribution-free, *i.e.*, the domain space $\mathscr{D}_{XY}$ consists of all domains. However, due to the absence of OOD data during the training process [8, 14, 24], it is obvious that the learnability of OOD detection is not distribution-free (*i.e.*, Theorem 4). In fact, we discover that the learnability of OOD detection is deeply correlated with the relationship between the domain space $\mathscr{D}_{XY}$ and the hypothesis space $\mathcal{H}$. That is, OOD detection is learnable only when the domain space $\mathscr{D}_{XY}$ and the hypothesis space $\mathcal{H}$ satisfy some special conditions, *e.g.,* Condition 1 and Condition 3. We present our goal as follows:

> ***Goal:*** *given a hypothesis space $\mathcal{H}$ and several representative domain spaces $\mathscr{D}_{XY}$, what are the* ***conditions*** *to ensure the learnability of OOD detection? Furthermore, if possible, we hope that these conditions are* ***necessary and sufficient*** *in some scenarios.*

Therefore, compared to the agnostic PAC learnability of supervised learning, our theory doesn't focus on the distribution-free case, but focuses on discovering essential conditions to guarantee the learnability of OOD detection in several representative and practical domain spaces $\mathscr{D}_{XY}$. By these essential conditions, we can know *when* OOD detection can be successful in real applications.

3 Learning in Priori-unknown Spaces

We first investigate a special space, called prior-unknown space. In such space, Definition 1 and Definition 2 are equivalent. Furthermore, we also prove that if OOD detection is strongly learnable in a space $\mathscr{D}_{XY}$, then one can discover a larger domain space, which is prior-unknown, to ensure the learnability of OOD detection. These results imply that it is enough to consider our theory in the prior-unknown spaces. The prior-unknown space is introduced as follows:

Definition 3. *Given a domain space $\mathscr{D}_{XY}$, we say $\mathscr{D}_{XY}$ is a priori-unknown space, if for any domain $D_{XY} \in \mathscr{D}_{XY}$ and any $\alpha \in [0,1)$, we have $D^{\alpha}_{XY} := (1-\alpha)D_{X_{\text{I}}Y_{\text{I}}} + \alpha D_{X_{\text{O}}Y_{\text{O}}} \in \mathscr{D}_{XY}$.*

Theorem 1. *Given domain spaces $\mathscr{D}_{XY}$ and $\mathscr{D}'_{XY} = \{D^{\alpha}_{XY} : \forall D_{XY} \in \mathscr{D}_{XY}, \forall\alpha \in [0,1)\}$, then*
1) $\mathscr{D}'_{XY}$ is a priori-unknown space and $\mathscr{D}_{XY} \subset \mathscr{D}'_{XY}$;
2) if $\mathscr{D}_{XY}$ is a priori-unknown space, then Definition 1 and Definition 2 are ***equivalent****;*
3) OOD detection is strongly learnable in $\mathscr{D}_{XY}$ ***if and only if*** *OOD detection is learnable in $\mathscr{D}'_{XY}$.*

The second result of Theorem 1 bridges the learnability and strong learnability, which implies that if an algorithm $\mathbf{A}$ is consistent with respect to a prior-unknown space, then this algorithm $\mathbf{A}$ can address the imbalance issue between ID and OOD distributions, and the priori-unknown issue well. Based on Theorem 1, we focus on our theory in the prior-unknown spaces. Furthermore, to demystify the learnability of OOD detection, we introduce five representative priori-unknown spaces:

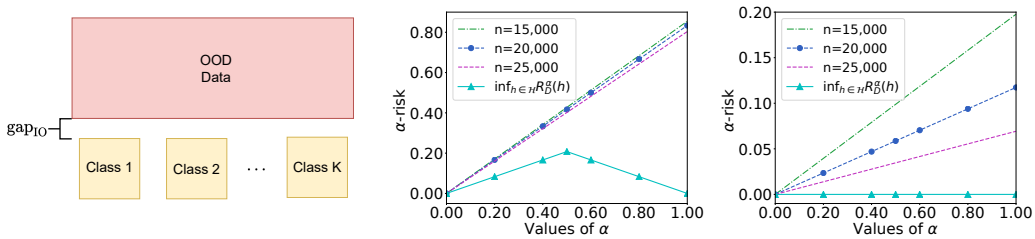
- Single-distribution space $\mathscr{D}^{D_{XY}}_{XY}$. For a domain D_{XY}, $\mathscr{D}^{D_{XY}}_{XY} := \{D^{\alpha}_{XY} : \forall\alpha \in [0,1)\}$.

- Total space $\mathscr{D}^{\text{all}}_{XY}$, which consists of all domains.

- Separate space $\mathscr{D}^{s}_{XY}$, which consists of all domains that satisfy the separate condition, that is for any $D_{XY} \in \mathscr{D}^{s}_{XY}$, $\text{supp}D_{X_{\text{O}}} \cap \text{supp}D_{X_{\text{I}}} = \emptyset$, where supp means the support set.

- Finite-ID-distribution space $\mathscr{D}^{F}_{XY}$, which is a prior-unknown space satisfying that the number of distinct ID joint distributions $D_{X_{\text{I}}Y_{\text{I}}}$ in $\mathscr{D}^{F}_{XY}$ is finite, *i.e.*, $|\{D_{X_{\text{I}}Y_{\text{I}}} : \forall D_{XY} \in \mathscr{D}^{F}_{XY}\}| < +\infty$.

- Density-based space $\mathscr{D}^{\mu,b}_{XY}$, which is a prior-unknown space consisting of some domains satisfying that: for any D_{XY}, there exists a density function f with $1/b \leq f \leq b$ in $\text{supp}\mu$ and $0.5 * D_{X_{\text{I}}} +$



(a) ID and OOD Distributions (b) Overlap Exists (c) No Overlap

Figure 1: Illustration of $\inf_{h\in\mathcal{H}} R_D^\alpha(h)$ (solid lines with triangle marks) and the estimated $\mathbb{E}_{S\sim D_{\mathrm{in}}^n} R_D^\alpha(\mathbf{A}(S))$ (dash lines) with $\alpha \in [0, 1)$ in different scenarios, where $D_{\mathrm{in}} = D_{X_{\mathrm{I}}Y_{\mathrm{I}}}$ and the algorithm $\mathbf{A}$ is the free-energy OOD detection method [23]. Subfigure (a) shows the ID and OOD distributions. In (a), $\mathrm{gap}_{\mathrm{IO}}$ represents the distance between the support sets of ID and OOD distributions. In (b), since there is an overlap between ID and OOD data, the solid line is a ployline. In (c), since there is no overlap between ID and OOD data, we can check that $\inf_{h\in\mathcal{H}} R_D^\alpha(h)$ forms a straight line (the solid line). However, since dash lines are always straight lines, two observations can be obtained from (b) and (c): 1) dash lines cannot approximate the solid ployline in (b), which implies the unlearnability of OOD detection; and 2) the solid line in (c) is a straight line and may be approximated by the dash lines in (c). The above observations motivate us to propose Condition 1.

$0.5 * D_{X_{\mathrm{O}}} = \int f \mathrm{d}\mu$, where μ is a measure defined over $\mathcal{X}$. Note that if μ is discrete, then D_X is a discrete distribution; and if μ is the Lebesgue measure, then D_X is a continuous distribution.

The above representative spaces widely exist in real applications. For example, 1) if the images from different semantic labels with different styles are clearly different, then those images can form a distribution belonging to a separate space $\mathscr{D}_{XY}^{s}$; and 2) when designing an algorithm, we only have finite ID datasets, *e.g.*, CIFAR-10, MNIST, SVHN, and ImageNet, to build a model. Then, finite-ID-distribution space $\mathscr{D}_{XY}^{F}$ can handle this real scenario. Note that the single-distribution space is a special case of the finite-ID-distribution space. In this paper, we mainly discuss these five spaces.

4 Impossibility Theorems for OOD Detection

In this section, we first give a necessary condition for the learnability of OOD detection. Then, we show this necessary condition does not hold in the total space $\mathscr{D}_{XY}^{\mathrm{all}}$ and the separate space $\mathscr{D}_{XY}^{s}$.

Necessary Condition. We find a necessary condition for the learnability of OOD detection, *i.e.*, Condition 1, motivated by the experiments in Figure 1. Details of Figure 1 can be found in Appendix C.2.

Condition 1 (Linear Condition)**.** *For any $D_{XY} \in \mathscr{D}_{XY}$ and any $\alpha \in [0, 1)$,*

$$\inf_{h\in\mathcal{H}} R_D^\alpha(h) = (1-\alpha) \inf_{h\in\mathcal{H}} R_D^{\mathrm{in}}(h) + \alpha \inf_{h\in\mathcal{H}} R_D^{\mathrm{out}}(h).$$

To reveal the importance of Condition 1, Theorem 2 shows that Condition 1 is a *necessary and sufficient* condition for the learnability of OOD detection if the $\mathscr{D}_{XY}$ is the single-distribution space.

Theorem 2. *Given a hypothesis space $\mathcal{H}$ and a domain D_{XY}, OOD detection is learnable in the single-distribution space $\mathscr{D}_{XY}^{D_{XY}}$ for $\mathcal{H}$* ***if and only if*** *linear condition (i.e., Condition 1) holds.*

Theorem 2 implies that Condition 1 is important for the learnability of OOD detection. Due to the simplicity of single-distribution space, Theorem 2 implies that Condition 1 is the necessary condition for the learnability of OOD detection in the prior-unknown space, see Lemma 1 in Appendix F.

Impossibility Theorems. Here, we first study whether Condition 1 holds in the total space $\mathscr{D}_{XY}^{\mathrm{all}}$. If Condition 1 does not hold, then OOD detection is not learnable. Theorem 3 shows that Condition 1 is not always satisfied, especially, when there is an overlap between the ID and OOD distributions:

Definition 4 (Overlap Between ID and OOD)**.** *We say a domain D_{XY} has overlap between ID and OOD distributions, if there is a σ-finite measure $\tilde{\mu}$ such that D_X is absolutely continuous with respect to $\tilde{\mu}$, and $\tilde{\mu}(A_{\mathrm{overlap}}) > 0$, where $A_{\mathrm{overlap}} = \{\mathbf{x} \in \mathcal{X} : f_{\mathrm{I}}(\mathbf{x}) > 0 \text{ and } f_{\mathrm{O}}(\mathbf{x}) > 0\}$. Here f_{I} and f_{O} are the representers of $D_{X_{\mathrm{I}}}$ and $D_{X_{\mathrm{O}}}$ in Radon–Nikodym Theorem [36],*

$$D_{X_{\mathrm{I}}} = \int f_{\mathrm{I}} \mathrm{d}\tilde{\mu}, \quad D_{X_{\mathrm{O}}} = \int f_{\mathrm{O}} \mathrm{d}\tilde{\mu}.$$

Theorem 3. *Given a hypothesis space $\mathcal{H}$ and a prior-unknown space $\mathscr{D}_{XY}$, if there is $D_{XY} \in \mathscr{D}_{XY}$, which has overlap between ID and OOD, and $\inf_{h\in\mathcal{H}} R_D^{\text{in}}(h) = 0$ and $\inf_{h\in\mathcal{H}} R_D^{\text{out}}(h) = 0$, then Condition 1 does not hold. Therefore, OOD detection is not learnable in $\mathscr{D}_{XY}$ for $\mathcal{H}$.*

Theorem 3 clearly shows that under proper conditions, Condition 1 does not hold, if there exists a domain whose ID and OOD distributions have overlap. By Theorem 3, we can obtain that the OOD detection is not learnable in the total space $\mathscr{D}_{XY}^{\text{all}}$ for any non-trivial hypothesis space $\mathcal{H}$.

Theorem 4 (Impossibility Theorem for Total Space)**.** *OOD detection is not learnable in the total space $\mathscr{D}_{XY}^{\text{all}}$ for $\mathcal{H}$, if $|\phi \circ \mathcal{H}| > 1$, where ϕ maps ID labels to 1 and maps OOD labels to 2.*

Since the overlaps between ID and OOD distributions may cause that Condition 1 does not hold, we then consider studying the learnability of OOD detection in the separate space $\mathscr{D}_{XY}^{s}$, where there are no overlaps between the ID and OOD distributions. However, Theorem 5 shows that even if we consider the separate space, the OOD detection is still not learnable in some scenarios. Before introducing the impossibility theorem for separate space, *i.e.*, Theorem 5, we need a mild assumption:

Assumption 1 (Separate Space for OOD)**.** *A hypothesis space $\mathcal{H}$ is separate for OOD data, if for each data point $\mathbf{x} \in \mathcal{X}$, there exists at least one hypothesis function $h_{\mathbf{x}} \in \mathcal{H}$ such that $h_{\mathbf{x}}(\mathbf{x}) = K + 1$.*

Assumption 1 means that every data point $\mathbf{x}$ has the possibility to be detected as OOD data. Assumption 1 is mild and can be satisfied by many hypothesis spaces, *e.g.*, the FCNN-based hypothesis space (Proposition 1 in Appendix K), score-based hypothesis space (Proposition 2 in Appendix K) and universal kernel space. Next, we use *Vapnik–Chervonenkis* (VC) dimension [22] to measure the size of hypothesis space, and study the learnability of OOD detection in $\mathscr{D}_{XY}^{s}$ based on the VC dimension.

Theorem 5 (Impossibility Theorem for Separate Space)**.** *If Assumption 1 holds, $\text{VCdim}(\phi \circ \mathcal{H}) < +\infty$ and $\sup_{h\in\mathcal{H}} |\{\mathbf{x} \in \mathcal{X} : h(\mathbf{x}) \in \mathcal{Y}\}| = +\infty$, then OOD detection is not learnable in separate space $\mathscr{D}_{XY}^{s}$ for $\mathcal{H}$, where ϕ maps ID labels to 1 and maps OOD labels to 2.*

The finite VC dimension normally implies the learnability of supervised learning. However, in our results, the finite VC dimension cannot guarantee the learnability of OOD detection in the separate space, which reveals the difficulty of the OOD detection. Although the above impossibility theorems are frustrating, there is still room to discuss the conditions in Theorem 5, and to find out the proper conditions for ensuring the learnability of OOD detection in the separate space (see Sections 5 and 6).

5 When OOD Detection Can Be Successful

Here, we discuss when the OOD detection can be learnable in the separate space $\mathscr{D}_{XY}^{s}$, finite-ID-distribution space $\mathscr{D}_{XY}^{F}$ and density-based space $\mathscr{D}_{XY}^{\mu,b}$. We first study the separate space $\mathscr{D}_{XY}^{s}$.

OOD Detection in the Separate Space. Theorem 5 has indicated that $\text{VCdim}(\phi \circ \mathcal{H}) = +\infty$ or $\sup_{h\in\mathcal{H}} |\{\mathbf{x} \in \mathcal{X} : h(\mathbf{x}) \in \mathcal{Y}\}| < +\infty$ is necessary to ensure the learnability of OOD detection in $\mathscr{D}_{XY}^{s}$ if Assumption 1 holds. However, generally, hypothesis spaces generated by feed-forward neural networks with proper activation functions have finite VC dimension [37, 38]. Therefore, we study the learnability of OOD detection in the case that $|\mathcal{X}| < +\infty$, which implies that $\sup_{h\in\mathcal{H}} |\{\mathbf{x} \in \mathcal{X} : h(\mathbf{x}) \in \mathcal{Y}\}| < +\infty$. Additionally, Theorem 10 also implies that $|\mathcal{X}| < +\infty$ is the necessary and sufficient condition for the learnability of OOD detection in separate space, when the hypothesis space is generated by FCNN. Hence, $|\mathcal{X}| < +\infty$ may be necessary in the space $\mathscr{D}_{XY}^{s}$.

For simplicity, we first discuss the case that $K = 1$, *i.e.*, the one-class novelty detection. We show the necessary and sufficient condition for the learnability of OOD detection in $\mathscr{D}_{XY}^{s}$, when $|\mathcal{X}| < +\infty$.

Theorem 6. *Let $K = 1$ and $|\mathcal{X}| < +\infty$. Suppose that Assumption 1 holds and the constant function $h^{\text{in}} := 1 \in \mathcal{H}$. Then OOD detection is learnable in $\mathscr{D}_{XY}^{s}$ for $\mathcal{H}$* ***if and only if*** *$\mathcal{H}_{\text{all}} - \{h^{\text{out}}\} \subset \mathcal{H}$, where $\mathcal{H}_{\text{all}}$ is the hypothesis space consisting of all hypothesis functions, and h^{out} is a constant function that $h^{\text{out}} := 2$, here 1 represents ID data and 2 represents OOD data.*

The condition $h^{\text{in}} \in \mathcal{H}$ presented in Theorem 6 is mild. Many practical hypothesis spaces satisfy this condition, *e.g.*, the FCNN-based hypothesis space (Proposition 1 in Appendix K), score-based hypothesis space (Proposition 2 in Appendix K) and universal kernel-based hypothesis space. Theorem 6 implies that if $K = 1$ and OOD detection is learnable in $\mathscr{D}_{XY}^{s}$ for $\mathcal{H}$, then the hypothesis space $\mathcal{H}$

should contain almost all hypothesis functions, implying that if the OOD detection can be learnable in the distribution-agnostic case, then a large-capacity model is necessary.

Next, we extend Theorem 6 to a general case, *i.e.*, $K > 1$. When $K > 1$, we will first use a binary classifier h^b to classify the ID and OOD data. Then, for the ID data identified by h^b, an ID hypothesis function h^{in} will be used to classify them into corresponding ID classes. We state this strategy as follows: given a hypothesis space $\mathcal{H}^{\text{in}}$ for ID distribution and a binary classification hypothesis space $\mathcal{H}^{\text{b}}$ introduced in Section 2, we use $\mathcal{H}^{\text{in}}$ and $\mathcal{H}^{\text{b}}$ to construct an OOD detection's hypothesis space $\mathcal{H}$, which consists of all hypothesis functions h satisfying the following condition: there exist $h^{\text{in}} \in \mathcal{H}^{\text{in}}$ and $h^{\text{b}} \in \mathcal{H}^{b}$ such that for any $\mathbf{x} \in \mathcal{X}$,

$$h(\mathbf{x}) = i, \quad \text{if } h^{\text{in}}(\mathbf{x}) = i \text{ and } h^{\text{b}}(\mathbf{x}) = 1; \text{ otherwise, } h(\mathbf{x}) = K + 1. \tag{4}$$

We use $\mathcal{H}^{\text{in}} \bullet \mathcal{H}^{\text{b}}$ to represent a hypothesis space consisting of all h defined in Eq. (4). In addition, we also need an additional condition for the loss function ℓ. This condition is shown as follows:

Condition 2. *$\ell(y_2, y_1) \leq \ell(K + 1, y_1)$, for any in-distribution labels y_1 and $y_2 \in \mathcal{Y}$.*

Theorem 7. *Let $|\mathcal{X}| < +\infty$ and $\mathcal{H} = \mathcal{H}^{\text{in}} \bullet \mathcal{H}^{\text{b}}$. If $\mathcal{H}_{\text{all}} - \{h^{\text{out}}\} \subset \mathcal{H}^{\text{b}}$ and Condition 2 holds, then OOD detection is learnable in $\mathscr{D}^{s}_{XY}$ for $\mathcal{H}$, where $\mathcal{H}_{\text{all}}$ and h^{out} are defined in Theorem 6.*

OOD Detection in the Finite-ID-Distribution Space. Since researchers can only collect finite ID datasets as the training data in the process of algorithm design, it is worthy to study the learnability of OOD detection in the finite-ID-distribution space $\mathscr{D}^{F}_{XY}$. We first show two necessary concepts below.

Definition 5 (ID Consistency)**.** *Given a domain space $\mathscr{D}_{XY}$, we say any two domains $D_{XY} \in \mathscr{D}_{XY}$ and $D'_{XY} \in \mathscr{D}_{XY}$ are ID consistency, if $D_{X_{\text{I}}Y_{\text{I}}} = D'_{X_{\text{I}}Y_{\text{I}}}$. We use the notation $\sim$ to represent the ID consistency, i.e., $D_{XY} \sim D'_{XY}$ if and only if D_{XY} and D'_{XY} are ID consistency.*

It is easy to check that the ID consistency $\sim$ is an equivalence relation. Therefore, we define the set $[D_{XY}] := \{D'_{XY} \in \mathscr{D}_{XY} : D_{XY} \sim D'_{XY}\}$ as the equivalence class with respect to space $\mathscr{D}_{XY}$.

Condition 3 (Compatibility)**.** *For any equivalence class $[D'_{XY}]$ with respect to $\mathscr{D}_{XY}$ and any $\epsilon > 0$, there exists a hypothesis function $h_\epsilon \in \mathcal{H}$ such that for any domain $D_{XY} \in [D'_{XY}]$,*

$$h_\epsilon \in \{h' \in \mathcal{H} : R^{\text{out}}_D(h') \leq \inf_{h \in \mathcal{H}} R^{\text{out}}_D(h) + \epsilon\} \cap \{h' \in \mathcal{H} : R^{\text{in}}_D(h') \leq \inf_{h \in \mathcal{H}} R^{\text{in}}_D(h) + \epsilon\}.$$

In Appendix F, Lemma 2 has implied that Condition 3 is a general version of Condition 1. Next, Theorem 8 indicates that Condition 3 is the *necessary and sufficient condition* in the space $\mathscr{D}^{F}_{XY}$.

Theorem 8. *Suppose that $\mathcal{X}$ is a bounded set. OOD detection is learnable in the finite-ID-distribution space $\mathscr{D}^{F}_{XY}$ for $\mathcal{H}$* ***if and only if*** *the compatibility condition (i.e., Condition 3) holds. Furthermore, the learning rate $\epsilon_{\text{cons}}(n)$ can attain $O(1/\sqrt{n^{1-\theta}})$, for any $\theta \in (0, 1)$.*

Theorem 8 shows that, in the process of algorithm design, OOD detection cannot be successful without the compatibility condition. Theorem 8 also implies that Condition 3 is essential for the learnability of OOD detection. This motivates us to study whether OOD detection can be successful in more general spaces (*e.g.*, the density-based space), when the compatibility condition holds.

OOD Detection in the Density-based Space. To ensure that Condition 3 holds, we consider a basic assumption in learning theory—*Realizability Assumption* (see Appendix D.2), *i.e.*, for any $D_{XY} \in \mathscr{D}_{XY}$, there exists $h^* \in \mathcal{H}$ such that $R_D(h^*) = 0$. We discover that in the density-based space $\mathscr{D}^{\mu,b}_{XY}$, Realizability Assumption can conclude the compatibility condition (*i.e.*, Condition 3). Based on this observation, we can prove the following theorem:

Theorem 9. *Given a density-based space $\mathscr{D}^{\mu,b}_{XY}$, if $\mu(\mathcal{X}) < +\infty$, the Realizability Assumption holds, then when $\mathcal{H}$ has finite Natarajan dimension [21], OOD detection is learnable in $\mathscr{D}^{\mu,b}_{XY}$ for $\mathcal{H}$. Furthermore, the learning rate $\epsilon_{\text{cons}}(n)$ can attain $O(1/\sqrt{n^{1-\theta}})$, for any $\theta \in (0, 1)$.*

To further investigate the importance and necessary of Realizability Assumption, Theorem 11 has indicated that in some practical scenarios, Realizability Assumption is the necessary and sufficient condition for the learnability of OOD detection in the density-based space. Therefore, Realizability Assumption may be indispensable for the learnability of OOD detection in some practical scenarios.

6 Connecting Theory to Practice

In Section 5, we have shown the successful scenarios where OOD detection problem can be addressed in theory. In this section, we will discuss how the proposed theory is applied to two representative hypothesis spaces—neural-network-based hypothesis spaces and score-based hypothesis spaces.

Fully-connected Neural Networks. Given a sequence $\mathbf{q} = (l_1, l_2, ..., l_g)$, where l_i and g are positive integers and $g > 2$, we use g to represent the ***depth*** of neural network and use l_i to represent the ***width*** of the i-th layer. After the activation function σ is selected[4], we can obtain the architecture of FCNN according to the sequence $\mathbf{q}$. Let $\mathbf{f}_{\mathbf{w},\mathbf{b}}$ be the function generated by FCNN with weights $\mathbf{w}$ and bias $\mathbf{b}$. An FCNN-based scoring function space is defined as: $\mathcal{F}_{\mathbf{q}}^{\sigma} := \{\mathbf{f}_{\mathbf{w},\mathbf{b}} : \forall \text{ weights } \mathbf{w}, \ \forall \text{ bias } \mathbf{b}\}$. In addition, for simplicity, given any two sequences $\mathbf{q} = (l_1, ..., l_g)$ and $\mathbf{q}' = (l'_1, ..., l'_{g'})$, we use the notation $\mathbf{q} \lesssim \mathbf{q}'$ to represent the following equations and inequalities:

$$1)\ g \leq g', l_1 = l'_1, l_g = l'_{g'}; \quad 2)\ l_i \leq l'_i, \ \forall i = 1, ..., g-1; \ \text{ and } \ 3)\ l_{g-1} \leq l'_i, \ \forall i = g, ..., g'-1.$$

In Appendix L, Lemma 10 shows $\mathbf{q} \lesssim \mathbf{q}' \Rightarrow \mathcal{F}_{\mathbf{q}}^{\sigma} \subset \mathcal{F}_{\mathbf{q}'}^{\sigma}$. We use $\lesssim$ to compare the sizes of FCNNs.

FCNN-based Hypothesis Space. Let $l_g = K + 1$. The FCNN-based scoring function space $\mathcal{F}_{\mathbf{q}}^{\sigma}$ can induce an FCNN-based hypothesis space. For any $\mathbf{f}_{\mathbf{w},\mathbf{b}} \in \mathcal{F}_{\mathbf{q}}^{\sigma}$, the induced hypothesis function is:

$$h_{\mathbf{w},\mathbf{b}} := \underset{k \in \{1,...,K+1\}}{\arg\max} \ f^k_{\mathbf{w},\mathbf{b}}, \text{ where } f^k_{\mathbf{w},\mathbf{b}} \text{ is the } k\text{-th coordinate of } \mathbf{f}_{\mathbf{w},\mathbf{b}}.$$

Then, the FCNN-based hypothesis space is defined as $\mathcal{H}_{\mathbf{q}}^{\sigma} := \{h_{\mathbf{w},\mathbf{b}} : \forall \text{ weights } \mathbf{w}, \ \forall \text{ bias } \mathbf{b}\}$.

Score-based Hypothesis Space. Many OOD detection algorithms detect OOD data by using a score-based strategy. That is, given a threshold λ, a scoring function space $\mathcal{F}_l \subset \{\mathbf{f} : \mathcal{X} \to \mathbb{R}^l\}$ and a scoring function $E : \mathcal{F}_l \to \mathbb{R}$, then $\mathbf{x}$ is regarded as ID data if and only if $E(\mathbf{f}(\mathbf{x})) \geq \lambda$. We introduce several representative scoring functions E as follows: for any $\mathbf{f} = [f^1, ..., f^l]^\top \in \mathcal{F}_l$,
• softmax-based function [7] and temperature-scaled function [8]: $\lambda \in (\frac{1}{l}, 1)$ and $T > 0$,

$$E(\mathbf{f}) = \max_{k \in \{1,...,l\}} \frac{\exp(f^k)}{\sum_{c=1}^{l} \exp(f^c)}, \qquad E(\mathbf{f}) = \max_{k \in \{1,...,l\}} \frac{\exp(f^k/T)}{\sum_{c=1}^{l} \exp(f^c/T)}; \tag{5}$$

• energy-based function [23]: $\lambda \in (0, +\infty)$ and $T > 0$,

$$E(\mathbf{f}) = T \log \sum_{c=1}^{l} \exp(f^c/T). \tag{6}$$

Using E, λ and $\mathbf{f} \in \mathcal{F}_{\mathbf{q}}^{\sigma}$, we have a classifier: $h^{\lambda}_{\mathbf{f},E}(\mathbf{x}) = 1$, if $E(\mathbf{f}(\mathbf{x})) \geq \lambda$; otherwise, $h^{\lambda}_{\mathbf{f},E}(\mathbf{x}) = 2$, where 1 represents the ID data and 2 represents the OOD data. Hence, a binary classification hypothesis space $\mathcal{H}^b$, which consists of all $h^{\lambda}_{\mathbf{f},E}$, is generated. We define $\mathcal{H}^{\sigma,\lambda}_{\mathbf{q},E} := \{h^{\lambda}_{\mathbf{f},E} : \forall \mathbf{f} \in \mathcal{F}_{\mathbf{q}}^{\sigma}\}$.

Learnability of OOD Detection in Different Hypothesis Spaces. Next, we present applications of our theory regarding the above two practical and important hypothesis spaces $\mathcal{H}_{\mathbf{q}}^{\sigma}$ and $\mathcal{H}^{\sigma,\lambda}_{\mathbf{q},E}$.

Theorem 10. *Suppose that Condition 2 holds and the hypothesis space $\mathcal{H}$ is FCNN-based or score-based, i.e., $\mathcal{H} = \mathcal{H}_{\mathbf{q}}^{\sigma}$ or $\mathcal{H} = \mathcal{H}^{\text{in}} \bullet \mathcal{H}^{\text{b}}$, where $\mathcal{H}^{\text{in}}$ is an ID hypothesis space, $\mathcal{H}^{\text{b}} = \mathcal{H}^{\sigma,\lambda}_{\mathbf{q},E}$ and $\mathcal{H} = \mathcal{H}^{\text{in}} \bullet \mathcal{H}^{\text{b}}$ is introduced below Eq. (4), here E is introduced in Eqs. (5) or (6). Then*

There is a sequence $\mathbf{q} = (l_1, ..., l_g)$ such that OOD detection is learnable in the separate space $\mathscr{D}^{s}_{XY}$ for $\mathcal{H}$ ***if and only if*** *$|\mathcal{X}| < +\infty$.*

Furthermore, if $|\mathcal{X}| < +\infty$, then there exists a sequence $\mathbf{q} = (l_1, ..., l_g)$ such that for any sequence $\mathbf{q}'$ satisfying that $\mathbf{q} \lesssim \mathbf{q}'$, OOD detection is learnable in $\mathscr{D}^{s}_{XY}$ for $\mathcal{H}$.

Theorem 10 states that 1) when the hypothesis space is FCNN-based or score-based, the finite feature space is the necessary and sufficient condition for the learnability of OOD detection in the separate space; and 2) a larger architecture of FCNN has a greater probability to achieve the learnability of

[4] We consider the *rectified linear unit* (ReLU) function as the default activation function σ, which is defined by $\sigma(x) = \max\{x, 0\}, \forall\, x \in \mathbb{R}$. *We will not repeatedly mention the definition of σ in the rest of our paper.*

OOD detection in the separate space. Note that when we select Eqs. (5) or (6) as the scoring function E, Theorem 10 also shows that the selected scoring functions E can guarantee the learnability of OOD detection, which is a theoretical support for the representative works [8, 23, 7]. Furthermore, Theorem 11 also offers theoretical supports for these works in the density-based space, when $K = 1$.

Theorem 11. *Suppose that each domain D_{XY} in $\mathscr{D}_{XY}^{\mu,b}$ is attainable, i.e., $\arg\min_{h\in\mathcal{H}} R_D(h) \neq \emptyset$ (the finite discrete domains satisfy this). Let $K = 1$ and the hypothesis space $\mathcal{H}$ be score-based ($\mathcal{H} = \mathcal{H}_{\mathbf{q},E}^{\sigma,\lambda}$, where E is in Eqs. (5) or (6)) or FCNN-based ($\mathcal{H} = \mathcal{H}_{\mathbf{q}}^{\sigma}$). If $\mu(\mathcal{X}) < +\infty$, then the following four conditions are* ***equivalent****:*

Learnability in $\mathscr{D}_{XY}^{\mu,b}$ for $\mathcal{H}$ $\iff$ Condition 1 $\iff$ Realizability Assumption $\iff$ Condition 3

Theorem 11 still holds if the function space $\mathcal{F}_{\mathbf{q}}^{\sigma}$ is generated by Convolutional Neural Network.

Overlap and Benefits of Multi-class Case. We investigate when the hypothesis space is FCNN-based or score-based, what will happen if there exists an overlap between the ID and OOD distributions?

Theorem 12. *Let $K = 1$ and the hypothesis space $\mathcal{H}$ be score-based ($\mathcal{H} = \mathcal{H}_{\mathbf{q},E}^{\sigma,\lambda}$, where E is in Eqs. (5) or (6)) or FCNN-based ($\mathcal{H} = \mathcal{H}_{\mathbf{q}}^{\sigma}$). Given a prior-unknown space $\mathscr{D}_{XY}$, if there exists a domain $D_{XY} \in \mathscr{D}_{XY}$, which has an overlap between ID and OOD distributions (see Definition 4), then OOD detection is not learnable in the domain space $\mathscr{D}_{XY}$ for $\mathcal{H}$.*

When $K = 1$ and the hypothesis space is FCNN-based or score-based, Theorem 12 shows that overlap between ID and OOD distributions is the sufficient condition for the unlearnability of OOD detection. Theorem 12 takes roots in the conditions $\inf_{h\in\mathcal{H}} R_D^{\text{in}}(h) = 0$ and $\inf_{h\in\mathcal{H}} R_D^{\text{out}}(h) = 0$. However, when $K > 1$, we can ensure $\inf_{h\in\mathcal{H}} R_D^{\text{in}}(h) > 0$ if ID distribution $D_{X_{\text{I}}Y_{\text{I}}}$ has overlap between ID classes. By this observation, we conjecture that when $K > 1$, OOD detection is learnable in some special cases where overlap exists, even if the hypothesis space is FCNN-based or score-based.

7 Discussion

Understanding Far-OOD Detection. Many existing works [7, 39] study the far-OOD detection issue. Existing benchmarks include 1) MNIST [40] as ID dataset, and Texture [41], CIFAR-10 [42] or Place365 [43] as OOD datasets; and 2) CIFAR-10 [42] as ID dataset, and MNIST [40], or Fashion-MNIST [43] as OOD datasets. In far-OOD case, we find that the ID and OOD datasets have different semantic labels and different styles. From the theoretical view, we can define far-OOD detection tasks as follows: for $\tau > 0$, a domain space $\mathscr{D}_{XY}$ is τ-far-OOD, if for any domain $D_{XY} \in \mathscr{D}_{XY}$,

$$\text{dist}(\text{supp}D_{X_{\text{O}}}, \text{supp}D_{X_{\text{I}}}) > \tau.$$

Theorems 7, 8 and 10 imply that under appropriate hypothesis space, τ-far-OOD detection is learnable. In Theorem 7, the condition $|\mathcal{X}| < +\infty$ is necessary for the separate space. However, one can prove that in the far-OOD case, when $\mathcal{H}^{\text{in}}$ is agnostic PAC learnable for ID distribution, the results in Theorem 7 still holds, if the condition $|\mathcal{X}| < +\infty$ is replaced by a weaker condition that $\mathcal{X}$ is compact. In addition, it is notable that when $\mathcal{H}^{\text{in}}$ is agnostic PAC learnable for ID distribution and $\mathcal{X}$ is compact, the KNN-based OOD detection algorithm [44] is consistent in the τ-far-OOD case.

Understanding Near-OOD Detection. When the ID and OOD datasets have similar semantics or styles, OOD detection tasks become more challenging. [45, 46] consider this issue and names it near-OOD detection. Existing benchmarks include 1) MNIST [40] as ID dataset, and Fashion-MNIST [43] or Not-MNIST [47] as OOD datasets; and 2) CIFAR-10 [42] as ID dataset, and CIFAR-100 [48] as OOD dataset. From the theoretical view, some near-OOD tasks may imply the overlap condition, *i.e.* Definition 4. Therefore, Theorems 3 and 12 imply that near-OOD detection may be not learnable. Developing a theory to understand the feasibility of near-OOD detection is still an *open question*.

Understanding One-class Novelty Detection. In one-class novelty detection and semantic anomaly detection (*i.e.* $K = 1$), Theorem 6 has revealed that it is necessary to use a large-capacity model to ensure the good generalization in the separate space. Theorem 3 and Theorem 12 suggest that we should try to avoid the overlap between ID and OOD distributions in the one-class case. If the overlap cannot be avoided, we suggest considering the multi-class OOD detection instead of the one-class case. Additionally, in the density-based space, Theorem 11 has shown that it is necessary to select a suitable hypothesis space satisfying the Realizability Assumption to ensure the learnability of OOD

detection in the density-based space. Generally, a large-capacity model can be helpful to guarantee that the Realizability Assumption holds.

FPR and AUROC Metrics. FPR (*i.e.*, False Positive Rate) and AUROC (*i.e.*, Area under the Receiver Operating Characteristic Curve) metrics are commonly used metrics in OOD detection [4]. TPR (*i.e.*, True Positive Rate) is a corresponding metric with respect to FPR. Generally, we hope that an OOD detection algorithm can approximate the optimal TPR and FPR performances. With minor adjustment, our theory can also be used to understand OOD detection under the TPR and FPR metrics. However, due to the complexity of AUROC, we still cannot give an OOD detection theory under AUROC metric. Understanding OOD detection under AUROC metric is still an *open question*.

8 Related Work

We briefly review the related theoretical works below. See Appendix A for detailed related works.

OOD Detection Theory. [49] understands the OOD detection via goodness-of-fit tests and typical set hypothesis, and argues that minimal density estimation errors can lead to OOD detection failures, when there exists an overlap between ID and OOD distributions. Beyond [49], [50] paves a new avenue to designing provable OOD detection algorithms. Compared to [50, 49], our theory focuses on the PAC learnable theory of OOD detection and identifies several necessary and sufficient conditions for the learnability of OOD detection, opening a door to study OOD detection in theory.

Open-set Learning Theory. [51] and [29, 52] propose the agnostic PAC learning bounds for open-set detection and open-set domain adaptation, respectively. Unfortunately, [29, 51, 52] all require that the test data are indispensable during the training process. To investigate open-set learning (OSL) *without accessing the test data* during training, [24] proposes and investigates the *almost* agnostic PAC learnability for OSL. However, the assumptions used in [24] are very strong and unpractical.

Learning Theory for Classification with Reject Option. Many works [53, 54] also investigate the *classification with reject option* (CwRO) problem, which is similar to OOD detection in some cases. [55, 56, 57, 58, 59] study the learning theory and propose the PAC learning bounds for CwRO. However, compared to our work regarding OOD detection, existing CwRO theories mainly focus on how the ID risk R_D^{in} (*i.e.*, the risk that ID data is wrongly classified) is influenced by special rejection rules. Our theory not only focuses on the ID risk, but also pays attention to the OOD risk.

PQ Learning Theory. Under some conditions, PQ learning theory [60, 61] can be regarded as the PAC theory for OOD detection in the semi-supervised or transductive learning cases, *i.e.*, test data are required during training. Besides, [60, 61] aim to give the PAC estimation under Realizability Assumption [21]. Our theory does not only study the PAC estimation in the realization cases, but also studies the other cases, which are more difficult than PAC theory under Realizability Assumption.

9 Conclusions and Future Works

Detecting OOD data has shown its significance in improving the reliability of machine learning. However, very few works discuss OOD detection in theory, which hinders real-world applications of OOD detection algorithms. In this paper, we are the *first* to provide the PAC theory for OOD detection. Our results imply that we cannot expect a universally consistent algorithm to handle all scenarios in OOD detection. Yet, it is still possible to make OOD detection learnable in certain scenarios. For example, when we design OOD detection algorithms, we normally only have finite ID datasets. In this real scenario, Theorem 8 provides a necessary and sufficient condition for the success of OOD detection. Our theory reveals many necessary and sufficient conditions for the learnability of OOD detection, hence *opening a door* to studying the learnability of OOD detection. In future, we will focus on studying the robustness of OOD detection based on robust statistics [62, 63].

Acknowledgment

JL and ZF were supported by the Australian Research Council (ARC) under FL190100149. YXL was supported by the Wisconsin Alumni Research Foundation (WARF). BH was supported by the RGC Early Career Scheme No. 22200720. The authors thank the reviewers and the meta-reviewer for their helpful comments. ZF would also like to thank Dr. Tongliang Liu for productive discussions.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[2] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *CVPR*, 2017.

[3] Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized ODIN: detecting out-of-distribution image without learning from out-of-distribution data. In *CVPR*, 2020.

[4] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *CoRR*, abs/2110.11334, 2021.

[5] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *The IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2016.

[6] Jiefeng Chen, Yixuan Li, Xi Wu, Yingyu Liang, and Somesh Jha. Atom: Robustifying out-of-distribution detection using outlier mining. *ECML*, 2021.

[7] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2017.

[8] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*, 2018.

[9] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *NeurIPS*, 2018.

[10] Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Dae-ki Cho, and Haifeng Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *ICLR*, 2018.

[11] Stanislav Pidhorskyi, Ranya Almohsen, and Gianfranco Doretto. Generative probabilistic novelty detection with adversarial autoencoders. In *NeurIPS*, 2018.

[12] Eric T. Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Görür, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *ICLR*, 2019.

[13] Dan Hendrycks, Mantas Mazeika, and Thomas G. Dietterich. Deep anomaly detection with outlier exposure. In *ICLR*, 2019.

[14] Jie Ren, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark A. DePristo, Joshua V. Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In *NeurIPS*, 2019.

[15] Ziqian Lin, Sreya Dutta Roy, and Yixuan Li. Mood: Multi-level out-of-distribution detection. In *CVPR*, 2021.

[16] Mohammadreza Salehi, Hossein Mirzaei, Dan Hendrycks, Yixuan Li, Mohammad Hossein Rohban, and Mohammad Sabokrou. A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges. *arXiv preprint arXiv:2110.14051*, 2021.

[17] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. In *NeurIPS*, 2021.

[18] Rui Huang, Andrew Geng, and Yixuan Li. On the Importance of Gradients for Detecting Distributional Shifts in the Wild. In *NeurIPS*, 2021.

[19] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the Limits of Out-of-Distribution Detection. In *NeurIPS*, 2021.

[20] Yifei Ming, Hang Yin, and Yixuan Li. On the impact of spurious correlation for out-of-distribution detection. *AAAI*, 2022.

[21] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[22] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[23] Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *NeurIPS*, 2020.

[24] Zhen Fang, Jie Lu, Anjin Liu, Feng Liu, and Guangquan Zhang. Learning bounds for open-set learning. In *ICML*, 2021.

[25] Guangyao Chen, Peixi Peng, Xiangqian Wang, and Yonghong Tian. Adversarial reciprocal points learning for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[26] Lukas Ruff, Nico Görnitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Robert A. Vandermeulen, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *ICML*, 2018.

[27] Sachin Goyal, Aditi Raghunathan, Moksh Jain, Harsha Vardhan Simhadri, and Prateek Jain. DROCC: deep robust one-class classification. In *ICML*, 2020.

[28] Lucas Deecke, Robert A. Vandermeulen, Lukas Ruff, Stephan Mandt, and Marius Kloft. Image anomaly detection with generative adversarial networks. In *ECML*, 2018.

[29] Z. Fang, Jie Lu, F. Liu, Junyu Xuan, and G. Zhang. Open set domain adaptation: Theoretical bound and algorithm. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[30] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *J. Mach. Learn. Res.*, 11:2635–2670, 2010.

[31] Guangyao Chen, Limeng Qiao, Yemin Shi, Peixi Peng, Jia Li, Tiejun Huang, Shiliang Pu, and Yonghong Tian. Learning open set network with discriminative reciprocal points. *ICCV*, 2020.

[32] Jiefeng Chen, Yixuan Li, Xi Wu, Yingyu Liang, and Somesh Jha. Informative outlier matters: Robustifying out-of-distribution detection using outlier mining. *ICML Workshop*, 2020.

[33] Jiefeng Chen, Yixuan Li, Xi Wu, Yingyu Liang, and Somesh Jha. Robust out-of-distribution detection for neural networks. *arXiv preprint arXiv:2003.09711*, 2020.

[34] Wentao Bao, Qi Yu, and Yu Kong. Evidential deep learning for open set action recognition. *ICCV*, 2021.

[35] Wentao Bao, Qi Yu, and Yu Kong. Opental: Towards open set temporal action localization. *CVPR*, 2022.

[36] Donald L Cohn. *Measure theory*. Springer, 2013.

[37] Peter L. Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019.

[38] Marek Karpinski and Angus Macintyre. Polynomial bounds for VC dimension of sigmoidal and general pfaffian neural networks. *J. Comput. Syst. Sci.*, 54(1):169–176, 1997.

[39] Jingkang Yang, Kaiyang Zhou, and Ziwei Liu. Full-spectrum out-of-distribution detection. *CoRR*, 2022.

[40] Li Deng. The MNIST database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Process. Mag.*, 2012.

[41] Gustaf Kylberg. *Kylberg texture dataset v. 1.0*. 2011.

[42] Alex Krizhevsky and Geoff Hinton. Convolutional deep belief networks on cifar-10. *Technical report, Citeseer*, 2009.

[43] Bolei Zhou, Àgata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2018.

[44] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *ICML*, 2022.

[45] Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-ood detection. *CoRR*, abs/2106.09022, 2021.

[46] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. In *NeurIPS*, 2021.

[47] Yaroslav Bulatov. Notmnist dataset. *Google (Books/OCR), Tech. Rep.[Online]. Available: http://yaroslavvb. blogspot. it/2011/09/notmnist-dataset. html,2*, 2011.

[48] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 and cifar-100 datasets. 2009.

[49] Lily H. Zhang, Mark Goldstein, and Rajesh Ranganath. Understanding failures in out-of-distribution detection with deep generative models. In *ICML*, 2021.

[50] Peyman Morteza and Yixuan Li. Provable guarantees for understanding out-of-distribution detection. *AAAI*, 2022.

[51] Si Liu, Risheek Garrepalli, Thomas G. Dietterich, Alan Fern, and Dan Hendrycks. Open category detection with PAC guarantees. In *ICML*, 2018.

[52] Yadan Luo, Zijian Wang, Zi Huang, and Mahsa Baktashmotlagh. Progressive graph learning for open-set domain adaptation. In *ICML*, 2020.

[53] C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 1970.

[54] Vojtech Franc, Daniel Průša, and V. Voracek. Optimal strategies for reject option classifiers. *CoRR*, abs/2101.12523, 2021.

[55] Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Learning with rejection. In *ALT*, 2016.

[56] Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Boosting with abstention. In *NeurIPS*, 2016.

[57] Chenri Ni, Nontawat Charoenphakdee, Junya Honda, and Masashi Sugiyama. On the calibration of multiclass classification with rejection. In *NeurIPS*, 2019.

[58] Nontawat Charoenphakdee, Zhenghang Cui, Yivan Zhang, and Masashi Sugiyama. Classification with rejection based on cost-sensitive classification. In *ICML*, 2021.

[59] Peter L. Bartlett and Marten H. Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 2008.

[60] Shafi Goldwasser, Adam Tauman Kalai, Yael Kalai, and Omar Montasser. Beyond perturbations: Learning guarantees with arbitrary adversarial test examples. In *NeurIPS*, 2020.

[61] Adam Tauman Kalai and Varun Kanade. Efficient learning with arbitrary covariate shift. In *ALT*, Proceedings of Machine Learning Research, 2021.

[62] Ilias Diakonikolas and Daniel M. Kane. Recent advances in algorithmic high-dimensional robust statistics. *A shorter version appears as an Invited Book Chapter in Beyond the Worst-Case Analysis of Algorithms*, 2020.

[63] Ilias Diakonikolas, Daniel M. Kane, Alistair Stewart, and Yuxin Sun. Outlier-robust learning of ising models under dobrushin's condition. In *COLT*, 2021.

[64] Akshay Raj Dhamija, Manuel Günther, and Terrance E. Boult. Reducing network agnostophobia. In *NeurIPS*, pages 9175–9186, 2018.

[65] Haoran Wang, Weitang Liu, Alex Bocchieri, and Yixuan Li. Can multi-label classification networks know what they don't know? In *NeurIPS*, 2021.

[66] Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *NeurIPS*, 2018.

[67] Zhisheng Xiao, Qing Yan, and Yali Amit. Likelihood regret: An out-of-distribution detection score for variational auto-encoder. In *NeurIPS*, 2020.

[68] Alireza Zaeemzadeh, Niccoló Bisagno, Zeno Sambugaro, Nicola Conci, Nazanin Rahnavard, and Mubarak Shah. Out-of-distribution detection using union of 1-dimensional subspaces. In *CVPR*, 2021.

[69] Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *ICML*, 2020.

[70] Sachin Vernekar, Ashish Gaurav, Vahdat Abdelzad, Taylor Denouden, Rick Salay, and Krzysztof Czarnecki. Out-of-distribution detection in classifiers via generation. In *NeurIPS Workshop*, 2019.

[71] Ryuichi Kiryo, Gang Niu, Marthinus Christoffel du Plessis, and Masashi Sugiyama. Positive-unlabeled learning with non-negative risk estimator. In *NeurIPS*, 2017.

[72] Takashi Ishida, Gang Niu, and Masashi Sugiyama. Binary classification from positive-confidence data. In *NeurIPS*, 2018.

[73] Shuo Chen, Gang Niu, Chen Gong, Jun Li, Jian Yang, and Masashi Sugiyama. Large-margin contrastive learning with distance polarization regularizer. In *ICML*, 2021.

[74] Jiahua Dong, Yang Cong, Gan Sun, Bineng Zhong, and Xiaowei Xu. What can be transferred: Unsupervised domain adaptation for endoscopic lesions segmentation. In *CVPR*, 2020.

[75] Zhen Fang, Jie Lu, Feng Liu, and Guangquan Zhang. Semi-supervised heterogeneous domain adaptation: Theory and algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[76] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[77] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander J. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 2012.

[78] Itay Safran and Ohad Shamir. Depth-width tradeoffs in approximating natural functions with neural networks. In *ICML*, 2017.

[79] Allan Pinkus. Approximation theory of the mlp model in neural networks. *Acta numerica*, 8:143–195, 1999.

[80] Peter L Bartlett and Wolfgang Maass. Vapnik-chervonenkis dimension of neural nets. *The handbook of brain theory and neural networks*, 2003.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Appendix B
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Appendix B
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Table of Contents of Appendix

A Detailed Related Work **17**

B Limitations and Potential Negative Societal Impacts **18**

C Discussions and Details about Experiments in Figure 1 **18**
C.1 Summary 18
C.2 Details of Experiments in Figure 1 19

D Notations **20**
D.1 Main Notations and Their Descriptions 20
D.2 Realizability Assumption 20
D.3 Learnability and PAC learnability 20
D.4 Explanations for Some Notations in Section 2 22

E Proof of Theorem 1 **24**

F Proof of Theorem 2 **26**

G Proofs of Theorem 3 and Theorem 4 **29**
G.1 Proof of Theorem 3 29
G.2 Proof of Theorem 4 31

H Proof of Theorem 5 **31**

I Proofs of Theorem 6 and Theorem 7 **35**
I.1 Proof of Theorem 6 35
I.2 Proof of Theorem 7 38

J Proofs of Theorems 8 and 9 **39**
J.1 Proof of Theorem 8 39
J.2 Proof of Theorem 9 41

K Proof of Proposition 1 and Proof of Proposition 2 **44**

L Proof of Theorem 10 **45**

M Proofs of Theorem 11 and Theorem 12 **51**
M.1 Proof of Theorem 11 51
M.2 Proof of Theorem 12 52

A Detailed Related Work

OOD Detection Algorithms. We will briefly review many representative OOD detection algorithms in three categories. 1) Classification-based methods use an ID classifier to detect OOD data [7][5]. Representative works consider using the maximum softmax score [7], temperature-scaled score [14] and energy-based score [23, 65] to identify OOD data. 2) Density-based methods aim to estimate an ID distribution and identify the low-density area as OOD data [10]. 3) The recent development of generative models provides promising ways to make them successful in OOD detection [11, 12, 14, 66, 67]. Distance-based methods are based on the assumption that OOD data should be relatively far away from the centroids of ID classes [9], including Mahalanobis distance [9, 45], cosine similarity [68], and kernel similarity [69].

Early works consider using the maximum softmax score to express the ID-ness [7]. Then, temperature scaling functions are used to amplify the separation between the ID and OOD data [14]. Recently, researchers propose hyperparameter-free energy scores to improve the OOD uncertainty estimation [23, 65]. Additionally, researchers also consider using the information contained in gradients to help improve the performance of OOD detection [18].

Except for the above algorithms, researchers also study the situation, where auxiliary OOD data can be obtained during the training process [13, 64]. These methods are called outlier exposure, and have much better performance than the above methods due to the appearance of OOD data. However, the exposure of OOD data is a strong assumption [4]. Thus, researchers also consider generating OOD data to help the separation of OOD and ID data [70]. In this paper, we do not make an assumption that OOD data are available during training, since this assumption may not hold in real world.

OOD Detection Theory. [49] understands the OOD detection via goodness-of-fit tests and typical set hypothesis, and discovers that relevant OOD distributions can lie in the high likelihood regions of a data distribution. [49] argues that minimal density estimation errors can lead to OOD detection failures, when there is an overlap between ID and OOD distributions. Compared to [49], our theory focuses on the PAC learnable theory of OOD detection. If detectors are generated by FCNN, our theory (Theorem 12) shows that the overlap is the sufficient condition to the failure of learnability of OOD detection, which is complementary to [49]. In addition, we identify several necessary and sufficient conditions for the learnability of OOD detection, which opens a door to studying OOD detection in theory. Beyond [49], [50] paves a new avenue to designing provable OOD detection algorithms. Compared to [50], our paper aims to characterize the learnability of OOD detection to answer the question: is OOD detection PAC learnable?

Open-set Learning Theory. [51] is the first to propose the agnostic PAC guarantees for open-set detection. Unfortunately, the test data must be used during the training process. [29] considers the open-set domain adaptation (OSDA) [52] and proposes the first learning bound for OSDA. [29] mainly depends on the positive-unlabeled learning techniques [71, 72, 73]. However, similar to [51], the test data must be available during training. To study open-set learning (OSL) *without accessing the test data* during training, [24] proposes and studies the almost PAC learnability for OSL, which is motivated by transfer learning [74, 75]. In our paper, we study the PAC learnability for OOD detection, which is an open problem proposed by [24].

Learning Theory for Classification with Reject Option. Many works [53, 54] also investigate the *classification with reject option* (CwRO) problem, which is similar to OOD detection in some cases. [55, 56, 57, 58, 59] study the learning theory and propose the agnostic PAC learning bounds for CwRO. However, compared to our work regarding OOD detection, existing CwRO theories mainly focus on how the ID risk (*i.e.*, the risk that ID data is wrongly classified) is influenced by special rejection rules. Our theory not only focuses on the ID risk, but also pays attention to the OOD risk.

PQ Learning Theory. Under some conditions, PQ learning theory [60, 61] can be regarded as the PAC theory for OOD detection in the semi-supervised or transductive learning cases, *i.e.*, test data are required during the training process. Additionally, PQ learning theory in [60, 61] aims to give the PAC estimation under Realizability Assumption [21]. Our theory focuses on the PAC theory in different cases, which is more difficult and more practical than PAC theory under Realizability Assumption.

[5] Note that, some methods assume that OOD data are available in advance [13, 64]. However, the exposure of OOD data is a strong assumption [4]. We do not consider this situation in our paper.

B Limitations and Potential Negative Societal Impacts

Limitations. The main limitation of our work lies in that we do not answer the most general question:

Given any hypothesis space $\mathcal{H}$ and space $\mathscr{D}_{XY}$, what is the necessary and sufficient condition to ensure the PAC learnability of OOD detection?

However, this question is still difficult to be addressed, due to limited mathematical skills. Yet, based on our observations and the main results in our paper, we believe the following result may hold:

Conjecture: *If $\mathcal{H}$ is agnostic learnable for supervised learning, then OOD detection is learnable in $\mathscr{D}_{XY}$* ***if and only if*** *compatibility condition (i.e., Condition 3) holds.*

We leave this question as a future work.

Potential Negative Societal Impacts. Since our paper is a theoretical paper and the OOD detection problem is significant to ensure the safety of deploying existing machine learning algorithms, there are no potential negative societal impacts in our paper.

C Discussions and Details about Experiments in Figure 1

In this section, we summarize our main results, then give the details of the experiments in Figure 1.

C.1 Summary

We summarize our main results as follows:

• A necessary condition (*i.e.*, Condition 1) for the learnability of OOD detection is proposed. Theorem 2 shows that Condition 1 is the *necessary and sufficient condition* for the learnability of OOD detection, when the domain space is the single-distribution space $\mathscr{D}_{XY}^{D_{XY}}$. This implies the Condition 1 is the necessary condition for the learnability of OOD detection.

• Theorem 3 has shown that the overlap between ID and OOD data can lead the failures of OOD detection under some mild assumptions. Furthermore, Theorem 12 shows that when $K = 1$, the overlap is the sufficient condition for the failures of OOD detection, when the hypothesis space is FCNN-based or score-based.

• Theorem 4 provides an impossibility theorem for the total space $\mathscr{D}_{XY}^{\text{all}}$. OOD detection is not learnable in $\mathscr{D}_{XY}^{\text{all}}$ for any non-trivial hypothesis space.

• Theorem 5 gives impossibility theorems for the separate space $\mathscr{D}_{XY}^{s}$. To ensure the impossibility theorems hold, mild assumptions are required. Theorem 5 also implies that OOD detection may be learnable in the separate space $\mathscr{D}_{XY}^{s}$, if the feature space is finite, *i.e.*, $|\mathcal{X}| < +\infty$. Additionally, Theorem 10 implies that the finite feature space may be the necessary condition to ensure the learnability of OOD detection in the separate space.

• When $|\mathcal{X}| < +\infty$ and $K = 1$, Theorem 6 provides the *necessary and sufficient condition* for the learnability of OOD detection in the separate space $\mathscr{D}_{XY}^{s}$. Theorem 6 implies that if the OOD detection can be learnable in the distribution-agnostic case, then a large-capacity model is necessary. Based on Theorem 6, Theorem 7 studies the learnability in the $K > 1$ case.

• The compatibility condition (*i.e.*, Condition 3) for the learnability of OOD detection is proposed. Theorem 8 shows that Condition 3 is the *necessary and sufficient condition* for the learnability of OOD detection in the finite-ID-distribution space $\mathscr{D}_{XY}^{F}$. This also implies Condition 3 is the necessary condition for any prior-unknown space. Note that we can only collect finite ID datasets to build models. Hence, Theorem 8 can handle the most practical scenarios.

• To further understand the importance of the compatibility condition (Condition 3). Theorem 9 considers the density-based space $\mathscr{D}_{XY}^{\mu,b}$. We discover that Realizability Assumption implies the compatibility condition in the density-based space. Based on this observation, we prove that OOD detection is learnable in $\mathscr{D}_{XY}^{\mu,b}$ under Realizability Assumption.

• Theorem 10 gives practical applications of our theory. In this theorem, we discover that the finite feature space is a *necessary and sufficient condition* for the learnability of OOD detection in the separate space $\mathscr{D}^{s}_{XY}$, when the hypothesis space is FCNN-based or score-based.

• Theorem 11 has shown that when $K = 1$ and the hypothesis space is FCNN-based or score-based, Realizability Assumption, Condition 3, Condition 1 and the learnability of OOD detection in the density-based space $\mathcal{D}^{\mu,b}_{XY}$ are all *equivalent*.

• **Meaning of Our Theory.** In classical statistical learning theory, the generalization theory guarantees that a well-trained classifier can be generalized well on the test set as long as the training and test sets are from the same distribution [21, 22]. However, since the OOD data are unseen during the training process, it is very difficult to determine whether the generalization theory holds for OOD detection.

Normally, OOD data are unseen and can be various. We hope that there exists an algorithm that can be used for the various OOD data instead of some certain OOD data, which is the reason why the generalization theory for OOD detection needs to be developed. In this paper, we investigate the generalization theory regarding OOD detection and point out when the OOD detection can be successful. Our theory is based on the PAC learning theory. The impossibility theorems and the given necessary and sufficient conditions outlined provide important perspectives from which to think about OOD detection.

C.2 Details of Experiments in Figure 1

In this subsection, we present details of the experiments in Figure 1, including data generation, configuration and OOD detection procedure.

Data Generation. ID and OOD data are drawn from the following *uniform* (U) distributions (note that we use $\mathrm{U}(\mathbf{I})$ to present the uniform distribution in region $\mathbf{I}$).
• The marginal distribution of ID distribution for class c: for any $c \in \{1, ..., 10\}$,

$$D_{X_{\mathrm{I}}|Y_{\mathrm{I}}=c} = \mathrm{U}(\mathbf{I}_c), \text{ where } \mathbf{I}_c = [d_c, d_c + 4] \times [1, 5], \tag{7}$$

here $d_i = 5 + \mathrm{gap}_{\mathrm{II}} * (i - 1) + 4(i - 2)$ and $\mathrm{gap}_{\mathrm{II}}$ is a positive constant.
• The class-prior probability for class c: for any $c \in \{1, ..., 10\}$,

$$D_{Y_{\mathrm{I}}}(y = c) = \frac{1 - \alpha}{10}.$$

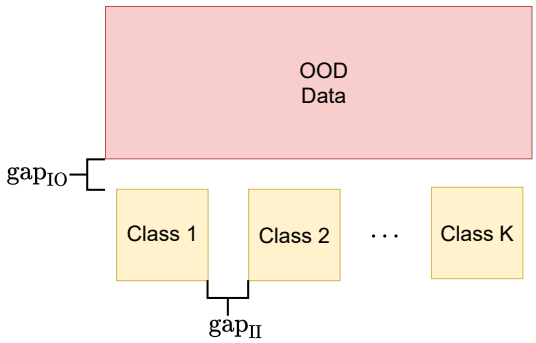
• The marginal distribution of OOD distribution:

$$D_{X_{\mathrm{O}}} = \mathrm{U}(\mathbf{I}_{\mathrm{out}}), \text{ where } \mathbf{I}_{\mathrm{out}} = [d_1 - 1, d_{10} + 5] \times [5 + \mathrm{gap}_{\mathrm{IO}}, 10 + \mathrm{gap}_{\mathrm{IO}}]. \tag{8}$$

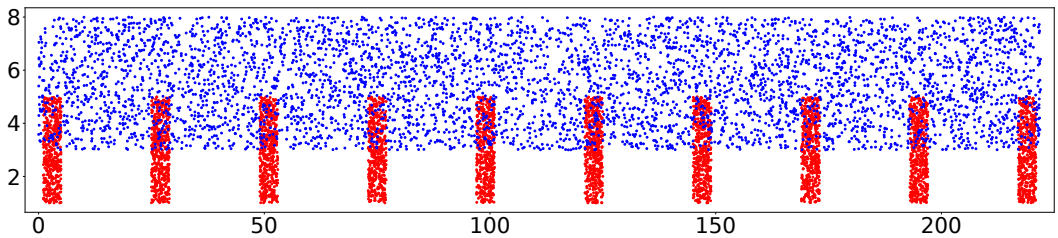
Figure 2 shows the OOD and ID distributions, when $\mathrm{gap}_{\mathrm{II}} = 20$ and $\mathrm{gap}_{\mathrm{IO}} = -2$. In Figure 1, we draw n data from ID distribution ($n = 15,000, 20,000, 25,000$) and $25,000$ data from the OOD distribution.

Configuration. The architecture of ID classifier is a four-layer FCNN. The number of neurons in hidden layers is set to 100, and the number of neurons of output layer is set to 10. These neurons use sigmoid activations. We use the Adam optimizer [76] to optimize the network's parameters (with the ℓ_2 loss). The learning rate is set to 0.001, and the max number of training iterations is set to $10,000$. Within each iteration, we use full batch to update the network's parameters. $\mathrm{gap}_{\mathrm{II}}$ is set to 20 in our experiments. In Figure 1b, $\mathrm{gap}_{\mathrm{IO}} = -2$ (the overlap exists, see Figure 2), and in Figure 1c, $\mathrm{gap}_{\mathrm{IO}} = 100$ (no overlap).

OOD Detection Procedure. We first train an ID classifier with n data drawn from the ID distribution. Then, according to [23], we apply the free-energy score to identify the OOD data and calculate the α-risk (with the 0-1 loss). We repeat the above detection procedure 20 times and report the average α-risk in Figure 1. Note that, following [23], we choose the threshold used by the free-energy method so that 95% of ID data are correctly identified as the ID classes by the OOD detector.



(a) ID and OOD Distributions



(b) Illustration of ID and OOD Data

Figure 2: ID and OOD distributions in Figure 1.

D Notations

D.1 Main Notations and Their Descriptions

In this section, we summarize important notations in Table 1.
Given $\mathbf{f} = [f^1, ..., f^l]^\top$, for any $\mathbf{x} \in \mathcal{X}$,

$$\arg\max_{k \in \{1,...,l\}} f^k(\mathbf{x}) := \max\{k \in \{1, ..., l\} : f^k(\mathbf{x}) \geq f^i(\mathbf{x}), \forall i = 1, ..., l\},$$

where f^k is the k-th coordinate of $\mathbf{f}$ and f^i is the i-th coordinate of $\mathbf{f}$. The above definition about $\arg\max$ aims to overcome some special cases. For example, there exist k_1, k_2 ($k_1 < k_2$) such that $f^{k_1}(\mathbf{x}) = f^{k_2}(\mathbf{x})$ and $f^{k_1}(\mathbf{x}) > f^i(\mathbf{x})$, $f^{k_2}(\mathbf{x}) > f^i(\mathbf{x})$, $\forall i \in \{1, ..., l\} - \{k_1, k_2\}$. Then, according to the above definition, $k_2 = \arg\max_{k \in \{1,...,l\}} f^k(\mathbf{x})$.

D.2 Realizability Assumption

Assumption 2 (Realizability Assumption)**.** *A domain space $\mathscr{D}_{XY}$ and hypothesis space $\mathcal{H}$ satisfy the Realizability Assumption, if for each domain $D_{XY} \in \mathscr{D}_{XY}$, there exists at least one hypothesis function $h^* \in \mathcal{H}$ such that $R_D(h^*) = 0$.*

D.3 Learnability and PAC learnability

Here we give a proof to show that Learnability given in Definition 1 and PAC learnability are equivalent.

First, we prove that Learnability concludes the PAC learnability.

According to Definition 1,

$$\mathbb{E}_{S \sim D^n_{X_{\text{I}} Y_{\text{I}}}} R_D(\mathbf{A}(S)) \leq \inf_{h \in \mathcal{H}} R_D(h) + \epsilon_{\text{cons}}(n),$$

Table 1: Main notations and their descriptions.

Notation	Description
• **Spaces and Labels**	
d and $\mathcal{X} \subset \mathbb{R}^d$	the feature dimension of data point and feature space
$\mathcal{Y}$	ID label space $\{1, ..., K\}$
$K+1$	$K+1$ represents the OOD labels
$\mathcal{Y}_{\text{all}}$	$\mathcal{Y} \cup \{K+1\}$
• **Distributions**	
X_{I}, X_{O}, Y_{I}, Y_{O}	ID feature, OOD feature, ID label, OOD label random variables
$D_{X_{\text{I}}Y_{\text{I}}}$, $D_{X_{\text{O}}Y_{\text{O}}}$	ID joint distribution and OOD joint distribution
D_{XY}^{α}	$D_{XY}^{\alpha} = (1-\alpha)D_{X_{\text{I}}Y_{\text{I}}} + \alpha D_{X_{\text{O}}Y_{\text{O}}}, \ \ \forall \alpha \in [0,1]$
π^{out}	class-prior probability for OOD distribution
D_{XY}	$D_{XY} = (1-\pi^{\text{out}})D_{X_{\text{I}}Y_{\text{I}}} + \pi^{\text{out}} D_{X_{\text{O}}Y_{\text{O}}}$, called domain
$D_{X_{\text{I}}}, D_{X_{\text{O}}}, D_X$	marginal distributions for $D_{X_{\text{I}}Y_{\text{I}}}$, $D_{X_{\text{O}}Y_{\text{O}}}$ and D_{XY}, respectively
• **Domain Spaces**	
$\mathscr{D}_{XY}$	domain space consisting of some domains
$\mathscr{D}_{XY}^{\text{all}}$	total space
$\mathscr{D}_{XY}^{s}$	seperate space
$\mathscr{D}_{XY}^{D_{XY}}$	single-distribution space
$\mathscr{D}_{XY}^{F}$	finite-ID-distribution space
$\mathscr{D}_{XY}^{\mu,b}$	density-based space
• **Loss Function, Function Spaces**	
$\ell(\cdot,\cdot)$	loss: $\mathcal{Y}_{\text{all}} \times \mathcal{Y}_{\text{all}} \to \mathbb{R}_{\geq 0}$: $\ell(y_1, y_2) = 0$ if and only if $y_1 = y_2$
$\mathcal{H}$	hypothesis space
$\mathcal{H}^{\text{in}}$	ID hypothesis space
$\mathcal{H}^{\text{b}}$	hypothesis space in binary classification
$\mathcal{F}_l$	scoring function space consisting some l dimensional vector-valued functions
• **Risks and Partial Risks**	
$R_D(h)$	risk corresponding to D_{XY}
$R_D^{\text{in}}(h)$	partial risk corresponding to $D_{X_{\text{I}}Y_{\text{I}}}$
$R_D^{\text{out}}(h)$	partial risk corresponding to $D_{X_{\text{O}}Y_{\text{O}}}$
$R_D^{\alpha}(h)$	α-risk corresponding to D_{XY}^{α}
• **Fully-Connected Neural Networks**	
$\mathbf{q}$	a sequence $(l_1, ..., l_g)$ to represent the architecture of FCNN
σ	activation function. In this paper, we use ReLU function
$\mathcal{F}_{\mathbf{q}}^{\sigma}$	FCNN-based scoring function space
$\mathcal{H}_{\mathbf{q}}^{\sigma}$	FCNN-based hypothesis space
$\mathbf{f}_{\mathbf{w},\mathbf{b}}$	FCNN-based scoring function, which is from $\mathcal{F}_{\mathbf{q}}^{\sigma}$
$h_{\mathbf{w},\mathbf{b}}$	FCNN-based hypothesis function, which is from $\mathcal{H}_{\mathbf{q}}^{\sigma}$
• **Score-based Hypothesis Space**	
E	scoring function
λ	threshold
$\mathcal{H}_{\mathbf{q},E}^{\sigma,\lambda}$	score-based hypothesis space—a binary classification space
$h_{\mathbf{f},E}^{\lambda}$	score-based hypothesis function—a binary classifier

which implies that

$$\mathbb{E}_{S\sim D_{X_{\text{I}}Y_{\text{I}}}^{n}}[R_D(\mathbf{A}(S)) - \inf_{h\in\mathcal{H}} R_D(h)] \leq \epsilon_{\text{cons}}(n).$$

Note that $R_D(\mathbf{A}(S)) - \inf_{h\in\mathcal{H}} R_D(h) \geq 0$. Therefore, by Markov's inequality, we have

$$\mathbb{P}(R_D(\mathbf{A}(S)) - \inf_{h\in\mathcal{H}} R_D(h) < \epsilon) > 1 - \mathbb{E}_{S\sim D_{X_{\text{I}}Y_{\text{I}}}^{n}}[R_D(\mathbf{A}(S)) - \inf_{h\in\mathcal{H}} R_D(h)]/\epsilon \geq 1 - \epsilon_{\text{cons}}(n)/\epsilon.$$

Because $\epsilon_{\text{cons}}(n)$ is monotonically decreasing, we can find a smallest m such that $\epsilon_{\text{cons}}(m) \geq \epsilon\delta$ and $\epsilon_{\text{cons}}(m-1) < \epsilon\delta$, for $\delta \in (0,1)$. We define that $m(\epsilon,\delta) = m$. Therefore, for any $\epsilon > 0$ and $\delta \in (0,1)$, there exists a function $m(\epsilon,\delta)$ such that when $n > m(\epsilon,\delta)$, with the probability at least $1-\delta$, we have

$$R_D(\mathbf{A}(S)) - \inf_{h\in\mathcal{H}} R_D(h) < \epsilon,$$

which is the definition of PAC learnability.

Second, we prove that the PAC learnability concludes Learnability.

PAC-learnability: for any $\epsilon > 0$ and $0 < \delta < 1$, there exists a function $m(\epsilon, \delta) > 0$ such that when the sample size $n > m(\epsilon, \delta)$, we have that with the probability at least $1 - \delta > 0$,

$$R_D(\mathbf{A}(S)) - \inf_{h \in \mathcal{H}} R_D(h) \leq \epsilon.$$

Note that the loss ℓ defined in Section 2 has upper bound (because $\mathcal{Y} \cup \{K+1\}$ is a finite set). We assume the upper bound of ℓ is M. Hence, according to the definition of PAC-learnability, when the sample size $n > m(\epsilon, \delta)$, we have that

$$\mathbb{E}_S[R_D(\mathbf{A}(S)) - \inf_{h \in \mathcal{H}} R_D(h)] \leq \epsilon(1-\delta) + 2M\delta < \epsilon + 2M\delta.$$

If we set $\delta = \epsilon$, then when the sample size $n > m(\epsilon, \epsilon)$, we have that

$$\mathbb{E}_S[R_D(\mathbf{A}(S)) - \inf_{h \in \mathcal{H}} R_D(h)] < (2M+1)\epsilon,$$

this implies that

$$\lim_{n \to +\infty} \mathbb{E}_S[R_D(\mathbf{A}(S)) - \inf_{h \in \mathcal{H}} R_D(h)] = 0,$$

which implies the Learnability in Definition 1. We have completed this proof.

D.4 Explanations for Some Notations in Section 2

First, we explain the concept that $S \sim D^n_{X_I Y_I}$ in Eq. (2).

$S = \{(\mathbf{x}^1, y^1), ..., (\mathbf{x}^n, y^n)\}$ is training data drawn independent and identically distributed from $D_{X_\mathrm{I} Y_\mathrm{I}}$.

$D^n_{X_\mathrm{I} Y_\mathrm{I}}$ denotes the probability over n-tuples induced by applying $D_{X_\mathrm{I} Y_\mathrm{I}}$ to pick each element of the tuple independently of the other members of the tuple.

Because these samples are i.i.d. drawn n times, researchers often use "$S \sim D^n_{X_\mathrm{I} Y_\mathrm{I}}$" to represent a sample set S (of size n) whose each element is drawn i.i.d. from $D_{X_\mathrm{I} Y_\mathrm{I}}$.

Second, we explain the concept "+" in $(1 - \pi^{\mathrm{out}}) D_{X_\mathrm{I}} + \pi^{\mathrm{out}} D_{X_\mathrm{O}}$.

For convenience, let $P = (1 - \pi^{\mathrm{out}}) D_{X_\mathrm{I}}$ and $Q = \pi^{\mathrm{out}} D_{X_\mathrm{O}}$. It is clear that P and Q are measures. Then $P + Q$ is also a measure, which is defined as follows: for any measurable set $A \subset \mathcal{X}$, we have

$$(P + Q)(A) = P(A) + Q(A).$$

For example, when P and Q are discrete measures, then $P + Q$ is also discrete measure: for any $\mathbf{x} \in \mathcal{X}$,

$$(P + Q)(\mathbf{x}) = P(\mathbf{x}) + Q(\mathbf{x}).$$

When P and Q are continuous measures with density functions f and g, then $P + Q$ is also continuous measure with density function $f + g$: for any measurable $A \subset \mathcal{X}$,

$$P(A) = \int_A f(\mathbf{x}) \mathrm{d}\mathbf{x}, \quad Q(A) = \int_A g(\mathbf{x}) \mathrm{d}\mathbf{x},$$

then

$$(P + Q)(A) = \int_A f(\mathbf{x}) + g(\mathbf{x}) \mathrm{d}\mathbf{x}.$$

Third, we explain the concept $\mathbb{E}_{(\mathbf{x}, y) \sim D_{XY}} \ell(h(\mathbf{x}), y)$.

The concept $\mathbb{E}_{(\mathbf{x},y)\sim D_{XY}}\ell(h(\mathbf{x}),y)$ can be computed as follows:

$$\mathbb{E}_{(\mathbf{x},y)\sim D_{XY}}\ell(h(\mathbf{x}),y) = \int_{\mathcal{X}\times\mathcal{Y}_{\text{all}}} \ell(h(\mathbf{x}),y)\mathrm{d}D_{XY}(\mathbf{x},y).$$

For example, when D_{XY} is a finite discrete distribution: let $\mathcal{Z} = \{(\mathbf{x}^1, y^1), ..., (\mathbf{x}^m, y^m)\}$ be the support set of D_{XY}, and assume that a^i is the probability for $(\mathbf{x}^i, y^i)$, *i.e.*, $a^i = D_{XY}(\mathbf{x}^i, y^i)$. Then

$$\begin{aligned}\mathbb{E}_{(\mathbf{x},y)\sim D_{XY}}\ell(h(\mathbf{x}),y) &= \int_{\mathcal{X}\times\mathcal{Y}_{\text{all}}} \ell(h(\mathbf{x}),y)\mathrm{d}D_{XY}(\mathbf{x},y) \\ &= \frac{1}{m}\sum_{i=1}^{m} a^i \ell(h(\mathbf{x}^i), y^i).\end{aligned}$$

When D_X is a continuous distribution with density f, and $D_{Y|X}(Y = k|X = \mathbf{x})$ (k-th class-conditional distribution for $\mathbf{x}$) is $a^k(\mathbf{x})$, then

$$\begin{aligned}\mathbb{E}_{(\mathbf{x},y)\sim D_{XY}}\ell(h(\mathbf{x}),y) &= \int_{\mathcal{X}\times\mathcal{Y}_{\text{all}}} \ell(h(\mathbf{x}),y)\mathrm{d}D_{XY}(\mathbf{x},y) \\ &= \int_{\mathcal{X}} \sum_{k=1}^{K+1} \ell(h(\mathbf{x}),k) f(\mathbf{x}) a^k(\mathbf{x})\mathrm{d}\mathbf{x},\end{aligned}$$

where $D_{Y|X}(Y = k|X = \mathbf{x})$ is the k-th class-conditional distribution.

E Proof of Theorem 1

Theorem 1. *Given domain spaces $\mathscr{D}_{XY}$ and $\mathscr{D}'_{XY} = \{D^{\alpha}_{XY} : \forall D_{XY} \in \mathscr{D}_{XY}, \forall \alpha \in [0,1)\}$, then*
1) $\mathscr{D}'_{XY}$ is a priori-unknown space and $\mathscr{D}_{XY} \subset \mathscr{D}'_{XY}$;
*2) if $\mathscr{D}_{XY}$ is a priori-unknown space, then Definition 1 and Definition 2 are **equivalent**;*
*3) OOD detection is strongly learnable in $\mathscr{D}_{XY}$ **if and only if** OOD detection is learnable in $\mathscr{D}'_{XY}$.*

Proof of Theorem 1.

Proof of the First Result.

To prove that $\mathscr{D}'_{XY}$ is a priori-unknown space, we need to show that for any $D^{\alpha'}_{XY} \in \mathscr{D}'_{XY}$, then $D^{\alpha}_{XY} \in \mathscr{D}'_{XY}$ for any $\alpha \in [0,1)$.

According to the definition of $\mathscr{D}'_{XY}$, for any $D^{\alpha'}_{XY} \in \mathscr{D}'_{XY}$, we can find a domain $D_{XY} \in \mathscr{D}_{XY}$, which can be written as $D_{XY} = (1-\pi^{\text{out}})D_{X_{\text{I}}Y_{\text{I}}} + \pi^{\text{out}} D_{X_{\text{O}}Y_{\text{O}}}$ (here $\pi^{\text{out}} \in [0,1)$) such that

$$D^{\alpha'}_{XY} = (1-\alpha')D_{X_{\text{I}}Y_{\text{I}}} + \alpha' D_{X_{\text{O}}Y_{\text{O}}}.$$

Note that $D^{\alpha}_{XY} = (1-\alpha)D_{X_{\text{I}}Y_{\text{I}}} + \alpha D_{X_{\text{O}}Y_{\text{O}}}$.

Therefore, based on the definition of $\mathscr{D}'_{XY}$, for any $\alpha \in [0,1)$, $D^{\alpha}_{XY} \in \mathscr{D}'_{XY}$, which implies that $\mathscr{D}'_{XY}$ is a prior-known space. Additionally, for any $D_{XY} \in \mathscr{D}_{XY}$, we can rewrite D_{XY} as $D^{\pi_{\text{out}}}_{XY}$, thus $D_{XY} = D^{\pi_{\text{out}}}_{XY} \in \mathscr{D}'_{XY}$, which implies that $\mathscr{D}_{XY} \subset \mathscr{D}'_{XY}$.

Proof of the Second Result.

First, we prove that Definition 1 concludes Definition 2, if $\mathscr{D}_{XY}$ is a prior-unknown space:

> The domain space $\mathscr{D}_{XY}$ is a priori-unknown space, and OOD detection is learnable in $\mathscr{D}_{XY}$ for $\mathcal{H}$.
> $\Downarrow$
> OOD detection is strongly learnable in $\mathscr{D}_{XY}$ for $\mathcal{H}$: there exist an algorithm $\mathbf{A} : \cup_{n=1}^{+\infty}(\mathcal{X}\times\mathcal{Y})^n \to \mathcal{H}$, and a monotonically decreasing sequence $\epsilon(n)$, such that $\epsilon(n) \to 0$, as $n \to +\infty$
> $$\mathbb{E}_{S\sim D^n_{X_{\text{I}}Y_{\text{I}}}}\left[R^{\alpha}_{D}(\mathbf{A}(S)) - \inf_{h\in\mathcal{H}} R^{\alpha}_{D}(h)\right] \le \epsilon(n), \quad \forall \alpha \in [0,1],\ \forall D_{XY} \in \mathscr{D}_{XY}.$$

In the priori-unknown space, for any $D_{XY} \in \mathscr{D}_{XY}$, we have that for any $\alpha \in [0,1)$,

$$D^{\alpha}_{XY} = (1-\alpha)D_{X_{\text{I}}Y_{\text{I}}} + \alpha D_{X_{\text{O}}Y_{\text{O}}} \in \mathscr{D}_{XY}.$$

Then, according to the definition of learnability of OOD detection, we have an algorithm $\mathbf{A}$ and a monotonically decreasing sequence $\epsilon_{\text{cons}}(n) \to 0$, as $n \to +\infty$, such that for any $\alpha \in [0,1)$,

$$\mathbb{E}_{S\sim D^n_{X_{\text{I}}Y_{\text{I}}}} R_{D^{\alpha}}(\mathbf{A}(S)) \le \inf_{h\in\mathcal{H}} R_{D^{\alpha}}(h) + \epsilon_{\text{cons}}(n), \quad \text{(by the property of priori-unknown space)}$$

where

$$R_{D^{\alpha}}(\mathbf{A}(S)) = \int_{\mathcal{X}\times\mathcal{Y}_{\text{all}}} \ell(\mathbf{A}(S)(\mathbf{x}), y)\mathrm{d}D^{\alpha}_{XY}(\mathbf{x}, y), \quad R_{D^{\alpha}}(h) = \int_{\mathcal{X}\times\mathcal{Y}_{\text{all}}} \ell(h(\mathbf{x}), y)\mathrm{d}D^{\alpha}_{XY}(\mathbf{x}, y).$$

Since $R_{D^{\alpha}}(\mathbf{A}(S)) = R^{\alpha}_{D}(\mathbf{A}(S))$ and $R_{D^{\alpha}}(h) = R^{\alpha}_{D}(h)$, we have that

$$\mathbb{E}_{S\sim D^n_{X_{\text{I}}Y_{\text{I}}}} R^{\alpha}_{D}(\mathbf{A}(S)) \le \inf_{h\in\mathcal{H}} R^{\alpha}_{D}(h) + \epsilon_{\text{cons}}(n), \quad \forall \alpha \in [0,1). \tag{9}$$

Next, we consider the case that $\alpha = 1$. Note that

$$\liminf_{\alpha\to 1} \inf_{h\in\mathcal{H}} R^{\alpha}_{D}(h) \ge \liminf_{\alpha\to 1} \alpha \inf_{h\in\mathcal{H}} R^{\text{out}}_{D}(h) = \inf_{h\in\mathcal{H}} R^{\text{out}}_{D}(h). \tag{10}$$

Then, we assume that $h_{\epsilon} \in \mathcal{H}$ satisfies that

$$R^{\text{out}}_{D}(h_{\epsilon}) - \inf_{h\in\mathcal{H}} R^{\text{out}}_{D}(h) \le \epsilon.$$

It is obvious that

$$R^{\alpha}_{D}(h_{\epsilon}) \ge \inf_{h\in\mathcal{H}} R^{\alpha}_{D}(h).$$

Let $\alpha \to 1$. Then, for any $\epsilon > 0$,

$$R_D^{\text{out}}(h_\epsilon) = \lim_{\alpha\to 1} R_D^\alpha(h_\epsilon) = \limsup_{\alpha\to 1} R_D^\alpha(h_\epsilon) \geq \limsup_{\alpha\to 1} \inf_{h\in\mathcal{H}} R_D^\alpha(h),$$

which implies that

$$\inf_{h\in\mathcal{H}} R_D^{\text{out}}(h) = \lim_{\epsilon\to 0} R_D^{\text{out}}(h_\epsilon) \geq \lim_{\epsilon\to 0} \limsup_{\alpha\to 1} \inf_{h\in\mathcal{H}} R_D^\alpha(h) = \limsup_{\alpha\to 1} \inf_{h\in\mathcal{H}} R_D^\alpha(h). \tag{11}$$

Combining Eq. (10) with Eq. (11), we have

$$\inf_{h\in\mathcal{H}} R_D^{\text{out}}(h) = \limsup_{\alpha\to 1} \inf_{h\in\mathcal{H}} R_D^\alpha(h) = \liminf_{\alpha\to 1} \inf_{h\in\mathcal{H}} R_D^\alpha(h), \tag{12}$$

which implies that

$$\inf_{h\in\mathcal{H}} R_D^{\text{out}}(h) = \lim_{\alpha\to 1} \inf_{h\in\mathcal{H}} R_D^\alpha(h). \tag{13}$$

Note that

$$\mathbb{E}_{S\sim D^n_{X_{\text{I}}Y_{\text{I}}}} R_D^\alpha(\mathbf{A}(S)) = (1-\alpha)\mathbb{E}_{S\sim D^n_{X_{\text{I}}Y_{\text{I}}}} R_D^{\text{in}}(\mathbf{A}(S)) + \alpha\mathbb{E}_{S\sim D^n_{X_{\text{I}}Y_{\text{I}}}} R_D^{\text{out}}(\mathbf{A}(S)).$$

Hence, Lebesgue's Dominated Convergence Theorem [36] implies that

$$\lim_{\alpha\to 1} \mathbb{E}_{S\sim D^n_{X_{\text{I}}Y_{\text{I}}}} R_D^\alpha(\mathbf{A}(S)) = \mathbb{E}_{S\sim D^n_{X_{\text{I}}Y_{\text{I}}}} R_D^{\text{out}}(\mathbf{A}(S)). \tag{14}$$

Using Eq. (9), we have that

$$\lim_{\alpha\to 1} \mathbb{E}_{S\sim D^n_{X_{\text{I}}Y_{\text{I}}}} R_D^\alpha(\mathbf{A}(S)) \leq \lim_{\alpha\to 1} \inf_{h\in\mathcal{H}} R_D^\alpha(h) + \epsilon_{\text{cons}}(n). \tag{15}$$

Combining Eq. (13), Eq. (14) with Eq. (15), we obtain that

$$\mathbb{E}_{S\sim D^n_{X_{\text{I}}Y_{\text{I}}}} R_D^{\text{out}}(\mathbf{A}(S)) \leq \inf_{h\in\mathcal{H}} R_D^{\text{out}}(h) + \epsilon_{\text{cons}}(n).$$

Since $R_D^{\text{out}}(\mathbf{A}(S)) = R_D^1(\mathbf{A}(S))$ and $R_D^{\text{out}}(h) = R_D^1(h)$, we obtain that

$$\mathbb{E}_{S\sim D^n_{X_{\text{I}}Y_{\text{I}}}} R_D^1(\mathbf{A}(S)) \leq \inf_{h\in\mathcal{H}} R_D^1(h) + \epsilon_{\text{cons}}(n). \tag{16}$$

Combining Eq. (9) and Eq. (16), we have proven that: if the domain space $\mathscr{D}_{XY}$ is a priori-unknown space, then OOD detection is learnable in $\mathscr{D}_{XY}$ for $\mathcal{H}$.

$$\Downarrow$$

OOD detection is strongly learnable in $\mathscr{D}_{XY}$ for $\mathcal{H}$: there exist an algorithm $\mathbf{A} : \cup_{n=1}^{+\infty}(\mathcal{X}\times\mathcal{Y})^n \to \mathcal{H}$, and a monotonically decreasing sequence $\epsilon(n)$, such that $\epsilon(n) \to 0$, as $n \to +\infty$,

$$\mathbb{E}_{S\sim D^n_{X_{\text{I}}Y_{\text{I}}}} R_D^\alpha(\mathbf{A}(S)) \leq \inf_{h\in\mathcal{H}} R_D^\alpha(h) + \epsilon(n), \quad \forall\alpha\in[0,1],\ \forall D_{XY}\in\mathscr{D}_{XY}.$$

Second, we prove that Definition 2 concludes Definition 1:

OOD detection is strongly learnable in $\mathscr{D}_{XY}$ for $\mathcal{H}$: there exist an algorithm $\mathbf{A} : \cup_{n=1}^{+\infty}(\mathcal{X}\times\mathcal{Y})^n \to \mathcal{H}$, and a monotonically decreasing sequence $\epsilon(n)$, such that $\epsilon(n) \to 0$, as $n \to +\infty$,

$$\mathbb{E}_{S\sim D^n_{X_{\text{I}}Y_{\text{I}}}} \left[R_D^\alpha(\mathbf{A}(S)) - \inf_{h\in\mathcal{H}} R_D^\alpha(h)\right] \leq \epsilon(n), \quad \forall\alpha\in[0,1],\ \forall D_{XY}\in\mathscr{D}_{XY}.$$

$$\Downarrow$$

OOD detection is learnable in $\mathscr{D}_{XY}$ for $\mathcal{H}$.

If we set $\alpha = \pi^{\text{out}}$, then $\mathbb{E}_{S\sim D^n_{X_{\text{I}}Y_{\text{I}}}} R_D^\alpha(\mathbf{A}(S)) \leq \inf_{h\in\mathcal{H}} R_D^\alpha(h) + \epsilon(n)$ implies that

$$\mathbb{E}_{S\sim D^n_{X_{\text{I}}Y_{\text{I}}}} R_D(\mathbf{A}(S)) \leq \inf_{h\in\mathcal{H}} R_D(h) + \epsilon(n),$$

which means that OOD detection is learnable in $\mathscr{D}_{XY}$ for $\mathcal{H}$. We have completed this proof.

Proof of the Third Result.

The third result is a simple conclusion of the second result. Hence, we omit it. □

F Proof of Theorem 2

Before introducing the proof of Theorem 2, we extend Condition 1 to a general version (Condition 4). Then, Lemma 1 proves that Conditions 1 and 4 are the necessary conditions for the learnability of OOD detection. First, we provide the details of Condition 4.

Let $\Delta_l^{\mathrm{o}} = \{(\lambda_1, ..., \lambda_l) : \sum_{j=1}^{l} \lambda_j < 1 \text{ and } \lambda_j \geq 0, \forall j = 1, ..., l\}$, where l is a positive integer. Next, we introduce an important definition as follows:

Definition 6 (OOD Convex Decomposition and Convex Domain)**.** *Given any domain $D_{XY} \in \mathscr{D}_{XY}$, we say joint distributions $Q_1, ..., Q_l$, which are defined over $\mathcal{X} \times \{K+1\}$, are the OOD convex decomposition for D_{XY}, if*

$$D_{XY} = (1 - \sum_{j=1}^{l} \lambda_j) D_{X_{\mathrm{I}} Y_{\mathrm{I}}} + \sum_{j=1}^{l} \lambda_j Q_j,$$

for some $(\lambda_1, ..., \lambda_l) \in \Delta_l^{\mathrm{o}}$. We also say domain $D_{XY} \in \mathscr{D}_{XY}$ is an OOD convex domain corresponding to OOD convex decomposition $Q_1, ..., Q_l$, if for any $(\alpha_1, ..., \alpha_l) \in \Delta_l^{\mathrm{o}}$,

$$(1 - \sum_{j=1}^{l} \alpha_j) D_{X_{\mathrm{I}} Y_{\mathrm{I}}} + \sum_{j=1}^{l} \alpha_j Q_j \in \mathscr{D}_{XY}.$$

We extend the linear condition (Condition 1) to a multi-linear scenario.

Condition 4 (Multi-linear Condition)**.** *For each OOD convex domain $D_{XY} \in \mathscr{D}_{XY}$ corresponding to OOD convex decomposition $Q_1, ..., Q_l$, the following function*

$$f_{D,Q}(\alpha_1, ..., \alpha_l) := \inf_{h \in \mathcal{H}} \Big((1 - \sum_{j=1}^{l} \alpha_j) R_D^{\mathrm{in}}(h) + \sum_{j=1}^{l} \alpha_j R_{Q_j}(h) \Big), \quad \forall (\alpha_1, ..., \alpha_l) \in \Delta_l^{\mathrm{o}}$$

satisfies that

$$f_{D,Q}(\alpha_1, ..., \alpha_l) = (1 - \sum_{j=1}^{l} \alpha_j) f_{D,Q}(\mathbf{0}) + \sum_{j=1}^{l} \alpha_j f_{D,Q}(\boldsymbol{\alpha}_j),$$

where $\mathbf{0}$ is the $1 \times l$ vector, whose elements are 0, and $\boldsymbol{\alpha}_j$ is the $1 \times l$ vector, whose j-th element is 1 and other elements are 0.

When $l = 1$ and the domain space $\mathscr{D}_{XY}$ is a priori-unknown space, Condition 4 degenerates into Condition 1. Lemma 1 shows that Condition 4 is necessary for the learnability of OOD detection.

Lemma 1. *Given a priori-unknown space $\mathscr{D}_{XY}$ and a hypothesis space $\mathcal{H}$, if OOD detection is learnable in $\mathscr{D}_{XY}$ for $\mathcal{H}$, then Conditions 1 and 4 hold.*

Proof of Lemma 1.

Since Condition 1 is a special case of Condition 4, we only need to prove that Condition 4 holds.

For any OOD convex domain $D_{XY} \in \mathscr{D}_{XY}$ corresponding to OOD convex decomposition $Q_1, ..., Q_l$, and any $(\alpha_1, ..., \alpha_l) \in \Delta_l^{\mathrm{o}}$, we set

$$Q^{\boldsymbol{\alpha}} = \frac{1}{\sum_{i=1}^{l} \alpha_i} \sum_{j=1}^{l} \alpha_j Q_j.$$

Then, we define

$$D_{XY}^{\boldsymbol{\alpha}} = (1 - \sum_{i=1}^{l} \alpha_i) D_{X_{\mathrm{I}} Y_{\mathrm{I}}} + (\sum_{i=1}^{l} \alpha_i) Q^{\boldsymbol{\alpha}}, \text{ which belongs to } \mathscr{D}_{XY}.$$

Let

$$R_D^{\boldsymbol{\alpha}}(h) = \int_{\mathcal{X} \times \mathcal{Y}_{\mathrm{all}}} \ell(h(\mathbf{x}), y) \mathrm{d} D_{XY}^{\boldsymbol{\alpha}}(\mathbf{x}, y).$$

Since OOD detection is learnable in $\mathscr{D}_{XY}$ for $\mathcal{H}$, there exist an algorithm $\mathbf{A} : \cup_{n=1}^{+\infty}(\mathcal{X} \times \mathcal{Y})^n \to \mathcal{H}$, and a monotonically decreasing sequence $\epsilon(n)$, such that $\epsilon(n) \to 0$, as $n \to +\infty$, and

$$0 \leq \mathbb{E}_{S\sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} R^{\boldsymbol{\alpha}}_D(\mathbf{A}(S)) - \inf_{h\in\mathcal{H}} R^{\boldsymbol{\alpha}}_D(h) \leq \epsilon(n).$$

Note that

$$\mathbb{E}_{S\sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} R^{\boldsymbol{\alpha}}_D(\mathbf{A}(S)) = (1 - \sum_{j=1}^{l} \alpha_j)\mathbb{E}_{S\sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} R^{\mathrm{in}}_D(\mathbf{A}(S)) + \sum_{j=1}^{l} \alpha_j \mathbb{E}_{S\sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} R_{Q_j}(\mathbf{A}(S)),$$

and

$$\inf_{h\in\mathcal{H}} R^{\boldsymbol{\alpha}}_D(h) = f_{D,Q}(\alpha_1, ..., \alpha_l),$$

where

$$R_{Q_j}(\mathbf{A}(S)) = \int_{\mathcal{X}\times\{K+1\}} \ell(\mathbf{A}(S)(\mathbf{x}), y) \mathrm{d}Q_j(\mathbf{x}, y).$$

Therefore, we have that for any $(\alpha_1, ..., \alpha_l) \in \Delta^{\mathrm{o}}_l$,

$$\left|(1 - \sum_{j=1}^{l} \alpha_j)\mathbb{E}_{S\sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} R^{\mathrm{in}}_D(\mathbf{A}(S)) + \sum_{j=1}^{l} \alpha_j \mathbb{E}_{S\sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} R_{Q_j}(\mathbf{A}(S)) - f_{D,Q}(\alpha_1, ..., \alpha_l)\right| \leq \epsilon(n). \tag{17}$$

Let

$$g_n(\alpha_1, ..., \alpha_l) = (1 - \sum_{j=1}^{l} \alpha_j)\mathbb{E}_{S\sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} R^{\mathrm{in}}_D(\mathbf{A}(S)) + \sum_{j=1}^{l} \alpha_j \mathbb{E}_{S\sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} R_{Q_j}(\mathbf{A}(S)).$$

Note that Eq. (17) implies that

$$\begin{aligned} \lim_{n\to+\infty} g_n(\alpha_1, ..., \alpha_l) &= f_{D,Q}(\alpha_1, ..., \alpha_l), \ \ \forall(\alpha_1, ..., \alpha_l) \in \Delta^{\mathrm{o}}_l, \\ \lim_{n\to+\infty} g_n(\mathbf{0}) &= f_{D,Q}(\mathbf{0}). \end{aligned} \tag{18}$$

Step 1. Since $\boldsymbol{\alpha}_j \notin \Delta^{\mathrm{o}}_l$, we need to prove that

$$\lim_{n\to+\infty} \mathbb{E}_{S\sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} R_{Q_j}(\mathbf{A}(S)) = f(\boldsymbol{\alpha}_j), \textit{i.e.}, \lim_{n\to+\infty} g_n(\boldsymbol{\alpha}_j) = f(\boldsymbol{\alpha}_j), \tag{19}$$

where $\boldsymbol{\alpha}_j$ is the $1 \times l$ vector, whose j-th element is 1 and other elements are 0.

Let $\tilde{D}_{XY} = 0.5 * D_{X_{\mathrm{I}}Y_{\mathrm{I}}} + 0.5 * Q_j$. The second result of Theorem 1 implies that

$$\mathbb{E}_{S\sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} R^{\mathrm{out}}_{\tilde{D}}(\mathbf{A}(S)) \leq \inf_{h\in\mathcal{H}} R^{\mathrm{out}}_{\tilde{D}}(h) + \epsilon(n).$$

Since $R^{\mathrm{out}}_{\tilde{D}}(\mathbf{A}(S)) = R_{Q_j}(\mathbf{A}(S))$ and $R^{\mathrm{out}}_{\tilde{D}}(h) = R_{Q_j}(h)$,

$$\mathbb{E}_{S\sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} R_{Q_j}(\mathbf{A}(S)) \leq \inf_{h\in\mathcal{H}} R_{Q_j}(h) + \epsilon(n).$$

Note that $\inf_{h\in\mathcal{H}} R_{Q_j}(h) \leq \mathbb{E}_{S\sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} R_{Q_j}(\mathbf{A}(S))$. We have

$$0 \leq \mathbb{E}_{S\sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} R_{Q_j}(\mathbf{A}(S)) - \inf_{h\in\mathcal{H}} R_{Q_j}(h) \leq \epsilon(n). \tag{20}$$

Eq. (20) implies that

$$\lim_{n\to+\infty} \mathbb{E}_{S\sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} R_{Q_j}(\mathbf{A}(S)) = \inf_{h\in\mathcal{H}} R_{Q_j}(h). \tag{21}$$

We note that $\inf_{h\in\mathcal{H}} R_{Q_j}(h) = f_{D,Q}(\boldsymbol{\alpha}_j)$. Therefore,

$$\lim_{n\to+\infty} \mathbb{E}_{S\sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} R_{Q_j}(\mathbf{A}(S)) = f_{D,Q}(\boldsymbol{\alpha}_j), \ \textit{i.e.}, \lim_{n\to+\infty} g_n(\boldsymbol{\alpha}_j) = f(\boldsymbol{\alpha}_j). \tag{22}$$

Step 2. It is easy to check that for any $(\alpha_1, ..., \alpha_l) \in \Delta_l^{\mathrm{o}}$,

$$\begin{aligned}\lim_{n\to+\infty} g_n(\alpha_1, ..., \alpha_l) &= \lim_{n\to+\infty} \big((1-\sum_{j=1}^{l}\alpha_j) g_n(\mathbf{0}) + \sum_{j=1}^{l}\alpha_j g_n(\boldsymbol{\alpha}_j)\big) \\ &= (1-\sum_{j=1}^{l}\alpha_j) \lim_{n\to+\infty} g_n(\mathbf{0}) + \sum_{j=1}^{l}\alpha_j \lim_{n\to+\infty} g_n(\boldsymbol{\alpha}_j).\end{aligned} \tag{23}$$

According to Eq. (18) and Eq. (22), we have

$$\begin{aligned}&\lim_{n\to+\infty} g_n(\alpha_1, ..., \alpha_l) = f_{D,Q}(\alpha_1, ..., \alpha_l), \ \ \forall(\alpha_1, ..., \alpha_l) \in \Delta_l^{\mathrm{o}}, \\ &\lim_{n\to+\infty} g_n(\mathbf{0}) = f_{D,Q}(\mathbf{0}), \\ &\lim_{n\to+\infty} g_n(\boldsymbol{\alpha}_j) = f(\boldsymbol{\alpha}_j),\end{aligned} \tag{24}$$

Combining Eq. (24) with Eq. (23), we complete the proof. □

Lemma 2.

$$\inf_{h\in\mathcal{H}} R_D^{\alpha}(h) = (1-\alpha)\inf_{h\in\mathcal{H}} R_D^{\mathrm{in}}(h) + \alpha \inf_{h\in\mathcal{H}} R_D^{\mathrm{out}}(h), \ \forall\alpha\in[0,1),$$

if and only if *for any* $\epsilon > 0$,

$$\{h' \in \mathcal{H} : R_D^{\mathrm{in}}(h') \le \inf_{h\in\mathcal{H}} R_D^{\mathrm{in}}(h) + 2\epsilon\} \cap \{h' \in \mathcal{H} : R_D^{\mathrm{out}}(h') \le \inf_{h\in\mathcal{H}} R_D^{\mathrm{out}}(h) + 2\epsilon\} \neq \emptyset.$$

Proof of Lemma 2. For the sake of convenience, we set $f_D(\alpha) = \inf_{h\in\mathcal{H}} R_D^{\alpha}(h)$, for any $\alpha \in [0,1]$.

First, we prove that $f_D(\alpha) = (1-\alpha)f_D(0) + \alpha f_D(1), \ \forall\alpha\in[0,1)$ implies

$$\{h' \in \mathcal{H} : R_D^{\mathrm{in}}(h') \le \inf_{h\in\mathcal{H}} R_D^{\mathrm{in}}(h) + 2\epsilon\} \cap \{h' \in \mathcal{H} : R_D^{\mathrm{out}}(h') \le \inf_{h\in\mathcal{H}} R_D^{\mathrm{out}}(h) + 2\epsilon\} \neq \emptyset.$$

For any $\epsilon > 0$ and $0 \le \alpha < 1$, we can find $h_\epsilon^\alpha \in \mathcal{H}$ satisfying that

$$R_D^{\alpha}(h_\epsilon^\alpha) \le \inf_{h\in\mathcal{H}} R_D^{\alpha}(h) + \epsilon.$$

Note that

$$\inf_{h\in\mathcal{H}} R_D^{\alpha}(h) = \inf_{h\in\mathcal{H}} \Big((1-\alpha)R_D^{\mathrm{in}}(h) + \alpha R_D^{\mathrm{out}}(h)\Big) \ge (1-\alpha)\inf_{h\in\mathcal{H}} R_D^{\mathrm{in}}(h) + \alpha\inf_{h\in\mathcal{H}} R_D^{\mathrm{out}}(h).$$

Therefore,

$$(1-\alpha)\inf_{h\in\mathcal{H}} R_D^{\mathrm{in}}(h) + \alpha\inf_{h\in\mathcal{H}} R_D^{\mathrm{out}}(h) \le \inf_{h\in\mathcal{H}} R_D^{\alpha}(h) \le R_D^{\alpha}(h_\epsilon^\alpha) \le \inf_{h\in\mathcal{H}} R_D^{\alpha}(h) + \epsilon. \tag{25}$$

Note that $f_D(\alpha) = (1-\alpha)f_D(0) + \alpha f_D(1), \forall\alpha\in[0,1)$, *i.e.*,

$$\inf_{h\in\mathcal{H}} R_D^{\alpha}(h) = (1-\alpha)\inf_{h\in\mathcal{H}} R_D^{\mathrm{in}}(h) + \alpha\inf_{h\in\mathcal{H}} R_D^{\mathrm{out}}(h), \forall\alpha\in[0,1). \tag{26}$$

Using Eqs. (25) and (26), we have that for any $0 \le \alpha < 1$,

$$\epsilon \ge \big|R_D^{\alpha}(h_\epsilon^\alpha) - \inf_{h\in\mathcal{H}} R_D^{\alpha}(h)\big| = \big|(1-\alpha)\big(R_D^{\mathrm{in}}(h_\epsilon^\alpha) - \inf_{h\in\mathcal{H}} R_D^{\mathrm{in}}(h)\big) + \alpha\big(R_D^{\mathrm{out}}(h_\epsilon^\alpha) - \inf_{h\in\mathcal{H}} R_D^{\mathrm{out}}(h)\big)\big|. \tag{27}$$

Since $R_D^{\mathrm{out}}(h_\epsilon^\alpha) - \inf_{h\in\mathcal{H}} R_D^{\mathrm{out}}(h) \ge 0$ and $R_D^{\mathrm{in}}(h_\epsilon^\alpha) - \inf_{h\in\mathcal{H}} R_D^{\mathrm{in}}(h) \ge 0$, Eq. (27) implies that: for any $0 < \alpha < 1$,

$$\begin{aligned}&R_D^{\mathrm{in}}(h_\epsilon^\alpha) \le \inf_{h\in\mathcal{H}} R_D^{\mathrm{in}}(h) + \epsilon/(1-\alpha), \\ &R_D^{\mathrm{out}}(h_\epsilon^\alpha) \le \inf_{h\in\mathcal{H}} R_D^{\mathrm{out}}(h) + \epsilon/\alpha.\end{aligned}$$

Therefore,

$$h_\epsilon^\alpha \in \{h' \in \mathcal{H} : R_D^{\mathrm{in}}(h') \le \inf_{h\in\mathcal{H}} R_D^{\mathrm{in}}(h) + \epsilon/(1-\alpha)\} \cap \{h' \in \mathcal{H} : R_D^{\mathrm{out}}(h') \le \inf_{h\in\mathcal{H}} R_D^{\mathrm{out}}(h) + \epsilon/\alpha\}.$$

If we set $\alpha = 0.5$, we obtain that for any $\epsilon > 0$,

$$\{h' \in \mathcal{H} : R_D^{\text{in}}(h') \leq \inf_{h\in\mathcal{H}} R_D^{\text{in}}(h) + 2\epsilon\} \cap \{h' \in \mathcal{H} : R_D^{\text{out}}(h') \leq \inf_{h\in\mathcal{H}} R_D^{\text{out}}(h) + 2\epsilon\} \neq \emptyset.$$

Second, we prove that for any $\epsilon > 0$, if

$$\{h' \in \mathcal{H} : R_D^{\text{in}}(h') \leq \inf_{h\in\mathcal{H}} R_D^{\text{in}}(h) + 2\epsilon\} \cap \{h' \in \mathcal{H} : R_D^{\text{out}}(h') \leq \inf_{h\in\mathcal{H}} R_D^{\text{out}}(h) + 2\epsilon\} \neq \emptyset,$$

then $f_D(\alpha) = (1-\alpha)f_D(0) + \alpha f_D(1)$, for any $\alpha \in [0, 1)$.

Let $h_\epsilon \in \{h' \in \mathcal{H} : R_D^{\text{in}}(h') \leq \inf_{h\in\mathcal{H}} R_D^{\text{in}}(h) + 2\epsilon\} \cap \{h' \in \mathcal{H} : R_D^{\text{out}}(h') \leq \inf_{h\in\mathcal{H}} R_D^{\text{out}}(h) + 2\epsilon\}$. Then,

$$\inf_{h\in\mathcal{H}} R_D^{\alpha}(h) \leq R_D^{\alpha}(h_\epsilon) \leq (1-\alpha) \inf_{h\in\mathcal{H}} R_D^{\text{in}}(h) + \alpha \inf_{h\in\mathcal{H}} R_D^{\text{out}}(h) + 2\epsilon \leq \inf_{h\in\mathcal{H}} R_D^{\alpha}(h) + 2\epsilon,$$

which implies that $|f_D(\alpha) - (1-\alpha)f_D(0) - \alpha f_D(1)| \leq 2\epsilon$.

As $\epsilon \to 0$, $|f_D(\alpha) - (1-\alpha)f_D(0) - \alpha f_D(1)| \leq 0$. We have completed the proof. □

Theorem 2. *Given a hypothesis space $\mathcal{H}$ and a domain D_{XY}, OOD detection is learnable in the single-distribution space $\mathscr{D}_{XY}^{D_{XY}}$ for $\mathcal{H}$* ***if and only if*** *linear condition (i.e., Condition 1) holds.*

Proof of Theorem 2. Based on Lemma 1, we obtain that Condition 1 is the necessary condition for the learnability of OOD detection in the single-distribution space $\mathscr{D}_{XY}^{D_{XY}}$. Next, it suffices to prove that Condition 1 is the sufficient condition for the learnability of OOD detection in the single-distribution space $\mathscr{D}_{XY}^{D_{XY}}$. We use Lemma 2 to prove the sufficient condition.

Let $\mathscr{F}$ be the infinite sequence set that consists of all infinite sequences, whose coordinates are hypothesis functions, *i.e.*,

$$\mathscr{F} = \{\boldsymbol{h} = (h_1, ..., h_n, ...) : \forall h_n \in \mathcal{H}, n = 1, ..., +\infty\}.$$

For each $\boldsymbol{h} \in \mathscr{F}$, there is a corresponding algorithm $\mathbf{A}_{\boldsymbol{h}}$[6]: $\mathbf{A}_{\boldsymbol{h}}(S) = h_n$, if $|S| = n$. $\mathscr{F}$ generates an algorithm class $\mathscr{A} = \{\mathbf{A}_{\boldsymbol{h}} : \forall \boldsymbol{h} \in \mathscr{F}\}$. We select a consistent algorithm from the algorithm class $\mathscr{A}$.

We construct a special infinite sequence $\tilde{\boldsymbol{h}} = (\tilde{h}_1, ..., \tilde{h}_n, ...) \in \mathscr{F}$. For each positive integer n, we select $\tilde{h}_n$ from $\{h' \in \mathcal{H} : R_D^{\text{in}}(h') \leq \inf_{h\in\mathcal{H}} R_D^{\text{in}}(h) + 2/n\} \cap \{h' \in \mathcal{H} : R_D^{\text{out}}(h') \leq \inf_{h\in\mathcal{H}} R_D^{\text{out}}(h) + 2/n\}$ (the existence of $\tilde{h}_n$ is based on Lemma 2). It is easy to check that

$$\mathbb{E}_{S\sim D_{X_1Y_1}^n} R_D^{\text{in}}(\mathbf{A}_{\tilde{\boldsymbol{h}}}(S)) \leq \inf_{h\in\mathcal{H}} R_D^{\text{in}}(h) + 2/n.$$
$$\mathbb{E}_{S\sim D_{X_1Y_1}^n} R_D^{\text{out}}(\mathbf{A}_{\tilde{\boldsymbol{h}}}(S)) \leq \inf_{h\in\mathcal{H}} R_D^{\text{out}}(h) + 2/n.$$

Since $(1-\alpha)\inf_{h\in\mathcal{H}} R_D^{\text{in}}(h) + \alpha \inf_{h\in\mathcal{H}} R_D^{\text{out}}(h) \leq \inf_{h\in\mathcal{H}} R_D^{\alpha}(h)$, we obtain that for any $\alpha \in [0, 1]$,

$$\mathbb{E}_{S\sim D_{X_1Y_1}^n} R_D^{\alpha}(\mathbf{A}_{\tilde{\boldsymbol{h}}}(S)) \leq \inf_{h\in\mathcal{H}} R_D^{\alpha}(h) + 2/n.$$

We have completed this proof. □

G Proofs of Theorem 3 and Theorem 4

G.1 Proof of Theorem 3

Theorem 3. *Given a hypothesis space $\mathcal{H}$ and a prior-unknown space $\mathscr{D}_{XY}$, if there is $D_{XY} \in \mathscr{D}_{XY}$, which has overlap between ID and OOD, and $\inf_{h\in\mathcal{H}} R_D^{\text{in}}(h) = 0$ and $\inf_{h\in\mathcal{H}} R_D^{\text{out}}(h) = 0$, then Condition 1 does not hold. Therefore, OOD detection is not learnable in $\mathscr{D}_{XY}$ for $\mathcal{H}$.*

[6]In this paper, we regard an algorithm as a mapping from $\cup_{n=1}^{+\infty}(\mathcal{X} \times \mathcal{Y})^n$ to $\mathcal{H}$. So we can design an algorithm like this.

Proof of Theorem 3. We **first** explain how we get f_{I} and f_{O} in Definition 4. Since D_X is absolutely continuous respect to μ ($D_X \ll \mu$), then $D_{X_{\mathrm{I}}} \ll \mu$ and $D_{X_{\mathrm{O}}} \ll \mu$. By Radon-Nikodym Theorem [36], we know there exist two non-negative functions defined over $\mathcal{X}$: f_{I} and f_{O} such that for any μ-measurable set $A \subset \mathcal{X}$,

$$D_{X_{\mathrm{I}}}(A) = \int_A f_{\mathrm{I}}(\mathbf{x})\mathrm{d}\mu(\mathbf{x}), \quad D_{X_{\mathrm{O}}}(A) = \int_A f_{\mathrm{O}}(\mathbf{x})\mathrm{d}\mu(\mathbf{x}).$$

Second, we prove that for any $\alpha \in (0,1)$, $\inf_{h\in\mathcal{H}} R_D^{\alpha}(h) > 0$.

We define $A_m = \{\mathbf{x} \in \mathcal{X} : f_{\mathrm{I}}(\mathbf{x}) \geq \frac{1}{m} \text{ and } f_{\mathrm{O}}(\mathbf{x}) \geq \frac{1}{m}\}$. It is clear that

$$\cup_{m=1}^{+\infty} A_m = \{\mathbf{x} \in \mathcal{X} : f_{\mathrm{I}}(\mathbf{x}) > 0 \text{ and } f_{\mathrm{O}}(\mathbf{x}) > 0\} = A_{\text{overlap}},$$

and

$$A_m \subset A_{m+1}.$$

Therefore,

$$\lim_{m\to+\infty} \mu(A_m) = \mu(A_{\text{overlap}}) > 0,$$

which implies that there exists m_0 such that

$$\mu(A_{m_0}) > 0.$$

For any $\alpha \in (0,1)$, we define $c_\alpha = \min_{y_1\in\mathcal{Y}_{\text{all}}} \big((1-\alpha)\min_{y_2\in\mathcal{Y}} \ell(y_1,y_2) + \alpha\ell(y_1,K+1)\big)$. It is clear that $c_\alpha > 0$ for $\alpha \in (0,1)$. Then, for any $h \in \mathcal{H}$,

$$\begin{aligned}
&R_D^{\alpha}(h)\\
=&\int_{\mathcal{X}\times\mathcal{Y}_{\text{all}}} \ell(h(\mathbf{x}),y)\mathrm{d}D_{XY}^{\alpha}(\mathbf{x},y)\\
=&\int_{\mathcal{X}\times\mathcal{Y}} (1-\alpha)\ell(h(\mathbf{x}),y)\mathrm{d}D_{X_{\mathrm{I}}Y_{\mathrm{I}}}(\mathbf{x},y) + \int_{\mathcal{X}\times\{K+1\}} \alpha\ell(h(\mathbf{x}),y)\mathrm{d}D_{X_{\mathrm{O}}Y_{\mathrm{O}}}(\mathbf{x},y)\\
\geq&\int_{A_{m_0}\times\mathcal{Y}} (1-\alpha)\ell(h(\mathbf{x}),y)\mathrm{d}D_{X_{\mathrm{I}}Y_{\mathrm{I}}}(\mathbf{x},y) + \int_{A_{m_0}\times\{K+1\}} \alpha\ell(h(\mathbf{x}),y)\mathrm{d}D_{X_{\mathrm{O}}Y_{\mathrm{O}}}(\mathbf{x},y)\\
=&\int_{A_{m_0}} \big((1-\alpha)\int_{\mathcal{Y}} \ell(h(\mathbf{x}),y)\mathrm{d}D_{Y_{\mathrm{I}}|X_{\mathrm{I}}}(y|\mathbf{x})\big)\mathrm{d}D_{X_{\mathrm{I}}}(\mathbf{x})\\
+&\int_{A_{m_0}} \alpha\ell(h(\mathbf{x}),K+1)\mathrm{d}D_{X_{\mathrm{O}}}(\mathbf{x})\\
\geq&\int_{A_{m_0}} (1-\alpha)\min_{y_2\in\mathcal{Y}} \ell(h(\mathbf{x}),y_2)\mathrm{d}D_{X_{\mathrm{I}}}(\mathbf{x}) + \int_{A_{m_0}} \alpha\ell(h(\mathbf{x}),K+1)\mathrm{d}D_{X_{\mathrm{O}}}(\mathbf{x})\\
\geq&\int_{A_{m_0}} (1-\alpha)\min_{y_2\in\mathcal{Y}} \ell(h(\mathbf{x}),y_2)f_{\mathrm{I}}(\mathbf{x})\mathrm{d}\mu(\mathbf{x}) + \int_{A_{m_0}} \alpha\ell(h(\mathbf{x}),K+1)f_{\mathrm{O}}(\mathbf{x})\mathrm{d}\mu(\mathbf{x})\\
\geq&\frac{1}{m_0}\int_{A_{m_0}} (1-\alpha)\min_{y_2\in\mathcal{Y}} \ell(h(\mathbf{x}),y_2)\mathrm{d}\mu(\mathbf{x}) + \frac{1}{m_0}\int_{A_{m_0}} \alpha\ell(h(\mathbf{x}),K+1)\mathrm{d}\mu(\mathbf{x})\\
=&\frac{1}{m_0}\int_{A_{m_0}} \big((1-\alpha)\min_{y_2\in\mathcal{Y}} \ell(h(\mathbf{x}),y_2) + \alpha\ell(h(\mathbf{x}),K+1)\big)\mathrm{d}\mu(\mathbf{x}) \geq \frac{c_\alpha}{m_0}\mu(A_{m_0}) > 0.
\end{aligned}$$

Therefore,

$$\inf_{h\in\mathcal{H}} R_D^{\alpha}(h) \geq \frac{c_\alpha}{m_0}\mu(A_{m_0}) > 0.$$

Third, Condition 1 indicates that $\inf_{h\in\mathcal{H}} R_D^{\alpha}(h) = (1-\alpha)\inf_{h\in\mathcal{H}} R_D^{\text{in}}(h) + \alpha\inf_{h\in\mathcal{H}} R_D^{\text{in}}(h) = 0$ (here we have used conditions $\inf_{h\in\mathcal{H}} R_D^{\text{in}}(h) = 0$ and $\inf_{h\in\mathcal{H}} R_D^{\text{out}}(h) = 0$), which contradicts with $\inf_{h\in\mathcal{H}} R_D^{\alpha}(h) > 0$ ($\alpha \in (0,1)$). Therefore, Condition 1 does not hold. Using Lemma 1, we obtain that OOD detection in $\mathscr{D}_{XY}$ is not learnable for $\mathcal{H}$. □

G.2 Proof of Theorem 4

Theorem 4 (Impossibility Theorem for Total Space)**.** *OOD detection is not learnable in the total space $\mathscr{D}_{XY}^{\text{all}}$ for $\mathcal{H}$, if $|\phi \circ \mathcal{H}| > 1$, where ϕ maps ID labels to 1 and maps OOD labels to 2.*

Proof of Theorem 4. We need to prove that OOD detection is not learnable in the total space $\mathscr{D}_{XY}^{\text{all}}$ for $\mathcal{H}$, if $\mathcal{H}$ is non-trivial, *i.e.*, $\{\mathbf{x} \in \mathcal{X} : \exists h_1, h_2 \in \mathcal{H}, \text{s.t. } h_1(\mathbf{x}) \in \mathcal{Y}, h_2(\mathbf{x}) = K+1\} \neq \emptyset$.

The main idea is to construct a domain D_{XY} satisfying that:
1) the ID and OOD distributions have overlap (Definition 4); and 2) $R_D^{\text{in}}(h_1) = 0$, $R_D^{\text{out}}(h_2) = 0$.

According to the condition that $\mathcal{H}$ is non-trivial, we know that there exist $h_1, h_2 \in \mathcal{H}$ such that $h_1(\mathbf{x}_1) \in \mathcal{Y}, h_2(\mathbf{x}_1) = K+1$, for some $\mathbf{x}_1 \in \mathcal{X}$. We set $D_{XY} = 0.5 * \delta_{(\mathbf{x}_1, h_1(\mathbf{x}_1))} + 0.5 * \delta_{(\mathbf{x}_1, h_2(\mathbf{x}_1))}$, where δ is the Dirac measure. It is easy to check that $R_D^{\text{in}}(h_1) = 0$, $R_D^{\text{out}}(h_2) = 0$, which implies that $\inf_{h \in \mathcal{H}} R_D^{\text{in}}(h) = 0$ and $\inf_{h \in \mathcal{H}} R_D^{\text{out}}(h) = 0$. In addition, the ID distribution $\delta_{(\mathbf{x}_1, h_1(\mathbf{x}_1))}$ and OOD distribution $\delta_{(\mathbf{x}_1, h_2(\mathbf{x}_1))}$ have overlap $\mathbf{x}_1$. By using Theorem 3, we have completed this proof. □

H Proof of Theorem 5

Before proving Theorem 5, we need three important lemmas.

Lemma 3. *Suppose that D_{XY} is a domain with OOD convex decomposition $Q_1, ..., Q_l$ (convex decomposition is given by Definition 6 in Appendix F), and D_{XY} is a finite discrete distribution, then (the definition of $f_{D,Q}$ is given in Condition 4)*

$$f_{D,Q}(\alpha_1, ..., \alpha_l) = (1 - \sum_{j=1}^{l} \alpha_j) f_{D,Q}(\mathbf{0}) + \sum_{j=1}^{l} \alpha_j f_{D,Q}(\boldsymbol{\alpha}_j), \quad \forall (\alpha_1, ..., \alpha_l) \in \Delta_l^{\text{o}},$$

if and only if

$$\arg\min_{h \in \mathcal{H}} R_D(h) = \bigcap_{j=1}^{l} \arg\min_{h \in \mathcal{H}} R_{Q_j}(h) \bigcap \arg\min_{h \in \mathcal{H}} R_D^{\text{in}}(h),$$

where $\mathbf{0}$ is the $1 \times l$ vector, whose elements are 0, and $\boldsymbol{\alpha}_j$ is the $1 \times l$ vector, whose j-th element is 1 and other elements are 0, and

$$R_{Q_j}(h) = \int_{\mathcal{X} \times \{K+1\}} \ell(h(\mathbf{x}), y) \mathrm{d}Q_j(\mathbf{x}, y).$$

Proof of Lemma 3. To better understand this proof, we recall the definition of $f_{D,Q}(\alpha_1, ..., \alpha_l)$:

$$f_{D,Q}(\alpha_1, ..., \alpha_l) = \inf_{h \in \mathcal{H}} \Big((1 - \sum_{j=1}^{l} \alpha_j) R_D^{\text{in}}(h) + \sum_{j=1}^{l} \alpha_j R_{Q_j}(h) \Big), \quad \forall (\alpha_1, ..., \alpha_l) \in \Delta_l^{\text{o}}$$

First, we prove that if

$$f_{D,Q}(\alpha_1, ..., \alpha_l) = (1 - \sum_{j=1}^{l} \alpha_j) f_{D,Q}(\mathbf{0}) + \sum_{j=1}^{l} \alpha_j f_{D,Q}(\boldsymbol{\alpha}_j), \quad \forall (\alpha_1, ..., \alpha_l) \in \Delta_l^{\text{o}},$$

then,

$$\arg\min_{h \in \mathcal{H}} R_D(h) = \bigcap_{j=1}^{l} \arg\min_{h \in \mathcal{H}} R_{Q_j}(h) \bigcap \arg\min_{h \in \mathcal{H}} R_D^{\text{in}}(h).$$

Let $D_{XY} = (1 - \sum_{j=1}^{l} \lambda_j) D_{X_1 Y_1} + \sum_{j=1}^{l} \lambda_j Q_j$, for some $(\lambda_1, ..., \lambda_l) \in \Delta_l^{\text{o}}$. Since D_{XY} has finite support set, we have

$$\arg\min_{h \in \mathcal{H}} R_D(h) = \arg\min_{h \in \mathcal{H}} \Big((1 - \sum_{j=1}^{l} \lambda_j) R_D^{\text{in}}(h) + \sum_{j=1}^{l} \lambda_j R_{Q_j}(h) \Big) \neq \emptyset.$$

We can find that $h_0 \in \arg\min_{h\in\mathcal{H}} \Big((1-\sum_{j=1}^{l}\lambda_j)R_D^{\text{in}}(h)+\sum_{j=1}^{l}\lambda_j R_{Q_j}(h)\Big)$. Hence,

$$(1-\sum_{j=1}^{l}\lambda_j)R_D^{\text{in}}(h_0)+\sum_{j=1}^{l}\lambda_j R_{Q_j}(h_0)=\inf_{h\in\mathcal{H}}\Big((1-\sum_{j=1}^{l}\lambda_j)R_D^{\text{in}}(h)+\sum_{j=1}^{l}\lambda_j R_{Q_j}(h)\Big). \quad (28)$$

Note that the condition $f_{D,Q}(\alpha_1,...,\alpha_l)=(1-\sum_{j=1}^{l}\alpha_j)f_{D,Q}(\mathbf{0})+\sum_{j=1}^{l}\alpha_j f_{D,Q}(\boldsymbol{\alpha}_j)$ implies

$$(1-\sum_{j=1}^{l}\lambda_j)\inf_{h\in\mathcal{H}}R_D^{\text{in}}(h)+\sum_{j=1}^{l}\lambda_j\inf_{h\in\mathcal{H}}R_{Q_j}(h)=\inf_{h\in\mathcal{H}}\Big((1-\sum_{j=1}^{l}\lambda_j)R_D^{\text{in}}(h)+\sum_{j=1}^{l}\lambda_j R_{Q_j}(h)\Big). \quad (29)$$

Therefore, Eq. (28) and Eq. (29) imply that

$$(1-\sum_{j=1}^{l}\lambda_j)\inf_{h\in\mathcal{H}}R_D^{\text{in}}(h)+\sum_{j=1}^{l}\lambda_j\inf_{h\in\mathcal{H}}R_{Q_j}(h)=(1-\sum_{j=1}^{l}\lambda_j)R_D^{\text{in}}(h_0)+\sum_{j=1}^{l}\lambda_j R_{Q_j}(h_0). \quad (30)$$

Since $R_D^{\text{in}}(h_0)\geq\inf_{h\in\mathcal{H}}R_D^{\text{in}}(h)$ and $R_{Q_j}(h_0)\geq\inf_{h\in\mathcal{H}}R_{Q_j}^{\text{in}}(h)$, for $j=1,...,l$, then using Eq. (30), we have that

$$\begin{aligned}R_D^{\text{in}}(h_0)&=\inf_{h\in\mathcal{H}}R_D^{\text{in}}(h),\\ R_{Q_j}(h_0)&=\inf_{h\in\mathcal{H}}R_{Q_j}(h),\quad\forall j=1,...,l,\end{aligned}$$

which implies that

$$h_0\in\bigcap_{j=1}^{l}\arg\min_{h\in\mathcal{H}}R_{Q_j}(h)\bigcap\arg\min_{h\in\mathcal{H}}R_D^{\text{in}}(h).$$

Therefore,

$$\arg\min_{h\in\mathcal{H}}R_D(h)\subset\bigcap_{j=1}^{l}\arg\min_{h\in\mathcal{H}}R_{Q_j}(h)\bigcap\arg\min_{h\in\mathcal{H}}R_D^{\text{in}}(h). \quad (31)$$

Additionally, using

$$f_{D,Q}(\alpha_1,...,\alpha_l)=(1-\sum_{j=1}^{l}\alpha_j)f_{D,Q}(\mathbf{0})+\sum_{j=1}^{l}\alpha_j f_{D,Q}(\boldsymbol{\alpha}_j),\ \forall(\alpha_1,...,\alpha_l)\in\Delta_l^{\text{o}},$$

we obtain that for any $h'\in\bigcap_{j=1}^{l}\arg\min_{h\in\mathcal{H}}R_{Q_j}(h)\bigcap\arg\min_{h\in\mathcal{H}}R_D^{\text{in}}(h)$,

$$\begin{aligned}\inf_{h\in\mathcal{H}}R_D(h)&=\inf_{h\in\mathcal{H}}\Big((1-\sum_{j=1}^{l}\lambda_j)R_D^{\text{in}}(h)+\sum_{j=1}^{l}\lambda_j R_{Q_j}(h)\Big)\\ &=(1-\sum_{j=1}^{l}\lambda_j)\inf_{h\in\mathcal{H}}R_D^{\text{in}}(h)+\sum_{j=1}^{l}\lambda_j\inf_{h\in\mathcal{H}}R_{Q_j}(h)\\ &=(1-\sum_{j=1}^{l}\lambda_j)R_D^{\text{in}}(h')+\sum_{j=1}^{l}\lambda_j R_{Q_j}(h')=R_D(h'),\end{aligned}$$

which implies that

$$h'\in\arg\min_{h\in\mathcal{H}}R_D(h).$$

Therefore,

$$\bigcap_{j=1}^{l}\arg\min_{h\in\mathcal{H}}R_{Q_j}(h)\bigcap\arg\min_{h\in\mathcal{H}}R_D^{\text{in}}(h)\subset\arg\min_{h\in\mathcal{H}}R_D(h). \quad (32)$$

Combining Eq. (31) with Eq. (32), we obtain that

$$\bigcap_{j=1}^{l} \arg\min_{h\in\mathcal{H}} R_{Q_j}(h) \bigcap \arg\min_{h\in\mathcal{H}} R_D^{\text{in}}(h) = \arg\min_{h\in\mathcal{H}} R_D(h).$$

Second, we prove that if

$$\arg\min_{h\in\mathcal{H}} R_D(h) = \bigcap_{j=1}^{l} \arg\min_{h\in\mathcal{H}} R_{Q_j}(h) \bigcap \arg\min_{h\in\mathcal{H}} R_D^{\text{in}}(h),$$

then,

$$f_{D,Q}(\alpha_1, ..., \alpha_l) = (1 - \sum_{j=1}^{l} \alpha_j) f_{D,Q}(\mathbf{0}) + \sum_{j=1}^{l} \alpha_j f_{D,Q}(\boldsymbol{\alpha}_j), \quad \forall (\alpha_1, ..., \alpha_l) \in \Delta_l^{\text{o}}.$$

We set

$$h_0 \in \bigcap_{j=1}^{l} \arg\min_{h\in\mathcal{H}} R_{Q_j}(h) \bigcap \arg\min_{h\in\mathcal{H}} R_D^{\text{in}}(h),$$

then, for any $(\alpha_1, ..., \alpha_l) \in \Delta_l^{\text{o}}$,

$$\begin{aligned}
(1 - \sum_{j=1}^{l} \alpha_j) \inf_{h\in\mathcal{H}} R_D^{\text{in}}(h) + \sum_{j=1}^{l} \alpha_j \inf_{h\in\mathcal{H}} R_{Q_j}(h) &\leq \inf_{h\in\mathcal{H}} \Big((1 - \sum_{j=1}^{l} \alpha_j) R_D^{\text{in}}(h) + \sum_{j=1}^{l} \alpha_j R_{Q_j}(h) \Big) \\
&\leq (1 - \sum_{j=1}^{l} \alpha_j) R_D^{\text{in}}(h_0) + \sum_{j=1}^{l} \alpha_j R_{Q_j}(h_0) \\
&= (1 - \sum_{j=1}^{l} \alpha_j) \inf_{h\in\mathcal{H}} R_D^{\text{in}}(h) + \sum_{j=1}^{l} \alpha_j \inf_{h\in\mathcal{H}} R_{Q_j}(h).
\end{aligned}$$

Therefore, for any $(\alpha_1, ..., \alpha_l) \in \Delta_l^{\text{o}}$,

$$(1 - \sum_{j=1}^{l} \alpha_j) \inf_{h\in\mathcal{H}} R_D^{\text{in}}(h) + \sum_{j=1}^{l} \alpha_j \inf_{h\in\mathcal{H}} R_{Q_j}(h) = \inf_{h\in\mathcal{H}} \Big((1 - \sum_{j=1}^{l} \alpha_j) R_D^{\text{in}}(h) + \sum_{j=1}^{l} \alpha_j R_{Q_j}(h) \Big),$$

which implies that: for any $(\alpha_1, ..., \alpha_l) \in \Delta_l^{\text{o}}$,

$$f_{D,Q}(\alpha_1, ..., \alpha_l) = (1 - \sum_{j=1}^{l} \alpha_j) f_{D,Q}(\mathbf{0}) + \sum_{j=1}^{l} \alpha_j f_{D,Q}(\boldsymbol{\alpha}_j).$$

We have completed this proof. □

Lemma 4. *Suppose that Assumption 1 holds. If there is a finite discrete domain $D_{XY} \in \mathscr{D}^s_{XY}$ such that $\inf_{h\in\mathcal{H}} R^{\text{out}}_D(\boldsymbol{h}) > 0$, then OOD detection is not learnable in $\mathscr{D}^s_{XY}$ for $\mathcal{H}$.*

Proof of Lemma 4. Suppose that $\text{supp}D_{X_{\text{O}}} = \{\mathbf{x}^{\text{out}}_1, ..., \mathbf{x}^{\text{out}}_l\}$, then it is clear that D_{XY} has OOD convex decomposition $\delta_{\mathbf{x}^{\text{out}}_1}, ..., \delta_{\mathbf{x}^{\text{out}}_l}$, where $\delta_{\mathbf{x}}$ is the dirac measure whose support set is $\{\mathbf{x}\}$.

Since $\mathcal{H}$ is the separate space for OOD (*i.e.*, Assumption 1 holds), then $\forall j = 1, ..., l$,

$$\inf_{h\in\mathcal{H}} R_{\delta_{\mathbf{x}^{\text{out}}_j}}(h) = 0,$$

where

$$R_{\delta_{\mathbf{x}^{\text{out}}_j}}(h) = \int_{\mathcal{X}} \ell(h(\mathbf{x}), K+1)\mathrm{d}\delta_{\mathbf{x}^{\text{out}}_j}(\mathbf{x}).$$

This implies that: if $\bigcap_{j=1}^{l} \arg\min_{h\in\mathcal{H}} R_{\delta_{\mathbf{x}^{\text{out}}_j}}(h) \neq \emptyset$, then for $\forall h' \in \bigcap_{j=1}^{l} \arg\min_{h\in\mathcal{H}} R_{\delta_{\mathbf{x}^{\text{out}}_j}}(h)$,

$$h'(\mathbf{x}^{\text{out}}_i) = K+1, \ \forall i = 1, ..., l.$$

Therefore, if $\bigcap_{j=1}^{l} \arg\min_{h\in\mathcal{H}} R_{\delta_{\mathbf{x}^{\text{out}}_j}}(h) \bigcap \arg\min_{h\in\mathcal{H}} R^{\text{in}}_D(h) \neq \emptyset$,

then for any $h^* \in \bigcap_{j=1}^{l} \arg\min_{h\in\mathcal{H}} R_{\delta_{\mathbf{x}^{\text{out}}_j}}(h) \bigcap \arg\min_{h\in\mathcal{H}} R^{\text{in}}_D(h)$, we have that

$$h^*(\mathbf{x}^{\text{out}}_i) = K+1, \ \forall i = 1, ..., l.$$

Proof by Contradiction: assume OOD detection is learnable in $\mathscr{D}^s_{XY}$ for $\mathcal{H}$, then Lemmas 1 and 3 imply that

$$\bigcap_{j=1}^{l} \arg\min_{h\in\mathcal{H}} R_{\delta_{\mathbf{x}^{\text{out}}_j}}(h) \bigcap \arg\min_{h\in\mathcal{H}} R^{\text{in}}_D(h) = \arg\min_{h\in\mathcal{H}} R_D(h) \neq \emptyset.$$

Therefore, for any $h^* \in \arg\min_{h\in\mathcal{H}} R_D(h)$, we have that

$$h^*(\mathbf{x}^{\text{out}}_i) = K+1, \ \forall i = 1, ..., l,$$

which implies that for any $h^* \in \arg\min_{h\in\mathcal{H}} R_D(h)$, we have $R^{\text{out}}_D(h^*) = 0$, which implies that $\inf_{h\in\mathcal{H}} R^{\text{out}}_D(h) = 0$.

It is clear that $\inf_{h\in\mathcal{H}} R^{\text{out}}_D(h) = 0$ is **inconsistent** with the condition $\inf_{h\in\mathcal{H}} R^{\text{out}}_D(h) > 0$. Therefore, OOD detection is not learnable in $\mathscr{D}^s_{XY}$ for $\mathcal{H}$. □

Lemma 5. *If Assumption 1 holds, $\text{VCdim}(\phi \circ \mathcal{H}) = v < +\infty$ and $\sup_{h\in\mathcal{H}} |\{\mathbf{x} \in \mathcal{X} : h(\mathbf{x}) \in \mathcal{Y}\}| > m$ such that $v < m$, then OOD detection is not learnable in $\mathscr{D}^s_{XY}$ for $\mathcal{H}$, where ϕ maps ID's labels to 1 and maps OOD's labels to 2.*

Proof of Lemma 5. Due to $\sup_{h\in\mathcal{H}} |\{\mathbf{x} \in \mathcal{X} : \boldsymbol{h}(\mathbf{x}) \in \mathcal{Y}\}| > m$, we can obtain a set

$$C = \{\mathbf{x}_1, ..., \mathbf{x}_m, \mathbf{x}_{m+1}\},$$

which satisfies that there exists $\tilde{h} \in \mathcal{H}$ such that $\tilde{h}(\mathbf{x}_i) \in \mathcal{Y}$ for any $i = 1, ..., m, m+1$.

Let $\mathcal{H}^{\phi}_C = \{(\phi \circ h(\mathbf{x}_1), ..., \phi \circ h(\mathbf{x}_m), \phi \circ h(\mathbf{x}_{m+1}) : h \in \mathcal{H}\}$. It is clear that

$$(1, 1, ..., 1) = (\phi \circ \tilde{h}(\mathbf{x}_1), ..., \phi \circ \tilde{h}(\mathbf{x}_m), \phi \circ \tilde{h}(\mathbf{x}_{m+1})) \in \mathcal{H}^{\phi}_C,$$

where $(1, 1, ..., 1)$ means all elements are 1.

Let $\mathcal{H}^{\phi}_{m+1} = \{(\phi \circ h(\mathbf{x}_1), ..., \phi \circ h(\mathbf{x}_m), \phi \circ h(\mathbf{x}_{m+1}) : h \text{ is any hypothesis function from } \mathcal{X} \text{ to } \mathcal{Y}_{\text{all}}\}$.

Clearly, $\mathcal{H}^{\phi}_C \subset \mathcal{H}^{\phi}_{m+1}$ and $|\mathcal{H}^{\phi}_{m+1}| = 2^{m+1}$. Sauer-Shelah-Perles Lemma (Lemma 6.10 in [21]) implies that

$$|\mathcal{H}^{\phi}_C| \leq \sum_{i=0}^{v} \binom{m+1}{i}.$$

Since $\sum_{i=0}^{v}\binom{m+1}{i} < 2^{m+1}-1$ (because $v < m$), we obtain that $|\mathcal{H}_C^{\phi}| \leq 2^{m+1}-2$. Therefore, $\mathcal{H}_C^{\phi} \cup \{(2,2...,2)\}$ is a proper subset of $\mathcal{H}_{m+1}^{\phi}$, where $(2,2,...,2)$ means that all elements are 2. Note that $(1,1...,1)$ (all elements are 1) also belongs to $\mathcal{H}_C^{\phi}$. Hence, $\mathcal{H}_C^{\phi} \cup \{(2,2...,2)\} \cup \{(1,1...,1)\}$ is a proper subset of $\mathcal{H}_{m+1}^{\phi}$, which implies that we can obtain a hypothesis function h' satisfying that:

$$
\begin{aligned}
&1)(\phi \circ h'(\mathbf{x}_1), ..., \phi \circ h'(\mathbf{x}_m), \phi \circ h'(\mathbf{x}_{m+1})) \notin \mathcal{H}_C^{\phi};\\
&2)\ \text{There exist } \mathbf{x}_j, \mathbf{x}_p \in C \text{ such that } \phi \circ h'(\mathbf{x}_j) = 2 \text{ and } \phi \circ h'(\mathbf{x}_p) = 1.
\end{aligned}
$$

Let $C_{\mathrm{I}} = C \cap \{\mathbf{x} \in \mathcal{X} : \phi \circ h'(\mathbf{x}) = 1\}$ and $C_{\mathrm{O}} = C \cap \{\mathbf{x} \in \mathcal{X} : \phi \circ h'(\mathbf{x}) = 2\}$;

Then, we construct a special domain D_{XY}:

$$D_{XY} = 0.5 * D_{X_{\mathrm{I}}} * D_{Y_{\mathrm{I}}|X_{\mathrm{I}}} + 0.5 * D_{X_{\mathrm{O}}} * D_{Y_{\mathrm{O}}|X_{\mathrm{O}}}, \text{ where}$$

$$D_{X_{\mathrm{I}}} = \frac{1}{|C_{\mathrm{I}}|}\sum_{\mathbf{x}\in C_{\mathrm{I}}} \delta_{\mathbf{x}} \quad \text{and} \quad D_{Y_{\mathrm{I}}|X_{\mathrm{I}}}(y|\mathbf{x}) = 1, \text{ if } \tilde{h}(\mathbf{x}) = y \text{ and } \mathbf{x} \in C_{\mathrm{I}};$$

and

$$D_{X_{\mathrm{O}}} = \frac{1}{|C_{\mathrm{O}}|}\sum_{\mathbf{x}\in C_{\mathrm{O}}} \delta_{\mathbf{x}} \quad \text{and} \quad D_{Y_{\mathrm{O}}|X_{\mathrm{O}}}(K+1|\mathbf{x}) = 1, \text{ if } \mathbf{x} \in C_{\mathrm{O}}.$$

Since D_{XY} is a finite discrete distribution and $(\phi \circ h'(\mathbf{x}_1), ..., \phi \circ h'(\mathbf{x}_m), \phi \circ h'(\mathbf{x}_{m+1})) \notin \mathcal{H}_C^{\phi}$, it is clear that $\arg\min_{h\in\mathcal{H}} R_D(h) \neq \emptyset$ and $\inf_{h\in\mathcal{H}} R_D(h) > 0$.

Additionally, $R_D^{\mathrm{in}}(\tilde{h}) = 0$. Therefore, $\inf_{h\in\mathcal{H}} R_D^{\mathrm{in}}(h) = 0$.

Proof by Contradiction: suppose that OOD detection is learnable in $\mathscr{D}_{XY}^{s}$ for $\mathcal{H}$, then Lemma 1 implies that

$$\inf_{h\in\mathcal{H}} R_D(h) = 0.5 * \inf_{h\in\mathcal{H}} R_D^{\mathrm{in}}(h) + 0.5 * \inf_{h\in\mathcal{H}} R_D^{\mathrm{out}}(h).$$

Therefore, if OOD detection is learnable in $\mathscr{D}_{XY}^{s}$ for $\mathcal{H}$, then $\inf_{h\in\mathcal{H}} R_D^{\mathrm{out}}(h) > 0$.

Until now, we have constructed a domain D_{XY} (defined over $\mathcal{X} \times \mathcal{Y}_{\mathrm{all}}$) with finite support and satisfying that $\inf_{h\in\mathcal{H}} R_D^{\mathrm{out}}(h) > 0$. Note that $\mathcal{H}$ is the separate space for OOD data (Assumption 1 holds). Using Lemma 4, we know that OOD detection is not learnable in $\mathscr{D}_{XY}^{s}$ for $\mathcal{H}$, which is **inconsistent** with our assumption that OOD detection is learnable in $\mathscr{D}_{XY}^{s}$ for $\mathcal{H}$. Therefore, OOD detection is not learnable in $\mathscr{D}_{XY}^{s}$ for $\mathcal{H}$. We have completed the proof. □

Theorem 5 (Impossibility Theorem for Separate Space)**.** *If Assumption 1 holds,* $\mathrm{VCdim}(\phi \circ \mathcal{H}) < +\infty$ *and* $\sup_{h\in\mathcal{H}} |\{\mathbf{x} \in \mathcal{X} : h(\mathbf{x}) \in \mathcal{Y}\}| = +\infty$*, then OOD detection is not learnable in separate space* $\mathscr{D}_{XY}^{s}$ *for* $\mathcal{H}$*, where* ϕ *maps ID labels to* 1 *and maps OOD labels to* 2*.*

Proof of Theorem 5. Let $\mathrm{VCdim}(\phi \circ \mathcal{H}) = v$. Since $\sup_{h\in\mathcal{H}} |\{\mathbf{x} \in \mathcal{X} : \boldsymbol{h}(\mathbf{x}) \in \mathcal{Y}\}| = +\infty$, it is clear that $\sup_{h\in\mathcal{H}} |\{\mathbf{x} \in \mathcal{X} : \boldsymbol{h}(\mathbf{x}) \in \mathcal{Y}\}| > v$. Using Lemma 5, we complete this proof. □

I Proofs of Theorem 6 and Theorem 7

I.1 Proof of Theorem 6

Firstly, we need two lemmas, which are motivated by Lemma 19.2 and Lemma 19.3 in [21].

Lemma 6. *Let* C_1*,...,*C_r *be a cover of space* $\mathcal{X}$*, i.e.,* $\sum_{i=1}^{r} C_i = \mathcal{X}$*. Let* $S_X = \{\mathbf{x}^1, ..., \mathbf{x}^n\}$ *be a sequence of* n *data drawn from* $D_{X_{\mathrm{I}}}$*, i.i.d. Then*

$$\mathbb{E}_{S_X \sim D_{X_{\mathrm{I}}}^n}\Big(\sum_{i:C_i\cap S_X=\emptyset} D_{X_{\mathrm{I}}}(C_i)\Big) \leq \frac{r}{en}.$$

Proof of Lemma 6.

$$\mathbb{E}_{S_X\sim D^n_{X_{\rm I}}}\Big(\sum_{i:C_i\cap S_X=\emptyset} D_{X_{\rm I}}(C_i)\Big)=\sum_{i=1}^r\Big(D_{X_{\rm I}}(C_i)\cdot\mathbb{E}_{S_X\sim D^n_{X_{\rm I}}}\big(\mathbf{1}_{C_i\cap S_X=\emptyset}\big)\Big),$$

where $\mathbf{1}$ is the characteristic function.

For each i,

$$\begin{aligned}\mathbb{E}_{S_X\sim D^n_{X_{\rm I}}}\big(\mathbf{1}_{C_i\cap S_X=\emptyset}\big)&=\int_{\mathcal{X}^n}\mathbf{1}_{C_i\cap S_X=\emptyset}\mathrm{d}D^n_{X_{\rm I}}(S_X)\\&=\Big(\int_{\mathcal{X}}\mathbf{1}_{C_i\cap\{\mathbf{x}\}=\emptyset}\mathrm{d}D_{X_{\rm I}}(\mathbf{x})\Big)^n\\&=\big(1-D_{X_{\rm I}}(C_i)\big)^n\le e^{-nD_{X_{\rm I}}(C_i)}.\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}_{S_X\sim D^n_{X_{\rm I}}}\Big(\sum_{i:C_i\cap S=\emptyset} D_{X_{\rm I}}(C_i)\Big)&\le\sum_{i=1}^r D_{X_{\rm I}}(C_i)e^{-nD_{X_{\rm I}}(C_i)}\\&\le r\max_{i\in\{1,\dots,r\}}D_{X_{\rm I}}(C_i)e^{-nD_{X_{\rm I}}(C_i)}\le\frac{r}{ne},\end{aligned}$$

here we have used inequality: $\max_{i\in\{1,\dots,r\}}a_ie^{-na_i}\le 1/(ne)$. The proof has been completed. □

Lemma 7. *Let $K=1$. When $\mathcal{X}\subset\mathbb{R}^d$ is a bounded set, there exists a monotonically decreasing sequence $\epsilon_{\rm cons}(m)$ satisfying that $\epsilon_{\rm cons}(m)\to 0$, as $m\to 0$, such that*

$$\mathbb{E}_{\mathbf{x}\sim D_{X_{\rm I}},S\sim D^n_{X_{\rm I}Y_{\rm I}}}\mathrm{dist}(\mathbf{x},\pi_1(\mathbf{x},S))<\epsilon_{\rm cons}(n),$$

where dist *is the Euclidean distance, $\pi_1(\mathbf{x},S)=\arg\min_{\tilde{\mathbf{x}}\in S_X}\mathrm{dist}(\mathbf{x},\tilde{\mathbf{x}})$, here S_X is the feature part of S, i.e., $S_X=\{\mathbf{x}^1,\dots,\mathbf{x}^n\}$, if $S=\{(\mathbf{x}^1,y^1),\dots,(\mathbf{x}^n,y^n)\}$.*

Proof of Lemma 7. Since $\mathcal{X}$ is bounded, without loss of generality, we set $\mathcal{X}\subset[0,1)^d$. Fix $\epsilon=1/T$, for some integer T. Let $r=T^d$ and $C_1,C_2,\dots,C_r$ be a cover of $\mathcal{X}$: for every $(a_1,\dots,a_T)\in[T]^d:=[1,\dots,T]^d$, there exists a $C_i=\{\mathbf{x}=(x_1,\dots,x_d):\forall j\in\{1,\dots,d\},x_j\in[(a_j-1)/T,a_j/T)\}$.

If $\mathbf{x},\mathbf{x}'$ belong to some C_i, then $\mathrm{dist}(\mathbf{x},\mathbf{x}')\le\sqrt{d}\epsilon$; otherwise, $\mathrm{dist}(\mathbf{x},\mathbf{x}')\le\sqrt{d}$. Therefore,

$$\begin{aligned}&\mathbb{E}_{\mathbf{x}\sim D_{X_{\rm I}},S\sim D^n_{X_{\rm I}Y_{\rm I}}}\mathrm{dist}(\mathbf{x},\pi_1(\mathbf{x},S))\\\le&\mathbb{E}_{S\sim D^n_{X_{\rm I}Y_{\rm I}}}\Big(\sqrt{d}\epsilon\sum_{i:C_i\cap S_X\neq\emptyset}D_{X_{\rm I}}(C_i)+\sqrt{d}\sum_{i:C_i\cap S_X=\emptyset}D_{X_{\rm I}}(C_i)\Big)\\\le&\mathbb{E}_{S_X\sim D^n_{X_{\rm I}}}\Big(\sqrt{d}\epsilon\sum_{i:C_i\cap S_X\neq\emptyset}D_{X_{\rm I}}(C_i)+\sqrt{d}\sum_{i:C_i\cap S_X=\emptyset}D_{X_{\rm I}}(C_i)\Big).\end{aligned}$$

Note that $C_1,\dots,C_r$ are disjoint.

Therefore, $\sum_{i:C_i\cap S_X\neq\emptyset}D_{X_{\rm I}}(C_i)\le D_{X_{\rm I}}(\sum_{i:C_i\cap S_X\neq\emptyset}C_i)\le 1$. Using Lemma 6, we obtain

$$\mathbb{E}_{\mathbf{x}\sim D_{X_{\rm I}},S\sim D^n_{X_{\rm I}Y_{\rm I}}}\mathrm{dist}(\mathbf{x},\pi_1(\mathbf{x},S))\le\sqrt{d}\epsilon+\frac{r\sqrt{d}}{ne}=\sqrt{d}\epsilon+\frac{\sqrt{d}}{ne\epsilon^d}.$$

If we set $\epsilon=2n^{-1/(d+1)}$, then

$$\mathbb{E}_{\mathbf{x}\sim D_{X_{\rm I}},S\sim D^n_{X_{\rm I}Y_{\rm I}}}\mathrm{dist}(\mathbf{x},\pi_1(\mathbf{x},S))\le\frac{2\sqrt{d}}{n^{1/(d+1)}}+\frac{\sqrt{d}}{2^den^{1/(d+1)}}.$$

If we set $\epsilon_{\rm cons}(n)=\frac{2\sqrt{d}}{n^{1/(d+1)}}+\frac{\sqrt{d}}{2^den^{1/(d+1)}}$, we complete this proof. □

Theorem 6. *Let $K = 1$ and $|\mathcal{X}| < +\infty$. Suppose that Assumption 1 holds and the constant function $h^{\text{in}} := 1 \in \mathcal{H}$. Then OOD detection is learnable in $\mathscr{D}^s_{XY}$ for $\mathcal{H}$* ***if and only if*** *$\mathcal{H}_{\text{all}} - \{h^{\text{out}}\} \subset \mathcal{H}$, where $\mathcal{H}_{\text{all}}$ is the hypothesis space consisting of all hypothesis functions, and h^{out} is a constant function that $h^{\text{out}} := 2$, here 1 represents ID data and 2 represents OOD data.*

Proof of Theorem 6. **First**, we prove that if the hypothesis space $\mathcal{H}$ is a separate space for OOD (*i.e.*, Assumption 1 holds), the constant function $h^{\text{in}} := 1 \in \mathcal{H}$, then that OOD detection is learnable in $\mathscr{D}^s_{XY}$ for $\mathcal{H}$ implies $\mathcal{H}_{\text{all}} - \{h^{\text{out}}\} \subset \mathcal{H}$.

Proof by Contradiction: suppose that there exists $h' \in \mathcal{H}_{\text{all}}$ such that $h' \neq h^{\text{out}}$ and $h' \notin \mathcal{H}$.

Let $\mathcal{X} = \{\mathbf{x}_1, ..., \mathbf{x}_m\}$, $C_{\text{I}} = \{\mathbf{x} \in \mathcal{X} : h'(\mathbf{x}) \in \mathcal{Y}\}$ and $C_{\text{O}} = \{\mathbf{x} \in \mathcal{X} : h'(\mathbf{x}) = K + 1\}$.

Because $h' \neq h^{\text{out}}$, we know that $C_{\text{I}} \neq \emptyset$.

We construct a special domain $D_{XY} \in \mathscr{D}^s_{XY}$: if $C_{\text{O}} = \emptyset$, then $D_{XY} = D_{X_{\text{I}}} * D_{Y_{\text{I}}|X_{\text{I}}}$; otherwise,

$$D_{XY} = 0.5 * D_{X_{\text{I}}} * D_{Y_{\text{I}}|X_{\text{I}}} + 0.5 * D_{X_{\text{O}}} * D_{Y_{\text{O}}|X_{\text{O}}}, \quad \text{where}$$

$$D_{X_{\text{I}}} = \frac{1}{|C_{\text{I}}|} \sum_{\mathbf{x} \in C_{\text{I}}} \delta_{\mathbf{x}} \ \text{ and } \ D_{Y_{\text{I}}|X_{\text{I}}}(y|\mathbf{x}) = 1, \ \text{ if } h'(\mathbf{x}) = y \text{ and } \mathbf{x} \in C_{\text{I}},$$

and

$$D_{X_{\text{O}}} = \frac{1}{|C_{\text{O}}|} \sum_{\mathbf{x} \in C_{\text{O}}} \delta_{\mathbf{x}} \ \text{ and } \ D_{Y_{\text{O}}|X_{\text{O}}}(K+1|\mathbf{x}) = 1, \ \text{ if } \mathbf{x} \in C_{\text{O}}.$$

Since $h' \notin \mathcal{H}$ and $|\mathcal{X}| < +\infty$, then $\arg\min_{h \in \mathcal{H}} R_D(h) \neq \emptyset$, and $\inf_{h \in \mathcal{H}} R_D(h) > 0$. Additionally, $R^{\text{in}}_D(h^{\text{in}}) = 0$ (here $h^{\text{in}} = 1$), hence, $\inf_{h \in \mathcal{H}} R^{\text{in}}_D(h) = 0$.

Since OOD detection is learnable in $\mathscr{D}^s_{XY}$ for $\mathcal{H}$, Lemma 1 implies that

$$\inf_{h \in \mathcal{H}} R_D(h) = (1 - \pi^{\text{out}}) \inf_{h \in \mathcal{H}} R^{\text{in}}_D(h) + \pi^{\text{out}} \inf_{h \in \mathcal{H}} R^{\text{out}}_D(h),$$

where $\pi^{\text{out}} = D_Y(Y = K + 1) = 1$ or 0.5. Since $\inf_{h \in \mathcal{H}} R^{\text{in}}_D(h) = 0$ and $\inf_{h \in \mathcal{H}} R_D(h) > 0$, we obtain that $\inf_{h \in \mathcal{H}} R^{\text{out}}_D(h) > 0$.

Until now, we have constructed a special domain $D_{XY} \in \mathscr{D}^s_{XY}$ satisfying that $\inf_{h \in \mathcal{H}} R^{\text{out}}_D(h) > 0$. Using Lemma 4, we know that OOD detection in $\mathscr{D}^s_{XY}$ is not learnable for $\mathcal{H}$, which is **inconsistent** with the condition that OOD detection is learnable in $\mathscr{D}^s_{XY}$ for $\mathcal{H}$. Therefore, the assumption (there exists $h' \in \mathcal{H}_{\text{all}}$ such that $h' \neq h^{\text{out}}$ and $h \notin \mathcal{H}$) doesn't hold, which implies that $\mathcal{H}_{\text{all}} - \{h^{\text{out}}\} \subset \mathcal{H}$.

Second, we prove that if $\mathcal{H}_{\text{all}} - \{h^{\text{out}}\} \subset \mathcal{H}$, then OOD detection is learnable in $\mathscr{D}^s_{XY}$ for $\mathcal{H}$.

To prove this result, we need to design a special algorithm. Let $d_0 = \min_{\mathbf{x}, \mathbf{x}' \in \mathcal{X} \text{ and } \mathbf{x} \neq \mathbf{x}'} \text{dist}(\mathbf{x}, \mathbf{x}')$, where dist is the Euclidean distance. It is clear that $d_0 > 0$. Let

$$\mathbf{A}(S)(\mathbf{x}) = \begin{cases} 1, & \text{if } \text{dist}(\mathbf{x}, \pi_1(\mathbf{x}, S)) < 0.5 * d_0; \\ 2, & \text{if } \text{dist}(\mathbf{x}, \pi_1(\mathbf{x}, S)) \geq 0.5 * d_0, \end{cases}$$

where $\pi_1(\mathbf{x}, S) = \arg\min_{\tilde{\mathbf{x}} \in S_X} \text{dist}(\mathbf{x}, \tilde{\mathbf{x}})$, here S_X is the feature part of S, *i.e.*, $S_X = \{\mathbf{x}^1, ..., \mathbf{x}^n\}$, if $S = \{(\mathbf{x}^1, y^1), ..., (\mathbf{x}^n, y^n)\}$.

For any $\mathbf{x} \in \text{supp} D_{X_{\text{I}}}$, it is easy to check that for almost all $S \sim D^n_{X_{\text{I}} Y_{\text{I}}}$,

$$\text{dist}(\mathbf{x}, \pi_1(\mathbf{x}, S)) > 0.5 * d_0,$$

which implies that

$$\mathbf{A}(S)(\mathbf{x}) = 2,$$

hence,

$$\mathbb{E}_{S \sim D^n_{X_{\text{I}} Y_{\text{I}}}} R^{\text{out}}_D(\mathbf{A}(S)) = 0. \tag{33}$$

Using Lemma 7, for any $\mathbf{x} \in \text{supp} D_{X_{\text{I}}}$, we have

$$\mathbb{E}_{\mathbf{x} \sim D_{X_{\text{I}}}, S \sim D^n_{X_{\text{I}} Y_{\text{I}}}} \text{dist}(\mathbf{x}, \pi_1(\mathbf{x}, S)) < \epsilon_{\text{cons}}(n),$$

where $\epsilon_{\text{cons}}(n) \to 0$, as $n \to 0$ and $\epsilon_{\text{cons}}(n)$ is a monotonically decreasing sequence.

Hence, we have that

$$D_{X_{\rm I}} \times D^n_{X_{\rm I}Y_{\rm I}}(\{(\mathbf{x}, S) : \mathrm{dist}(\mathbf{x}, \pi_1(\mathbf{x}, S)) \geq 0.5 * d_0\}) \leq 2\epsilon_{\rm cons}(n)/d_0,$$

where $D_{X_{\rm I}} \times D^n_{X_{\rm I}Y_{\rm I}}$ is the product measure of $D_{X_{\rm I}}$ and $D^n_{X_{\rm I}Y_{\rm I}}$ [36]. Therefore,

$$D_{X_{\rm I}} \times D^n_{X_{\rm I}Y_{\rm I}}(\{(\mathbf{x}, S) : \mathbf{A}(S)(\mathbf{x}) = 1\}) > 1 - 2\epsilon_{\rm cons}(n)/d_0,$$

which implies that

$$\mathbb{E}_{S\sim D^n_{X_{\rm I}Y_{\rm I}}} R^{\rm in}_D(\mathbf{A}(S)) \leq 2B\epsilon_{\rm cons}(n)/d_0, \tag{34}$$

where $B = \max\{\ell(1,2), \ell(2,1)\}$. Using Eq. (33) and Eq. (34), we have proved that

$$\mathbb{E}_{S\sim D^n_{X_{\rm I}Y_{\rm I}}} R_D(\mathbf{A}(S)) \leq 0 + 2B\epsilon_{\rm cons}(m)/d_0 \leq \inf_{h\in\mathcal{H}} R_D(h) + 2B\epsilon_{\rm cons}(m)/d_0. \tag{35}$$

It is easy to check that $\mathbf{A}(S) \in \mathcal{H}_{\rm all} - \{h^{\rm out}\}$. Therefore, we have constructed a consistent algorithm $\mathbf{A}$ for $\mathcal{H}$. We have completed this proof. □

I.2 Proof of Theorem 7

Theorem 7. *Let $|\mathcal{X}| < +\infty$ and $\mathcal{H} = \mathcal{H}^{\rm in} \bullet \mathcal{H}^{\rm b}$. If $\mathcal{H}_{\rm all} - \{h^{\rm out}\} \subset \mathcal{H}^{\rm b}$ and Condition 2 holds, then OOD detection is learnable in $\mathscr{D}^s_{XY}$ for $\mathcal{H}$, where $\mathcal{H}_{\rm all}$ and $h^{\rm out}$ are defined in Theorem 6.*

Proof of Theorem 7. Since $|\mathcal{X}| < +\infty$, we know that $|\mathcal{H}| < +\infty$, which implies that $\mathcal{H}^{\rm in}$ is agnostic PAC learnable for supervised learning in classification. Therefore, there exist an algorithm $\mathbf{A}^{\rm in} : \cup_{n=1}^{+\infty}(\mathcal{X}\times\mathcal{Y})^n \to \mathcal{H}^{\rm in}$ and a monotonically decreasing sequence $\epsilon(n)$, such that $\epsilon(n) \to 0$, as $n \to +\infty$, and for any $D_{XY} \in \mathscr{D}^s_{XY}$,

$$\mathbb{E}_{S\sim D^n_{X_{\rm I}Y_{\rm I}}} R^{\rm in}_D(\mathbf{A}^{\rm in}(S)) \leq \inf_{h\in\mathcal{H}^{\rm in}} R^{\rm in}_D(h) + \epsilon(n).$$

Since $|\mathcal{X}| < +\infty$ and $\mathcal{H}^{\rm b}$ almost contains all binary classifiers, then using Theorem 6 and Theorem 1, we obtain that there exist an algorithm $\mathbf{A}^{\rm b} : \cup_{n=1}^{+\infty}(\mathcal{X}\times\{1,2\})^n \to \mathcal{H}^{\rm b}$ and a monotonically decreasing sequence $\epsilon'(n)$, such that $\epsilon'(n) \to 0$, as $n \to +\infty$, and for any $D_{XY} \in \mathscr{D}^s_{XY}$,

$$\mathbb{E}_{S\sim D^n_{X_{\rm I}Y_{\rm I}}} R^{\rm in}_{\phi(D)}(\mathbf{A}^{\rm b}(\phi(S))) \leq \inf_{h\in\mathcal{H}^{\rm b}} R^{\rm in}_{\phi(D)}(h) + \epsilon'(n),$$
$$\mathbb{E}_{S\sim D^n_{X_{\rm I}Y_{\rm I}}} R^{\rm out}_{\phi(D)}(\mathbf{A}^{\rm b}(\phi(S))) \leq \inf_{h\in\mathcal{H}^{\rm b}} R^{\rm out}_{\phi(D)}(h) + \epsilon'(n),$$

where ϕ maps ID's labels to 1 and OOD's label to 2,

$$R^{\rm in}_{\phi(D)}(\mathbf{A}^{\rm b}(\phi(S))) = \int_{\mathcal{X}\times\mathcal{Y}} \ell(\mathbf{A}^{\rm b}(\phi(S))(\mathbf{x}), \phi(y)) {\rm d}D_{X_{\rm I}Y_{\rm I}}(\mathbf{x}, y), \tag{36}$$

$$R^{\rm in}_{\phi(D)}(h) = \int_{\mathcal{X}\times\mathcal{Y}} \ell(h(\mathbf{x}), \phi(y)) {\rm d}D_{X_{\rm I}Y_{\rm I}}(\mathbf{x}, y), \tag{37}$$

$$R^{\rm out}_{\phi(D)}(\mathbf{A}^{\rm b}(\phi(S))) = \int_{\mathcal{X}\times\{K+1\}} \ell(\mathbf{A}^{\rm b}(\phi(S))(\mathbf{x}), \phi(y)) {\rm d}D_{X_{\rm O}Y_{\rm O}}(\mathbf{x}, y), \tag{38}$$

and

$$R^{\rm out}_{\phi(D)}(h) = \int_{\mathcal{X}\times\{K+1\}} \ell(h(\mathbf{x}), \phi(y)) {\rm d}D_{X_{\rm O}Y_{\rm O}}(\mathbf{x}, y), \tag{39}$$

here $\phi(S) = \{(\mathbf{x}^1, \phi(y^1)), ..., (\mathbf{x}^n, \phi(y^n))\}$, if $S = \{(\mathbf{x}^1, y^1), ..., (\mathbf{x}^n, y^n)\}$.

Note that $\mathcal{H}^b$ almost contains all classifiers, and $\mathscr{D}^s_{XY}$ is the separate space. Hence,

$$\mathbb{E}_{S\sim D^n_{X_{\rm I}Y_{\rm I}}} R^{\rm in}_{\phi(D)}(\mathbf{A}^{\rm b}(\phi(S))) \leq \epsilon'(n), \quad \mathbb{E}_{S\sim D^n_{X_{\rm I}Y_{\rm I}}} R^{\rm out}_{\phi(D)}(\mathbf{A}^{\rm b}(\phi(S))) \leq \epsilon'(n).$$

Next, we construct an algorithm $\mathbf{A}$ using $\mathbf{A}^{\rm in}$ and $\mathbf{A}^{\rm out}$.

$$\mathbf{A}(S)(\mathbf{x}) = \begin{cases} K+1, & \text{if } \mathbf{A}^{\mathrm{b}}(\phi(S))(\mathbf{x}) = 2; \\ \mathbf{A}^{\mathrm{in}}(S)(\mathbf{x}), & \text{if } \mathbf{A}^{\mathrm{b}}(\phi(S))(\mathbf{x}) = 1. \end{cases}$$

Since $\inf_{h\in\mathcal{H}} R^{\mathrm{in}}_{\phi(D)}(\phi\circ h) = 0$, $\inf_{h\in\mathcal{H}} R^{\mathrm{out}}_{D}(h) = 0$, then by Condition 2, it is easy to check that

$$\inf_{h\in\mathcal{H}^{\mathrm{in}}} R^{\mathrm{in}}_{D}(h) = \inf_{h\in\mathcal{H}} R^{\mathrm{in}}_{D}(h).$$

Additionally, the risk $R^{\mathrm{in}}_{D}(\mathbf{A}(S))$ is from two parts: 1) ID data are detected as OOD data; 2) ID data are detected as ID data, but are classified as incorrect ID classes. Therefore, we have the inequality:

$$\begin{aligned} \mathbb{E}_{S\sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} R^{\mathrm{in}}_{D}(\mathbf{A}(S)) &\leq \mathbb{E}_{S\sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} R^{\mathrm{in}}_{D}(\mathbf{A}^{\mathrm{in}}(S)) + c\mathbb{E}_{S\sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} R^{\mathrm{in}}_{\phi(D)}(\mathbf{A}^{\mathrm{b}}(\phi(S))) \\ &\leq \inf_{h\in\mathcal{H}^{\mathrm{in}}} R^{\mathrm{in}}_{D}(h) + \epsilon(n) + c\epsilon'(n) = \inf_{h\in\mathcal{H}} R^{\mathrm{in}}_{D}(h) + \epsilon(n) + c\epsilon'(n), \end{aligned} \tag{40}$$

where $c = \max_{y_1,y_2\in\mathcal{Y}} \ell(y_1,y_2)/\min\{\ell(1,2),\ell(2,1)\}$.

Note that the risk $R^{\mathrm{out}}_{D}(\mathbf{A}(S))$ is from the case that OOD data are detected as ID data. Therefore,

$$\begin{aligned} \mathbb{E}_{S\sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} R^{\mathrm{out}}_{D}(\mathbf{A}(S)) &\leq c\mathbb{E}_{S\sim D^n_{X_{|rmI}Y_{\mathrm{I}}}} R^{\mathrm{out}}_{\phi(D)}(\mathbf{A}^{\mathrm{b}}(\phi(S))) \\ &\leq c\epsilon'(n) \leq \inf_{h\in\mathcal{H}} R^{\mathrm{out}}_{D}(h) + c\epsilon'(n). \end{aligned} \tag{41}$$

Note that $(1-\alpha)\inf_{h\in\mathcal{H}} R^{\mathrm{in}}_{D}(h) + \alpha\inf_{h\in\mathcal{H}} R^{\mathrm{out}}_{D}(h) \leq \inf_{h\in\mathcal{H}} R^{\alpha}_{D}(h)$. Then, using Eq. (40) and Eq. (41), we obtain that for any $\alpha\in[0,1]$,

$$\mathbb{E}_{S\sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} R^{\alpha}_{D}(\mathbf{A}(S)) \leq \inf_{h\in\mathcal{H}} R^{\alpha}_{D}(h) + \epsilon(n) + c\epsilon'(n).$$

According to Theorem 1 (the second result), we complete the proof. □

J Proofs of Theorems 8 and 9

J.1 Proof of Theorem 8

Lemma 8. *Given a prior-unknown space $\mathscr{D}_{XY}$ and a hypothesis space $\mathcal{H}$, if Condition 3 holds, then for any equivalence class $[D'_{XY}]$ with respect to $\mathscr{D}_{XY}$, OOD detection is learnable in the equivalence class $[D'_{XY}]$ for $\mathcal{H}$. Furthermore, the learning rate can attain $O(1/n)$.*

Proof. Let $\mathscr{F}$ be a set consisting of all infinite sequences, whose coordinates are hypothesis functions, *i.e.*,

$$\mathscr{F} = \{\boldsymbol{h} = (h_1, ..., h_n, ...) : \forall h_n \in \mathcal{H}, n = 1, ..., +\infty\}.$$

For each $\boldsymbol{h}\in\mathscr{F}$, there is a corresponding algorithm $\mathbf{A}_{\boldsymbol{h}}$: $\mathbf{A}_{\boldsymbol{h}}(S) = h_n$, if $|S| = n$. $\mathscr{F}$ generates an algorithm class $\mathscr{A} = \{\mathbf{A}_{\boldsymbol{h}} : \forall \boldsymbol{h}\in\mathscr{F}\}$. We select a consistent algorithm from the algorithm class $\mathscr{A}$.

We construct a special infinite sequence $\tilde{\boldsymbol{h}} = (\tilde{h}_1, ..., \tilde{h}_n, ...) \in \mathscr{F}$. For each positive integer n, we select $\tilde{h}_n$ from

$$\bigcap_{\forall D_{XY}\in[D'_{XY}]} \{h'\in\mathcal{H} : R^{\mathrm{out}}_{D}(h') \leq \inf_{h\in\mathcal{H}} R^{\mathrm{out}}_{D}(h) + 2/n\} \bigcap \{h'\in\mathcal{H} : R^{\mathrm{in}}_{D}(h') \leq \inf_{h\in\mathcal{H}} R^{\mathrm{in}}_{D}(h) + 2/n\}.$$

The existence of $\tilde{h}_n$ is based on Condition 3. It is easy to check that for any $D_{XY}\in[D'_{XY}]$,

$$\mathbb{E}_{S\sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} R^{\mathrm{in}}_{D}(\mathbf{A}_{\tilde{\boldsymbol{h}}}(S)) \leq \inf_{h\in\mathcal{H}} R^{\mathrm{in}}_{D}(h) + 2/n.$$

$$\mathbb{E}_{S\sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} R^{\mathrm{out}}_{D}(\mathbf{A}_{\tilde{\boldsymbol{h}}}(S)) \leq \inf_{h\in\mathcal{H}} R^{\mathrm{out}}_{D}(h) + 2/n.$$

Since $(1-\alpha)\inf_{h\in\mathcal{H}} R^{\mathrm{in}}_{D}(h) + \alpha\inf_{h\in\mathcal{H}} R^{\mathrm{out}}_{D}(h) \leq \inf_{h\in\mathcal{H}} R^{\alpha}_{D}(h)$, we obtain that for any $\alpha\in[0,1]$,

$$\mathbb{E}_{S\sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} R^{\alpha}_{D}(\mathbf{A}_{\tilde{\boldsymbol{h}}}(S)) \leq \inf_{h\in\mathcal{H}} R^{\alpha}_{D}(h) + 2/n.$$

Using Theorem 1 (the second result), we have completed this proof. □

Theorem 8. *Suppose that $\mathcal{X}$ is a bounded set. OOD detection is learnable in the finite-ID-distribution space $\mathscr{D}_{XY}^{F}$ for $\mathcal{H}$* ***if and only if*** *the compatibility condition (i.e., Condition 3) holds. Furthermore, the learning rate $\epsilon_{\text{cons}}(n)$ can attain $O(1/\sqrt{n^{1-\theta}})$, for any $\theta \in (0,1)$.*

Proof of Theorem 8.

First, we prove that if OOD detection is learnable in $\mathscr{D}_{XY}^{F}$ for $\mathcal{H}$, then Condition 3 holds.

Since $\mathscr{D}_{XY}^{F}$ is the prior-unknown space, by Theorem 1, there exist an algorithm $\mathbf{A} : \cup_{n=1}^{+\infty}(\mathcal{X}\times\mathcal{Y})^n \to \mathcal{H}$ and a monotonically decreasing sequence $\epsilon_{\text{cons}}(n)$, such that $\epsilon_{\text{cons}}(n) \to 0$, as $n \to +\infty$, and for any $D_{XY} \in \mathscr{D}_{XY}^{F}$,

$$\mathbb{E}_{S\sim D^n_{X_{\text{I}}Y_{\text{I}}}}\left[R_D^{\text{in}}(\mathbf{A}(S)) - \inf_{h\in\mathcal{H}} R_D^{\text{in}}(h)\right] \leq \epsilon_{\text{cons}}(n),$$
$$\mathbb{E}_{S\sim D^n_{X_{\text{I}}Y_{\text{I}}}}\left[R_D^{\text{out}}(\mathbf{A}(S)) - \inf_{h\in\mathcal{H}} R_D^{\text{out}}(h)\right] \leq \epsilon_{\text{cons}}(n).$$

Then, for any $\epsilon > 0$, we can find n_ϵ such that $\epsilon \geq \epsilon_{\text{cons}}(n_\epsilon)$, therefore, if $n = n_\epsilon$, we have

$$\mathbb{E}_{S\sim D^{n_\epsilon}_{X_{\text{I}}Y_{\text{I}}}}\left[R_D^{\text{in}}(\mathbf{A}(S)) - \inf_{h\in\mathcal{H}} R_D^{\text{in}}(h)\right] \leq \epsilon,$$
$$\mathbb{E}_{S\sim D^{n_\epsilon}_{X_{\text{I}}Y_{\text{I}}}}\left[R_D^{\text{out}}(\mathbf{A}(S)) - \inf_{h\in\mathcal{H}} R_D^{\text{out}}(h)\right] \leq \epsilon,$$

which implies that there exists $S_\epsilon \sim D^{n_\epsilon}_{X_{\text{I}}Y_{\text{I}}}$ such that

$$R_D^{\text{in}}(\mathbf{A}(S_\epsilon)) - \inf_{h\in\mathcal{H}} R_D^{\text{in}}(h) \leq \epsilon,$$
$$R_D^{\text{out}}(\mathbf{A}(S_\epsilon)) - \inf_{h\in\mathcal{H}} R_D^{\text{out}}(h) \leq \epsilon.$$

Therefore, for any equivalence class $[D'_{XY}]$ with respect to $\mathscr{D}_{XY}^{F}$ and any $\epsilon > 0$, there exists a hypothesis function $\mathbf{A}(S_\epsilon) \in \mathcal{H}$ such that for any domain $D_{XY} \in [D'_{XY}]$,

$$\mathbf{A}(S_\epsilon) \in \{h' \in \mathcal{H} : R_D^{\text{out}}(h') \leq \inf_{h\in\mathcal{H}} R_D^{\text{out}}(h) + \epsilon\} \cap \{h' \in \mathcal{H} : R_D^{\text{in}}(h') \leq \inf_{h\in\mathcal{H}} R_D^{\text{in}}(h) + \epsilon\},$$

which implies that Condition 3 holds.

Second, we prove Condition 3 implies the learnability of OOD detection in $\mathscr{D}_{XY}^{F}$ for $\mathcal{H}$.

For convenience, we assume that all equivalence classes are $[D^1_{XY}], ..., [D^m_{XY}]$. By Lemma 8, for every equivalence class $[D^i_{XY}]$, we can find a corresponding algorithm $\mathbf{A}_{D^i}$ such that OOD detection is learnable in $[D^i_{XY}]$ for $\mathcal{H}$. Additionally, we also set the learning rate for $\mathbf{A}_{D^i}$ is $\epsilon^i(n)$. By Lemma 8, we know that $\epsilon^i(n)$ can attain $O(1/n)$.

Let $\mathcal{Z}$ be $\mathcal{X}\times\mathcal{Y}$. Then, we consider a bounded universal kernel $K(\cdot,\cdot)$ defined over $\mathcal{Z}\times\mathcal{Z}$. Consider the *maximum mean discrepancy* (MMD) [77], which is a metric between distributions: for any distributions P and Q defined over $\mathcal{Z}$, we use $\text{MMD}_K(Q,P)$ to represent the distance.

Let $\mathscr{F}$ be a set consisting of all finite sequences, whose coordinates are labeled data, *i.e.*,

$$\mathscr{F} = \{\mathbf{S} = (S_1, ..., S_i, ..., S_m) : \forall i = 1, ..., m \text{ and } \forall \text{ labeled data } S_i\}.$$

Then, we define an algorithm space as follows:

$$\mathscr{A} = \{\mathbf{A_S}^{7} : \forall\, \mathbf{S} \in \mathscr{F}\},$$

where

$$\mathbf{A_S}(S) = \mathbf{A}_{D^i}(S), \text{ if } i = \arg\min_{i\in\{1,...m\}} \text{MMD}_K(P_{S_i}, P_S),$$

here

$$P_S = \frac{1}{n}\sum_{(\mathbf{x},y)\in S} \delta_{(\mathbf{x},y)}, \quad P_{S_i} = \frac{1}{n}\sum_{(\mathbf{x},y)\in S_i}, \delta_{(\mathbf{x},y)}$$

[7] In this paper, we regard an algorithm as a mapping from $\cup_{n=1}^{+\infty}(\mathcal{X}\times\mathcal{Y})^n$ to $\mathcal{H}$. So we can design an algorithm like this.

and $\delta_{(\mathbf{x},y)}$ is the Dirac measure. Next, we prove that we can find an algorithm $\mathbf{A}$ from the algorithm space $\mathscr{A}$ such that $\mathbf{A}$ is the consistent algorithm.

Since the number of different equivalence classes is finite, we know that there exists a constant $c > 0$ such that for any different equivalence classes $[D^i_{XY}]$ and $[D^j_{XY}]$ $(i \neq j)$,

$$\mathrm{MMD}_K(D^i_{X_\mathrm{I}Y_\mathrm{I}}, D^j_{X_\mathrm{I}Y_\mathrm{I}}) > c.$$

Additionally, according to [77] and the property of $\mathscr{D}^F_{XY}$ (the number of different equivalence classes is finite), there exists a monotonically decreasing $\epsilon(n) \to 0$, as $n \to +\infty$ such that for any $D_{XY} \in \mathscr{D}$,

$$\mathbb{E}_{S\sim D^n_{X_\mathrm{I}Y_\mathrm{I}}}\mathrm{MMD}_K(D_{X_\mathrm{I}Y_\mathrm{I}}, P_S) \leq \epsilon(n), \text{ where } \epsilon(n) = O(\frac{1}{\sqrt{n^{1-\theta}}}). \tag{42}$$

Therefore, for every equivalence class $[D^i_{XY}]$, we can find data points S_{D^i} such that

$$\mathrm{MMD}_K(D^i_{X_\mathrm{I}Y_\mathrm{I}}, P_{S_{D^i}}) < \frac{c}{100}.$$

Let $\mathbf{S}' = \{S_{D^1}, ..., S_{D^i}, ..., S_{D^m}\}$. Then, we prove that $\mathbf{A}_{\mathbf{S}'}$ is a consistent algorithm. By Eq. (42), it is easy to check that for any $i \in \{1, ..., m\}$ and any $0 < \delta < 1$,

$$\mathbb{P}_{S\sim D^{i,n}_{X_\mathrm{I}Y_\mathrm{I}}}\big[\mathrm{MMD}_K(D^i_{X_\mathrm{I}Y_\mathrm{I}}, P_S) \leq \frac{\epsilon(n)}{\delta}\big] > 1 - \delta,$$

which implies that

$$\mathbb{P}_{S\sim D^{i,n}_{X_\mathrm{I}Y_\mathrm{I}}}\big[\mathrm{MMD}_K(P_{S_{D^i}}, P_S) \leq \frac{\epsilon(n)}{\delta} + \frac{c}{100}\big] > 1 - \delta.$$

Therefore, (here we set $\delta = 200\epsilon(n)/c$)

$$\mathbb{P}_{S\sim D^{i,n}_{X_\mathrm{I}Y_\mathrm{I}}}\big[\mathbf{A}_{\mathbf{S}'}(S) \neq \mathbf{A}_{D^i}(S)\big] \leq \frac{200\epsilon(n)}{c}.$$

Because $\mathbf{A}_{D^i}$ is a consistent algorithm for $[D^i_{XY}]$, we conclude that for all $\alpha \in [0, 1]$,

$$\mathbb{E}_{S\sim D^{i,n}_{X_\mathrm{I}Y_\mathrm{I}}}\big[R^\alpha_D(\mathbf{A}_{\mathbf{S}'}(S)) - \inf_{h\in\mathcal{H}} R^\alpha_D(h)\big] \leq \epsilon^i(n) + \frac{200B\epsilon(n)}{c},$$

where $\epsilon^i(n) = O(1/n)$ is the learning rate of $\mathbf{A}_{D^i}$ and B is the upper bound of the loss ℓ.

Let $\epsilon^{\max}(n) = \max\{\epsilon^1(n), ..., \epsilon^m(n)\} + \frac{200B\epsilon(n)}{c}$.

Then, we obtain that for any $D_{XY} \in \mathscr{D}^F_{XY}$ and all $\alpha \in [0, 1]$,

$$\mathbb{E}_{S\sim D^n_{X_\mathrm{I}Y_\mathrm{I}}}\big[R^\alpha_D(\mathbf{A}_{\mathbf{S}'}(S)) - \inf_{h\in\mathcal{H}} R^\alpha_D(h)\big] \leq \epsilon^{\max}(n) = O(\frac{1}{\sqrt{n^{1-\theta}}}).$$

According to Theorem 1 (the second result), $\mathbf{A}_{\mathbf{S}'}$ is the consistent algorithm. This proof is completed. □

J.2 Proof of Theorem 9

Theorem 9. *Given a density-based space $\mathscr{D}^{\mu,b}_{XY}$, if $\mu(\mathcal{X}) < +\infty$, the Realizability Assumption holds, then when $\mathcal{H}$ has finite Natarajan dimension [21], OOD detection is learnable in $\mathscr{D}^{\mu,b}_{XY}$ for $\mathcal{H}$. Furthermore, the learning rate $\epsilon_{\mathrm{cons}}(n)$ can attain $O(1/\sqrt{n^{1-\theta}})$, for any $\theta \in (0, 1)$.*

Proof of Theorem 9. **First**, we consider the case that the loss ℓ is the zero-one loss.

Since $\mu(\mathcal{X}) < +\infty$, without loss of generality, we assume that $\mu(\mathcal{X}) = 1$. We also assume that f_I is D_{X_I}'s density function and f_O is D_{X_O}'s density function. Let f be the density function for $0.5 * D_{X_\mathrm{I}} + 0.5 * D_{X_\mathrm{O}}$. It is easy to check that $f = 0.5 * f_\mathrm{I} + 0.5 * f_\mathrm{O}$. Additionally, due to

Realizability Assumption, it is obvious that for any samples $S = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\} \sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}$, i.i.d., we have that there exists $h^* \in \mathcal{H}$ such that

$$\frac{1}{n}\sum_{i=1}^{n} \ell(h^*(\mathbf{x}_i), y_i) = 0.$$

Given m data points $S_m = \{\mathbf{x}'_1, ..., \mathbf{x}'_m\} \subset \mathcal{X}^m$. We consider the following learning rule:

$$\min_{h\in\mathcal{H}} \frac{1}{m}\sum_{j=1}^{m} \ell(h(\mathbf{x}'_j), K+1), \text{ subject to } \frac{1}{n}\sum_{i=1}^{n} \ell(h(\mathbf{x}_i), y_i) = 0.$$

We denote the algorithm, which solves the above rule, as $\mathbf{A}_{S_m}$. For different data points S_m, we have different algorithm $\mathbf{A}_{S_m}$. Let $\mathcal{S}$ be the infinite sequence set that consists of all infinite sequences, whose coordinates are data points, *i.e.*,

$$\mathcal{S} := \{\mathbf{S} := (S_1, S_2, ..., S_m, ...) : S_m \text{ are any } m \text{ data points}, \ m = 1, ..., +\infty\}. \tag{43}$$

Using $\mathcal{S}$, we construct an algorithm space as follows:

$$\mathscr{A} := \{\mathbf{A}_{\mathbf{S}} : \forall\, \mathbf{S} \in \mathcal{S}\}, \text{ where } \mathbf{A}_{\mathbf{S}}(S) = \mathbf{A}_{S_n}(S), \text{ if } |S| = n.$$

Next, we prove that there exists an algorithm $\mathbf{A}_{\mathbf{S}} \in \mathscr{A}$, which is a consistent algorithm. Given data points $S_n \sim \mu^n$, i.i.d., using the Natarajan dimension theory and Empirical risk minimization principle [21], it is easy to obtain that there exists a uniform constant C_θ such that (we mainly use the uniform bounds to obtain the following bounds)

$$\mathbb{E}_{S\sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} \sup_{h\in\mathcal{H}_S} R^{\mathrm{in}}_D(h) \leq \inf_{h\in\mathcal{H}} R^{\mathrm{in}}_D(h) + \frac{C_\theta}{\sqrt{n^{1-\theta}}},$$

and

$$\mathbb{E}_{S_n\sim\mu^n} R_\mu(\mathbf{A}_{S_n}(S), K+1) \leq \inf_{h\in\mathcal{H}_S} R_\mu(h, K+1) + \frac{C_\theta}{\sqrt{n^{1-\theta}}}, \tag{44}$$

where

$$\mathcal{H}_S = \{h \in \mathcal{H} : \sum_{i=1}^{n} \ell(h(\mathbf{x}_i), y_i) = 0\}, \text{ here } S = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\} \sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}},$$

and

$$R_\mu(h, K+1) = \mathbb{E}_{\mathbf{x}\sim\mu}\ell(h(\mathbf{x}), K+1) = \int_{\mathcal{X}} \ell(h(\mathbf{x}), K+1)\mathrm{d}\mu(\mathbf{x}).$$

Due to Realizability Assumption, we obtain that $\inf_{h\in\mathcal{H}} R^{\mathrm{in}}_D(h) = 0$. Therefore,

$$\mathbb{E}_{S\sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} \sup_{h\in\mathcal{H}_S} R^{\mathrm{in}}_D(h) \leq \frac{C_\theta}{\sqrt{n^{1-\theta}}}, \tag{45}$$

which implies that (in following inequalities, g is the groundtruth labeling function, *i.e.*, $R_D(g) = 0$)

$$\begin{aligned}\frac{C_\theta}{\sqrt{n}} \geq \mathbb{E}_{S\sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} \sup_{h\in\mathcal{H}_S} R^{\mathrm{in}}_D(h) =& \mathbb{E}_{S\sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} \sup_{h\in\mathcal{H}_S} \int_{g<K+1} \ell(h(\mathbf{x}), g(\mathbf{x}))f_{\mathrm{I}}(\mathbf{x})\mathrm{d}\mu(\mathbf{x}) \\ \geq& \frac{2}{b}\mathbb{E}_{S\sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} \sup_{h\in\mathcal{H}_S} \int_{g<K+1} \ell(h(\mathbf{x}), g(\mathbf{x}))\mathrm{d}\mu(\mathbf{x}).\end{aligned}$$

This implies that (here we have used the property of zero-one loss)

$$\mathbb{E}_{S\sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} \inf_{h\in\mathcal{H}_S} \int_{g<K+1} \ell(h(\mathbf{x}), K+1)\mathrm{d}\mu(\mathbf{x}) \geq \mu(\mathbf{x}\in\mathcal{X} : g(\mathbf{x}) < K+1) - \frac{C_\theta b}{2\sqrt{n^{1-\theta}}}.$$

Therefore,

$$\mathbb{E}_{S\sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} \inf_{h\in\mathcal{H}_S} R_\mu(h, K+1) \geq \mu(\mathbf{x}\in\mathcal{X} : g(\mathbf{x}) < K+1) - \frac{C_\theta b}{2\sqrt{n^{1-\theta}}}. \tag{46}$$

Additionally, $R_\mu(g, K+1) = \mu(\mathbf{x} \in \mathcal{X} : g(\mathbf{x}) < K+1)$ and $g \in \mathcal{H}_S$, which implies that

$$\inf_{h \in \mathcal{H}_S} R_\mu(h, K+1) \leq \mu(\mathbf{x} \in \mathcal{X} : g(\mathbf{x}) < K+1). \tag{47}$$

Combining inequalities (46) and (47), we obtain that

$$\left|\mathbb{E}_{S \sim D^n_{X_{\mathrm{I}} Y_{\mathrm{I}}}} \inf_{h \in \mathcal{H}_S} R_\mu(h, K+1) - \mu(\mathbf{x} \in \mathcal{X} : g(\mathbf{x}) < K+1)\right| \leq \frac{C_\theta b}{2\sqrt{n^{1-\theta}}}. \tag{48}$$

Using inequalities (44) and (48), we obtain that

$$\left|\mathbb{E}_{S \sim D^n_{X_{\mathrm{I}} Y_{\mathrm{I}}}} \mathbb{E}_{S_n \sim \mu^n} R_\mu(\mathbf{A}_{S_n}(S), K+1) - \mu(\mathbf{x} \in \mathcal{X} : g(\mathbf{x}) < K+1)\right| \leq \frac{C_\theta (b+1)}{\sqrt{n^{1-\theta}}}.$$

By Fubini theorem, we have that

$$\left|\mathbb{E}_{S_n \sim \mu^n} \mathbb{E}_{S \sim D^n_{X_{\mathrm{I}} Y_{\mathrm{I}}}} R_\mu(\mathbf{A}_{S_n}(S), K+1) - \mu(\mathbf{x} \in \mathcal{X} : g(\mathbf{x}) < K+1)\right| \leq \frac{C_\theta (b+1)}{\sqrt{n^{1-\theta}}}. \tag{49}$$

Using inequality (45), we have

$$\mathbb{E}_{S_n \sim \mu^n} \mathbb{E}_{S \sim D^n_{X_{\mathrm{I}} Y_{\mathrm{I}}}} R^{\mathrm{in}}_D(\mathbf{A}_{S_n}(S)) \leq \frac{C_\theta}{\sqrt{n^{1-\theta}}}, \tag{50}$$

which implies that (here we use the property of zero-one loss)

$$\left|\mathbb{E}_{S_n \sim \mu^n} \mathbb{E}_{S \sim D^n_{X_{\mathrm{I}} Y_{\mathrm{I}}}} \int_{g < K+1} \ell(\mathbf{A}_{S_n}(S)(\mathbf{x}), K+1) \mathrm{d}\mu(\mathbf{x}) - \mu(\mathbf{x} \in \mathcal{X} : g(\mathbf{x}) < K+1)\right| \leq \frac{2bC_\theta}{\sqrt{n^{1-\theta}}}. \tag{51}$$

Combining inequalities (49) and (51), we have

$$\left|\mathbb{E}_{S_n \sim \mu^n} \mathbb{E}_{S \sim D^n_{X_{\mathrm{I}} Y_{\mathrm{I}}}} \int_{g = K+1} \ell(\mathbf{A}_{S_n}(S)(\mathbf{x}), K+1) \mathrm{d}\mu(\mathbf{x})\right| \leq \frac{2bC_\theta}{\sqrt{n^{1-\theta}}} + \frac{C_\theta (b+1)}{\sqrt{n^{1-\theta}}}.$$

Therefore, there exist data points S'_n such that

$$\begin{aligned}
&\mathbb{E}_{S \sim D^n_{X_{\mathrm{I}} Y_{\mathrm{I}}}} R^{\mathrm{out}}_D(\mathbf{A}_{S'_n}) \\
=&\mathbb{E}_{S \sim D^n_{X_{\mathrm{I}} Y_{\mathrm{I}}}} \int_{g = K+1} \ell(\mathbf{A}_{S_n}(S)(\mathbf{x}), K+1) f_{\mathrm{O}}(\mathbf{x}) \mathrm{d}\mu(\mathbf{x}) \\
\leq& 2b \mathbb{E}_{S \sim D^n_{X_{\mathrm{I}} Y_{\mathrm{I}}}} \int_{g = K+1} \ell(\mathbf{A}_{S_n}(S)(\mathbf{x}), K+1) \mathrm{d}\mu(\mathbf{x}) \leq \frac{4b^2 C_\theta}{\sqrt{n^{1-\theta}}} + \frac{2C_\theta (b^2+b)}{\sqrt{n^{1-\theta}}}.
\end{aligned} \tag{52}$$

Combining inequalities (45) and (52), we obtain that for any n, there exists data points S'_n such that

$$\mathbb{E}_{S \sim D^n_{X_{\mathrm{I}} Y_{\mathrm{I}}}} R^\alpha_D(\mathbf{A}_{S'_n}) \leq \max\left\{\frac{4b^2 C_\theta}{\sqrt{n^{1-\theta}}} + \frac{2C_\theta (b^2+b)}{\sqrt{n^{1-\theta}}}, \frac{C_\theta}{\sqrt{n^{1-\theta}}}\right\}.$$

We set data point sequences $\mathbf{S}' = (S'_1, S'_2, ..., S'_n, ...)$. Then, $\mathbf{A}_{\mathbf{S}'} \in \mathscr{A}$ is the universally consistent algorithm, *i.e.*, for any $\alpha \in [0, 1]$

$$\mathbb{E}_{S \sim D^n_{X_{\mathrm{I}} Y_{\mathrm{I}}}} R^\alpha_D(\mathbf{A}_{\mathbf{S}}) \leq \max\left\{\frac{4b^2 C_\theta}{\sqrt{n^{1-\theta}}} + \frac{2C_\theta (b^2+b)}{\sqrt{n^{1-\theta}}}, \frac{C_\theta}{\sqrt{n^{1-\theta}}}\right\}.$$

We have completed this proof when ℓ is the zero-one loss.

Second, we prove the case that ℓ is not the zero-one loss. We use the notation ℓ_{0-1} as the zero-one loss. According the definition of loss introduced in Section 2, we know that there exists a constant $M > 0$ such that for any $y_1, y_2 \in \mathcal{Y}_{\mathrm{all}}$,

$$\frac{1}{M} \ell_{0-1}(y_1, y_2) \leq \ell(y_1, y_2) \leq M \ell_{0-1}(y_1, y_2).$$

Hence,

$$\frac{1}{M}R_D^{\alpha,\ell_{0-1}}(h) \leq R_D^{\alpha,\ell}(h) \leq MR_D^{\alpha,\ell_{0-1}}(h),$$

where $R_D^{\alpha,\ell_{0-1}}$ is the α-risk with zero-one loss, and $R_D^{\alpha,\ell}$ is the α-risk for loss ℓ.
Above inequality tells us that Realizability Assumption holds with zero-one loss if and only if Realizability Assumption holds with the loss ℓ. Therefore, we use the result proven in first step. We can find a consistent algorithm $\mathbf{A}$ such that for any $\alpha \in [0,1]$,

$$\mathbb{E}_{S\sim D^n_{X_1Y_1}} R_D^{\alpha,\ell_{0-1}}(\mathbf{A}) \leq O(\frac{1}{\sqrt{n^{1-\theta}}}),$$

which implies that for any $\alpha \in [0,1]$,

$$\frac{1}{M}\mathbb{E}_{S\sim D^n_{X_1Y_1}} R_D^{\alpha,\ell}(\mathbf{A}) \leq O(\frac{1}{\sqrt{n^{1-\theta}}}).$$

We have completed this proof. □

K Proof of Proposition 1 and Proof of Proposition 2

To better understand the contents in Appendices K-M, we introduce the important notations for FCNN-based hypothesis space and score-based hypothesis space detaily.

FCNN-based Hypothesis Space. Given a sequence $\mathbf{q} = (l_1, l_2, ..., l_g)$, where l_i and g are positive integers and $g > 2$, we use g to represent the depth of neural network and use l_i to represent the width of the i-th layer. After the activation function σ is selected, we can obtain the architecture of FCNN according to the sequence $\mathbf{q}$. Given any weights $\mathbf{w}_i \in \mathbb{R}^{l_i \times l_{i-1}}$ and bias $\mathbf{b}_i \in \mathbb{R}^{l_i \times 1}$, the output of the i-layer can be written as follows: for any $\mathbf{x} \in \mathbb{R}^{l_1}$,

$$\mathbf{f}_i(\mathbf{x}) = \sigma(\mathbf{w}_i \mathbf{f}_{i-1}(\mathbf{x}) + \mathbf{b}_i),\ \forall i = 2, ..., g-1,$$

where $\mathbf{f}_{i-1}(\mathbf{x})$ is the i-th layer output and $\mathbf{f}_1(\mathbf{x}) = \mathbf{x}$. Then, the output of FCNN is $\mathbf{f}_{\mathbf{w},\mathbf{b}}(\mathbf{x}) = \mathbf{w}_g \mathbf{f}_{g-1}(\mathbf{x}) + \mathbf{b}_g$, where $\mathbf{w} = \{\mathbf{w}_2, ..., \mathbf{w}_g\}$ and $\mathbf{b} = \{\mathbf{b}_2, ..., \mathbf{b}_g\}$.

An FCNN-based scoring function space is defined as:

$$\mathcal{F}_{\mathbf{q}}^{\sigma} := \{\mathbf{f}_{\mathbf{w},\mathbf{b}} : \forall \mathbf{w}_i \in \mathbb{R}^{l_i \times l_{i-1}},\ \forall \mathbf{b}_i \in \mathbb{R}^{l_i \times 1},\ i = 2, ..., g\}.$$

Additionally, given two sequences $\mathbf{q} = (l_1, ..., l_g)$ and $\mathbf{q}' = (l'_1, ..., l'_{g'})$, we use the notation $\mathbf{q} \lesssim \mathbf{q}'$ to represent the following equations and inequalities:

$$\begin{aligned} g \leq g',\quad l_1 = l'_1,\quad l_g = l'_{g'}, \\ l_i \leq l'_i,\quad \forall i = 1, ..., g-1, \\ l_{g-1} \leq l'_i,\quad \forall i = g, ..., g'-1. \end{aligned}$$

Given a sequence $\mathbf{q} = (l_1, ...l_g)$ satisfying that $l_1 = d$ and $l_g = K+1$, the FCNN-based scoring function space $\mathcal{F}_{\mathbf{q}}^{\sigma}$ can induce an FCNN-based hypothesis space. Before defining the FCNN-based hypothesis space, we define the induced hypothesis function. For any $\mathbf{f}_{\mathbf{w},\mathbf{b}} \in \mathcal{F}_{\mathbf{q}}^{\sigma}$, the induced hypothesis function is:

$$h_{\mathbf{w},\mathbf{b}}(\mathbf{x}) := \underset{k\in\{1,...,K+1\}}{\arg\max}\ f^k_{\mathbf{w},\mathbf{b}}(\mathbf{x}),\ \forall \mathbf{x} \in \mathcal{X},$$

where $f^k_{\mathbf{w},\mathbf{b}}(\mathbf{x})$ is the k-th coordinate of $\mathbf{f}_{\mathbf{w},\mathbf{b}}(\mathbf{x})$. Then, we define the FCNN-based hypothesis space as follows:

$$\mathcal{H}_{\mathbf{q}}^{\sigma} := \{h_{\mathbf{w},\mathbf{b}} : \forall \mathbf{w}_i \in \mathbb{R}^{l_i \times l_{i-1}},\ \forall \mathbf{b}_i \in \mathbb{R}^{l_i \times 1},\ i = 2, ..., g\}.$$

Score-based Hypothesis Space. Many OOD algorithms detect OOD data using a score-based strategy. That is, given a threshold λ, a scoring function space $\mathcal{F}_l \subset \{\mathbf{f} : \mathcal{X} \to \mathbb{R}^l\}$ and a scoring function $E : \mathcal{F}_l \to \mathbb{R}$, then $\mathbf{x}$ is regarded as ID, if $E(\mathbf{f}(\mathbf{x})) \geq \lambda$; otherwise, $\mathbf{x}$ is regarded as OOD.

Using E, λ and $\mathbf{f} \in \mathcal{F}_{\mathbf{q}}^{\sigma}$, we can generate a binary classifier $h_{\mathbf{f},E}^{\lambda}$:

$$h_{\mathbf{f},E}^{\lambda}(\mathbf{x}) := \begin{cases} 1, & \text{if } E(\mathbf{f}(\mathbf{x})) \geq \lambda; \\ 2, & \text{if } E(\mathbf{f}(\mathbf{x})) < \lambda, \end{cases}$$

where 1 represents ID data, and 2 represents OOD data. Hence, a binary classification hypothesis space $\mathcal{H}^b$, which consists of all $h_{\mathbf{f},E}^{\lambda}$, is generated. We define the score-based hypothesis space $\mathcal{H}_{\mathbf{q},E}^{\sigma,\lambda} := \{h_{\mathbf{f},E}^{\lambda} : \forall \mathbf{f} \in \mathcal{F}_{\mathbf{q}}^{\sigma}\}$.

Next, we introduce two important propositions.

Proposition 1. *Given a sequence $\mathbf{q} = (l_1, ...l_g)$ satisfying that $l_1 = d$ and $l_g = K+1$ (note that d is the dimension of input data and $K+1$ is the dimension of output), then the constant functions h_1, h_2,...,h_{K+1} belong to $\mathcal{H}_{\mathbf{q}}^{\sigma}$, where $h_i(\mathbf{x}) = i$, for any $\mathbf{x} \in \mathcal{X}$. Therefore, Assumption 1 holds for $\mathcal{H}_{\mathbf{q}}^{\sigma}$.*

Proof of Proposition 1. Note that the output of FCNN can be written as

$$\mathbf{f}_{\mathbf{w},\mathbf{b}}(\mathbf{x}) = \mathbf{w}_g \mathbf{f}_{g-1}(\mathbf{x}) + \mathbf{b}_g,$$

where $\mathbf{w}_g \in \mathbb{R}^{(K+1)\times l_{g-1}}$, $\mathbf{b}_g \in \mathbb{R}^{(K+1)\times 1}$ and $\mathbf{f}_{g-1}(\mathbf{x})$ is the output of the l_{g-1}-th layer. If we set $\mathbf{w}_g = \mathbf{0}$, and set $\mathbf{b}_g = \mathbf{y}_i$, where $\mathbf{y}_i$ is the one-hot vector corresponding to label i. Then $\mathbf{f}_{\mathbf{w},\mathbf{b}}(\mathbf{x}) = \mathbf{y}_i$, for any $\mathbf{x} \in \mathcal{X}$. Therefore, $h_i(\mathbf{x}) \in \mathcal{H}_{\mathbf{q}}^{\sigma}$, for any $i = 1, ..., K, K+1$. □

Note that in some works [78], $\mathbf{b}_g$ is fixed to $\mathbf{0}$. In fact, it is easy to check that when $g > 2$ and activation function σ is not a constant, Proposition 1 still holds, even if $\mathbf{b}_g = \mathbf{0}$.

Proposition 2. *For any sequence $\mathbf{q} = (l_1, ..., l_g)$ satisfying that $l_1 = d$ and $l_g = l$ (note that d is the dimension of input data and l is the dimension of output), if $\{\mathbf{v} \in \mathbb{R}^l : E(\mathbf{v}) \geq \lambda\} \neq \emptyset$ and $\{\mathbf{v} \in \mathbb{R}^l : E(\mathbf{v}) < \lambda\} \neq \emptyset$, then the functions h_1 and h_2 belong to $\mathcal{H}_{\mathbf{q},E}^{\sigma,\lambda}$, where $h_1(\mathbf{x}) = 1$ and $h_2(\mathbf{x}) = 2$, for any $\mathbf{x} \in \mathcal{X}$, where 1 represents the ID labels, and 2 represents the OOD labels. Therefore, Assumption 1 holds.*

Proof of Proposition 2. Since $\{\mathbf{v} \in \mathbb{R}^l : E(\mathbf{v}) \geq \lambda\} \neq \emptyset$ and $\{\mathbf{v} \in \mathbb{R}^l : E(\mathbf{v}) < \lambda\} \neq \emptyset$, we can find $\mathbf{v}_1 \in \{\mathbf{v} \in \mathbb{R}^l : E(\mathbf{v}) \geq \lambda\}$ and $\mathbf{v}_2 \in \{\mathbf{v} \in \mathbb{R}^l : E(\mathbf{v}) < \lambda\}$.

For any $\mathbf{f}_{\mathbf{w},\mathbf{b}} \in \mathcal{F}_{\mathbf{q}}^{\sigma}$, we have

$$\mathbf{f}_{\mathbf{w},\mathbf{b}}(\mathbf{x}) = \mathbf{w}_g \mathbf{f}_{g-1}(\mathbf{x}) + \mathbf{b}_g,$$

where $\mathbf{w}_g \in \mathbb{R}^{l\times l_{g-1}}$, $\mathbf{b}_g \in \mathbb{R}^{l\times 1}$ and $\mathbf{f}_{g-1}(\mathbf{x})$ is the output of the l_{g-1}-th layer.

If we set $\mathbf{w}_g = \mathbf{0}_{l\times l_{g-1}}$ and $\mathbf{b}_g = \mathbf{v}_1$, then $\mathbf{f}_{\mathbf{w},\mathbf{b}}(\mathbf{x}) = \mathbf{v}_1$ for any $\mathbf{x} \in \mathcal{X}$, where $\mathbf{0}_{l\times l_{g-1}}$ is $l \times l_{g-1}$ zero matrix. Hence, h_1 can be induced by $\mathbf{f}_{\mathbf{w},\mathbf{b}}$. Therefore, $h_1 \in \mathcal{H}_{\mathbf{q},E}^{\sigma,\lambda}$.

Similarly, if we set $\mathbf{w}_g = \mathbf{0}_{l\times l_{g-1}}$ and $\mathbf{b}_g = \mathbf{v}_2$, then $\mathbf{f}_{\mathbf{w},\mathbf{b}}(\mathbf{x}) = \mathbf{v}_2$ for any $\mathbf{x} \in \mathcal{X}$, where $\mathbf{0}_{l\times l_{g-1}}$ is $l \times l_{g-1}$ zero matrix. Hence, h_2 can be induced by $\mathbf{f}_{\mathbf{w},\mathbf{b}}$. Therefore, $h_2 \in \mathcal{H}_{\mathbf{q},E}^{\sigma,\lambda}$. □

It is easy to check that when $g > 2$ and activation function σ is not a constant, Proposition 2 still holds, even if $\mathbf{b}_g = \mathbf{0}$.

L Proof of Theorem 10

Before proving Theorem 10, we need several lemmas.

Lemma 9. *Let σ be ReLU function: $\max\{x, 0\}$. Given $\mathbf{q} = (l_1, ..., l_g)$ and $\mathbf{q}' = (l'_1, ..., l'_g)$ such that $l_g = l'_g$ and $l_1 = l'_1$, and $l_i \leq l'_i$ $(i = 1, ..., g-1)$, then $\mathcal{F}_{\mathbf{q}}^{\sigma} \subset \mathcal{F}_{\mathbf{q}'}^{\sigma}$ and $\mathcal{H}_{\mathbf{q}}^{\sigma} \subset \mathcal{H}_{\mathbf{q}'}^{\sigma}$.*

Proof of Lemma 9. Given any weights $\mathbf{w}_i \in \mathbb{R}^{l_i\times l_{i-1}}$ and bias $\mathbf{b}_i \in \mathbb{R}^{l_i\times 1}$, the i-layer output of FCNN with architecture $\mathbf{q}$ can be written as

$$\mathbf{f}_i(\mathbf{x}) = \sigma(\mathbf{w}_i \mathbf{f}_{i-1}(\mathbf{x}) + \mathbf{b}_i), \quad \forall \mathbf{x} \in \mathbb{R}^{l_1}, \forall i = 2, ..., g-1,$$

where $\mathbf{f}_{i-1}(\mathbf{x})$ is the i-th layer output and $\mathbf{f}_1(\mathbf{x}) = \mathbf{x}$. Then, the output of last layer is

$$\mathbf{f}_{\mathbf{w},\mathbf{b}}(\mathbf{x}) = \mathbf{w}_g \mathbf{f}_{g-1}(\mathbf{x}) + \mathbf{b}_g.$$

We will show that $\mathbf{f}_{\mathbf{w},\mathbf{b}} \in \mathcal{F}^{\sigma}_{\mathbf{q}'}$. We construct $\mathbf{f}_{\mathbf{w}',\mathbf{b}'}$ as follows: for every $\mathbf{w}'_i \in \mathbb{R}^{l'_i \times l'_{i-1}}$, if $l'_i - l_i > 0$ and $l'_{i-1} - l_{i-1} > 0$, we set

$$\mathbf{w}'_i = \begin{bmatrix} \mathbf{w}_i & \mathbf{0}_{l_i \times (l'_{i-1} - l_{i-1})} \\ \mathbf{0}_{(l'_i - l_i) \times l'_{i-1}} & \mathbf{0}_{(l'_i - l_i) \times (l'_{i-1} - l_{i-1})} \end{bmatrix}, \quad \mathbf{b}'_i = \begin{bmatrix} \mathbf{b}_i \\ \mathbf{0}_{(l'_i - l_i) \times 1} \end{bmatrix}$$

where $\mathbf{0}_{pq}$ means the $p \times q$ zero matrix. If $l'_i - l_i = 0$ and $l'_{i-1} - l_{i-1} > 0$, we set

$$\mathbf{w}'_i = \begin{bmatrix} \mathbf{w}_i & \mathbf{0}_{l_i \times (l'_{i-1} - l_{i-1})} \end{bmatrix}, \quad \mathbf{b}'_i = \mathbf{b}_i.$$

If $l'_{i-1} - l_{i-1} = 0$ and $l'_i - l_i > 0$, we set

$$\mathbf{w}'_i = \begin{bmatrix} \mathbf{w}_i \\ \mathbf{0}_{(l'_i - l_i) \times l'_{i-1}} \end{bmatrix}, \quad \mathbf{b}'_i = \begin{bmatrix} \mathbf{b}_i \\ \mathbf{0}_{(l'_i - l_i) \times 1} \end{bmatrix}.$$

If $l'_{i-1} - l_{i-1} = 0$ and $l'_i - l_i = 0$, we set

$$\mathbf{w}'_i = \mathbf{w}_i, \quad \mathbf{b}'_i = \mathbf{b}_i.$$

It is easy to check that if $l'_i - l_i > 0$

$$\mathbf{f}'_i = \begin{bmatrix} \mathbf{f}_i \\ \mathbf{0}_{(l'_i - l_i) \times 1} \end{bmatrix}.$$

If $l'_i - l_i = 0$,

$$\mathbf{f}'_i = \mathbf{f}_i.$$

Since $l'_g - l_g = 0$,

$$\mathbf{f}'_g = \mathbf{f}_g, \ \textit{i.e.}, \ \mathbf{f}_{\mathbf{w}',\mathbf{b}'} = \mathbf{f}_{\mathbf{w},\mathbf{b}}.$$

Therefore, $f_{\mathbf{w},\mathbf{b}} \in \mathcal{F}^{\sigma}_{\mathbf{q}'}$, which implies that $\mathcal{F}^{\sigma}_{\mathbf{q}} \subset \mathcal{F}^{\sigma}_{\mathbf{q}'}$. Therefore, $\mathcal{H}^{\sigma}_{\mathbf{q}} \subset \mathcal{H}^{\sigma}_{\mathbf{q}'}$. □

Lemma 10. *Let σ be the ReLU function: $\sigma(x) = \max\{x, 0\}$. Then, $\mathbf{q} \lesssim \mathbf{q}'$ implies that $\mathcal{F}^{\sigma}_{\mathbf{q}} \subset \mathcal{F}^{\sigma}_{\mathbf{q}'}$, $\mathcal{H}^{\sigma}_{\mathbf{q}} \subset \mathcal{H}^{\sigma}_{\mathbf{q}'}$, where $\mathbf{q} = (l_1, ..., l_g)$ and $\mathbf{q}' = (l'_1, ..., l'_{g'})$.*

Proof of Lemma 10. Given $l'' = (l''_1, ..., l''_{g''})$ satisfying that $g \leq g''$, $l''_i = l_i$ for $i = 1, ..., g-1$, $l''_i = l_{g-1}$ for $i = g, ..., g''-1$, and $l''_{g''} = l_g$, we first prove that $\mathcal{F}^{\sigma}_{\mathbf{q}} \subset \mathcal{F}^{\sigma}_{\mathbf{q}''}$ and $\mathcal{H}^{\sigma}_{\mathbf{q}} \subset \mathcal{H}^{\sigma}_{\mathbf{q}''}$.

Given any weights $\mathbf{w}_i \in \mathbb{R}^{l_i \times l_{i-1}}$ and bias $\mathbf{b}_i \in \mathbb{R}^{l_i \times 1}$, the i-th layer output of FCNN with architecture $\mathbf{q}$ can be written as

$$\mathbf{f}_i(\mathbf{x}) = \sigma(\mathbf{w}_i \mathbf{f}_{i-1}(\mathbf{x}) + \mathbf{b}_i), \ \ \forall \mathbf{x} \in \mathbb{R}^{l_1}, \forall i = 2, ..., g-1,$$

where $\mathbf{f}_{i-1}(\mathbf{x})$ is the i-th layer output and $\mathbf{f}_1(\mathbf{x}) = \mathbf{x}$. Then, the output of the last layer is

$$\mathbf{f}_{\mathbf{w},\mathbf{b}}(\mathbf{x}) = \mathbf{w}_g \mathbf{f}_{g-1}(\mathbf{x}) + \mathbf{b}_g.$$

We will show that $\mathbf{f}_{\mathbf{w},\mathbf{b}} \in \mathcal{F}^{\sigma}_{\mathbf{q}''}$. We construct $\mathbf{f}_{\mathbf{w}'',\mathbf{b}''}$ as follows: if $i = 2, ..., g-1$, then $\mathbf{w}''_i = \mathbf{w}$ and $\mathbf{b}''_i = \mathbf{b}_i$; if $i = g, ..., g''-1$, then $\mathbf{w}''_i = \mathbf{I}_{l_{g-1} \times l_{g-1}}$ and $\mathbf{b}''_i = \mathbf{0}_{l_{g-1} \times 1}$, where $\mathbf{I}_{l_{g-1} \times l_{g-1}}$ is the $l_{g-1} \times l_{g-1}$ identity matrix, and $\mathbf{0}_{l_{g-1} \times 1}$ is the $l_{g-1} \times 1$ zero matrix; and if $i = g''$, then $\mathbf{w}''_{g''} = \mathbf{w}_g$, $\mathbf{b}''_{g''} = \mathbf{b}_g$. Then it is easy to check that the output of the i-th layer is

$$\mathbf{f}''_i = \mathbf{f}_{g-1}, \forall i = g-1, g, ..., g''-1.$$

Therefore, $\mathbf{f}_{\mathbf{w}'',\mathbf{b}''} = \mathbf{f}_{\mathbf{w},\mathbf{b}}$, which implies that $\mathcal{F}^{\sigma}_{\mathbf{q}} \subset \mathcal{F}^{\sigma}_{\mathbf{q}''}$. Hence, $\mathcal{H}^{\sigma}_{\mathbf{q}} \subset \mathcal{H}^{\sigma}_{\mathbf{q}''}$.

When $g'' = g'$, we use Lemma 9 ($\mathbf{q}''$ and $\mathbf{q}$ satisfy the condition in Lemma 9), which implies that $\mathcal{F}^{\sigma}_{\mathbf{q}''} \subset \mathcal{F}^{\sigma}_{\mathbf{q}'}$, $\mathcal{H}^{\sigma}_{\mathbf{q}''} \subset \mathcal{H}^{\sigma}_{\mathbf{q}'}$. Therefore, $\mathcal{F}^{\sigma}_{\mathbf{q}} \subset \mathcal{F}^{\sigma}_{\mathbf{q}'}$, $\mathcal{H}^{\sigma}_{\mathbf{q}} \subset \mathcal{H}^{\sigma}_{\mathbf{q}'}$. □

Lemma 11. *[79] If the activation function σ is not a polynomial, then for any continuous function f defined in $\mathbb{R}^d$, and any compact set $C \subset \mathbb{R}^d$, there exists a fully-connected neural network with architecture $\mathbf{q}$ $(l_1 = d, l_g = 1)$ such that*

$$\inf_{f_{\mathbf{w},\mathbf{b}} \in \mathcal{F}_{\mathbf{q}}^{\sigma}} \max_{\mathbf{x} \in C} |f_{\mathbf{w},\mathbf{b}}(\mathbf{x}) - f(\mathbf{x})| < \epsilon.$$

Proof of Lemma 11. The proof of Lemma 11 can be found in Theorem 3.1 in [79]. □

Lemma 12. *If the activation function σ is the ReLU function, then for any continuous vector-valued function $\mathbf{f} \in C(\mathbb{R}^d; \mathbb{R}^l)$, and any compact set $C \subset \mathbb{R}^d$, there exists a fully-connected neural network with architecture $\mathbf{q}$ $(l_1 = d, l_g = l)$ such that*

$$\inf_{\mathbf{f}_{\mathbf{w},\mathbf{b}} \in \mathcal{F}_{\mathbf{q}}^{\sigma}} \max_{\mathbf{x} \in C} \|\mathbf{f}_{\mathbf{w},\mathbf{b}}(\mathbf{x}) - \mathbf{f}(\mathbf{x})\|_2 < \epsilon,$$

where $\|\cdot\|_2$ is the ℓ_2 norm. (Note that we can also prove the same result, if σ is not a polynomial.)

Proof of Lemma 12. Let $\mathbf{f} = [f_1, ..., f_l]^\top$, where f_i is the i-th coordinate of $\mathbf{f}$. Based on Lemma 11, we obtain l sequences $\mathbf{q}^1$, $\mathbf{q}^2$,...,$\mathbf{q}^l$ such that

$$\inf_{g_1 \in \mathcal{F}_{\mathbf{q}^1}^{\sigma}} \max_{\mathbf{x} \in C} |g_1(\mathbf{x}) - f_1(\mathbf{x})| < \epsilon/\sqrt{l},$$

$$\inf_{g_2 \in \mathcal{F}_{\mathbf{q}^2}^{\sigma}} \max_{\mathbf{x} \in C} |g_2(\mathbf{x}) - f_2(\mathbf{x})| < \epsilon/\sqrt{l},$$

$$...$$

$$...$$

$$\inf_{g_l \in \mathcal{F}_{\mathbf{q}^l}^{\sigma}} \max_{\mathbf{x} \in C} |g_l(\mathbf{x}) - f_l(\mathbf{x})| < \epsilon/\sqrt{l}.$$

It is easy to find a sequence $\mathbf{q} = (l_1, ..., l_g)$ $(l_g = 1)$ such that $\mathbf{q}^i \lesssim \mathbf{q}$, for all $i = 1, ..., l$. Using Lemma 10, we obtain that $\mathcal{F}_{\mathbf{q}^i}^{\sigma} \subset \mathcal{F}_{\mathbf{q}}^{\sigma}$. Therefore,

$$\inf_{g \in \mathcal{F}_{\mathbf{q}}^{\sigma}} \max_{\mathbf{x} \in C} |g(\mathbf{x}) - f_1(\mathbf{x})| < \epsilon/\sqrt{l},$$

$$\inf_{g \in \mathcal{F}_{\mathbf{q}}^{\sigma}} \max_{\mathbf{x} \in C} |g(\mathbf{x}) - f_2(\mathbf{x})| < \epsilon/\sqrt{l},$$

$$...$$

$$...$$

$$\inf_{g \in \mathcal{F}_{\mathbf{q}}^{\sigma}} \max_{\mathbf{x} \in C} |g(\mathbf{x}) - f_l(\mathbf{x})| < \epsilon/\sqrt{l}.$$

Therefore, for each i, we can find $g_{\mathbf{w}^i,\mathbf{b}^i}$ from $\mathcal{F}_{\mathbf{q}}^{\sigma}$ such that

$$\max_{\mathbf{x} \in C} |g_{\mathbf{w}^i,\mathbf{b}^i}(\mathbf{x}) - f_i(\mathbf{x})| < \epsilon/\sqrt{l},$$

where $\mathbf{w}^i$ represents weights and $\mathbf{b}^i$ represents bias.

We construct a larger FCNN with $\mathbf{q}' = (l'_1, l'_2, ..., l'_g)$ satisfying that $l'_1 = d$, $l'_i = l * l_i$, for $i = 2, ..., g$. We can regard this larger FCNN as a combinations of l FCNNs with architecture $\mathbf{q}$, that is: there are m disjoint sub-FCNNs with architecture $\mathbf{q}$ in the larger FCNN with architecture $\mathbf{q}'$. For i-th sub-FCNN, we use weights $\mathbf{w}^i$ and bias $\mathbf{b}^i$. For weights and bias which connect different sub-FCNNs, we set these weights and bias to $\mathbf{0}$. Finally, we can obtain that $\mathbf{g}_{\mathbf{w},\mathbf{b}} = [g_{\mathbf{w}^1,\mathbf{b}^1}, g_{\mathbf{w}^2,\mathbf{b}^2}, ..., g_{\mathbf{w}^l,\mathbf{b}^l}]^\top \in \mathcal{F}_{\mathbf{q}'}^{\sigma}$, which implies that

$$\max_{\mathbf{x} \in C} \|\mathbf{g}_{\mathbf{w},\mathbf{b}}(\mathbf{x}) - \mathbf{f}(\mathbf{x})\|_2 < \epsilon.$$

We have completed this proof. □

Given a sequence $\mathbf{q} = (l_1, ..., l_g)$, we are interested in following function space $\mathcal{F}^{\sigma}_{\mathbf{q},\mathbf{M}}$:

$$\mathcal{F}^{\sigma}_{\mathbf{q},\mathbf{M}} := \{\mathbf{M} \cdot (\sigma \circ \mathbf{f}) : \forall\, \mathbf{f} \in \mathcal{F}^{\sigma}_{\mathbf{q}}\},$$

where $\circ$ means the composition of two functions, $\cdot$ means the product of two matrices, and

$$\mathbf{M} = \begin{bmatrix} \mathbf{1}_{1\times(l_g-1)} & 0 \\ \mathbf{0}_{1\times(l_g-1)} & 1 \end{bmatrix},$$

here $\mathbf{1}_{1\times(l_g-1)}$ is the $1 \times (l_g - 1)$ matrix whose all elements are 1, and $\mathbf{0}_{1\times(l_g-1)}$ is the $1 \times (l_g - 1)$ zero matrix. Using $\mathcal{F}^{\sigma}_{\mathbf{q},\mathbf{M}}$, we can construct a binary classification space $\mathcal{H}^{\sigma}_{\mathbf{q},\mathbf{M}}$, which consists of all classifiers satisfying the following condition:

$$h(\mathbf{x}) = \underset{k=\{1,2\}}{\arg\min}\, f^k_{\mathbf{M}}(\mathbf{x}),$$

where $f^k_{\mathbf{M}}(\mathbf{x})$ is the k-th coordinate of $\mathbf{M} \cdot (\sigma \circ \mathbf{f})$.

Lemma 13. *Suppose that σ is the ReLU function: $\max\{x, 0\}$. Given a sequence $\mathbf{q} = (l_1, ..., l_g)$ satisfying that $l_1 = d$ and $l_g = K + 1$, then the space $\mathcal{H}^{\sigma}_{\mathbf{q},\mathbf{M}}$ contains $\phi \circ \mathcal{H}^{\sigma}_{\mathbf{q}}$, and $\mathcal{H}^{\sigma}_{\mathbf{q},\mathbf{M}}$ has finite VC dimension (Vapnik–Chervonenkis dimension), where ϕ maps ID data to 1 and OOD data to 2. Furthermore, if given $\mathbf{q}' = (l'_1, ..., l'_g)$ satisfying that $l'_g = K$ and $l'_i = l_i$, for $i = 1, ..., g - 1$, then $\mathcal{H}^{\sigma}_{\mathbf{q}} \subset \mathcal{H}^{\sigma}_{\mathbf{q}'} \bullet \mathcal{H}^{\sigma}_{\mathbf{q},\mathbf{M}}$.*

Proof of Lemma 13. For any $h_{\mathbf{w},\mathbf{b}} \in \mathcal{H}^{\sigma}_{\mathbf{q}}$, then there exists $\mathbf{f}_{\mathbf{w},\mathbf{b}} \in \mathcal{F}^{\sigma}_{\mathbf{q}}$ such that $h_{\mathbf{w},\mathbf{b}}$ is induced by $\mathbf{f}_{\mathbf{w},\mathbf{b}}$. We can write $\mathbf{f}_{\mathbf{w},\mathbf{b}}$ as follows:

$$\mathbf{f}_{\mathbf{w},\mathbf{b}}(\mathbf{x}) = \mathbf{w}_g \mathbf{f}_{g-1}(\mathbf{x}) + \mathbf{b}_g,$$

where $\mathbf{w}_g \in \mathbb{R}^{(K+1)\times l_{g-1}}$, $\mathbf{b}_g \in \mathbb{R}^{(K+1)\times 1}$ and $\mathbf{f}_{g-1}(\mathbf{x})$ is the output of the l_{g-1}-th layer.

Suppose that

$$\mathbf{w}_g = \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ ... \\ \mathbf{v}_K \\ \mathbf{v}_{K+1} \end{bmatrix}, \quad \mathbf{b}_g = \begin{bmatrix} b_1 \\ b_2 \\ ... \\ b_K \\ b_{K+1} \end{bmatrix},$$

where $\mathbf{v}_i \in \mathbb{R}^{1\times l_{g-1}}$ and $b_i \in \mathbb{R}$.

We set

$$\mathbf{f}_{\mathbf{w}',\mathbf{b}'}(\mathbf{x}) = \mathbf{w}'_g \mathbf{f}_{g-1}(\mathbf{x}) + \mathbf{b}'_g,$$

where

$$\mathbf{w}'_g = \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ ... \\ \mathbf{v}_K \end{bmatrix}, \quad \mathbf{b}'_g = \begin{bmatrix} b_1 \\ b_2 \\ ... \\ b_K \end{bmatrix},$$

It is obvious that $\mathbf{f}_{\mathbf{w}',\mathbf{b}'} \in \mathcal{F}^{\sigma}_{\mathbf{q}'}$. Using $\mathbf{f}_{\mathbf{w}',\mathbf{b}'} \in \mathcal{F}^{\sigma}_{\mathbf{q}'}$, we construct a classifier $h_{\mathbf{w}',\mathbf{b}'} \in \mathcal{H}^{\sigma}_{\mathbf{q}'}$:

$$h_{\mathbf{w}',\mathbf{b}'} = \underset{k\in\{1,...,K\}}{\arg\max}\, f^k_{\mathbf{w}',\mathbf{b}'},$$

where $f^k_{\mathbf{w}',\mathbf{b}'}$ is the k-th coordinate of $\mathbf{f}_{\mathbf{w}',\mathbf{b}'}$.

Additionally, we consider

$$\mathbf{f}_{\mathbf{w},\mathbf{b},\mathbf{B}} = \mathbf{M} \cdot \sigma(\mathbf{B} \cdot \mathbf{f}_{\mathbf{w},\mathbf{b}}) \in \mathcal{F}^{\sigma}_{\mathbf{q},\mathbf{M}},$$

where

$$\mathbf{B} = \begin{bmatrix} \mathbf{I}_{(l_g-1)\times(l_g-1)} & -\mathbf{1}_{(l_g-1)\times 1} \\ \mathbf{0}_{1\times(l_g-1)} & 0 \end{bmatrix},$$

here $\mathbf{I}_{(l_g-1)\times(l_g-1)}$ is the $(l_g - 1) \times (l_g - 1)$ identity matrix, $\mathbf{0}_{1\times(l_g-1)}$ is the $1 \times (l_g - 1)$ zero matrix, and $\mathbf{1}_{(l_g-1)\times 1}$ is the $(l_g - 1) \times 1$ matrix, whose all elements are 1.

Then, we define that for any $\mathbf{x} \in \mathcal{X}$,

$$h_{\mathbf{w},\mathbf{b},\mathbf{B}}(\mathbf{x}) := \arg\max_{k \in \{1,2\}} f^k_{\mathbf{w},\mathbf{b},\mathbf{B}}(\mathbf{x}),$$

where $f^k_{\mathbf{w},\mathbf{b},\mathbf{B}}(\mathbf{x})$ is the k-th coordinate of $\mathbf{f}_{\mathbf{w},\mathbf{b},\mathbf{B}}(\mathbf{x})$. Furthermore, we can check that $h_{\mathbf{w},\mathbf{b},\mathbf{B}}$ can be written as follows: for any $\mathbf{x} \in \mathcal{X}$,

$$h_{\mathbf{w},\mathbf{b},\mathbf{B}}(\mathbf{x}) = \begin{cases} 1, & \text{if } f^1_{\mathbf{w},\mathbf{b},\mathbf{B}}(\mathbf{x}) > 0; \\ 2, & \text{if } f^1_{\mathbf{w},\mathbf{b},\mathbf{B}}(\mathbf{x}) \leq 0. \end{cases}$$

It is easy to check that

$$h_{\mathbf{w},\mathbf{b},\mathbf{B}} = \phi \circ h_{\mathbf{w},\mathbf{b}},$$

where ϕ maps ID labels to 1 and OOD labels to 2.

Therefore, $h_{\mathbf{w},\mathbf{b}}(\mathbf{x}) = K+1$ if and only if $h_{\mathbf{w},\mathbf{b},\mathbf{B}} = 2$; and $h_{\mathbf{w},\mathbf{b}}(\mathbf{x}) = k$ ($k \neq K+1$) if and only if $h_{\mathbf{w},\mathbf{b},\mathbf{B}} = 1$ and $h_{\mathbf{w}',\mathbf{b}'}(\mathbf{x}) = k$. This implies that $\mathcal{H}^\sigma_{\mathbf{q}} \subset \mathcal{H}^\sigma_{\mathbf{q}'} \bullet \mathcal{H}^\sigma_{\mathbf{q},\mathbf{M}}$ and $\phi \circ \mathcal{H}^\sigma_{\mathbf{q}} \subset \mathcal{H}^\sigma_{\mathbf{q},\mathbf{M}}$.

Let $\tilde{\mathbf{q}}$ be $(l_1, ..., l_g, 2)$. Then $\mathcal{F}^\sigma_{\mathbf{q},\mathbf{M}} \subset \mathcal{F}^\sigma_{\tilde{\mathbf{q}}}$. Hence, $\mathcal{H}^\sigma_{\mathbf{q},\mathbf{M}} \subset \mathcal{H}^\sigma_{\tilde{\mathbf{q}}}$. According to the VC dimension theory [37] for feed-forward neural networks, $\mathcal{H}^\sigma_{\tilde{\mathbf{q}}}$ has finite VC dimension. Hence, $\mathcal{H}^\sigma_{\mathbf{q},\mathbf{M}}$ has finite VC-dimension. We have completed the proof. □

Lemma 14. *Let $|\mathcal{X}| < +\infty$ and σ be the ReLU function: $\max\{x, 0\}$. Given r hypothesis functions $h_1, h_2, ..., h_r \in \{h : \mathcal{X} \to \{1, ..., l\}\}$, then there exists a sequence $\mathbf{q} = (l_1, ..., l_g)$ with $l_1 = d$ and $l_g = l$, such that $h_1, ..., h_r \in \mathcal{H}^\sigma_{\mathbf{q}}$.*

Proof of Lemma 14. For each h_i ($i = 1, ..., r$), we introduce a corresponding $\mathbf{f}_i$ (defined over $\mathcal{X}$) satisfying that for any $\mathbf{x} \in \mathcal{X}$, $\mathbf{f}_i(\mathbf{x}) = \mathbf{y}_k$ if and only if $h_i(\mathbf{x}) = k$, where $\mathbf{y}_k \in \mathbb{R}^l$ is the one-hot vector corresponding to the label k. Clearly, $\mathbf{f}_i$ is a continuous function in $\mathcal{X}$, because $\mathcal{X}$ is a discrete set. Tietze Extension Theorem implies that $\mathbf{f}_i$ can be extended to a continuous function in $\mathbb{R}^d$.

Since $\mathcal{X}$ is a compact set, then Lemma 12 implies that there exist a sequence $\mathbf{q}^i = (l^i_1, ..., l^i_{g^i})$ ($l^i_1 = d$ and $l^i_{g^i} = l$) and $\mathbf{f}_{\mathbf{w},\mathbf{b}} \in \mathcal{F}^\sigma_{\mathbf{q}^i}$ such that

$$\max_{\mathbf{x} \in \mathcal{X}} \|\mathbf{f}_{\mathbf{w},\mathbf{b}}(\mathbf{x}) - \mathbf{f}_i(\mathbf{x})\|_{\ell_2} < \frac{1}{10 \cdot l},$$

where $\|\cdot\|_{\ell_2}$ is the ℓ_2 norm in $\mathbb{R}^l$. Therefore, for any $\mathbf{x} \in \mathcal{X}$, it easy to check that

$$\arg\max_{k \in \{1,...,l\}} f^k_{\mathbf{w},\mathbf{b}}(\mathbf{x}) = h_i(\mathbf{x}),$$

where $f^k_{\mathbf{w},\mathbf{b}}(\mathbf{x})$ is the k-th coordinate of $\mathbf{f}_{\mathbf{w},\mathbf{b}}(\mathbf{x})$. Therefore, $h_i(\mathbf{x}) \in \mathcal{H}^\sigma_{\mathbf{q}^i}$.

Let $\mathbf{q}$ be $(l_1, ..., l_g)$ ($l_1 = d$ and $l_g = l$) satisfying that $\mathbf{q}^i \lesssim \mathbf{q}$. Using Lemma 10, we obtain that $\mathcal{H}^\sigma_{\mathbf{q}^i} \subset \mathcal{H}^\sigma_{\mathbf{q}}$, for each $i = 1, ..., r$. Therefore, $h_1, ..., h_r \in \mathcal{H}^\sigma_{\mathbf{q}}$. □

Lemma 15. *Let the activation function σ be the ReLU function. Suppose that $|\mathcal{X}| < +\infty$. If $\{\mathbf{v} \in \mathbb{R}^l : E(\mathbf{v}) \geq \lambda\}$ and $\{\mathbf{v} \in \mathbb{R}^l : E(\mathbf{v}) < \lambda\}$ both contain nonempty open sets of $\mathbb{R}^l$ (here, open set is a topological terminology). There exists a sequence $\mathbf{q} = (l_1, ..., l_g)$ ($l_1 = d$ and $l_g = l$) such that $\mathcal{H}^{\sigma,\lambda}_{\mathbf{q},E}$ consists of all binary classifiers.*

Proof of Lemma 15. Since $\{\mathbf{v} \in \mathbb{R}^l : E(\mathbf{v}) \geq \lambda\}$, $\{\mathbf{v} \in \mathbb{R}^l : E(\mathbf{v}) < \lambda\}$ both contain nonempty open sets, we can find $\mathbf{v}_1 \in \{\mathbf{v} \in \mathbb{R}^l : E(\mathbf{v}) \geq \lambda\}$, $\mathbf{v}_2 \in \{\mathbf{v} \in \mathbb{R}^l : E(\mathbf{v}) < \lambda\}$ and a constant $r > 0$ such that $B_r(\mathbf{v}_1) \subset \{\mathbf{v} \in \mathbb{R}^l : E(\mathbf{v}) \geq \lambda\}$ and $B_r(\mathbf{v}_2) \subset \{\mathbf{v} \in \mathbb{R}^l : E(\mathbf{v}) < \lambda\}$, where $B_r(\mathbf{v}_1) = \{\mathbf{v} : \|\mathbf{v} - \mathbf{v}_1\|_{\ell_2} < r\}$ and $B_r(\mathbf{v}_2) = \{\mathbf{v} : \|\mathbf{v} - \mathbf{v}_2\|_{\ell_2} < r\}$, here $\|\cdot\|_{\ell_2}$ is the ℓ_2 norm.

For any binary classifier h over $\mathcal{X}$, we can induce a vector-valued function as follows: for any $\mathbf{x} \in \mathcal{X}$,

$$\mathbf{f}(\mathbf{x}) = \begin{cases} \mathbf{v}_1, & \text{if } h(\mathbf{x}) = 1; \\ \mathbf{v}_2, & \text{if } h(\mathbf{x}) = 2. \end{cases}$$

Since $\mathcal{X}$ is a finite set, then Tietze Extension Theorem implies that $\mathbf{f}$ can be extended to a continuous function in $\mathbb{R}^d$. Since $\mathcal{X}$ is a compact set, Lemma 12 implies that there exists a sequence $\mathbf{q}^h = (l_1^h, ..., l_{g^h}^h)$ ($l_1^h = d$ and $l_{g^h}^h = l$) and $\mathbf{f}_{\mathbf{w},\mathbf{b}} \in \mathcal{F}_{\mathbf{q}^h}^{\sigma}$ such that

$$\max_{\mathbf{x} \in \mathcal{X}} \|\mathbf{f}_{\mathbf{w},\mathbf{b}}(\mathbf{x}) - \mathbf{f}(\mathbf{x})\|_{\ell_2} < \frac{r}{2},$$

where $\|\cdot\|_{\ell_2}$ is the ℓ_2 norm in $\mathbb{R}^l$. Therefore, for any $\mathbf{x} \in \mathcal{X}$, it is easy to check that $E(\mathbf{f}_{\mathbf{w},\mathbf{b}}(\mathbf{x})) \geq \lambda$ if and only if $h(\mathbf{x}) = 1$, and $E(\mathbf{f}_{\mathbf{w},\mathbf{b}}(\mathbf{x})) < \lambda$ if and only if $h(\mathbf{x}) = 2$.

For each h, we have found a sequence $\mathbf{q}^h$ such that h is induced by $\mathbf{f}_{\mathbf{w},\mathbf{b}} \in \mathcal{F}_{\mathbf{q}^h}^{\sigma}$, E and λ. Since $|\mathcal{X}| < +\infty$, only finite binary classifiers are defined over $\mathcal{X}$. Using Lemma 14, we can find a sequence $\mathbf{q}$ such that $\mathcal{H}_{\text{all}}^b = \mathcal{H}_{\mathbf{q},E}^{\sigma,\lambda}$, where $\mathcal{H}_{\text{all}}^b$ consists of all binary classifiers. □

Lemma 16. *Suppose the hypothesis space is score-based. Let $|\mathcal{X}| < +\infty$. If $\{\mathbf{v} \in \mathbb{R}^l : E(\mathbf{v}) \geq \lambda\}$ and $\{\mathbf{v} \in \mathbb{R}^l : E(\mathbf{v}) < \lambda\}$ both contain nonempty open sets, and Condition 2 holds, then there exists a sequence $\mathbf{q} = (l_1, ..., l_g)$ ($l_1 = d$ and $l_g = l$) such that for any sequence $\mathbf{q}'$ satisfying $\mathbf{q} \lesssim \mathbf{q}'$ and any ID hypothesis space $\mathcal{H}^{\text{in}}$, OOD detection is learnable in the separate space $\mathscr{D}_{XY}^s$ for $\mathcal{H}^{\text{in}} \bullet \mathcal{H}^{\text{b}}$, where $\mathcal{H}^{\text{b}} = \mathcal{H}_{\mathbf{q}',E}^{\sigma,\lambda}$ and $\mathcal{H}^{\text{in}} \bullet \mathcal{H}^{\text{b}}$ is defined below Eq. (4).*

Proof of Lemma 16. Note that we use the ReLU function as the activation function in this lemma. Using Lemma 10, Lemma 15 and Theorem 7, we can prove this result. □

Theorem 10. *Suppose that Condition 2 holds and the hypothesis space $\mathcal{H}$ is FCNN-based or score-based, i.e., $\mathcal{H} = \mathcal{H}_{\mathbf{q}}^{\sigma}$ or $\mathcal{H} = \mathcal{H}^{\text{in}} \bullet \mathcal{H}^{\text{b}}$, where $\mathcal{H}^{\text{in}}$ is an ID hypothesis space, $\mathcal{H}^{\text{b}} = \mathcal{H}_{\mathbf{q},E}^{\sigma,\lambda}$ and $\mathcal{H} = \mathcal{H}^{\text{in}} \bullet \mathcal{H}^{\text{b}}$ is introduced below Eq. (4), here E is introduced in Eqs. (5) or (6). Then*

There is a sequence $\mathbf{q} = (l_1, ..., l_g)$ such that OOD detection is learnable in the separate space $\mathscr{D}_{XY}^s$ for $\mathcal{H}$ ***if and only if*** *$|\mathcal{X}| < +\infty$.*

Furthermore, if $|\mathcal{X}| < +\infty$, then there exists a sequence $\mathbf{q} = (l_1, ..., l_g)$ such that for any sequence $\mathbf{q}'$ satisfying that $\mathbf{q} \lesssim \mathbf{q}'$, OOD detection is learnable in $\mathscr{D}_{XY}^s$ for $\mathcal{H}$.

Proof of Theorem 10. Note that we use the ReLU function as the activation function in this theorem.

• The Case that $\mathcal{H}$ is FCNN-based.

First, we prove that if $|\mathcal{X}| = +\infty$, then OOD detection is not learnable in $\mathscr{D}_{XY}^s$ for $\mathcal{H}_{\mathbf{q}}^{\sigma}$, for any sequence $\mathbf{q} = (l_1, ..., l_g)$ ($l_1 = d$ and $l_g = K + 1$).

By Lemma 13, Theorems 5 and 8 in [80], we know that $\text{VCdim}(\phi \circ \mathcal{H}_{\mathbf{q}}^{\sigma}) < +\infty$, where ϕ maps ID data to 1 and maps OOD data to 2. Additionally, Proposition 1 implies that Assumption 1 holds and $\sup_{h \in \mathcal{H}_{\mathbf{q}}^{\sigma}} |\{\mathbf{x} \in \mathcal{X} : h(\mathbf{x}) \in \mathcal{Y}\}| = +\infty$, when $|\mathcal{X}| = +\infty$. Therefore, Theorem 5 implies that OOD detection is not learnable in $\mathscr{D}_{XY}^s$ for $\mathcal{H}_{\mathbf{q}}^{\sigma}$, when $|\mathcal{X}| = +\infty$.

Second, we prove that if $|\mathcal{X}| < +\infty$, there exists a sequence $\mathbf{q} = (l_1, ..., l_g)$ ($l_1 = d$ and $l_g = K+1$) such that OOD detection is learnable in $\mathscr{D}_{XY}^s$ for $\mathcal{H}_{\mathbf{q}}^{\sigma}$.

Since $|\mathcal{X}| < +\infty$, it is clear that $|\mathcal{H}_{\text{all}}| < +\infty$, where $\mathcal{H}_{\text{all}}$ consists of all hypothesis functions from $\mathcal{X}$ to $\mathcal{Y}_{\text{all}}$. According to Lemma 14, there exists a sequence $\mathbf{q}$ such that $\mathcal{H}_{\text{all}} \subset \mathcal{H}_{\mathbf{q}}^{\sigma}$. Additionally, Lemma 13 implies that there exist $\mathcal{H}^{\text{in}}$ and $\mathcal{H}^{\text{b}}$ such that $\mathcal{H}_{\mathbf{q}}^{\sigma} \subset \mathcal{H}^{\text{in}} \bullet \mathcal{H}^{\text{b}}$. Since $\mathcal{H}_{\text{all}}$ consists all hypothesis space, $\mathcal{H}_{\text{all}} = \mathcal{H}_{\mathbf{q}}^{\sigma} = \mathcal{H}^{\text{in}} \bullet \mathcal{H}^{\text{b}}$. Therefore, $\mathcal{H}^{\text{b}}$ contains all binary classifiers from $\mathcal{X}$ to $\{1, 2\}$. Theorem 7 implies that OOD detection is learnable in $\mathscr{D}_{XY}^s$ for $\mathcal{H}_{\mathbf{q}}^{\sigma}$.

Third, we prove that if $|\mathcal{X}| < +\infty$, then there exists a sequence $\mathbf{q} = (l_1, ..., l_g)$ ($l_1 = d$ and $l_g = K+1$) such that for any sequence $\mathbf{q}' = (l'_1, ..., l'_{g'})$ satisfying that $\mathbf{q} \lesssim \mathbf{q}'$, OOD detection is learnable in $\mathscr{D}^s_{XY}$ for $\mathcal{H}^\sigma_{\mathbf{q}'}$.

We can use the sequence $\mathbf{q}$ constructed in the second step of the proof. Therefore, $\mathcal{H}^\sigma_{\mathbf{q}} = \mathcal{H}_{\text{all}}$. Lemma 10 implies that $\mathcal{H}^\sigma_{\mathbf{q}} \subset \mathcal{H}^\sigma_{\mathbf{q}'}$. Therefore, $\mathcal{H}^\sigma_{\mathbf{q}'} = \mathcal{H}_{\text{all}} = \mathcal{H}^\sigma_{\mathbf{q}}$. The proving process (second step of the proof) has shown that if $|\mathcal{X}| < +\infty$, Condition 2 holds and hypothesis space $\mathcal{H}$ consists of all hypothesis functions, then OOD detection is learnable in $\mathscr{D}^s_{XY}$ for $\mathcal{H}$. Therefore, OOD detection is learnable in $\mathscr{D}^s_{XY}$ for $\mathcal{H}^\sigma_{\mathbf{q}'}$. We complete the proof when the hypothesis space $\mathcal{H}$ is FCNN-based.

• The Case that $\mathcal{H}$ is score-based

Fourth, we prove that if $|\mathcal{X}| = +\infty$, then OOD detection is not learnable in $\mathscr{D}^s_{XY}$ for $\mathcal{H}^{\text{in}} \bullet \mathcal{H}^{\text{b}}$, where $\mathcal{H}^{\text{b}} = \mathcal{H}^{\sigma,\lambda}_{\mathbf{q},E}$ for any sequence $\mathbf{q} = (l_1, ..., l_g)$ ($l_1 = d$, $l_g = l$), where E is in Eqs. (5) or (6).

By Theorems 5 and 8 in [80], we know that $\text{VCdim}(\mathcal{H}^{\sigma,\lambda}_{\mathbf{q},E}) < +\infty$. Additionally, Proposition 2 implies that Assumption 1 holds and $\sup_{h\in\mathcal{H}^\sigma_{\mathbf{q}}} |\{\mathbf{x} \in \mathcal{X} : h(\mathbf{x}) \in \mathcal{Y}\}| = +\infty$, when $|\mathcal{X}| = +\infty$. Hence, Theorem 5 implies that OOD detection is not learnable in $\mathscr{D}^s_{XY}$ for $\mathcal{H}^\sigma_{\mathbf{q}}$, when $|\mathcal{X}| = +\infty$.

Fifth, we prove that if $|\mathcal{X}| < +\infty$, there exists a sequence $\mathbf{q} = (l_1, ..., l_g)$ ($l_1 = d$ and $l_g = l$) such that OOD detection is learnable in $\mathscr{D}^s_{XY}$ for for $\mathcal{H}^{\text{in}} \bullet \mathcal{H}^{\text{b}}$, where $\mathcal{H}^{\text{b}} = \mathcal{H}^{\sigma,\lambda}_{\mathbf{q},E}$ for any sequence $\mathbf{q} = (l_1, ..., l_g)$ ($l_1 = d$, $l_g = l$), where E is in Eq. (5) or Eq. (6).

Based on Lemma 16, we only need to show that $\{\mathbf{v} \in \mathbb{R}^l : E(\mathbf{v}) \geq \lambda\}$ and $\{\mathbf{v} \in \mathbb{R}^l : E(\mathbf{v}) < \lambda\}$ both contain nonempty open sets for different scoring functions E.

Since $\max_{k\in\{1,...,l\}} \frac{\exp(v^k)}{\sum_{c=1}^l \exp(v^c)}$, $\max_{k\in\{1,...,l\}} \frac{\exp(v^k/T)}{\sum_{c=1}^K \exp(v^c/T)}$ and $T \log \sum_{c=1}^l \exp(v^c/T)$ are continuous functions, whose ranges contain $(\frac{1}{l}, 1)$, $(\frac{1}{l}, 1)$, $(0, +\infty)$ and $(0, +\infty)$, respectively.

Based on the property of continuous function ($E^{-1}(A)$ is an open set, if A is an open set), we obtain that $\{\mathbf{v} \in \mathbb{R}^l : E(\mathbf{v}) \geq \lambda\}$ and $\{\mathbf{v} \in \mathbb{R}^l : E(\mathbf{v}) < \lambda\}$ both contain nonempty open sets. Using Lemma 16, we complete the fifth step.

Sixth, we prove that if $|\mathcal{X}| < +\infty$, then there exists a sequence $\mathbf{q} = (l_1, ..., l_g)$ ($l_1 = d$ and $l_g = l$) such that for any sequence $\mathbf{q}' = (l'_1, ..., l'_{g'})$ satisfying that $\mathbf{q} \lesssim \mathbf{q}'$, OOD detection is learnable in $\mathscr{D}^s_{XY}$ for for $\mathcal{H}^{\text{in}} \bullet \mathcal{H}^{\text{b}}$, where $\mathcal{H}^{\text{b}} = \mathcal{H}^{\sigma,\lambda}_{\mathbf{q}',E}$, where E is in Eq. (5) or Eq. (6).

In the fifth step, we have proven that Eq. (5) and Eq. (6) meet the condition in Lemma 16. Therefore, Lemma 16 implies this result. We complete the proof when the hypothesis space $\mathcal{H}$ is score-based. □

M Proofs of Theorem 11 and Theorem 12

M.1 Proof of Theorem 11

Theorem 11. *Suppose that each domain D_{XY} in $\mathscr{D}^{\mu,b}_{XY}$ is attainable, i.e., $\arg\min_{h\in\mathcal{H}} R_D(h) \neq \emptyset$ (the finite discrete domains satisfy this). Let $K = 1$ and the hypothesis space $\mathcal{H}$ be score-based ($\mathcal{H} = \mathcal{H}^{\sigma,\lambda}_{\mathbf{q},E}$, where E is in Eqs. (5) or (6)) or FCNN-based ($\mathcal{H} = \mathcal{H}^\sigma_{\mathbf{q}}$). If $\mu(\mathcal{X}) < +\infty$, then the following four conditions are **equivalent**:*

Learnability in $\mathscr{D}^{\mu,b}_{XY}$ for $\mathcal{H}$ $\iff$ Condition 1 $\iff$ Realizability Assumption $\iff$ Condition 3

Proof of Theorem 11.

1) By Lemma 1, we conclude that Learnability in $\mathscr{D}^{\mu,b}_{XY}$ for $\mathcal{H} \Rightarrow$ Condition 1.

2) By Proposition 1 and Proposition 2, we know that when $K = 1$, there exist $h_1, h_2 \in \mathcal{H}$, where $h_1 = 1$ and $h_2 = 2$, here 1 represents ID, and 2 represent OOD. Therefore, we know that when $K = 1$, $\inf_{h\in\mathcal{H}} R^{\text{in}}_D(h) = 0$ and $\inf_{h\in\mathcal{H}} R^{\text{out}}_D(h) = 0$, for any $D_{XY} \in \mathscr{D}^{\mu,b}_{XY}$.

By Condition 1, we obtain that $\inf_{h\in\mathcal{H}} R_D = 0$, for any $D_{XY} \in \mathscr{D}^{\mu,b}_{XY}$. Because each domain D_{XY} in $\mathscr{D}^{\mu,b}_{XY}$ is attainable, we conclude that Realizability Assumption holds.

We have proven that Condition 1$\Rightarrow$ Realizability Assumption.

3) By Theorems 5 and 8 in [80], we know that $\mathrm{VCdim}(\phi \circ \mathcal{H}_{\mathbf{q}}^{\sigma}) < +\infty$ and $\mathrm{VCdim}(\mathcal{H}_{\mathbf{q},E}^{\sigma,\lambda}) < +\infty$. Then, using Theorem 9, we conclude that Realizability Assumption$\Rightarrow$ Learnability in $\mathscr{D}_{XY}^{\mu,b}$ for $\mathcal{H}$.

4) According to the results in 1), 2) and 3), we have proven that

$$\text{Learnability in } \mathscr{D}_{XY}^{\mu,b} \text{ for } \mathcal{H} \Leftrightarrow \text{Condition 1} \Leftrightarrow \text{Realizability Assumption.}$$

5) By Lemma 2, we conclude that Condition 3$\Rightarrow$Condition 1.

6) **Here we prove that Learnability in $\mathscr{D}_{XY}^{\mu,b}$ for $\mathcal{H}$ $\Rightarrow$Condition 3.** Since $\mathscr{D}_{XY}^{\mu,b}$ is the prior-unknown space, by Theorem 1, there exist an algorithm $\mathbf{A} : \cup_{n=1}^{+\infty}(\mathcal{X} \times \mathcal{Y})^n \to \mathcal{H}$ and a monotonically decreasing sequence $\epsilon_{\text{cons}}(n)$, such that $\epsilon_{\text{cons}}(n) \to 0$, as $n \to +\infty$, and for any $D_{XY} \in \mathscr{D}_{XY}^{\mu,b}$,

$$\mathbb{E}_{S\sim D_{X_{\mathrm{I}}Y_{\mathrm{I}}}^{n}}\left[R_D^{\text{in}}(\mathbf{A}(S)) - \inf_{h\in\mathcal{H}} R_D^{\text{in}}(h)\right] \leq \epsilon_{\text{cons}}(n),$$
$$\mathbb{E}_{S\sim D_{X_{\mathrm{I}}Y_{\mathrm{I}}}^{n}}\left[R_D^{\text{out}}(\mathbf{A}(S)) - \inf_{h\in\mathcal{H}} R_D^{\text{out}}(h)\right] \leq \epsilon_{\text{cons}}(n).$$

Then, for any $\epsilon > 0$, we can find n_ϵ such that $\epsilon \geq \epsilon_{\text{cons}}(n_\epsilon)$, therefore, if $n = n_\epsilon$, we have

$$\mathbb{E}_{S\sim D_{X_{\mathrm{I}}Y_{\mathrm{I}}}^{n_\epsilon}}\left[R_D^{\text{in}}(\mathbf{A}(S)) - \inf_{h\in\mathcal{H}} R_D^{\text{in}}(h)\right] \leq \epsilon,$$
$$\mathbb{E}_{S\sim D_{X_{\mathrm{I}}Y_{\mathrm{I}}}^{n_\epsilon}}\left[R_D^{\text{out}}(\mathbf{A}(S)) - \inf_{h\in\mathcal{H}} R_D^{\text{out}}(h)\right] \leq \epsilon,$$

which implies that there exists $S_\epsilon \sim D_{X_{\mathrm{I}}Y_{\mathrm{I}}}^{n_\epsilon}$ such that

$$R_D^{\text{in}}(\mathbf{A}(S_\epsilon)) - \inf_{h\in\mathcal{H}} R_D^{\text{in}}(h) \leq \epsilon,$$
$$R_D^{\text{out}}(\mathbf{A}(S_\epsilon)) - \inf_{h\in\mathcal{H}} R_D^{\text{out}}(h) \leq \epsilon.$$

Therefore, for any equivalence class $[D'_{XY}]$ with respect to $\mathscr{D}_{XY}^{\mu,b}$ and any $\epsilon > 0$, there exists a hypothesis function $\mathbf{A}(S_\epsilon) \in \mathcal{H}$ such that for any domain $D_{XY} \in [D'_{XY}]$,

$$\mathbf{A}(S_\epsilon) \in \{h' \in \mathcal{H} : R_D^{\text{out}}(h') \leq \inf_{h\in\mathcal{H}} R_D^{\text{out}}(h) + \epsilon\} \cap \{h' \in \mathcal{H} : R_D^{\text{in}}(h') \leq \inf_{h\in\mathcal{H}} R_D^{\text{in}}(h) + \epsilon\},$$

which implies that Condition 3 holds. Therefore, Learnability in $\mathscr{D}_{XY}^{\mu,b}$ for $\mathcal{H}$ $\Rightarrow$Condition 3.

7) Note that in 4), 5) and 6), we have proven that

Learnability in $\mathscr{D}_{XY}^{\mu,b}$ for $\mathcal{H}$ $\Rightarrow$Condition 3$\Rightarrow$Condition 1, and Learnability in $\mathscr{D}_{XY}^{\mu,b}$ for $\mathcal{H}$ $\Leftrightarrow$Condition 1, thus, we conclude that Learnability in $\mathscr{D}_{XY}^{\mu,b}$ for $\mathcal{H}$ $\Leftrightarrow$Condition 3$\Leftrightarrow$Condition 1.

8) Combining 4) and 7), we have completed the proof.

□

M.2 Proof of Theorem 12

Theorem 12. *Let $K = 1$ and the hypothesis space $\mathcal{H}$ be score-based ($\mathcal{H} = \mathcal{H}_{\mathbf{q},E}^{\sigma,\lambda}$, where E is in Eqs. (5) or (6)) or FCNN-based ($\mathcal{H} = \mathcal{H}_{\mathbf{q}}^{\sigma}$). Given a prior-unknown space $\mathscr{D}_{XY}$, if there exists a domain $D_{XY} \in \mathscr{D}_{XY}$, which has an overlap between ID and OOD distributions (see Definition 4), then OOD detection is not learnable in the domain space $\mathscr{D}_{XY}$ for $\mathcal{H}$.*

Proof of Theorem 12. Using Proposition 1 and Proposition 2, we obtain that $\inf_{h\in\mathcal{H}} R_D^{\text{in}}(h) = 0$ and $\inf_{h\in\mathcal{H}} R_D^{\text{out}}(h) = 0$. Then, Theorem 3 implies this result. □

Note that if we replace the activation function σ (ReLU function) in Theorem 12 with any other activation functions, Theorem 12 still hold.