A Provable Decision Rule for Out-of-Distribution Detection

Xinsong Ma [1] **Xin Zou** [1] **Weiwei Liu** [1]

Abstract

Out-of-distribution (OOD) detection task plays the key role in reliable and safety-critical applications. Existing researches mainly devote to designing or training the powerful score function but overlook investigating the decision rule based on the proposed score function. Different from previous work, this paper aims to design a decision rule with rigorous theoretical guarantee and well empirical performance. Specifically, we provide a new insight for the OOD detection task from a hypothesis testing perspective and propose a novel generalized Benjamini Hochberg (g-BH) procedure with empirical p-values to solve the testing problem. Theoretically, the g-BH procedure controls false discovery rate (FDR) at pre-specified level. Furthermore, we derive an upper bound of the expectation of false positive rate (FPR) for the g-BH procedure based on the tailed generalized Gaussian distribution family, indicating that the FPR of g-BH procedure converges to zero in probability. Finally, the extensive experimental results verify the superiority of g-BH procedure over the traditional threshold-based decision rule on several OOD detection benchmarks.

1. Introduction

Deep Neural Networks (DNNs) have attained remarkable achievements in a broad range of challenging problems from image classification (He et al., 2016b), speech recognition (Amodei et al., 2016), to machine translation (Dankers et al., 2022). The excellent performance of these models lie in the promising generalization on the in-distribution (ID) inputs that are drawn from the same distribution as the examples used to train the model. Nevertheless, in the open real-world scenarios, these models often struggle with out-of-distribution (OOD) inputs from a different distribution that the network has not been exposed to during training. To ensure the reliability and safety of sensitive applications, such as medical diagnosis (Frolova et al., 2022) and finance (Özbayoglu et al., 2020), the OOD inputs should be identified and not be predicted with high confidence during testing time (Nguyen et al., 2015). Such a task is referred to as OOD detection (Hendrycks & Gimpel, 2017).

OOD detection has gained significant attention recently and a plethora of literature has emerged to address this issue (Liu et al., 2020; Yang et al., 2021; Wang et al., 2022; Hendrycks et al., 2022; Djurisic et al., 2023; Liu et al., 2023). In the current literature, the OOD detection task is formulated as the following decision problem. For an input x, the decision rule is:

$$\phi(x) = \begin{cases} ID, & \text{if } s(x) \geq s^* \\ OOD, & \text{if } s(x) < s^* \end{cases} \tag{1}$$

where $s(\cdot)$ denotes the score function and the threshold s^* is empirically selected so that the ture positive rate (TPR) on ID validation set [1] is 95% before testing. Given the choice of threshold s^*, we call the decision rule in Eq. (1) empirical decision rule (e-DR). Existing OOD detection methods mainly devote to obtaining a powerful score function and then apply the e-DR in Eq. (1) to identify the OOD examples directly. Factually, the decision rule is extremely significant for the OOD detection task, since it may be directly used in sensitive applications such as self-driving (Huang et al., 2020; Li et al., 2022). However, these *state-of-the-art* approaches are lack of the systematic investigation into the decision rule. Besides, the e-DR is empirical, thus its outputs are not covered by any rigorous theoretical guarantee. Then a natural question arise:

> *How to design an decision rule with rigorous theoretical guarantee and superior empirical performance?*

This paper aims to systematically study the above question. Different from previous OOD detection literature, we investigate the OOD detection problem from the statistical

[1]School of Computer Science, National Engineering Research Center for Multimedia Software, Institute of Artificial Intelligence and Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University, Wuhan, China. Correspondence to: Weiwei Liu <liuweiwei863@gmail.com>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

[1]In practical application, the testing data is unknown and we can not guarantee that the TPR on test set is 95%. Therefore, the threshold is selected using the ID validation set.

perspective. Specifically, we consider the OOD detection task as a multiple hypothesis testing problem. It is known that Benjamini Hochberg (BH) procedure is one of the most popular and widely used algorithms for multiple hypothesis testing (Benjamini & Hochberg, 1995). To control the false discovery rate (FDR) [2], the BH procedure demands that the p-values corresponding to different null hypotheses are mutually independent or follow certain patterns of dependence, such as the *positive regression dependence on subset* (PRDS) (Benjamini & Yekutieli, 2001) or the *dependency control* (DC) condition (Blanchard & Roquain, 2008). Following these literature, we propose a novel generalized BH (g-BH) procedure with empirical p-values to solve the OOD detection problem, which can control FDR at prespecified level. Moreover, we derive an upper bound of the false positive rate (FPR) expectation for the g-BH procedure based on the tailed generalized Gaussian distribution family. This upper bound indicates that the FPR of g-BH procedure converges to 0 in probability.

Extensive experiments demonstrate the superiority of the g-BH procedure over the traditional threshold-based decision rule from practical perspective (focusing on TPR, FPR and F1 -score) and classical perspective (focusing on FPR95, AUROC and AUPR). For example, using CIFAR-10 as ID and Place365 as OOD, our method reduces the FPR from 48.21% to 16.79%, and improves the F1-score from 49.83% to 67.86% compared with the e-DR based on KNN (Sun et al., 2022), a direct improvement of **31.43%** and **18.03%**. In addition, combining the MSP (Hendrycks & Gimpel, 2017) with the g-BH procedure, the FPR95 is reduced by **13.65%** on average compared with the vanilla MSP.

Overall, the g-BH procedure achieves promising performance on OOD detection and it is easily adopted based on the existing score functions, without any sophisticated changes to the loss or training scheme. We summarize our contributions below:

(1) We frame the OOD detection task as the multiple hypothesis testing problem and propose a novel g-BH procedure to solve it. Our method is distribution-free, easy to implement, and does not rely on the extra information of OOD data. Besides, any score-based methods can be plugged in the g-BH procedure.

(2) We develop the theoretical results of classical BH procedure about FDR control. Besides, we derive an upper bound of the FPR expectation for the g-BH procedure based on the tailed generalized Gaussian distribution family. This indicates that the FPR of g-BH procedure converges to zero in probability.

(3) Finally, we conduct extensive experiments to demonstrate the superiority of the g-BH procedure over traditional threshold-based decision rule on several OOD detection benchmarks. The results show that our method improves the OOD detection performance of the existing methods.

[2]FDR is related to the concept of *Precision* (see Section 2.2), and can be considered as the generalization of type-I error for single hypothesis testing. Therefore, the FDR should be first controlled for a hypothesis testing algorithm.

The rest of this paper is organized as follows. Section 2 introduces the background of OOD detection and related concepts. Section 3 proposes the g-BH procedure and proves that it controls FDR at a prescribed level. Next, Section 4 derives an upper bound of the FPR expectation for the g-BH procedure under the tailed generalized Gaussian distribution family. The experimental results are presented at Section 5. We discuss the related works in Appendix A. Besides, the omited proofs and in main context are presented in Appendices B.

2. Preliminaries

2.1. Background

We denote by $\mathcal{X} \subseteq \mathbb{R}^d$ the feature space and $\mathcal{Y} = \{1, 2, 3, \ldots, L\}$ the label space with the joint distribution $\mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$, and $\mathcal{X}$ has marginal distribution $\mathcal{P}_{in}$. Let (x, y) be the feature-label pair, where instance $x \in \mathcal{X}$ and label $y \in \mathcal{Y}$. Let $f(\theta; x) : \mathcal{X} \to \mathbb{R}^{|\mathcal{Y}|}$ be a neural network, which is parameterized by θ and outputs a logit vector used to predict the label of an input sample. For simplicity, we denote $\ell(\theta, x, y) = \ell(f_\theta(x), y)$ where $\ell(\cdot)$ is a loss function. We attain a perform-well multi-class classifier by minimizing the following risk:

$$R(f) = \mathbb{E}_{(x,y)\sim\mathcal{P}_{\mathcal{X}\times\mathcal{Y}}}[\ell(\theta, x, y)].$$

In practice, we generally use cross-entropy loss with the softmax activation function:

$$\ell(\theta, x, y) = -\log p(y|x) = -\log \frac{e^{f_y(\theta;x)}}{\sum_{j=1}^{L} e^{f_j(\theta;x)}},$$

where $p(y|x)$ is the softmax probability and $f_j(\theta, x)$ denotes the j-th element of $f(\theta; , x)$ corresponding to the ground-truth label j.

During the prediction phase, we usually assume that the test data are drawn from the same distribution $\mathcal{P}_{in}$ as the training data. Nevertheless, practical situations may introduce inputs from unfamiliar distributions, with label space potentially lacking any intersection with $\mathcal{Y}$. These inputs are referred to as OOD data and should not be predicted. The goal of OOD detection is to identify the OOD examples in testing set. Existing literature for OOD detection mainly devotes to obtaining a powerful score function and directly applies the e-DR in (1) to determine whether a input is ID or OOD.

2.2. A Perspective of hypothesis testing for OOD Detection

Different from existing OOD detection methods which directly apply e-DR in Eq. (1), we aim to design a new decision rule with rigorous theoretical guarantee and superior empirical performance. Firstly, we provide a new insight for the OOD detection problem from a hypothesis testing perspective. Specifically, for a testing set $\mathcal{T}^{test} = \{X_1^{test}, X_2^{test}, \ldots, X_n^{test}\}$, OOD detection task can be formulated as the following hypothesis testing problem:

$$\begin{aligned} H_{1;0} &: X_1^{test} \sim \mathcal{D}_{in}, \quad H_{1;1} : X_1^{test} \nsim \mathcal{D}_{in} \\ H_{2;0} &: X_2^{test} \sim \mathcal{D}_{in}, \quad H_{2;1} : X_2^{test} \nsim \mathcal{D}_{in} \\ &\vdots \\ H_{n;0} &: X_n^{test} \sim \mathcal{D}_{in} \quad H_{n;1} : X_n^{test} \nsim \mathcal{D}_{in} \end{aligned} \tag{2}$$

where $H_{i;0}$ and $H_{i;1}$ are called null hypothesis and alternative hypothesis, respectively. If $H_{i;0}$ is rejected, we declare that X_i^{test} is OOD.

Whether we reject null hypothesis $H_{i;0}$ or not depends on the significant concepts: *p-value*, defined as follows.

Definition 2.1 (p-value (Casella & Berger, 2002)). Given a sample $\widetilde{X}$ [3]. A statistic $p(\widetilde{X})$ is called p-value corresponding to the null hypothesis H_0, if $0 \leq p(\widetilde{X}) \leq 1$, and for every $0 \leq t \leq 1$,

$$\mathcal{P}[p(\widetilde{X}) \leq t | H_0] \leq t \tag{3}$$

Usually, a small p-value means strong evidence against the null hypothesis.

Another critical concept is *false discovery rate* (FDR), which can be considered as the generalization of the probability of type-I error. Denote by $\mathcal{R}$ the set of indices for the rejected hypotheses. Let $\mathcal{N}_0$ and $\mathcal{N}_1$ be the set of indices for the true null hypotheses and false null hypothesis, respectively. Denote $n_0 = |\mathcal{N}_0|$ the number of true null hypotheses. The hypotheses in Eq. (2) rejected by the detection algorithms are called discoveries. The FDR is the expected proportion of erroneous discoveries among all discoveries. Its rigorous definition is as follows.

Definition 2.2 (FDR (Benjamini & Hochberg, 1995)). False discovery proportion (FDP) is the ratio of the number of false discoveries to that of all claimed discoveries:

$$\mathrm{FDP} = \frac{|\mathcal{R} \cap \mathcal{N}_0|}{\max\{|\mathcal{R}|, 1\}}.$$

The FDR is the expectation of FDP, namely $\mathrm{FDR} = \mathbb{E}(\mathrm{FDP})$ where the expectation is taken over the true probability distribution.

[3] A sample means a sequence of examples.

Denote by $\mathcal{R}^c$ the complement of set $\mathcal{R}$. Using the confusion matrix notations [4], the FDP can be also expressed as $\mathrm{FDP} = \frac{FN}{FN+TN}$, which is the "dual" concept of the *Precision*, where

$$\mathrm{Precision} = \frac{TP}{TP+FP} = \frac{|\mathcal{R}^c \cap \mathcal{N}_0|}{\max\{|\mathcal{R}^c|, 1\}}.$$

2.3. BH Procedure

We first introduce the classical BH procedure, which is one of the most popular and heavily studied algorithms for problem (2) (Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001; Blanchard & Roquain, 2008; Basu et al., 2018).

Definition 2.3 (BH Procedure (Benjamini & Hochberg, 1995)). Given the p-values $p_1, p_2, \cdots, p_n$ corresponding to the null hypotheses $H_{1;0}, H_{2;0}, \cdots, H_{n;0}$, and let $p_{(i)}$ be the i-th order statistics from the smallest to the largest. For a pre-specified level $\alpha \in (0,\ 1)$, define

$$i^*_{BH} = \max\{i \in [n] : p_{(i)} \leq \frac{i}{n}\alpha\}. \tag{4}$$

Then, the null hypothesis $H_{(i);0}$ is rejected if $i \leq i^*_{BH}$.

In statistics, α is usually specified as 0.05.Similar to the type-I error in single hypothesis testing, a testing algorithm for problem (2) should make as many discoveries as possible while maintaining the FDR at a prescribed level. We present some known theoretical results of the BH procedure under the dependence between p-values below.

Theorem 2.4. *(Benjamini & Yekutieli, 2001) Given the dependent p-values $p_1, p_2, \ldots, p_n$, the BH procedure applied at level $\alpha \in (0,1)$ controls the FDR at level αC_n:*

$$\mathrm{FDR}_{BH} \leq \alpha C_n. \tag{5}$$

where $C_n = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}$. Particularly, there exists a joint distribution of dependent p-values for which $\mathrm{FDR}_{BH} = \min\{\alpha C_n, 1\}$.

Note that C_n is monotonically increasing and converges to ∞ as $n \to \infty$. In many applications, the number of examples in the test set $\mathcal{T}^{test}$ is large, then $\alpha C_n > 1$ when $C_n > \frac{1}{\alpha}$, implying that $\mathrm{FDR}_{BH} = 1 > \alpha$. Hence, the BH procedure can not control FDR at level α for arbitrarily dependent p-values. Then, many researches (Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001; Blanchard & Roquain, 2008) have established various conditions on p-values to control FDR for BH procedure. Besides, Clarke & Hall (2009) investigated the difficulties caused by dependence of p-values for FDR control.

[4] Based on the notations of $\mathcal{R}$, $\mathcal{N}_0$ and $\mathcal{N}_1$, we have the following relations: $TP = |\mathcal{R}^c \cap \mathcal{N}_0|$, $FN = |\mathcal{R} \cap \mathcal{N}_0|$, $FP = |\mathcal{R}^c \cap \mathcal{N}_1|$ and $TN = |\mathcal{R} \cap \mathcal{N}_1|$.

3. Generalized BH Procedure

In this section, we modify the BH procedure and then propose a novel generalized BH (g-BH) procedure as the decision rule for the OOD detection problem. We first denote $f_+(0) = \lim_{x\to 0+} f(x)$, and define two function classes:

$$\mathcal{F}_1 = \{f(x) : f_+(0) = 0, f'(x) > 0, \int_0^1 \frac{1}{f(x)}\, dx \le 1\}$$
$$\mathcal{F}_2 = \{f(x) : f_+(0) = 0, f'(x) \ge 1\},$$

for $x \in (0, 1)$. The g-BH procedure is defined as follows:

Definition 3.1 (g-BH Procedure). Given the p-values $p_1, p_2, \cdots, p_n$ corresponding to the null hypotheses $H_{1;0}, H_{2;0}, \cdots, H_{n;0}$, let $p_{(i)}$ be the i-th order statistics from the smallest to the largest. For a pre-specified level $\alpha \in (0,\ 1)$, define

$$i^*_{g-BH} = \max\{i \in [n] : f(p_{(i)}) \le \frac{i}{n}\alpha\}, \tag{6}$$

where $f(\cdot) \in \mathcal{F}_1 \cup \mathcal{F}_2$. Then, the null hypothesis $H_{(i);0}$ is rejected if $i \le i^*_{g-BH}$.

Eq. (6) indicates that the g-BH procedure rejects the null hypothesis $H_{i,0}$ if $f(p_i) \le \frac{\alpha}{n} i^*_{g-BH}$ and $|\mathcal{R}| = i^*_{g-BH}$.

We first investigate the theoretical properties of the g-BH procedure about FDR control. It is well known that if p-values $p_1, p_2, \ldots, p_n$ are mutually independent or PRDS, the BH procedure can control FDR at level $\frac{n_0}{n}\alpha$. Factually, the g-BH procedure also enjoys this theoretical results. We begin with the following definitions.

Definition 3.2 (Increasing Set). A subset $\mathcal{D} \subset \mathbb{R}^n$ is said to be increasing if for all $x \in \mathcal{D}$, $x \le y$ implies $y \in \mathcal{D}$, where the comparison of x and y is component-wise.

Definition 3.3 (PRDS(Benjamini & Yekutieli, 2001)). A family of random variables $\{X_1, X_2, \ldots, X_n\}$ is said to be PRDS on a subset $I_0 \subset \{1, 2, \ldots, n\}$ if for all $i \in I_0$, the function $\mathbb{P}((X_1, X_2, \ldots, X_n) \in \mathcal{D} | X_i = x)$ is an increasing function in x for any increasing subset $\mathcal{D}$.

According to the definition of PRDS, we obtain two useful propositions.

Proposition 3.4. *Suppose that the p-values $p_1, p_2, \ldots, p_n$ are PRDS on $\mathcal{N}_0$ and denote $p_i^* := f(p_i)$ for all $i \in \{1, 2, \ldots, n\}$ where $f(\cdot)$ is strictly increasing or decreasing. Then $\{p_1^*, p_2^*, \ldots, p_n^*\}$ is PRDS on N_0 as well.*

Proposition 3.5. *If the p-values $p_1, p_2, \ldots, p_n$ are PRDS on the set $\mathcal{N}_0$, then for any $i \in \mathcal{N}_0$, the function*

$$\mathbb{P}((p_1, p_2, \ldots, p_n) \in \mathcal{D} \mid p_i \le x)$$

is increasing in x for any increasing set $\mathcal{D}$.

Proposition 3.4 indicates that PRDS is invariant for monotonic transformation. Proposition 3.5 gives another form of PRDS in some degree. Then, the first core theorem is presented as follows:

Theorem 3.6. *Given the p-values $p_1, p_2, \ldots, p_n$ corresponding to various null hypotheses $H_{1;0}, H_{2;0}, \cdots, H_{n;0}$ and level $\alpha \in (0,\ 1)$.*

(1) For $f(\cdot) \in \mathcal{F}_1 \cup \mathcal{F}_2$, if $p_1, p_2, \ldots, p_n$ are mutually independent, then the g-BH procedure satisfies

$$FDR_{g\text{-}BH} \le \frac{n_0}{n}\alpha \le \alpha.$$

(2) For $f(\cdot) \in \mathcal{F}_2$, if $p_1, p_2, \ldots, p_n$ are PRDS on $\mathcal{N}_0$, then the g-BH procedure satisfies

$$FDR_{g\text{-}BH} \le \frac{n_0}{n}\alpha \le \alpha.$$

The proof of Theorem 3 is presented in Appendix B.1. Theorem 3.6 indicates that the g-BH procedure controls FDR at a prescribed level for at least one function class. Therefore, we can choose appropriate function in $\mathcal{F}_1$ or $\mathcal{F}_2$ for different OOD detection task according to the condition of p-values. Note that if we choose $f(x) = x \in \mathcal{F}_2$, the g-BH procedure degenerates to the classical BH procedure. Hence, our proposed method is called the generalized BH procedure.

4. FPR Control of g-BH Procedure

False positive rate (FPR) is a significant evaluation criterion for OOD detection. Using the notations $\mathcal{R}$ and $\mathcal{N}_1$, the FPR can be expressed as Although FDR control has been widely studied, relatively little is known about the theoretical properties of FPR. For example, for a pre-specified FDR control level, what is the worst expectation of FPR attainable with finite samples? In this section, we investigate the nonasymptotic behavior of FPR for the g-BH procedure. Our analytical framework is similar to that of Donoho & Jin (2004); Neuvial & Roquain (2012).

4.1. Analytical Framework

Many theoretical studies (Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001; Storey, 2002; Blanchard & Roquain, 2008) assume that p-values are available, implying that the null distribution of test statistic is known. In this case, p-values can be expressed as $p_i = \Psi(T_i)$, where $\Psi(\cdot)$ is the survival function of test statistic and T_i is the observation of test statistic corresponding to H_i. For mathematical convenience, we impose that $T_1, T_2, \ldots, T_n$ are independent continuous random variables. Reasonably, working with p-values $p_1, p_2, \ldots, p_n$ is equivalent to working with observations $T_1, T_2, \ldots, T_n$.

In this section, we describe the distribution of the observation T_i in terms of the tailed generalized Gaussian model, which is a variant of the generalized Gaussian model.

Definition 4.1 (**Tailed Generalized Gaussian Distribution Family**). A random variable X is said to follow the tailed generalized Gaussian distribution family with the location μ and the degree $\lambda > 1$, denoted by $X \sim G(\mu, \lambda)$, if its cumulative distribution function $F(\cdot)$ and the survival function $\Psi(\cdot)$ satisfy $F(0) = \Psi(0) = \frac{1}{2}$, and for constants $C_l > C_u > 0$, we have

i, there exists the positive real number X_l such that

$$\frac{e^{\frac{-|x-\mu|^\lambda}{\lambda}}}{C_l} \leq F(x-\mu) \leq \frac{e^{\frac{-|x-\mu|^\lambda}{\lambda}}}{C_u}$$

for $x - \mu < -\mathsf{X}_l$.

ii, there exists the positive real number X_u such that

$$\frac{e^{\frac{-|x-\mu|^\lambda}{\lambda}}}{C_l} \leq \Psi(x-\mu) \leq \frac{e^{\frac{-|x-\mu|^\lambda}{\lambda}}}{C_u}$$

for $x - \mu > \mathsf{X}_u$.

It is easy to verify that Gaussian distribution satisfies these conditions. Note that $\Psi(0) = \frac{1}{2}$ implies that $C_l \geq 2$ and $C_u \leq 2$. Since $\Psi(\cdot)$ is a decreasing function, $\Psi(x-\mu) > F(x-\mu)$ for $x < \mu$.

Our inspiration for considering the tail generalized Gaussian distribution comes from the previous works (Donoho & Jin, 2004; Ingster & Suslina, 2012) on global testing. In terms of the notation $G(\mu, \lambda)$, we assume that the observation T_i is distributed as

$$T_i \sim \begin{cases} G(0, \lambda) & \text{if } i \in \mathcal{N}_0 \\ G(\mu, \lambda) & \text{if } i \in \mathcal{N}_1, \end{cases} \tag{7}$$

where $\mu > 0$ is allowed to vary with the number of hypotheses n. Eq. (7) shows that the non-null hypothesis is distinguished frome null hypothesis by a positive mean shift μ. In Donoho & Jin (2004), μ is set to $\sqrt{2r\log(n)}$. In this section, μ is parameterized as $\mu = \left(\lambda r \log n\right)^{1/\lambda}$ where $r > 0$.

In Donoho & Jin (2004), we find that if $r < \rho$, where

$$\rho(\beta) = \begin{cases} \beta - \frac{1}{2} & \frac{1}{2} < \beta \leq \frac{3}{4} \\ (1 - \sqrt{1-\beta})^2 & \frac{3}{4} < \beta < 1 \end{cases}, \tag{8}$$

then the sum of probability of type-I error and type-II error for likelihood ratio test is at least 1 in asymptotic regime (non-null hypothesis can not be detected reliably), while if $r > \rho$, the sum of type-I error and type-II tends to 0 (non-null hypothesis can be detected reliably). Therefore, this paper considers the pair (r, β), where $\beta < r < r_{\max}$ for some positive constant $r_{\max} < 1$.

Following the past works (Donoho & Jin, 2004; 2006; Jin & Ke, 2016), we focus on so-called sparse regime in which the number of non-null hypotheses $|\mathcal{N}_1|$ is relatively smaller than that of the null hypotheses $|\mathcal{N}_0|$. Thus, we set $|\mathcal{N}_1| = n_1 = n^{1-\beta} < \frac{n}{2}$, suggesting that $\beta > \frac{\log 2}{\log n}$.

Recall the g-BH procedure, the set of indices of rejected hypotheses $\mathcal{R}$ can be expressed as:

$$\mathcal{R} = \{i \in [n] : i \leq i^*(p_1, p_2, \ldots, p_n)\},$$

where

$$i^*(p_1, p_2, \ldots, p_n) = \max\{i \in [n] : \nu \cdot p_{(i)}^\delta \geq \frac{i\alpha}{n}\}.$$

For any $i \in [n]$, denote by $T_{(i)}$ the i-th order statistic of $T_1, T_2, \ldots, T_n$ from the largest to the smallest. Following Barber & Candès (2015), in terms of the survival function $\Psi(T_i) = p_i$, $i^*(p_1, p_2, \ldots, p_n)$ can be presented as

$$\begin{aligned} i^*(T_1, T_2, \ldots, T_n) &= \max\{i \in [n] : f(\Psi(T_{(i)})) \geq \frac{n}{\alpha i}\} \\ &= \max\{i \in [n] : \frac{1}{f(\Psi(T_{(i)}))} \leq \frac{i}{n}\alpha\}. \end{aligned}$$

Further, we have $\mathcal{R} = \{i \in [n] : T_i \geq T^*\}$, where

$$T^* := min\{T_{(i)} : i \in [n], f(\Psi(T_{(i)})) \leq \frac{i\alpha}{n}\}. \tag{9}$$

Obviously, T^* is a data-dependent threshold. For simplicity, we denote by $\mathrm{FDR}(T^*)$ the FDR of g-BH procedure.

4.2. Upper Bound of Expectations of FPR for g-BH Procedure

In this section, we derive an upper bound of the FPR expectation for g-BH procedure based on the settings and assumptions in Section 4.1. First, we define some critical notations as follows:

$$r_{\min}(k\alpha) := \beta + \frac{\log \frac{1}{12C_l k\alpha}}{\log n},$$

and $\zeta(k\alpha) = \frac{\log \frac{1}{k\alpha}}{\log n}$. Besides, we define $n_{k\alpha} := (\frac{1}{k\alpha})^{\frac{1}{r_{\max}-\beta}}$ and

$$n_{\min} := \min\left\{n \in \mathbb{N} : n \geq [36\log(C_u n^2)]^{\frac{1}{1-r_{\max}}}\right\}.$$

For simplicity, we denote $d_\lambda(x, y) := \left|x^{1/\lambda} - y^{1/\lambda}\right|^\lambda$ and ω will be used to simplify the upper bound of $\Psi(\eta(\omega\alpha))$ in the proof of Theorem 4.2.

Then, we show the upper bound of the expectation of FPR for g-BH procedure.

Theorem 4.2. *Given the p-values* $p_1 = \Psi(T_1), p_2 = \Psi(T_2), \ldots, p_n = \Psi(T_n)$ *satisfing the condition* (7). *For* $r_{\min}(\omega\alpha) < r < r_{max}$, *if* $n \geq \max\{n_{\omega\alpha}, n_{\min}\}$, *then the FPR expectation for the g-BH procedure satisfies*

$$\mathbb{E}[FPR_{\text{g-BH}}] \leq \frac{2(\frac{108C_l}{C_u^2})^{(\frac{\beta}{r_{\max}-\beta})^{\frac{1-\lambda}{\lambda}}}}{C_u} \cdot n^{-d_\lambda(\beta+\zeta(\alpha),r)}.$$

The proof of Theorem 4.2 is presented in Section B.2. Note that $r > \beta$ and $\zeta(\alpha) \to 0$ as $n \to \infty$, then we have $n^{-d_\lambda(\beta+\zeta(\alpha),r)} < n^{-d_\lambda(\beta,r)/2} \to 0$ as $n \to \infty$. Therefore, Theorem 4.2 indicates that the FPR expectation for the g-BH procedure tends to zero as $n \to \infty$. Based on this conclusion, we obtain the following corollary.

Corollary 4.3. *Under the conditions in Theorem 4.2, the FPR of g-BH procedure satisfies*

$$\text{FPR}_{g-BH} \to 0 \quad \textit{in probability}$$

The proof of Corollary 4.3 is presented in Appendix B.3 Corollary 4.3 shows that large-scale test set is beneficial for reducing the FPR of g-BH procedure under the conditions in Theorem 4.2.

In many literature (Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001; Blanchard & Roquain, 2008), the p-values corresponding to various null hypotheses are assumed to be accessible. However, in practice, often we have little information about potential marginal distribution $\mathcal{P}_{in}$ of ID data and thus we cannot obtain the precise p-values. Instead, we just compute the empirical p-values. Specifically, given a calibrated set $\mathcal{T}^{cal} = \{X_1^{cal}, X_2^{cal}, \ldots, X_m^{cal}\}$ consisting of ID data, the empirical p-value corresponding to the testing example X_i^{test} is defined as

$$p_i = p(X_i^{test}) = \frac{|\{j \in [m] : \hat{s}(X_j^{cal}) \leq \hat{s}(X_i^{test})\}| + 1}{m+1},$$

where $\hat{s}(\cdot)$ is a certain score function. Yu et al. (2023) demonstrate that the empirical p-values based on neural networks are PRDS. Then, the practical g-BH procedure with the empirical p-values is presented in Algorithm 1, called g-BH. Obviously, Algorithm 1 does not rely on any extra information of OOD data and enables to directly utilities the existing score functions.

In a nutshell, our method has two compelling strengths. First, Algorithm 1 is distribution-free and all score-based methods can be plugged into our proposed algorithm. Besides, many theoretical results provide rigorous statistical guarantee for Algorithm 1. We examine the performance of Algorithm 1 in Section 5.

5. Experiment

In this section, we conduct extensive experiments to demonstrate the superiority of the g-BH procedure over traditional threshold-based decision rule from two perspectives: practical perspective (focusing on TPR, FPR and F1-score) and classical perspective (focusing on FPR95, AUROC and AUPR). The experimental results show that our method enables to improve the OOD detection performance of the existing methods.

Algorithm 1 g-BH

1: **Input:** Training set $\mathcal{T}$, calibrated set $\mathcal{T}^{cal} = \{X_1^{cal}, X_2^{cal}, \ldots, X_m^{cal}\}$ testing set $\mathcal{T}^{test} = \{X_1^{test}, X_2^{test}, \ldots, X_n^{test}\}$, prescribed level $\alpha \in (0, 1)$.
2: Train the score function $\hat{s}(x)$ on $\mathcal{T}$.
3: Calculate the p-values corresponding to X_i^{test}:

$$p_i = p(X_i^{test}) = \frac{|\{j \in [m] : \hat{s}(X_j^{cal}) \leq \hat{s}(X_i^{test})\}| + 1}{m+1}.$$

4: Compute $i^*_{g-BH} = \max\{i \in [n] : f(p_{(i)}) \leq \frac{i}{n}\alpha\}$.
5: **Output:** Declare that $X_{(i)}^{test}$ is OOD if $i \leq i^*_{g-BH}$, and the rests are ID.

5.1. Experimental Settings

We mainly follow the experimental settings in Yang et al. (2022); Zhang et al. (2023b), and our codes are based on Zhang et al. (2023b). We next introduce some necessary settings in our experiments.

Baselines. We choose eight popular OOD detection methods as our baselines, including MSP (Hendrycks & Gimpel, 2017), KLM (Hendrycks et al., 2022), KNN (Sun et al., 2022), LogitNorm (Wei et al., 2022), RankFeat (Song et al., 2022), ASH (Djurisic et al., 2023), Cider (Ming et al., 2023) and GEN (Liu et al., 2023).

Benchmarks. We use CIFAR-10 (Krizhevsky et al., 2009) as ID data, and use CIFAR-100, TinyImageNet (Krizhevsky et al., 2017), SVHN (Netzer et al., 2011), Texture (Kylberg, 2011), Places365 (Zhou et al., 2018) and MNIST (Deng, 2012), as OOD data.

Metrics. For the comparison between e-DR in (1) and g-BH from practical perspective, we report the following metrics: (1) ture positive rate (**TPR**), (2) false positive rate (**FPR**), (3) **F1-score**. For the comparison between e-DR and g-BH from classical perspective, we report the following metrics: (1) the FPR of OOD samples when the TPR of ID samples is at 95% (**FPR95**), (2) the area under the receiver operating characteristic curve (**AUROC**), (3) the area under the Precision-Recall curve (**AUPR**). In this paper, we regard ID as positive [5].

Models. We train a ResNet-18 (He et al., 2016a) to construct

[5] In the code of (Zhang et al., 2023b), OOD is set to positive

Table 1. Experimental results (%) of practical perspective on **CIFAR-10** as ID data. We compare the performance between e-DR and g-BH based on the same trained score function. For each baseline method, we report results after using our framework in the next line. Due to the space limitations, the results of MNIST are presented in Appendix E. ↑ indicates larger values are better and vice versa.

Model	CIFAR-100			TinyImageNet			SVHN			Texture			Place365		
	TPR ↑	FPR↓	F1 ↑	TPR↑	FPR↓	F1↑	TPR ↑	FPR ↓	F1 ↑	TPR ↑	FPR↓	F1 ↑	TPR ↑	FPR ↓	F1 ↑
ASH	95.16	63.61	73.58	95.16	60.11	77.01	95.16	65.68	49.86	95.16	59.56	81.9	95.16	55.71	46.11
ASH + g-BH	92.51	51.51	79.71	92.96	47.96	82.56	93.04	49.24	60.19	92.89	53.68	88.11	92.37	44.98	53.18
Cider	93.27	86.8	64.39	93.27	74.16	75.51	93.27	22.8	81	93.27	70.3	78.6	93.27	76.63	38.79
Cider + g-BH	90.58	80.35	65.29	91.27	68.91	78.39	91.56	18.79	86.68	90.91	59.69	86.74	90.35	69.51	42.77
GEN	95.5	57.78	75.42	95.57	52.91	79.16	95.5	48.49	57.08	95.5	51.28	83.92	95.5	51.62	48.1
GEN + g-BH	90.68	31.01	82.03	90.56	25.65	85.35	90.18	21.49	75.43	89.71	25.54	89.12	91.05	26.76	66.12
KLM	94.79	59.45	74.59	94.79	55.47	78.1	94.79	54.88	53.72	94.79	53.62	76.02	94.79	54.2	46.66
KLM + g-BH	90.42	38.51	78.67	90.74	33.76	82.19	90.21	31.72	63.54	90.38	31.34	86.75	90.72	32.12	57.81
KNN	93.67	55.17	76.29	93.67	50.59	79.91	93.67	52.75	54.97	93.67	45.98	77.25	93.67	48.21	49.83
KNN + g-BH	90.26	28.32	85.33	89.74	21.11	86.28	88.47	14.26	79.32	89.58	19.55	90.28	90.13	16.79	67.86
LogitNorm	94.83	45.3	78.98	94.83	34.88	84.28	94.83	16.68	78.03	94.83	27.66	84.4	94.83	27.57	62.67
LogitNorm + g-BH	93.21	38.43	81.32	93.72	25.92	87.39	91.49	10.06	82.66	93.28	22.49	89.72	92.41	16.53	71.39
MSP	95.27	62.83	73.87	95.27	59.25	77.3	95.27	55.45	53.7	95.27	57.28	82.45	95.27	58.86	44.87
MSP + g-BH	90.64	35.81	83.61	90.72	34.05	85.11	91.16	32.46	76.79	92.01	40.44	88.24	91.24	37.56	62.26
RankFeat	95.15	75.26	63.23	95.15	71.77	72.72	95.15	72.94	41.02	95.15	69.87	75.68	95.15	74.29	34.73
RankFeat + g-BH	93.69	70.55	65.25	92.72	68.87	74.71	92.49	67.75	45.88	93.29	64.53	78.28	92.38	69.84	37.77

Table 2. The Experimental results (%) of classical perspective on **CIFAR-10** as ID data. We compare the performance between e-DR and g-BH procedure based on the same score function. For each baseline method, we report results after using our framework in the next line. Due to the space limitations, we simplify **AUROC** as **AUC** and the results of MNIST are presented in Appendix E. ↑ indicates larger values are better and vice versa.

Model	CIFAR-100			TinyImageNet			SVHN			Texture			Place365		
	FPR95↓	AUC↑	AUPR↑	FPR95↓	AUC↑	AUPR↑	FPR95↓	AUC↑	AUPR↑	FPR95↓	AUC↑	AUPR↑	FPR95↓	AUC↑	AUPR↑
ASH	63.15	74.11	68.56	59.75	76.44	73.51	65.21	73.46	46.14	59.27	77.45	79.66	59.23	79.89	45.79
ASH + g-BH	60.12	74.58	68.72	55.21	77.65	74.23	57.18	77.21	49.12	53.58	77.58	79.74	50.67	80.42	45.73
Cider	56.18	89.19	90.58	46.76	92.04	94.01	19.14	98.25	96.21	42.06	92.73	94.59	38.49	93.18	85.27
Cider + g-BH	53.02	89.78	90.68	43.84	92.69	94.64	18.73	98.58	96.21	37.82	93.39	95.24	32.63	93.77	86.04
GEN	54.85	87.21	84.99	49.95	88.2	88.82	45.4	91.87	82.31	48.06	90.14	92.02	48.66	89.46	67.54
GEN + g-BH	43.87	88.59	85.31	38.97	89.84	89.16	36.78	92.92	82.98	33.48	91.88	92.25	40.48	89.99	67.97
KLM	60.21	77.89	69.2	56.47	80.49	75.14	55.97	84.99	63.52	54.75	82.35	82.33	55.11	78.37	36
KLM + g-BH	50.96	78.56	69.59	47.76	81.12	74.71	46.53	85.78	63.48	45.96	81.97	81.64	45.54	77.86	35.69
KNN	51.96	89.73	90.15	50.92	91.56	93.3	49.5	92.67	88.47	47.46	93.16	96.13	49.8	91.77	80.39
KNN + g-BH	46.76	90.46	90.75	41.92	91.63	93.23	44.07	92.75	88.36	38.08	93.42	95.93	40.65	92.49	80.87
LogitNorm	45.93	90.86	91.39	35.62	93.68	95.03	17.08	96.93	93.51	28.12	94.89	96.89	28.15	95.14	89.07
LogitNorm + g-BH	43.58	90.94	91.37	33.09	93.76	95.34	15.47	97.23	93.76	26.11	95.19	96.85	25.95	95.34	89.47
MSP	60.83	87.19	85.85	57.39	88.87	89.27	53.22	91.46	83.37	55.11	89.89	92.5	56.71	88.92	68.86
MSP + g-BH	47.73	88.97	86.96	43.79	89.87	90.17	38.75	92.35	84.34	41.49	90.44	93.47	43.24	90.51	70.84
RankFeat	78.41	77.98	79.11	75.56	80.94	84.25	94.86	68.15	57.04	89.47	73.46	84.15	65.71	85.99	71.68
RankFeat g-BH	72.84	78.96	79.85	68.08	81.92	84.94	87.45	69.28	57.91	85.26	74.51	84.87	57.77	86.08	72.73

the score function. More details of the experimental settings can be found in Zhang et al. (2023b).

5.2. Comparison between e-DR and g-BH: Practical Perspective

In this experiment, we apply the e-DR and g-BH to identify the OOD examples from practical perspective, respectively. The experimental results on CIFAR-10 as ID data are presented in Table 1. As the Table 1 shows, the g-BH dramatically improves the FPR and F1-score of OOD detection in all cases, despite a slight decrease in TPR. For example, using CIFAR-10 as ID and Place365 as OOD, our method reduces the FPR from 48.21% to 16.79%, and improves the F1-score from 49.83% to 67.86% compared to the e-DR based on KNN (Sun et al., 2022), a direct improvement of **31.43%** and **18.03%** at the cost of 3.54% decrease in TPR. Similarly, using TinyImageNet as OOD data, our method reduces the FPR from 52.91% to $25.65\$$, and improves the F1-score from 79.16% to 85.35% at the cost of 5.01% decrease in TPR based on GEN (Liu et al., 2023).

g-BH procedure is free from the distribution assumptions and the type of score function. As can be seen from Table 1, the g-BH consistently achieves superior performance under the different OOD data set and various type of score functions. For example, using Texture as OOD data, our method reduces the FPR by **22.19%**, **26.43%** and **16.87%** based on KLM, KNN and MSP, compared with the e-DR. Simultaneously, corresponding F1-scores are improved by **10.73%**, **13.03%** and **5.78%**, respectively. This observation suggests that the g-BH does not depend on the distributional assumptions of OOD data and the types of

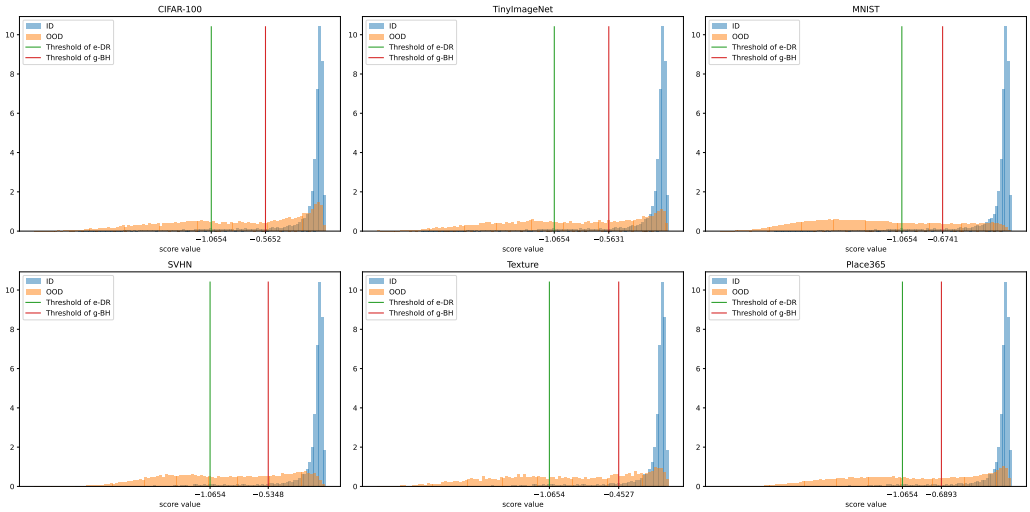

Figure 1. The comparison of threshold between e-DR and g-BH based on the same trained score function in GEN Liu et al. (2023). ID data is CIFAR-10.

score function.

Why does the g-BH performs better than e-DR? To explore the underlying reasons of the superior performance of g-BH, we compare the thresholds determined by e-DR and g-BH based on the score function in GEN. The results are presented in Figure 1. As is shown in Figure 1, the thresholds of e-DR for different OOD data are fixed because they rely on the ID validation set and can not be adjusted for the coming OOD data. Besides, the e-DR prefers to guarantee the performance in ID data (high TPR) but sacrifices the performance in OOD data (high FPR). Thus, the e-DR can not trade off between TPR and FPR well (low F1-score). By contrast, our method enables to adaptively modify thresholds in response to the various OOD data. furthermore, our method can control FDR. To control FDR, the g-BH tends to reject more null hypotheses (to classify more testing example as OOD) while maintaining small false rejections of the null hypothesis (to control the number of falsely classifying the ID as the OOD). Hence, as is shown in Table 1, our method achieves the smaller FPR and the larger F1-score than e-DR. The above analysis demonstrates that our method enables to achieve a better tradeoff between TPR and FPR.

5.3. Comparison between e-DR and g-BH: Classical Perspective

In this experiment, we compare the OOD detection performance in FPR95, AUROC and AUPR between the e-DR and g-BH. The experimental results on CIFAR-10 as ID data are presented in Table 2. From the table, we find that after adding g-BH to the existing methods, the AUROC and AUPR achieve a certain degree of improvement or are comparable with the vanilla methods. What's even more exciting, combining the existing methods with g-BH improves the FPR95 obviously. For example, plugging the MSP into g-BH and using SVHN as OOD data, the FPR95 is reduced from 53.22% to 38.75% compared with the MSP, a direct improvement of **14.47%** while maintaining the improvements of 0.89% and 0.97% in terms of AUROC and AUPR, respectively. Averaged over a diverse collection of OOD datasets, our method reduces the FPR95 by **13.65%** compared with the MSP. When using Texture as OOD data, our method reduces the FPR95 by **14.58%** compared with the GEN. Notably, these improvements consistently exist for different OOD data and various type of score functions. Therefore, our proposed method enables to enhance the OOD detection performance of the existing methods without the dependence on the distribution assumptions of OOD data and the type of the score functions.

6. Conclusion

In this paper, we systematically investigate the decision rule for OOD detection, which is overlooked by the existing literature. Concretely, we formulate the OOD detection task as the multiple hypothesis testing problem and propose a novel g-BH procedure to solve it. Theoretically, g-BH procedure controls FDR at pre-specified level. Besides, we derive an upper bound of the expectation of FPR under the tailed generalized Gaussian distribution family. Finally, we

conduct the extensive experiments to verify the superiority of g-BH over the traditional score-based decision rule.

Impact Statement

To our best knowledge, this work has no negative social impact. This work mainly provides a solid theoretical support for the field of the OOD detection and improves the performance of existing methods obviously. Hence, our work may promote the development of the related applications.

Acknowledge

This work is supported by the National Key R&D Program of China under Grant 2023YFC3604702, and the Fundamental Research Fund Program of LIESMARS.

References

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Chen, J., Chrzanowski, M., Coates, A., Diamos, G., Elsen, E., Engel, J. H., Fan, L., Fougner, C., Hannun, A. Y., Jun, B., Han, T., LeGresley, P., Li, X., Lin, L., Narang, S., Ng, A. Y., Ozair, S., Prenger, R., Qian, S., Raiman, J., Satheesh, S., Seetapun, D., Sengupta, S., Wang, C., Wang, Y., Wang, Z., Xiao, B., Xie, Y., Yogatama, D., Zhan, J., and Zhu, Z. Deep speech 2 : End-to-end speech recognition in english and mandarin. In *ICML*, volume 48, pp. 173–182, 2016.

Barber, R. F. and Candès, E. J. Controlling the false discovery rate via knockoffs. *Annals of Statistics*, 43(5), 2015.

Basu, P., Cai, T. T., Das, K., and Sun, W. Weighted false discovery rate control in large-scale multiple testing. *Journal of the American Statistical Association*, 113(523):1172–1183, 2018.

Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Benjamini, Y. and Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pp. 1165–1188, 2001.

Blanchard, G. and Roquain, E. Two simple sufficient conditions for fdr control. *Electronic Journal of Statistics*, 2: 963–992, 2008.

Casella, G. and Berger, R. L. *Statistical inference*. Cengage Learning, 2002.

Chen, Y. and Liu, W. A theory of transfer-based black-box attacks: Explanation and implications. In *NeurIPS*, 2023.

Clarke, S. and Hall, P. Robustness of multiple testing procedures against dependence. *Annals of Statistics*, 37(1):332 – 358, 2009.

Dankers, V., Bruni, E., and Hupkes, D. The paradox of the compositionality of natural language: A neural machine translation case study. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *ACL*, pp. 4154–4175, 2022.

Deng, L. The MNIST database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Process. Mag.*, 29(6):141–142, 2012.

DeVries, T. and Taylor, G. W. Learning confidence for out-of-distribution detection in neural networks. *arXiv*, abs/1802.04865, 2018.

Djurisic, A., Bozanic, N., Ashok, A., and Liu, R. Extremely simple activation shaping for out-of-distribution detection. In *ICLR*, 2023.

Donoho, D. and Jin, J. Higher criticism for detecting sparse heterogeneous mixtures. *Annals of statistics*, 32(3):962–994, 2004.

Donoho, D. and Jin, J. Asymptotic minimaxity of false discovery rate thresholding for sparse exponential data. *Annals of Statistics*, pp. 2980–3018, 2006.

Frolova, D., Vasiliuk, A., Belyaev, M., and Shirokikh, B. Solving sample-level out-of-distribution detection on 3d medical images. *arXiv preprint arXiv:2212.06506*, 2022.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016a.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016b.

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

Hendrycks, D., Basart, S., Mazeika, M., Zou, A., Kwon, J., Mostajabi, M., Steinhardt, J., and Song, D. Scaling out-of-distribution detection for real-world settings. In *ICML*, volume 162, pp. 8759–8773, 2022.

Huang, R., Geng, A., and Li, Y. On the importance of gradients for detecting distributional shifts in the wild. In *NeurIPS*, pp. 677–689, 2021.

Huang, X., Kroening, D., Ruan, W., Sharp, J., Sun, Y., Thamo, E., Wu, M., and Yi, X. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Computer Science Review*, 37:100270, 2020.

Ingster, Y. and Suslina, I. A. Nonparametric goodness-of-fit testing under gaussian models. volume 169. Springer Science & Business Media, 2012.

Jin, J. and Ke, Z. T. Rare and weak effects in large-scale inference: methods and phase diagrams. *Statistica Sinica*, pp. 1–34, 2016.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, 2017.

Kylberg, G. *Kylberg texture dataset v. 1.0*. Centre for Image Analysis, Swedish University of Agricultural Sciences and . . . , 2011.

Lee, K., Lee, K., Lee, H., and Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *NeurIPS 2018*, pp. 7167–7177, 2018a.

Lee, K., Lee, K., Lee, H., and Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *NeurIPS*, pp. 7167–7177, 2018b.

Li, K., Chen, K., Wang, H., Hong, L., Ye, C., Han, J., Chen, Y., Zhang, W., Xu, C., Yeung, D., Liang, X., Li, Z., and Xu, H. CODA: A real-world road corner case dataset for object detection in autonomous driving. In *ECCV*, volume 13698, pp. 406–423, 2022.

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

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

Liu, X., Lochman, Y., and Zach, C. GEN: pushing the limits of softmax-based out-of-distribution detection. In *CVPR*, pp. 23946–23955, 2023.

Ma, X., Wang, Z., and Liu, W. On the tradeoff between robustness and fairness. In *NeurIPS 2022*, 2022.

Ming, Y., Fan, Y., and Li, Y. POEM: out-of-distribution detection with posterior sampling. In *ICML*, volume 162, pp. 15650–15665, 2022.

Ming, Y., Sun, Y., Dia, O., and Li, Y. How to exploit hyperspherical embeddings for out-of-distribution detection? In *ICLR*, 2023.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. 2011.

Neuvial, P. and Roquain, E. On false discovery rate thresholding for classification under sparsity. *Annals of Statistics*, 40(5):2572–2600, 2012.

Nguyen, A. M., Yosinski, J., and Clune, J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *CVPR*, pp. 427–436, 2015.

Özbayoglu, A. M., Gudelek, M. U., and Sezer, O. B. Deep learning for financial applications : A survey. *Applied Soft Computing*, 93:106384, 2020.

Shi, L. and Liu, W. Adversarial self-training improves robustness and generalization for gradual domain adaptation. In *NeurIPS*, 2023.

Song, Y., Sebe, N., and Wang, W. Rankfeat: Rank-1 feature removal for out-of-distribution detection. In *NeurIPS*, 2022.

Storey, J. D. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002.

Storey, J. D., Taylor, J. E., and Siegmund, D. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):187–205, 2004.

Sun, Y., Ming, Y., Zhu, X., and Li, Y. Out-of-distribution detection with deep nearest neighbors. In *ICML*, volume 162, pp. 20827–20840, 2022.

Wang, H., Li, Z., Feng, L., and Zhang, W. Vim: Out-of-distribution with virtual-logit matching. In *CVPR*, pp. 4911–4920, 2022.

Wei, H., Xie, R., Cheng, H., Feng, L., An, B., and Li, Y. Mitigating neural network overconfidence with logit normalization. In *ICML*, volume 162, pp. 23631–23644, 2022.

Yang, J., Wang, H., Feng, L., Yan, X., Zheng, H., Zhang, W., and Liu, Z. Semantically coherent out-of-distribution detection. In *ICCV*, pp. 8281–8289, 2021.

Yang, J., Wang, P., Zou, D., Zhou, Z., Ding, K., Peng, W., Wang, H., Chen, G., Li, B., Sun, Y., Du, X., Zhou, K., Zhang, W., Hendrycks, D., Li, Y., and Liu, Z. Openood: Benchmarking generalized out-of-distribution detection. In *NeurIPS*, 2022.

Yu, C., Ma, X., and Liu, W. Delving into noisy label detection with clean data. In *ICML*, volume 202, pp. 40290–40305, 2023.

Zhang, J., Fu, Q., Chen, X., Du, L., Li, Z., Wang, G., Liu, X., Han, S., and Zhang, D. Out-of-distribution detection based on in-distribution data patterns memorization with modern hopfield energy. In *ICLR*, 2023a.

Zhang, J., Yang, J., Wang, P., Wang, H., Lin, Y., Zhang, H., Sun, Y., Du, X., Zhou, K., Zhang, W., Li, Y., Liu, Z., Chen, Y., and Li, H. Openood v1.5: Enhanced benchmark for out-of-distribution detection. *CoRR*, abs/2306.09301, 2023b.

Zhou, B., Lapedriza, À., Khosla, A., Oliva, A., and Torralba, A. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(6): 1452–1464, 2018.

Zou, X. and Liu, W. On the adversarial robustness of out-of-distribution generalization models. In *NeurIPS*, 2023.

Zou, X. and Liu, W. Coverage-guaranteed prediction sets for out-of-distribution data. In *AAAI*, pp. 17263–17270, 2024.

A. Related Work

OOD Detection AI safety has become critical in machine learning community (Ma et al., 2022; Chen & Liu, 2023; Shi & Liu, 2023), in which OOD detection and generalization are two hot topics (Zou & Liu, 2024; 2023). Lots of algorithms (Hendrycks & Gimpel, 2017; Liu et al., 2020; Huang et al., 2021; Hendrycks et al., 2022; Djurisic et al., 2023; Zhang et al., 2023a) have been proposed to solve these problem. Existing methods for OOD detection can be roughly divided into two categories: post-hoc and training-based. Post-hoc methods (Liang et al., 2018; Liu et al., 2020; 2023) directly obtain confidence from the classifier with some beneficial design to OOD detection. These designs mainly focus on the loss function (Liu et al., 2020; Huang et al., 2021; Liu et al., 2023), classifier architecture (Lee et al., 2018b; Djurisic et al., 2023), and some post-hoc processing techniques (Hendrycks & Gimpel, 2017; Liang et al., 2018). The distance-based methods (Lee et al., 2018a; Hendrycks et al., 2022; Sun et al., 2022), which usually compute distance in the high-dimension space such as feature space and gradient space to distinguish ID and OOD example, also belongs to this category of methods. Training-based methods (DeVries & Taylor, 2018; Liang et al., 2018; Wei et al., 2022) are allowed to specifically retrain new auxiliary networks specifically for OOD detection rather than directly using already trained models. Additionally, some methods need access to the OOD example to train the new networks (Yang et al., 2021; Ming et al., 2022). All these researches focus on designing or training a powerful score functions but overlook the systematic investigation of decision rule. Different from previous literature, this paper aims to design a new decision rule with rigorous theoretical guarantee and well empirical performance.

FDR Control Similar to control type-I error in single hypothesis testing, how to control FDR is the core problem for multiple hypothesis testing algorithms. For this purpose, early work (Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001; Storey, 2002; Storey et al., 2004) assumes that the p-values corresponding to the hypotheses are mutually independent or follow certain patterns such as PRDS (Benjamini & Yekutieli, 2001), self-consistent condition and dependency control condition (Blanchard & Roquain, 2008). Besides, Clarke & Hall (2009) show that the difficulties caused by dependence of p-values are less serious than in classical cases when the null distributions of testing statistics are relatively light-tailed. Based on the thought of variable selection, Barber & Candès (2015) use knockoff filter to controls FDR. Recently, Yu et al. (2023) utilities the BH procedure with empirical p-values to tackle the noisy label detection problem and achieves SOTA performance.

B. Omited Proof in Main Context

B.1. Proof of Theorem 3.6

Proof. We will prove Theorem 3.6 in two cases.

Case 1: p-values are mutually independent.

Recall the definition of FDP, we have

$$\frac{|\mathcal{R}\cap\mathcal{N}_0|}{\max\{|\mathcal{R}|,1\}}=\sum_{i\in\mathcal{N}_0}\frac{R_i}{\max\{|\mathcal{R}|,1\}}=\sum_{i\in\mathcal{N}_0}\sum_{j=1}^{n}\frac{R_i\mathbb{1}_{\{|\mathcal{R}|=j\}}}{j}.$$

where $R_i=\mathbb{1}_{\{H_{i;0}\text{ is rejected}\}}$.

Define $|\mathcal{R}(p_i\to 0)|$ to be the number of rejected hypotheses from g-BH procedure if we changed p_i to 0, keeping all the rest the same. If $R_i=1$, denote $|\mathcal{R}|=m$ where $m\in[n]$, then we have $\sum_{j=1}^{n}\frac{R_i\mathbb{1}_{\{|\mathcal{R}|=j\}}}{j}=\frac{1}{m}$. According to the definition of R_i, for $f(\cdot)\in\mathcal{F}_1\cup\mathcal{F}_2$, $R_i=1$ means $H_{i,0}$ is rejected, and further $f(p_i)$ already is below the rejection threshold. Hence, changing p_i to 0 (equivalently, changing $f(p_i)$ to 0) does not increase the number of rejected hypotheses and $|\mathcal{R}(p_i\to 0)|=|\mathcal{R}|=m$. We conclude

$$\sum_{j=1}^{n}\frac{R_i\mathbb{1}_{\{|\mathcal{R}|=j\}}}{j}=\sum_{j=1}^{n}\frac{R_i\mathbb{1}_{\{|\mathcal{R}(p_i\to 0)|=j\}}}{j}. \tag{10}$$

If $R_i=0$, Eq. (10) still holds. Let σ_i be the σ-algebra generated by $p_1,\dots,p_{i-1},p_{i+1},\dots,p_n$. Recall the g-BH procedure, if $|\mathcal{R}|=j$, then rejecting $H_{i,0}$ is equivalent to $f(p_i)\le\frac{j\alpha}{n}$. Since $|\mathcal{R}(p_i=0)|$ is σ_i-measurable and R_i is independent of

σ_i, we have

$$\begin{aligned}\mathbb{E}(R_i \mathbb{1}_{\{|\mathcal{R}(p_i=0)|=j\}}|\sigma_i) &= \mathbb{E}(\mathbb{1}_{\{f(p_i)\leq \frac{j\alpha}{n}\}} \mathbb{1}_{\{|\mathcal{R}(p_i=0)|=j\}}|\sigma_i) \\ &= \mathbb{1}_{\{|\mathcal{R}(p_i=0)|=j\}} \mathbb{P}(f(p_i) \leq \frac{j\alpha}{n})\end{aligned}$$

We now consider the situation where $f(\cdot) \in \mathcal{F}_1$. Recall that the definition of p-value, for any $c \in (0,\ 1)$, we have $\mathbb{P}(p_i \leq c) \leq c$. Denote by X the random variable uniformly distributed on $(0, 1)$. If $c \geq 1$, then $\mathbb{P}(p_i \leq c) = \mathbb{P}(X \leq c) = 1$. Otherwise, if $0 < c < 1$, we get $\mathbb{P}(p_i \leq c) \leq c = \mathbb{P}(X \leq c)$. Therefore, we have

$$\mathbb{P}(p_i \leq c) \leq \mathbb{P}(X \leq c).$$

for $c > 0$. Then, it follows that

$$\mathbb{P}\left(\frac{1}{f(p_i)} > c\right) \leq \mathbb{P}\left(\frac{1}{f(X)} > c\right).$$

Note that for any non-negative random variable Y, its expectation satisfies

$$\mathbb{E}(Y) = \int_0^\infty y f(y)\, dy = \int_0^\infty \mathbb{P}(Y > y)\, dy.$$

Then we obtain

$$\begin{aligned}\mathbb{E}\left[\frac{1}{f(p_i)}\right] &= \int_0^\infty \mathbb{P}\left(\frac{1}{f(p_i)} \geq x\right) dx \leq \int_0^\infty \mathbb{P}\left(\frac{1}{f(X)} \geq x\right) dx \\ &= \mathbb{E}\left[\frac{1}{f(X)}\right] = \int_0^1 \frac{1}{f(x)} \cdot 1\, dx \leq 1.\end{aligned} \tag{11}$$

By Markov's inequality, we have

$$\mathbb{P}(f(p_i) \leq \frac{j\alpha}{n}) = \mathbb{P}(\frac{1}{f(p_i)} \geq \frac{1}{\frac{j\alpha}{n}}) \leq \mathbb{E}\left[\frac{1}{f(p_i)}\right] \cdot \frac{j\alpha}{n} \leq \frac{j\alpha}{n}. \tag{12}$$

It is easy to verify that the argument in Eq. (12) is still valid if $f(\cdot) \in \mathcal{F}_2$. The above analysis demonstrates that

$$\begin{aligned}\mathbb{E}\Big[\frac{R_i}{\max\{|\mathcal{R}|, 1\}}|\sigma_i\Big] &= \sum_{j=1}^n \mathbb{E}\Big[\frac{R_i \mathbb{1}_{\{|\mathcal{R}(p_i=0)|=j\}}}{j}|\sigma_i\Big] \\ &= \sum_{j=1}^n \mathbb{E}\Big[\frac{\mathbb{1}_{\{f(p_i)\leq \frac{j\alpha}{n}\}} \mathbb{1}_{\{|\mathcal{R}(p_i=0)|=j\}}}{j}|\sigma_i\Big] \\ &= \sum_{j=1}^n \frac{\mathbb{1}_{\{|\mathcal{R}(p_i=0)|=j\}}}{j} \mathbb{P}(f(p_i) \leq \frac{j\alpha}{n}) \\ &\leq \frac{\alpha}{n} \sum_{j=1}^n \mathbb{1}_{\{|\mathcal{R}(p_i=0)|=j\}} = \frac{\alpha}{n}.\end{aligned}$$

Note that $n_0 = |\mathcal{R}|$ and $\frac{R_i}{\max\{|\mathcal{R}|,1\}}$ are identically distributed for all $i \in \mathcal{N}_0$ since the null p-values have the same distribution under the null hypotheses in Eq. (2). Then we have

$$\mathbb{E}(\frac{|\mathcal{R} \cap \mathcal{N}_0|}{\max\{|\mathcal{R}|, 1\}}) = \sum_{i \in \mathcal{N}_0} \mathbb{E}\Big[\mathbb{E}\Big[\frac{R_i}{\max\{|\mathcal{R}|, 1\}}|\sigma_i\Big]\Big] \leq \frac{n_0 \alpha}{n}.$$

Case 2: p-values are PRDS.

Since p-values $p_1, p_2, \cdots, p_n$ are PRDS, then $f(p_1), f(p_2), \cdots, f(p_n)$ are still PRDS according to Proposition 3.4. According to the definition of g-BH procedure, we have $|\mathcal{R}| = i^*_{g-BH}$. Without loss of generality, we assume that $i^*_{g-BH} \geq 1$.

Using the notation i^*_{g-BH}, the FDR can be expressed as

$$\begin{aligned}\mathbb{E}(\frac{|\mathcal{R}\cap\mathcal{N}_0|}{\max\{|\mathcal{R}|,1\}}) &= \mathbb{E}\left[\sum_{i\in\mathcal{H}_0}\frac{\mathbb{1}(f(p_i)\le\frac{i^*_{g-BH}}{n}\alpha)}{i^*_{g-BH}}\right]\\ &= \mathbb{E}\left[\sum_{i\in\mathcal{H}_0}\frac{\mathbb{1}(f(p_i)\le\frac{i^*_{g-BH}}{n}\alpha)}{\frac{i^*_{g-BH}}{n}\alpha}\cdot\frac{\alpha}{n}\right]\end{aligned}$$

For simplicity, we denote $\varpi_{g-BH}=\frac{i^*_{g-BH}}{n}\alpha$. Obviously, ϖ_{g-BH} is the function with respect to $f(p_1),f(p_2),\cdots,f(p_n)$ and $\varpi_{g-BH}<1$.

For a positive number $\epsilon\in(0,\ 1)$, we choose a positive integer m such that $\varpi_{g-BH}>\epsilon^m$. Denote $s_j=\epsilon^{m+1-j}$ for $j\in[m+1]$. Note that

$$\mathbb{P}\left(f(p_i)\le\varpi_{g-BH}\right)=\mathbb{P}\left(f(p_i)\le\varpi_{g-BH},\varpi_{g-BH}\in\cup_{j=1}^m(s_j,\ s_{j+1}]\right).$$

Then, for $i\in\mathcal{H}_0$, the following chain of inequities hold:

$$\begin{aligned}\mathbb{E}\left[\frac{\mathbb{1}(f(p_i)\le\varpi_{g-BH})}{\varpi_{g-BH}}\right] &\le \sum_{j=1}^m\frac{\mathbb{P}(f(p_i)\le s_{j+1},\varpi_{g-BH}\in(s_j,\ s_{j+1}])}{s_j}\\ &= \sum_{j=1}^m\frac{\mathbb{P}(f(p_i)\le s_{j+1})\cdot\mathbb{P}(\varpi_{g-BH}\in(s_j,\ s_{j+1}]|f(p_i)\le s_{j+1})}{\mathbb{P}(f(p_i)\le s_{j+1})}\cdot\frac{\mathbb{P}(f(p_i)\le s_{j+1})}{s_j}\\ &\le \sum_{j=1}^m P(\varpi_{g-BH}\in(s_j,\ s_{j+1}]|f(p_i)\le s_{j+1})\cdot\frac{s_{j+1}}{s_j}\\ &\le \epsilon^{-1}\sum_{j=1}^m\left(\mathbb{P}(\varpi_{g-BH}\le s_{j+1}|f(p_i)\le s_{j+1})-\mathbb{P}(\varpi_{g-BH}\le s_j|f(p_i)\le s_{j+1})\right)\\ &= \epsilon^{-1}\sum_{j=1}^{m-1}\left(\mathbb{P}(\varpi_{g-BH}\le s_{j+1}|f(p_i)\le s_{j+1})-\mathbb{P}(\varpi_{g-BH}\le s_{j+1}|f(p_i)\le s_{j+2})\right)\\ &\quad+\epsilon^{-1}\left(\mathbb{P}(\varpi_{g-BH}\le s_{m+1}|f(p_i)\le s_{m+1})-\mathbb{P}(\varpi_{g-BH}\le s_1|f(p_i)\le s_2)\right)\\ &\le \epsilon^{-1}\cdot\mathbb{P}(\varpi_{g-BH}\le s_{m+1}|f(p_i)\le s_{m+1})\qquad\text{(by Proposition 3.5)}\\ &\le \epsilon^{-1}.\end{aligned}$$

Letting $\epsilon\to1$ and applying the monotone convergence theorem, we have

$$\mathbb{E}\left[\frac{\mathbb{1}(f(p_i)\le\varpi_{g-BH})}{\varpi_{g-BH}}\right]\le1.$$

Then, we have

$$\mathbb{E}(\frac{|\mathcal{R}\cap\mathcal{N}_0|}{\max\{|\mathcal{R}|,1\}})=\frac{\alpha}{n}\sum_{i\in\mathcal{H}_0}\mathbb{E}\left[\frac{\mathbb{1}(f(p_i)\le\varpi_{g-BH})}{\varpi_{g-BH}}\right]\le\frac{n_0\alpha}{n}.$$

which completes the proof of Theorem 3.6.

□

B.2. Proof of Theorem 4.2

Before our proofs, we first recall some important notations. Motivated by Donoho & Jin (2004), we define a critical threshold

$$r_{\min}(k\alpha):=\beta+\frac{\log\frac{1}{12C_lk\alpha}}{\log n},\tag{13}$$

where $k > 0$. Denote $\zeta(k\alpha) = \frac{\log \frac{1}{k\alpha}}{\log n}$. In terms of the notation $\zeta(k\alpha)$, we have $k\alpha = n^{-\zeta(\alpha)}$ and $r_{\min}(k\alpha)$ can be expressed as

$$r_{\min}(k\alpha) := \beta + \zeta(k\alpha) + \frac{\log \frac{1}{12C_l}}{\log n}. \tag{14}$$

For analytical simplicity, we denote $n_{k\alpha} := (\frac{1}{k\alpha})^{\frac{1}{r_{\max}-\beta}}$. If $n \geq n_{k\alpha}$, then we have $\beta + \zeta(k\alpha) < r_{\max}$. Associated with T^*, we define another important threshold:

$$\eta(k\alpha) := \left(\lambda r_{\min}\left(k\alpha\right) \log n\right)^{1/\lambda}. \tag{15}$$

More generally, for a testing procedure $g(\cdot)$, if $\mathcal{R}$ corresponding to g can be expressed as:

$$\mathcal{R}_g = \{i \in [n] : T_i \geq t\},$$

where t is a positive threshold (data-dependent or fixed), we use the notations $\mathrm{FDP}(t)$, $\mathrm{FDR}(t)$ and $\mathrm{FPR}(t)$ to denote the metrics associated with the procedure g. In our proof, the following lemmas are used.

Note that $\mu = (\lambda r \log n)^{1/\lambda}$. As $n \to \infty$, then $\eta(k\alpha) \to \infty$ and $\eta(k\alpha) - \mu \to -\infty$ if $r > r_{\min}(k\alpha)$. Therefore, for sufficiently large n, we have $\eta(k\alpha) > \mathsf{X}_u$ and $\eta(k\alpha) - \mu < -\mathsf{X}_l$.

Lemma B.1. *Suppose random variable X follows the binomial distribution $B(n, p)$. For any $0 < \gamma < 1$, we have*

i. Upper tail bound:

$$\mathbb{P}\left(X \geq (1+\gamma)\mathbb{E}(X)\right) \leq \exp\left(-\frac{\delta^2 \mathbb{E}(X)}{3}\right);$$

ii. Lower tail bound:

$$\mathbb{P}\left(X \leq (1-\gamma)\mathbb{E}(X)\right) \leq \exp\left(-\frac{\delta^2 \mathbb{E}(X)}{2}\right);$$

Lemma B.2. *Suppose that the observations $T_1, T_2, \ldots, T_n$ satisfy the condition in Eq. (7). For $r_{\min}(k\alpha) < r < r_{\max}$ and $n \geq n_{k\alpha}$, if threshold $t \leq \eta(k\alpha)$, we have $\mathbb{E}[\mathrm{FPR}(t)] \leq \mathbb{E}[\mathrm{FPR}(\eta(k\alpha))]$. and*

$$\mathbb{E}[\mathrm{FPR}(\eta(k\alpha))] \leq \frac{\rho^{(\frac{\beta}{r_{\max}-\beta})^{\frac{1-\lambda}{\lambda}}}}{C_u} \cdot n^{-d_\lambda(\beta+\zeta(\alpha),r)},$$

where

$$\rho = \begin{cases} \frac{1}{12C_l k^2} & 0 < k < \frac{1}{12C_l} \\ 12C_l & \frac{1}{12C_l} \leq k < 1 \\ 12C_l k^2 & k \geq 1. \end{cases}$$

The proof of Lemma B.2 is presented in Appendix C.3.

Proof of Theorem 4.2. In order complete our proof, we need following argument:

$$\mathbb{P}\big(T^* > \eta(\omega\alpha)\big) \leq \exp\big(-\frac{n^{1-r_{\max}}}{24}\big). \tag{16}$$

We now prove the bound (16). we define the empirical survival function $\hat{\Psi}$ of survival function Ψ as

$$\hat{\Psi}\big(t\big) = \big(1 - \frac{1}{n^\beta}\big) \cdot \hat{\Psi}_0\big(t\big) + \frac{1}{n^\beta} \cdot \hat{\Psi}_1\big(t\big), \tag{17}$$

where

$$\hat{\Psi}_0\big(t\big) = \frac{1}{n - n^{1-\beta}} \sum_{i \in \mathcal{N}_0} \mathbf{1}\big(T_i \geq t\big) \text{ and } \hat{\Psi}_1\big(t\big) = \frac{1}{n^{1-\beta}} \sum_{i \in \mathcal{N}_1} \mathbf{1}\big(T_i \geq t\big).$$

According to the settings in Section 4.1, we have

$$p_{(i)} = \Psi\left(T_{(i)}\right) \quad \text{and} \quad \hat{\Psi}\left(T_{(i)}\right) = \frac{i}{n}, \tag{18}$$

and $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(n)}$. Then, the hypothesis H_i is rejected by g-BH procedure if $T_i \geq T^* = T_{(i^*)}$ (For simplicity, we use i^* to represent i^*_{g-BH}), where

$$i^* = \max\left\{i \in [n] \colon f(\Psi\left(T_{(i)}\right)) \leq \alpha\hat{\Psi}\left(T_{(i)}\right)\right\}. \tag{19}$$

We define $A = \{T^* \leq \eta(\omega\alpha)\}$. Recall the definition of $r_{\min}(\omega\alpha)$, we have

$$\log\Psi\big(\eta(\alpha\omega)\big) \leq -r_{\min}\left(\alpha\omega\right)\log n + \log\frac{1}{C_u} \leq \log\frac{\alpha}{3n^\beta}$$

namely, $\Psi\big(\eta(\alpha\omega)\big) \leq \frac{\alpha}{3n^\beta}$. Since $f(\cdot)$ in g-BH procedure is increasing, we get

$$f(\Psi\big(\eta(\alpha\omega)\big)) \leq f(\frac{\alpha}{3n^\beta}).$$

Without loss of generality, we assume that $f(x) \leq 9x$ for $0 \leq x \leq 1$. It follows that

$$\begin{aligned}
A = \{T^* \leq \eta(\alpha\omega)\} &\supset \left\{f(\Psi\big(\eta(\alpha\omega)\big)) \leq \alpha\hat{\Psi}\big(\eta(\alpha\omega)\big)\right\} \\
&\supset \left\{f(\Psi\big(\eta(\alpha\omega)\big)) \leq \frac{\alpha}{n^\beta}\cdot\frac{S_1}{n^{1-\beta}}\right\} \\
&\supset \left\{\frac{f(\frac{\alpha}{3n^\beta})}{\frac{9\alpha}{3n^\beta}} \leq \frac{9}{3}\cdot\frac{S_1}{n^{1-\beta}}\right\} \\
&\supset \left\{\frac{n^{1-\beta}}{3} \leq S_1\right\},
\end{aligned} \tag{20}$$

where $S_1 = \sum_{i\in\mathcal{N}_1}\mathbf{1}\big(T_i \geq \eta(\alpha\omega)\big) \sim B\big(\Psi\big(\eta(\alpha\omega) - \mu\big),\ n^{1-\beta}\big)$. Hence, we have

$$\mathbb{P}(A) \geq \mathbb{P}\left(S_1 > \frac{n^{1-\beta}}{3}\right).$$

Further, we conclude

$$\mathbb{P}\left(T^* > \eta(\alpha\omega)\right) \ \leq\ 1 - \mathbb{P}\left(S_1 > \frac{n^{1-\beta}}{3}\right) \ =\ \mathbb{P}\left(S_1 \leq \frac{n^{1-\beta}}{3}\right).$$

Since $r > r_{\min}(\omega\alpha)$, then we have $\eta(\alpha\omega) \leq \mu$, $\Psi\left(\eta(\alpha\omega) - \mu\right) \geq \frac{1}{2}$ and $\mathbb{E}(S_1) \geq \frac{n^{1-\beta}}{2}$. By Lemma B.1, we obtain

$$\begin{aligned}
\mathbb{P}\left(S_1 \leq \frac{n^{1-\beta}}{3}\right) &\leq \mathbb{P}\left(S_1 \leq \frac{2}{3}\mathbb{E}(S_1)\right) \leq \exp\big(-\frac{\mathbb{E}(S_1)}{18}\big) \\
&\leq \exp\big(-\frac{n^{1-\beta}}{36}\big) \leq \exp\big(-\frac{n^{1-r_{\max}}}{36}\big),
\end{aligned}$$

Therefore, we have established the required claim (16).

Now we derive the upper bound of expectation of FPR for g-BH procedure. Denote $\mathbb{E}(\mathrm{FPR}(\cdot \mid A))$ and $E(\mathrm{FPR}(\cdot \mid A^c))$ the conditional expectations of FPR on A and its complement A^c, respectively. Observe that

$$\omega = \frac{C_u}{36C_l} < \frac{1}{12C_l}(C_u \leq 2), \quad (\frac{108C_l}{C_u^2})^{(\frac{\beta}{r_{\max}-\beta})^{\frac{1-\lambda}{\lambda}}} \geq 1, \quad d_\lambda(\beta + \zeta(\omega\alpha), r) \leq 2.$$

Then, if $n \geq n_{\min}$, we have

$$\frac{(\frac{108C_l}{C_u^2})^{(\frac{\beta}{r_{\max}-\beta})^{\frac{1-\lambda}{\lambda}}}}{C_u}\cdot n^{-d_\lambda(\beta+\zeta(\alpha),r)} \geq \frac{1}{C_u n^2} \geq \exp\big(-\frac{n^{1-r_{\max}}}{36}\big).$$

Based on the bound (16) and Lemma B.2, we have

$$\begin{aligned}\mathbb{E}(\mathrm{FPR}(T^*)) &= \mathbb{P}\big(A\big)\cdot\mathbb{E}(\mathrm{FPR}(T^*)|A) + \mathbb{P}\big(A^c\big)\cdot\mathbb{E}(\mathrm{FPR}(T^*)|A^c)\\ &\le \mathbb{P}\big(A\big)\cdot\mathbb{E}(\mathrm{FPR}(\eta(\omega\alpha))|A) + \mathbb{P}\big(A^c\big) \le \mathbb{E}(\mathrm{FPR}(\eta(\omega\alpha))) + \mathbb{P}\big(A^c\big)\\ &\le \mathbb{E}(\mathrm{FPR}(\eta(\omega\alpha))) + \exp\big(-\frac{n^{1-r_{\max}}}{36}\big)\\ &\le \frac{(\frac{108C_l}{C_u^2})^{(\frac{\beta}{r_{\max}-\beta})^{\frac{1-\lambda}{\lambda}}}}{C_u}\cdot n^{-d_\lambda(\beta+\zeta(\alpha),r)} + \exp\big(-\frac{n^{1-r_{\max}}}{36}\big)\\ &\le \frac{2(\frac{108C_l}{C_u^2})^{(\frac{\beta}{r_{\max}-\beta})^{\frac{1-\lambda}{\lambda}}}}{C_u}\cdot n^{-d_\lambda(\beta+\zeta(\alpha),r)},\end{aligned}$$

which completes the proof. □

B.3. Proof of Corollary 4.3

Proof. Theorem 4.2 shows

$$\lim_{n\to\infty}\mathbb{E}(\mathrm{FPR}_{g-BH}) = 0.$$

By Markov's inequality, for any $\epsilon > 0$, we have

$$\mathbb{P}(\mathrm{FPR}_{g-BH} > \epsilon) \le \frac{\mathbb{E}(\mathrm{FPR}_{g-BH})}{\epsilon} \to 0$$

as $n \to \infty$, namely

$$\mathrm{FPR}_{g-BH} \to 0 \quad \text{in probability.}$$

□

C. Proof of Propositions and Lemmas

C.1. Proof of Proposition 3.4 and 3.5

Proof. (proof of Proposition 3.4) Without loss of generality, we assume that $f(\cdot)$ is strictly increasing. For any strictly increasing set $\mathcal{D}$, denote $f^{-1}(\mathcal{D}) = \{f^{-1}(x) : x \in \mathcal{D}\}$. We claim that $f^{-1}(\mathcal{D})$ is increasing set. Given a, b satisfying $a \le b$ and $a \in f^{-1}(\mathcal{D})$, we have $f(a) \in \mathcal{D}$ and $f(a) \le f(b)$. Since $\mathcal{D}$ is increasing, then $f(b) \in \mathcal{D}$ and $b \in f^{-1}(\mathcal{D})$. Hence, $f^{-1}(\mathcal{D})$ is increasing. If $p_1, p_2, \ldots, p_n$ is PRDS on $\mathcal{N}_0$, it follows that

$$\mathbb{P}((p_1^*, p_2^*, \ldots, p_n^*) \in \mathcal{D} \mid p_i^* = x) = \mathbb{P}((p_1, p_2, \ldots, p_n) \in f^{-1}(\mathcal{D}) \mid p_i = f^{-1}(x))$$

is increasing in x. Therefore, $\{p_1^*, p_1^*, \ldots, p_n^*\}$ is PRDS on $\mathcal{N}_0$. □

Proof. (proof of Proposition 3.5) Denote $\mathbf{p} = \{p_1, p_2, \ldots, p_n\}$. For any x, $\mathbb{P}(\mathbf{p} \in \mathcal{D} | p_i \le x) = \frac{\mathbb{P}(\mathbf{p}\in\mathcal{D}, p_i \le x)}{\mathbb{P}(p_i \le x)}$. For $y > x$, we have

$$\mathbb{P}\left(\mathbf{p} \in \mathcal{D} \mid p_i \le y\right) = \frac{\mathbb{P}\left(\mathbf{p} \in \mathcal{D}, p_i \le x\right) + \mathbb{P}\left(\mathbf{p} \in \mathcal{D}, p_i \in (x, y]\right)}{\mathbb{P}\left(p_i \le x\right) + \mathbb{P}\left(p_i \in (x, y]\right)}.$$

It suffices to show that

$$\frac{\mathbb{P}\left(\mathbf{p} \in \mathcal{D}, p_i \le x\right)}{\mathbb{P}\left(p_i \le x\right)} \le \frac{\mathbb{P}\left(\mathbf{p} \in \mathcal{D}, p_i \in (x, y]\right)}{\mathbb{P}\left(p_i \in (x, y]\right)}. \tag{21}$$

Denote F_i the cumulative distribution function of p_i, then

$$\begin{aligned}\mathbb{P}\left(\mathbf{p} \in D, p_i \le x\right) &= \int_0^x \mathbb{P}\left(\mathbf{p} \in \mathcal{D} \mid p_i = s\right) dF_i(s)\\ &\le \int_0^x \mathbb{P}\left(\mathbf{p} \in \mathcal{D} \mid p_i = x\right) dF_i(s)\\ &= \mathbb{P}\left(\mathbf{p} \in \mathcal{D} \mid p_i = x\right)\mathbb{P}\left(p_i \le x\right)\\ &\Longrightarrow \frac{\mathbb{P}\left(\mathbf{p} \in \mathcal{D}, p_i \le x\right)}{\mathbb{P}\left(p_i \le x\right)} \le \mathbb{P}\left(\mathbf{p} \in \mathcal{D} \mid p_i = x\right).\end{aligned} \tag{22}$$

Similarly, we have

$$\begin{aligned}\mathbb{P}\left(\mathbf{p}\in\mathcal{D}, p_i\in(x,y]\right) &= \int_x^y \mathbb{P}\left(\mathbf{p}\in D \mid p_i = s\right) dF_i(s)\\ &\geq \int_x^y \mathbb{P}\left(\mathbf{p}\in\mathcal{D} \mid p_i = x\right) dF_i(s)\\ &= \mathbb{P}\left(\mathbf{p}\in\mathcal{D} \mid p_i = x\right)\mathbb{P}\left(p_i\in(x,y]\right)\\ &\Longrightarrow \frac{\mathbb{P}\left(\mathbf{p}\in\mathcal{D}, p_i\in(x,y]\right)}{\mathbb{P}\left(p_i\in(x,y]\right)} \geq \mathbb{P}\left(\mathbf{p}\in\mathcal{D} \mid p_i = x\right),\end{aligned} \tag{23}$$

which completes the Proof. □

C.2. The proof of Lemma B.1

Proof. For simplicity, we denote $X = \sum_{i=1}^n X_I$ where $X_1, \ldots, X_n$ are i.i.d. and $X_i \sim \mathrm{Ber(p)}$ for $i \in [n]$. Besides, let $\mu = \mathbb{E}(X)np$. The moment-generating function (MGF) of X_i is

$$M_{X_i}(t) = \mathbb{E}(e^{tX_i}) = pe^t + 1 - p = 1 + p(e^t - 1) \leq e^{p(e^t-1)}.$$

Then, we have

$$M_X(t) = \prod_{i=1}^n M_{X_i}(t) \leq e^{np(e^t-1)}.$$

For any $t > 0, a = (1+\gamma)\mu$, by Markov's inequality, we have

$$\mathbb{P}(X \geq a) = \mathbb{P}(e^{tX} \geq e^{ta}) \leq \frac{\mathbb{E}(e^{tX})}{e^{ta}} = \frac{M_X(t)}{e^{ta}}.$$

It follows that

$$\begin{aligned}\mathbb{P}(X \geq (1+\gamma)\mu) &\leq \min_{t>0} \frac{e^{\mu(e^t-1)}}{e^{ta}} \leq \frac{e^{\mu(\log(1+\gamma)-1)}}{e^{(1+\gamma)\mu\log(1+\gamma)}}\\ &= \frac{e^{\mu\gamma}}{(1+\gamma)^{\mu(1+\gamma)}} \leq \exp(\frac{-\gamma^2\mu}{2+\gamma})\\ &\leq \exp(\frac{-\gamma^2\mu}{3}),\end{aligned}$$

where we use the fact that $(2+\gamma)\log(1+\gamma) > 2\gamma$. The proof of lower tail bound is analogous. Let $t = \log(1-\gamma)$, then we have

$$\mathbb{P}(X \leq (1-\gamma)\mu) \leq \frac{e^{-\mu\gamma}}{(1-\gamma)^{\mu(1-\gamma)}} \leq \frac{e^{-\mu\gamma}}{e^{-\mu\gamma+\frac{\mu\gamma^2}{2}}} = \exp(\frac{-\gamma^2\mu}{2}),$$

which completes teh proof. □

C.3. Proof of Lemma B.2

Proof. Since $\mathrm{FPR}(\cdot)$ is increasing, then for any $t \leq \eta(k\alpha)$, we have $\mathbb{E}(\mathrm{FPR}(t)) \leq \mathbb{E}(\mathrm{FPR}(\eta(k\alpha)))$. Hence, our gaol is to seek upper bound of $\mathbb{E}(\mathrm{FPR}(\eta(k\alpha)))$. We first consider $\mathbb{E}(\mathrm{FPR}(\eta(\alpha)))$. Recall that FPR at threshold $\eta(\alpha)$ can be expressed as

$$\mathrm{FPR}(\eta(\alpha)) = \frac{\sum_{i\in\mathcal{N}_1} \mathbf{1}\left(T_i < \eta(\alpha) - \mu\right)}{n^{1-\beta}}$$

To simplify notations, we denote $S_1 = \sum_{i\in\mathcal{N}_1} \mathbf{1}\left(T_i < \eta(\alpha) - \mu\right)$. It is easy to verify that

$$S_1 \sim B(n^{1-\beta}, 1 - \Psi(\eta(\alpha) - \mu))$$

and further we have

$$\mathbb{E}(\mathrm{FPR}(\eta(\alpha))) = \mathbb{E}\left(\frac{\sum_{i\in\mathcal{N}_1}\mathbf{1}\left(T_i < \eta(\alpha)-\mu\right)}{n^{1-\beta}}\right) = 1-\Psi(\eta(\alpha)-\mu).$$

If $r_{\min}(\alpha) < r < r_{\max}$, in terms of the definitions of $\eta(\alpha)$ and μ, we have

$$\begin{aligned}\mu-\eta(\alpha) &= (\lambda\log n)^{1/\lambda}\left(r^{1/\lambda} - r_{\min}^{1/\lambda}(\alpha)\right)\\ &= \left(\lambda d_\lambda(r, r_{\min}(\alpha))\log n\right)^{1/\lambda}.\end{aligned}$$

According to the assumption about $\Psi(\cdot)$, we have

$$\begin{aligned}1-\Psi(\eta(\alpha)-\mu) &\le \frac{1}{C_u}\exp\left(-d_\lambda(r, r_{\min}(\alpha))\log n\right)\\ &= \frac{n^{-d_\lambda(r, r_{\min}(\alpha))}}{C_l}.\end{aligned}$$

The definition of $r_{\min}(\alpha)$ implies that $r_{\min}(\alpha) < \beta+\zeta(\alpha)$. Now we consider the function $g(t) = d_\lambda(r_{\min}(\alpha)+t, r)$ and its derivative

$$g'(t) = \begin{cases} -\left(r_{\min}(\alpha)+t\right)^{\frac{1-\lambda}{\lambda}} d_\lambda(r_{\min}(\alpha)+t, r)^{\frac{\lambda-1}{\lambda}} & 0 < t \le r - r_{\min}(\alpha),\\ \left(r_{\min}(\alpha)+t\right)^{\frac{1-\lambda}{\lambda}} d_\lambda(r_{\min}(\alpha)+t, r)^{\frac{\lambda-1}{\lambda}} & t > r - r_{\min}(\alpha).\end{cases}$$

When $t \in (0, \beta+\zeta(\alpha)-r_{\min}(\alpha))$, the simple calculation gives

$$\left(r_{\min}(\alpha)+t\right)^{\frac{1-\lambda}{\lambda}} \le \beta^{\frac{1-\lambda}{\lambda}}$$

and

$$d_\lambda(r_{\min}(\alpha)+t, r)^{\frac{\lambda-1}{\lambda}} \le |r-r_{\min}(\alpha)-t|^{\frac{\lambda-1}{\lambda}} \le (r_{\max}-\beta)^{\frac{\lambda-1}{\lambda}}$$

It follows that

$$|g'(t)| \le \beta^{\frac{1-\lambda}{\lambda}}\cdot(r_{\max}-\beta)^{\frac{\lambda-1}{\lambda}} = \left(\frac{\beta}{r_{\max}-\beta}\right)^{\frac{1-\lambda}{\lambda}}$$

for $t \in (0, \beta+\zeta(\alpha)-r_{\min}(\alpha))$. According to Lagrange mean value theorem, we conclude

$$\begin{aligned}|d_\lambda(r_{\min}(\alpha), r) - d_\lambda(\beta+\zeta(\alpha), r)| &\le \left(\frac{\beta}{r_{\max}-\beta}\right)^{\frac{1-\lambda}{\lambda}}|r_{\min}(\alpha)-\beta-\zeta(\alpha)|\\ &= \left(\frac{\beta}{r_{\max}-\beta}\right)^{\frac{1-\lambda}{\lambda}}\frac{\log 12C_l}{\log n}.\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}n^{-d_\lambda(r, r_{\min}(\alpha))} &\le n^{-d_\lambda(r,\beta+\zeta(\alpha))}\cdot n^{\left(\frac{\beta}{r_{\max}-\beta}\right)^{\frac{1-\lambda}{\lambda}}\frac{\log 12C_l}{\log n}}\\ &= (12C_l)^{\left(\frac{\beta}{r_{\max}-\beta}\right)^{\frac{1-\lambda}{\lambda}}} n^{-d_\lambda(r,\beta+\zeta(\alpha))}.\end{aligned}$$

For $\alpha' = k\alpha$, we impose $n > (\min\{\alpha, k\alpha\}^{\frac{-1}{r_{\max}-\beta}})$, implying $\beta+\max\{\zeta(k\alpha), \zeta(\alpha)\} < r_{\max}$. If $r_{\min}(k\alpha) < r < r_{\max}$, the same reasons above shows

$$|d_\lambda(r_{\min}(k\alpha), r) - d_\lambda(\beta+\zeta(\alpha), r)| \le \left(\frac{\beta}{r_{\max}-\beta}\right)^{\frac{1-\lambda}{\lambda}}\frac{|\log 12C_l k|\cdot|\log k|}{\log n}.$$

It follows that

$$\begin{aligned}n^{-d_\lambda(r, r_{\min}(k\alpha))} &\le n^{-d_\lambda(r,\beta+\zeta(\alpha))}\cdot n^{\left(\frac{\beta}{r_{\max}-\beta}\right)^{\frac{1-\lambda}{\lambda}}\frac{|\log 12C_l k|\cdot|\log k|}{\log n}}\\ &= \rho^{\left(\frac{\beta}{r_{\max}-\beta}\right)^{\frac{1-\lambda}{\lambda}}}\cdot n^{-d_\lambda(r,\beta+\zeta(\alpha))},\end{aligned}$$

where

$$\rho = \begin{cases} \frac{1}{12C_l k^2} & 0 < k < \frac{1}{12C_l} \\ 12C_l & \frac{1}{12C_l} \leq k < 1 \\ 12C_l k^2 & k \geq 1, \end{cases}$$

which completes the proof. □

D. Some Known Results

Lemma D.1. ***(Bernstein's inequality)****. Let $X_1, X_2, \cdots, X_n$ be independent zero-mean random variables.Suppose $|X_i| \leq M$ almost surely, then for all positive t,*

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq t\right) \leq \exp\left(-\frac{\frac{1}{2}t^2}{\sum_{i=1}^{n}\mathbb{E}\left[X_i^2\right] + \frac{1}{3}Mt}\right). \tag{24}$$

E. Additional Experimental Results

Table 3. Additional experimental results (%) of practical perspective on **CIFAR-10** as ID data and **MNIST** as OOD data. We compare the performance between e-DR and g-BH based on the same trained score function. For each baseline method, we report results after using our framework in the next line. ↑ indicates larger values are better and vice versa.

ID	CIFAR-10		
OOD	MNIST		
Model	TPR↑	FPR↓	F1↑
ASH	95.16	48.59	33.39
ASH + g-BH	93.23	35.37	41.34
Cider	93.27	88.99	21.07
Cider + g-BH	91.77	74.06	27.45
GEN	95.5	37.83	39.67
GEN + g-BH	91.01	9.32	67.23
KLM	94.79	46.58	34.04
KLM + g-KNN	90.88	23.68	48.27
KNN	93.67	40.86	37.27
KNN + g-BH	89.43	18.69	58.24
LogitNorm	94.83	24.62	82.2
LogitNorm + g-BH	91.49	19.51	83.49
MSP	95.24	40.41	38.69
MSP + g-BH	90.53	23.98	52.75
RankFeat	95.15	95.54	20.29
RankFeat + g-BH	91.49	82.58	26.33

Table 4. Additional experimental results (%) of classical perspective on **CIFAR-10** as ID data and **MNIST** as OOD data. We compare the performance between e-DR and g-BH based on the same trained score function. For each baseline method, we report results after using our framework in the next line. ↑ indicates larger values are better and vice versa.

ID	CIFAR-10		
OOD	MNIST		
Model	FPR95↓	AUC↑	AUPR↑
KLM	47.55	85	36.57
KLM + g-KNN	38.45	86.02	36.98
KNN	36.96	94.26	83.73
KNN + g-BH	31.79	94.88	83.76
LogitNorm	24.83	94.91	96.02
LogitNorm + g-BH	15.64	95.58	96.87
ASH	48.08	83.16	40.64
ASH + g-BH	42.82	84.48	41.84
Cider	38.32	93.99	83.56
Cider + g-BH	29.37	94.95	84.46
GEN	35.49	92.83	76.04
GEN + g-BH	22.96	94.52	77.33
RankFeat	85.1	75.87	43.91
RankFeat + g-BH	83.49	75.98	43.98
MSP	47.29	90.63	75.57
MSP + g-BH	33.41	92.43	76.28