Journal of Machine Learning Research 25 (2024) 1-83 Submitted 9/23; Published 4/24

On the Learnability of Out-of-distribution Detection

Zhen Fang ZHEN.FANG@UTS.EDU.AU
Australian Artificial Intelligence Institute
University of Technology Sydney
61 Broadway, Ultimo NSW 2007, Australia

Yixuan Li SHARONLI@CS.WISC.EDU
Department of Computer Sciences
The University of Wisconsin Madison
1210 W Dayton St, Madison, WI 53706, USA

Feng Liu ✉ FENG.LIU1@UNIMELB.EDU.AU
School of Computing and Information Systems
The University of Melbourne
700 Swanston Street, Carlton VIC 3053, Australia

Bo Han BHANML@COMP.HKBU.EDU.HK
Department of Computer Science
Hong Kong Baptist University
Kowloon Tong, Hong Kong SAR

Jie Lu ✉ JIE.LU@UTS.EDU.AU
Australian Artificial Intelligence Institute
University of Technology Sydney
61 Broadway, Ultimo NSW 2007, Australia

Editor: Amos Storkey

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, and corresponding learning theory is still an *open problem*. To study the generalization of OOD detection, this paper investigates the *probably approximately correct* (PAC) learning theory of OOD detection that fits the commonly used evaluation metrics in the literature. 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 offer theoretical support for representative OOD detection works based on our OOD theory.

Keywords: out-of-distribution detection, weakly supervised learning, learnability

1. Introduction

The success of supervised learning is established on an *in-distribution* (ID) assumption that training and test data share the same distribution (Dosovitskiy et al., 2021; Huang et al., 2017; Hsu et al., 2020; Yang et al., 2021). However, in many real-world scenarios, the distribution of test data violates the assumption and, instead, contains *out-of-distribution* (OOD) data whose labels have not been seen during the training process (Bendale and Boult, 2016; Chen et al., 2021a). To mitigate the risk brought by OOD data, a more practical learning scenario is considered in the machine learning field: OOD detection, which determines whether an input is ID/OOD, while classifying the ID data into respective classes.

OOD detection can significantly increase the reliability of machine learning models when deploying them in the real world. Many seminar algorithms have been developed to empirically address the OOD detection problem (Hendrycks and Gimpel, 2017; Liang et al., 2018; Lee et al., 2018; Zong et al., 2018; Pidhorskyi et al., 2018; Nalisnick et al., 2019; Hendrycks et al., 2019; Ren et al., 2019; Lin et al., 2021; Salehi et al., 2021; Sun et al., 2021). A common solution paradigm to OOD detection is to propose a new learning objective or/and a score function to identify if one upcoming data point is OOD data. When evaluating algorithms under this solution paradigm, both threshold-dependent metrics (*e.g.*, risk) and threshold-independent metrics (*e.g.*, AUC) will be used to see to what extent the algorithms can successfully identify OOD data. 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, a theoretical framework is proposed to understand the learnability of OOD detection problem in view of threshold-dependent metrics and threshold-independent metrics[1]. We investigate the probably approximately correct (PAC) learning theory of OOD detection when the evaluation metrics are risk and AUC, 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. Because OOD data can be diverse in many real-world scenarios, we want to study whether there exists an algorithm that can be used to detect data from various OOD distributions instead of merely data from some specified OOD distributions. Such is the significance of studying the learning theory for OOD detection (Yang et al., 2021). 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 under two common metrics: risk and AUC?*

To answer the above research question and investigate the learning theory, we mainly focus on two basic spaces: domain space and function space. The domain space is a space consisting of some distributions, and the function space is a space consisting of some classifiers or ranking functions. Existing agnostic PAC theories in supervised learning (Shalev-Shwartz and Ben-David, 2014; Mohri et al., 2018) are distribution-free, *i.e.*, the domain space consists of all domains. Yet, in Theorem 5 and Theorem 8, we show that the learning theory of OOD detection is not distribution-free. Furthermore, we find that OOD detection is learnable only if the domain space and the function space satisfy some special conditions, *e.g.,* Conditions

1. This paper is an extended version of our previous conference paper (Fang et al., 2022a). In Section 7, we discuss the main difference between this paper and Fang et al. (2022a).

1, 2, and 4. 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 under risk and AUC metrics. From these essential conditions, we can know *when* OOD detection can be successful in practical scenarios. We restate our question and goal in the following:

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

Main Results. We start to study the learnability of OOD detection in the largest space—the total space, and give two necessary conditions for the learnability of OOD detection under risk and AUC (Condition 1 for risk and Condition 2 for AUC). However, we find that the overlap between ID and OOD data may result in that both necessary conditions do not hold. Therefore, we give two impossibility theorems to demonstrate that OOD detection fails in the total space (Theorem 5 under risk and Theorem 8 under AUC). Then, we investigate OOD detection in a separate space, where the ID and OOD data do not overlap. Unfortunately, there still exists impossibility theorems (Theorem 6 under risk and Theorem 9 under AUC), meaning that we cannot expect OOD detection is learnable under risk and AUC in the separate space under some conditions of the developed theorems.

It is frustrating to find the impossibility theorems regarding OOD detection in a separate space, but we find that some conditions of these impossibility theorems may not hold in several practical scenarios. Stemming from this observation, we give several *necessary and sufficient* conditions to characterize the learnability of OOD detection under risk and AUC in the separate space (Theorems 10 and 16 under risk, and Theorems 12 and 17 under AUC). Especially, when our function space is based on *fully-connected neural network* (FCNN), OOD detection is learnable under risk and AUC in the separate space if and only if the feature space is finite. Then, we focus on other more practical domain spaces, *e.g.*, the finite-ID-distribution space and the density-based space and investigate the learnability of OOD detection in both spaces. Theorem 13 shows a necessary and sufficient condition of learnability of OOD detection under risk. Theorems 14 and 15 show two sufficient conditions for the learnability of OOD detection under risk and AUC, respectively. It should be noted that when studying learnability of OOD detection in the finite-ID-distribution space, we discover a compatibility condition (Condition 4) that is a necessary and sufficient condition of learnability of OOD detection under risk for this space. Then, we explore 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 18).

Implications and Impacts of Theory. Our study is not of purely theoretical interest; it has also practical impacts. (i) From the perspective of domain space, we consider the finite-ID-distribution space that fits the common scenarios in the real world: we normally only have finite ID datasets. In this case, Theorem 13 gives a necessary and sufficient condition to the success of OOD detection under risk. More importantly, 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. (ii) From the perspective of function space, we investigate the learnability of OOD detection under risk and AUC for commonly used FCNN-based function spaces. Our theory provides theoretical support (Theorems 16 and 18 under risk, and Theorems 17 and 19 under AUC) for several representative OOD detection works (Hendrycks and Gimpel, 2017; Liang et al., 2018; Liu et al., 2020). (iii) From the perspective of evaluation metrics, our paper studies the learnability of OOD detection under risk and the learnability of OOD detection under AUC, which covers the major evaluation metrics used in OOD detection evaluation and provides theoretical guidance when users have different requirement in evaluating OOD detection performance. Based on all of our theoretical results, they suggest we should not expect a universally working OOD detection algorithm. It is necessary to design different algorithms in different scenarios.

2. Learning Setups

We begin 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 $\mathcal{Y} := \{1, \ldots, K\}$, 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 encounter 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 we 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 $\pi^{\mathrm{out}} \in [0, 1)$ represents an unknown class-prior probability. Next, we provide the formal definition of the OOD detection problem and key concepts used in this paper.

2.1 Problem Setting and Concepts

Problem 1 (OOD Detection (Yang et al., 2021)) *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:*

- *if $\mathbf{x}$ is an observation from D_{X_I}, f can classify $\mathbf{x}$ into correct ID classes;*
- *if $\mathbf{x}$ is an observation from D_{X_O}, f can detect $\mathbf{x}$ as OOD data.*

According to Yang et al. (2021), when $K = 1$, OOD detection reduces to one-class novelty detection or semantic anomaly detection (Ruff et al., 2018; Goyal et al., 2020; Deecke et al., 2018). Next, we introduce some basic and important concepts and notations.

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.*, $Y_\mathrm{O} = K + 1$ (Fang et al., 2021, 2020). 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 $\mathcal{H}^{\text{in}} \subset \{h : \mathcal{X} \to \mathcal{Y}\}$ to the ID hypothesis space. We also define $\mathcal{H}^{\text{b}} \subset \{h : \mathcal{X} \to \{1, 2\}\}$ 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.

Ranking Function Spaces. Most representative OOD detection algorithms (Liu et al., 2020) output a ranking function from a given ranking function space $\mathcal{R} \subset \mathcal{R}_{\text{all}} = \{r : \mathcal{X} \to \mathbb{R}\}$. If the ranking function $r(\mathbf{x})$ has a higher value, then $\mathbf{x}$ is from $D_{X_{\text{I}}}$ with a higher probability. A perfect ranking function r^* fulfills the condition $r^*(\mathbf{x}) > r^*(\mathbf{x}')$ for all $\mathbf{x}$ from $D_{X_{\text{I}}}$ and all $\mathbf{x}'$ from $D_{X_{\text{O}}}$, indicating that rankings of ID data are always higher than rankings of OOD data. The general strategy to construct the ranking function space $\mathcal{R}$ is to design a scoring function $E : \mathbb{R}^l \to \mathbb{R}$ and integrate it with the scoring function space $\mathcal{F}_l$, i.e., $\mathcal{R} = E \circ \mathcal{F}_l$.

Loss, Risks and AUC Metric. Let $\mathcal{Y}_{\text{all}} = \mathcal{Y} \cup \{K+1\}$. Given a loss function $\ell : \mathcal{Y}_{\text{all}} \times \mathcal{Y}_{\text{all}} \to \mathbb{R}_{\geq 0}$ 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 $R_D^{\alpha}(h) := (1-\alpha)R_D^{\text{in}}(h) + \alpha R_D^{\text{out}}(h), \forall \alpha \in [0,1]$, where $R_D^{\text{in}}(h)$ and $R_D^{\text{out}}(h)$ are

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

Except for using risk to evaluate the OOD detection performance, AUC is also a promising metric to see if a ranking function r can separate the ID and OOD data:

$$\text{AUC}(r; D_{XY}) = \mathbb{E}_{\mathbf{x}\sim D_{X_{\text{I}}}} \mathbb{E}_{\mathbf{x}'\sim D_{X_{\text{O}}}} \big[\mathbf{1}_{r(\mathbf{x})>r(\mathbf{x}')} + \frac{1}{2}\mathbf{1}_{r(\mathbf{x})=r(\mathbf{x}')}\big]. \tag{2}$$

Note that since value of AUC only denpends on the marginal distributions $D_{X_{\text{I}}}$ and $D_{X_{\text{O}}}$, therefore, it is also convientent for us to rewrite $\text{AUC}(r; D_{XY})$ as $\text{AUC}(r; D_{X_{\text{I}}}, D_{X_{\text{O}}})$.

2.2 Learnability under Risk

Based on risk defined in Eq. (1), OOD detection aims 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 under risk. Formally, we have:

Condition 1 (Learnability of OOD Detection under Risk) *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 **learnable** in $\mathscr{D}_{XY}$ for $\mathcal{H}$ under risk, if there exists an algorithm $\mathbf{A}$[2] $: \cup_{n=1}^{+\infty}(\mathcal{X} \times \mathcal{Y})^n \to \mathcal{H}$ and a*

2. Similar to Shalev-Shwartz et al. (2010), in this paper, we regard an algorithm as a mapping from $\cup_{n=1}^{+\infty}(\mathcal{X} \times \mathcal{Y})^n$ to $\mathcal{H}$ or $\mathcal{R}$.

monotonically decreasing sequence $\epsilon_{\text{cons}}(n)$*, such that* $\epsilon_{\text{cons}}(n) \to 0$*, as* $n \to +\infty$*, and for any domain* $D_{XY} \in \mathscr{D}_{XY}$*,*

$$\mathbb{E}_{S\sim D^n_{X_{\text{I}}Y_{\text{I}}}}\left[R_D(\mathbf{A}(S)) - \inf_{h\in\mathcal{H}} R_D(h)\right] \leq \epsilon_{\text{cons}}(n), \tag{3}$$

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 (Shalev-Shwartz et al., 2010). If for any $D_{XY} \in \mathscr{D}_{XY}$, $\pi^{\text{out}} = 0$, 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 Shalev-Shwartz and Ben-David (2014), one can prove that they are equivalent if ℓ is bounded, see Appendix A.3. Since OOD data are unavailable, it is impossible to obtain any 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 mitigate this challenge, we revise Eq. (3) as follows:

$$\mathbb{E}_{S\sim D^n_{X_{\text{I}}Y_{\text{I}}}}\left[R^\alpha_D(\mathbf{A}(S)) - \inf_{h\in\mathcal{H}} R^\alpha_D(h)\right] \leq \epsilon_{\text{cons}}(n),\ \forall\alpha \in [0, 1]. \tag{4}$$

If an algorithm $\mathbf{A}$ satisfies Eq. (4), 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 under risk as follows:

Condition 2 (Strong Learnability of OOD Detection under Risk) *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* $D_{XY} \in \mathscr{D}_{XY}$*,*

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

Remark. In Theorem 1, we have shown that the strong learnability of OOD detection under risk is equivalent to the learnability of OOD detection under risk, if the domain space $\mathscr{D}_{XY}$ is a *prior-unknown space* (see Definition 4). In this paper, we mainly discuss the learnability in the prior-unknown space. Therefore, *when we mention that OOD detection is learnable under risk, we also mean that OOD detection is strongly learnable under risk.*

2.3 Learnability under AUC

Based on AUC defined in Eq. (2), OOD detection aims to select a ranking function $r \in \mathcal{R}$ with approximately maximal AUC, 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 under AUC. Formally, we have:

Condition 3 (Learnability of OOD Detection under AUC) *Given a domain space* $\mathscr{D}_{XY}$*, a ranking function space* $\mathcal{R} \subset \{r : \mathcal{X} \to \mathbb{R}\}$*, we say OOD detection is* ***learnable***

in $\mathscr{D}_{XY}$ *for* $\mathcal{R}$ *under AUC, if there exists an algorithm* $\mathbf{A} : \cup_{n=1}^{+\infty}(\mathcal{X} \times \mathcal{Y})^n \to \mathcal{R}$ *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* $D_{XY} \in \mathscr{D}_{XY}$*,*

$$\mathbb{E}_{S\sim D^n_{X_\text{I}Y_\text{I}}}\big[\sup_{r\in\mathcal{R}} \text{AUC}(r; D_{XY}) - \text{AUC}(\mathbf{A}(S); D_{XY})\big] \leq \epsilon_{\text{cons}}(n). \tag{5}$$

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

Definition 3 is another version of Definition 1. Here, we use AUC instead of risk to evaluate the performance of the OOD detection. Note that the learnability of OOD detection under AUC is not influenced by the π_{out}, as AUC is directly calculated by using D_{X_I} and D_{X_O}.

2.4 Goal of Our 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 (Liang et al., 2018; Ren et al., 2019; Fang et al., 2021), it is obvious that the learnability of OOD detection is not distribution-free (*i.e.*, Theorem 5 and Theorem 8). 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 $\mathcal{H}$ (or the ranking function space $\mathcal{R}$). That is, OOD detection is learnable only when the domain space $\mathscr{D}_{XY}$ and the hypothesis space $\mathcal{H}$ (or the ranking function space $\mathcal{R}$) satisfy some special conditions, *e.g.,* Conditions 1, 4 (under risk), Conditions 2 (under AUC). We present our goal as follows:

> ***Goal:*** *given a hypothesis space* $\mathcal{H}$ *(or a ranking function space* $\mathcal{R}$*), 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 and prove that if OOD detection is strongly learnable under risk or learnable under AUC 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 under risk or AUC. These results imply that it is enough to study learnabiligy of OOD detection in the prior-unknown spaces. The prior-unknown space is as follows:

Condition 4 *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* $D^{\alpha}_{XY} := (1-\alpha)D_{X_\text{I}Y_\text{I}} + \alpha D_{X_\text{O}Y_\text{O}} \in \mathscr{D}_{XY}$*.*

Then the following theorem presents importance and necessity of priori-unknown space.

Theorem 1 *Given 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 $\mathscr{D}_{XY}$ under risk ***if and only if*** *OOD detection is learnable in $\mathscr{D}'_{XY}$ under risk;*
4) OOD detection is learnable in $\mathscr{D}_{XY}$ under AUC ***if and only if*** *OOD detection is learnable in $\mathscr{D}'_{XY}$ under AUC.*

The second result of Theorem 1 bridges the learnability and strong learnability under risk, which implies that if an algorithm **A** is consistent with respect to a prior-unknown space, then this algorithm **A** can address the imbalance issue between ID and OOD distributions, and the priori-unknown issue well. The fourth result of Theorem 1 shows that the learnability of OOD detection under AUC is not influenced by the unknown class-prior probability π^{out}. Based on Theorem 1, we focus on our theory in the prior-unknown spaces. 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 $\text{supp}D$ means the support set of a distribution D.
- 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 $0.5 * D_{X_{\text{I}}} + 0.5 * D_{X_{\text{O}}} = \int f \text{d}\mu$, 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 $\mathscr{D}^{s}_{XY}$; 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 $\mathscr{D}^{F}_{XY}$ 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 $\mathscr{D}^{\text{all}}_{XY}$ and the separate space $\mathscr{D}^{s}_{XY}$.

4.1 Necessary Conditions for Learnability of OOD Detection

We first find a necessary condition for the learnability of OOD detection under risk (AUC), *i.e.*, Condition 1 (Condition 2).

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

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

The importance of Condition 1 is reflected by Theorem 2, showing that Condition 1 is a *necessary and sufficient* condition for the learnability of OOD detection under risk 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 under risk in the single-distribution space $\mathscr{D}_{XY}^{D_{XY}}$ for $\mathcal{H}$* ***if and only if*** *Condition 1 holds.*

Theorem 2 implies that Condition 1 is important for the learnability of OOD detection under risk. Due to the simplicity of single-distribution space, Theorem 2 implies that Condition 1 is the necessary condition for the learnability of OOD detection under risk in the prior-unknown space, see Lemma 1 in Appendix C. Then, we focus on finding a necessary condition for the learnability of OOD detection under AUC. The condition is similar to Condition 1 but replacing risk with AUC. Note that, for simplicity, in the following of this paper, we use $\text{AUC}(r; D_{X_\text{I}}, D_{X_\text{O}})$ to present $\text{AUC}(r; D_{XY})$.

Condition 2 (Linear Condition under AUC) *For any $D_{XY} = \beta D_{X_\text{I}Y_\text{I}} + (1-\beta) D_{X_\text{O}Y_\text{O}}$, $D'_{XY} = \beta' D_{X_\text{I}Y_\text{I}} + (1-\beta') D'_{X_\text{O}Y_\text{O}} \in \mathscr{D}_{XY}$, then for any $\alpha \in [0,1)$,*

$$\alpha \sup_{r\in\mathcal{R}} \text{AUC}(r; D_{X_\text{I}}, D_{X_\text{O}}) + (1-\alpha) \sup_{r\in\mathcal{R}} \text{AUC}(r; D_{X_\text{I}}, D'_{X_\text{O}}) = \sup_{r\in\mathcal{R}} \text{AUC}(r; D_{X_\text{I}}, D^{\alpha}_{X_\text{O}}),$$

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

The importance of Condition 2 is reflected in Theorem 3, showing that Condition 2 is a *necessary* condition for the learnability of OOD detection under AUC if the $\mathscr{D}_{XY}$ is a simple distribution space.

Theorem 3 *Given a ranking function space $\mathcal{R}$ and a domain space $\mathscr{D}_{XY}$, if OOD detection is learnable under AUC for $\mathcal{R}$ in $\mathscr{D}_{XY}$, then for any $D_{XY}, D'_{XY} \in \mathscr{D}_{XY}$, the linear condition under AUC (i.e., Condition 2) holds.*

Since the Condition 2 is a necessary condition for the learnability of OOD detection under AUC, this condition provides a new way to check if an OOD detection is learnable under AUC. Namely, if Condition 2 does not hold, OOD detection is not learnable under AUC.

4.2 Impossibility Theorems under Risk

In this subsection, we first study whether Condition 1 holds in the total space $\mathscr{D}_{XY}^{\text{all}}$. If Condition 1 does not hold, then OOD detection is not learnable under risk. Theorem 4 shows that Condition 1 is not always satisfied, especially, when there is an overlap between the ID and OOD distributions:

Condition 5 (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 $\tilde{\mu}(A_{\text{overlap}}) > 0$, where $A_{\text{overlap}} = \{\mathbf{x} \in \mathcal{X} : f_{\text{I}}(\mathbf{x}) > 0 \text{ and } f_{\text{O}}(\mathbf{x}) > 0\}$. Here f_{I} and f_{O} are the representers of $D_{X_{\text{I}}}$ and $D_{X_{\text{O}}}$ in Radon–Nikodym Theorem (Cohn, 2013),*

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

Lemma 4 *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 $\inf_{h\in\mathcal{H}} R_D^{\text{in}}(h) = 0$, $\inf_{h\in\mathcal{H}} R_D^{\text{out}}(h) = 0$, then Condition 1 does not hold. Therefore, OOD detection is not learnable under risk in $\mathscr{D}_{XY}$ for $\mathcal{H}$.*

Lemma 4 clearly shows that under proper conditions, Condition 1 does not hold, if there exists a domain whose ID and OOD distributions have overlap. By Lemma 4, 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 5 (Impossibility Theorem for Total Space under Risk) *OOD detection is not learnable under risk 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 6 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 6, we need a mild assumption:

Assumption 1 (Separate Space for OOD under Risk) *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 3 in Appendix N), score-based hypothesis space (Proposition 4 in Appendix N) and universal kernel space. Next, we use *Vapnik–Chervonenkis* (VC) dimension (Mohri et al., 2018) 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 6 (Impossibility Theorem for Separate Space under Risk) *If Assumption 1 holds, $\text{VCdim}(\phi \circ \mathcal{H}) < +\infty$ and $\sup_{h\in\mathcal{H}} |\{\mathbf{x} \in \mathcal{X} : h(\mathbf{x}) \in \mathcal{Y}\}| = +\infty$, OOD detection is not learnable under risk in the 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 under risk in the separate space, which reveals the difficulty of the OOD detection.

4.3 Impossibility Theorems under AUC

We then study whether Condition 2 holds in the total space $\mathscr{D}_{XY}^{\text{all}}$. If Condition 2 does not hold, then OOD detection is not learnable under AUC. We first present Lemma 7 to point out when Condition 2 does not hold.

Lemma 7 *Given a ranking function space $\mathcal{R}$, a domain space $\mathscr{D}_{XY}$ and $D_{XY} = \beta D_{X_{\text{I}}Y_{\text{I}}} + (1-\beta)D_{X_{\text{O}}Y_{\text{O}}}$, $D'_{XY} = \beta' D_{X_{\text{I}}Y_{\text{I}}} + (1-\beta')D'_{X_{\text{O}}Y_{\text{O}}} \in \mathscr{D}_{XY}$, let P be the overlap set between $D_{X_{\text{I}}}$ and $D_{X_{\text{O}}}$ and P' be the overlap set between $D_{X_{\text{I}}}$ and $D'_{X_{\text{O}}}$ based on the Definition 5. If*

$$\sup_{r\in\mathcal{R}} \text{AUC}(r; D_{X_{\text{I}}}, D_{X_{\text{O}}}) = \sup_{r\in\mathcal{R}_{\text{all}}} \text{AUC}(r; D_{X_{\text{I}}}, D_{X_{\text{O}}})$$
$$\sup_{r\in\mathcal{R}} \text{AUC}(r; D_{X_{\text{I}}}, D'_{X_{\text{O}}}) = \sup_{r\in\mathcal{R}_{\text{all}}} \text{AUC}(r; D_{X_{\text{I}}}, D'_{X_{\text{O}}}),$$

and $D_{X_{\text{I}}}(P \cap P') < \min\{D_{X_{\text{I}}}(P), D_{X_{\text{I}}}(P')\}$, then Condition 2 does not hold, where $\mathcal{R}_{\text{all}}$ is a ranking function space consisting of all ranking functions from $\mathcal{X}$ to $\mathbb{R}$. Therefore, OOD detection is not learnable under AUC in $\mathscr{D}_{XY}$ for $\mathcal{R}$.

Based on Lemma 7, we know that, under proper conditions, Condition 2 does not hold once there is one domain whose ID and OOD distributions overlap. Then, based on Lemma 7, we can obtain that the OOD detection is not learnable in the total space $\mathscr{D}_{XY}^{\text{all}}$ for any non-trivial ranking function space $\mathcal{R}$.

Theorem 8 (Impossibility Theorem for Total Space under AUC) *Given ranking function space $\mathcal{R}$, if there exist $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ and $r, r' \in \mathcal{R}$ such that*

$$r(\mathbf{x}) > r(\mathbf{x}') \text{ and } r'(\mathbf{x}') > r'(\mathbf{x}),$$

then the learnability of OOD detection under AUC is not distribution-free for $\mathcal{R}$.

From Lemma 7, we know that the overlap between $D_{X_{\text{I}}}$ and $D_{X_{\text{O}}}$ is an important factor to influence the learnability of OOD detection under AUC. Thus, similar to the situation under risk, we want to study the learnability of OOD detection under AUC in separate space $\mathscr{D}_{XY}^{s}$ first. Before introducing the impossibility theorem for separate space, we need a mild assumption demonstrated below.

Assumption 2 (Separate Space for OOD under AUC) *A ranking function space $\mathcal{R}$ is called separate ranking function space, if for any $\mathbf{x} \in \mathcal{X}$, there exists $r_{\mathbf{x}} \in \mathcal{R}$ such that $r_{\mathbf{x}}(\mathbf{x}) < r_{\mathbf{x}}(\mathbf{x}')$, for any $\mathbf{x}' \in \mathcal{X} - \{\mathbf{x}\}$.*

Note that, the above assumption is weak and can be satisfied by some well-known spaces (see Propositions 1 and 2). The above assumption means that, for any data point $\mathbf{x}$, its ranking can be the lowest one compared to other data points in the space $\mathcal{X}$. Finally, we use *Vapnik–Chervonenkis* (VC) dimension (Mohri et al., 2018) to help measure the size of ranking function space, and study the learnability of OOD detection under AUC in $\mathscr{D}_{XY}^{s}$ with the help the VC dimension.

Theorem 9 (Impossibility Theorem for Separate Space under AUC) *Given a separate ranking function space $\mathcal{R}$, if* $\mathrm{VC}[\phi\circ\mathcal{R}] = d < +\infty$ *and* $|\mathcal{X}| \geq (28d{+}14)\log(14d{+}7)$*, then OOD detection is not learnable under AUC in $\mathscr{D}^s_{XY}$ for $\mathcal{R}$, where $\phi\circ\mathcal{R} = \{\mathbf{1}_{r(\mathbf{x})>r(\mathbf{x}')} : r \in \mathcal{R}\}$.*

Based on Theorem 9, we obtain a similar result to the learnability of OOD detection under risk: the finite VC dimension cannot guarantee the learnability of OOD detection under AUC in the separate space, which further reveals the difficulty of OOD detection. Although the above impossibility theorems (under risk and AUC) are frustrating, there is still room to discuss the conditions in Theorem 6 and Theorem 9, and to find out the proper conditions for ensuring the learnability of OOD detection under risk and AUC in the separate space (see the following section).

5. When OOD Detection Can Be Successful

Here, we discuss when the OOD detection can be learnable under risk/AUC in different spaces. We first study the separate space $\mathscr{D}^s_{XY}$.

5.1 OOD Detection in the Separate Space

Both Theorem 6 and Theorem 9 have indicated that $\mathrm{VCdim}(\phi \circ \mathcal{H}) = +\infty$ or $\sup_{h\in\mathcal{H}} |\{\mathbf{x} \in \mathcal{X} : h(\mathbf{x}) \in \mathcal{Y}\}| < +\infty$ (or $\sup_{r\in\mathcal{R}} |\{\mathbf{x} \in \mathcal{X} : r(\mathbf{x}) \in \mathcal{R}\}| < +\infty$ under AUC metric) is necessary to ensure the learnability of OOD detection under risk or AUC in $\mathscr{D}^s_{XY}$ if Assumption 1 or Assumption 2 holds. However, generally, hypothesis spaces generated by feed-forward neural networks with proper activation functions have finite VC dimension (Bartlett et al., 2019; Karpinski and Macintyre, 1997). Therefore, we study the learnability of OOD detection in the case that $|\mathcal{X}| < +\infty$, which implies that $\sup_{h\in\mathcal{H}} |\{\mathbf{x} \in \mathcal{X} : h(\mathbf{x}) \in \mathcal{Y}\}| < +\infty$ under risk metric or $\sup_{r\in\mathcal{R}} |\{\mathbf{x} \in \mathcal{R} : r(\mathbf{x}) \in \mathcal{X}\}| < +\infty$ under AUC metric. Additionally, Theorem 16 also implies that $|\mathcal{X}| < +\infty$ is the necessary and sufficient condition for the learnability of OOD detection under risk in a separate space, when the hypothesis space is generated by FCNN. Hence, $|\mathcal{X}| < +\infty$ may be necessary in the space $\mathscr{D}^s_{XY}$.

Learnability under Risk. 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 under risk in $\mathscr{D}^s_{XY}$, when $|\mathcal{X}| < +\infty$.

Theorem 10 *Let $K = 1$ and $|\mathcal{X}| < +\infty$. Suppose that Assumption 1 holds and the constant function $h^{\mathrm{in}} := 1 \in \mathcal{H}$. Then OOD detection is learnable under risk in $\mathscr{D}^s_{XY}$ for $\mathcal{H}$* ***if and only if*** *$\mathcal{H}_{\mathrm{all}} - \{h^{\mathrm{out}}\} \subset \mathcal{H}$, where $\mathcal{H}_{\mathrm{all}}$ is the hypothesis space consisting of all hypothesis functions, and h^{out} is a constant function that $h^{\mathrm{out}} := 2$, here 1 represents ID data and 2 represents OOD data.*

The condition $h^{\mathrm{in}} \in \mathcal{H}$ presented in Theorem 10 is mild. Many practical hypothesis spaces satisfy this condition, *e.g.*, the FCNN-based hypothesis space (Proposition 3 in Appendix N), score-based hypothesis space (Proposition 4 in Appendix N) and universal kernel-based hypothesis space. Theorem 10 implies that if $K = 1$ and OOD detection is learnable under risk in $\mathscr{D}^s_{XY}$ for $\mathcal{H}$, then the hypothesis space $\mathcal{H}$ should contain almost all hypothesis

functions, implying that if the OOD detection can be learnable under risk in the distribution-agnostic case, then a large-capacity model is necessary.

Next, we extend Theorem 10 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 $\forall \mathbf{x} \in \mathcal{X}$,

$$h(\mathbf{x}) = i, \quad \text{if } h^{\text{in}}(\mathbf{x}) = i \ \text{ and } \ h^{\text{b}}(\mathbf{x}) = 1; \text{ otherwise, } h(\mathbf{x}) = K + 1. \tag{6}$$

We use $\mathcal{H}^{\text{in}} \bullet \mathcal{H}^{\text{b}}$ to represent a hypothesis space consisting of all h defined in Eq. (6). In addition, we also need an additional condition for the loss function ℓ, shown as follows:

Condition 3 *$\ell(y_2, y_1) \leq \ell(K + 1, y_1)$, for any in-distribution labels y_1 and $y_2 \in \mathcal{Y}$.*

Theorem 11 *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 3 holds, then OOD detection is learnable under risk in $\mathscr{D}^s_{XY}$ for $\mathcal{H}$, where $\mathcal{H}_{\text{all}}$ and h^{out} are defined in Theorem 10.*

Learnability under AUC. Then, we study the learnability of OOD detection under AUC in the separate space. Here we require to introduce a basic assumption in learning theory for AUC—*AUC-based Realizability Assumption*, *i.e.*, for any $D_{XY} \in \mathscr{D}_{XY}$, there exists $r^* \in \mathcal{R}$ such that $\text{AUC}(r^*; D_{X_{\text{I}}}, D_{X_{\text{O}}}) = 1$ (see Appendix A.2). Based on this AUC-based Realizability Assumption, we prove the following theorem.

Theorem 12 *Given a separate ranking function space $\mathcal{R}$, if $|\mathcal{X}| < +\infty$, then OOD detection is learnable under AUC in the separate space $\mathscr{D}^s_{XY}$ for $\mathcal{R}$ **if and only if** AUC-based Realizability Assumption holds.*

Theorem 12 indicates the significance of AUC-based Realizability Assumption in OOD detection under AUC, which also means that a large ranking function space is essential for the success of OOD detection under AUC.

5.2 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 under risk in the finite-ID-distribution space $\mathscr{D}^F_{XY}$. We first show two necessary concepts below.

Condition 6 (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 $\sim$ 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 regarding $\mathscr{D}_{XY}$.

Condition 4 (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 $D_{XY} \in [D'_{XY}]$,*

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

In Appendix C, Lemma 2 has implied that Condition 4 is a general version of Condition 1. Next, Theorem 13 shows that Condition 4 is the *necessary and sufficient condition* in $\mathscr{D}_{XY}^F$.

Theorem 13 *Suppose that $\mathcal{X}$ is bounded. OOD detection is learnable under risk in $\mathscr{D}_{XY}^F$ for $\mathcal{H}$* ***if and only if*** *the compatibility condition (i.e., Condition 4) holds. Furthermore, the learning rate $\epsilon_{\text{cons}}(n)$ can attain $O(1/\sqrt{n^{1-\theta}})$, for any $\theta \in (0, 1)$.*

Theorem 13 shows that, in the process of algorithm design, OOD detection cannot be successful without the compatibility condition if we use risk to evaluate the performance. Theorem 13 also implies that Condition 4 is essential for the learnability of OOD detection under risk. 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.

As for the learnability of OOD detection under AUC in the finite-ID-distribution space, since Condition 2 only considers linearity between OOD distributions instead of OOD and ID distributions as shown in Condition 1. To further reveal the learnability of OOD detection under AUC in the finite-ID-distribution space, we might need to discover a new condition for compatibility w.r.t. OOD and ID distributions to extend Condition 2.

5.3 OOD Detection in the Density-based Space

Learnability under Risk. To ensure that Condition 4 holds, we consider a basic assumption in learning theory—*Risk-based Realizability Assumption* (see Appendix A.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 $\mathscr{D}_{XY}^{\mu,b}$, Risk-based Realizability Assumption can conclude the compatibility condition (Condition 4). Based on this observation, we prove the following theorem:

Theorem 14 *Given a density-based space $\mathscr{D}_{XY}^{\mu,b}$, if $\mu(\mathcal{X}) < +\infty$, the Risk-based Realizability Assumption holds, then when $\mathcal{H}$ has finite Natarajan dimension (Shalev-Shwartz and Ben-David, 2014), OOD detection is learnable in $\mathscr{D}_{XY}^{\mu,b}$ 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 Risk-based Realizability Assumption, Theorem 18 has indicated that in some practical scenarios, Risk-based Realizability Assumption is the necessary and sufficient condition for the learnability of OOD detection under risk in the density-based space. Therefore, Risk-based Realizability Assumption may be indispensable for the learnability of OOD detection under risk in some practical scenarios.

Learnability under AUC. To study the learnability of OOD detection under AUC in the density-based space, we first need to introduce a constant-closure assumption for $\mathcal{R}$.

Assumption 3 *We say a ranking function space $\mathcal{R}$ is constant closure, if for any $r \in \mathcal{R}$, the constant function space $\overline{r(\mathcal{X})} := \{c : r(\mathbf{x}) = c,\ \forall \mathbf{x} \in \mathcal{X}, r \in \mathcal{R}\} \subset \mathcal{R}$.*

Note that, the above assumption is weak and can be satisfied by some well-known ranking function space (see Propositions 1 and 2). Based on this assumption, we give a sufficient condition for learnability of OOD detection under AUC in the density-based space:

Theorem 15 *Suppose that $\mathcal{R}$ is constant closure, separate, and $\mu(\mathcal{X}) < +\infty$. Given a density-based space $\mathscr{D}_{XY}^{\mu,b}$, if the AUC-based Realizability Assumption holds, then when $\mathrm{VC}[\phi \circ \mathcal{R}] < +\infty$, OOD detection is learnable under AUC in $\mathscr{D}_{XY}^{\mu,b}$ for $\mathcal{R}$, where $\phi \circ \mathcal{R} = \{\mathbf{1}_{r_1(\mathbf{x})>r_2(\mathbf{x}')} : r_1, r_2 \in \mathcal{R}\}$. Furthermore, the learning rate $\epsilon_{\mathrm{cons}}(n)$ can attain $O(1/\sqrt{n^{1-\theta}})$, for any $\theta \in (0,1)$.*

Based on Theorem 15, we find that the AUC-based Realizability Assumption is also important for the learnability of OOD detection under AUC in the density-based space.

6. Connecting Theory to Practice

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

6.1 Key Concepts Regarding Fully-connected Neural Networks

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[3], 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:

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

Lemma 14 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:

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

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

3. 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.*

FCNN-based Ranking Function Space. Then, based on the definition of FCNN, we show that, given a specific $\mathbf{q}$, under some mild conditions, $\mathcal{F}_{\mathbf{q}}^{\sigma}$ is the separate and constant closure ranking function space.

Proposition 1 *Let $\mathcal{X}$ be a bounded feature space. Given $\mathbf{q} = (l_1, ..., l_{g-1}, 1)$, then*

- *if some s with $1 < s < g$, $d = l_1 \leq l_2 \leq ... \leq l_s$, and $l_s \geq 2d$, $\mathcal{F}_{\mathbf{q}}^{\sigma}$ is the separate ranking function space;*
- *$\mathcal{F}_{\mathbf{q}}^{\sigma}$ is constant closure;*
- *$\{\mathbf{1}_{f_1(\mathbf{x}) < f_2(\mathbf{x}')} : f_1, f_2 \in \mathcal{F}_{\mathbf{q}}^{\sigma}\}$ has finite VC dimension.*

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 score 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 score functions E as follows: for any $\mathbf{f} = [f^1, ..., f^l]^\top \in \mathcal{F}_l$,

- softmax-based function (Hendrycks and Gimpel, 2017): $\lambda \in (\frac{1}{l}, 1)$ and $T > 0$,

$$E(\mathbf{f}) = \max_{k \in \{1,...,l\}} \frac{\exp(f^k)}{\sum_{c=1}^{l} \exp(f^c)}; \tag{7}$$

- temperature-scaled function (Liang et al., 2018): $\lambda \in (\frac{1}{l}, 1)$ and $T > 0$,

$$E(\mathbf{f}) = \max_{k \in \{1,...,l\}} \frac{\exp(f^k/T)}{\sum_{c=1}^{l} \exp(f^c/T)}; \tag{8}$$

- energy-based function (Liu et al., 2020): $\lambda \in (0, +\infty)$ and $T > 0$,

$$E(\mathbf{f}) = T \log \sum_{c=1}^{l} \exp(f^c/T). \tag{9}$$

Using E, λ and $\mathbf{f} \in \mathcal{F}_{\mathbf{q}}^{\sigma}$, we have a classifier: $h_{\mathbf{f},E}^{\lambda}(\mathbf{x}) = 1$, if $E(\mathbf{f}(\mathbf{x})) \geq \lambda$; otherwise, $h_{\mathbf{f},E}^{\lambda}(\mathbf{x}) = 2$, 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 $h_{\mathbf{f},E}^{\lambda}$, is generated. We define $\mathcal{H}_{\mathbf{q},E}^{\sigma,\lambda} := \{h_{\mathbf{f},E}^{\lambda} : \forall \mathbf{f} \in \mathcal{F}_{\mathbf{q}}^{\sigma}\}$.

Score-based Ranking Function Space. Similar to the FCNN-based ranking function space, for several representative score functions E (*e.g.,*, Eqs (7), (8), and (9)), the FCNN-based score ranking function space $E \circ \mathcal{F}_{\mathbf{q}}^{\sigma}$ is separate, which is evidence that Assumption 2 is weak and can be easily satisfied.

Proposition 2 *Given $\mathbf{q} = (l_1, ..., l_{g-1}, l)$ and $\mathbf{q}' = (l_1, ..., l_{g-1}, 1)$, let $\mathcal{R} = E \circ \mathcal{F}_{\mathbf{q}}^{\sigma}$, then*

- *if $\mathcal{F}_{\mathbf{q}'}^{\sigma}$ is a separate ranking function space, $\mathcal{R}$ is the separate ranking function space;*

- $\mathcal{R}$ *is constant closure;*
- $\{\mathbf{1}_{r_1(\mathbf{x})<r_2(\mathbf{x}')}, r_1, r_2 \in \mathcal{R}\}$ *has finite VC dimension,*

where E *is Eq.* (7), (8) *or* (9).

According to the previous section, we find that FCNN-based ranking function space and FCNN-based score ranking function space can satisfy almost all conditions in theorems.

6.2 Learnability of OOD Detection in Different Spaces

Next, we present applications of our theory regarding the above practical and important hypothesis spaces and ranking function spaces.

Theorem 16 *Suppose that Condition 3 holds and the hypothesis space* $\mathcal{H}$ *is FCNN-based or score-based, i.e.,* $\mathcal{H} = \mathcal{H}^{\sigma}_{\mathbf{q}}$ *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.* (6)*, here* E *is Eq.* (7), (8) *or* (9)*. Then*

There is a sequence $\mathbf{q} = (l_1, ..., l_g)$ *such that OOD detection is learnable under risk 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 under risk in* $\mathscr{D}^{s}_{XY}$ *for* $\mathcal{H}$*.*

If we consider the ranking function space, we can obtain a similar theoretical result.

Theorem 17 *Suppose the ranking function space* $\mathcal{R}$ *is separate, and FCNN-based or score-based, i.e.,* $\mathcal{R} = \mathcal{F}^{\sigma}_{\mathbf{q}}$ *or* $\mathcal{R} = E \circ \mathcal{F}^{\sigma}_{\mathbf{q}}$*, where* E *is Eq.* (7), (8) *or* (9)*. Then*

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

Furthermore, if $|\mathcal{X}| < +\infty$*, then there is 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 under AUC in* $\mathscr{D}^{s}_{XY}$ *for* $\mathcal{R}$*.*

Theorems 16 and 17 state that 1) when the hypothesis space or ranking function space is FCNN-based or score-based, the finite feature space is the necessary and sufficient condition for the learnability of OOD detection (under risk or AUC) in the separate space; and 2) a larger architecture of FCNN has a greater probability to achieve the learnability of OOD detection in the separate space. Note that when we select Eqs. (7), (8), or (9) as the score function E, Theorems 16 and 17 also show that the selected score functions E can guarantee the learnability of OOD detection (under risk or AUC), which is a theoretical support for the representative works (Liang et al., 2018; Liu et al., 2020; Hendrycks and Gimpel, 2017). Furthermore, Theorems 18 and 19 also offer theoretical supports for these works in the density-based space.

Theorem 18 *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 Eq.* (7), (8)*, or* (9)) *or FCNN-based* ($\mathcal{H} = \mathcal{H}^{\sigma}_{\mathbf{q}}$)*. If*

$\mu(\mathcal{X}) < +\infty$, *then the following four conditions are* ***equivalent****:*

Learnability in $\mathscr{D}_{XY}^{\mu,b}$ *for* $\mathcal{H} \iff$ *Condition 1* $\iff$ *Risk-based Realizability Assumption* $\iff$ *Condition 4*

Theorem 19 *Suppose that the ranking function space* $\mathcal{R}$ *is separate and score-based* ($\mathcal{R} = E \circ \mathcal{F}_{\mathbf{q}}^{\sigma}$) *or FCNN-based* ($\mathcal{R} = \mathcal{F}_{\mathbf{q}}^{\sigma}$)*, where* E *is Eq.* (7), (8) *or* (9)*. If* $\mu(\mathcal{X}) < +\infty$*, then the following three conditions satisfy:*

AUC-based Realizability Assumption $\Rightarrow$ *Learnability in* $\mathscr{D}_{XY}^{\mu,b}$ *for* $\mathcal{R} \Rightarrow$ *Condition 2*

Compared to Theorem 18, Theorem 19 cannot obtain the equivalence among Realizability Assumption, Learnability in $\mathscr{D}_{XY}^{\mu,b}$ and linear condition under AUC. The main reason is that Condition 2 is a weaker necessary condition for learnability of OOD detection under AUC than Condition 1 for learnability of OOD detection under risk. We need to discover a strong necessary condition for learnability of OOD detection under AUC to obtain a similar equivalence that appeared in Theorem 18[4].

6.3 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 20 *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 Eq.* (7), (8)*, or* (9)) *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 5), then OOD detection is not learnable under risk in* $\mathscr{D}_{XY}$ *for* $\mathcal{H}$*.*

When $K = 1$ and the hypothesis space is FCNN-based or score-based, Theorem 20 shows that overlap between ID and OOD distributions is the sufficient condition for the unlearnability of OOD detection under risk. Theorem 20 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. As for the ranking function space, we can obtain a corresponding but weaker theoretical result shown below.

Theorem 21 *Let the separate ranking function space* $\mathcal{R}$ *be FCNN-based or score-based (where the score function* E *is Eq.* (7), (8)*, or* (9))*. Suppose that* $D_{XY}, D'_{XY} \in \mathscr{D}_{XY}$ *are discrete distributions with* $D_{X_{\text{I}}Y_{\text{I}}} = D_{X_{\text{I}}Y_{\text{I}}}$ *and* $D_{X_{\text{O}}} = \delta_{\mathbf{x}}, D'_{X_{\text{O}}} = \delta_{\mathbf{x}'}$*. If* $D_{X_{\text{O}}} = \delta_{\mathbf{x}}, D'_{X_{\text{O}}} = \delta_{\mathbf{x}'}$ *have overlaps with* $D_{X_{\text{I}}Y_{\text{I}}}$ *and* $D_{X_{\text{O}}} \neq D'_{X_{\text{O}}}$*, then OOD detection is not learnable under AUC in* $\mathscr{D}_{XY}$ *for* $\mathcal{R}$*.*

4. A stronger necessary condition normally means that this necessary condition is closer to the necessary and sufficient condition.

7. Discussion

Connections between Theorems under Risk and AUC. In our previous conference paper (Fang et al., 2022a), we mainly focus on the learnability of OOD detection under risk. However, in practice, AUC-related metrics are often used to evaluate the performance of OOD detection algorithms (Lin et al., 2021; Huang et al., 2021; Fort et al., 2021b; Ming et al., 2022; Yang et al., 2022). To fill this gap, we take a further step: investigating the learnability of OOD detection under AUC. Figure 1 illustrates the connections between the main theoretical results of learnability of OOD detection under risk and AUC. In the most of domain spaces considered in this paper, we can obtain similar theoretical results under AUC compared to theoretical results under risk. However, when considering the finite-ID-distribution space, we cannot get a corresponding theorem for the learnability of OOD detection under AUC. The main reason is that Condition 2 is a weaker necessary condition for learnability of OOD detection under AUC than Condition 1 for learnability of OOD detection under risk. Thus, additional information might be required to obtain a stronger necessary condition for AUC learnability. The influence of Condition 2 also appears in Theorem 19 where we cannot obtain a strong theoretical result like Theorem 18. To conclude, since Condition 1 is stronger (i.e., it is closer to necessary and sufficient condition) than Condition 2, the theoretical results under risk are stronger than those under AUC. In the future, we need to discover a stronger necessary condition for learnability of OOD detection under AUC.

Understanding Far-OOD Detection. Many existing works (Hendrycks and Gimpel, 2017; Yang et al., 2022) study the far-OOD detection issue. Existing benchmarks include 1) MNIST (Deng, 2012) as ID dataset, and Texture (Kylberg, 2011), CIFAR-10 (Krizhevsky and Hinton, 2009) or Place365 (Zhou et al., 2018) as OOD datasets; and 2) CIFAR-10 (Krizhevsky and Hinton, 2009) as ID dataset, and MNIST (Deng, 2012), or Fashion-MNIST (Zhou et al., 2018) 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 $D_{XY} \in \mathscr{D}_{XY}$,

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

Theorems 11, 13 and 16 imply that under appropriate hypothesis space, τ-far-OOD detection is learnable under risk. Theorem 17 implies that under appropriate hypothesis space, τ-far-OOD detection is learnable under AUC. In Theorem 11, the condition $|\mathcal{X}| < +\infty$ is necessary for the separate space. However, one can prove that in the far-OOD case, when $\mathcal{H}^{\mathrm{in}}$ is agnostic PAC learnable for ID distribution, the results in Theorem 11 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 $\mathcal{H}^{\mathrm{in}}$ is agnostic PAC learnable for ID distribution and $\mathcal{X}$ is compact, the KNN-based OOD detection algorithm (Sun et al., 2022) 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. (Ren et al., 2021; Fort et al., 2021a) consider this issue and name it near-OOD detection. Existing benchmarks

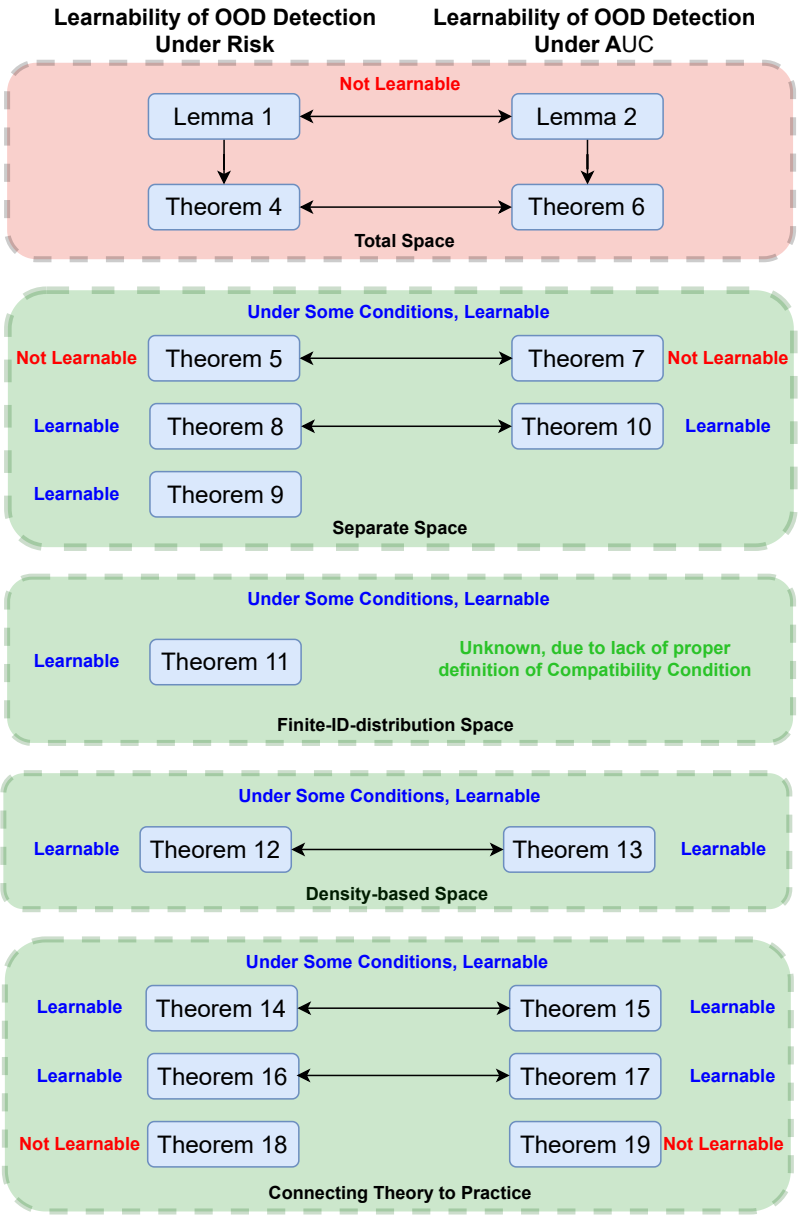

Figure 1: Connections among main theoretical results in this paper. Compared to risk, AUC has a more strict requirement for the classification. Perfect classification (i.e., accuracy is 100%) does not imply perfect AUC, which is the reason why theories regarding risk and AUC are very different.

include 1) MNIST (Deng, 2012) as ID dataset, and Fashion-MNIST (Zhou et al., 2018) or Not-MNIST (Bulatov, 2011) as OOD datasets; and 2) CIFAR-10 (Krizhevsky and Hinton, 2009) as ID dataset, and CIFAR-100 (Krizhevsky et al., 2009) as OOD dataset. From the theoretical view, some near-OOD tasks may imply the overlap condition, *i.e.* Definition 5. Therefore, Lemma 4 and Theorem 20 imply that near-OOD detection may be not

learnable under risk, and Lemma 7 and Theorem 21 imply that near-OOD detection may be not learnable under AUC. Developing a theory to understand the feasibility of near-OOD detection is an *open question*.

8. 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 (Hendrycks and Gimpel, 2017)[5]. Representative works consider using the maximum softmax score (Hendrycks and Gimpel, 2017), temperature-scaled score (Ren et al., 2019) and energy-based score (Liu et al., 2020; Wang et al., 2021) to identify OOD data. 2) Density-based methods aim to estimate an ID distribution and identify the low-density area as OOD data (Zong et al., 2018). 3) The recent development of generative models provides promising ways to make them successful in OOD detection (Pidhorskyi et al., 2018; Nalisnick et al., 2019; Ren et al., 2019; Kingma and Dhariwal, 2018; Xiao et al., 2020). Distance-based methods are based on the assumption that OOD data should be relatively far away from the centroids of ID classes (Lee et al., 2018), including Mahalanobis distance (Lee et al., 2018; Ren et al., 2021), cosine similarity (Zaeemzadeh et al., 2021), and kernel similarity (Amersfoort et al., 2020).

Early works consider using the maximum softmax score to express the ID-ness (Hendrycks and Gimpel, 2017). Then, temperature scaling functions are used to amplify the separation between the ID and OOD data (Ren et al., 2019). Recently, researchers propose hyperparameter-free energy scores to improve the OOD uncertainty estimation (Liu et al., 2020; Wang et al., 2021). Additionally, researchers also consider using the gradients to help improve the performance of OOD detection (Huang et al., 2021).

Except for the above algorithms, researchers also study the situation, where auxiliary OOD data can be obtained during the training process (Hendrycks et al., 2019; Dhamija et al., 2018). 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 (Yang et al., 2021). Thus, researchers also consider generating OOD data to help the separation of OOD and ID data (Vernekar et al., 2019). 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. Zhang et al. (2021) rejects the typical set hypothesis, the claim that relevant OOD distributions can lie in high likelihood regions of data distribution, as implausible. Zhang et al. (2021) argues that minimal density estimation errors can lead to OOD detection failures without assuming an overlap between ID and OOD distributions. Compared to Zhang et al. (2021), our theory focuses on the PAC learnable theory of OOD detection. If detectors are generated by FCNN, our theory (Theorem 20) shows that the overlap is the sufficient condition to the failure of learnability of OOD detection, which is complementary to Zhang et al. (2021). In addition, we identify several necessary and

5. Note that, some methods assume that OOD data are available in advance (Hendrycks et al., 2019; Dhamija et al., 2018). However, the exposure of OOD data is a strong assumption (Yang et al., 2021). We do not consider this situation in our paper.

sufficient conditions for the learnability of OOD detection, which opens a door to studying OOD detection in theory. Beyond Zhang et al. (2021), Morteza and Li (2022) paves a new avenue to designing provable OOD detection algorithms. Compared to Morteza and Li (2022), our paper aims to characterize the learnability of OOD detection to answer the question: is OOD detection PAC learnable?

Open-set Learning Theory. Liu et al. (2018) is the first to propose the agnostic PAC guarantees for open-set detection. Unfortunately, the test data must be used during the training process. Fang et al. (2020) considers the open-set domain adaptation (OSDA) (Luo et al., 2020) and proposes the first learning bound for OSDA. Fang et al. (2020) mainly depends on the positive-unlabeled learning techniques (Kiryo et al., 2017; Ishida et al., 2018; Chen et al., 2021b). However, similar to Liu et al. (2018), the test data must be available during training. To study open-set learning (OSL) *without accessing the test data* during training, Fang et al. (2021) proposes and studies the almost PAC learnability for OSL, which is motivated by transfer learning (Dong et al., 2020; Fang et al., 2022b). Recently, Wang et al. (2022) proposes a novel AUC-based OOD detection objective named OpenAUC (Yang et al., 2023; Jiang et al., 2023) as objective function to learn open-set predictors, and builds a corresponding AUC-based open-set learning theory. In our paper, we study the PAC learnability for OOD detection, which is an open problem proposed by Fang et al. (2021).

Learning Theory for Classification with Reject Option. Many works (Chow, 1970; Franc et al., 2021) also investigate the *classification with reject option* (CwRO) problem, which is similar to OOD detection in some cases. Cortes et al. (2016b,a); Ni et al. (2019); Charoenphakdee et al. (2021); Bartlett and Wegkamp (2008) 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.

Robust Statistics. In the field of robust statistics (Rousseeuw et al., 2011), researchers aim to propose estimators and testers that can mitigate the negative effects of outliers (similar to OOD data). The proposed estimators are supposed to be independent of the potentially high dimensionality of the data (Ronchetti and Huber, 2009; Diakonikolas et al., 2020, 2019). Existing works (Diakonikolas et al., 2021; Cheng et al., 2021; Diakonikolas et al., 2022) in the field have identified and resolved the statistical limits of outlier robust statistics by constructing estimators and proving impossibility results. In the future, it is a promising and interesting research direction to study the robustness of OOD detection based on robust statistics.

PQ Learning Theory. Under some conditions, PQ learning theory (Goldwasser et al., 2020; Kalai and Kanade, 2021) 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 Goldwasser et al. (2020); Kalai and Kanade (2021) aims to give the PAC estimation under Realizability Assumption (Shalev-Shwartz and Ben-David, 2014). Our theory focuses on the PAC theory in different cases, which is more difficult and more practical than PAC theory under Realizability Assumption.

9. Conclusions and Future Works

OOD detection has become an important technique to increase the reliability of machine learning. However, its theoretical foundation is merely investigated, which hinders real-world applications of OOD detection algorithms. This paper is the *first* to provide the PAC theory for OOD detection in terms of two commonly used metrics: risk and AUC. Our results imply that a universally consistent algorithm might not exist for all scenarios in OOD detection. Yet, we still discover some scenarios where OOD detection is learnable under risk or AUC metrics. Our theory reveals many necessary and sufficient conditions for the learnability of OOD detection under risk or AUC, hence *paving a road* to studying the learnability of OOD detection. In the future, we will focus on studying the robustness of OOD detection based on robust statistics (Diakonikolas et al., 2021; Diakonikolas and Kane, 2020).

Acknowledgments

JL and ZF were supported by the Australian Research Council (ARC) under FL190100149. YL is supported by National Science Foundation (NSF) Award No. IIS-2237037. FL was supported by the ARC with grant numbers DP230101540 and DE240101089, and the NSF&CSIRO Responsible AI program with grant number 2303037. ZF would also like to thank Prof. Peter Bartlett, Dr. Tongliang Liu and Dr. Zhiyong Yang for productive discussions.

References

J. V. Amersfoort, L. Smith, Y. W. Teh, and Y. Gal. Uncertainty estimation using a single deep deterministic neural network. In *ICML*, 2020.

P. L. Bartlett and W. Maass. Vapnik-chervonenkis dimension of neural nets. *The handbook of brain theory and neural networks*, 2003.

P. L. Bartlett and M. H. Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 2008.

P. L. Bartlett, N. Harvey, C. Liaw, and A. Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019.

A. Bendale and T. E. Boult. Towards open set deep networks. In *CVPR*, 2016.

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

N. Charoenphakdee, Z. Cui, Y. Zhang, and M. Sugiyama. Classification with rejection based on cost-sensitive classification. In *ICML*, 2021.

J. Chen, Y. Li, X. Wu, Y. Liang, and S. Jha. Atom: Robustifying out-of-distribution detection using outlier mining. *ECML*, 2021a.

S. Chen, G. Niu, C. Gong, J. Li, J. Yang, and M. Sugiyama. Large-margin contrastive learning with distance polarization regularizer. In *ICML*, 2021b.

Y. Cheng, I. Diakonikolas, D. M. Kane, R. Ge, S. Gupta, and M. Soltanolkotabi. Outlier-robust sparse estimation via non-convex optimization. In *NeurIPS*, 2021.

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

D. L. Cohn. *Measure theory*. Springer, 2013.

C. Cortes, G. DeSalvo, and M. Mohri. Boosting with abstention. In *NeurIPS*, 2016a.

C. Cortes, G. DeSalvo, and M. Mohri. Learning with rejection. In *ALT*, 2016b.

L. Deecke, R. A. Vandermeulen, L. Ruff, S. Mandt, and M. Kloft. Image anomaly detection with generative adversarial networks. In *ECML*, 2018.

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

A. R. Dhamija, M. Günther, and T. E. Boult. Reducing network agnostophobia. In *NeurIPS*, pages 9175–9186, 2018.

I. Diakonikolas and D. 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.

I. Diakonikolas, D. Kane, S. Karmalkar, E. Price, and A. Stewart. Outlier-robust high-dimensional sparse estimation via iterative filtering. In *NeurIPS*, 2019.

I. Diakonikolas, D. M. Kane, and A. Pensia. Outlier robust mean estimation with subgaussian rates via stability. In *NeurIPS*, 2020.

I. Diakonikolas, D. M. Kane, A. Stewart, and Y. Sun. Outlier-robust learning of ising models under dobrushin's condition. In *COLT*, 2021.

I. Diakonikolas, D. M. Kane, J. C. Lee, and A. Pensia. Outlier-robust sparse mean estimation for heavy-tailed distributions. In *NeurIPS*, 2022.

J. Dong, Y. Cong, G. Sun, B. Zhong, and X. Xu. What can be transferred: Unsupervised domain adaptation for endoscopic lesions segmentation. In *CVPR*, 2020.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

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

Z. Fang, J. Lu, A. Liu, F. Liu, and G. Zhang. Learning bounds for open-set learning. In *ICML*, 2021.

Z. Fang, Y. Li, J. Lu, J. Dong, B. Han, and F. Liu. Is out-of-distribution detection learnable? In *NeurIPS*, 2022a.

Z. Fang, J. Lu, F. Liu, and G. Zhang. Semi-supervised heterogeneous domain adaptation: Theory and algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022b.

S. Fort, J. Ren, and B. Lakshminarayanan. Exploring the limits of out-of-distribution detection. In *NeurIPS*, 2021a.

S. Fort, J. Ren, and B. Lakshminarayanan. Exploring the Limits of Out-of-Distribution Detection. In *NeurIPS*, 2021b.

V. Franc, D. Průša, and V. Voracek. Optimal strategies for reject option classifiers. *CoRR*, abs/2101.12523, 2021.

S. Goldwasser, A. T. Kalai, Y. Kalai, and O. Montasser. Beyond perturbations: Learning guarantees with arbitrary adversarial test examples. In *NeurIPS*, 2020.

S. Goyal, A. Raghunathan, M. Jain, H. V. Simhadri, and P. Jain. DROCC: deep robust one-class classification. In *ICML*, 2020.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. J. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 2012.

D. Hendrycks and K. Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2017.

D. Hendrycks, M. Mazeika, and T. G. Dietterich. Deep anomaly detection with outlier exposure. In *ICLR*, 2019.

Y. Hsu, Y. Shen, H. Jin, and Z. Kira. Generalized ODIN: detecting out-of-distribution image without learning from out-of-distribution data. In *CVPR*, 2020.

G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *CVPR*, 2017.

R. Huang, A. Geng, and Y. Li. On the Importance of Gradients for Detecting Distributional Shifts in the Wild. In *NeurIPS*, 2021.

T. Ishida, G. Niu, and M. Sugiyama. Binary classification from positive-confidence data. In *NeurIPS*, 2018.

Y. Jiang, Q. Xu, Y. Zhao, Z. Yang, P. Wen, X. Cao, and Q. Huang. Positive-unlabeled learning with label distribution alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

A. T. Kalai and V. Kanade. Efficient learning with arbitrary covariate shift. In *ALT*, Proceedings of Machine Learning Research, 2021.

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

D. P. Kingma and P. Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *NeurIPS*, 2018.

R. Kiryo, G. Niu, M. C. du Plessis, and M. Sugiyama. Positive-unlabeled learning with non-negative risk estimator. In *NeurIPS*, 2017.

A. Krizhevsky and G. Hinton. Convolutional deep belief networks on cifar-10. *Technical report, Citeseer*, 2009.

A. Krizhevsky, V. Nair, and G. Hinton. Cifar-10 and cifar-100 datasets. 2009.

G. Kylberg. *Kylberg texture dataset v. 1.0.* 2011.

K. Lee, K. Lee, H. Lee, and J. Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *NeurIPS*, 2018.

S. Liang, Y. Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*, 2018.

Z. Lin, S. D. Roy, and Y. Li. Mood: Multi-level out-of-distribution detection. In *CVPR*, 2021.

S. Liu, R. Garrepalli, T. G. Dietterich, A. Fern, and D. Hendrycks. Open category detection with PAC guarantees. In *ICML*, 2018.

W. Liu, X. Wang, J. D. Owens, and Y. Li. Energy-based out-of-distribution detection. In *NeurIPS*, 2020.

Y. Luo, Z. Wang, Z. Huang, and M. Baktashmotlagh. Progressive graph learning for open-set domain adaptation. In *ICML*, 2020.

Y. Ming, H. Yin, and Y. Li. On the impact of spurious correlation for out-of-distribution detection. *AAAI*, 2022.

M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of machine learning*. MIT press, 2018.

P. Morteza and Y. Li. Provable guarantees for understanding out-of-distribution detection. *AAAI*, 2022.

E. T. Nalisnick, A. Matsukawa, Y. W. Teh, D. Görür, and B. Lakshminarayanan. Do deep generative models know what they don't know? In *ICLR*, 2019.

C. Ni, N. Charoenphakdee, J. Honda, and M. Sugiyama. On the calibration of multiclass classification with rejection. In *NeurIPS*, 2019.

S. Pidhorskyi, R. Almohsen, and G. Doretto. Generative probabilistic novelty detection with adversarial autoencoders. In *NeurIPS*, 2018.

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

J. Ren, P. J. Liu, E. Fertig, J. Snoek, R. Poplin, M. A. DePristo, J. V. Dillon, and B. Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In *NeurIPS*, 2019.

J. Ren, S. Fort, J. Liu, A. G. Roy, S. Padhy, and B. Lakshminarayanan. A simple fix to mahalanobis distance for improving near-ood detection. *CoRR*, abs/2106.09022, 2021.

E. M. Ronchetti and P. J. Huber. *Robust statistics*. John Wiley & Sons, 2009.

P. J. Rousseeuw, F. R. Hampel, E. M. Ronchetti, and W. A. Stahel. *Robust statistics: the approach based on influence functions*. John Wiley & Sons, 2011.

L. Ruff, N. Görnitz, L. Deecke, S. A. Siddiqui, R. A. Vandermeulen, A. Binder, E. Müller, and M. Kloft. Deep one-class classification. In *ICML*, 2018.

I. Safran and O. Shamir. Depth-width tradeoffs in approximating natural functions with neural networks. In *ICML*, 2017.

M. Salehi, H. Mirzaei, D. Hendrycks, Y. Li, M. H. Rohban, and M. Sabokrou. A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges. *arXiv preprint arXiv:2110.14051*, 2021.

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

S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Learnability, stability and uniform convergence. *J. Mach. Learn. Res.*, 11:2635–2670, 2010.

Y. Sun, C. Guo, and Y. Li. React: Out-of-distribution detection with rectified activations. In *NeurIPS*, 2021.

Y. Sun, Y. Ming, X. Zhu, and Y. Li. Out-of-distribution detection with deep nearest neighbors. In *ICML*, 2022.

S. Vernekar, A. Gaurav, V. Abdelzad, T. Denouden, R. Salay, and K. Czarnecki. Out-of-distribution detection in classifiers via generation. In *NeurIPS Workshop*, 2019.

H. Wang, W. Liu, A. Bocchieri, and Y. Li. Can multi-label classification networks know what they don't know? In *NeurIPS*, 2021.

Z. Wang, Q. Xu, Z. Yang, Y. He, X. Cao, and Q. Huang. Openauc: Towards auc-oriented open-set recognition. *Advances in Neural Information Processing Systems*, 35:25033–25045, 2022.

Z. Xiao, Q. Yan, and Y. Amit. Likelihood regret: An out-of-distribution detection score for variational auto-encoder. In *NeurIPS*, 2020.

J. Yang, K. Zhou, Y. Li, and Z. Liu. Generalized out-of-distribution detection: A survey. *CoRR*, abs/2110.11334, 2021.

J. Yang, K. Zhou, and Z. Liu. Full-spectrum out-of-distribution detection. *CoRR*, 2022.

Z. Yang, Q. Xu, W. Hou, S. Bao, Y. He, X. Cao, and Q. Huang. Revisiting auc-oriented adversarial training with loss-agnostic perturbations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–18, 2023.

A. Zaeemzadeh, N. Bisagno, Z. Sambugaro, N. Conci, N. Rahnavard, and M. Shah. Out-of-distribution detection using union of 1-dimensional subspaces. In *CVPR*, 2021.

L. H. Zhang, M. Goldstein, and R. Ranganath. Understanding failures in out-of-distribution detection with deep generative models. In *ICML*, 2021.

B. Zhou, À. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2018.

B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho, and H. Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *ICLR*, 2018.

Table of Contents of Appendix

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
$\mathcal{R}$	ranking function space consisting some ranking functions
• Risks, Partial Risks and AUC	
$R_D(h)$	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}
$\text{AUC}(r; D_{XY})$	AUC corresponding to D_{XY}
$\text{AUC}(r; D_{X_{\text{I}}}, D_{X_{\text{O}}})$	AUC corresponding to $D_{X_{\text{I}}}, D_{X_{\text{O}}}$
• 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

Appendix A. Notations

A.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, $k_2 = \arg\max_{k\in\{1,...,l\}} f^k(\mathbf{x})$.

A.2 Realizability Assumptions

Assumption 4 (Risk-based Realizability Assumption) *A domain space $\mathscr{D}_{XY}$ and hypothesis space $\mathcal{H}$ satisfy the Risk-based 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 $R_D(h^*) = 0$.*

Assumption 5 (AUC-based Realizability Assumption) *A domain space $\mathscr{D}_{XY}$ and ranking function space $\mathcal{R}$ satisfy the AUC-based Realizability Assumption under AUC, if for each domain $D_{XY} \in \mathscr{D}_{XY}$, there exists at least one hypothesis function $r^* \in \mathcal{R}$ such that* $\mathrm{AUC}(h^*; D_{XY}) = 1$.

A.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,

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

which implies that

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

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

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

Because $\epsilon_{\mathrm{cons}}(n)$ is monotonically decreasing, we can find a smallest m such that $\epsilon_{\mathrm{cons}}(m) \geq \epsilon\delta$ and $\epsilon_{\mathrm{cons}}(m-1) < \epsilon\delta$, 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.

Appendix B. Proof of Theorem 1

Theorem 1 *Given 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 $\mathscr{D}_{XY}$ under risk ***if and only if*** *OOD detection is learnable in $\mathscr{D}'_{XY}$ under risk;*
4) OOD detection is learnable in $\mathscr{D}_{XY}$ under AUC ***if and only if*** *OOD detection is learnable in $\mathscr{D}'_{XY}$ under AUC.*

Proof [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:

$\mathscr{D}_{XY}$ 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 $n \to +\infty$

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

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 $\alpha \in [0,1)$,

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

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

$$\mathbb{E}_{S\sim D^n_{X_\mathrm{I}Y_\mathrm{I}}} R_D^\alpha(\mathbf{A}(S)) \leq \inf_{h\in\mathcal{H}} R_D^\alpha(h) + \epsilon_{\mathrm{cons}}(n), \quad \forall \alpha \in [0,1). \tag{10}$$

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

$$\liminf_{\alpha\to 1} \inf_{h\in\mathcal{H}} R_D^\alpha(h) \geq \liminf_{\alpha\to 1} \alpha \inf_{h\in\mathcal{H}} R_D^{\mathrm{out}}(h) = \inf_{h\in\mathcal{H}} R_D^{\mathrm{out}}(h). \tag{11}$$

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

$$R_D^{\mathrm{out}}(h_\epsilon) - \inf_{h\in\mathcal{H}} R_D^{\mathrm{out}}(h) \leq \epsilon.$$

It is obvious that

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

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

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

which implies that

$$\inf_{h\in\mathcal{H}} R_D^{\mathrm{out}}(h) = \lim_{\epsilon\to 0} R_D^{\mathrm{out}}(h_\epsilon) \geq \lim_{\epsilon\to 0} \limsup_{\alpha\to 1} \inf_{h\in\mathcal{H}} R_D^\alpha(h) = \limsup_{\alpha\to 1} \inf_{h\in\mathcal{H}} R_D^\alpha(h). \tag{12}$$

Combining Eq. (11) with Eq. (12), we have

$$\inf_{h\in\mathcal{H}} R_D^{\mathrm{out}}(h) = \limsup_{\alpha\to 1} \inf_{h\in\mathcal{H}} R_D^\alpha(h) = \liminf_{\alpha\to 1} \inf_{h\in\mathcal{H}} R_D^\alpha(h), \tag{13}$$

which implies that

$$\inf_{h\in\mathcal{H}} R_D^{\mathrm{out}}(h) = \lim_{\alpha\to 1} \inf_{h\in\mathcal{H}} R_D^\alpha(h). \tag{14}$$

Note that

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

Hence, Lebesgue's Dominated Convergence Theorem (Cohn, 2013) implies that

$$\lim_{\alpha\to 1} \mathbb{E}_{S\sim D^n_{X_\mathrm{I}Y_\mathrm{I}}} R_D^\alpha(\mathbf{A}(S)) = \mathbb{E}_{S\sim D^n_{X_\mathrm{I}Y_\mathrm{I}}} R_D^{\mathrm{out}}(\mathbf{A}(S)). \tag{15}$$

Using Eq. (10), we have that

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

Combining Eq. (14), Eq. (15) with Eq. (16), we obtain that

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

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

$$\mathbb{E}_{S\sim D^n_{X_\mathrm{I}Y_\mathrm{I}}} R_D^1(\mathbf{A}(S)) \leq \inf_{h\in\mathcal{H}} R_D^1(h) + \epsilon_{\mathrm{cons}}(n). \tag{17}$$

Combining Eq. (10) and Eq. (17), 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 $n \to +\infty$,

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

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 $n \to +\infty$,

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

$$\Downarrow$$

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

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

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

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.

Proof of the Fourth Result.
The fourth result is a simple conclusion of the property of AUC metric. Hence, we omit it. ■

Appendix C. Proof of Theorem 2

Before introducing the proof of Theorem 2, we extend Condition 1 to a general version (Condition 5). Then, Lemma 1 proves that Conditions 1 and 5 are the necessary conditions for the learnability of OOD detection. First, we provide the details of Condition 5.
Let $\Delta_l^{\circ} = \{(\lambda_1, ..., \lambda_l) : \sum_{j=1}^{l} \lambda_j < 1 \text{ and } \lambda_j \geq 0, \forall j = 1, ..., l\}$, where l is a positive integer. Next, we introduce an important definition as follows:

Condition 7 (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 $(\lambda_1, ..., \lambda_l) \in \Delta_l^{\circ}$. *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* $(\alpha_1, ..., \alpha_l) \in \Delta_l^{\circ}$,

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

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

Condition 5 (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*

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

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.

This is a more general condition compared to Condition 1. When $l = 1$ and the domain space $\mathscr{D}_{XY}$ is a priori-unknown space, Condition 5 degenerates into Condition 1. Lemma 1 shows that Condition 5 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 5 hold.*

Proof [Proof of Lemma 1]
Since Condition 1 is a special case of Condition 5, we only need to prove that Condition 5 holds.

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

$$\left|(1-\sum_{j=1}^{l}\alpha_j)\mathbb{E}_{S\sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} R^{\mathrm{in}}_D(\mathbf{A}(S)) + \sum_{j=1}^{l}\alpha_j \mathbb{E}_{S\sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} R_{Q_j}(\mathbf{A}(S)) - f_{D,Q}(\alpha_1,...,\alpha_l)\right| \leq \epsilon(n). \tag{18}$$

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. (18) implies that

$$\begin{aligned} \lim_{n\to+\infty} g_n(\alpha_1,...,\alpha_l) &= f_{D,Q}(\alpha_1,...,\alpha_l), \quad \forall(\alpha_1,...,\alpha_l)\in\Delta^{\mathrm{o}}_l, \\ \lim_{n\to+\infty} g_n(\mathbf{0}) &= f_{D,Q}(\mathbf{0}). \end{aligned} \tag{19}$$

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

$$\lim_{n\to+\infty} \mathbb{E}_{S\sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} R_{Q_j}(\mathbf{A}(S)) = f(\boldsymbol{\alpha}_j), \textit{i.e.}, \lim_{n\to+\infty} g_n(\boldsymbol{\alpha}_j) = f(\boldsymbol{\alpha}_j), \tag{20}$$

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

$$0 \leq \mathbb{E}_{S\sim D^n_{X_{\rm I}Y_{\rm I}}} R_{Q_j}(\mathbf{A}(S)) - \inf_{h\in\mathcal{H}} R_{Q_j}(h) \leq \epsilon(n). \tag{21}$$

Eq. (21) implies that

$$\lim_{n\to+\infty} \mathbb{E}_{S\sim D^n_{X_{\rm I}Y_{\rm I}}} R_{Q_j}(\mathbf{A}(S)) = \inf_{h\in\mathcal{H}} R_{Q_j}(h). \tag{22}$$

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

$$\lim_{n\to+\infty} \mathbb{E}_{S\sim D^n_{X_{\rm I}Y_{\rm I}}} R_{Q_j}(\mathbf{A}(S)) = f_{D,Q}(\boldsymbol{\alpha}_j),\ \textit{i.e.},\ \lim_{n\to+\infty} g_n(\boldsymbol{\alpha}_j) = f(\boldsymbol{\alpha}_j). \tag{23}$$

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

$$\begin{aligned}\lim_{n\to+\infty} g_n(\alpha_1, ..., \alpha_l) &= \lim_{n\to+\infty} \big((1-\sum_{j=1}^{l}\alpha_j) g_n(\mathbf{0}) + \sum_{j=1}^{l}\alpha_j g_n(\boldsymbol{\alpha}_j)\big) \\ &= (1-\sum_{j=1}^{l}\alpha_j) \lim_{n\to+\infty} g_n(\mathbf{0}) + \sum_{j=1}^{l}\alpha_j \lim_{n\to+\infty} g_n(\boldsymbol{\alpha}_j).\end{aligned} \tag{24}$$

According to Eq. (19) and Eq. (23), we have

$$\begin{aligned}&\lim_{n\to+\infty} g_n(\alpha_1, ..., \alpha_l) = f_{D,Q}(\alpha_1, ..., \alpha_l),\quad \forall(\alpha_1, ..., \alpha_l) \in \Delta^{\rm o}_l, \\ &\lim_{n\to+\infty} g_n(\mathbf{0}) = f_{D,Q}(\mathbf{0}), \\ &\lim_{n\to+\infty} g_n(\boldsymbol{\alpha}_j) = f(\boldsymbol{\alpha}_j),\end{aligned} \tag{25}$$

Combining Eq. (25) with Eq. (24), we complete the proof. ■

Lemma 2

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

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

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

Proof [Proof of Lemma 2] For the sake of convenience, we set $f_D(\alpha) = \inf_{h\in\mathcal{H}} R^{\alpha}_D(h)$, 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

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

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

$$R^{\alpha}_D(h^{\alpha}_{\epsilon}) \leq \inf_{h\in\mathcal{H}} R^{\alpha}_D(h) + \epsilon.$$

Note that

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

Therefore,

$$(1-\alpha)\inf_{h\in\mathcal{H}} R_D^{\text{in}}(h) + \alpha \inf_{h\in\mathcal{H}} R_D^{\text{out}}(h) \leq \inf_{h\in\mathcal{H}} R_D^{\alpha}(h) \leq R_D^{\alpha}(h_\epsilon^{\alpha}) \leq \inf_{h\in\mathcal{H}} R_D^{\alpha}(h) + \epsilon. \tag{26}$$

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

$$\inf_{h\in\mathcal{H}} R_D^{\alpha}(h) = (1-\alpha)\inf_{h\in\mathcal{H}} R_D^{\text{in}}(h) + \alpha \inf_{h\in\mathcal{H}} R_D^{\text{out}}(h), \forall \alpha \in [0,1). \tag{27}$$

Using Eqs. (26) and (27), we have that for any $0 \leq \alpha < 1$,

$$\epsilon \geq \big|R_D^{\alpha}(h_\epsilon^{\alpha}) - \inf_{h\in\mathcal{H}} R_D^{\alpha}(h)\big| = \big|(1-\alpha)\big(R_D^{\text{in}}(h_\epsilon^{\alpha}) - \inf_{h\in\mathcal{H}} R_D^{\text{in}}(h)\big) + \alpha\big(R_D^{\text{out}}(h_\epsilon^{\alpha}) - \inf_{h\in\mathcal{H}} R_D^{\text{out}}(h)\big)\big|. \tag{28}$$

Since $R_D^{\text{out}}(h_\epsilon^{\alpha}) - \inf_{h\in\mathcal{H}} R_D^{\text{out}}(h) \geq 0$ and $R_D^{\text{in}}(h_\epsilon^{\alpha}) - \inf_{h\in\mathcal{H}} R_D^{\text{in}}(h) \geq 0$, Eq. (28) implies that: for any $0 < \alpha < 1$,

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

Therefore,

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

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 under risk in the single-distribution space $\mathscr{D}_{XY}^{D_{XY}}$ for $\mathcal{H}$* ***if and only if*** *Condition 1 holds.*

Proof [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

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

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 $\alpha \in [0, 1]$,

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

We have completed this proof. ■

Appendix D. Proof of Theorem 3

Theorem 3 *Given a ranking function space $\mathcal{R}$ and a domain space $\mathscr{D}_{XY}$, if OOD detection is learnable under AUC for $\mathcal{R}$ in $\mathscr{D}_{XY}$, then for any $D_{XY}, D'_{XY} \in \mathscr{D}_{XY}$, the linear condition under AUC (i.e., Condition 2) holds.*

Proof [Proof of Theorem 3] According to Definition 3, we assume that $\mathbf{A}$ is the AUC learnable algorithm: there exists learning rate $\epsilon(n)$ such that for any $D_{XY} \in \mathscr{D}_{XY}$,

$$\mathbb{E}_{S\sim D^n_{X_{\text{I}}Y_{\text{I}}}} \big[\sup_{r\in\mathcal{R}} \text{AUC}(r; D_{XY}) - \text{AUC}(\mathbf{A}(S); D_{XY})\big] \leq \epsilon(n). \tag{29}$$

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

For any $D_{XY} = \beta D_{X_{\mathrm{I}}Y_{\mathrm{I}}} + (1-\beta) D_{X_{\mathrm{O}}Y_{\mathrm{O}}}$, $D'_{XY} = \beta' D_{X_{\mathrm{I}}Y_{\mathrm{I}}} + (1-\beta') D'_{X_{\mathrm{O}}Y_{\mathrm{O}}} \in \mathscr{D}_{XY}$, we set $D^{\alpha}_{X_{\mathrm{O}}} = \alpha D_{X_{\mathrm{O}}} + (1-\alpha) D'_{X_{\mathrm{O}}}$. Then it is clear that for any $\tilde{\beta} \in (0,1]$,

$$\begin{aligned}
&\mathrm{AUC}(\mathbf{A}(S); \tilde{\beta} D_{X_{\mathrm{I}}Y_{\mathrm{I}}} + (1-\tilde{\beta}) D^{\alpha}_{X_{\mathrm{O}}Y_{\mathrm{O}}}) \\
=&\mathrm{AUC}(\mathbf{A}(S); D_{X_{\mathrm{I}}}, D^{\alpha}_{X_{\mathrm{O}}}) \\
=&\alpha \mathrm{AUC}(\mathbf{A}(S); D_{X_{\mathrm{I}}}, D_{X_{\mathrm{O}}}) + (1-\alpha) \mathrm{AUC}(\mathbf{A}(S); D_{X_{\mathrm{I}}}, D'_{X_{\mathrm{O}}}) \\
=&\alpha \mathrm{AUC}(\mathbf{A}(S); D_{XY}) + (1-\alpha) \mathrm{AUC}(\mathbf{A}(S); D'_{XY}).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}_{S \sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} \big[\sup_{r \in \mathcal{R}} \mathrm{AUC}(r; \tilde{\beta} D_{X_{\mathrm{I}}Y_{\mathrm{I}}} + (1-\tilde{\beta}) D^{\alpha}_{X_{\mathrm{O}}Y_{\mathrm{O}}}) - \mathrm{AUC}(\mathbf{A}(S); \tilde{\beta} D_{X_{\mathrm{I}}Y_{\mathrm{I}}} + (1-\tilde{\beta}) D^{\alpha}_{X_{\mathrm{O}}Y_{\mathrm{O}}}) \big] \\
\leq &\mathbb{E}_{S \sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} \big[\alpha \sup_{r \in \mathcal{R}} \mathrm{AUC}(r; D_{XY}) + (1-\alpha) \sup_{r \in \mathcal{R}} \mathrm{AUC}(r; D'_{XY}) \\
&\quad - \alpha \mathrm{AUC}(\mathbf{A}(S); D_{XY}) - (1-\alpha) \mathrm{AUC}(\mathbf{A}(S); D'_{XY}) \big] \leq \epsilon(n),
\end{aligned}$$

which implies that

$$\sup_{r \in \mathcal{R}} \mathrm{AUC}(r; D_{X_{\mathrm{I}}}, D^{\alpha}_{X_{\mathrm{O}}}) = \lim_{n \to +\infty} \mathbb{E}_{S \sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} \mathrm{AUC}(\mathbf{A}(S); D_{X_{\mathrm{I}}}, D^{\alpha}_{X_{\mathrm{O}}}).$$

Note that

$$\begin{aligned}
&\lim_{n \to +\infty} \mathbb{E}_{S \sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} \mathrm{AUC}(\mathbf{A}(S); D_{X_{\mathrm{I}}}, D^{\alpha}_{X_{\mathrm{O}}}) \\
=&\alpha \lim_{n \to +\infty} \mathbb{E}_{S \sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} \mathrm{AUC}(\mathbf{A}(S); D_{X_{\mathrm{I}}}, D_{X_{\mathrm{O}}}) + (1-\alpha) \lim_{n \to +\infty} \mathbb{E}_{S \sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} \mathrm{AUC}(\mathbf{A}(S); D_{X_{\mathrm{I}}}, D'_{X_{\mathrm{O}}}) \\
=&\alpha \sup_{r \in \mathcal{R}} \mathrm{AUC}(r; D_{X_{\mathrm{I}}}, D_{X_{\mathrm{O}}}) + (1-\alpha) \sup_{r \in \mathcal{R}} \mathrm{AUC}(r; D_{X_{\mathrm{I}}}, D'_{X_{\mathrm{O}}}).
\end{aligned}$$

Therefore,

$$\sup_{r \in \mathcal{R}} \mathrm{AUC}(r; D_{X_{\mathrm{I}}}, D^{\alpha}_{X_{\mathrm{O}}}) = \alpha \sup_{r \in \mathcal{R}} \mathrm{AUC}(r; D_{X_{\mathrm{I}}}, D_{X_{\mathrm{O}}}) + (1-\alpha) \sup_{r \in \mathcal{R}} \mathrm{AUC}(r; D_{X_{\mathrm{I}}}, D'_{X_{\mathrm{O}}}).$$

■

Appendix E. Proofs of Theorem 4 and Theorem 5

E.1 Proof of Theorem 4

Lemma 22 *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 $\inf_{h \in \mathcal{H}} R^{\mathrm{in}}_D(h) = 0$, $\inf_{h \in \mathcal{H}} R^{\mathrm{out}}_D(h) = 0$, then Condition 1 does not hold. Therefore, OOD detection is not learnable under risk in $\mathscr{D}_{XY}$ for $\mathcal{H}$.*

Proof [Proof of Theorem 4] We **first** explain how we get f_{I} and f_{O} in Definition 5. 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 (Cohn, 2013), 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 $\mathcal{H}$. ■

E.2 Proof of Theorem 5

Theorem 5 (Impossibility Theorem for Total Space under Risk) *OOD detection is not learnable under risk 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 [Proof of Theorem 5] 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 5);

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 4, we have completed this proof. ■

Appendix F. Proof of Theorem 6

Before proving Theorem 6, 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 7 in Appendix C), and D_{XY} is a finite discrete distribution, then (the definition of $f_{D,Q}$ is given in Condition 5)*

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

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

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

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

$$(1 - \sum_{j=1}^{l} \lambda_j) \inf_{h \in \mathcal{H}} R_D^{\text{in}}(h) + \sum_{j=1}^{l} \lambda_j \inf_{h \in \mathcal{H}} R_{Q_j}(h) = \inf_{h \in \mathcal{H}} \Big((1 - \sum_{j=1}^{l} \lambda_j) R_D^{\text{in}}(h) + \sum_{j=1}^{l} \lambda_j R_{Q_j}(h) \Big). \quad (31)$$

Therefore, Eq. (30) and Eq. (31) imply that

$$(1 - \sum_{j=1}^{l} \lambda_j) \inf_{h \in \mathcal{H}} R_D^{\text{in}}(h) + \sum_{j=1}^{l} \lambda_j \inf_{h \in \mathcal{H}} R_{Q_j}(h) = (1 - \sum_{j=1}^{l} \lambda_j) R_D^{\text{in}}(h_0) + \sum_{j=1}^{l} \lambda_j R_{Q_j}(h_0). \quad (32)$$

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. (32), we have that

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

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,

$$\arg\min_{h\in\mathcal{H}} R_D(h) \subset \bigcap_{j=1}^{l} \arg\min_{h\in\mathcal{H}} R_{Q_j}(h) \bigcap \arg\min_{h\in\mathcal{H}} R_D^{\text{in}}(h). \tag{33}$$

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,

$$\bigcap_{j=1}^{l} \arg\min_{h\in\mathcal{H}} R_{Q_j}(h) \bigcap \arg\min_{h\in\mathcal{H}} R_D^{\text{in}}(h) \subset \arg\min_{h\in\mathcal{H}} R_D(h). \tag{34}$$

Combining Eq. (33) with Eq. (34), 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 $D_{XY} \in \mathscr{D}_{XY}^s$ such that $\inf_{h\in\mathcal{H}} R_D^{\text{out}}(\boldsymbol{h}) > 0$, then OOD detection is not learnable in $\mathscr{D}_{XY}^s$ for $\mathcal{H}$.*

Proof [Proof of Lemma 4] Suppose that $\text{supp}D_{X_{\text{O}}} = \{\mathbf{x}_1^{\text{out}}, ..., \mathbf{x}_l^{\text{out}}\}$, then it is clear that D_{XY} has OOD convex decomposition $\delta_{\mathbf{x}_1^{\text{out}}}, ..., \delta_{\mathbf{x}_l^{\text{out}}}$, 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 $\forall j = 1, ..., l$,

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

where

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

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

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

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

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

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

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

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

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

which implies that for any $h^* \in \arg\min_{h\in\mathcal{H}} R_D(h)$, we have $R_D^{\text{out}}(h^*) = 0$, which implies that $\inf_{h\in\mathcal{H}} R_D^{\text{out}}(h) = 0$. It is clear that $\inf_{h\in\mathcal{H}} R_D^{\text{out}}(h) = 0$ is **inconsistent** with the condition $\inf_{h\in\mathcal{H}} R_D^{\text{out}}(h) > 0$. Therefore, OOD detection is not learnable in $\mathscr{D}_{XY}^s$ 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 $\mathscr{D}_{XY}^s$ for $\mathcal{H}$, where ϕ maps ID's labels to* 1 *and maps OOD's labels to* 2.

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

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

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

Let $\mathcal{H}_{m+1}^{\phi}=\{(\phi\circ h(\mathbf{x}_1),...,\phi\circ h(\mathbf{x}_m),\phi\circ h(\mathbf{x}_{m+1}):h \text{ is any hypothesis function from } \mathcal{X} \text{ to } \mathcal{Y}_{\text{all}}\}$. Clearly, $\mathcal{H}_C^{\phi}\subset\mathcal{H}_{m+1}^{\phi}$ and $|\mathcal{H}_{m+1}^{\phi}|=2^{m+1}$. Sauer-Shelah-Perles Lemma (Lemma 6.10 in (Shalev-Shwartz and Ben-David, 2014)) implies that

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

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:

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) There exist $\mathbf{x}_j,\mathbf{x}_p\in C$ such that $\phi\circ h'(\mathbf{x}_j)=2$ and $\phi\circ h'(\mathbf{x}_p)=1$.

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

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

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

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

and

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

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

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

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

Therefore, if OOD detection is learnable in $\mathscr{D}^s_{XY}$ for $\mathcal{H}$, then $\inf_{h\in\mathcal{H}} R^{\text{out}}_D(h) > 0$.
Until now, we have constructed a domain D_{XY} (defined over $\mathcal{X}\times\mathcal{Y}_{\text{all}}$) with finite support and satisfying that $\inf_{h\in\mathcal{H}} R^{\text{out}}_D(h) > 0$. 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 $\mathscr{D}^s_{XY}$ for $\mathcal{H}$, which is **inconsistent** with our assumption that OOD detection is learnable in $\mathscr{D}^s_{XY}$ for $\mathcal{H}$. Therefore, OOD detection is not learnable in $\mathscr{D}^s_{XY}$ for $\mathcal{H}$. We have completed the proof. ■

Theorem 6 (Impossibility Theorem for Separate Space under Risk) *If Assumption 1 holds,* $\text{VCdim}(\phi\circ\mathcal{H}) < +\infty$ *and* $\sup_{h\in\mathcal{H}}|\{\mathbf{x}\in\mathcal{X}: h(\mathbf{x})\in\mathcal{Y}\}| = +\infty$*, OOD detection is not learnable under risk in the separate space* $\mathscr{D}^s_{XY}$ *for* $\mathcal{H}$*, where* ϕ *maps ID labels to* 1 *and maps OOD labels to* 2*.*

Proof [Proof of Theorem 6] Let $\text{VCdim}(\phi\circ\mathcal{H}) = v$. 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. ■

Appendix G. Proofs of Lemma 7 and Theorem 8

G.1 Proof of Lemma 7

Condition 8 *Given a ranking function space* $\mathcal{R}$*, the corresponding hypothesis space* $\mathcal{G}$ *consists of all* g_r *satisfying that there exists a* $r\in\mathcal{R}$ *such that*

$$g_r(\mathbf{x},\mathbf{x}') = \text{sign}(r(\mathbf{x}) - r(\mathbf{x}')).$$

Condition 9 (Equivalence) *Given measures* μ_1,μ_2*, we say two ranking functions* $r\in\mathcal{R}$ *and* $r'\in\mathcal{R}$ *are AUC equivalent over* μ_1,μ_2*, i.e.,* $f\sim f'$ *w.r.t.* μ_1,μ_2*, if and only if the corresponding hypothesis functions* $g_f = g_{f'}$ *a.e.* $\mu_1\times\mu_2$*.*

Lemma 6 *Assume that* $D_{X_{\text{I}}} = \int g_{X_{\text{I}}}\text{d}\mu$ *and* $D_{X_{\text{O}}} = \int g_{X_{\text{O}}}\text{d}\mu$*, then*

$$\sup_{r\in\mathcal{R}_{\text{all}}} \text{AUC}(f; D_{X_{\text{I}}}, D_{X_{\text{O}}}) = \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim\mu}\mathbb{E}_{\mathbf{x}'\sim\mu}\max\{g_{X_{\text{I}}}(\mathbf{x})g_{X_{\text{O}}}(\mathbf{x}'), g_{X_{\text{I}}}(\mathbf{x}')g_{X_{\text{O}}}(\mathbf{x})\}.$$

Proof Let $D(X) = \int g\text{d}\mu$, $D_{X_{\text{I}}} = \int g_{X_{\text{I}}}\text{d}\mu$ and $D_{X_{\text{O}}} = \int g_{X_{\text{O}}}\text{d}\mu$ satisfying that $g_{X_{\text{I}}} + g_{X_{\text{O}}} = 2g$. Let $P = \{\mathbf{x}\in\mathcal{X}: g(\mathbf{x}) > 0\}$, $P_{X_{\text{I}}} = \{\mathbf{x}\in\mathcal{X}: g_{X_{\text{I}}}(\mathbf{x}) > 0\}$ and $P_{X_{\text{O}}} = \{\mathbf{x}\in\mathcal{X}: g_{X_{\text{O}}}(\mathbf{x}) > 0\}$. To any two points $\mathbf{x}$ and $\mathbf{x}'$, we consider that

$$\begin{aligned} R(r,\mathbf{x},\mathbf{x}') = &\big[\mathbf{1}_{r(\mathbf{x})>r(\mathbf{x}')} + \frac{1}{2}\mathbf{1}_{r(\mathbf{x})=r(\mathbf{x}')}\big]g_{X_{\text{I}}}(\mathbf{x})g_{X_{\text{O}}}(\mathbf{x}') \\ &+ \big[\mathbf{1}_{r(\mathbf{x}')>r(\mathbf{x})} + \frac{1}{2}\mathbf{1}_{r(\mathbf{x})=r(\mathbf{x}')}\big]g_{X_{\text{I}}}(\mathbf{x}')g_{X_{\text{O}}}(\mathbf{x}). \end{aligned}$$

We set $r^*(\mathbf{x}) = \text{sigmoid}\big(g_{X_{\text{I}}}(\mathbf{x})/g_{X_{\text{O}}}(\mathbf{x})\big)$, if $g_{X_{\text{O}}}(\mathbf{x}) > 0$; otherwise, $r^*(\mathbf{x}) = 1$. It is clear that we have that

$$R(r^*,\mathbf{x},\mathbf{x}') = \max_{r\in\mathcal{R}_{\text{all}}} R(r,\mathbf{x},\mathbf{x}') = \max\{g_{X_{\text{I}}}(\mathbf{x})g_{X_{\text{O}}}(\mathbf{x}'), g_{X_{\text{I}}}(\mathbf{x}')g_{X_{\text{O}}}(\mathbf{x})\}.$$

Due to

$$2\mathrm{AUC}(r; D_{X_\mathrm{I}}, D_{X_\mathrm{O}}) = \int_P \int_P R(r, \mathbf{x}, \mathbf{x}') \mathrm{d}\mu(\mathbf{x}) \mathrm{d}\mu(\mathbf{x}'),$$

then

$$\begin{aligned} 2\mathrm{AUC}(r^*; D_{X_\mathrm{I}}, D_{X_\mathrm{O}}) &= \int_P \int_P \max_{r \in \mathcal{R}_{\mathrm{all}}} R(r, \mathbf{x}, \mathbf{x}') \mathrm{d}\mu(\mathbf{x}) \mathrm{d}\mu(\mathbf{x}') \\ \geq \max_{r \in \mathcal{R}_{\mathrm{all}}} \int_P \int_P R(r, \mathbf{x}, \mathbf{x}') \mathrm{d}\mu(\mathbf{x}) \mathrm{d}\mu(\mathbf{x}') &= \max_{r \in \mathcal{R}_{\mathrm{all}}} 2\mathrm{AUC}(r; D_{X_\mathrm{I}}, D_{X_\mathrm{O}}). \end{aligned}$$

Therefore, r^* is the optimal solution, and

$$\mathrm{AUC}(r^*; D_{X_\mathrm{I}}, D_{X_\mathrm{O}}) = \frac{1}{2} \mathbb{E}_{\mathbf{x} \sim \mu} \mathbb{E}_{\mathbf{x}' \sim \mu} \max\{g_{X_\mathrm{I}}(\mathbf{x}) g_{X_\mathrm{O}}(\mathbf{x}'), g_{X_\mathrm{I}}(\mathbf{x}') g_{X_\mathrm{O}}(\mathbf{x})\}.$$

We have completed this proof. ■

Lemma 7 *Given a ranking function space $\mathcal{R} \subset \mathcal{R}_{\mathrm{all}}$, $D_{X_\mathrm{I}} = \int g_{X_\mathrm{I}} \mathrm{d}\mu$, $D_{X_\mathrm{O}} = \int g_{X_\mathrm{O}} \mathrm{d}\mu$ and $D'_{X_\mathrm{O}} = \int g'_{X_\mathrm{O}} \mathrm{d}\mu$, if*

$$\begin{aligned} \sup_{r \in \mathcal{R}} \mathrm{AUC}(r; D_{X_\mathrm{I}}, D_{X_\mathrm{O}}) &= \sup_{r \in \mathcal{R}_{\mathrm{all}}} \mathrm{AUC}(r; D_{X_\mathrm{I}}, D_{X_\mathrm{O}}), \\ \sup_{r \in \mathcal{R}} \mathrm{AUC}(r; D_{X_\mathrm{I}}, D'_{X_\mathrm{O}}) &= \sup_{r \in \mathcal{R}_{\mathrm{all}}} \mathrm{AUC}(r; D_{X_\mathrm{I}}, D'_{X_\mathrm{O}}), \end{aligned}$$

and there exists $\alpha \in (0, 1)$ such that

$$\alpha \sup_{r \in \mathcal{R}} \mathrm{AUC}(r; D_{X_\mathrm{I}}, D_{X_\mathrm{O}}) + (1 - \alpha) \sup_{r \in \mathcal{R}} \mathrm{AUC}(r; D_{X_\mathrm{I}}, D'_{X_\mathrm{O}}) = \sup_{r \in \mathcal{R}} \mathrm{AUC}(r; D_{X_\mathrm{I}}, D^{\alpha}_{X_\mathrm{O}}),$$

where $D^{\alpha}_{X_\mathrm{O}} = \alpha D_{X_\mathrm{O}} + (1 - \alpha) D'_{X_\mathrm{O}}$. Then

$$\frac{g_{X_\mathrm{I}}}{g_{X_\mathrm{I}} + g_{X_\mathrm{O}}} \sim \frac{g_{X_\mathrm{I}}}{g_{X_\mathrm{I}} + g'_{X_\mathrm{O}}} \ \ w.r.t. \ \ D_{X_\mathrm{I}}|_{P'_{X_\mathrm{O}} - P_{X_\mathrm{O}}}, D_{X_\mathrm{I}}|_{P_{X_\mathrm{O}} - P'_{X_\mathrm{O}}},$$

where $P_{X_\mathrm{O}} = \{\mathbf{x} : g_{X_\mathrm{O}}(\mathbf{x}) > 0\}$ and $P'_{X_\mathrm{O}} = \{\mathbf{x} : g'_{X_\mathrm{O}}(\mathbf{x}) > 0\}$

Proof Let $\eta_{X_\mathrm{I}} = \frac{g_{X_\mathrm{I}}}{g_{X_\mathrm{I}} + g_{X_\mathrm{O}}}$ and $\eta'_{X_\mathrm{I}} = \frac{g_{X_\mathrm{I}}}{g_{X_\mathrm{I}} + g'_{X_\mathrm{O}}}$. It is easy to check that the proof process of Lemma 6 implies that to each $i = 1, 2$,

$$\eta^i_{X_\mathrm{I}} \in \arg\max_{r \in \mathcal{R}_{\mathrm{all}}} \mathrm{AUC}(f; D_{X_\mathrm{I}}, D^i_{X_\mathrm{O}}).$$

Additionally,

$$\begin{aligned} \sup_{r \in \mathcal{R}} \mathrm{AUC}(r; D_{X_\mathrm{I}}, D^{\alpha}_{X_\mathrm{O}}) =& \alpha \sup_{r \in \mathcal{R}} \mathrm{AUC}(r; D_{X_\mathrm{I}}, D_{X_\mathrm{O}}) + (1 - \alpha) \sup_{r \in \mathcal{R}} \mathrm{AUC}(r; D_{X_\mathrm{I}}, D'_{X_\mathrm{O}}) \\ =& \alpha \sup_{r \in \mathcal{R}_{\mathrm{all}}} \mathrm{AUC}(r; D_{X_\mathrm{I}}, D_{X_\mathrm{O}}) + (1 - \alpha) \sup_{r \in \mathcal{R}_{\mathrm{all}}} \mathrm{AUC}(r; D_{X_\mathrm{I}}, D'_{X_\mathrm{O}}) \\ \geq& \sup_{r \in \mathcal{R}_{\mathrm{all}}} \mathrm{AUC}(r; D_{X_\mathrm{I}}, D^{\alpha}_{X_\mathrm{O}}) \geq \sup_{r \in \mathcal{R}} \mathrm{AUC}(r; D_{X_\mathrm{I}}, D^{\alpha}_{X_\mathrm{O}}). \end{aligned}$$

Therefore, under the condition that

$$\sup_{r\in\mathcal{R}} \mathrm{AUC}(r; D_{X_\mathrm{I}}, D_{X_\mathrm{O}}) = \sup_{r\in\mathcal{R}_\mathrm{all}} \mathrm{AUC}(r; D_{X_\mathrm{I}}, D_{X_\mathrm{O}}),$$
$$\sup_{r\in\mathcal{R}} \mathrm{AUC}(r; D_{X_\mathrm{I}}, D'_{X_\mathrm{O}}) = \sup_{r\in\mathcal{R}_\mathrm{all}} \mathrm{AUC}(r; D_{X_\mathrm{I}}, D'_{X_\mathrm{O}}),$$

we obtain that

$$\sup_{r\in\mathcal{R}_\mathrm{all}} \mathrm{AUC}(r; D_{X_\mathrm{I}}, D^{\alpha}_{X_\mathrm{O}}) = \alpha \sup_{r\in\mathcal{R}_\mathrm{all}} \mathrm{AUC}(r; D_{X_\mathrm{I}}, D_{X_\mathrm{O}}) + (1-\alpha) \sup_{r\in\mathcal{R}_\mathrm{all}} \mathrm{AUC}(r; D_{X_\mathrm{I}}, D'_{X_\mathrm{O}}).$$

Because $\sup_{r\in\mathcal{R}_\mathrm{all}} \mathrm{AUC}(r; D_{X_\mathrm{I}}, D_{X_\mathrm{O}})$ and $\sup_{r\in\mathcal{R}_\mathrm{all}} \mathrm{AUC}(r; D_{X_\mathrm{I}}, D_{X_\mathrm{O}})$ are attainable (the proof process of Lemma 6 implies this), it is easy to check (similar the proof of Lemma 3) that

$$\arg\max_{r\in\mathcal{R}_\mathrm{all}} \mathrm{AUC}(r; D_{X_\mathrm{I}}, D_{X_\mathrm{O}}) \cap \arg\max_{r\in\mathcal{R}_\mathrm{all}} \mathrm{AUC}(r; D_{X_\mathrm{I}}, D'_{X_\mathrm{O}}) \neq \emptyset.$$

Combining with Lemma 6, above equality implies there exists r^* such that

$$r^* \sim \eta_{X_\mathrm{I}}, \quad \text{w.r.t. } D_{X_\mathrm{I}}, D_{X_\mathrm{O}}, \quad r^* \sim \eta'_{X_\mathrm{I}}, \quad \text{w.r.t. } D_{X_\mathrm{I}}, D'_{X_\mathrm{O}}$$

Therefore,

$$\eta_{X_\mathrm{I}} \sim \eta'_{X_\mathrm{I}} \text{ w.r.t. } D_{X_\mathrm{I}}|_{P'_{X_\mathrm{O}} - P_{X_\mathrm{O}}}, D_{X_\mathrm{I}}|_{P_{X_\mathrm{O}} - P'_{X_\mathrm{O}}}$$

We have completed the proof. ■

Lemma 23 *Given a ranking function space $\mathcal{R}$, a domain space $\mathscr{D}_{XY}$ and $D_{XY} = \beta D_{X_\mathrm{I}Y_\mathrm{I}} + (1-\beta) D_{X_\mathrm{O}Y_\mathrm{O}}$, $D'_{XY} = \beta' D_{X_\mathrm{I}Y_\mathrm{I}} + (1-\beta') D'_{X_\mathrm{O}Y_\mathrm{O}} \in \mathscr{D}_{XY}$, let P be the overlap set between D_{X_I} and D_{X_O} and P' be the overlap set between D_{X_I} and D'_{X_O} based on the Definition 5. If*

$$\sup_{r\in\mathcal{R}} \mathrm{AUC}(r; D_{X_\mathrm{I}}, D_{X_\mathrm{O}}) = \sup_{r\in\mathcal{R}_\mathrm{all}} \mathrm{AUC}(r; D_{X_\mathrm{I}}, D_{X_\mathrm{O}})$$
$$\sup_{r\in\mathcal{R}} \mathrm{AUC}(r; D_{X_\mathrm{I}}, D'_{X_\mathrm{O}}) = \sup_{r\in\mathcal{R}_\mathrm{all}} \mathrm{AUC}(r; D_{X_\mathrm{I}}, D'_{X_\mathrm{O}}),$$

and $D_{X_\mathrm{I}}(P \cap P') < \min\{D_{X_\mathrm{I}}(P), D_{X_\mathrm{I}}(P')\}$, then Condition 2 does not hold, where $\mathcal{R}_\mathrm{all}$ is a ranking function space consisting of all ranking functions from $\mathcal{X}$ to $\mathbb{R}$. Therefore, OOD detection is not learnable under AUC in $\mathscr{D}_{XY}$ for $\mathcal{R}$.

Proof [Proof of Lemma 7] By the condition that $D_{X_\mathrm{I}}(P_1 \cap P_2) < \min\{D_{X_\mathrm{I}}(P_1), D_{X_\mathrm{I}}(P_2)\}$, we can ensure that

$$D_{X_\mathrm{I}}(P_1 - P_2) > 0, \quad D_{X_\mathrm{I}}(P_2 - P_1) > 0.$$

By this, one can easily check that

$$\frac{g_{X_\mathrm{I}}}{g_{X_\mathrm{I}} + g_{X_\mathrm{O}}} \sim \frac{g_{X_\mathrm{I}}}{g_{X_\mathrm{I}} + g'_{X_\mathrm{O}}} \text{ w.r.t. } D_{X_\mathrm{I}}|_{P_2 - P_1}, D_{X_\mathrm{I}}|_{P_1 - P_2} \text{ does not hold.}$$

Therefore, Lemma 7 implies that Condition 2 does not hold. ■

G.2 Proof of Theorem 8

Theorem 8 (Impossibility Theorem for Total Space under AUC) *Given ranking function space $\mathcal{R}$, if there exist $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ and $r, r' \in \mathcal{R}$ such that*

$$r(\mathbf{x}) > r(\mathbf{x}') \text{ and } r'(\mathbf{x}') > r'(\mathbf{x}),$$

then the learnability of OOD detection under AUC is not distribution-free for $\mathcal{R}$.

Proof [Proof of Theorem 8] Let $D_{X_{\mathrm{I}}} = \frac{1}{2}\delta_{\mathbf{x}} + \frac{1}{2}\delta_{\mathbf{x}'}$, $D_{X_{\mathrm{O}}} = \delta_{\mathbf{x}}$ and $D_{X'_{\mathrm{O}}} = \delta_{\mathbf{x}'}$, where δ is the Dirac measure. Then the condition in Lemma 7 holds, which implies the result of Theorem 8. ■

Appendix H. Proof of Theorem 9

Lemma 8 *Given a separate ranking function space $\mathcal{R}$, if there exists finite discrete $D_{XY} \subset \mathscr{D}^s_{XY}$ such that*

$$\sup_{r\in\mathcal{R}} \mathrm{AUC}(r; D_{XY}) < 1,$$

then OOD detection is not learnable under AUC in $\mathscr{D}^s_{XY}$ for $\mathcal{R}$.

Proof [Proof of Lemma 8] Assume that if OOD detection is learnable under AUC in $\mathscr{D}_{XY}$ for $\mathcal{R}$, then Condition 2 implies that: let $D_{X_{\mathrm{O}}} = \sum_{i=1}^m \lambda_i \delta_{\mathbf{x}_i}$ ($\sum_{i=1}^m = 1$),

$$\sup_{r\in\mathcal{R}} \mathrm{AUC}(r; D_{XY}) = \sum_{i=1}^m \lambda_i \sup_{r\in\mathcal{R}} \mathrm{AUC}(r; D_{X_{\mathrm{I}}}, \delta_{\mathbf{x}_i}).$$

Because $\mathcal{R}$ is the separate ranking function space, we know that

$$\sup_{r\in\mathcal{R}} \mathrm{AUC}(r; D_{XY}) = \sum_{i=1}^m \lambda_i \sup_{r\in\mathcal{R}} \mathrm{AUC}(r; D_{X_{\mathrm{I}}}, \delta_{\mathbf{x}_i}) = 1,$$

which is conflict with the condition that $\sup_{r\in\mathcal{R}} \mathrm{AUC}(r; D_{XY}) < 1$. We have completed this proof. ■

Theorem 9 (Impossibility Theorem for Separate Space under AUC) *Given a separate ranking function space $\mathcal{R}$, if $\mathrm{VC}[\phi\circ\mathcal{R}] = d < +\infty$ and $|\mathcal{X}| \geq (28d{+}14)\log(14d{+}7)$, then OOD detection is not learnable under AUC in $\mathscr{D}^s_{XY}$ for $\mathcal{R}$, where $\phi\circ\mathcal{R} = \{\mathbf{1}_{r(\mathbf{x})>r(\mathbf{x}')} : r \in \mathcal{R}\}$.*

Proof [Proof of Theorem 9] Given disjoint samples $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_m\}$. Consider the following matrix and set

$$\mathbf{B}_r[\mathbf{X}] = \left[\mathbf{1}_{r(\mathbf{x}_i)>r(\mathbf{x}_j)}\right], \quad \mathbf{B}_{\mathcal{R}}[\mathbf{X}] = \{\mathbf{B}_r[\mathbf{X}] : r \in \mathcal{R}\}.$$

Let $\mathcal{D}_{X_{\mathrm{I}},X_{\mathrm{O}}}[\mathbf{X}]$ be the set consisting of all $(D_{X_{\mathrm{I}}}, D_{X_{\mathrm{O}}})$ which satisfies the following conditions:

- $D_{X_{\mathrm{I}}}, D_{X_{\mathrm{O}}}$ are from the separate space;

- $D_{X_\mathrm{I}}, D_{X_\mathrm{O}}$ are uniform distributions;
- $\mathrm{supp}D_{X_\mathrm{I}} \cup \mathrm{supp}D_{X_\mathrm{O}} = \mathbf{X}$.

Additionally, let

$$\mathcal{D}_{\mathcal{R}}[\mathbf{X}] = \{(D_{X_\mathrm{I}}, D_{X_\mathrm{O}}) \in \mathcal{D}_{X_\mathrm{I},X_\mathrm{O}}[\mathbf{X}] : \exists r \in \mathcal{R}, \mathrm{AUC}(r; D_{X_\mathrm{I}}, D_{X_\mathrm{O}}) = 1\}.$$

It is clear that

$$|\mathbf{B}_{\mathcal{R}}[\mathbf{X}]| \leq \frac{e^d}{d^d} m^{2d}; \quad |\mathcal{D}_{\mathcal{R}}[\mathbf{X}]| \leq (m-1)|\mathbf{B}_{\mathcal{R}}[\mathbf{X}]| \leq \frac{e^d}{d^d} m^{2d+1}; \quad |\mathcal{D}_{X_\mathrm{I},X_\mathrm{O}}[\mathbf{X}]| = 2^m - 2.$$

When m is large enough ($m \geq (28d+14)\log(14d+7)$), we have that

$$|\mathcal{D}_{\mathcal{R}}[\mathbf{X}]| < |\mathcal{D}_{X_\mathrm{I},X_\mathrm{O}}[\mathbf{X}]|.$$

Therefore, we can find

$$(D_{X_\mathrm{I}}, D_{X_\mathrm{O}}) \in \mathcal{D}_{X_\mathrm{I},X_\mathrm{O}}[\mathbf{X}] \text{ such that } \sup_{r \in \mathcal{R}} \mathrm{AUC}(r; D_{X_\mathrm{I}}, D_{X_\mathrm{O}}) < 1.$$

By Lemma 8, we have completed this proof. ■

Appendix I. Proofs of Theorem 10 and Theorem 11

I.1 Proof of Theorem 10

Firstly, we need two lemmas, which are motivated by Lemma 19.2 and Lemma 19.3 in (Shalev-Shwartz and Ben-David, 2014).

Lemma 9 *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_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 [Proof of Lemma 9]

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

where $\mathbf{1}$ is the characteristic function.
For each i,

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

Therefore,

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

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

Lemma 10 *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*

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

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

Proof [Proof of Lemma 10] 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 $C_1, C_2, ..., C_r$ be a cover of $\mathcal{X}$: for every $(a_1, ..., a_T) \in [T]^d := [1, ..., T]^d$, there exists a $C_i = \{\mathbf{x} = (x_1, ..., x_d) : \forall j \in \{1, ..., d\}, x_j \in [(a_j - 1)/T, a_j/T)\}$.

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,

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

Note that $C_1, ..., C_r$ 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 9, we obtain

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

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

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

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

Theorem 10 *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 under risk 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.*

Proof [Proof of Theorem 10] **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 $\mathscr{D}_{XY}^s$ 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 $D_{XY} \in \mathscr{D}_{XY}^s$: if $C_{\text{O}} = \emptyset$, then $D_{XY} = D_{X_{\text{I}}} * D_{Y_{\text{I}}|X_{\text{I}}}$; otherwise,

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

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

and

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

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, $R_D^{\text{in}}(h^{\text{in}}) = 0$ (here $h^{\text{in}} = 1$), hence, $\inf_{h \in \mathcal{H}} R_D^{\text{in}}(h) = 0$.
Since OOD detection is learnable in $\mathscr{D}_{XY}^s$ for $\mathcal{H}$, Lemma 1 implies that

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

where $\pi^{\text{out}} = D_Y(Y = K + 1) = 1$ or 0.5. Since $\inf_{h \in \mathcal{H}} R_D^{\text{in}}(h) = 0$ and $\inf_{h \in \mathcal{H}} R_D(h) > 0$, we obtain that $\inf_{h \in \mathcal{H}} R_D^{\text{out}}(h) > 0$.
Until now, we have constructed a special domain $D_{XY} \in \mathscr{D}_{XY}^s$ satisfying that $\inf_{h \in \mathcal{H}} R_D^{\text{out}}(h) > 0$. Using Lemma 4, we know that OOD detection in $\mathscr{D}_{XY}^s$ is not learnable for $\mathcal{H}$, which is **inconsistent** with the condition that OOD detection is learnable in $\mathscr{D}_{XY}^s$ 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 $\mathscr{D}_{XY}^s$ 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,

$$\mathbb{E}_{S\sim D^n_{X_{\text{I}}Y_{\text{I}}}} R^{\text{out}}_D(\mathbf{A}(S)) = 0. \tag{35}$$

Using Lemma 10, 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

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

where $D_{X_{\text{I}}} \times D^n_{X_{\text{I}}Y_{\text{I}}}$ is the product measure of $D_{X_{\text{I}}}$ and $D^n_{X_{\text{I}}Y_{\text{I}}}$ (Cohn, 2013). Therefore,

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

which implies that

$$\mathbb{E}_{S\sim D^n_{X_{\text{I}}Y_{\text{I}}}} R^{\text{in}}_D(\mathbf{A}(S)) \leq 2B\epsilon_{\text{cons}}(n)/d_0, \tag{36}$$

where $B = \max\{\ell(1, 2), \ell(2, 1)\}$. Using Eq. (35) and Eq. (36), we have proved that

$$\mathbb{E}_{S\sim D^n_{X_{\text{I}}Y_{\text{I}}}} R_D(\mathbf{A}(S)) \leq 0 + 2B\epsilon_{\text{cons}}(m)/d_0 \leq \inf_{h\in\mathcal{H}} R_D(h) + 2B\epsilon_{\text{cons}}(m)/d_0. \tag{37}$$

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

I.2 Proof of Theorem 11

Theorem 11 *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 3 holds, then OOD detection is learnable under risk in $\mathscr{D}^s_{XY}$ for $\mathcal{H}$, where $\mathcal{H}_{\text{all}}$ and h^{out} are defined in Theorem 10.*

Proof [Proof of Theorem 11] Since $|\mathcal{X}| < +\infty$, we know that $|\mathcal{H}| < +\infty$, which implies that $\mathcal{H}^{\text{in}}$ is agnostic PAC learnable for supervised learning in classification. Therefore, there exist an algorithm $\mathbf{A}^{\text{in}} : \cup_{n=1}^{+\infty}(\mathcal{X} \times \mathcal{Y})^n \to \mathcal{H}^{\text{in}}$ and a monotonically decreasing sequence $\epsilon(n)$, such that $\epsilon(n) \to 0$, as $n \to +\infty$, and for any $D_{XY} \in \mathscr{D}^s_{XY}$,

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

Since $|\mathcal{X}| < +\infty$ and $\mathcal{H}^{\text{b}}$ almost contains all binary classifiers, then using Theorem 10 and Theorem 1, we obtain that there exist an algorithm $\mathbf{A}^{\text{b}} : \cup_{n=1}^{+\infty}(\mathcal{X} \times \{1, 2\})^n \to \mathcal{H}^{\text{b}}$ and a

monotonically decreasing sequence $\epsilon'(n)$, such that $\epsilon'(n) \to 0$, as $n \to +\infty$, and for any $D_{XY} \in \mathscr{D}^s_{XY}$,

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

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

$$R^{\mathrm{in}}_{\phi(D)}(\mathbf{A}^{\mathrm{b}}(\phi(S))) = \int_{\mathcal{X}\times\mathcal{Y}} \ell(\mathbf{A}^{\mathrm{b}}(\phi(S))(\mathbf{x}), \phi(y))\mathrm{d}D_{X_{\mathrm{I}}Y_{\mathrm{I}}}(\mathbf{x}, y), \tag{38}$$

$$R^{\mathrm{in}}_{\phi(D)}(h) = \int_{\mathcal{X}\times\mathcal{Y}} \ell(h(\mathbf{x}), \phi(y))\mathrm{d}D_{X_{\mathrm{I}}Y_{\mathrm{I}}}(\mathbf{x}, y), \tag{39}$$

$$R^{\mathrm{out}}_{\phi(D)}(\mathbf{A}^{\mathrm{b}}(\phi(S))) = \int_{\mathcal{X}\times\{K+1\}} \ell(\mathbf{A}^{\mathrm{b}}(\phi(S))(\mathbf{x}), \phi(y))\mathrm{d}D_{X_{\mathrm{O}}Y_{\mathrm{O}}}(\mathbf{x}, y), \tag{40}$$

and

$$R^{\mathrm{out}}_{\phi(D)}(h) = \int_{\mathcal{X}\times\{K+1\}} \ell(h(\mathbf{x}), \phi(y))\mathrm{d}D_{X_{\mathrm{O}}Y_{\mathrm{O}}}(\mathbf{x}, y), \tag{41}$$

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,

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

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

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

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 3, 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:

$$\begin{aligned} \mathbb{E}_{S\sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} R^{\mathrm{in}}_D(\mathbf{A}(S)) &\leq \mathbb{E}_{S\sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} R^{\mathrm{in}}_D(\mathbf{A}^{\mathrm{in}}(S)) + c\mathbb{E}_{S\sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} R^{\mathrm{in}}_{\phi(D)}(\mathbf{A}^{\mathrm{b}}(\phi(S))) \\ &\leq \inf_{h\in\mathcal{H}^{\mathrm{in}}} R^{\mathrm{in}}_D(h) + \epsilon(n) + c\epsilon'(n) = \inf_{h\in\mathcal{H}} R^{\mathrm{in}}_D(h) + \epsilon(n) + c\epsilon'(n), \end{aligned} \tag{42}$$

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 $R_D^{\text{out}}(\mathbf{A}(S))$ is from the case that OOD data are detected as ID data. Therefore,

$$\begin{aligned}\mathbb{E}_{S\sim D^n_{X_{\text{I}}Y_{\text{I}}}} R_D^{\text{out}}(\mathbf{A}(S)) &\leq c\mathbb{E}_{S\sim D^n_{X_{|rmIY_{\text{I}}}}} R^{\text{out}}_{\phi(D)}(\mathbf{A}^{\text{b}}(\phi(S))) \\ &\leq c\epsilon'(n) \leq \inf_{h\in\mathcal{H}} R_D^{\text{out}}(h) + c\epsilon'(n).\end{aligned} \tag{43}$$

Note that $(1-\alpha)\inf_{h\in\mathcal{H}} R_D^{\text{in}}(h) + \alpha \inf_{h\in\mathcal{H}} R_D^{\text{out}}(h) \leq \inf_{h\in\mathcal{H}} R_D^{\alpha}(h)$. Then, using Eq. (42) and Eq. (43), we obtain that for any $\alpha \in [0,1]$,

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

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

Appendix J. Proof of Theorem 12

Lemma 11 *Suppose that $|\mathcal{X}| < +\infty$. If AUC-based Realizability Assumption holds for AUC metric, then OOD detection is learnable under AUC in separate space for $\mathcal{R}$.*

Proof [Proof of Lemma 11] Without loss of generality, we assume that $K = 1$, and any $r \in \mathcal{R}$ satisfies that $0 < r < 1$ (one can achieve this by using sigmoid function). Given m data points $S_m = \{\mathbf{x}'_1, ..., \mathbf{x}'_m\} \subset \mathcal{X}^m$. We consider the following learning rule

$$\max_{r\in\mathcal{R},\tau\in(0,1)} \sum_{i=1}^{m} \mathbf{1}_{r(\mathbf{x}'_i)\leq\tau}\mathbf{1}_{\mathbf{x}'_i\notin S}, \text{ subject to } \frac{1}{n}\sum_{j=1}^{n}\mathbf{1}_{r(\mathbf{x}_j)\leq\tau} = 0.$$

We denote the algorithm, which solves the above rule, as $\mathbf{A}_{S_m}$ ($\mathbf{A}_{S_m}$ outputs r and a corresponding τ). For different data points S_m, we have different algorithm $\mathbf{A}_{S_m}$. Let $\mathcal{S}$ be the set that consists of all data points, *i.e.*,

$$\mathcal{S} := \{S_m : S_m \text{ are any } m \text{ data points, } m = 1, ..., +\infty\}. \tag{44}$$

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

$$\mathscr{A} := \{\mathbf{A}_{S'} : \forall\ S' \in \mathcal{S}\}.$$

We will find an algorithm $\mathbf{A}$ from $\mathscr{A}$, which is learnable. Let $S' = \mathcal{X}$. We will show that $\mathbf{A}_{\mathcal{X}}$ can guarantee the learnability. Suppose that r_S and τ_S is the output of $\mathbf{A}_{\mathcal{X}}(S)$, then the realizability assumption implies that there exists learning rate $\epsilon(n)$ such that

$$\mathbb{E}_{S\sim D^n_{X_{\text{I}}}}\mathbb{E}_{\mathbf{x}\sim D_{X_{\text{I}}}}\mathbf{1}_{r_S(\mathbf{x})\leq\tau_S} \leq \epsilon(n). \tag{45}$$

Additionally, due to $\text{supp}(D_{X_{\text{O}}}) \subset \mathcal{X} - S$, the AUC-based Realizability Assumption implies that

$$\mathbb{E}_{S\sim D^n_{X_{\text{I}}}}\mathbb{E}_{\mathbf{x}\sim D_{X_{\text{O}}}}\mathbf{1}_{r_S(\mathbf{x})\leq\tau_S} = 1. \tag{46}$$

Then,

$$\begin{aligned}
&\mathbb{E}_{S\sim D_{X_{\mathrm{I}}}^n}\mathrm{AUC}(f_S; D_{X_{\mathrm{I}}}, D_{X_{\mathrm{O}}})\\
\geq&\mathbb{E}_{S\sim D_{X_{\mathrm{I}}}^n}\mathbb{E}_{\mathbf{x}\sim D_{X_{\mathrm{O}}}}\mathbb{E}_{\mathbf{x}'\sim D_{X_{\mathrm{I}}}}\mathbf{1}_{r_S(\mathbf{x})<r_S(\mathbf{x}')}\\
\geq&\mathbb{E}_{S\sim D_{X_{\mathrm{I}}}^n}\mathbb{E}_{\mathbf{x}\sim D_{X_{\mathrm{O}}}}\mathbb{E}_{\mathbf{x}'\sim D_{X_{\mathrm{I}}}}\mathbf{1}_{r_S(\mathbf{x})\leq\tau_S}\mathbf{1}_{r_S(\mathbf{x}')>\tau_S}\\
\geq&1-\epsilon(n).
\end{aligned}$$

Then the ranking function part r_S of $\mathbf{A}_{\mathcal{X}} \in \mathscr{A}$ is the universally consistent algorithm, *i.e.*,

$$\mathbb{E}_{S\sim D_{X_{\mathrm{I}}}^n}\mathrm{AUC}(r_S; D_{X_{\mathrm{I}}}, D_{X_{\mathrm{O}}}) \geq 1-\epsilon(n).$$

■

Theorem 12 *Given a separate ranking function space $\mathcal{R}$, if $|\mathcal{X}| < +\infty$, then OOD detection is learnable under AUC in the separate space $\mathscr{D}_{XY}^s$ for $\mathcal{R}$* ***if and only if*** *AUC-based Realizability Assumption holds.*

Proof [Proof of Theorem 12] Lemmas 8 and 11 imply this result. ■

Appendix K. Proofs of Theorems 13 and 14

K.1 Proof of Theorem 13

Lemma 12 *Given a prior-unknown space $\mathscr{D}_{XY}$ and a hypothesis space $\mathcal{H}$, if Condition 4 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

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

The existence of $\tilde{h}_n$ is based on Condition 4. It is easy to check that for any $D_{XY} \in [D'_{XY}]$,

$$\begin{aligned}
&\mathbb{E}_{S\sim D_{X_{\mathrm{I}}Y_{\mathrm{I}}}^n} R_D^{\mathrm{in}}(\mathbf{A}_{\tilde{\boldsymbol{h}}}(S)) \leq \inf_{h\in\mathcal{H}} R_D^{\mathrm{in}}(h) + 2/n.\\
&\mathbb{E}_{S\sim D_{X_{\mathrm{I}}Y_{\mathrm{I}}}^n} R_D^{\mathrm{out}}(\mathbf{A}_{\tilde{\boldsymbol{h}}}(S)) \leq \inf_{h\in\mathcal{H}} R_D^{\mathrm{out}}(h) + 2/n.
\end{aligned}$$

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 $\alpha \in [0,1]$,

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

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

Theorem 13 *Suppose that $\mathcal{X}$ is bounded. OOD detection is learnable under risk in $\mathscr{D}_{XY}^F$ for $\mathcal{H}$ **if and only if** the compatibility condition (i.e., Condition 4) holds. Furthermore, the learning rate $\epsilon_{\text{cons}}(n)$ can attain $O(1/\sqrt{n^{1-\theta}})$, for any $\theta \in (0,1)$.*

Proof [Proof of Theorem 13] **First**, we prove that if OOD detection is learnable in $\mathscr{D}_{XY}^F$ for $\mathcal{H}$, then Condition 4 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 4 holds.

Second, we prove Condition 4 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 12, 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 12, 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) (Gretton et al., 2012), which is a metric between distributions: for any distributions P and Q defined over $\mathcal{Z}$, we use $\mathrm{MMD}_K(Q, P)$ 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

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

here

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

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 (Gretton et al., 2012) 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 $D_{XY} \in \mathscr{D}$,

$$\mathbb{E}_{S \sim D^n_{X_\mathrm{I} Y_\mathrm{I}}} \mathrm{MMD}_K(D_{X_\mathrm{I} Y_\mathrm{I}}, P_S) \leq \epsilon(n), \text{ where } \epsilon(n) = O(\frac{1}{\sqrt{n^{1-\theta}}}). \tag{47}$$

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. (47), 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

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

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

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

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

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

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

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 $\alpha \in [0,1]$,

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

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

K.2 Proof of Theorem 14

Theorem 14 *Given a density-based space $\mathscr{D}^{\mu,b}_{XY}$, if $\mu(\mathcal{X}) < +\infty$, the Risk-based Realizability Assumption holds, then when $\mathcal{H}$ has finite Natarajan dimension (Shalev-Shwartz and Ben-David, 2014), 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 [Proof of Theorem 14] **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 $D_{X_{\mathrm{I}}}$'s density function and f_{O} is $D_{X_{\mathrm{O}}}$'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 Risk-based 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:

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

We denote the algorithm, which solves the above rule, as $\mathbf{A}_{S_m}$ [8]. 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.*,

$$\mathcal{S} := \{\mathbf{S} := (S_1, S_2, ..., S_m, ...) : S_m \text{ are any } m \text{ data points, } m = 1, ..., +\infty\}. \tag{48}$$

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

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

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

Next, we prove that there exists an algorithm $\mathbf{A_S} \in \mathscr{A}$, 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 (Shalev-Shwartz and Ben-David, 2014), 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)

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

and because of $\mathcal{H}_S \subset \mathcal{H}$,

$$\mathbb{E}_{S_n\sim\mu^n} \sup_{S\in(\mathcal{X}\times\mathcal{Y})^n} [R_\mu(\mathbf{A}_{S_n}(S), K+1) - \inf_{h\in\mathcal{H}_S} R_\mu(h, K+1)] \le \frac{C_\theta}{\sqrt{n^{1-\theta}}}, \tag{49}$$

where

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

and

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

We set $\mathscr{D}_{\mathrm{I}} = \{D_{X_{\mathrm{I}}Y_{\mathrm{I}}} : \text{there exists } D_{X_{\mathrm{O}}Y_{\mathrm{O}}} \text{such that } (1-\alpha)D_{X_{\mathrm{I}}Y_{\mathrm{I}}} + \alpha D_{X_{\mathrm{O}}Y_{\mathrm{O}}} \in \mathscr{D}^{\mu,b}_{XY}\}$. Then by Eq. (49), we have

$$\mathbb{E}_{S_n\sim\mu^n} \sup_{D_{X_{\mathrm{I}}Y_{\mathrm{I}}}\in\mathscr{D}_{\mathrm{I}}} \mathbb{E}_{S\sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} [R_\mu(\mathbf{A}_{S_n}(S), K+1) - \inf_{h\in\mathcal{H}_S} R_\mu(h, K+1)] \le \frac{C_\theta}{\sqrt{n^{1-\theta}}}. \tag{50}$$

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

$$\mathbb{E}_{S\sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} \sup_{h\in\mathcal{H}_S} R^{\mathrm{in}}_D(h) \le \frac{C_\theta}{\sqrt{n^{1-\theta}}}, \tag{51}$$

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

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

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

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

Therefore,

$$\mathbb{E}_{S\sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} \inf_{h\in\mathcal{H}_S} R_\mu(h, K+1) \geq \mu(\mathbf{x}\in\mathcal{X} : g(\mathbf{x}) < K+1) - \frac{C_\theta b}{2\sqrt{n^{1-\theta}}}. \tag{52}$$

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

$$\inf_{h\in\mathcal{H}_S} R_\mu(h, K+1) \leq \mu(\mathbf{x}\in\mathcal{X} : g(\mathbf{x}) < K+1). \tag{53}$$

Combining inequalities (52) and (53), we obtain that

$$\big|\mathbb{E}_{S\sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} \inf_{h\in\mathcal{H}_S} R_\mu(h, K+1) - \mu(\mathbf{x}\in\mathcal{X} : g(\mathbf{x}) < K+1)\big| \leq \frac{C_\theta b}{2\sqrt{n^{1-\theta}}}. \tag{54}$$

Using inequalities (50) and (54), we obtain that

$$\mathbb{E}_{S_n\sim\mu^n} \sup_{D_{X_{\mathrm{I}}Y_{\mathrm{I}}}\in\mathscr{D}_{\mathrm{I}}} \big[\mathbb{E}_{S\sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} R_\mu(\mathbf{A}_{S_n}(S), K+1) - \mu(\mathbf{x}\in\mathcal{X} : g(\mathbf{x}) < K+1)\big] \leq \frac{C_\theta(b+1)}{\sqrt{n^{1-\theta}}}. \tag{55}$$

Using inequality (51), we have

$$\mathbb{E}_{S_n\sim\mu^n} \sup_{D_{X_{\mathrm{I}}Y_{\mathrm{I}}}\in\mathscr{D}_{\mathrm{I}}} \mathbb{E}_{S\sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} R^{\mathrm{in}}_D(\mathbf{A}_{S_n}(S)) \leq \frac{C_\theta}{\sqrt{n^{1-\theta}}}, \tag{56}$$

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

$$\begin{aligned}\mathbb{E}_{S_n\sim\mu^n} \sup_{D_{X_{\mathrm{I}}Y_{\mathrm{I}}}\in\mathscr{D}_{\mathrm{I}}} \mathbb{E}_{S\sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} \big[&- \int_{g<K+1} \ell(\mathbf{A}_{S_n}(S)(\mathbf{x}), K+1)\mathrm{d}\mu(\mathbf{x}) \\ &+ \mu(\mathbf{x}\in\mathcal{X} : g(\mathbf{x}) < K+1)\big] \leq \frac{2bC_\theta}{\sqrt{n^{1-\theta}}}.\end{aligned} \tag{57}$$

Combining inequalities (55) and (57), we have

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

Therefore, there exist data points S'_n such that

$$\begin{aligned}&\sup_{D_{X_{\mathrm{I}}Y_{\mathrm{I}}}\in\mathscr{D}_{\mathrm{I}}} \mathbb{E}_{S\sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} R^{\mathrm{out}}_D(\mathbf{A}_{S'_n}) \\ =&\sup_{D_{X_{\mathrm{I}}Y_{\mathrm{I}}}\in\mathscr{D}_{\mathrm{I}}} \mathbb{E}_{S\sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} \int_{g=K+1} \ell(\mathbf{A}_{S'_n}(S)(\mathbf{x}), K+1) f_{\mathrm{O}}(\mathbf{x})\mathrm{d}\mu(\mathbf{x}) \\ \leq& 2b \sup_{D_{X_{\mathrm{I}}Y_{\mathrm{I}}}\in\mathscr{D}_{\mathrm{I}}} \mathbb{E}_{S\sim D^n_{X_{\mathrm{I}}Y_{\mathrm{I}}}} \int_{g=K+1} \ell(\mathbf{A}_{S'_n}(S)(\mathbf{x}), K+1)\mathrm{d}\mu(\mathbf{x}) \leq \frac{4b^2C_\theta}{\sqrt{n^{1-\theta}}} + \frac{2C_\theta(b^2+b)}{\sqrt{n^{1-\theta}}}.\end{aligned} \tag{58}$$

Combining inequalities (51) and (58), we obtain that for any n, there exists data points S'_n such that

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

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 $\alpha \in [0,1]$

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

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 $y_1, y_2 \in \mathcal{Y}_{\text{all}}$,

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

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 Risk-based Realizability Assumption holds with zero-one loss if and only if Risk-based 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 $\alpha \in [0,1]$,

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

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

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

We have completed this proof. ■

Appendix L. Proof of Theorem 15

Theorem 15 *Suppose that $\mathcal{R}$ is constant closure, separate, and $\mu(\mathcal{X}) < +\infty$. Given a density-based space $\mathscr{D}_{XY}^{\mu,b}$, if the AUC-based Realizability Assumption holds, then when $\mathrm{VC}[\phi \circ \mathcal{R}] < +\infty$, OOD detection is learnable under AUC in $\mathscr{D}_{XY}^{\mu,b}$ for $\mathcal{R}$, where $\phi \circ \mathcal{R} = \{\mathbf{1}_{r_1(\mathbf{x})>r_2(\mathbf{x}')} : r_1, r_2 \in \mathcal{R}\}$. Furthermore, the learning rate $\epsilon_{\text{cons}}(n)$ can attain $O(1/\sqrt{n^{1-\theta}})$, for any $\theta \in (0,1)$.*

Proof [Proof of Theorem 15] Without loss of generality, we assume that $K = 1$, and any $r \in \mathcal{R}$ satisfies that $0 < r < 1$ (one can achieve this by using sigmoid function). Then it is clear that $\mathcal{R}_{(0,1)} = \{\mathbf{1}_{r(\mathbf{x})\leq\tau} : \forall r \in \mathcal{R}, \forall\tau \in (0,1)\}$ has finite VC-dimension by the condition constant closure and $\mathrm{VC}[\phi \circ \mathcal{F}] < +\infty$. Given m data points $S_m = \{\mathbf{x}'_1, ..., \mathbf{x}'_m\} \subset \mathcal{X}^m$. We consider the following learning rule:

$$\max_{r\in\mathcal{R},\tau\in(0,1)} \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}_{r(\mathbf{x}'_i)\leq\tau}, \text{ subject to } \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}_{r(\mathbf{x}_j)\leq\tau} = 0.$$

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.*,

$$\mathcal{S} := \{\mathbf{S} := (S_1, S_2, ..., S_m, ...) : S_m \text{ are any } m \text{ data points, } m = 1, ..., +\infty\}. \tag{59}$$

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

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

Then we can check that

$$\mathbb{E}_{S\sim D_{X_{\mathrm{I}}}^n} \sup_{\mathbf{1}_{r\le\tau}\in\mathcal{G}_S} \mathbb{E}_{\mathbf{x}\sim D_{X_{\mathrm{I}}}} \mathbf{1}_{r(\mathbf{x})\le\tau} \le \inf_{\mathbf{1}_{r\le\tau}\in\mathcal{R}_{(0,1)}} \mathbb{E}_{\mathbf{x}\sim D_{X_{\mathrm{I}}}} \mathbf{1}_{r(\mathbf{x})\le\tau} + \frac{C_\theta}{\sqrt{n^{1-\theta}}},$$

and because of $\mathcal{G}_S \subset \mathcal{R}_{(0,1)}$,

$$\mathbb{E}_{S_n\sim\mu^n} \sup_{S\in\mathcal{X}^n} [\sup_{\mathbf{1}_{r\le\tau}\in\mathcal{G}_S} R_\mu(\mathbf{1}_{r\le\tau}) - R_\mu(\mathbf{A}_{S_n}(S))] \le \frac{C_\theta}{\sqrt{n^{1-\theta}}}, \tag{60}$$

where

$$\mathcal{G}_S = \{\mathbf{1}_{r(\mathbf{x})\le\tau} \in \mathcal{R}_{(0,1)} : \sum_{i=1}^n \mathbf{1}_{r(\mathbf{x}_j)\le\tau} = 0 < 0\}, \text{ here } S = \{\mathbf{x}_1, ..., \mathbf{x}_n\},$$

and

$$R_\mu(\mathbf{1}_{r\le\tau}) = \mathbb{E}_{\mathbf{x}\sim\mu}\mathbf{1}_{r(\mathbf{x})\le\tau}.$$

Let $\mathcal{D}_I$ be the set consisting of all ID distribution in the density-based space. Then we have

$$\mathbb{E}_{S_n\sim\mu^n} \sup_{D_{X_{\mathrm{I}}}\in\mathcal{D}_I} \mathbb{E}_{S\sim D_{X_{\mathrm{I}}}^n} [\sup_{\mathbf{1}_{r\le\tau}\in\mathcal{G}_S} R_\mu(\mathbf{1}_{r\le\tau}) - R_\mu(\mathbf{A}_{S_n}(S))] \le \frac{C_\theta}{\sqrt{n^{1-\theta}}}, \tag{61}$$

Due to AUC-based Realizability Assumption, we obtain that $\inf_{\mathbf{1}_{r\le\tau}\in\mathcal{R}_{(0,1)}} \mathbb{E}_{\mathbf{x}\sim D_{X_{\mathrm{I}}}} \mathbf{1}_{r(\mathbf{x})\le\tau} = 0$, therefore,

$$\mathbb{E}_{S\sim D_{X_{\mathrm{I}}}^n} \sup_{\mathbf{1}_{r\le\tau}\in\mathcal{G}_S} \mathbb{E}_{\mathbf{x}\sim D_{X_{\mathrm{I}}}} \mathbf{1}_{r(\mathbf{x})\le\tau} \le \frac{C_\theta}{\sqrt{n^{1-\theta}}}. \tag{62}$$

Let $r^* \in \mathcal{R}$ be the optimal ranking function satisfying that

$$\mathrm{AUC}(r^*; D_{X_{\mathrm{I}}}, D_{X_{\mathrm{O}}}) = 1,$$

which implies that there exists τ^* satisfying that for any $\epsilon > 0$

$$D_{X_{\mathrm{I}}}(\mathbf{x} : r^*(\mathbf{x}) < \tau^* - \epsilon) = 0, \quad D_{X_{\mathrm{I}}}(\mathbf{x} : r^*(\mathbf{x}) < \tau^* + \epsilon) > 0.$$

Then we consider set $\Omega_{X_{\mathrm{I}}} := \{\mathbf{x} \in \mathcal{X} : r^*(\mathbf{x}) > \tau^*\}$, if $D_{X_{\mathrm{I}}}(\mathbf{x} : r^*(\mathbf{x}) = \tau^*) = 0$; otherwise, $\Omega_{X_{\mathrm{I}}} := \{\mathbf{x} \in \mathcal{X} : r^*(\mathbf{x}) \ge \tau^*\}$. $\Omega_{X_{\mathrm{O}}} := \mathcal{X} - \Omega_{X_{\mathrm{I}}}$. Then we have that

$$\frac{C_\theta}{\sqrt{n^{1-\theta}}} \ge \mathbb{E}_{S\sim D_{X_{\mathrm{I}}}^n} \sup_{\mathbf{1}_{r\le\tau}\in\mathcal{G}_S} \mathbb{E}_{\mathbf{x}\sim D_{X_{\mathrm{I}}}} \mathbf{1}_{r(\mathbf{x})\le\tau} \ge \frac{2}{b}\mathbb{E}_{S\sim D_{X_{\mathrm{I}}}^n} \sup_{\mathbf{1}_{r\le\tau}\in\mathcal{G}_S} \int_{\Omega_{X_{\mathrm{I}}}} \mathbf{1}_{r(\mathbf{x})\le\tau} \mathrm{d}\mu(\mathbf{x}),$$

which implies that

$$\mathbb{E}_{S\sim D_{X_{\mathrm{I}}}^n} \inf_{\mathbf{1}_{r\le\tau}\in\mathcal{G}_S} \int_{\Omega_{X_{\mathrm{I}}}} \mathbf{1}_{r(\mathbf{x})>\tau} \mathrm{d}\mu(\mathbf{x}) + \frac{bC_\theta}{2\sqrt{n^{1-\theta}}} \ge \mu(\Omega_{X_{\mathrm{I}}}).$$

Therefore,

$$\mathbb{E}_{S\sim D_{X_{\mathrm{I}}}^n} \inf_{\mathbf{1}_{r\le\tau}\in\mathcal{G}_S} \int_{\mathcal{X}} 1 - \mathbf{1}_{r(\mathbf{x})\le\tau} \mathrm{d}\mu(\mathbf{x}) + \frac{bC_\theta}{2\sqrt{n^{1-\theta}}} \ge \mu(\Omega_{X_{\mathrm{I}}}),$$

which implies that

$$\mu(\Omega_{X_{\mathrm{O}}}) + \frac{bC_\theta}{2\sqrt{n^{1-\theta}}} \ge \mathbb{E}_{S\sim D_{X_{\mathrm{I}}}^n} \sup_{\mathbf{1}_{r\le\tau}\in\mathcal{G}_S} R_\mu(\mathbf{1}_{r(\mathbf{x})\le\tau}). \tag{63}$$

Additionally, due to $\mathbf{1}_{r^*(\mathbf{x})\le\tau-\epsilon} \in \mathcal{G}_S$, it is clear that

$$\sup_{\mathbf{1}_{r\le\tau}\in\mathcal{G}_S} R_\mu(\mathbf{1}_{r\le\tau}) \ge \mu(\Omega_{X_{\mathrm{O}}}). \tag{64}$$

Combining Eq. (63) with Eq. (64), we have

$$|\mathbb{E}_{S\sim D_{X_{\mathrm{I}}}^n} \sup_{\mathbf{1}_{r\le\tau}\in\mathcal{G}_S} R_\mu(\mathbf{1}_{r\le\tau}) - \mu(\Omega_{X_{\mathrm{O}}})| \le \frac{bC_\theta}{2\sqrt{n^{1-\theta}}}.$$

By using Eq. (61) and Eq. (64), we have

$$\mathbb{E}_{S_n\sim\mu^n} \sup_{D_{X_{\mathrm{I}}}\in\mathcal{D}_I} \mathbb{E}_{S\sim D_{X_{\mathrm{I}}}^n}[\mu(\Omega_{X_{\mathrm{O}}}) - R_\mu(\mathbf{A}_{S_n}(S))] \le \frac{C_\theta}{\sqrt{n^{1-\theta}}} + \frac{bC_\theta}{2\sqrt{n^{1-\theta}}}. \tag{65}$$

Using inequality (62), we have

$$\mathbb{E}_{S_n\sim\mu^n} \sup_{D_{X_{\mathrm{I}}}\in\mathcal{D}_I} \mathbb{E}_{S\sim D_{X_{\mathrm{I}}}^n} \mathbb{E}_{\mathbf{x}\sim D_{X_{\mathrm{I}}}} \mathbf{A}_{S_n}(S)(\mathbf{x}) \le \frac{C_\theta}{\sqrt{n^{1-\theta}}}, \tag{66}$$

which implies that

$$\mathbb{E}_{S_n\sim\mu^n} \sup_{D_{X_{\mathrm{I}}}\in\mathcal{D}_I} \mathbb{E}_{S\sim D_{X_{\mathrm{I}}}^n} \int_{\Omega_{X_{\mathrm{I}}}} \mathbf{A}_{S_n}(S)(\mathbf{x}) \mathrm{d}\mu(\mathbf{x}) \le \frac{bC_\theta}{2\sqrt{n^{1-\theta}}}. \tag{67}$$

Then inequalities (65) and (67) imply that

$$\mathbb{E}_{S_n\sim\mu^n} \sup_{D_{X_{\mathrm{I}}}\in\mathcal{D}_I} \mathbb{E}_{S\sim D_{X_{\mathrm{I}}}^n} \int_{\Omega_{X_{\mathrm{O}}}} 1 - \mathbf{A}_{S_n}(S)(\mathbf{x}) \mathrm{d}\mu(\mathbf{x}) \le \frac{(1+b)C_\theta}{\sqrt{n^{1-\theta}}}.$$

Therefore,

$$\mathbb{E}_{S_n\sim\mu^n} \sup_{D_{X_{\mathrm{I}}}\in\mathcal{D}_I} \mathbb{E}_{S\sim D_{X_{\mathrm{I}}}^n} \int_{\mathcal{X}} 1 - \mathbf{A}_{S_n}(S)(\mathbf{x}) \mathrm{d}D_{X_{\mathrm{O}}}(\mathbf{x}) \le \frac{2b(1+b)C_\theta}{\sqrt{n^{1-\theta}}}.$$

We assume that $\mathbf{A}_{S_n}(S) = \mathbf{1}_{r_{S_n,S}\le\tau_{S_n,S}}$. Then above inequality implies that

$$\mathbb{E}_{S_n\sim\mu^n} \inf_{D_{X_{\mathrm{I}}}\in\mathcal{D}_I} \mathbb{E}_{S\sim D_{X_{\mathrm{I}}}^n} \mathbb{E}_{\mathbf{x}\sim D_{X_{\mathrm{O}}}} \mathbf{1}_{r_{S_n,S}(\mathbf{x})\le\tau_{S_n,S}} \ge 1 - \frac{2b(1+b)C_\theta}{\sqrt{n^{1-\theta}}}.$$

Inequality (66) implies that

$$\mathbb{E}_{S_n\sim\mu^n}\inf_{D_{X_{\mathrm{I}}}\in\mathcal{D}_I}\mathbb{E}_{S\sim D_{X_{\mathrm{I}}}^n}\mathbb{E}_{\mathbf{x}\sim D_{X_{\mathrm{I}}}}\mathbf{1}_{r_{S_n,S}(\mathbf{x})>\tau_{S_n,S}}\geq 1-\frac{C_\theta}{\sqrt{n^{1-\theta}}},$$

which shows that there exists $S_n'\sim\mu^n$ and C' such that

$$\begin{aligned}
&\inf_{D_{X_{\mathrm{I}}}\in\mathcal{D}_I}\mathbb{E}_{S\sim D_{X_{\mathrm{I}}}^n}\mathrm{AUC}(f_{S_n',S};D_{X_{\mathrm{I}}},D_{X_{\mathrm{O}}})\\
\geq&\inf_{D_{X_{\mathrm{I}}}\in\mathcal{D}_I}\mathbb{E}_{S\sim D_{X_{\mathrm{I}}}^n}\mathbb{E}_{\mathbf{x}\sim D_{X_{\mathrm{O}}}}\mathbb{E}_{\mathbf{x}'\sim D_{X_{\mathrm{I}}}}\mathbf{1}_{r_{S_n',S}(\mathbf{x})<r_{S_n',S}(\mathbf{x}')}\\
\geq&\inf_{D_{X_{\mathrm{I}}}\in\mathcal{D}_I}\mathbb{E}_{S\sim D_{X_{\mathrm{I}}}^n}\mathbb{E}_{\mathbf{x}\sim D_{X_{\mathrm{O}}}}\mathbb{E}_{\mathbf{x}'\sim D_{X_{\mathrm{I}}}}\mathbf{1}_{r_{S_n',S}(\mathbf{x})\leq\tau_{S_n',S}}\mathbf{1}_{r_{S_n',S}(\mathbf{x}')>\tau_{S_n',S}}\\
\geq&1-\frac{\max\{C_\theta,C'\}}{\sqrt{n^{1-\theta}}}-\frac{2b(1+b)C_\theta}{\sqrt{n^{1-\theta}}}.
\end{aligned}$$

We set data point sequences $\mathbf{S}'=(S_1',S_2',...,S_n',...)$. Then, the function part $r_{S_n',S}$ of $\mathbf{A}_{\mathbf{S}'}\in\mathscr{A}$ is the universally consistent algorithm, *i.e.*,

$$\inf_{D_{X_{\mathrm{I}}}\in\mathcal{D}_I}\mathbb{E}_{S\sim D_{X_{\mathrm{I}}}^n}\mathrm{AUC}(r_{S_n',S};D_{X_{\mathrm{I}}},D_{X_{\mathrm{O}}})\geq 1-\frac{\max\{C_\theta,C'\}}{\sqrt{n^{1-\theta}}}-\frac{2b(1+b)C_\theta}{\sqrt{n^{1-\theta}}}.$$

We have completed this proof. ■

Appendix M. Proofs of Proposition 1 and Proposition 2

Proposition 1 *Let $\mathcal{X}$ be a bounded feature space. Given $\mathbf{q}=(l_1,...,l_{g-1},1)$, then*

- *if some s with $1<s<g$, $d=l_1\leq l_2\leq...\leq l_s$, and $l_s\geq 2d$, $\mathcal{F}_{\mathbf{q}}^{\sigma}$ is the separate ranking function space;*
- *$\mathcal{F}_{\mathbf{q}}^{\sigma}$ is constant closure;*
- *$\{\mathbf{1}_{f_1(\mathbf{x})<f_2(\mathbf{x}')}:f_1,f_2\in\mathcal{F}_{\mathbf{q}}^{\sigma}\}$ has finite VC dimension.*

Proof [Proof of Proposition 1] **Firstly**, we prove that if some s with $1<s<g$, $d=l_1\leq l_2\leq...\leq l_s$, and $l_s\geq 2d$, $\mathcal{F}_{\mathbf{q}}^{\sigma}$ is the separate ranking function space.

First, we show that if $\mathcal{X}\subset\mathbb{R}^1$, and $\mathbf{q}=(1,2,1)$, then $\mathcal{F}_{\mathbf{q}}^{\sigma}$ is the separate ranking space. For any $x\in\mathcal{X}$,

$$f_x(x')=[1,1]\,\sigma(\begin{bmatrix}1\\-1\end{bmatrix}x'+\begin{bmatrix}-x\\x\end{bmatrix})+0.$$

It is easy to check that for any $x'\neq x$,

$$f_x(x)<f_x(x').$$

Next, we prove that if $\mathcal{X} \subset \mathbb{R}^d$, and $\mathbf{q} = (d, 2d, 1)$, then $\mathcal{F}_{\mathbf{q}}^{\sigma}$ is the separate ranking space. Let $\mathbf{v}_i \in \mathbb{R}^{2d\times 1}$ is a vector whose $2i$-th and $2i+1$-th coordinates are -1; otherwise, other coordinates are 0. For any $\mathbf{x} = [x_1, ..., x_d]^\top \in \mathcal{X}$,

$$f_{\mathbf{x}}(\mathbf{x}') = [1, 1..., 1]_{1\times(2d)}\, \sigma([\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_d]\, \mathbf{x}' + \mathbf{M}_{\mathbf{d+1}}) + 0,$$

where $\mathbf{M}_{d+1} \in \mathbb{R}^{2d\times 1}$ is a vector whose $2i$-th coordinate is $-x_i$ and $(2i+1)$-th coordinate is x_i. It is easy to check that for any $\mathbf{x}' \neq \mathbf{x}$,

$$f_{\mathbf{x}}(\mathbf{x}) < f_{\mathbf{x}}(\mathbf{x}').$$

Thirdly, we prove that if $d = l_1 = l_2 = ... = l_{r-1}$ and $l_r = 2d$, then $\mathcal{F}_{\mathbf{q}}^{\sigma}$ is the separate ranking space. Due to $\mathcal{X}$ is bounded, we can find b such that any $\mathbf{x} = [x_1, ..., x_d]^\top \in \mathcal{X}$ satisfies that $x_i + b \geq 0$. Then, we set $\mathbf{w}_2, \mathbf{w}_3, ..., \mathbf{w}_{r-1}$ are identity matrices, $\mathbf{b}_2$ is the matrix whose all coordinates are b, and $\mathbf{b}_3, ..., \mathbf{b}_{r-1}$ are $\mathbf{0}$. Then the result in second step implies the result. Finally, using the result that $\mathbf{q} \lesssim \mathbf{q}' \Rightarrow \mathcal{F}_{\mathbf{q}}^{\sigma} \subset \mathcal{F}_{\mathbf{q}'}^{\sigma}$ (Lemma 14) implies the final result.

Secondly, it is clear that $\mathcal{F}_{\mathbf{q}}^{\sigma}$ is constant closure. We omit the proof.
Thirdly, by Theorems 5 and 8 in (Bartlett and Maass, 2003), we can obtain the third result. ■

Proposition 2 *Given* $\mathbf{q} = (l_1, ..., l_{g-1}, l)$ *and* $\mathbf{q}' = (l_1, ..., l_{g-1}, 1)$*, let* $\mathcal{R} = E \circ \mathcal{F}_{\mathbf{q}}^{\sigma}$*, then*

- *if* $\mathcal{F}_{\mathbf{q}'}^{\sigma}$ *is a separate ranking function space,* $\mathcal{R}$ *is the separate ranking function space;*
- $\mathcal{R}$ *is constant closure;*
- $\{\mathbf{1}_{r_1(\mathbf{x})<r_2(\mathbf{x}')}, r_1, r_2 \in \mathcal{R}\}$ *has finite VC dimension,*

where E *is Eq.* (7)*,* (8) *or* (9)*.*

Proof [Proof of Proposition 2] **Firstly**, we prove that if $\mathcal{F}_{\mathbf{q}'}^{\sigma}$ is a separate ranking function space, $\mathcal{R}$ is the separate ranking function space.

For the softmax-based function:

$$E(\mathbf{f}) = \max_{k\in\{1,...,l\}} \frac{\exp(f^k)}{\sum_{c=1}^{l} \exp(f^c)}.$$

Note that $\mathcal{F}_{\mathbf{q}'}^{\sigma}$ is separate ranking space. Then for any $\mathbf{x} \in \mathcal{X}$, there exists $f_{\mathbf{x}} \in \mathcal{F}_{\mathbf{q}'}^{\sigma}$ such that $0 = f_{\mathbf{x}}(\mathbf{x}) < f_{\mathbf{x}}(\mathbf{x}')$, for any $\mathbf{x}' \in \mathcal{X}$ and $\mathbf{x}' \neq \mathbf{x}$. Then $\mathbf{f}_{\mathbf{x}} = [f_{\mathbf{x}}, -f_{\mathbf{x}}, .., -f_{\mathbf{x}}] \in \mathcal{F}_{\mathbf{q}}^{\sigma}$ can ensure that for any $\mathbf{x}' \in \mathcal{X}$ and $\mathbf{x}' \neq \mathbf{x}$,

$$E(\mathbf{f}_{\mathbf{x}}(\mathbf{x})) < E(\mathbf{f}_{\mathbf{x}}(\mathbf{x}')).$$

Using the same strategy, we can prove that the temperature-scaled function is the separate ranking space. For the energy-based function:

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

Note that $\mathcal{F}_{\mathbf{q}'}^{\sigma}$ is separate ranking space. Then for any $\mathbf{x} \in \mathcal{X}$, there exists $f_{\mathbf{x}} \in \mathcal{F}_{\mathbf{q}'}^{\sigma}$ such that $0 = f_{\mathbf{x}}(\mathbf{x}) < f_{\mathbf{x}}(\mathbf{x}')$, for any $\mathbf{x}' \in \mathcal{X}$ and $\mathbf{x}' \neq \mathbf{x}$. Then $\mathbf{f}_{\mathbf{x}} = [f_{\mathbf{x}}, f_{\mathbf{x}}, .., f_{\mathbf{x}}] \in \mathcal{F}_{\mathbf{q}}^{\sigma}$ can ensure that for any $\mathbf{x}' \in \mathcal{X}$ and $\mathbf{x}' \neq \mathbf{x}$,

$$E(\mathbf{f}_{\mathbf{x}}(\mathbf{x})) < E(\mathbf{f}_{\mathbf{x}}(\mathbf{x}')).$$

Secondly, it is easy to show that R is constant closure. We omit it.
Thirdly, by Theorems 5 and 8 in (Bartlett and Maass, 2003), we can obtain the third result.
■

Appendix N. Proofs of Proposition 3 and Proposition 4

To better understand the contents in Appendices N-Q, 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:

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

where $f_{\mathbf{w},\mathbf{b}}^{k}(\mathbf{x})$ 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 3 *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 [Proof of Proposition 3] 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 (Safran and Shamir, 2017), $\mathbf{b}_g$ 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 4 *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 [Proof of Proposition 4] 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, $h_1 \in \mathcal{H}^{\sigma,\lambda}_{\mathbf{q},E}$.

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, $h_2 \in \mathcal{H}^{\sigma,\lambda}_{\mathbf{q},E}$. ■

It is easy to check that when $g > 2$ and activation function σ is not a constant, Proposition 4 still holds, even if $\mathbf{b}_g = \mathbf{0}$.

Appendix O. Proof of Theorem 16

Before proving Theorem 16, we need several lemmas.

Lemma 13 *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 $l_i \le l'_i$ $(i = 1, ..., g-1)$, then $\mathcal{F}^{\sigma}_{\mathbf{q}} \subset \mathcal{F}^{\sigma}_{\mathbf{q}'}$ and $\mathcal{H}^{\sigma}_{\mathbf{q}} \subset \mathcal{H}^{\sigma}_{\mathbf{q}'}$.*

Proof [Proof of Lemma 13] 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 $l'_g - l_g = 0$,

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

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 14 *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 [Proof of Lemma 14] Given $l'' = (l''_1, ..., l''_{g''})$ satisfying that $g \le g''$, $l''_i = l_i$ 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

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

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 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 13 ($\mathbf{q}''$ and $\mathbf{q}$ satisfy the condition in Lemma 13), 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 15 *(Pinkus, 1999) 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*

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

Proof [Proof of Lemma 15] The proof of Lemma 15 can be found in Theorem 3.1 in (Pinkus, 1999). ■

Lemma 16 *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*

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

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

Proof [Proof of Lemma 16] Let $\mathbf{f} = [f_1, ..., f_l]^\top$, where f_i is the i-th coordinate of $\mathbf{f}$. Based on Lemma 15, we obtain l sequences $\mathbf{q}^1$, $\mathbf{q}^2$,...,$\mathbf{q}^l$ such that

$$\inf_{g_1 \in \mathcal{F}^{\sigma}_{\mathbf{q}^1}} \max_{\mathbf{x} \in C} |g_1(\mathbf{x}) - f_1(\mathbf{x})| < \epsilon/\sqrt{l},$$
$$\inf_{g_2 \in \mathcal{F}^{\sigma}_{\mathbf{q}^2}} \max_{\mathbf{x} \in C} |g_2(\mathbf{x}) - f_2(\mathbf{x})| < \epsilon/\sqrt{l},$$
$$...$$
$$...$$
$$\inf_{g_l \in \mathcal{F}^{\sigma}_{\mathbf{q}^l}} \max_{\mathbf{x} \in C} |g_l(\mathbf{x}) - f_l(\mathbf{x})| < \epsilon/\sqrt{l}.$$

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 14, we obtain that $\mathcal{F}^{\sigma}_{\mathbf{q}^i} \subset \mathcal{F}^{\sigma}_{\mathbf{q}}$. Therefore,

$$\inf_{g \in \mathcal{F}^{\sigma}_{\mathbf{q}}} \max_{\mathbf{x} \in C} |g(\mathbf{x}) - f_1(\mathbf{x})| < \epsilon/\sqrt{l},$$
$$\inf_{g \in \mathcal{F}^{\sigma}_{\mathbf{q}}} \max_{\mathbf{x} \in C} |g(\mathbf{x}) - f_2(\mathbf{x})| < \epsilon/\sqrt{l},$$
$$...$$
$$...$$
$$\inf_{g \in \mathcal{F}^{\sigma}_{\mathbf{q}}} \max_{\mathbf{x} \in C} |g(\mathbf{x}) - f_l(\mathbf{x})| < \epsilon/\sqrt{l}.$$

Therefore, for each i, we can find $g_{\mathbf{w}^i,\mathbf{b}^i}$ from $\mathcal{F}^{\sigma}_{\mathbf{q}}$ 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 $\mathbf{g}_{\mathbf{w},\mathbf{b}} = [g_{\mathbf{w}^1,\mathbf{b}^1}, g_{\mathbf{w}^2,\mathbf{b}^2}, ..., g_{\mathbf{w}^l,\mathbf{b}^l}]^\top \in \mathcal{F}^{\sigma}_{\mathbf{q}'}$, 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:

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

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

Lemma 17 *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 [Proof of Lemma 17] 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 $h_{\mathbf{w}',\mathbf{b}'} \in \mathcal{H}^{\sigma}_{\mathbf{q}'}$:

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

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 $\mathbf{x}\in\mathcal{X}$,

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

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 $\mathbf{x}\in\mathcal{X}$,

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

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 $(l_1,\ldots,l_g,2)$. 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 (Bartlett et al., 2019) 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 18 *Let $|\mathcal{X}| < +\infty$ and σ be the ReLU function:* $\max\{x,0\}$. *Given r hypothesis functions $h_1, h_2, \ldots, h_r \in \{h : \mathcal{X}\to\{1,\ldots,l\}\}$, then there exists a sequence $\mathbf{q} = (l_1,\ldots,l_g)$ with $l_1 = d$ and $l_g = l$, such that $h_1,\ldots,h_r \in \mathcal{H}^{\sigma}_{\mathbf{q}}$.*

Proof [Proof of Lemma 18] For each h_i $(i=1,\ldots,r)$, 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 16 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 14, 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 19 *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 [Proof of Lemma 19] 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 16 implies that there exists a sequence $\mathbf{q}^h = (l^h_1, ..., l^h_{g^h})$ ($l^h_1 = d$ and $l^h_{g^h} = l$) and $\mathbf{f}_{\mathbf{w},\mathbf{b}} \in \mathcal{F}^\sigma_{\mathbf{q}^h}$ 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 $\mathbf{f}_{\mathbf{w},\mathbf{b}} \in \mathcal{F}^\sigma_{\mathbf{q}^h}$, E and λ. Since $|\mathcal{X}| < +\infty$, only finite binary classifiers are defined over $\mathcal{X}$. Using Lemma 18,

we can find a sequence $\mathbf{q}$ such that $\mathcal{H}^b_{\text{all}} = \mathcal{H}^{\sigma,\lambda}_{\mathbf{q},E}$, where $\mathcal{H}^b_{\text{all}}$ consists of all binary classifiers. ■

Lemma 20 *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 3 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 $\mathscr{D}^s_{XY}$ for $\mathcal{H}^{\text{in}} \bullet \mathcal{H}^{\text{b}}$, where $\mathcal{H}^{\text{b}} = \mathcal{H}^{\sigma,\lambda}_{\mathbf{q}',E}$ and $\mathcal{H}^{\text{in}} \bullet \mathcal{H}^{\text{b}}$ is defined below Eq. (6).*

Proof [Proof of Lemma 20] Note that we use the ReLU function as the activation function in this lemma. Using Lemma 14, Lemma 19 and Theorem 11, we can prove this result. ■

Theorem 16 *Suppose that Condition 3 holds and the hypothesis space $\mathcal{H}$ is FCNN-based or score-based, i.e., $\mathcal{H} = \mathcal{H}^\sigma_{\mathbf{q}}$ 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. (6), here E is Eq. (7), (8) or (9). Then*

There is a sequence $\mathbf{q} = (l_1, ..., l_g)$ such that OOD detection is learnable under risk 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 under risk in $\mathscr{D}^s_{XY}$ for $\mathcal{H}$.

Proof [Proof of Theorem 16] 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 $\mathscr{D}^s_{XY}$ for $\mathcal{H}^\sigma_{\mathbf{q}}$, for any sequence $\mathbf{q} = (l_1, ..., l_g)$ ($l_1 = d$ and $l_g = K + 1$).

By Lemma 17, Theorems 5 and 8 in (Bartlett and Maass, 2003), we know that $\text{VCdim}(\phi \circ \mathcal{H}^\sigma_{\mathbf{q}}) < +\infty$, where ϕ maps ID data to 1 and maps OOD data to 2. Additionally, Proposition 3 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 $|\mathcal{X}| = +\infty$. Therefore, Theorem 6 implies that OOD detection is not learnable in $\mathscr{D}^s_{XY}$ for $\mathcal{H}^\sigma_{\mathbf{q}}$, 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 $\mathscr{D}^s_{XY}$ for $\mathcal{H}^\sigma_{\mathbf{q}}$.

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 18, there exists a sequence $\mathbf{q}$ such that $\mathcal{H}_{\text{all}} \subset \mathcal{H}^\sigma_{\mathbf{q}}$. Additionally, Lemma 17 implies that there exist $\mathcal{H}^{\text{in}}$ and $\mathcal{H}^{\text{b}}$ such that $\mathcal{H}^\sigma_{\mathbf{q}} \subset \mathcal{H}^{\text{in}} \bullet \mathcal{H}^{\text{b}}$. Since $\mathcal{H}_{\text{all}}$ consists all hypothesis space, $\mathcal{H}_{\text{all}} = \mathcal{H}^\sigma_{\mathbf{q}} = \mathcal{H}^{\text{in}} \bullet \mathcal{H}^{\text{b}}$. Therefore, $\mathcal{H}^{\text{b}}$ contains all binary classifiers from $\mathcal{X}$ to $\{1, 2\}$. Theorem 11 implies that OOD detection is learnable in $\mathscr{D}^s_{XY}$ for $\mathcal{H}^\sigma_{\mathbf{q}}$.

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 14 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 3 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 Eq. (7), (8), or (9).

By Theorems 5 and 8 in (Bartlett and Maass, 2003), we know that $\text{VCdim}(\mathcal{H}^{\sigma,\lambda}_{\mathbf{q},E}) < +\infty$. Additionally, Proposition 4 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 6 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. (7), (8), or Eq. (9).

Based on Lemma 20, 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 score 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 20, 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 $\mathcal{H}^{\text{in}} \bullet \mathcal{H}^{\text{b}}$, where $\mathcal{H}^{\text{b}} = \mathcal{H}^{\sigma,\lambda}_{\mathbf{q}',E}$, where E is in Eq. (7), (8), or Eq. (9).

In the fifth step, we have proven that Eqs. (7), (8), and (9) meet the condition in Lemma 20. Therefore, Lemma 20 implies this result. We complete the proof when the hypothesis space $\mathcal{H}$ is score-based. ■

Appendix P. Proof of Theorem 17

Theorem 17 *Suppose the ranking function space $\mathcal{R}$ is separate, and FCNN-based or score-based, i.e., $\mathcal{R} = \mathcal{F}^{\sigma}_{\mathbf{q}}$ or $\mathcal{R} = E \circ \mathcal{F}^{\sigma}_{\mathbf{q}}$, where E is Eq. (7), (8) or (9). Then*

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

Furthermore, if $|\mathcal{X}| < +\infty$, then there is 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 under AUC in $\mathscr{D}^{s}_{XY}$ for $\mathcal{R}$.

Proof [Proof of Theorem 17] Using a similar strategy of Theorem 16, we can prove this theorem by Theorem 9 and Lemma 11. ■

Appendix Q. Proof of Theorem 18

Theorem 18 *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 Eq. (7), (8), or (9)) 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 $\iff$ Risk-based Realizability Assumption $\iff$ Condition 4

Proof [Proof of Theorem 18]
1) By Lemma 1, we conclude that Learnability in $\mathscr{D}_{XY}^{\mu,b}$ for $\mathcal{H} \Rightarrow$ Condition 1.
2) By Proposition 3 and Proposition 4, 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 $K = 1$, $\inf_{h\in\mathcal{H}} R_D^{\text{in}}(h) = 0$ and $\inf_{h\in\mathcal{H}} R_D^{\text{out}}(h) = 0$, for any $D_{XY} \in \mathscr{D}_{XY}^{\mu,b}$.
By Condition 1, we obtain that $\inf_{h\in\mathcal{H}} R_D(h) = 0$, for any $D_{XY} \in \mathscr{D}_{XY}^{\mu,b}$. Because each domain D_{XY} in $\mathscr{D}_{XY}^{\mu,b}$ is attainable, we conclude that Risk-based Realizability Assumption holds.
We have proven that Condition 1$\Rightarrow$ Risk-based Realizability Assumption.
3) By Theorems 5 and 8 in (Bartlett and Maass, 2003), we know that $\text{VCdim}(\phi \circ \mathcal{H}_{\mathbf{q}}^{\sigma}) < +\infty$ and $\text{VCdim}(\mathcal{H}_{\mathbf{q},E}^{\sigma,\lambda}) < +\infty$. Then, using Theorem 14, we conclude that Risk-based 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
Learnability in $\mathscr{D}_{XY}^{\mu,b}$ for $\mathcal{H}$ $\Leftrightarrow$Condition 1$\Leftrightarrow$ Risk-based Realizability Assumption.
5) By Lemma 2, we conclude that Condition 4$\Rightarrow$Condition 1.
6) **Here we prove that Learnability in $\mathscr{D}_{XY}^{\mu,b}$ for $\mathcal{H}$ $\Rightarrow$Condition 4.** 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 $D_{XY} \in \mathscr{D}_{XY}^{\mu,b}$,

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

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

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

which implies that there exists $S_\epsilon \sim D_{X_{\mathrm{I}}Y_{\mathrm{I}}}^{n_\epsilon}$ such that

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

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 $D_{XY} \in [D'_{XY}]$,

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

which implies that Condition 4 holds. Therefore, Learnability in $\mathscr{D}_{XY}^{\mu,b}$ for $\mathcal{H}$ $\Rightarrow$Condition 4.
7) Note that in 4), 5) and 6), we have proven that
Learnability in $\mathscr{D}_{XY}^{\mu,b}$ for $\mathcal{H}$ $\Rightarrow$Condition 4$\Rightarrow$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 4$\Leftrightarrow$Condition 1.
8) Combining 4) and 7), we have completed the proof.

■

Appendix R. Proof of Theorem 19

Theorem 19 *Suppose that the ranking function space $\mathcal{R}$ is separate and score-based ($\mathcal{R} = E \circ \mathcal{F}_{\mathbf{q}}^{\sigma}$) or FCNN-based ($\mathcal{R} = \mathcal{F}_{\mathbf{q}}^{\sigma}$), where E is Eq. (7), (8) or (9). If $\mu(\mathcal{X}) < +\infty$, then the following three conditions satisfy:*

AUC-based Realizability Assumption $\Rightarrow$ Learnability in $\mathscr{D}_{XY}^{\mu,b}$ for $\mathcal{R}$ $\Rightarrow$ Condition 2

Proof [Proof of Theorem 19] The result can be obtained by Theorems 15 and 3. ■

Appendix S. Proof of Theorem 20

Theorem 20 *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 Eq. (7), (8), or (9)) 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 5), then OOD detection is not learnable under risk in $\mathscr{D}_{XY}$ for $\mathcal{H}$.*

Proof [Proof of Theorem 20] Using Proposition 3 and Proposition 4, we obtain that $\inf_{h\in\mathcal{H}} R_D^{\mathrm{in}}(h) = 0$ and $\inf_{h\in\mathcal{H}} R_D^{\mathrm{out}}(h) = 0$. Then, Theorem 4 implies this result. ■

Note that if we replace the activation function σ (ReLU function) in Theorem 20 with any other activation functions, Theorem 20 still hold.

Appendix T. Proof of Theorem 21

Theorem 21 *Let the separate ranking function space $\mathcal{R}$ be FCNN-based or score-based (where the score function E is Eq. (7), (8), or (9)). Suppose that $D_{XY}, D'_{XY} \in \mathscr{D}_{XY}$ are discrete distributions with $D_{X_\mathrm{I}Y_\mathrm{I}} = D_{X_\mathrm{I}Y_\mathrm{I}}$ and $D_{X_\mathrm{O}} = \delta_{\mathbf{x}}, D'_{X_\mathrm{O}} = \delta_{\mathbf{x}'}$. If $D_{X_\mathrm{O}} = \delta_{\mathbf{x}}, D'_{X_\mathrm{O}} = \delta_{\mathbf{x}'}$ have overlaps with $D_{X_\mathrm{I}Y_\mathrm{I}}$ and $D_{X_\mathrm{O}} \neq D'_{X_\mathrm{O}}$, then OOD detection is not learnable under AUC in $\mathscr{D}_{XY}$ for $\mathcal{R}$.*

Proof [Proof of Theorem 21] This is a conclusion of Lemma 7. ■